Downloaded From

sslecturenotes.in

one night study

DATA WAREHOUSING

Dt. 9.8.11

Startagic Information :-

The information needed for formulation & executing business startagic & objectives is meant for the entire organisation.

Such type of organisational reffered to as startagy information.

The operational system could not provide this startagy information to its upper user.

Business were compeled to term to new ways of getting startagic information.

The new way is called data warehousing.

Data warehousing is a new program that provides startagic information.

Bussiness organisation can achive considerable competative advantages by analysing their historical data.

Increasing Demand for Startagic Information :-

It is not required for only day-today operation of business & neither is required to produce an invoice setel a claim or post a withdrawal from a bank account.

Managers need information to make right decission at right time to formulate business startagy, establish goal, set objectives & monitor results.

Some ex: of business objectives :-

Retain the current customer of the business

Enhance the market shair by 15%.

Lunch new & better product in the market in the next year.

Improve product quality of top 5 semi-product.

Increase sales in the north-reasion.

For making decission about business objectives executive & managers need different kind of information :-

Analyse how key business factors affect each other
Monitors how business factors change with time.
compare the company performance with that of their competators.

Characterstics of Startagic Information :-

→ Integrated
→ Data Integrity
→ Accessable
→ Timely

Integrated :-
It must have an overall enterprise widely.

Data Integrity :- Data in all levels must be accurate.

Accessable :- Easily accessable by the users.

Timely :- The information must be available within stipulated time.

④ Information crises :-
Companies have an information crises not because of lack of sufficient data but due to unavailability of data ie raidly useful for startagy decission making.

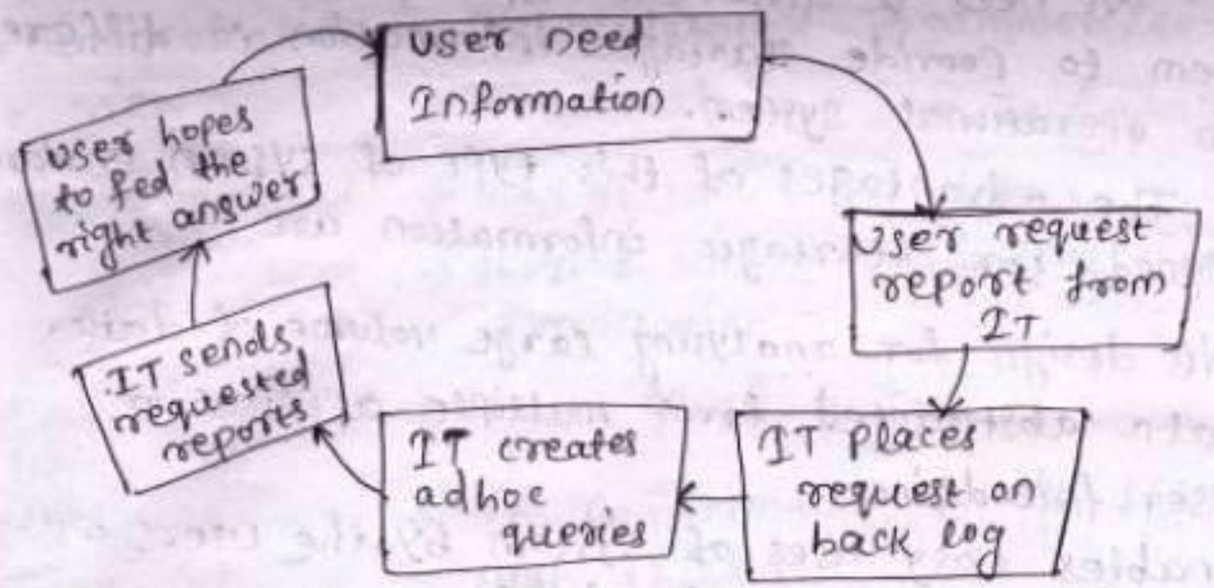
Information crices persist because of meny reason.

Data is coporation this side in various

Multiple platforms & divers structure for proper decission making on coporate startagy, we need information integrated from all systems.

Data needed for making startagic decission must be available in a format that enables execution & managers to analyse trains inorder to lead their companies in right direction.

The tremendous amount of operational data can't be readily used for reviewing data from different angles.



→ IT receives too many adoh queries for a variety of I/ps but with limitted resource, IT was not able to generate all the reports in the requested manner & within assigned time frame.

→ Request were not only numerous but also kept changing over time with user (large enough) wanting more reports subsequently to expand & understand earlier reports.

→ The users dependent upon IT to provide an Information as they could not access information directly in an interactive manner. As a result IT was unable to provide an enviornment for flexible analysis to the managers & executive for making startagic decission.

(3) Presence of Better Technology :-

→ The power of microprocessor has been doubling every two years.

→ The processing speed of the microprocessor has been increased while its cost has declined.

→ The price of digital storage has been coming down for several years.

→ There has been a constant increased in the n/w bandwidth & decrease in its cost to access that bandwidth.

→ Technology advances in data storage & reduction in storage cost fulfills the data storage needs for startagic decission system.

We need a different type of decision system to provide startagic information ie different from operational system.

The advantages of this type of system environment designed for startagic information are :-

(1) dB design for analysing large volume of data.
(2) Data abstracted from multiple application.
(3) User friendliness
(4) Enables easy uses of system by the users without asistance of IT proffectional.
(5) periodically updating of data ~~containent~~ content.
(6) content current as well as historical data.
(7) Ability for users to formulate & execute queries & get results online.

Difference Between operational & decission support system :-

Attributes	Operational	Decission support system
→ Data content	current values	→ summerized data
→ Data structure	optimized for transcation	→ optimized for complex queries
→ Access frequency	High	→ medium to low
→ Access TYPE	Read, update, delete	→ Read (only)
→ usage	Repetative	→ Random & adhoc queries
→ Response Time	subseconds	→ several seconds to minutes.
→ User Number	→ more users (large no)	→ Less users
→ Characterstics	Opertional Processing	→ Informational Processing
→ users	DBA, System Analysis, clerk	→ Executives, managers Business Executives
→ Function	Day today operation	→ Long term Information Requirement
→ dB design	ER based or application based	→ snow flake or star or object oriented

→ summerization	→ Highly detailed	→ summerized
→ view	→ Relational	→ Multidimensional
→ Record accessed	→ may be 10	→ Millions
→ unit of work	→ Short & simple transcation	→ complex query
→ dB size	→ 100 MB to GB	→ 100GB to TB & 240
→ priority	→ High Performance High ability	→ High flexibility
→ Indexes	→ Few	→ More
→ Join	→ many	→ some
→ duplicate data	→ Normilised dB	→ Denormalisation (make a single table)

In the data warehousing model operational system are not accessed directly to perform information processing.

De. 12 & 11

Data ware House Defined :-

Data ware house were develop to meet the growing demand for information analysis that couldn't be meet by operational system for range of reasons.

The processing load of reporting affecting the response time of operational data

→ The dB design of opertional system were not optimized for information analysis & startagic decission making.

→ Generally all big orgⁿ had a no. of operating system so enterprise wide reporting would not be supported for a single system.

The data ware house is thus an informational enviornment which does the following:-

Provides an integrated view of an enterprise.

→ facilitated the enterprise current as well as historical data readily available for making startagic information.

→ It makes decession making possible without hindering / abstracting operational system.

→ It makes the orgⁿ inforⁿ consistant & easily available

→ It provides a flexible & interactive source of startagic decission

What can a data ware House do?

① Immediat Inforⁿ delivery :-

② Integration of data from within & outside the orgⁿ:-

Data warehouse combine data from multiple sources. The data is collected from different departments like sales, marketing, finance & accounting etc.

Provides an insight into the feture:-

Data warehouse store large ammount of historical inforⁿ that enables the decission maker to analyse the previling trends in the market & produce good ~~action~~ according to the customer demand.

Enables User to look at the same data in diff. ways

→ A data warehouse provides its user with tool for providing, analysing & manipulating in diff. ways.

→ It facilitate the users to drill down/detailed down into the detailed data with the click of mouse that could have otherwise taken a few days with the traditional approach.

Provides freedom from the dependancy on IT :-

With data warehouse managers needs an ~~ad~~ adhoc report he can himself form it without the asistance of any computer.

What Data Ware House Can't be ?

① → It acts as an information ~~reposiotary~~ repositary that collects & reports data that already exist.

→ It can't create an additional data on its own.

② If an orgⁿ has dirty data in the source system the data ware house will not be able to give correct results unless & unless the data is first clean.

Data Ware House an enviornment or a product :-

It is a computing enviornment where users can find startagic information, an enviroment where the users are put directly in touch with the data they need to make better decision.

The characteristics of new computing enviornment called the data ware house are :-

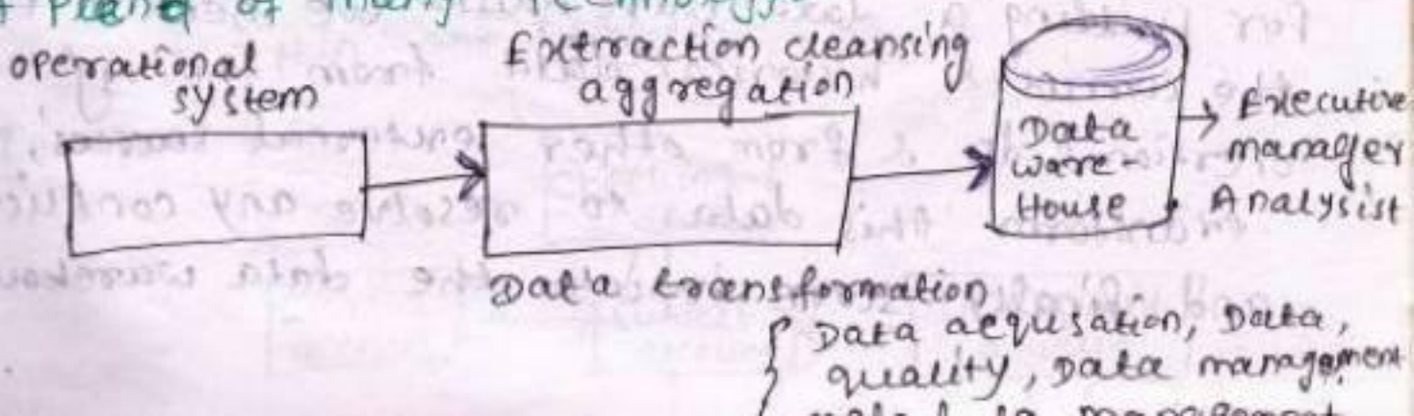
① An ideal enviornment for data analysis & decission making.

② Flexible & interactive

③ User friendly

(4) Enables the user to discover answers to complex queries.

A plant of many Technology :-



warehouse is to bring infor^n from desparate sources & put the infor^n into a format ie conducive for business decission.

Data warehousing requires both business & technical enterprise. ~~& involves in following~~

Basic operation of Data Ware House :-

Extract data from operational system

Include relevant data from outside sources like magagines, journals, reports of other org^n in the same industry,

Remove the inconsistancies & transform & clean the data.

It stored the data in such a way that, it is easy access for decission making.

Data warehouse Building Blocks :-

D1.16.8.11

Father of data warehouse is Bill inmon

The data warehouse is an information delivery system.

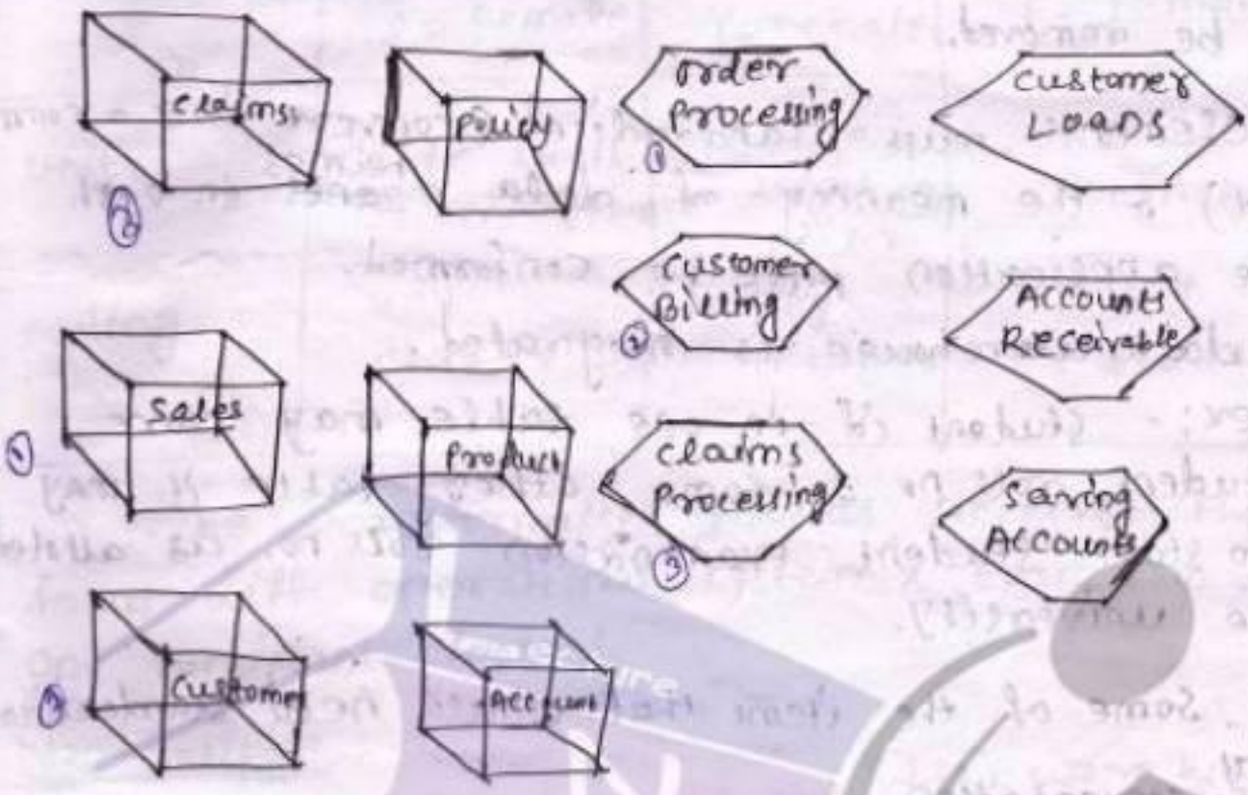
In this system the data warehouse integrate & transform enterprise data into information suitable for startagic decission making.

In other words it is meant to provide information to its user so that they can make use of this information to make their business better.

For building a data warehouse we need to extract the current & historical data from the org^n, operational DB & from other external sources, transform this data to resolve any conflicts and finally store it into the data warehouse.

Def:
Data warehouse is a subject oriented, integrated, non-volatile & time varient collection of data in support of management decission. (By Bill inmon)

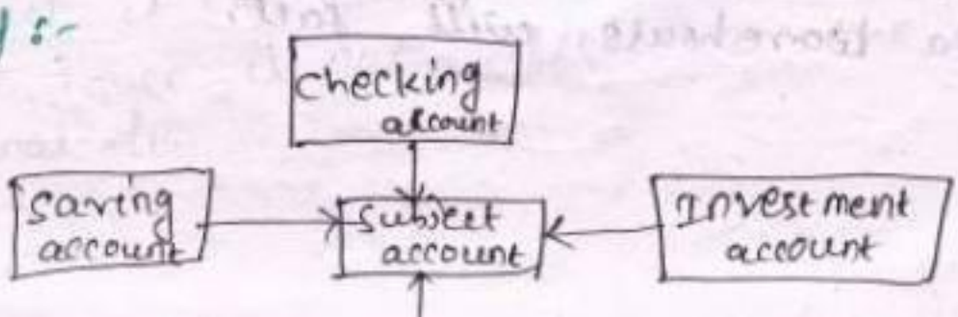
① Subject-oriented data :-



In data warehouse data is stored by subject not by application. Business subjects differ from enterprise to enterprise.
For a manufacturing company sales, inventory are critical business subjects.
For a ~~written~~ retail store sales at a check out counter is a critical subject.

The data warehouse gathers all the data in one place. The structure of the data is such that all the data for a perticular subject is a content within one table.

② Integrated :-



from multiple sources.

operational data is the basis of the data ware house.

Before the data from various distinct sources can be stored in data ware house. All such inconsistan must be removed.

Data elements must standardised (convert into a common format) & the meaning of data ~~meants~~ names on each source application must be conformed.

The data ware house is integrated.

for ex:- student id in one table may store the student roll no & some other table it may use to store student examination roll no. as alloted by the university.

Some of the items that would need standardization

→ ~~Mai~~ many conventation
→ codes
→ Data attributes
→ Measurement

The integration process consists of two task:-

Data cleansing

Data Transformation

Data cleanshing :-

It is the process of removing errors from i/p data & it is the part of integration process.

If the cleanshing process defaulting the decission maker will not trust the data & purpose of having data warehouse will fail.

Data Transformations :-

Description	customer name	cname	custome Name
Encoding	sex 1 = male 2 = female	sex α = male γ = female	sex m = male f = female
units	cable length in centimeter	cable length in meter	cable length in inches
coding	key char (10)	key integer	key pic 99999

The transformation process receives the i/p data from diff. operational system & transform them into one consistant format.

Non-volitile :-

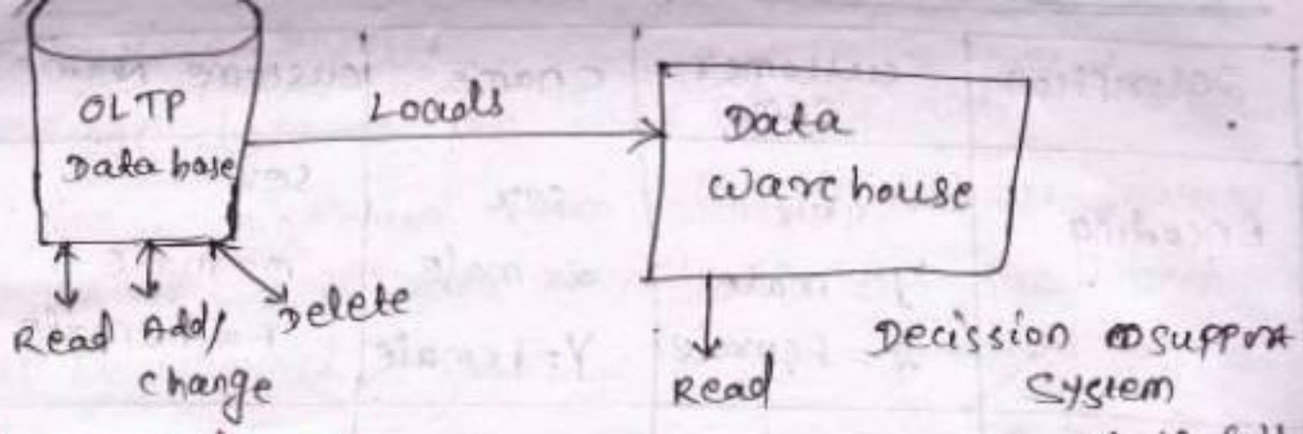
For storing the data in the data ware house we extract them from multiple operational system & from external sources.

These data after being extracted has to be transformed, integrated then store in a data warehouse.

Business transcation do not update data in data ware house. They update only the operational system.

In real time with each transcation that occurs the user may add, change (delete) the data from operational system but they do not update the data in data ware house.

Thus it can said that the data in data ware house is not as volitile as the data in an operational dB.



Dt. 18.8.11

Time varient :-

For an operational system the stored data contain the current values becoz these system support day-to-day transcation.

In an order entry system the status of an order is the current status of the order.

The data in data warehouse is being for startagic analysis & decision making.

A data warehouse becoz of the nature of its varient purpose has to find historical as well as of appart from current value.

The key benifit of the time varient nature of the data in a d.w is that data allows for study & analysis of the past correlate the information to the operation present cenario & finally enables forecasting

Therefore every record stores in dw has an associated time element which specified the Date-Mon-year when the sale transcation took place.

Data Granularity :-

It refers to the level of detail or level of summerization of data in DW. More detail data means lower level of granularity

similarity less detail indicates higher level of granularity.

The lower level of granularity, the larger is the amount of data stored in the warehouse.

The higher level of Granularity, the lesser is the level of detail for which queries can be answered.

Detailed data allows the user for the DW to view same data from diff. angles in different prospectives.

Advantages & Disadvantages:

Lower level Granularity

→ The amount of computing is needed to process a large volume of data in order to access the data.

→ The volume of data & storage space is another factor.

Higher level of Granularity:

→ There is a corresponding loss in the ability to answer queries using the data as a high level of granularity limits no. of questions that data can handle.

Dual level Granularity:-

When the DW is designed to handel dual level of granularity the DW will contain two types of data:-

→ summerized data

→ detail data

The detail data can include details of even 10 years of history

Data mart :- A data mart is a decision support system that stores a no. of subjects area based on the needs of the user in the department.

A data mart can thus be the thought of as a subset of enterprise wide DW.

Ex:- The finance dept. of orgⁿ has their own DM, marketing dept. has their separate DM & so on.

Every individual department owns the h/w, s/w, data & programs that are needed for that DM. DMs are usually implemented on low cost departmental servers. Each department wants customization of data that will be stored in their DM.

There are two types of DM:-

① Dependent DM
② Independent DM

Used in UNIX/ LINUX or window based. The implementation cycle of DM measured in weeks rather than months / years

Dependent DM

A dependent DM is one that takes the data from the DW.

Independent DM

It is one that depends upon for legacy application environment for its data source.

Dependent DM :-

→ All dependent DM are fed by the same source which is the DW.

→ Each DM are easier to use because they have only inforⁿ that the specific user do within the particular department needs.

→ The dependent DM are created with the subset of information in the DW.

Data WareHouse	Data Mart
→ The data mart is designed to meet the requirement of users of a perticular dept.	
→ on the otherhand the DW is designed to meet the collective requirement of the users of the entire orgⁿ.	
→ While dm contains fabricated / summerized data, the DW contains most detail data,	
→ While DW are designed to contain robust amount of history	→ DM contain only modest amount of history.
	→ Subject area found in the DM may or may not be related to the subject area found in the DW.
	→ The types of queries that are executed in DM are diff. from those are executed in the DW.
The kinds of user we use the DM are diff. ~~we use~~ from those DW.	
→ user is business excutives, high level managers, Board of directors etc.	→ user is the departmental manager.
	Dt. 19.8.11
→ corporate wide, enterprise wise	→ Departmental scope.
→ Lower level of granularity	→ High level of granularity
→ Normalised structure & robust amount of Historical data	→ modest / less amount of Historical data.

→ Technology optimal ~~for~~ for holding & managing massive amount of data.	→ Technology optimal for data access & analysis.
→ Lightly indexed	→ Highly indexed
→ combination of more than one business process.	→ A single business process
→ structure for corporate view of data	structure to suit the dept.
→ Takes month to year for implementation	Takes ~~month~~ weeks for implementation
→ DW size varies from 100GB to few ~~ten~~ TB	→ Less than 100GB.
→ Flexible query & analysis.	→ Restrictive query & analysis

Advantages of DM :

→ cost is low

→ The lead time to ~~increase~~ implementation is short often under 90 days.

→ They are controlled locaty rather than centrally.

→ Rapid Response.

Disadvantages:

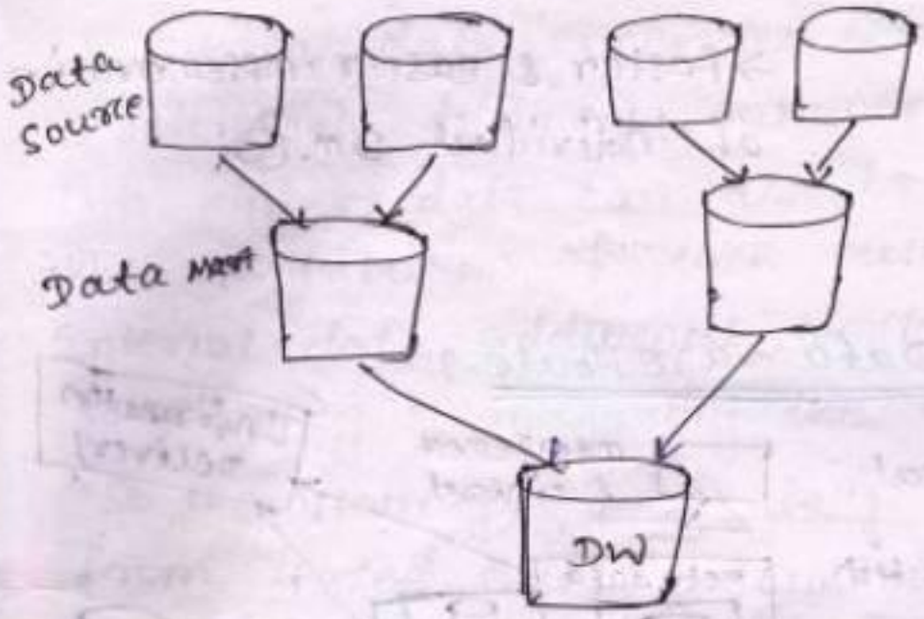
Performance degradation occurs as the size of DM increases.

Administration of DM becomes difficult.

Building Data Mart :

There are two main approach for building DM:-

The ~~to~~ DW Project ~~p~~ looks at the larger picture of the orgⁿ & builds a huge DW first, ~~&~~ that will feed the individual DM.



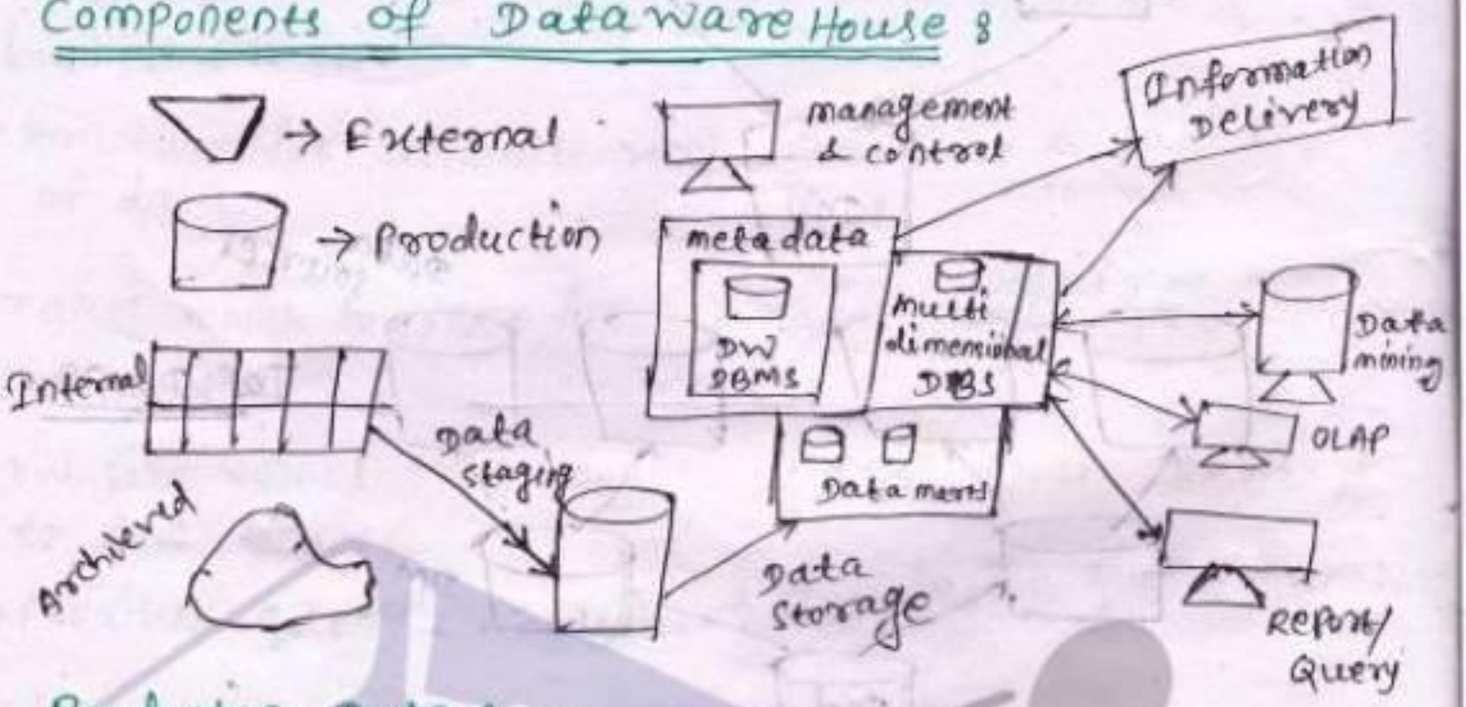
The DW project team captures to the requirement of individual dept & build DM first that will feed data to the corporate wide DW.

Top Down	Bottom UP
① Data is extracted from operational system transform, clean & integrated finally stored in DW.	① Data is extracted from operational system transform, clean & integrated, finally stored in DM.
② Presents an enterprise view of data.	Presents data only at the department level.
→ single central storage of data.	Inherently incremented as the key considered schedule important DM first.
→ It takes longer time to build the overall	→ Implementation of

~~faster~~ → faster & easier implementation of individual DM.

~~External~~

Components of DataWare House :



Production Data :

→ It is the main source of data in DW.
→ This is the data ie collected from various operational system of the org.
→ In operational system same peice of data is scattered after several application like employee table may be stored in acc. dept, HR dept & payroll dept. For data to be stored in DW it must be collected from operational system, this is called Production data.

Before Production data is stored in DW all the related data ~~has~~ must to be collected & standardized into ~~simple~~ single format.

Internal Data :

Internal data is taken from Private files & put ~~in~~ into

Features :

Internal data compersies of various personal sprade sheet, document, customer profiles etc. This kind of data can also be useful in a DW When contribution of each customer is significant. Internal data additional complexity to the process of transforming & integrating the data

So the startagy have to design to collect the data from spread sheet, textual document & tie them into departmental DB.

Archieved Data :-

An opeartionalsystem use to run day today transcation & for we need if only current information in dB.

So every operational system periodically the old data is taken in an archived file.

Basically what infor is archived & when it will be archive will vary one orgⁿ to another orgⁿ.

(1) Recent data is archived to separte archival dB that may still be online.

(2) The older data is archived to flat file or disk storage.

The older data is archived to take & flat file or even kept upside.

Dt. 23.8.11

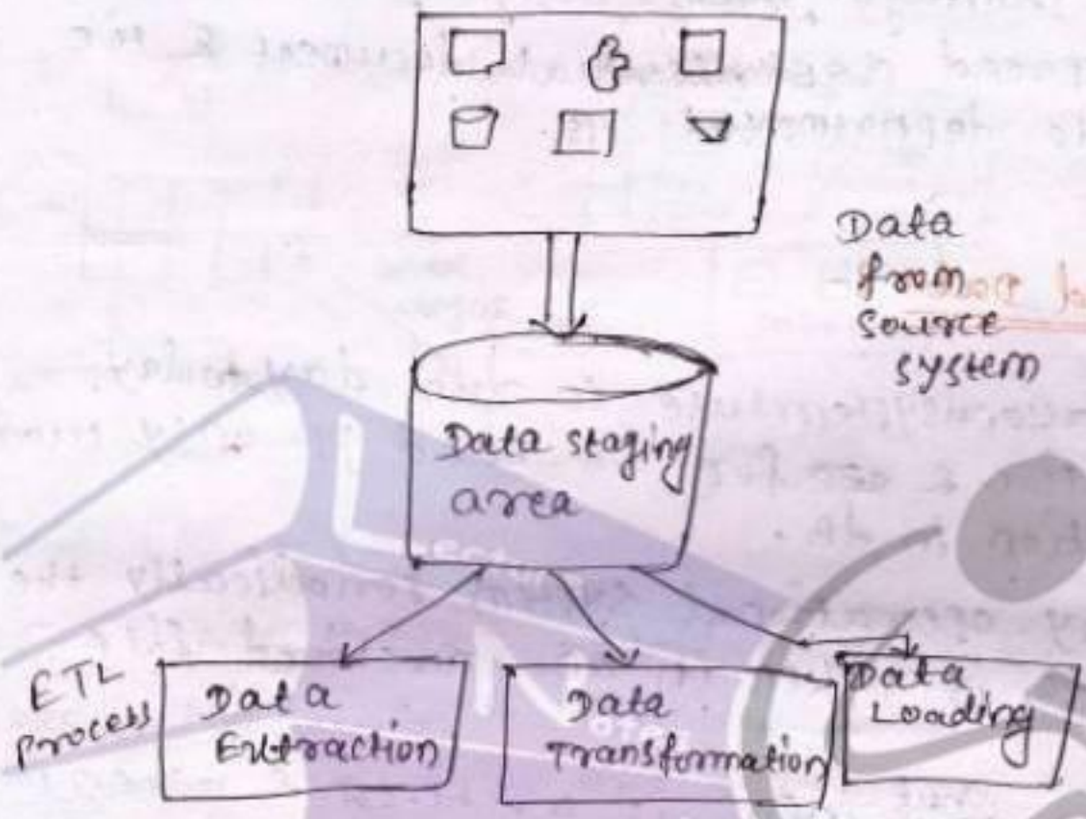
External Data :-

→ Most excutives depend on data from external sources for a high percentage of the information they used.

→ The purpose served by external data

→ The external data mainly collected from business magagines, Industry news letter, Tech reports, Reports generate by the consult expesially for the enterprise, sells & marketing analysis report etc.

Data staging :-



Data staging area is the place where all the extracted data is temporarily stored & prepared for loading into DW.

It is compared with the construction area where the extracted files are examined, business rules are review, the data transformation func^n are performed, data is sorted & merged, inconsistancy are resolved & the data is cleansed. Finally when the data is processed & prepair for an enterprise wide DW, it temporarily reside in the staging area.

Repositary waiting to be loaded into DW.

ETL Process

(i) Data Extraction :-

→ This function has to deal with numerous data sources source data may be from diff. source, m/c, indiverse data formats.

→ Part of the source data may be in relational dB may be hierarchical dB system. Many data source may still be in flat files.

→ Data extraction may become quite complex. Diff. tools are available in the market for extraction

(ii) Data Transformation :-

→ A no. of Individual task as part of data transformation, clean the data extracted from each source.

→ Cleaning may just be correction of mispelling or may include resolution of conflicts bet^n state code & zip code in the source data or eliminating duplicates.

→ Standardization of data elements form a large part of DT.

(iii) Data Loading :-

→ When we complete the design & construction of the DW & go leave for the 1st time, we do the initial loading of the data into DW storage.

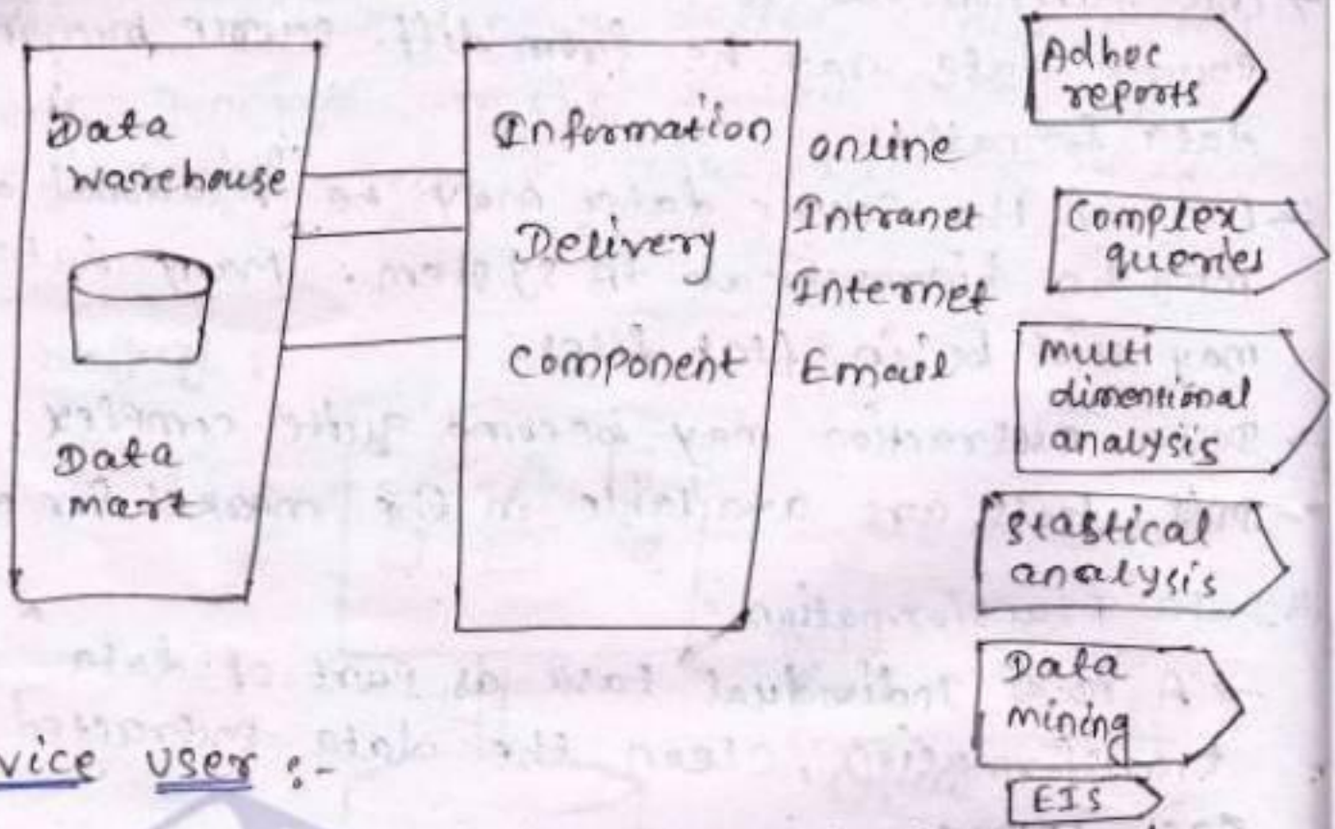
→ The initial loading moves large volume of data using substential amount of time.

Data Storage component :,

The data storage for the DW is a separate repositary.

In data repositary of a DW, we need to keep large volume of historical data for analysis. When the analysist use the data in the DW for analysis they need to know that the data

Information Delivery component :-



Novice User :-

No training & therefore needs prefabricated reports & present queries.

Casual Users :-

Needs Information once in a while not regularly.

Business Analysist :-

Looks for ability to do complex analysis using the information in DW.

Adhoc reports are predefined reports primarily need for novice & casual user.

Complex queries :-

complex queries, multidimensional analysis & statistical analysis to the need of business analysis.

Executing Infor^n system (EIS) is need for senior executive & high level managers.

Data mining :-

Data mining application ~~knowledge & know~~ are Knowledge discovery system where mining algorithm helps us discover traints & patterns from the uses of our data.

Information Delivery components :- Dt. 25.8.11

Online queries & Reports :-

The user will enter their request online & receives their results online.

We may set up delivery of schedule reports through e-mail or we may make ~~of~~ mail adquate use of our org^ intranet for information delivery.

Management & control component :-

The mang. & control component co-ordinates the services & activities within the DW. It controls

(1) The data transformation & data transfer into the DW storage.

(2) It works with DBMS & enables data to be properly stored in the reprositary.

(3) It monitors the movement of data into the storage area & from their into the DW storage itself.

(4) It intracts with the metadata component to perform the management & control function.

Meta Data Component (Data about Data) :-

Types of meta Data :-

→ Operational Meta Data

→ Extraction & Transformation

→ End-user Meta Data

Operational Meta Data :

→ The data for the DW comes from several operational system of the enterprise.

→ selecting data from the source system for DW we split the records, combine parts of records from diff. source files & deal with multiple coding scheme & field length.

able to try back to the original source data

Extraction & Transformation Meta Data :-

→ It contain data about the extraction of data from source system, extraction frequencies, extraction method & business rule for the data extraction

→ This catagory of MD contains infr^n about all the data transformation that take place in the Data staging area.

End-User Meta Data :-

→ It enables the end user to find information from DW, but end-user meta data allows the end-users to use their own business terminology & look for information in those ways in which they normally think of their business.

TRENDS IN DATAWAREHOUSE

- What is the current senario & state of the market.
- What business ~~are~~ have adopted Data warehousing
- Whata are the technology Advances

continued Growth in DATA Warehousing :-

→ Now every company that has a DW is realising enmorous benifits that get translated into the results at the bottom line.

→ Many of the companies now in cooperating wave-based Tech. are enhancing the potential form & easier delivery of ~~vatial~~ vaital information.

Data Warehousing Main Streams :

(1) competation
(2) Govt. deregulation
(3) Need to change in internal process
(4) compulsory for customized market

DW Expanssion :-

Earlier DW concentrated on keeping summery data for high level analysis.
Now companies have the ability to capture, cleanish, mentain & use the ~~vast~~ vast amount of data generated by their business transcation.

vendor solution & Products :-

→ How to classify vendors & products & thereby make the sence of the market.
It is based to separate market broadly into 2-catagories :-

① → The first group consists of DW vendors & products providing the needs of corporate data warehouses in which all enterprise data is integrated & transform.

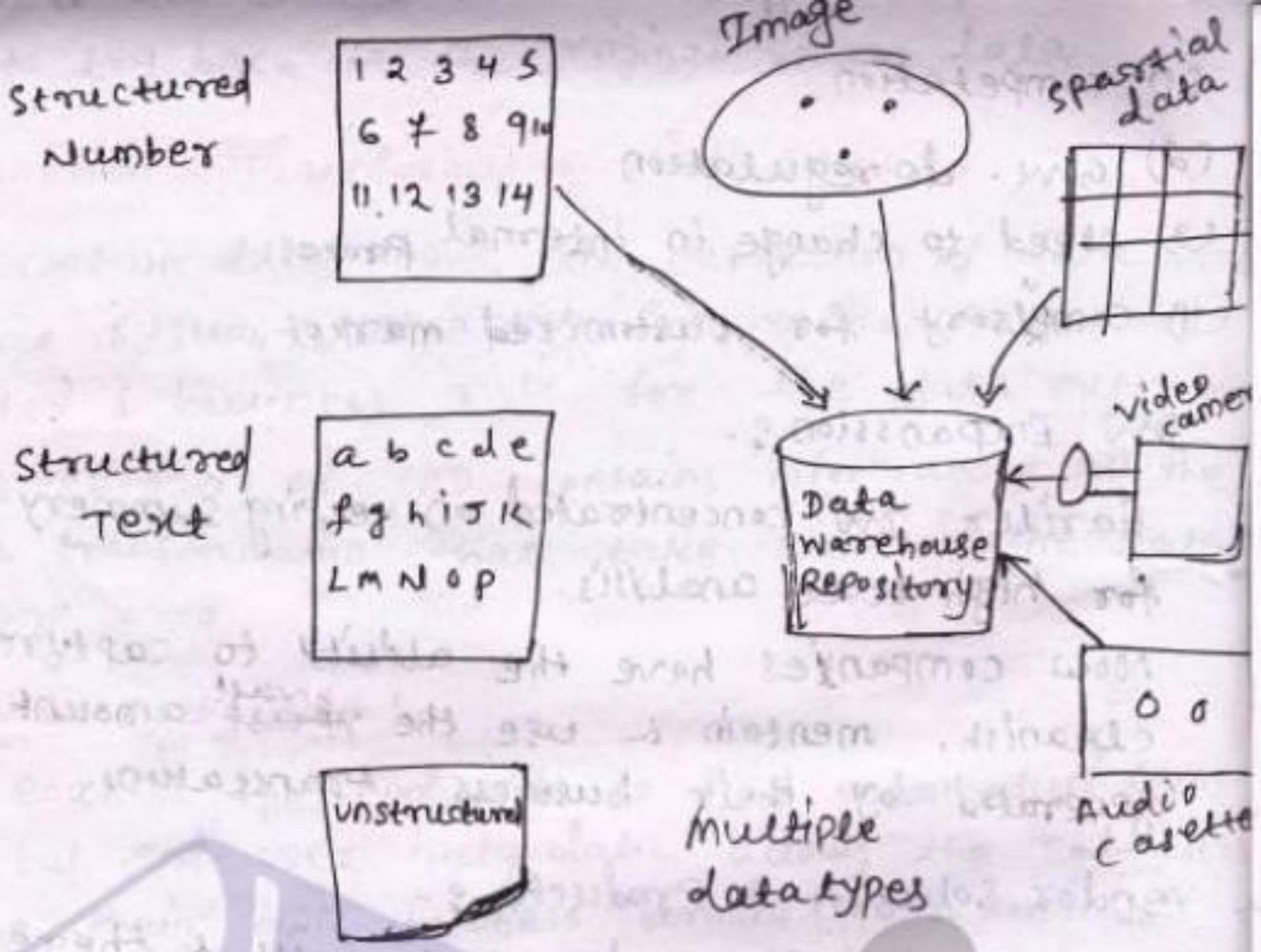
→ This segment has been reffered to as the market for stratagic DW.

② The second segment is consisting of departmental data mart, fragmented db marketing system & wide range of decission support system.

Significant Trends :-

Multiple Data Types :

Traditional companies included structured data mostly numery in their DW. The decession support system devided into two catagories :-



Data warehousing dealed with structured Data
Knowledge management involve in Unstructured Data.

Most marketing data consists of structured data in form of numeric values. Marketing data also contains unstructure data in terms of images

Adding Unstructure Data:-

Some vendors are addressing for resolution Unstructured data especially Text, image by ~~treating~~ printing such multi ~~numeric~~ media data & just another data type.

These are defined as a part of relational data & stored as binary large operater (BLO) upto 2GB.

Spartial Data (Geographical Data) :-

Dt. 26.8.11

What is the average income of people ~~leaving~~ living in the neighbourhood of the store.

Ex:- Address, street, city, country, state are example of spartial data.

Data visualisation :-

It helps the user to interprete query results quickly & easily.

(1) More Chart type

→ Most data visualisation are in the form of some standard chart type. The numerical results are converted into a pie chart, scattered plot.

Interactive visualisation :-

visulisation are no longer static, dynamic chart type are themselve user interfaces. The user can review a result chart, manipulate it & see the new views online.

visualisation Types :-

visualisation s/w supports a large array of chart types. The business user ~~sup~~ demand pie chart & bar chart, Technical & scientific user needs scattered plot, analysist looking for spartial data needs MAP, executives & managers who need to monitor performance like digital ~~tel~~ dash board. to allow ~~to~~ them to visualise the metrics as speedo meter, thermo meter or traffic light.

Advance visualisation Technic :-

The most remarkable advance in visualisation technic is the transition from static chart to dynamic interactive Presentation.

A user can roated a chart or dynamical change the chart type to get a clear view of results with complex visualisation types like scattered plot, a user can select data point with a mouse & move the point arround to clarify the view.

Drill Down

The visualisation first presents the results at the summery level.

Advance interaction :-

→ The user simply double click a part of visualisation & then drag & drop representation of data entities. or the user simply right click & choose the option from menu.

→ visual query is the most advance of user interaction feature.

Parallel processing :-

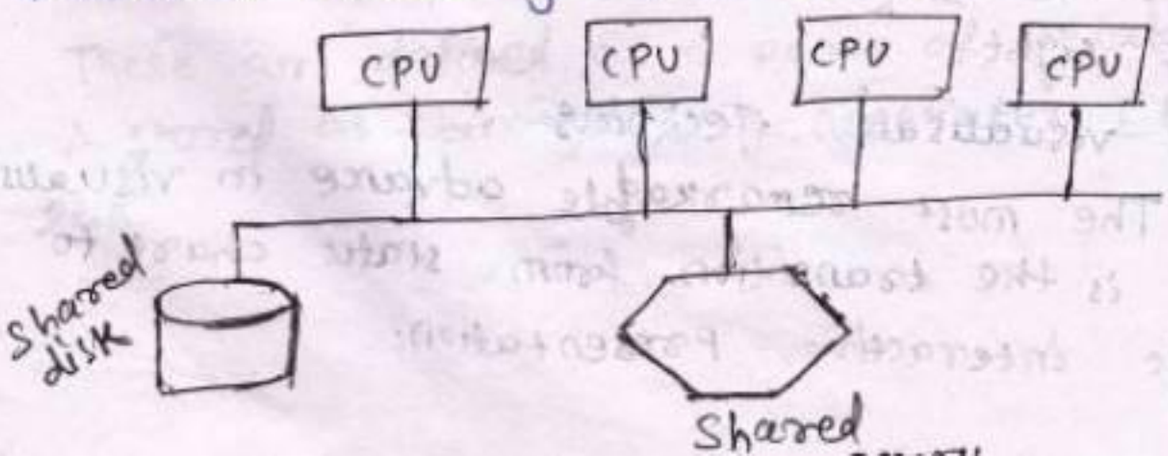
→ The user will constantly be executing complex queries to perform all type of analysis.

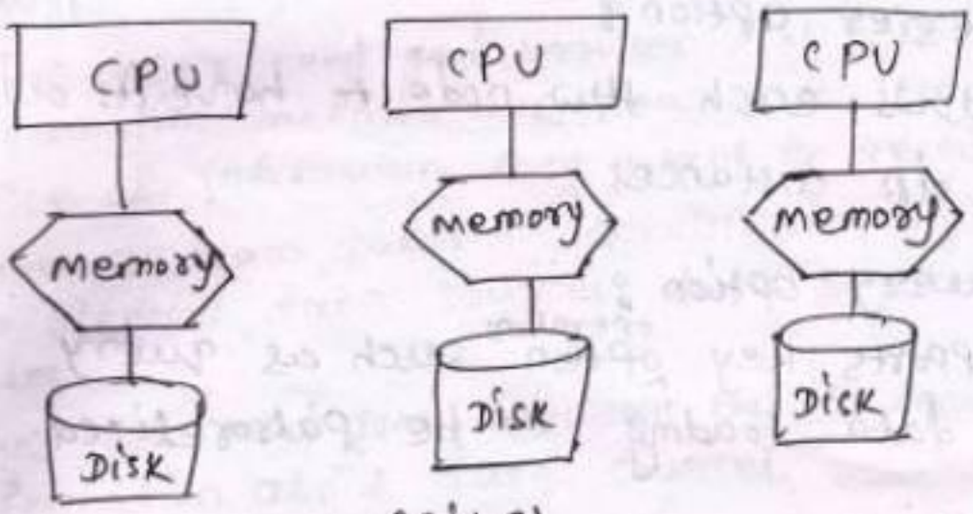
→ Each query would need to read large volume of data to produce result.

→ A very effective way to do to acomplinish this is to use parallel processing.

→ A task is divided into smaller units & though smaller units are executed concurrently.

Parallel Processing H/W options :-





(Fig. 2)

In a parallel processing environment we will find those characterstics, multiple CPU, memory modules, one or more server nodes, & high speed comm^n links between interconnected nodes.

Parallel Processing S/W Implementation :-

H/W alone would be worthless, if the O.S & DB S/W can't make use of parallel feature of the H/W. We will have to ensure that the S/W can allocate units of larger task to the H/W components appropriately.

Parallel Processing S/W must be capable of performing the following step:-

(1) Analysing a larger task to identify independent units that can be executed in parallel.

(2) Identifing which of the smaller units must be executed one after another.

(3) Executing the independent units in parallel & these units in proper sequence.

(4) collecting, collatating (combing) & consolidating (minimizing) the results by smaller units.

DB vendors provides two option for parallel processing:-

(i) parallel server option

(ii) parallel query option

It allows each H/W node to have its own separate DB instances.

Parallel query option :

It supports key ~~option~~ operations such as query processing, data loading to be parallelized.

Advantages :

Dt. 30.8.11

Query Tools :-

→ The success of DW depends on query tools. The following functions for which vendors have greatly enhanced their query tools.

(1) Flexible Presentation :-
Easy to use & able to present results online & reports in many diff. format.

(2) Crossing Subject Area :-
Able to cross-over from one subject data mart to another automatically.

(3) Multiple Heterogeneous Sources :-
Able of accessing heterogeneous data sources on different platform.

(4) Overcoming SQL limitation :-
Provide SQL extension to handle request that can't be usually done through standard SQL.

BROWSER TOOL :-

The user need good browser tool to browse through information metadata & search to locate the specific piece of information they want to receive.

→ when IT team wants to develop company DW they need to identify data sources, data structure & business rules.

→ They also need good browser tool to browse through information about data sources.

Data Fusion :-

Data fusion is a technology dealing with the merging of data from disparate sources.

Multi-diamensional analysis :-

Provision of multi-diamensional analysis to users simply means that they will be able to analyse business measurement in many different ways.

Syndicated Data :-

→ The value of data content is derived not only from the int. opertational system but from suitable ext. data as well.

→ Some of the earlier DW were incorporating syndicated data from such traditional supplier to enrich the data content.

DW & ERP (Enterprise Resource Planning) :-

→ Most of the companies are pluged by numerous disparate application that can't present a single unified view of co-operate information.

→ The data are retrived from various system to produce meanigful & correct information is extremely difficult & at large co-oporate almost impossible.

DW & Knowledge Management :-

→ Knowledge management takes enpowerment to high level.

→ It completes the process of providing users with knowledge to use the right information, at right time & at right place.

→ A knowledge management system must store all such knowledge in a knowledge repositary sometimes called knowledge warehouse.

→ If a DW contains structured information, the knowledge warehouse holds unstructured information.

Emergence on standards :-

→ Now, in industry such as DW there is no scarcity of vendors & products.

→ The multiple vendor products have to co-operate & work together in DW.

→ Standards are critical in 2 areas :-

① Metadata Exchange

② OLAP function

(1) Metadata Exchange :-

two separate bodies are working on standards for metadata :-

(i) Metadata coalition

(ii) object management group

Metadata coalition :-

→ Formed as a constorium of vendors & interested parties in oct. 1995 to launch a metadata standard. initiative the coalition has been working on standard known as open information model.

object management Group :-

→ Another group of vendors including oracle, hp & sun sought for metadata standard through the object management group, a large established forum dealing with wider array of standard in object technology.

② OLAP Function :-

→ The OLAP council was established in Jan 1995 as a customer Advocacy group to serve as an industry guide over the years.

→ The council was worked on OLAP standard for multi-diamensional application program interface & has come-up with revisions.

2.8 Web Enabled DW :-

→ Now no business can compete or survive without web
→ We have to transform occure DW into a web-enabled DW.

The warehouse to the web :-

→ Now-a-days, web is the primary information delivery mechanism.
→ This new delivery method will radically change the way users will retrive, analyze & share information from the DW.
→ The components of information delivery will be different.
→ The internet interface will include browser, search engine, push technology, home page, information content, hypertext links, & downloaded Java or Active x applets.
→ The warehouse to the web includes key requirements i.e. self-service data access, interactive analysis, high availability & performance, zero-administration client (thin client technology such as Java applets), tight security, & unified metadata.

The web to the warehouse :-

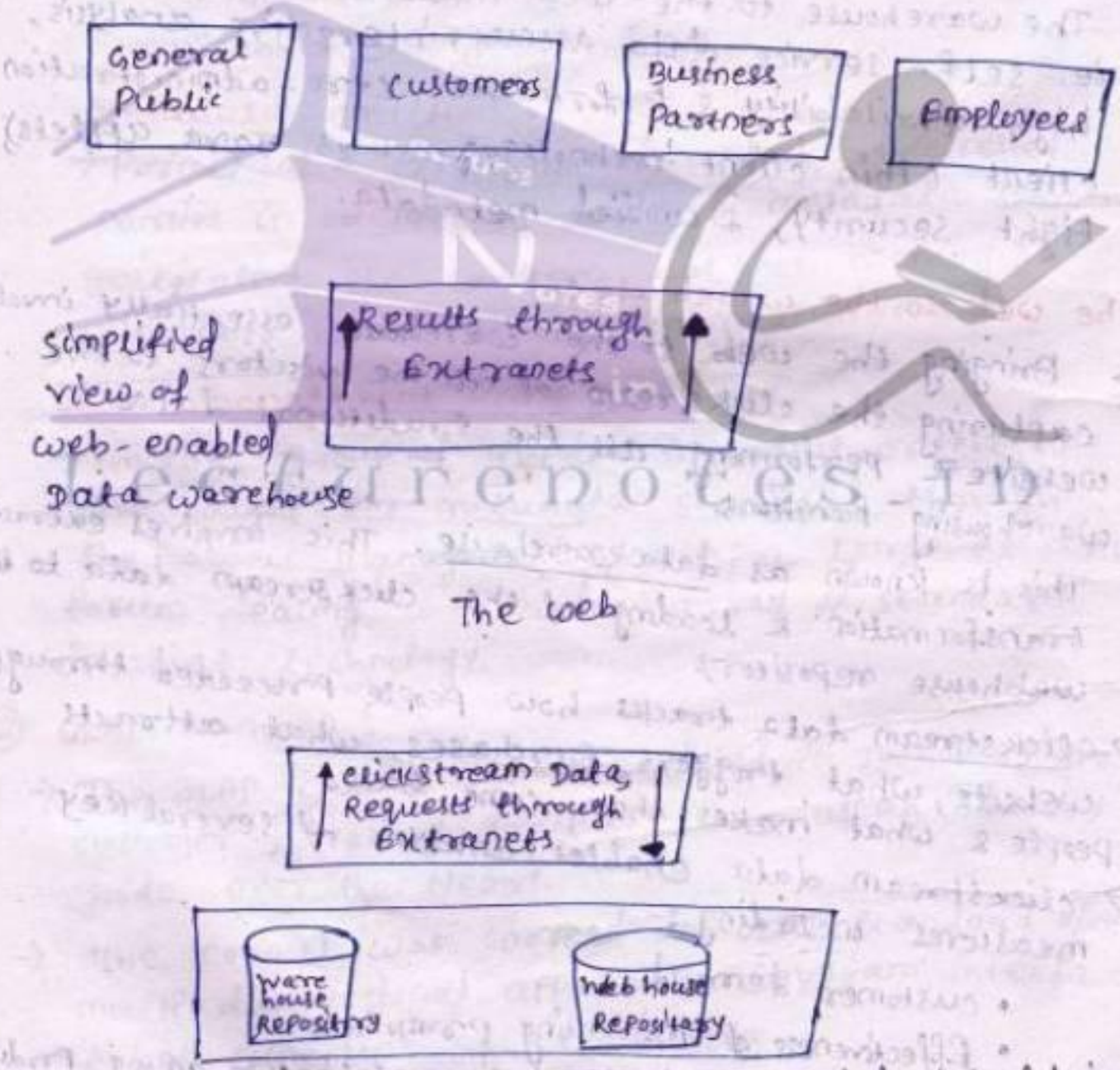
→ Bringing the web to the warehouse essentially involve capturing the clickstream of all the visitors to the website & performing all the traditional data warehousing functions.
→ This is known as <u>data warehouse</u>. This involves extracting transformation & loading of the click stream data to the webhouse repository.
→ <u>clickstream</u> data tracks how people proceeded through website, what triggers purchases, what attracts people & what makes them come back.
→ clickstream data enables analysis of several key measures including :-

- customer demand
- Effectiveness of marketing promotions
- Effectiveness of affiliate relationship among Products
- demographic data collection

important tool for identifing, prioritizing & retaining e-commerce customers. The webhouse can produce the following useful information:

- Site statistics
- visitor conversions
- Ad metrics
- Referring partner links
- site navigation resulting in orders
- site navigation at " " "
- Pages that are session killers
- Rel^ship bet^n customer profiles & page activities
- Best customer & worst customer analysis

The web-enabled configuration :-



→ web-enabling the DW means using the web for inform^n delivery & integrating the clickstream data from the corporate website for analysis.

→ The convergence of data warehousing & the web tech.

DW & OLAP rules are based on multi-dimensional data model. This model views data in the form of data queue.

Data Queue :- A data queue allows data to be modelled in multiple dimension. It is defined by dimensions & facts.

Dimension :- These are the prospectives or entities w.r.t which an orgⁿ wants to keep the record.

Eg :- Sales data warehouse, in order to keep records of stores sales w.r.t dimension, time, item, branch & locⁿ.

These diamension allow the store the keep track of things like monthly sales of item, branches & locⁿ at which the items are sold. Each dimension may have a table associated with it called dimension table, which further describe the dimension.

Facts :-

→ A multi-dimensional data is typically organized around a central theme like sales. This thing is represented by fact table.

→ Facts are numerical measures

→ The fact table contains name of the facts or measures as well as keys to each of the related dimension table.

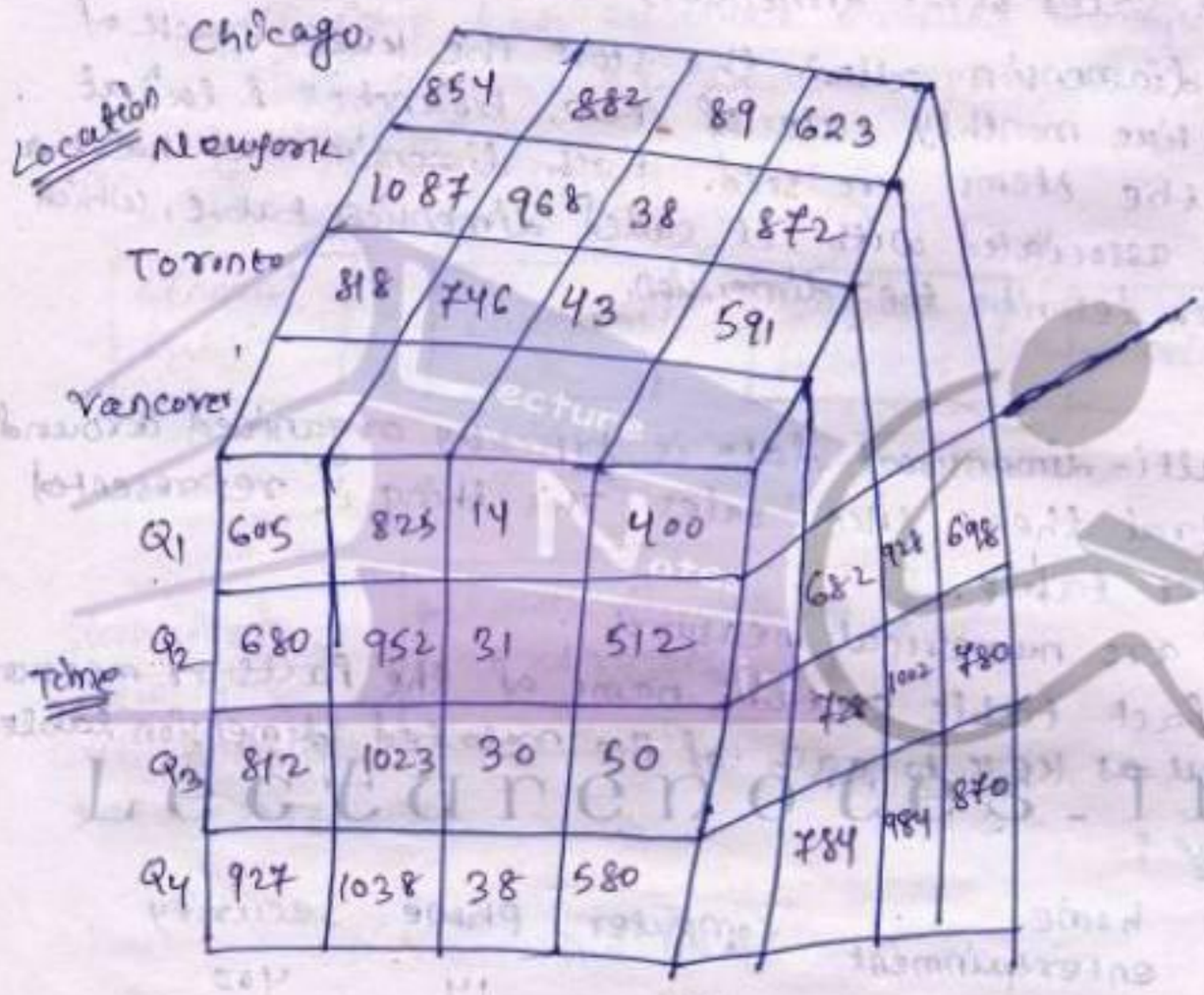
2-D view :-

Time	home entertainment	computer	Phone	security
Q_1	605	825	14	400
Q_2	680	952	31	512
Q_3	812	1023	30	501
Q_4	927	1038	38	580

Time	Home entertainment	computer	Phone	security
Q_1	854	882	89	623
Q_2	943	890	64	698
Q_3	1032	924	59	789
Q_4	1129	992	63	870

Location = "Chicago"

Time	Home entertainment	computer	phone	security
Q_1	1087	968	38	872
Q_2	1130	1024	41	925
Q_3	1034	1048	45	1002
Q_4	1142	1091	54	984



Location = "Toronto"

Time	Home entertainment	Computer	Phone	Security
Q_1	818	746	43	591
Q_2	894	769	52	682
Q_3	940	795	58	728
Q_4	978	864	59	784

Location = "vancover"

Time	home entertainment	computer	Phone	security
Q_1	605	825	14	400
Q_2	680	952	31	512
Q_3	812	1023	30	501
Q_4	927	1038	38	580

Representation of model :-

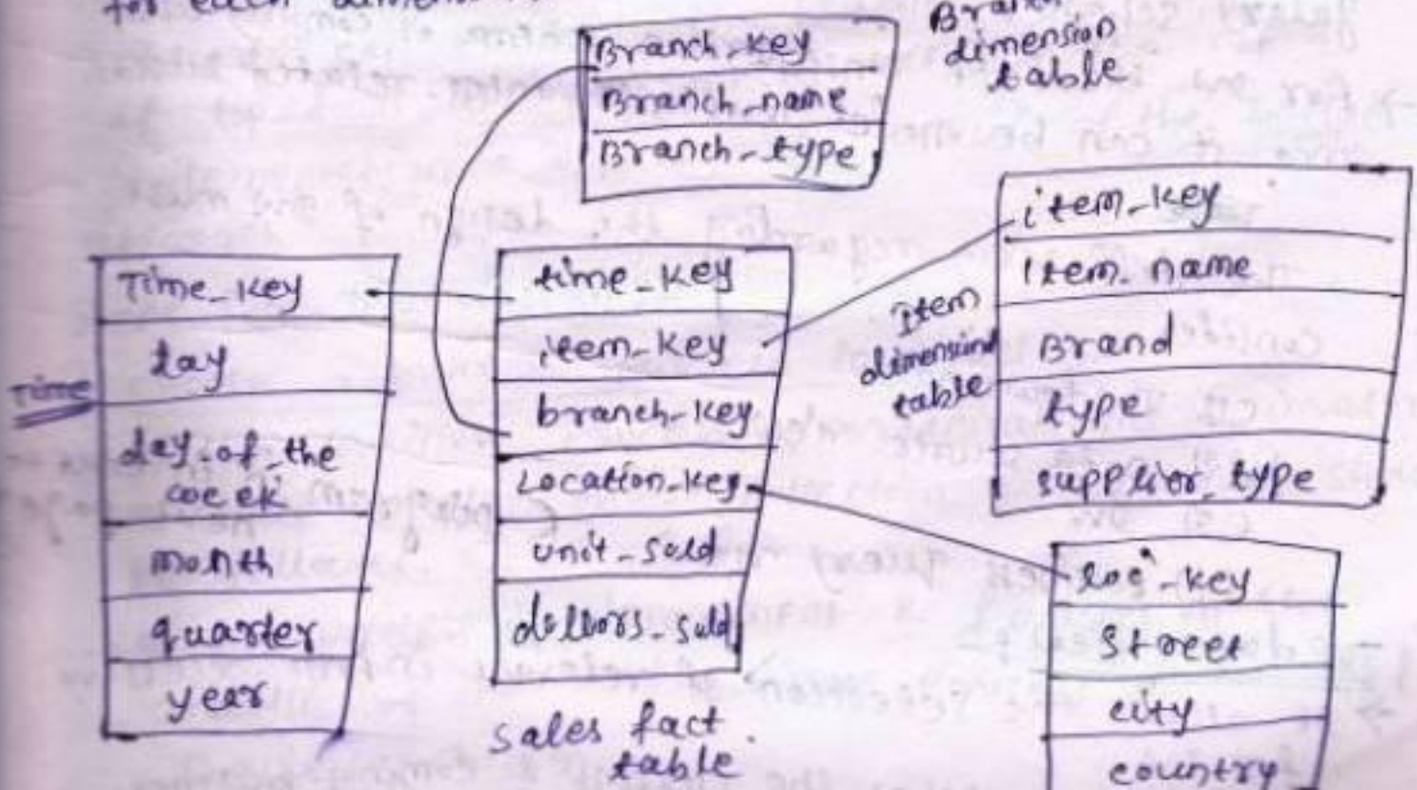
① star
② snowflake
③ fact constillation schema for multidimensional dB.

The most popular data model for a DW is a multi-dimensional model, such model can exist in form of star schema, snowflake schema & fact constillation schema.

Star :-

The most common model paradigm in the star schema in which DW contains.

(i) a large central table (fact table) containing the bulk of data with ~~more~~ no redudancy.
(ii) a set of smaller attendant tables (dimensional table for each dimension)

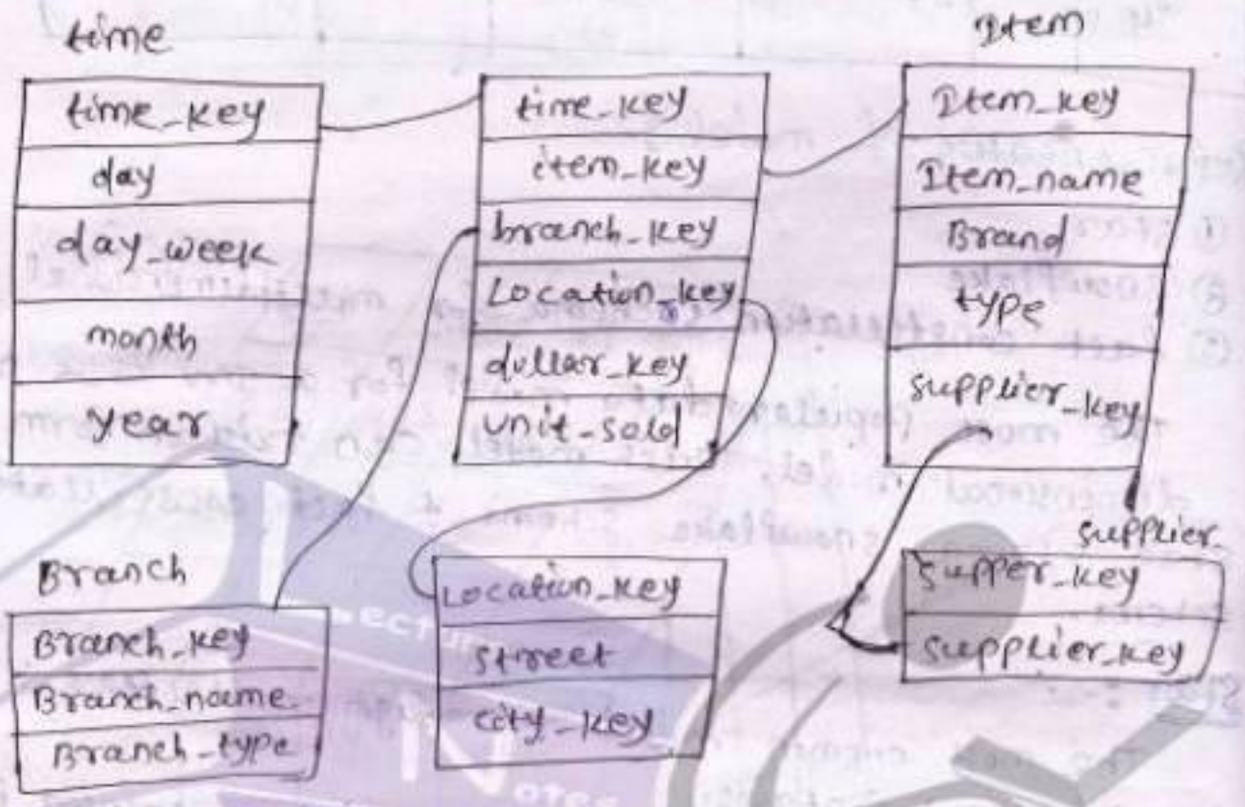


→ It is a varient of star schema table where some dimension tables are normalized.

→ They are further splitting the data into additional table.

→ The dimensional table or snowflake model may kept normalized form to reduce the redudancy.

Dt. 6.9.11



Fact Constellation Schema :-

→ Sophisticated appl^n may require multiple fact tables to share dimension table. This kind of schema can be viewed as collection of star & hence it is called galaxy schema.

→ For DW the fact constellation schema is comonly used since it can be model "multiple inter-related subjects

The diff view regarding the design of DW must consider

(1) Topdown view
(2) Data source view
(3) DW
(4) Business query view (Diagram in in next to next page)

Top down view :-

→ It allow the selection of relevant inform^n necessary for DW.

→ This inform^n matches the current & coming buschess needs.

Data source view :-

→ It expose the inform^n being captured, stored & managed by operational system.

→ Data source are often modelled by traditional data modelling techniques such as ER model.

DW view :-

→ It includes & dimension tables.

→ It represents the inform^n i.e, stored inside DW

Business Query view :-

→ It is the perspective of data in DW from the view point of the end user view.

Process of DW design :-

→ A DW can be build using top-down approach, bottom-up approach & mixed approach.

Top-Down Approach :-

→ It starts with overall design & planning.

→ It is useful in case where the technology is well known & where the business problems that must be solved & cleared & well-understood.

Bottom-up Approach :-

→ It starts with expt & prototypes.

→ It is useful in early stage & business module & technology development.

Mixed Approach :-

→ An org^n can exploid the planned & stratagic nature of top-down approach while retaining the rapid implementation & opportunity, applic^n of the bottom-up approach

Steps for DW design :-

① choose a business process to model :-

→ If the business process is organisational & involves multiple complex object collection, a DW model should be followed.

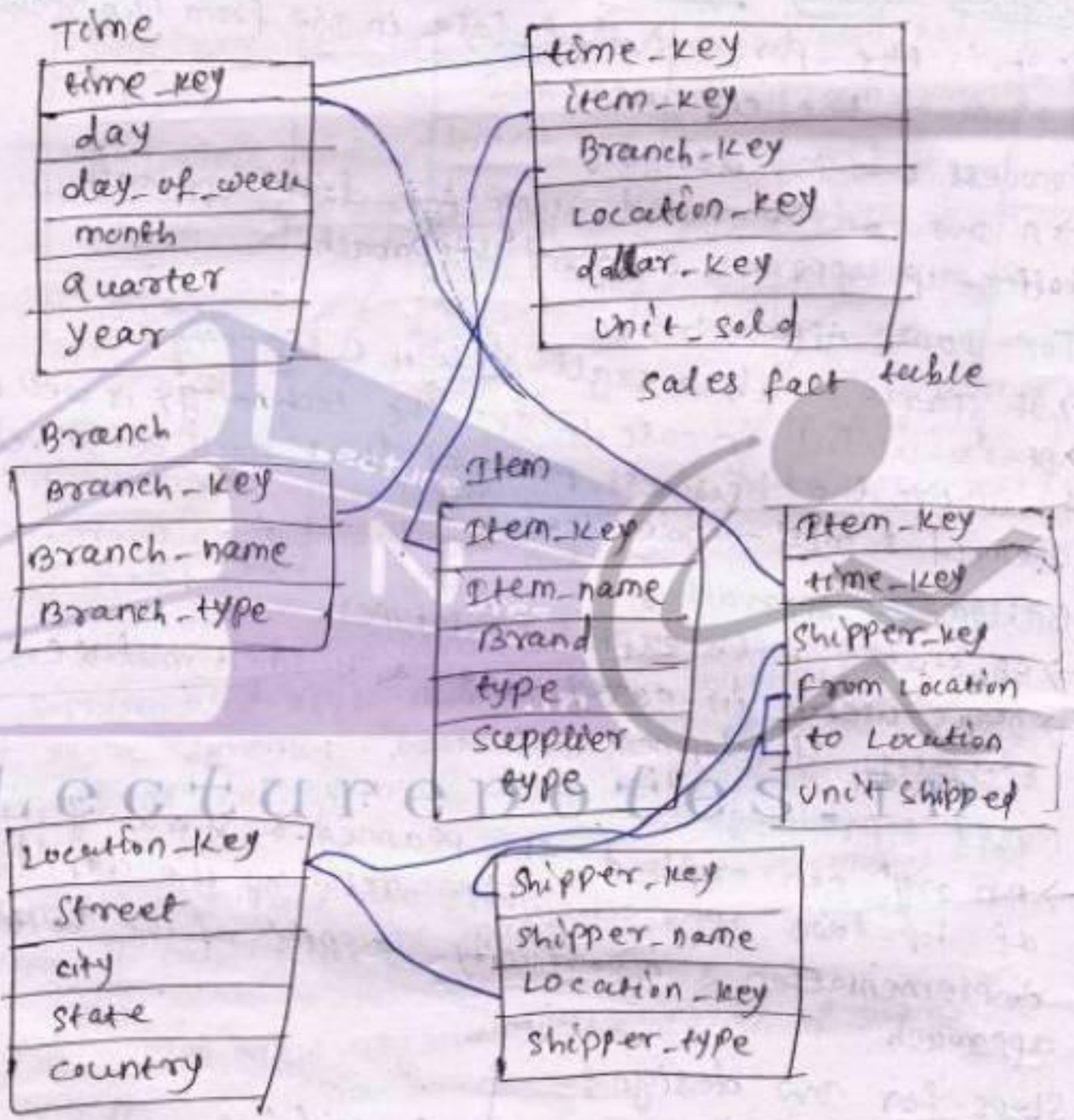
→ If the process is department & focuses on the analysis of one kind business process, a DM model should be choosen.

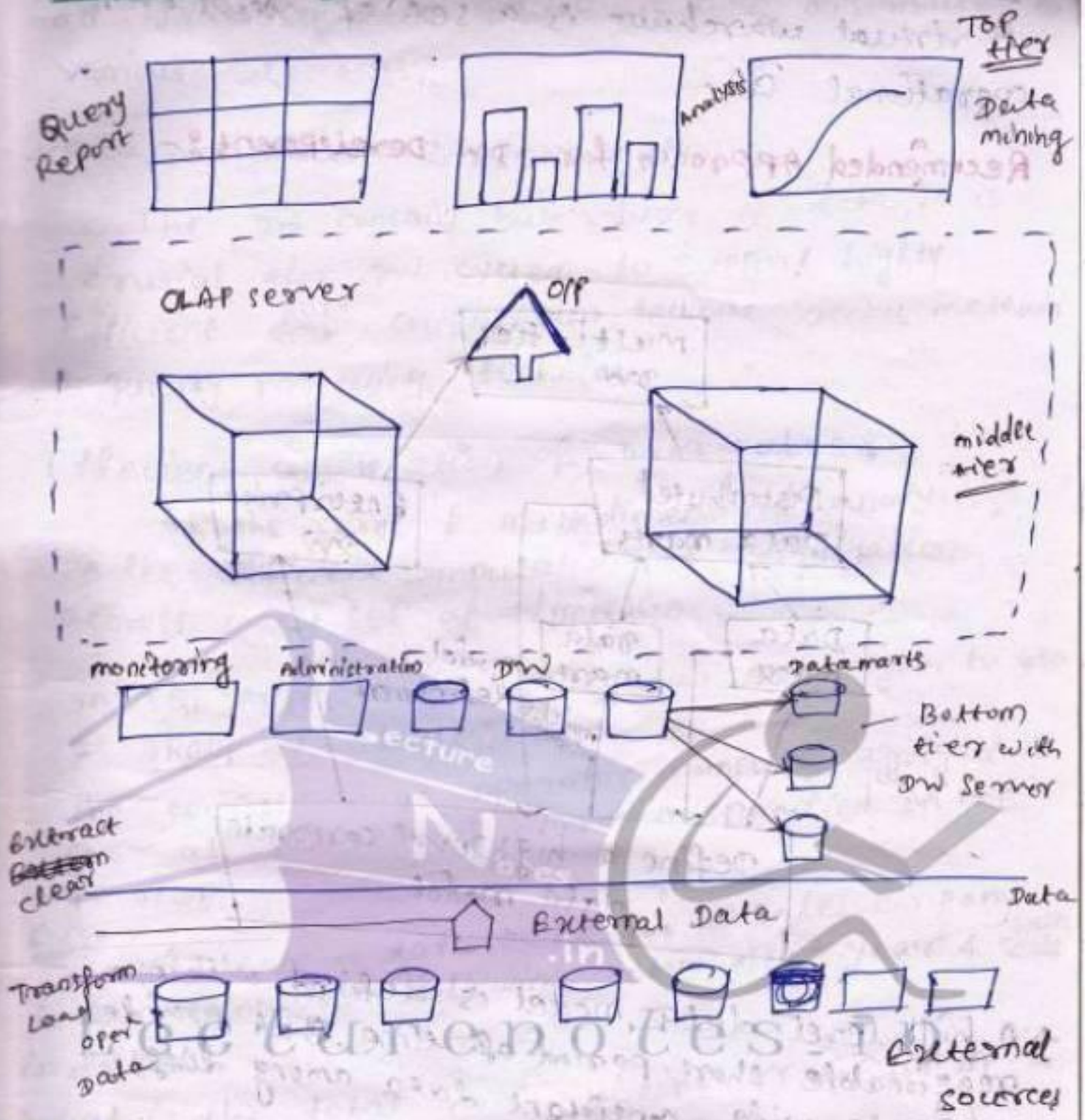
→ The grain is the fundamental atomic tabel of data to be represented in fact table for this process.

③ Choose the dimension that will apply to each fact table record.

④ Choose the measures that will populate each fact table.

Diagram of fact constellation schema :-





Dt. 8.9.11

From the architecture point of view there are three DW model :-

- Enterprise warehouse
- Data mart
- virtual warehouse

Enterprise warehouse :-

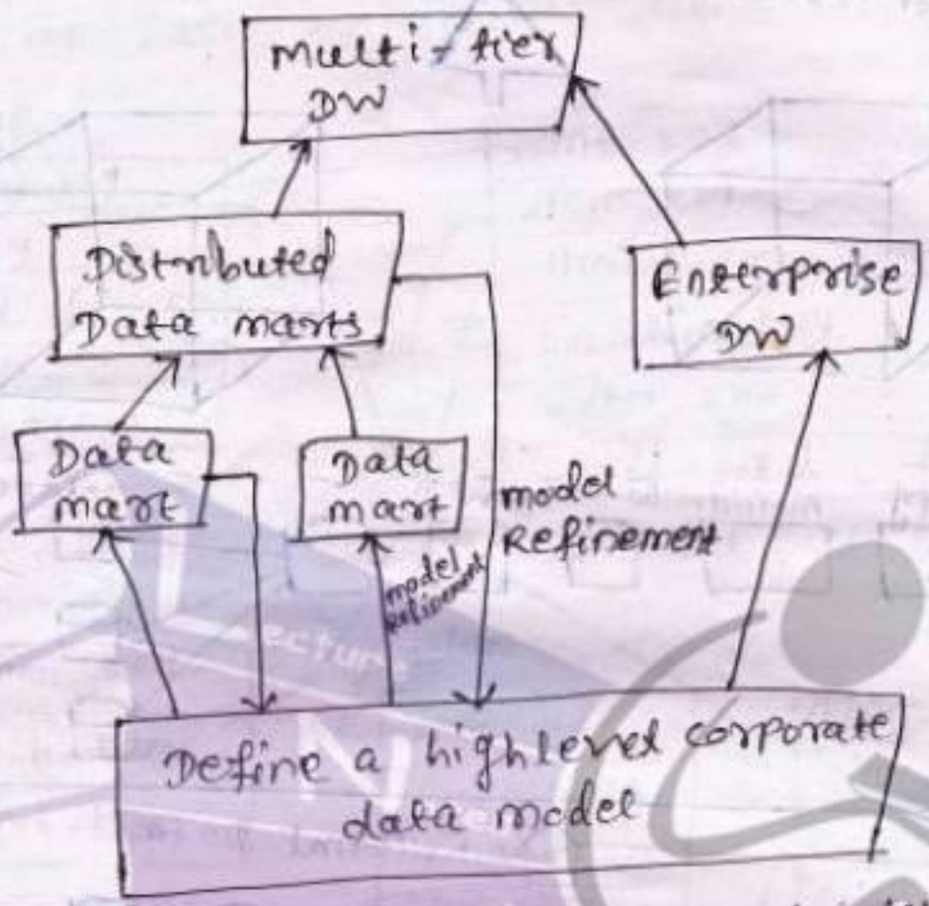
An enterprise warehose collects all of the infor^n about subjects spanning the entire org^n.

Datamart :-

A datamart contain a subset of corporate wild data ie of value to a specific user of

A virtual warehouse is a set of views over operational DB.

Recommended Approach for DW Development:-



→ A high level data model is defined within a reagionable short period of time that provides corporate wide consistant data among diff. subjects.

→ The high level model, it will need to be refined in the further development of enterprise DW & departmental data mart greatly reduced further integration problem.

→ Independent data marts implemented in parallel with the enterprise warehouse.

→ Distributed data marts can be constructed to integrate diff. data marts via server.

→ Finally a multi-tier DW is constructed, where the enterprise warehouse is constructed.

→ The enterprise warehouse is the sole custodian of all warehouse data which is then distributed to various datamarts.

Data WareHouse Implementation :-

The DW contain huge volume of data, it is crusial for DW system to support highly efficient ~~few~~ cube computation technics, access methods & querry processing technics.

Effecient computation of Data cubes :-

At the core of multidimensional analysis, is the efficient computation of aggrigation accross many set of dimension.

In SQL term these aggrigation are refer to ~~as~~ as GROUP BY clause.

The compute cube operator computes aggrigate over all sub-set of dimension specified in the operation.

Ex:- To create a data cube for a sales company that contains the following :- item, city, year & ~~sales~~ sales in dollar.

Q- What is the total no. of cuboids or GROUP-BY can be computed to the data cube ?

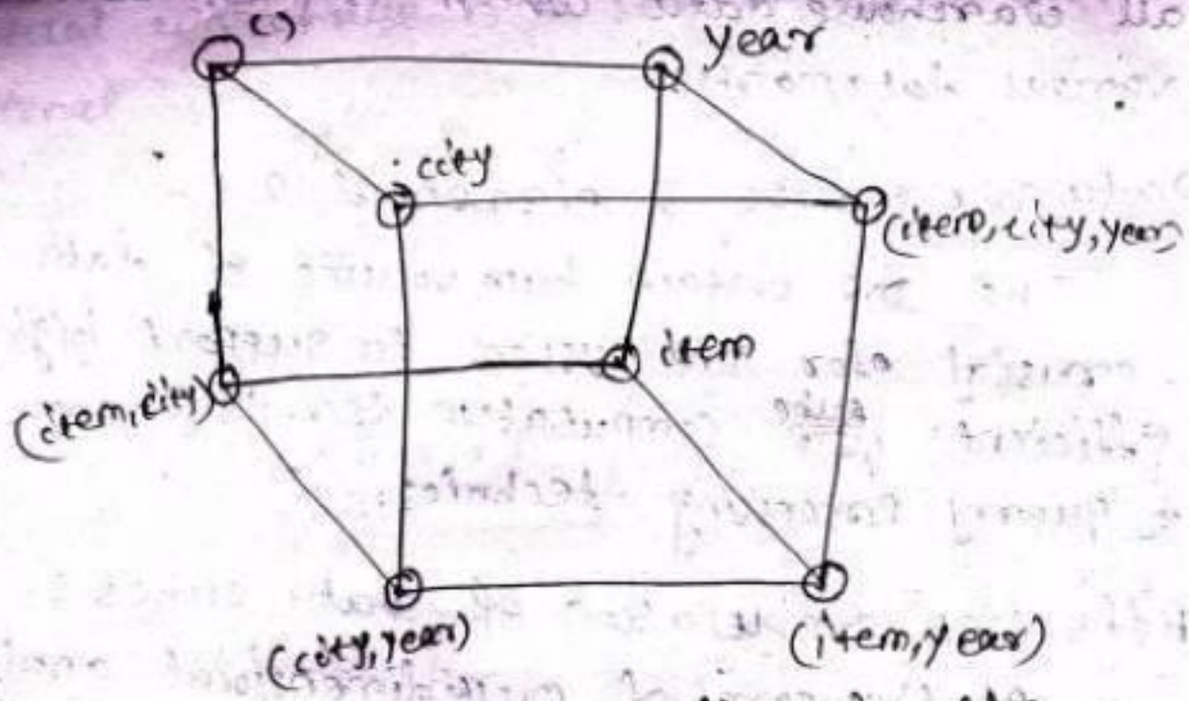
Sales

Item, city, year } Three dimension

Sol:- Total no. of cuboids = {(Item, city, year), ~~(Item,~~

(Item, city), (city, year), (Item), (city), (year), ()}

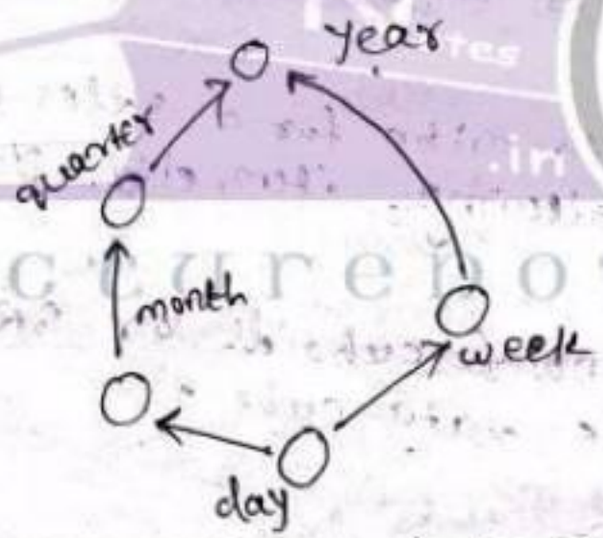
$\boxed{\text{Formul} = 2^n}$, where n = dimension



→ Taking 3 attributes as 3 dimension, the total no. of cuboid can be computed as

$2^3 = 8$

→ For a cube with 'n' dimensional a total of 2^n cuboids, including the base cuboid



For 'n' dimensional data cube the total no. of cuboids is denoted by T, &

$$T = \prod_{i=1}^{n} (l_i + 1)$$

l_i = no. of levels associated with dimension i.

→ For ex:- If the cube has '10' dimension & each dimension have '4' levels, the total no. of cuboids are generated

$= \prod_{i=1}^{10} (4+1) = 5^{10}$ cuboids to be formed (Ans)

STAR Schema:- Dt. 15.9.11

Syntax for creating fact table :-

define cube (cube_name) [(dimension_list)] : (measure_list);

define cube sales_star [time, item, branch, location] : dollar_sold = sum (sales_in_dollar) unit_sold = count(*);

Creation of dimension Table:-

Syntax

define dimension (dimension_name) as (attribute_on_subdimension list)

Ex:- define dimension time as (time_key, day, day_of_week, month, quater, 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, state, country)

Syntax for creation of snowflake schema :-

Define cube sales_snowflake [time, item, branch, location] : dollar_sold_sum(sales_in_dollar) unit_sold = count(*);

Define dimension time as (time_key, day_of_week, month, quater, 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)

city (city_key, city, sta
country))

FACT constellation schema :-

Define cube sales [time, item, branch, Location]:
dollar_sold = sum(sales-in-dollars
unit_sold = cound (*)

Define dimension time as (time_key, day, day_of_we
month, quater, year)

Define dimension item as (item_key, item_name, brand,
type, supper_type)

Define dimension branch as (branch_key, branch_n
branch_type)

Define dimension location as (location_key, street,
city, state, country)

Define cube shipping [time, item, shipper, from_locat
to_location]:
dollar_cost = sum (cost_in_dollers),
unit_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
cuble sales

Define dimension to_location as location in
cube sales

OLAP operation in Multidimensional Data model :-

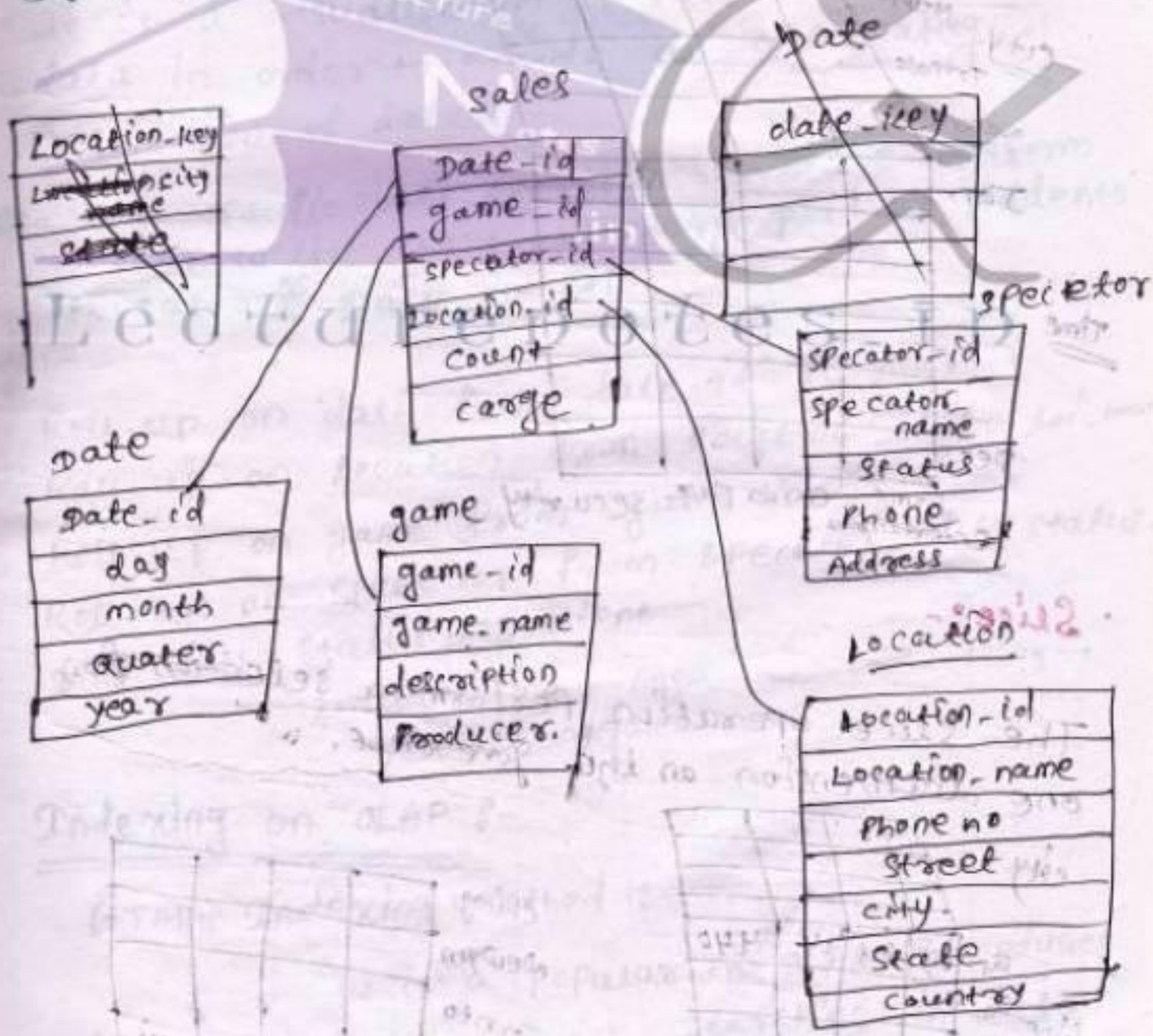
The roll-up operation (also called drill up oper^n)
Performs aggregation on a data cube either by
climbing up a concept of hierarchy for a
diamension or by diamension reduction.

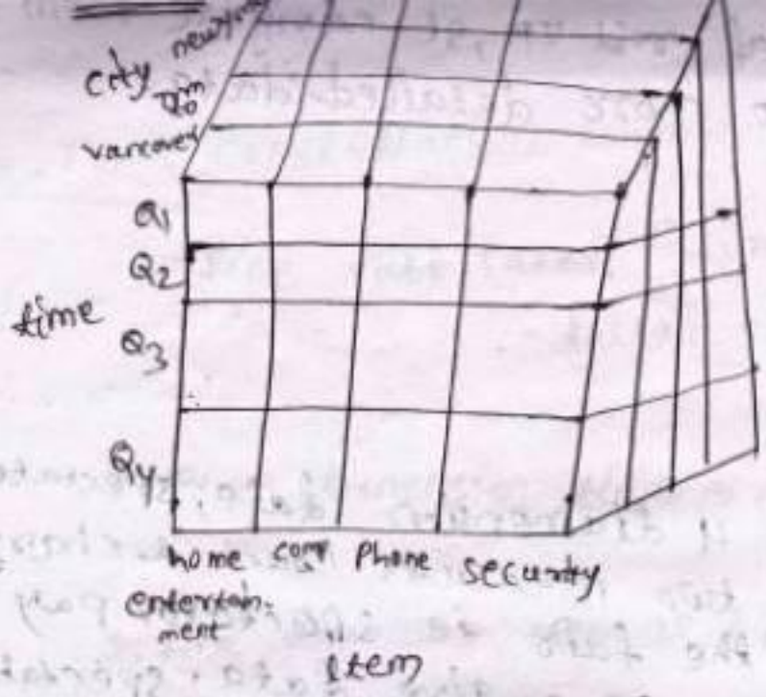
Drill Down

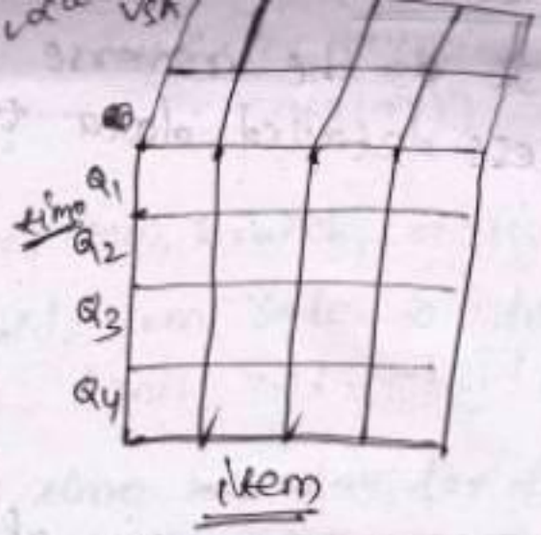
It is the reverse of roll-up, it navigate from less detailed data to more detailed data.

Q:- Suppose a DW consist of 4 diamension date, spectator, location & a game & two measurs count & charge (cost), where charge is the fair ie spectator pay when watching a game on a give data. spectator may be a student, adult or a senior with each catagory its own charge rate.

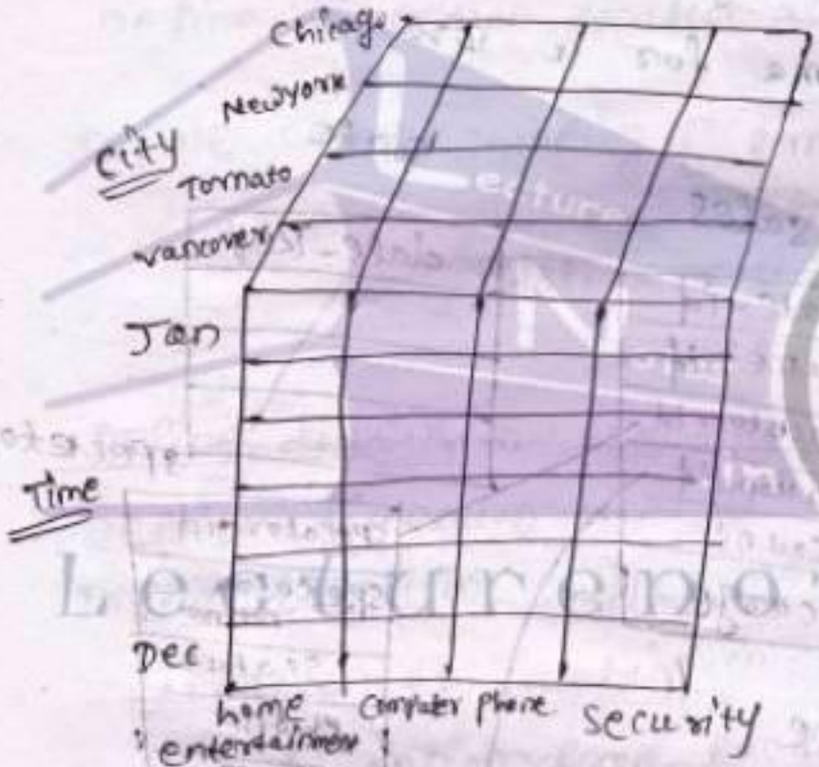
Draw a star schema for a DW.





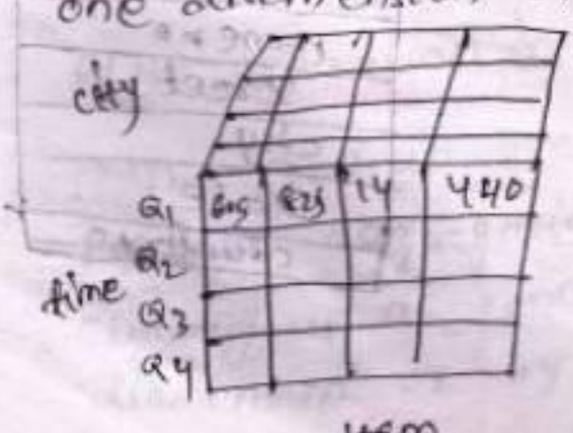


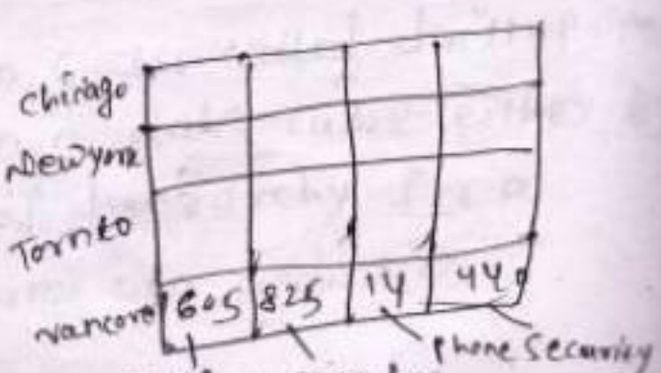
Drill Down



Slice:-

The slice operation perform a selection on one diamension on the give cube.

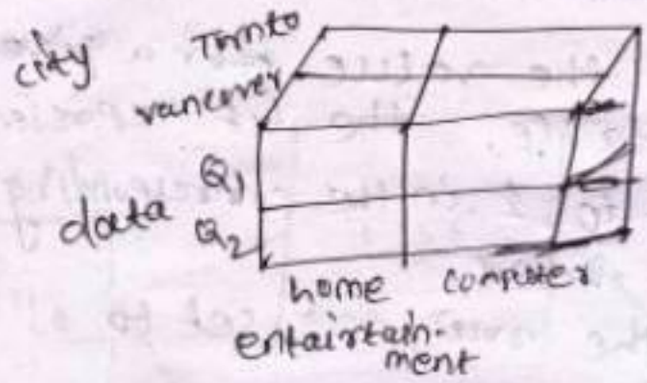




Dice

The dice operation defines a sub-cube by performing a selection on two or more diamension.

Ex:- Location is torrto & vancover & time is given Q_1 or Q_2 & item is home entertainment & computer.



Pivot

It is a visualisation operation that rotate the data in order to provide an alternating Presentation of data.

Q:- What specific OLAP operation should be perform inorder to list the total charge paid by students spectator at BBSR in 2010.

Ans:- Roll-up on date from date_id to year.
Roll-up on location from location_id to loc_name
Roll-up on game from game_id to all.
Roll-up on spetector from specator_id to status.
Dice with status = student
location_name = BBSR
year = 2010

Indexing on OLAP :-

BITMAP Indexing method :-

It is a popular in OLAP Procedures because it allows quick searching in data values. The bitmap index is an alternating

representation of the record_id (r_id) list.

→ In bitmap index for a given attribute there is a distinct bit vector 'bv' for each value 'v' in the domain of the attribute.

→ If the domain of a given attribute consists of 'n' values, then 'n' bits are needed for each entry in the bit-map index.

→ If the attribute has the value for a given row in the data table, the bit representing that value is set to 1 in the corresponding row of the bitmap index.

→ All other bits for the row are set to 0.

RID	Item	city
R_1	H	V
R_2	C	V
R_3	P	V
R_4	S	V
R_5	H	T
R_6	C	T
R_7	P	T
R_8	S	T

RID	H	C	P	S
R_1	1	0	0	0
R_2	0	1	0	0
R_3	0	0	1	0
R_4	0	0	0	1
R_5	1	0	0	0
R_6	0	1	0	0
R_7	0	0	1	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

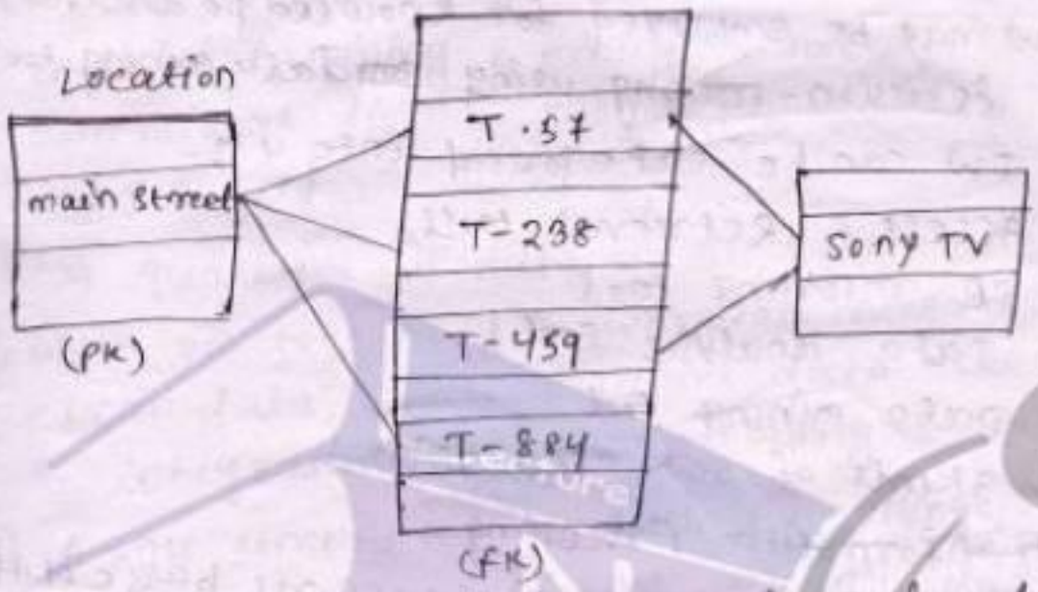
RID	V	T
R_1	1	0
R_2	1	0
R_3	1	0
R_4	1	0
R_5	0	1
R_6	0	1
R_7	0	1

JOIN INDEXING :-

Dt. 21.9.11

→ Join-indexing is especially used for maintaing relationship betn a foreign key & its matching primary key from the joinable relation.

→ The star-schema model of DW makes join-indexing attractive for cross-table search because the linkage betn a fact-table & corresponding dimension table are the foreign key of fact table & primary key of dimension-table.



JOIN-INDEX table for location & sales

mainstreet	T-57
main street	T238
mainstreet	T884

Join-Index table for sales & Item

T57	Sony TV
T459	Sony TV

Location	sales	Item
mainstreet	T57	Sony TV

From Data warehousing to Data Mining :-

Datawarehouse usage :-

→ Initially DW is mainly used for generating reports & answering pre-

→ progressively it is used to analyse summerized & detailed data where the results are presented inform of reports & charts.

→ Later DW is used for strategic purposes performing multi-dimensional analysis.

→ Finally, DW may be embloyed for knowledge discovery & strategic decission-making using ~~and~~ data mining tools.

→ Tools for DW can be categorised into 4 :-

① Access & Retrieval tools.
② DB reporting tool
③ Data analysis tool
④ Data mining tool

There are 3 kinds of DW application

① Information Processing
② Analytical processing (supports basic OLAP operations)
③ Data mining

It supports knowledge discovery by finding hidden patterns performing classification, specification and prediction & presenting the mining results using visualization tools).

Difference :- Data mining & Information Processing :-

→ based on queries, confined to useful information.

→ Information is stored in dB.

→ They do not reflect sophisticated pattern in dB.

→ So, information is not data mining

OLAP	Data Mining
→ OLAP is a data summerization or aggregation tools that simplify data analysis.	→ Data mining allows automated discovery of implicit patterns interesting knowledge hidden in large amount

→ It presents description of data process DW.
OLAP functions are essentially for user-defined data summary.

→ DW covers a much broader spectrum that OLAP becoz it not only performs data summery & comparisons but also performs associations but also performs association, classification, prediction, clustering & other DW

From OLAP to OLAM (Online Analytical Mining) :-

→ OLAM which integrates online Analytical processing with datamining & mining knowledge in multi-dimensional dB.

→ Why Should

Five qualities of data in DH :-

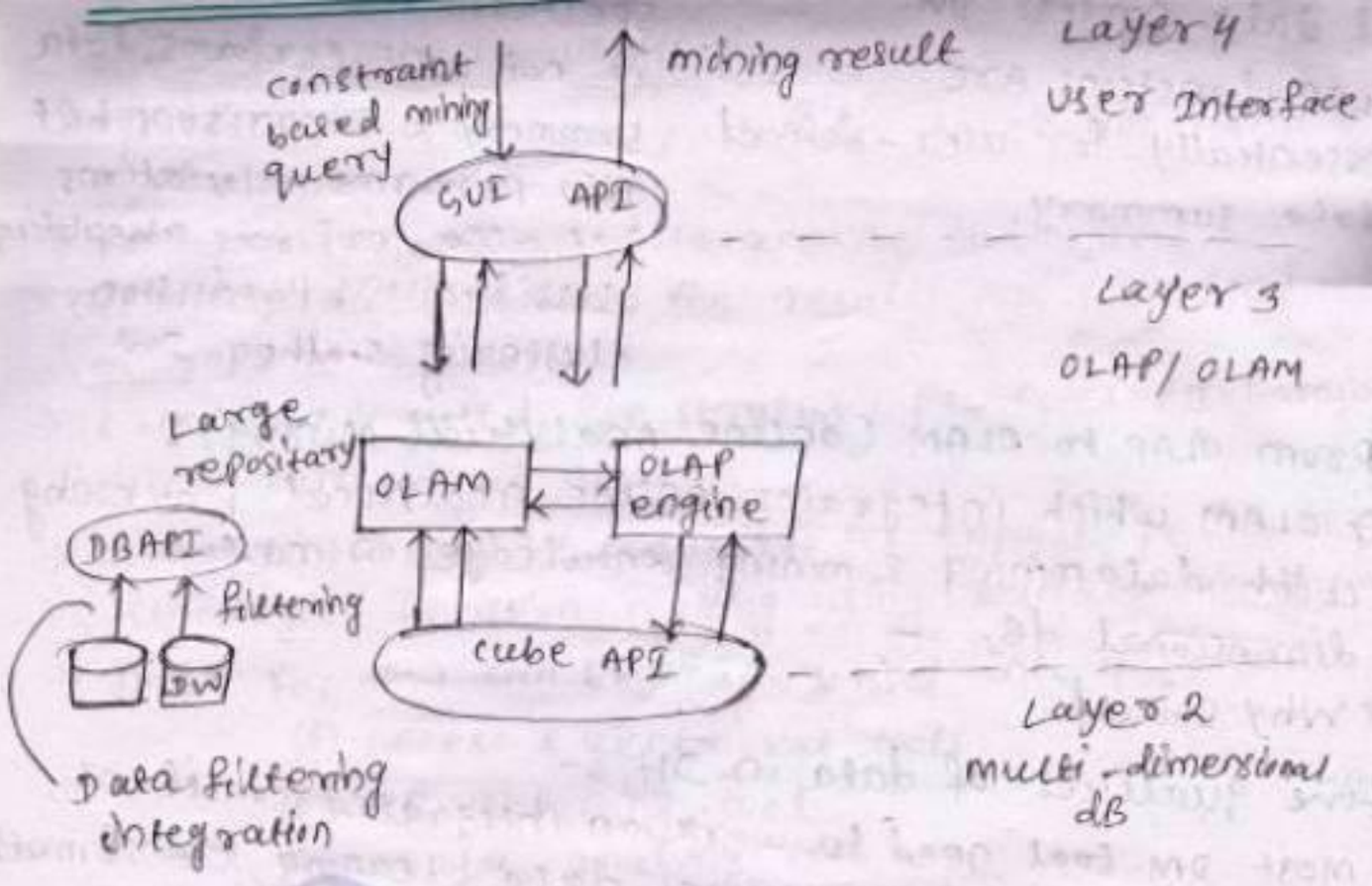
① Most DM tool need to work on integrated, consitent, clean data, which requires data cleaning, transformation & integration as pre-processing step.

② A DH constructed by such processing serves as a valuable source of high quality data for OLAP as well as OLAM.

③ Available information processing infrastructure surrounding DW. Accessing, Integration & Transform^n of multiple heterogenous dB, ODBC connection, web accessing & service facilities & reporting & OLAP analysis tools.

④ OLAP based exploiratory DBS :-
OLAM provides facilities for DM on different subset of data & different level of abstraction by pivoting, dicing, drilling, slicing on a data cube & intermediate DM results.

⑤ online selection of DM function :-
By integration OLAP with multiple DM functions OLAM provide user with flexibility to select desired DM function & swap DM task dynamically.



Both OLAP and OLAM :-

→ Accept user online queries via GUI & work with data cube in data analysis via a cube API.

→ A metadata directory is used to guide access of data cube.

→ The data cube can be constructed by accessing or integrating multiple dB via MDDB API or by filtering a DW via dB API that make support of ODBC connection.

DATA MINING

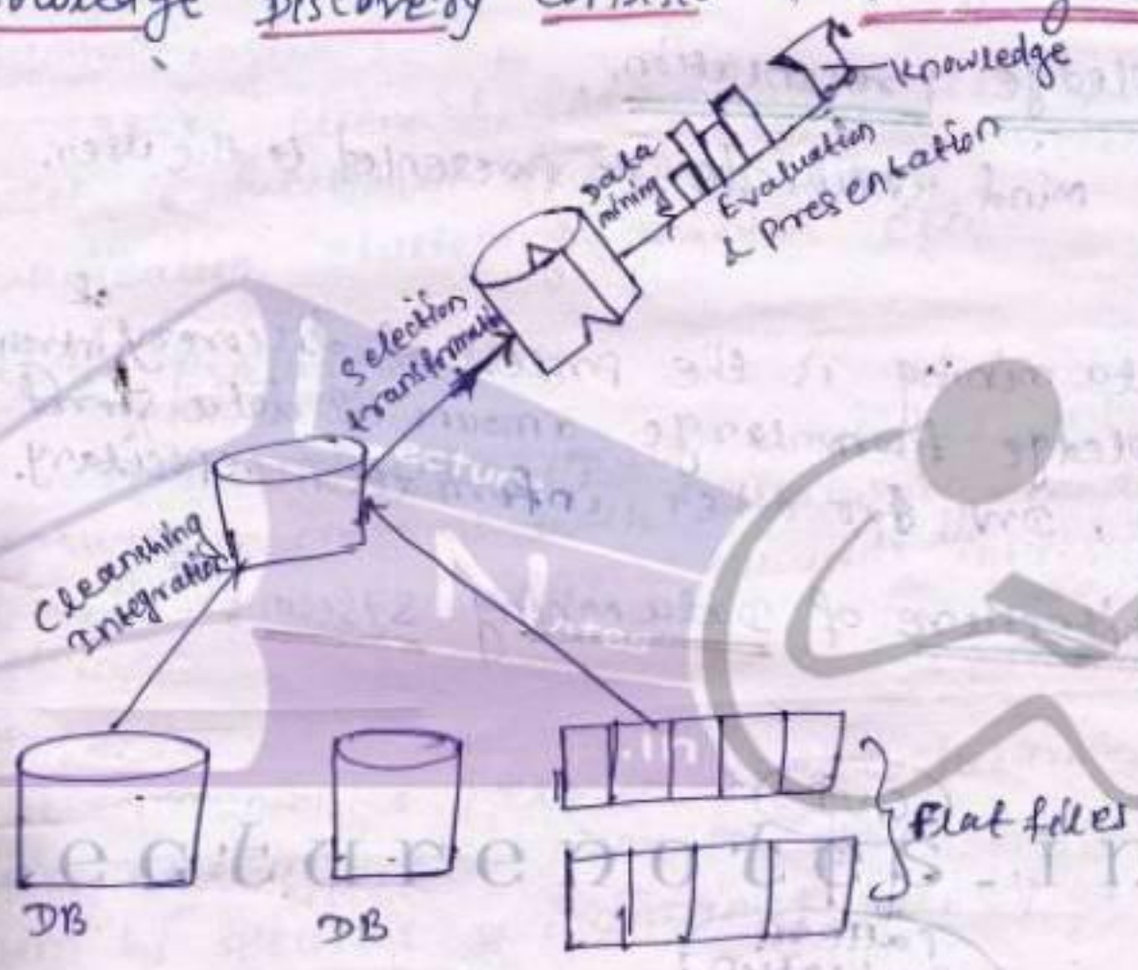
Dt. 22-9-11

Def:-

Data mining refers to extracting or mining knowledge from large amount data.

The term data mining is also refers to as knowledge mining, knowledge extraction or knowledge discovery in dB (Kdd)

Knowledge Discovery consists of following steps:-



Cleanshing :-

Removing inconsitant & noisy data.

· Integration :-

Multiple data sources are combine together.

Data selection :-

Data relevant to the analysis task are retrive from the dB.

Data Transformation :-

Data are transformed or consolidated

Data mining :-

Essential step when inteligent methods are applied to extract intresting data patterns.

Pattern Evaluation :-

Intresting patterns are identified by appling some weighted measures.

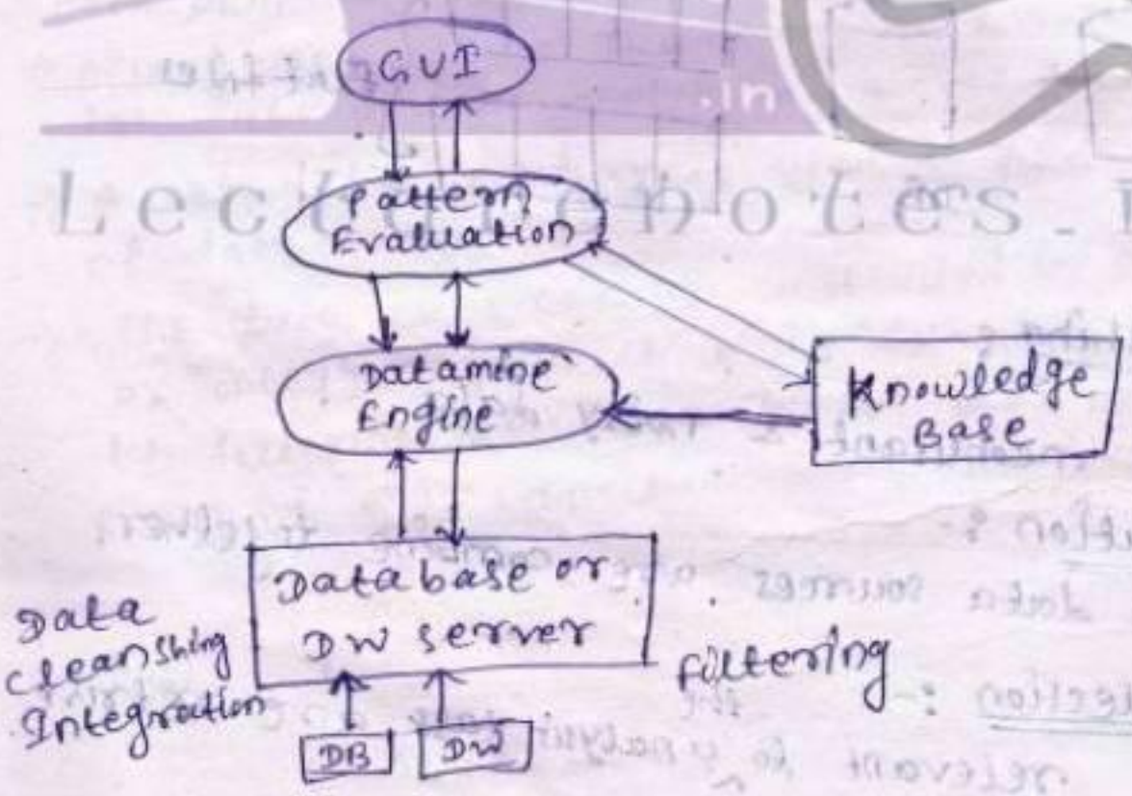
Knowledge Presentation.

Mind knowlede is presented to the user.

Def^

Data mining is the process of discovery of intresting knowledge from large amount of data stored in dB, DW for other information repositary.

Architecture of Data Mining System :-



DB & DW :-
These are information repositary.

DW server :-
It is responsible for fetching relevant data.

Knowledge Base :-
This is the domain knowledge used to guide the search or evaluate the resulting patterns.

Datamine Engine :-
It is essential to datamining & ideally consists of set of functional modules such as association, classification, clustering analysis etc.

Pattern Evaluation :-
This component typically employees interesting measures & interact with datamining modules so as to focus the search towards intresting pattern.

GUI :-
It communicates between user & datamining system allowing the user to interact with the system by specifing a datamining query or a task.

Datamining Functionality :-
Datamining task can be classified into two catagories :-
Descriptive
Predictive

Descriptive
Descriptive mining task characterise the general properties of data in the db.

Predictive :-
Predictive mining task perform inference on current data inorder to make prediction.

Description can be derived through two process

Data characterization

Data Discrimination

Data characterization :-

It is a summerization of general characteristics or feature of a target class of data.

for example :- the customer of store who spend more than ~~Rs~~ 1000/-

Data Discrimination :-

It is a comparision of general feature of target class data object with the general feature of objects from one or set of contrasting classes.

For example :- user may like to compare general feature of product whose sale increased by 10% last year with those ~~sale~~ sale decreased by 30% in the same year.

Various Datamining function :-

Mining frequent patterns :-

frequent patterns are patterns that occur frequently on data.

There are many frequent patterns such as

1) frequent item set
2) subsequences
3) sub- structures

Frequent - item set :-

Typically refers to set of items that occur together in transactional data set.

Frequent subsequences :-

It is a set of items that occurs or tend to occur in a subsequent

transcation from same class of customer.

Sub-structure :-

A sub-structure refers to different ~~structure~~ structural ~~from~~ form such as graph, tree etc which may be combine with item set.

Ex :- If a substructure occurs frequently, then it is called frequent structure pattern.

Dt. 23.9.11

Association Analysis :-

An association rule determine which item or items are closely associated with other item. This analysis is widely used for market basket or transcation data analysis.

An association rule is the form of $X \rightarrow Y$, where

$X = \{X_1 - - X_n\}$

$Y = \{Y_1 - - Y_n\}$ are set of items.

The association rule states that if a customer buy X_i he/she also likely to buy Y_i, where X & Y are set of items.

The set $X \cup Y$ is called item set.

The association rule has a support & confidence.

Support →

The support for a rule $X \cup Y$ is w.r.t item set ~~of~~ refers to how frequently a specific item set occurs in the DB.

Confidence →

It is the percentage of transcation that contain X as well as Y.

$$\text{support } X \rightarrow Y = \frac{\text{support count } (X \cup Y)}{N}$$

N — total no. of transcation

$$\text{confidence } X \rightarrow Y = \frac{\text{support count } (X \cup Y)}{\text{support count } (X)}$$

Problem

Transaction	Item
101	milk, Bread, cookies, juice
792	milk, Juice
1130	milk, egg
1735	Bread, cookies, coffee

~~Support~~	~~confidence~~ support	confidence
Bread → cookies	2/4	2/2 – 100%
milk → Juice	2/4	2/3
cookies → coffee	1/4	1/2
milk → egg	1/4	1/3 – 100%

~~Find out support~~

customer buys a bread, there is a 100% chance

A support of 50% means that out of all transcation, 50% transcation shows purchase of bread & cookies.

	confidence
cookies → Bread	2/2

Lecturenotes.in

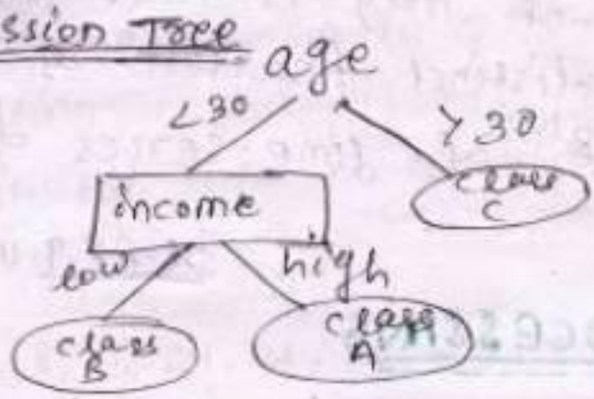
classification & Predication:-

classification is the process of finding a model or function which can be used to predict the class of object whose class is unknown. The model is derived by analysis of set of objects whose class level is known (training data).

once existing data is analysed & a model is prepared subsequent objects can be predicted to belong to which class.

age(x, <30) And income (x, high) ⇒ class(x, A)
age(x, >30) class (x, C)
age (x, <30) and income (x, low) ⇒ class(x, B)

Decission Tree



Clustering Analysis :-

In classification & prediction we create a model based on object with known class level. However in clustering data objects are analyse without consulting any model with known class level. The class level are not known in the begining. The objects are clustered or grouped together based on the principle of maximizing similarity between objects of the same cluster & minimizing similarity of object of diff. groups.

Outlayer Analysis :-

A dB may contain data object which does not complay with general behaviour or characteristics of data. such data objects are called out layers.

for example however in some application such as fraud detection, rare events are more intresting than regular events.

A credit card fraud can be detected when a transcation doesnot fall into the known behaviour pattern that transcation fall on outer layer.

Evaluation Analysis :-

This analysis describes and models regulation or trends for objects whose behaviour change over time. Although it may include all other analysis method, a distinct feature of evaluation analysis includes analysis of time series data.

Dt-27.9.11

DATA PREPROCESSING

There are no. of data processing technic :-

- Data cleaning
- Data Integration
- Data Transformation
- Data reduction

Data cleaning :-

→ To clean the data filling in missing values, smoothing out noisy data & identifing or removing out layers & resolving inconsistancies.

How to Handle the missing values :-

~~Ignore~~ Ignore the tupple

→ Filling in the missing value mannualy.

→ use a global constant to fill missing value

→ Use the attribute mean to fill in the missing value.

→ Use the attribute mean for all samples belong in the same class as given tupple.

→ Use the most probable value to fill missing value.

This may be determined with regression or decession tree.

Evaluation Analysis :-

This analysis describes and models regulation or trends for objects whose behaviour change over time. Although it may include all other analysis method, a distinct feature of evaluation analysis includes analysis of time series data.

Dt-27.9.11

DATA PREPROCESSING

There are no. of data processing technic :-

- Data cleaning
- Data Integration
- Data Transformation
- Data reduction

Data cleaning :-

→ To clean the data filling in missing values, smoothing out noisy data & identifing or removing out layers & resolving inconsistancies.

How to Handle the missing values :-

~~Ignore~~ Ignore the tupple

→ Filling in the missing value mannualy.

→ use a global constant to fill missing value

→ Use the attribute mean to fill in the missing value.

→ Use the attribute mean for all samples belong in the same class as given tupple.

→ Use the most probable value to fill missing value.

This may be determined with regression or decession tree.

How to Handle Noisy data?

Noisy means random error.
Remove the noise by using binning method.

Binning method smooth a sorted data value by consulting neghibourhood. The sorted values are distributed into no. of buckets or bins.

Ex:-

4, 8, 15, 21, 21, 24, 25, 28, 34 → It is already in sorted order

Bin = 3 (Given)

Bin 1 = 4, 8, 15
Bin 2 = 21, 21, 24
Bin 3 = 25, 28, 34

method

① Smooth By Bin mean

Find the mean & then replace it.

bin 1	9	9	9
bin 2	22	22	22
bin 3	29	29	29

② Smoothing by bin boundaries

Find mean & max^m values as bin boundaries
Each bin value is replaced by bin boundaries

Bin 1	4	4	15
Bin 2	21	21	24
Bin 3	25	25	34

③ smoothing by median

If there are 14 datas

$\frac{14}{2} = 7+1 = 8 \rightarrow \frac{\text{content of } 8^{th} \text{ position}}{2}$ = median

$q_2 = \frac{4 \text{ to } 5}{} = \frac{21+21}{2} = 21$

→ outlayers may be detected by clustering where similar values are organised into groups or cluster.

Regression :-

Data can be smoothed by fitting the data into a function such as regression. Linear regression involves finding the best line to fit two variables, so that one variable can be used to predict other.

Data Integration & Transformation :-

Problem

Entity Identification Problem

Redundancies can be detected by co-relation analysis.

$$r_{A,B} = \frac{\Sigma (A-\bar{A})(B-\bar{B})}{(n-1)\,\sigma_A \sigma_B}$$

$\bar{A} \rightarrow$ mean of A

$n \rightarrow$ total no. of tuples

$\sigma_A \rightarrow$ standard ~~derivation~~ deviation

$$\bar{A} = \frac{\Sigma A}{n}$$

$$\sigma A = \sqrt{\frac{\Sigma (A-\bar{A})^2}{n-1}}$$

cond^n :-

If the resulting value is greater than '0' then A & B are +vely correlated. It means that the value of A increases with the change of the value of B.

If result = 0, then A & B are independent to each other, there is no co-relation between them.

If resulting value < 0, A & B are -vely correlated, where the values of one attribute increases as the value of other attribute decreases.

Data Transformation :-

In data transformation the data are transformed or consolidated in the form of appropriate for mining. DT can involve the

smoothing (avoid noisy data)

Aggregation, where summery or aggregation operation are applied to data.

Generalization

In this method lower level data are placed by higher level concept through the use of concept of hierarchy.

Normalisation :-

Where the attribute data are scaled so as to within a small specified range such as 0.02 to 1.0 or -1 to 1.

Dt. 29.9.11

Q:- Suppose that the data for analysis include the attribute age. The age values for the data tupple are 13, 15, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 34, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

(i) use the smoothing by bin means to smooth the above data using a bin depth of 3.

Bin 1 = 13, 15, 16, 19, 20, 20, 21, 22, 22

Bin 2 = 25, 25, 25, 25, 30, 33, 33, 34, 35

Bin 3 = 35, 35, 35, 36, 40, 45, 46, 52, 70

Normalisation :-

① min max Normalization :-

It performs a linear transformation on the original data. suppose that min of A & max A at the min^m & max^m values of the attributes.

A min max normalization maps a value 'v' of A to v' in the range [new-min$_A$, new-max$_A$]

$$\cancel{v' = \frac{v - min_A}{max_A - min_A}}$$

$$\boxed{v' = \frac{v - min_A}{max_A - min_A}(new\text{-}max_A - new\text{-}min_A) + new\text{-}min_A}$$

Ex:- min^m income for the attribute is 12,000/-
max^m " " " is 98,000/-
we want to map bet^n [0.0, 1.0]
v = 73600

$$v' = \frac{73600 - 12,000}{98,000 - 12,000}(1.0 - 0.0) + 0.0$$

$$= \frac{61600}{86000}(1.0) = 0.716$$

Q:- suppose the value for age is
{20, 23, 29, 32, 22, 27}
you want to map bet^n (1.0, -0.0)

$$\frac{20 - 20}{32 - 20}(1.0 - 0.0) + 0.0$$

$$= 0$$

$$\frac{29 - 20}{12}(1.0) = \frac{9}{12}(1.0) = 0.75$$

23 ⇒

32 ⇒

22 ⇒

27 ⇒

Z-score Normalisation :-

The values for an attribute 'a' are normalised based on the mean & standard deviation of 'a'. 'a' value of v of A is normalised (v') by computing

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Ex:- min = 54000

σ_A = ~~73600~~ 16,000

v = 73600

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

$$= \frac{73600 - 54000}{16,000} = 1.225$$

Q:- The value for age is {20, 23, 29, 32, 22, 27}

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

$$\bar{A} = \frac{20+23+29+32+22+27}{6} = 25.5$$

$$\cancel{\sigma_A = \sqrt{\frac{(A-\bar{A})^2}{6-1}}}$$

$$\cancel{= \sqrt{\frac{9}{5}} =}$$

$$\sigma_A = \sqrt{\frac{\Sigma (A-\bar{A})^2}{n-1}}$$

$$\cancel{= \sqrt{\frac{20-25.5}{6-1}}}$$

$$\cancel{\sigma_{Age} = \frac{\Sigma (20-25.5)^2 + (23-25.5)^2 + (29-25.5)^2 + (32-25.5)}{}}$$

$$\sigma_{Age} = \sqrt{\frac{(20-25.5)^2 + (23-25.5)^2 + (29-25.5)^2 + (32-25.5)^2 + (22-25.5)^2 + (27-25.5)}{5}}$$

$$= \sqrt{\frac{30.25 + 6.25 + 12.25 + 42.25 + 12.25 + 2.25}{5}}$$

$$= \sqrt{21.1} = 4.593$$

$$20 \Rightarrow v' = \frac{v - \bar{A}}{\sigma_A}$$

$$= \frac{20 - 25.5}{4.593} = -1.197$$

$$23 \Rightarrow v' = \frac{23 - 25.5}{4.593} = -0.54$$

$$29 \Rightarrow v' = \frac{29 - 25.5}{4.593} = 0.76$$

Lecturenotes.in

$32 \Rightarrow v' = \frac{32-25.5}{4.59} = 1.41$

$22 \Rightarrow v' = \frac{22-25.5}{4.593} = -0.76$

$27 \Rightarrow v' = \frac{27-25.5}{4.593} = 0.326$

Normalisation by decimal scaling :-

It normalizes by moving the decimal point of values of attribute 'A'. The no. of decimal point move depends on max^m absolute value of A. A value of 'v' is normalized to v' by computing

$v' = \frac{v}{10^j}$, where j is the smallest integer.

Ex:- suppose that the recorded values of A range from -986 to 917.

The normalised by decimal scalling will devide each value by 1000.

$v' = \frac{-986}{1000} = -0.986$, $j=3$

Ex:- suppose the data range is given

{-879, -213 ... 114, 186}

v' of -879 : $v'(-879) = \frac{-879}{1000} = -0.879$, $j=3$

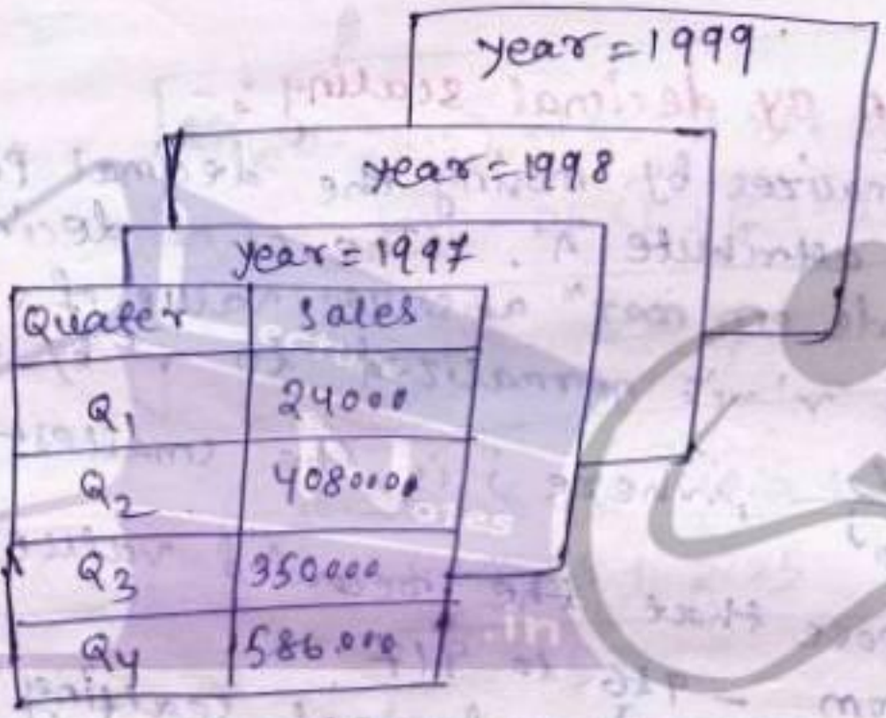
$10^j = 10^3 = 1000$

Data Reduction :-

Data reduction technics can be applied to obtain a reduced representation of data set ie smaller in volume.

Data cube Aggregation :-

Where Aggregation operation are applied to the data in the construction of a data cube.



Year	Sales
1997	1568000
1998	2356000
1999	3594000

Dt: 30.9.11

Diamension Reduction :-

In diamension reduction, where reduldant attributes, irrelevant attributes or diemension detected or removed.

In diamensionality reduction reduces the dataset size by removing such attributes from it. It reduces the no. of attributes appearing in the discovered pattern helping to make the pattern easier to understand.

Basic huristic method of attribute sub-set selection including the following technics:-

① → Step-wise forward selection

The procedure starts with an empty set of attribute the base top original attribute is determined & added to the set. At each sub-sequent step, the basetop remaining original attribute is added to the set.

$\{A_1, A_2, \ldots A_6\}$

initial set

$\{\ \}$

$\{A_1\}$ → Best

$\{A_1, A_4\}$ → Best

$\{A_1, A_4, A_6\}$

→ ② Step-wise Backward elimination

The procedure starts with full set of attributes. At each step, it removes the worst attribute remaining in the set.

$\{A_1, A_2, A_3, A_4, A_5, A_6\}$ → Initial set

$\{A_1, A_3, A_4, A_5, A_6\}$

$\{A_1, A_4, A_5, A_6\}$

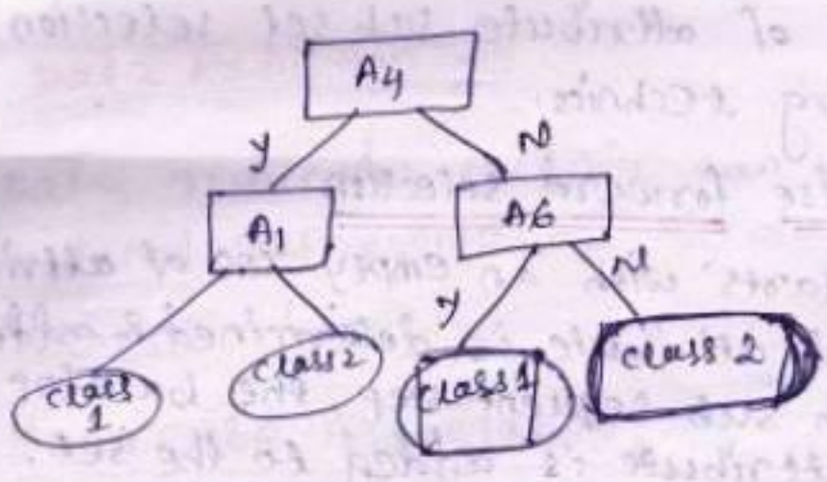
$\{A_1, A_4, A_6\}$

* ③ Combination of forward selection & Backward Elimination:-

The procedure selects the best attribute & removes the worst among the remaing attribute.

④ Decission Tree Induction:-

It is used for attribute subset selection. A tree is constructed from the given data. All attributes that donot appear in the tree assume to be irrelevant. The set of attributes appearing in tree from the reduced



Numerisity Reduction

It reduced data volume by choosing alternativ smaller form data representation.

(a) Clustering :-

In clustering they partition the data object into ~~data groups~~ objects into groups or clustores so that object in a cluster are similiar to one another & disimilar to the object in another cluster..

(b) Sampling :-

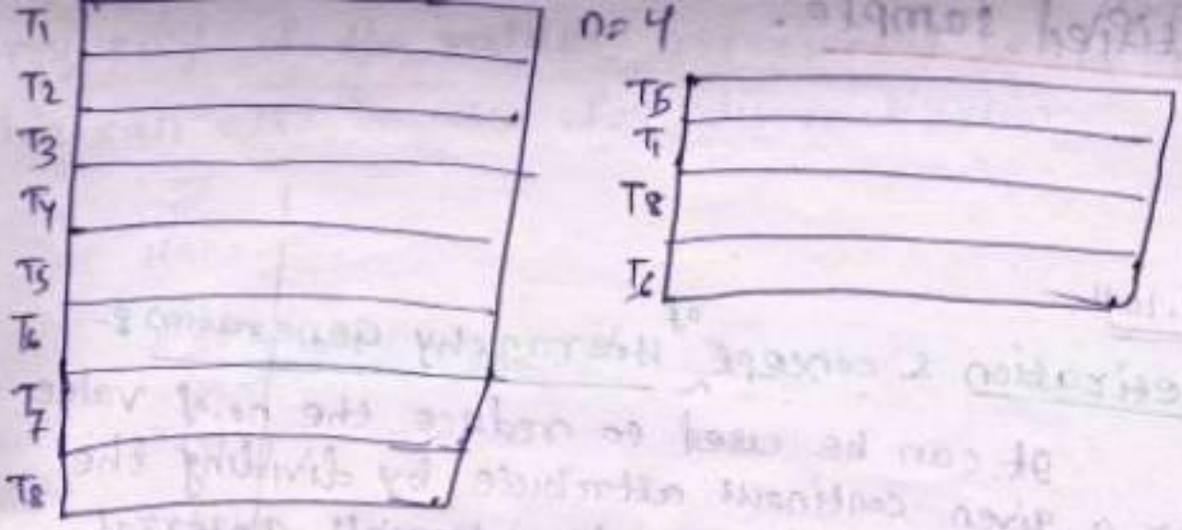
It can be used as a data reduction tech since it allows a large data cell to be represented by a smaller random sample of data.

① Simple random sample without replacement :-

Dataset = D

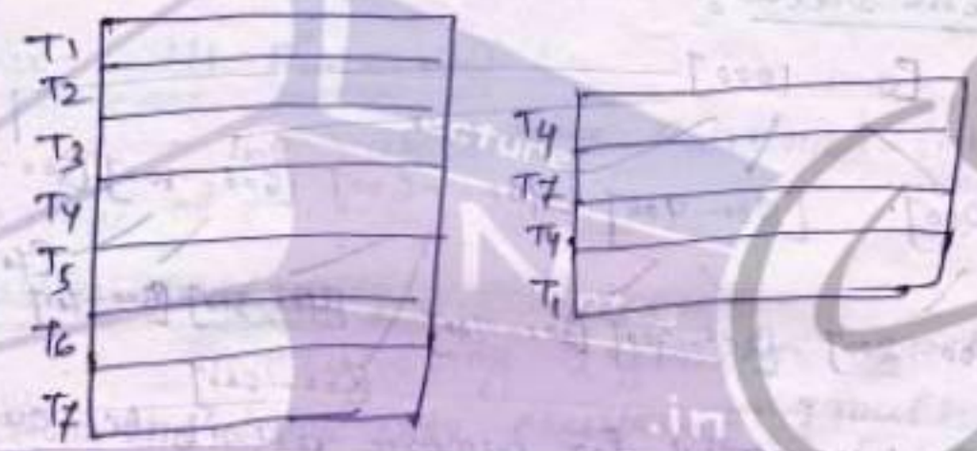
Total No. of tupples = N

This is created by drawing 'n' of the N tup from D (n < N) where the probability of drawing any tupple in D is 1/N.



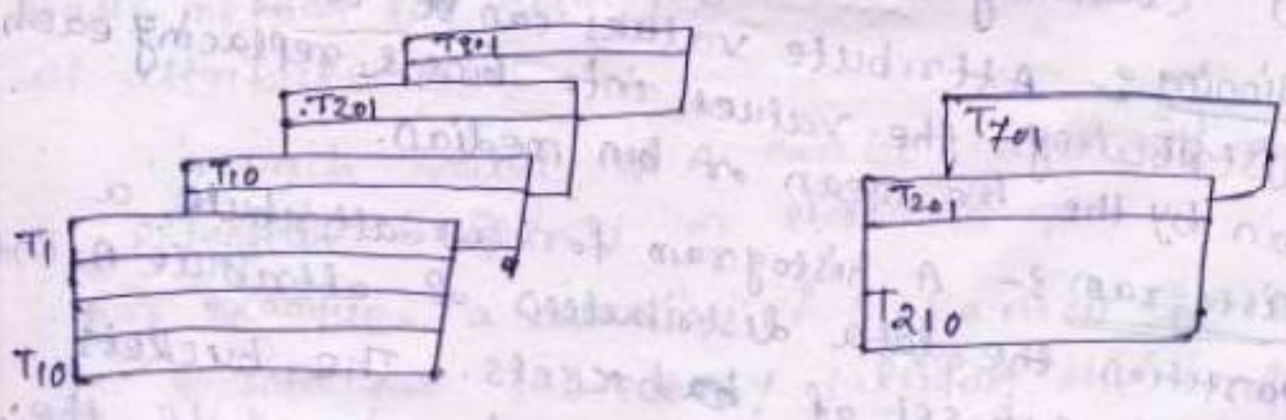
② → Simple Random sample with replacement of size 'n'.

This is similar to ~~except~~ 1'st method except each time, a tupple is drawn from D, it is recorded & then replaced ie after a tupple is drawn it is placed back in D, so that it may be drawn again.



③ Cluster sample :-

If the tupples in D are grouped into m mutually disjoint cluster, then simple random sample in cluster can be obtained, where n<M for exm Tupples in DB are usually retrieved a page at a time, so that it can be considered as a cluster.



height of bucket typically reflects the average frequency of the values represented by the bucket. This can also be used for discreatization.

Equi-width Histogram :-

The values are partioned into equal size partion of ranges.

cluster analysis :-

A clustoring algorithm can be applied to partion the data into clusters or groups. Each cluster form a node of a concept of hierarchy where all nodes are at some conceptual level.

Entropy Based Discreatization :-

An information based measure called entropy, can be used to recursively partioned the values of a numeric attribute A, given a set of data tupples 'S'.

Basic method for entropy based ~~discred~~ descretization of attribute A is :-

(i) Each value of A can be considered a potential boundary or threshold T.

For example- a value 'v' of A partion the sample in S into two sub-set by satisfing the condition

one that maximizes the information gain resulting from subsequent partitioning.

Formula for information Gain :-

$$I(S,T) = \frac{|S_1|}{|S|} * Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

Information gain | Sample | Threshold voltage

where S_1 & S_2 correspond the sample in s that satisfing the condition $A < T$ & $A \geq T$

The entropy function Ent for a given set is calculated based on class discreatization of the sample in set s.

$$\boxed{Ent(S_1) = -\sum_{i=1}^{m} P_i \log_2(P_i)}$$

where m is the no. of ~~classes~~ classes the entropy of S_1.

$P_i \rightarrow$ Probability of class i in S_1

$$\boxed{P_i = \frac{\text{The no. of samples of class } i \text{ in } S_1}{\text{Total no. of samples in } S_1}}$$

$$Ent(S_2) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

The process of determining threshold value is recursively applied to each ~~partition~~ partition obtain until some stopping critiera is made, when $\boxed{Ent(S) - I(S,T) - \delta}$

New Chapter

Minning Frequent Patterns Association & Correlation

Dt. 14/10/11

Frequent Pattern :-

Frequent pattern are patterns (item set, sub-sequences, sub-structure) that appear frequently in a data set.

Sub-sequence :-

A sub-sequence such as buying PC, then a digital camera, then a memory card. If it occurs frequently in a shopping dB, then it is a sequential pattern.

Sub-structure :-

It can reffers to diff. structural forms such as sub-graphs, sub-tree which may be combined with item-set.

Market Basket Analysis :-

A typical example of forequent item-set minning is market-basket analysis. This process analyzes customer buying habits by finding association between the diff. items, the customer place in their shopping basket.

The pattern can be represented in the form of association rule for example the customer who purchased computer, also tends to buy anti-virus s/w at the same time.

computer ⇒ Anti-virus s/w

[support = 2% & confidence 60%]

Support (A ⇒ B) :-

Holds in transcation set 'D' with support 's', where 's' is the % of transcation in set D that contain A∪B.

Confidence (A ⇒ B) :-

has confidence 'c', in the ~~in the trans~~ transcation set 'D', where 'c' is the percentage of transcation in 'D' containing A that also

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A)}{\text{Support}(A)}$$

Q:-

Transcation	Item Bought
101	milk, Bread, cookies, Juice
792	milk, Juice
1130	milk, eggs
1735	Bread, cookies, coffee

	Support	confidence
Bread ⇒ cookies	2/4	2/2
milk ⇒ Juice	2/4	2/3
cookies ⇒ coffee	1/4	1/2
milk ⇒ egg	1/4	1/3

A priori Algorithm for finding frequent Item-set

T_{ID}	List of Item-ID
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

minimum count = 2

Step-1

Itemset	Sup count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2
I_6	0
I_7	0
I_8	0
I_9	0

Step-2

Comparing with minimum count = 2

$L_4 \bowtie L_4$

Itemset	Sup count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Step-3

$L_1 \bowtie L_2$

$I_1 I_2$	4
$I_1 I_3$	4
$I_1 I_4$	1
$I_1 I_5$	2
$I_2 I_3$	4
$I_2 I_4$	2
$I_2 I_5$	2
$I_3 I_4$	0
$I_3 I_5$	1
$I_4 I_5$	0

Step-4

$I_1 I_2$	4
$I_1 I_3$	4
$I_1 I_5$	2
$I_2 I_3$	4
$I_2 I_4$	2
$I_2 I_5$	2
	4

Step-5

$I_1 I_2 I_3$	2	$I_3 I_4 I_5 \to 0$
$I_1 I_4 I_5$	0	$I_2 I_3 I_5 \to 1$.
$I_1 I_3 I_4$	0	
$I_1 I_4 I_5$	0	
$I_1 I_2 I_4$	1	
$I_1 I_2 I_5$	2	
$I_1 I_3 I_5$	1	

Step-6

$L_3 \bowtie L_3$

$I_1 I_2 I_3$	2
$I_1 I_2 I_5$	2

Lecture Notes .in
Lecturenotes.in

NULL

$I_1 I_2 I_3 I_5 - 1$

Rest are zero

Q:- 6.6

A dB has five transcations. Minimum support is & minimum confidence is 80%

TID	Item-bought
T_{100}	{M, O, N, K, E, Y}
T_{200}	{D, O, N, K, E, Y}
T_{300}	{M, A, K, E}
T_{400}	{M, U, C, K, Y}
T_{500}	{C, O, O, K, I, E}

Find all frequent item-set using A Priori alg.

Dt. 18.10.11

Apriori Algorithm :-

In general finding frequent item set from large data set is an exponential task.

The Apriori algorithm proposed by R. Agrawal, R. Srikant in 1994 makes it easier by generating cand set. It is named so because it uses prior know of frequent item-set in every iteration. It employ an iterative method where knowledge of k-items is used to determine (k+1) itemset.

Steps

All non-empty sub-sets of a frequent item-set must also be frequent. The iterative process of apriori works in two steps :-

(i) Join-step :- frequent k-item se are joined with itself to generate (k+1) item set tha are candidate for next iteration.

(ii) Prune step :- Any item-set below the minimum support count is ignore. for the next addration.
In addition only k-item set whose sub-set is not frequent item-set is also discarded.

Imp Rule for joinging

By defeault apriori assume that item in transcation are sorted in lexicographic order.
The join $L_{k-1} \bowtie L_{k-1}$ is performed where the members of L_{k-1} are joinable of their first $(k-2)$ items are same.

Generating Association Rule :-
once the frequent item set from the transcational dataset 'd' is determined, the next thing is to generate strong association rule s.t they satisfy both minimum support & min^m confidence.
Association rules are generate as follows :-
for each maximal frequent itemset 'l', generate all non-empty sub-set of l :

Ex :- $I_1 I_2 I_3$
$I_1 I_2 I_5$
subsets are $\{I_1\}, \{I_2\}, \{I_5\}$ $\{I_1 I_2\}$ $\{I_1 I_5\}$ $\{I_2 I_5\}$

For every non-empty subset s(l), the o/p of the rule $s \rightarrow l-s$, if $\frac{\text{support count of}(l)}{\text{support count of}(s)} >$ min^m count

$\Rightarrow I_1 \rightarrow \{I_1 I_2 I_5\} - \{I_1\}$
$I_1 \rightarrow \{I_2 I_5\}$
$I_2 \rightarrow \{I_1 I_5\}$
$I_5 \rightarrow \{I_1 I_2\}$
$I_1 I_2 \rightarrow I_5$
$I_1 I_5 \rightarrow I_2$
$I_2 I_5 \rightarrow I_1$

$$= \frac{\text{Support of } (I_1) \cup (I_2 I_5)}{\text{Support of } I_1}$$

$$= 2/6 = 1/3 = 33\%$$

$I_2 \rightarrow I_1 I_5 = 2/7 = 29\%$

$I_5 \rightarrow I_1 I_2 = 2/2 = 1 = 100\%$ ✓

$I_1 I_2 \rightarrow I_5 = \frac{2}{4} = 50\%$

$I_2 I_5 \rightarrow I_1 = 2/2 = 100\%$ ✓

$I_1 I_5 \rightarrow I_2 = 2/2 = 100\%$ ✓

Q:-6.6

Solution

$$\text{Support} = \frac{A \cup B}{\text{No. of transcation}} = 60\%$$

$$\Rightarrow \frac{\sup(A \cup B)}{5} = 0.6$$

$$\Rightarrow \sup(A \cup B) = 3$$

L1

M	3
N	2
O	4
K	5
E	4
Y	3
D	1
A	1
I	1
U	1
C	2

L

M	3
O	4
K	5
E	4
Y	3

L ⋈ L1

MO	1
MK	3
ME	2
MY	2
OK	3
OE	3
OY	2
KE	4
KY	3
EY	2

Lecturenotes.in

① Hash base Algorithm :-

Hashing item-set into bucket of a hash table structure using a simple hash function.

Step-1

→ First all frequent item set are generated in exactly in same manner as in apriori.

→ While generating frequent two item-set from L_1 each two item-set is mapped into a bucket of hash table structure.

→ Every time a two item-set is hashed unto a bucket the bucket count is increased.

If the bucket count is below the minimum suppo then the two item set is discarded & removed from the bucket.

$$H(x,y) = [\text{order}(x) * 10 + \text{order}(y)] \% 7$$

(For generating two item set)

Ex:- 6.5

TID	Items	order it in lexiographic order
T_{100}	I_1, I_2, I_3	$(I_1, I_2), (I_1, I_5), (I_2, I_5)$
T_{101}	I_2, I_4	I_2, I_4
T_{102}	I_2, I_3	(I_2, I_3)
T_{103}	I_1, I_2, I_4	$(I_1, I_2), (I_2, I_4), (I_1, I_4)$
T_{104}	I_1, I_3	(I_1, I_3)
T_{105}	I_2, I_3	(I_2, I_3)
T_{106}	I_1, I_3	(I_1, I_3)
T_{107}	I_1, I_2, I_3, I_5	$(I_1, I_2), (I_1, I_3), (I_1, I_5), (I_2, I_3), (I_2, I_5), (I_3, I_5)$
T_{108}	I_1, I_2, I_3	$(I_1, I_2), (I_2, I_3), (I_1, I_3)$

$H(I_1, I_2) = [1 * 10 + 2] \% 7$

$= 12 \% 7$

$= 5$

bucket no	0	1	2	3	4	5	6
bucket count		1 2	1 4	1 2	1 2	1 4	1 4
		$I_1 I_5$	$I_2 I_3$	$I_2 I_4$	$I_2 I_5$	$I_1 I_2$	I_1, I_3
		$I_1 I_5$	$I_2 I_3$	$I_2 I_4$	$I_2 I_5$	$I_1 I_2$	I_1, I_3
			$I_2 I_3$			$I_1 I_2$	I_1, I_3
			$I_2 I_3$			$I_1 I_2$	I_1, I_3

$H(I_1, I_5) = (1 * 10 + 5) \% 7$

$= 15 \% 7$

$= 1$

$H(I_2, I_5) = (2 * 10 + 5) \% 7$

$= 25 \% 7$

$= 4$

$H(I_2, I_4) = (2 * 10 + 4) \% 7$

$= 24 \% 7$

$= 3$

$H(I_2, I_3) = (2 * 10 + 3) \% 7$

$= 23 \% 7$

$= 2$

$H(I_1, I_3) = (1 * 10 + 3) \% 7$

$= 13 \% 7$

$= 6$

Lecture Notes .in
Lecturenotes.in

set in the join step.

Through the set, the set is pruned in the set

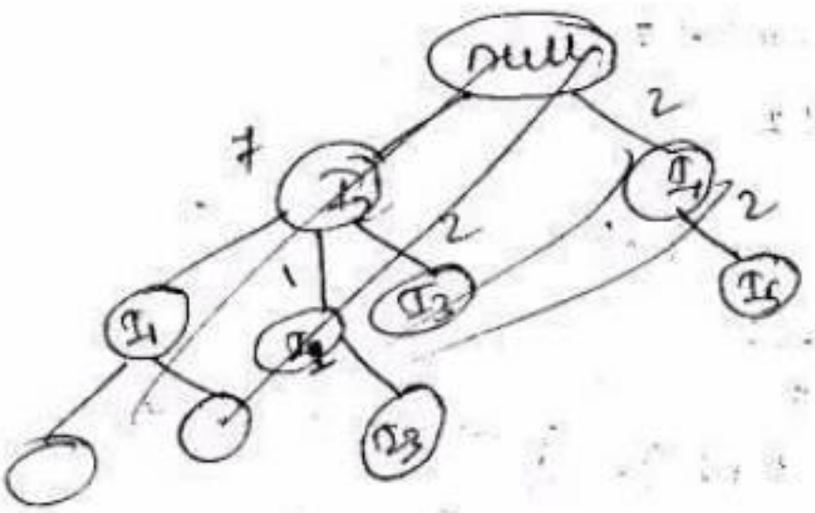
(ii) Apriori algorithm needs to scan the d repeatedly to determine the support count of the generated candidate item set.

Frequent pattern Growth Algorithm :-

Online apriori, the FP growth algorithm mine FP without generating candidate set. In this algorithm a tree is constructed & from the tree, the FP item sets are determined.

Ex:- L.5

Step-1	② Sort it	③ Item sorted in order
I_1 – 6	I_2	I_2, I_1, I_5
I_2 – 7	I_1	I_2, I_4
I_3 – 6	I_3	I_2, I_3
I_4 – 2	I_4	I_2, I_1, I_4
I_5 – 2	I_5	I_1, I_3
		I_2, I_3
		I_1, I_3
		I_2, I_1, I_3, I_5
		I_2, I_1, I_3



Start with last item in l. for ex:- all the path from the root node branches of I_5 in the FP tree.

$\langle I_2, I_1, I_5 : 1 \rangle$

$\langle I_2, I_1, I_3, I_5 : 1 \rangle$

Remove the suffix & keep only the prefix path.

$\langle I_2, I_1 : 1 \rangle$
$\langle I_2, I_1, I_3 : 1 \rangle$ } conditional Pattern base

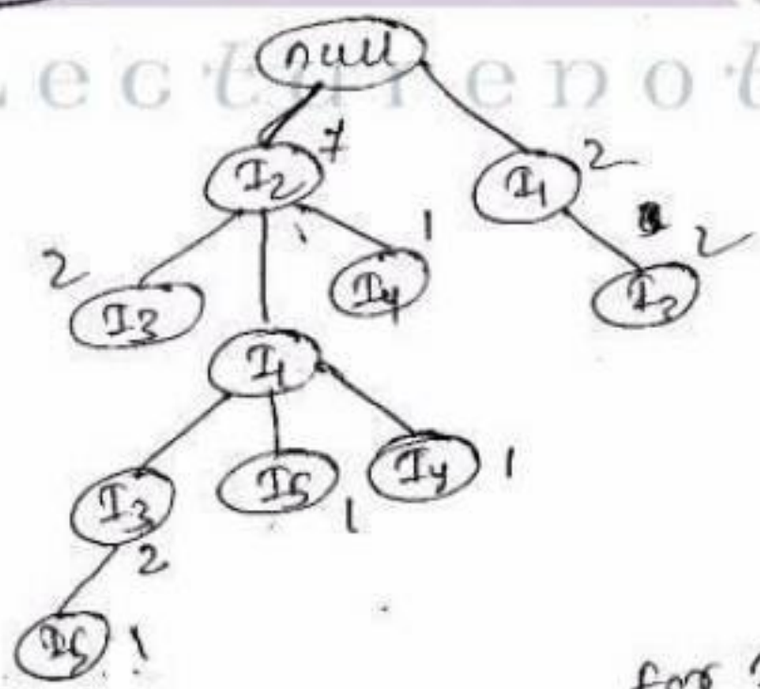
Add of prefix path & record the sum of each node.

$\langle I_2 : 2, I_1 : 2, I_3 : 1 \rangle$ → conditional FP Tree
$\langle I_2 : 2, I_1 : 2 \rangle$ [because support count = 2]

Add the suffix I_5 in all combination to generate frequent pattern.

$\langle I_2, I_5 : 2 \rangle \langle I_1, I_5 : 2 \rangle \langle I_2, I_1, I_5 : 2 \rangle$

Dt. 21.10.11



for I_4

$\langle I_2, I_4 : 1 \rangle$

$\langle I_2, I_1, I_4 : 1 \rangle$

for I_3

$\langle I_2, I_3 : 2 \rangle$

$\Rightarrow \langle I_2 : 2 \rangle$ (removing I_3)

$\langle I_1, I_3 : 2 \rangle$

$\langle I_1 : 2 \rangle$

Start with last item in l, for ex:- I_5. find out all the path from the root node to all the branches of I_5 in the FP tree.

$\langle I_2, I_1, I_5 : 1 \rangle$

$\langle I_2, I_1, I_3, I_5 : 1 \rangle$

Remove the suffix & keep only the prefix part.

$\langle I_2, I_1 : 1 \rangle$
$\langle I_2, I_1, I_3 : 1 \rangle$ } conditional Pattern base

Add of prefix path & record the sum of each node.

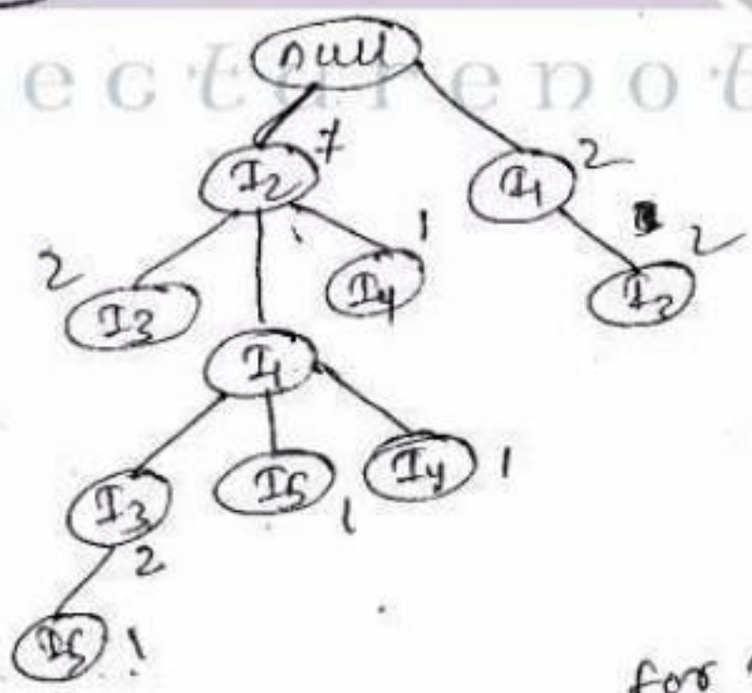
$\langle I_2 : 2, I_1 : 2, I_3 : 1 \rangle$ → conditional FP Tree

$\langle I_2 : 2, I_1 : 2 \rangle$ [because support count = 2]

Add the suffix I_5 in all combination to generate frequent pattern.

$\langle I_2, I_5 : 2 \rangle \langle I_1, I_5 : 2 \rangle \langle I_2, I_1, I_5 : 2 \rangle$

Dt. 21.10.11



for I_4

$\langle I_2, I_4 : 1 \rangle$

$\langle I_2, I_1, I_4 : 1 \rangle$

for I_3

$\langle I_2 I_3 : 2 \rangle$

$\Rightarrow \langle I_2 : 2 \rangle$ (removing I_3)

$\langle I_1, I_3 : 2 \rangle$

$\langle I_1 : 2 \rangle$

	Pattern Base	FP Tree	Pattern generate
I_5	$I_2, I_1 : 1$ $I_2, I_1, I_3 : 1$	$\{I_1 : 2, I_2 : 2\}$	$\{I_2, I_5 : 2\}$ $\{I_1, I_5 : 2\}$
I_4	$\{I_2, I_1 : 1\}$ $\{I_2 : 1\}$	$\{I_1 : 2\}$	$\{I_2, I_4 : 2\}$
I_3	$\{I_2 : 2\}$ $\{I_1 : 2\}$ $\{I_2, I_1 : 2\}$	$\{I_2 : 4\}$ $\{I_1 : 4\}$	$\{I_2, I_3 : 4\}$ $\{I_1, I_3 : 4\}$
I_1	$I_2 : 4$	$I_2 : 4$	$\{I_2, I_1 : 4\}$

Lecture Notes .in
Lecturenotes.in

Ex

min^m support count = 2

TID	Items
100	I_1, I_2, I_5
101	I_2, I_4
102	I_2, I_3
103	I_1, I_2, I_4
104	I_1, I_3
105	I_2, I_3
106	I_1, I_3
107	I_1, I_2, I_3, I_5
108	I_1, I_2, I_3

Itemset	TID
I_1	$T_{100}, I_{103}, T_{104}, T_{106}, T_{107}, T_{108}$
I_2	$T_{100}, T_{101}, T_{102}, T_{103}, T_{105}, T_{107}, T_{108}$
I_3	$T_{102}, T_{104}, T_{103}, T_{106}, T_{107}, T_{100}$
I_4	T_{101}, T_{103}
I_5	T_{100}, T_{107}

for 2-items

I_1I_2	$T_{100}, T_{103}, T_{107}, T_{108}$
I_1I_3	$T_{104}, T_{106}, T_{107}, T_{108}$
I_1I_4	T_{103} ✗
I_1I_5	T_{100}, T_{107}
I_2I_3	$T_{102}, T_{105}, T_{107}, T_{108}$
I_2I_4	T_{101}, T_{103}
I_2I_5	T_{100}, T_{107}
I_3I_4	0

Lecture Notes
Lecturenotes.in

$I_3 I_5$ T107 ×
$I_4 I_5$ 0 ×

3 Item set

: $I_1 I_2 I_3$. T107, T108
$I_1 I_2 I_5$ T100, T107

$I_1 \rightarrow I_2 I_3$

$$\frac{\text{support of } (I_1) \cup (I_2 I_3)}{\text{support of } I_1} = \frac{4}{6} = 2/3$$

$\{I_1\}, \{I_2\}, \{I_3\}, \{I_1, I_2\}, \{I_1, I_3\}, \{I_2, I_3\}$

$\{I_1\} \rightarrow \{I_1, I_2, I_3\}$

~~$I_2 \rightarrow \{I_1, I_2, I_3\}$~~

~~support~~ $s \rightarrow l - s$

$I_1 \rightarrow \{I_1, I_2, I_3\} - \{I_1\}$

$I_1 \rightarrow \{I_2, I_3\}$

$$\text{confidence} = \frac{\text{support of } (I_1) \cup (I_2 I_3)}{\text{support of } I_1}$$

$$= \frac{\text{support } (A \cup B)}{\text{support } (A)}$$

$s \rightarrow l - s$

$I_1 \rightarrow \{I_1, I_2, I_5\} - \{I_1\}$

$I_2 \Rightarrow l - s$

$\{I_1, I_2, I_3\} - \{I_2\}$

$= \{I_1, I_3\}$

$$\frac{\{I_2\} \cup \{I_1, I_3\}}{\{I_2\}} = 2/7$$

Dt. 27.10.

Correlation Analysis :-

In association rule mining we findout the strong rule by using support & confidence however strong rules are not necessarily interesting.

To findout wheather two items are truly co-related we measure by lift.

$$\text{lift}(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$$

$$= \underbrace{\frac{P(A \cup B)}{P(A)}}_{\text{confidence}} \times \frac{1}{P(B)}$$

If lift $(A,B) < 1$ then A & B are negatively co-related.

If lift $(A,B) > 1$, then A & B are positively correlated.

If lift $(A,B) = 1$, then A & B are independent.

$$\text{lift}(A,B) = \frac{P(A \cup B)}{P(A)} \times \frac{1}{P(B)}$$

$$= \frac{2/9}{2/9} \times \frac{1}{7/9}$$

$$= 9/7 = 1.29$$

I_1, I_2, I_5

$\underbrace{I_1, I_5}_{A} \rightarrow \underbrace{I_2}_{B}$

$5 \rightarrow 2-5$

$P(A \cup B) \rightarrow 2/9$

$P(A) \rightarrow 2/9$

$P(B) \rightarrow 7/9$

$I_1, I_5 \rightarrow I_2$ is truly correlated.

Lecturenotes.in

Def^n:- The data analysis task is called classification, where a model or classifier is constructed to predict class ~~lebel~~ label.

Data classification is a two-step process:-

- Learning step (where a classification model is constructed)
- classification steps (where the model is used to predict class level for a given data)

Technics for Data classification :-

Decission-Tree classifier
~~Bay~~ Bayesian classifier
Rule-Based classifier

class label attribute

name	age	income	Loan-decission
sandy	youth	Low	Risky
Bill	youth	Low	Risky
Fox	middle-age	high	safe
Rich	middle-age	Low	Risky
Susant	senior	Low	safe
gopal	senior	medium	safe
Smith	middle-age	high	safe

$\boxed{\text{classification algorithm}}$

↓ classification rules

if age = youth then loan-decission = risky
if income = high, then loan-decission = safe
if age = middle-aged and income = low then loan-decission = risky

Lecture Notes .in
Lecturenotes.in

test data

Ram senior low safe
Sita middle low risky
age

new data

John middle-age low loan decission risky

Supervised Learning :-

The class label of each training tuple is provided, this step is known as supervised learning. This is used in classification.

Unsupervised Learning :-

In which the class label of each training tuple is not known & ~~no~~ number or set of classes to be learned may not be known in advance. It is used in case of Clustering.

Steps

① → Learning step

→ Training data are analysied by classification algorithm.

② → Test data are used to estimate the accuracy of the classification rule.

→ If the accuracy is considered acceptable, the rule can be applied to the classifier of new data tuples.

Decission - Tree Induction :-

It is a flow-chart like tree structure, each internal node denotes a test on an attribute, each branch represents outcome of the test, each leaf node ~~rep~~ holds a class label. The topmost node in a tree is the

root-node. Internal nodes are denoted by rectangle.
Leaf nodes are denoted by ovals.

Importance of Decission-Tree Induction :-

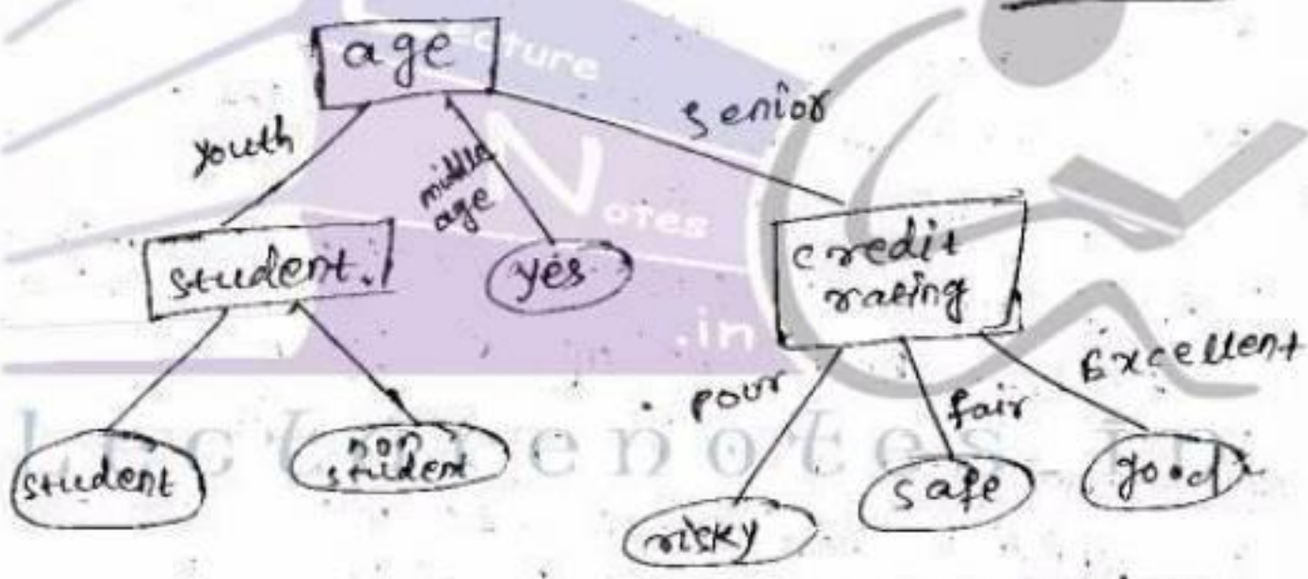
→ It doesnot require any domain knowledge or parameter setting, therefore it is appropriate for exploratory knowledge discovery.

→ It can handle multidimensional data

→ The learning & classification step of decission-tree induction are simple & fast.

→ Decission tree classifier have good accuracy.

Dt 1.11.11



First decission tree is given by Ross Quinlan in 1970-1980.

This is called ID3

2nd one is C4.5

3rd one is CART

↓ classification & Regression Tree

Algorithm

input

① Data partition D

set of training tupples with known class label

(2) Attribute list :

A set of candidate attributes

(3) Attribute selection Method

(1) A procedure to determine the spliting critir ~~the best~~ that best partion the data ~~tup~~ tup into individual classes.

The critira consists of spliting attribute.
splitting point / spliting sub-se

Method : generate - Decision - tree (D, attribut

1. create a node N

(2) If tuples of D are all of same class c the

(3) return ~~N~~ as Leaf node Labeled as c.

(4) If attribute - list = ϕ then

(5) return N as Leaf node Labeled with majority class in D.

(6) Apply attribute - selection - Method (D, attribu list) to find best spliting critira)

(7) Label node N with spliting critira

(8) If spliting - attribute is discreate value & multiway split equal to allow then
attribute - list ← attribut - list - splitting attr

(9) for each out come j of spliting critira
Dj ← set of tupples in D satisfing outcomes j

(10) If Dj = ϕ, then attach a leaf labeled with majority class in D with node N

else
attach node returned by generate decisi tree (Dj, attribute - list) to N

(11) end

(12) return N

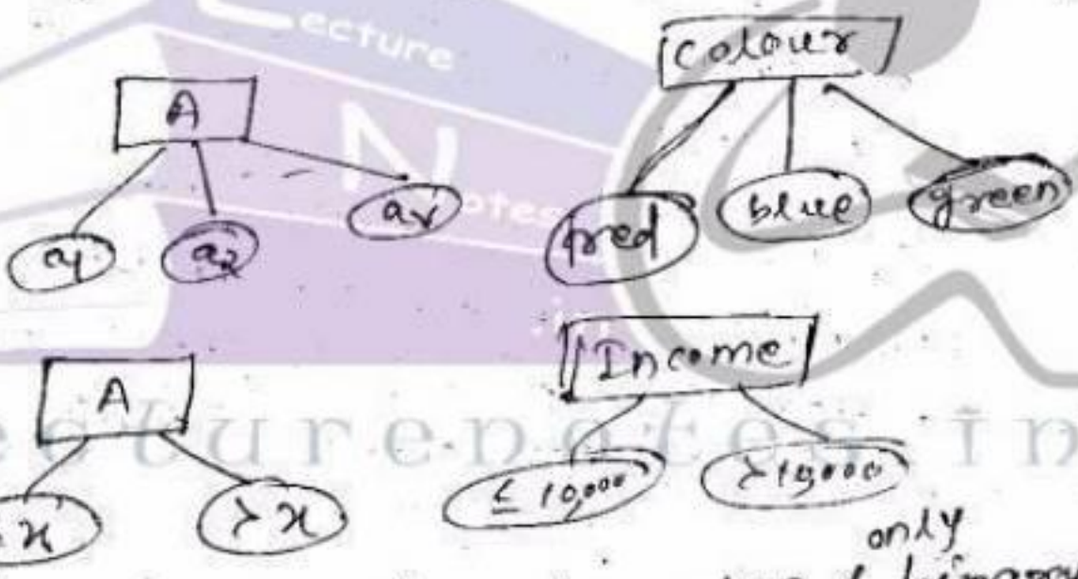
The attribute selection method determines splitting criteria i.e the splitting criteria indicates the splitting attribute & may also indicate a splitting point / subset.

The splitting criteria is determined such that the resulting partion at each branch are pure. (belongs to one class.)

Let A, B are splitting attribute & A having $\{a_1, a_2 -- a_v\}$, then are three possible

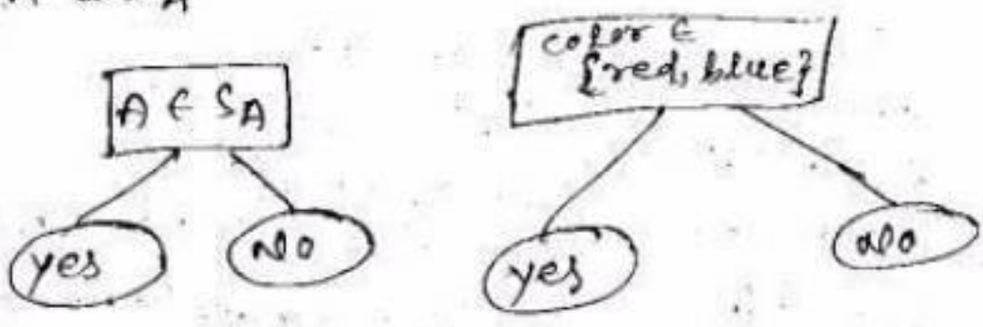
Let A, B are the splitting attribute
Attribute A has 'v' distinct attributes on the training data

$\{a_1, a_2 -- a_v\}$



If attribute A has discreate value & only binary tree is allowed the splitting sub-set is needed.

$A \in S_A$



Attribute selection measure :-

→ It is a heuristic for selecting the splitting crit that best separates the data partition D of c level training tuples into individual classes

→ Attribute selection measure is also known as spl rules, because they determine how the tuples at given node are to be split.

The popular attribute selection measure are:-

- information gain
- gain ratio
- gini index

Notation used

Let D be the data partition be a training set of class ~~tuples~~ labels tuples.

The class label has 'n' distinct values defini 'm' distinct classes.

→ ID3 uses infor^n gain as its attributes selection mea

Let node N represent the tuples of partiti D. The attributes with highest infor^n gain has choosen. The expected infor^n needed to classify a tuple in D is given by

$$info(D) = \sum_{i=1}^{m} P_i \log_2(P_i)$$

P_i is the probability ie an arbitary tuple in

$$P_i = \frac{|C_{iD}|}{|D|}$$

→ Information (D) is the avg. amount of infor^n nee to identify the class level of tuple in D.

→ info (D) is also known as entropy of D.

How much more infor^n would we still need after the partition to arrive at exact classification.

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * info(D_j)$$

where $\frac{|D_j|}{|D|}$ acts as weight of jth position.

1:-

RID	age	income	student	credit rating	buying-capacity
1	youth	high	no.	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes.	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Ans

$$info(D) = \frac{-9}{14} \log_2(9/14) - 5/14 \log_2(5/14) = +0.940$$

$$info_{age}(D) = \frac{5}{14} * (-2/5 \log_2(2/5) - 3/5 \log_2(3/5))$$

$$+ 4/14 * (-4/14 \log_2(4/4))$$

$$+ 5/14 * (-2/5 \log_2(2/5) - 3/5 \log_2(3/5))$$

$$= 0.694$$

$$gain(age) = 0.940 - 0.694$$

$$= 0.246$$

$info_{income}(D) =$

$gain(income) = 0.029$

$gain(student) = 0.151$

$gain(credit_rating) = 0.048$

② GAIN RATIO

Dt - 4.11.11

→ The information gain measure is biased towards test with many outcomes. It prefers to select attributes having large no. of values.

→ It applies a kind of normalization to information gain using split

$$\boxed{Split_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|}}$$

→ These values represents the potential information generated by splitting the data set D into v partition.

→ The attributes with max^m gain ratio is selected as the splitting attribute.

Ex:- $Split_{income}(D) =$

low = 4.

high = 4, medium = 6.

$$\boxed{\text{gain ratio}(A) = \frac{Gain(A)}{Split\ info_A(D)}}$$

$$Split_{income}(D) = -\frac{4}{14}\log_2\frac{4}{14} - \frac{4}{14}\log_2\frac{4}{14} - 6/14\log_2 6/14$$
$$= 1.557$$

$$\text{gain ratio} = 0.029/1.557 = 0.019$$

(b) GINI INDEX

It measures the impurity of D, a data partition or set of training tuples.

→ A data partition or set of training tuples as

$$\boxed{Gini(D) = 1 - \sum_{i=1}^{m} p_i^2}$$

Where p_i is the Probability that a tuple in D belongs to class C_i & is estimated as

$$\boxed{p_i = \frac{|C_{iD}|}{|D|}}$$

m is the no. of classes.

The gini index consider a binary split for each attribute.

$$\boxed{Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)}$$

For Descreate value :-

considering all subset that can be formed using known values of A. Each subset S_A.

→ Each sub-set S_A can be considered as binary test for attribute A of the form.

→ If 'A' has 'v' possible values, then there are 2^v possible subsets.

The attribute that minimize the reduc^n in impurity (min^m the gini index) is selected as splitting attributes.

$$\boxed{\Delta Gini(A) = Gini(D) - Gini_A(D)}$$

$Gini_{(income)}$

$$Gini(D) = 1 - (7/14)^2 - (5/14)^2 = 0.459$$

$Gini_{income \in \{low,\ medium\}}(D)$

$$= 10/14\, Gini(D_1) + \frac{4}{14} Gini(D_2)$$

$$= \frac{10}{14}\left\{1 - (7/10)^2 - \left(\frac{3}{10}\right)^2\right\} + \frac{4}{14}\left\{1 - (3/4)^2 - (2/4)^2\right\}$$

$$= 0.443$$

DE.8.11.11

$Gini_{income \in \{low,\ high\}\{medium\}}$

$$= \frac{8}{14} Gini(D_1) + \frac{6}{14} Gini(D_2)$$

$$= \frac{8}{14}\left\{1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2\right\} + 6/14\left\{1 - (4/6)^2 - (2/6)^2\right\}$$

$$= \frac{8}{14}(0.46) + 6/14 \times (0.444)$$

$$= 0.267 + 0.190$$

$$=$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

$$= 0.459 - 0.357$$

$$= 0.102$$

age

low, medium — high

Rule Based classification

At rule based classifier where learned model is represented as a set of if then else.

An "if then else" is an expression of the form

if(cond^) - then(cond^)

antecedent → consequence

Rule-1

If age = youth
student = yes
then buy computer = yes

The if part of the rule is known as rule antecedent/ pre condition.

The then part of the rule is known as consequence.

or if (age = youth) ∧ (student = yes)
⇒ buy_computer = yes

⇒ If the condition (all the attribute test in a rule) ancident holds true for a given tuple, we say that antecedent is satisfied (that rule is satisfied & that rule covers the tuple).

⇒ A rule can be accessed by a coverage & accuracy given a tuple yes from a class level dataset D.

⇒ $n_{correct}$ = no. of tuples correctly classified by R.

⇒ $|D|$ = no. of tuples in D.

⇒ $coverge = \frac{n_{covers}}{|D|}$

⇒ $accuracy (R) = \frac{n_{correct}}{n_{covers}}$

Where rules coverage is % of tuple that are covered by the rule i.e their attribute values hold true for their root antecendent.

→ The tuples that covers anci what % of them, the rule can correctly classified, is called

$$\text{coverage}(R_1) = \frac{n\,\text{covers}}{|D|} = \frac{2}{14} = 14.28\%$$

$$\text{accuracy}(R_1) = \frac{n\,\text{correct}}{n\,\text{covers}} = \frac{2}{2} = 100\%$$

Dt. 11.11.11

Rule Based Classification :-

① **Trigger** :- The tuple 'x' satisfies a rule R_1 can trigger the rule.

If rule R_1 is the only rule satisfied the rule fi by returning the class prediction for x.

If more than one rule is triggered, if we have potential problem.

② What if they each satisfied a different class.

(3) What if, no rule is satisfied by x.

Solution for

① There are many possible startagies :-

(i) size ordering

(ii) Rule ordering

It assignes the highest priority to the triggering rule that has the toughest requirement, where toughness is measured by the rule antecident size. ie triggering rule will be the most attribute test is fire.

Ex:- $R_1 : (a, b, c, d)$ ← 4 attributes

$R_2 : (a, c, d)$ ← 3 attributes

Rule ordering

This sheme priorities the rules, the ordering may be

1) class based
2) Rule based

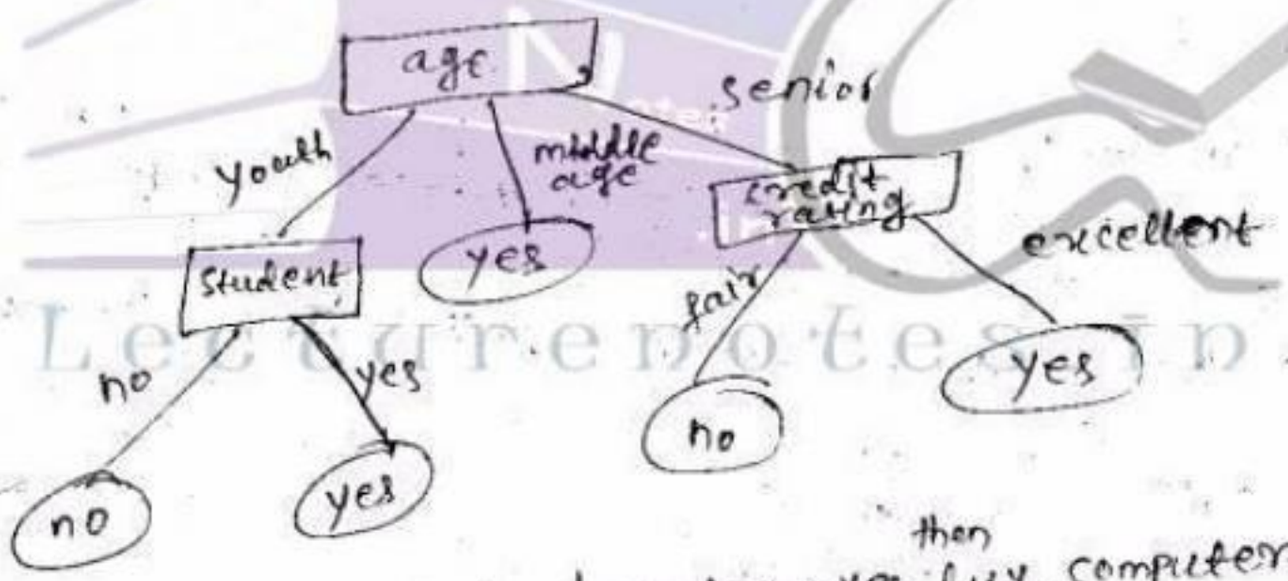
Class based

The classes are sorted in order of decreasing i.e. all the rules for the most ~~relevant~~ ~~Prevalent~~ Prelevant (or most frequent) ~~class~~ class come first, the rules for the next Prelevant class come next & so on.

Rule based ordering

The rules are organized into one long priority list according to the ~~some~~ some rule quality such as accuracy, coverage or size (no. of attribute test in the rule ancident) or based on advised from domain export.

Extracting classification Rule from the decission Tree:-



If age = youth and student = yes then buy_computer = yes
If age = youth and student = no buy_computer = No
If age = middle age buy_computer = yes
If age = senior and credit_rating = fair
buy_computer = no
If age = senior and credit_rating = excellent
buy_computer = yes

then we have to find the accuracy & coverage.

Clustering :- It is the process of grouping a data objects into multiple groups or clusters so object within a cluster have high ~~similitra~~ similarity but very disimilar to object in other cluster.

Disimilarity & similarity are accessed based on attribute values describing the object & often involved distance. ~~nature~~ ~~method~~ measure.

Clustering techniques are organized into following catagories:-

Partioning method
Hierarchical method
Density based method
Grid based method

Requirement of clustering Analysis :-

① Scalability :- A clustering algorithm should be able to work with a large dataset equally well as with small data set.

~~Ability~~ Ability to deal with diff. types of attribute

Attributes are:-
numerical
catagorical
ordinal (mixture of)

② Catagorical :- ~~Ex~~ The values donot have any ~~mining~~ meaningful order.

Ex:- Hair color

Dt. 15.11.11

i) Binary Attribute

Either 0 or 1
True or false.

Ex:- sex - male female

(ii) Ordinal Attribute

It is an attribute with possible values that have a meaningfull order or ranking among them.

③ Discovery of cluster with Arbitary shape :-

→ Algorithm based on such distant measure tend to find sperical cluster with similar size & density. It is important to develop algorithm that can detect cluster of arbitary shape.

④ Requirement for domain knowledge to determine i/p parameter.

Requiring the specification of domain knowledge not only burden user but also makes the quality of measuring difficult to control.

⑤ Ability to deal with Noisy data

Most real world data set contain out layers missing or unknown data.
Clustering algorithm can be sensitive to such noise & may produce co-quality clusters. So we need clustering method that are robust.

⑥ Incremental clustoring & insensitivity to I/p order

A good algo should be able to handle incremental clustoring ie ability to cluster newly inserted data or updated data.

⑦ High Diamensionality :-

Real world DB or DW can contain data with a large no. of attribute or diamension.
A good Algorithm should be able to handle high diamensional data.

Real world application may need to perform clust under various kind of ~~clustering~~ constraint.

A good ~~task~~ challenging task is find data groups w good clustering behaviour that satisfy specified constraint.

⑦ Interpretability & usability

clustering results to be interpretable, comprehensible & useable.

Dt 17.11.11

Measuring Data Similarity & Data Dissimilarity :-

Similarity measure for two objects :-

If i & j will return 0 if objects are unlike.

If higher the similar value, the greater the similarity between objects.

Dissimilarity measure for two objects :-

If it returns a value of 0, if the objects are same.

If higher the disimilar ~~two~~ value, the more disimilar two objects are.

Two ~~Architec~~ D.S are used for measurment :-

① Data matrix

② Dissimilarity Matrix

Data Matrix :-

This structure store the n data objects in the form of a relational table.

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity Matrix :-

Lower Trangular matrix

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1), & d(3,2) & 0 & \\ \vdots & & & \\ d(n,1) & \cdots & & 0 \end{bmatrix}$$

where $d(i,j)$ is the diff. between the objects of i & j.
If $d(i,j) = 0$, when obj i & j are highly similar & becomes larger, the more they differ.

Proximity measure of Binary Attribute :-

Binary Attribute catagorized into two types:-

- Symmetric
- Asymmetric

Ex:- Sex (symmetric)

m f (both the o/p is important)

Asymmetric

HIV

HIV+ HIV-

Dissimilarity between two binary Attribute :-

object J

	1	0	sum
1	q	r	q+r
object i 0	s	t	s+t
sum	q+s	r+t	p → total sum

2×2

where q is the no. of attribute = 1 for object i & j.
'r' is the no. of attribute = 1 for object i & = 0 for object j.

Dissimilarity betⁿ i & j if they are ~~Antisymmetric~~ Symmetric

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

Disimilarity for assymetric

$$d(i,j) = \frac{r+s}{q+r+s}$$

Ex:-

Name	gender	fever	Cough	test1	test2	test3	test4
Jack	M	Y 1	N 0	P 1	N 0	N 0	N 0
Jim	M	Y 1	Y 1	N 0	N 0	N 0	N 0
Mary	F	Y 1	N 0	P 1	N 0	P 1	N 0
Bond	M	Y 1	N 0	P 1	N 0	N 0	N 0

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$d(\text{jack}, \text{jim}) = \frac{1+1}{1+1+1} = 2/3$$
$$= 0.67$$

for yes & ~~yes~~ positive put 1 all other are 0

$$d(\text{jim}, \text{mary}) = \frac{1+2}{1+1+2} = 3/4 = 0.75 \text{ max}^m$$

$$d(\text{jack}, \text{mary}) = \frac{0+1}{2+0+1} = 1/3 = 0.33$$

If the value is max^m, ~~they~~ ~~both have~~ (jim, mary) ~~both have~~ unlikely to have common ~~(common)~~ disease, becoz they have ~~higher~~ higest disimilarity value

$d(\text{marry}, \text{bond}) = \frac{1+0}{1+1}$

$d(\text{jack}, \text{bond}) = \frac{0+0}{1+0+0} = 0$

If the value is 0, then also ~~the~~ both have common disease.

<u>Ans</u>
d(jim, mary) & d(jack, bond) have the common disease.

Dissimilarity of Numeric Data :-

We can findout the dissimilarity of data by using :-

1. Euclidean distance
2. Manhattan "

Euclidean distance

$i = x_{i1}, x_{i2} \ldots x_{ip}$

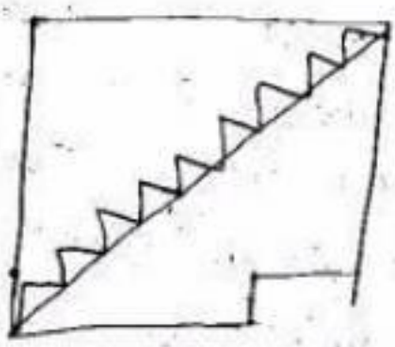
$j = x_{j1}, x_{j2} \ldots x_{jp}$

$$\boxed{d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots (x_{ip} - x_{jp})^2}}$$

Manhattan Distance

It is also called city block distance.
It is the distance in blocks between two points in a city.

$$\boxed{d_{ij} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots |x_{ip} - x_{jp}|}$$



Ex:-

Let $x_1 = (1,2)$, $x_2 = (3,5)$

find Euclidean & manhattan distance.

Ans:- for Euclidean

① $d_{ij} = \sqrt{(1-3)^2 + (2-5)^2}$

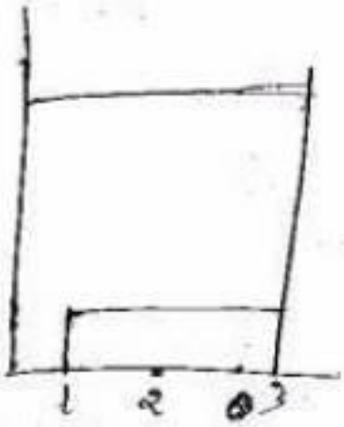
$= \sqrt{4+9} = \sqrt{13}$

② for manhattan

$d_{ij} = |1-3| + |2-5|$

$= 2+3$

$= 5$



(Manhattan)

(Euclidean)

Proximity measure of ordinal Attribute Dt. 18. 11. 11

Ex:- small, medium, large for a size of attribute.

Let M represent the no. of possible states, that an ordinal attribute can have.

$M = 3$

These states define the ranking from 1 to M_f.

Suppose 'f' is an attribute from a set of ordinal attributes describing 'n' objects

The dissimilarity computation wrt 'f' involves following step:-

The value of f for the object is x_{if} & 'f' has M_f order state representing ranking from 1 to M_f.

Lecture Notes
Lecturenotes.in

Replace each r_{if} by its corresponding rank

$r_{if} \in \{1, \ldots, m_f\}$

② Since each ordinal attribute can have diff. no. of state, it is often necessary to map the range of each attribute unto $[0.0, 0.1]$, so that each attribute has equal weight. We perform such that data normalization by replacing the rank r_{if} of the ith object in the fth attribut by

$$\boxed{z_{if} = \frac{r_{if} - 1}{m_f - 1}}$$

③ Dissimilarity can be computed using any of the distance measure (as in the numerical attribute) using z_{if} to represent the f value for the ith object.

Object Identifier	test-1
1	excellent
2	fair
3	good
4	excellent

No. of states, $m_f = 3$

fair - rank-1
good rank-2
excellent - rank-3

for fair

$$z_{1f} = \frac{r_{if} - 1}{m_f - 1} = \frac{1-1}{3-1} = 0$$

for good

$$z_{2f} = \frac{2-1}{3-1} = 1/2$$

for excellent

$$z_{3f} = \frac{3-1}{3-1} = 1$$

Lecturenotes.in

$$\begin{bmatrix} 1 & & & \\ 0.5 & 0.5 & & \\ 0 & 0.1 & 0.5 & 0 \end{bmatrix}$$

$d_{21} = \sqrt{(0-1)^2} = 1$

$d_{31} = \sqrt{(0.5-1)^2} = \sqrt{(0.5)^2}$
$= 0.5$

$d_{32} = \sqrt{(0.5-1)^2} = \sqrt{(0.5)^2}$

$d_{41} = \sqrt{(0)^2}$

$d_{42} = \sqrt{(1-0)^2} = 1$

$d_{43} = \sqrt{(0.5-1)^2} = 0.5$

Proximity measure of nominal Attribute:-

Ex:- Color

Let the no. of states of a nominal attribute be 'N'. The dissimilarity between two objects i & j can be computed as $\boxed{d(i,j) = \frac{p-m}{p}}$, where

p = Total no. of attributes describing the object

m = no. of matches (no. of attributes for i & j are same)

Ex:-

object	Test1 (nominal)
1	code A
2	code B
3	code C
4	code A

solu:-

$p = 1$

$d_{21} = \frac{p-m}{p} = \frac{1-0}{1} = 1$

$d_{31} = \frac{1-0}{1} = 1$

$d_{32} = \frac{1-0}{1} = 1$

$d_{41} = \frac{1-1}{1} = 0$

$d_{42} = \frac{1-0}{1} = 1$

$d_{43} = \frac{1-0}{1} = 1$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Mixed Attribute :-

Suppose that, the data set contains p attributes of mixed type.

Dissimilarity $d(i,j)$ betn objects i & j is defined as

$$\boxed{d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}(f)\, d_{ij}(f)}{\sum_{f=1}^{p} \delta_{ij}(f)}}$$

$\delta_{ij}(f) = 0$, if x_{if} or x_{jf} is missing (i.e. there is no measurment of attribute 'f' for object i or object j)

indicator

$x_{if} = x_{jf} = 0$ & 'f' is asymmetric binary, otherwise $\delta_{ij}(f) = 1$.

① If 'f' is numeric $\boxed{d_{ij}(f) = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}}$

where 'h' runs over all normalising object for attribute 'f'.

② If 'f' is nominal or binary $d_{ij}(f) = 0$, if $x_{if} = x_{jf}$, otherwise $d(i,j)(f) = 1$

If attribute is ordinal, compute the rank r_{if} & find $\boxed{z_{if} = \frac{r_{if} - 1}{M_f - 1}}$

object	Test1 (nominal)	Test2 (ordinal)	Test3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Solu:-

for numeric

$$d_{ij}(f) = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$

$$d_{21} = \frac{|22-45|}{64-22} = 0.55$$

$$d_{31} = \frac{|64-45|}{64-22} = 0.45$$

$$d_{32} = \frac{|64-22|}{64-22} = 1.0$$

$$d_{41} = \frac{|28-45|}{64-22} = \frac{17}{42} = 0.40$$

$$d_{42} = \frac{|28-22|}{64-22} = 0.14$$

$$d_{43} = \frac{|28-64|}{64-22} = \frac{36}{42} = 0.86$$

numeric

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.0 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

nominal

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

ordinal

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Lecture Notes
lecturenotes.in

Dissimilarity

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}(f)\, d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}(f)}$$

$$d(2,1) = \frac{1*1 + 1*1.0 + 0.55*1}{1+1+1} = 0.85$$

$$d(3,1) = \frac{1*1 + 0.5*1 + 0.45*1}{1+1+1} = 0.65$$

$$d(3,2) = \frac{1*1 + 0.5*1 + 1*1}{1+1+1} = 0.83$$

$$d(4,1) = \frac{0*1 + 0*1 + 0.40*1}{1+1+1} = 0.13$$

$$d(4,2) = \frac{1*1 + 1*1 + 0.14*1}{1+1+1} = 0.71$$

$$d(4,3) = \frac{1*1 + 0.5*1 + 0.86*1}{1+1+1} = 0.79$$

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

Clustering Algorithm :-

clustering method can be classified into following catagories :-

1. Partioning Method
2. Hierarchical Method
3. Density based Method
4. Grid based Method

Lecture Notes .in
Lecturenotes.in

Given a set of 'n' objects, a partitioning method construct K ~~partion~~ partation of the data where each partation represents a cluster where $K \leq n$ i.e it divides the data into 'K' groups such that each group must contain atleast one object.

Some algorithms are:-

① K-means algorithm
↙ no. of cluster

② K-mediod

K-means Algorithm:-

Step-1 gt randomly select K of object in D, each of which initially represents a cluster mean or center.

Step-2

For each remaining object, an object is assigned to the cluster to which it is most similar based on ~~Eclidian~~ Euclidean distance between the object & cluster-mean.

Step-3

For each cluster it computes the new mean using the objects assign to the cluster in the previous iteration.

④ All the objects are reassigned using the updated mean at the new cluster center.

⑤ The iteration continue untill the assignment is stable i.e the cluster formed in the current round are same as they formed in the previous round

Q:-

	Object	attribute X	attribute Y
medicine	A	1	1
"	B	2	1
"	C	4	3
"	D	5	4

K=2 (Given)

Solu^n :-

Step-1 $c_1 = (1,1)$

$c_2 = (2,1)$

find Euclidean distance for A $= \sqrt{(1-1)^2 + (1-1)^2}$ $\rightarrow c_1$

$= 0$

~~$c_2 = 2$~~

for center $c_2 = \sqrt{(1-2)^2 + (1-1)^2}$

$= 1$

for B

$\sqrt{(2-1)^2 + (1-1)^2} = 1$

$\sqrt{(2-2)^2 + (1-1)^2} = 0$

for C

$\sqrt{(4-1)^2 + (3-1)^2} = \sqrt{(3)^2 + (2)^2}$

$= 3.6$

$\sqrt{(4-2)^2 + (3-1)^2} = 2.82$

for D

$\sqrt{(5-1)^2 + (4-1)^2} = 5$

$\sqrt{(5-2)^2 + (4-1)^2} = 4.24$

$$D^0 = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1 & 3.6 & 5 \\ 1 & 0 & 2.82 & 4.24 \end{bmatrix} \end{array}$$

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$ for min^m put 1 & max^m put 0

$c_1 = (1,1)$

$c_2 = \left(\frac{2+4+5}{3} , \frac{1+3+4}{3} \right)$

$= (3.6 , 2.6)$

Lecturenotes.in

$C_2 = (11/3, 8/3)$

<u>For A</u>

$$\sqrt{(1-1)^2 + (1-1)^2} = 0$$

$$\sqrt{(1-11/3)^2 + (1-8/3)^2} = \sqrt{2.56 + 6.}$$

$$= 3.14$$

<u>For B</u>

$$\sqrt{(2-1)^2 + (1-1)^2} = 1$$

$$\sqrt{(2-3.67)^2 + (1-2.6)^2} = \sqrt{2.56 + 2.56} = 2.36$$

<u>For C</u>

$$\sqrt{(4-1)^2 + (3-1)^2} = 3.6$$

$$\sqrt{(4-3.67)^2 + (3-2.67)^2} = \sqrt{0.1089 + 0.1089}$$

$$= 0.47$$

<u>For D</u>

$$\sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$\sqrt{(5-3.67)^2 + (4-3.67)^2} = \sqrt{1.7689 + 1.7689}$$

$$= 1.89$$

$$D' = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix}$$

$$G' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$C_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (3/2, 1)$$

$$C_2 = (9/2, 7/2)$$

for A

$$\sqrt{(1-3/2)^2+(1-1)^2} = \sqrt{0.25} = 0.5$$

$$\sqrt{(1-9/2)^2+(1-7/2)^2} = \sqrt{12.25+6.25} = 4.30$$

for B

$$\sqrt{(2-1.5)^2+(1-1)^2} = \sqrt{0.25} = 0.5$$

$$\sqrt{(2-4.5)^2+(1-3.5)^2} = \sqrt{6.25+6.25} = 3.53$$

for C

$$\sqrt{(4-1.5)^2+(3-1)^2} = \sqrt{6.25+4} = 3.20$$

$$\sqrt{(4-4.5)^2+(3-3.5)^2} = \sqrt{0.25+0.25} = 0.70$$

for D

$$\sqrt{(5-1.5)^2+(4-1)^2} = \sqrt{12.25+9} = 4.60$$

$$\sqrt{(5-4.5)^2+(4-3.5)^2} = \sqrt{0.25+0.25} = 0.70$$

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.60 \\ 4.30 & 3.53 & 0.70 & 0.70 \end{bmatrix}$$

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

If similar results came, then we stop.

Lecture Notes .in

Input

K: the no. of cluster

D: a data set containing no objects

Step-1

Arbitarily choose K objects from D as initial cluster contains.

Step-2

repeat

Step-3

Reassign each object as the cluster to which the object is the most similar based on the mean value of the objects in the cluster.

Step-4

update the cluster mean that is cluster the mean value of the objects for each cluster.

Step-5

until no change.

Disadvantages

Depending upon outlayer value cluster center changes / distucts.

The k-mean algorithm is sensitive to outlayer be such object are far away from the majority of data.

Dt. 25.11.11

How to measure Error?

Quality of the cluster of C_i can be measured with cluster variation.

$$E = \sum_{i=1}^{K} \sum_{P \in C_i} dist(P, C_i)$$

where, E is the sum of squared error for objects the data set D.

p_i is the point in the space representing a given [illegible]
& c_i is the centroid of cluster C_i.

Ex:-

1, 2, 3. 8 9 10 25

mean :- $\frac{1+2+3}{3} = 2$ → one cluster

$\frac{8+9+10+25}{4} = 13$ → another cluster

$$E = (1-2)^2 + (2-2)^2 + (3-2)^2 + (8-13)^2 + (9-13)^2 + (10-13)^2 + (25-13)^2$$

$$= 196$$

1, 2, 3, 8 9, 10, 25

one cluster

$$c_1 = \frac{1+2+3+8}{4} = 3.5$$

$$c_2 = \frac{9+10+25}{3} = 14.66$$

$$E = (1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (8-3.5)^2 + (9-14.66)^2 + (10-14.66)^2 + (25-14.66)^2$$

$$= 189.66$$ → Ans

If the error is minimum clusters are faraway from each other.

K-medoid Algorithm:-

A K medoid is a classical partitioning technic of clustering that clusters the data set of n-object into k- cluster.

A midoid can be defined as it is a most centrally located point in the cluster.

Algorithm

Input : K: is the no. of clusters
D: a data set containing n objects

Step-1
Arbitarily choose k-objects in D as initial representative objects.

Step-2
repeat

Step-3
Associate each data point to the closest medoid (closest here is using valid distance method)

Step-4
Randomly select a nonrepresentative object o_{random},

Step-5
Compute the total cost S of swapping representative object o_j with o_{random}.

Step-6
If $S < 0$, then swap o_j with o_{random} to form the new set of k representative object

Step-7
until unchanged

Ex:- Data set: D

x_1	2	6
x_2	3	4
x_3	3	8
x_4	4	7
x_5	6	2
x_6	6	4
x_7	7	3
x_8	7	4
x_9	8	5
x_{10}	7	6

$K = 2$

Solu:-

Given $k=2$

$c_1 = (3,4)$

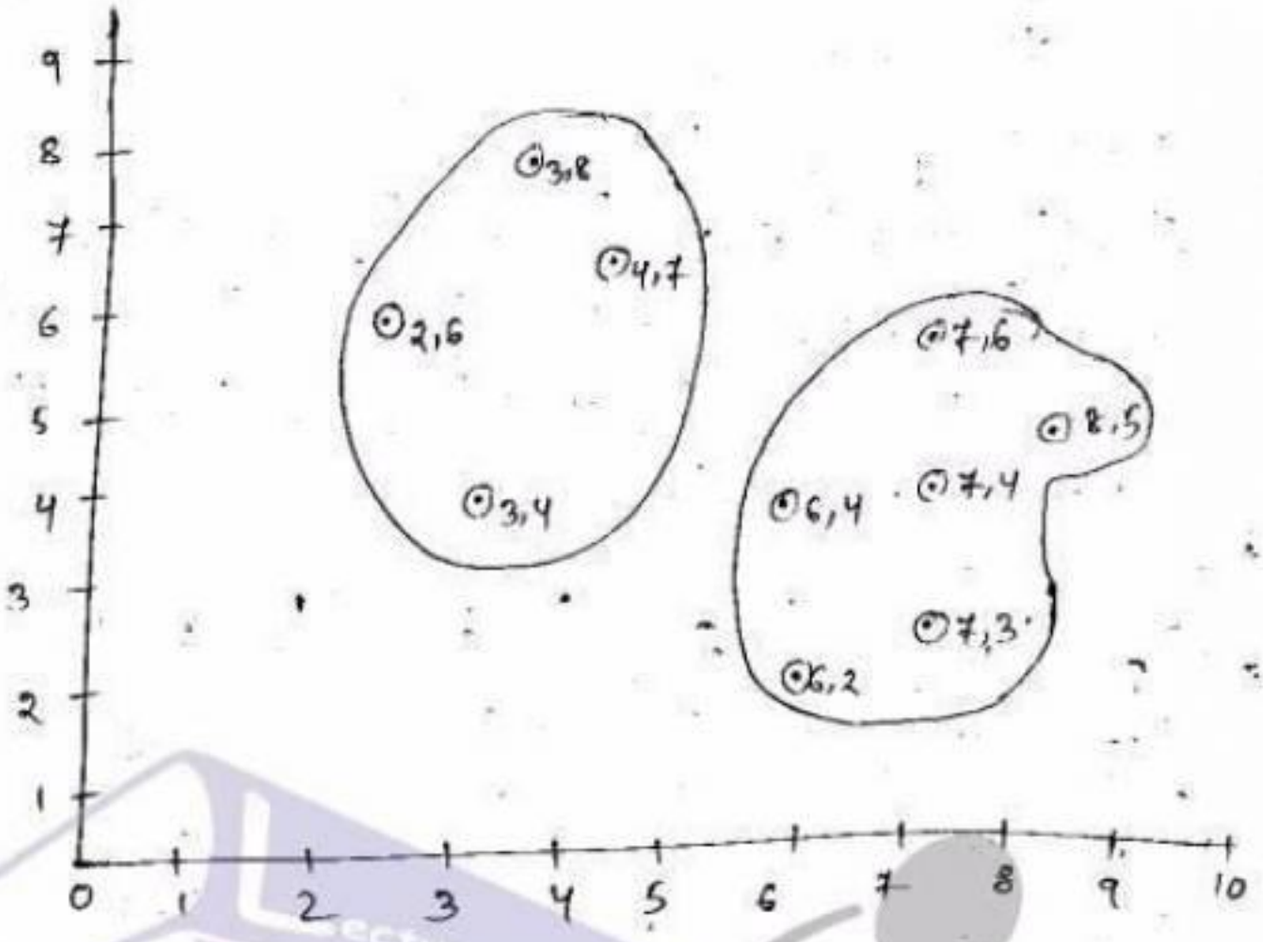
$c_2 = (7,4)$

considering $c_1(3,4)$

$$\sqrt{(3-2)^2+(4-6)^2+(3-3)^2+(4-8)^2+(3-4)^2+(4-7)^2+(3-6)^2+(4-2)^2+(3-4)^2+(4-4)^2+(3-7)^2+(4-3)^2+(3-7)^2+(4-4)^2+(3-8)^2+(4-5)^2+(3-7)^2+(4-6)^2}$$

3	4	2	6	2.23
3	4	3	8	4.0
3	4	4	7	3.16
3	4	6	2	3.60
3	4	6	4	3.00
3	4	7	3	4.12
3	4	8	5	5.09
3	4	7	6	4.47
3	4			
3	4			
3	4			

7	4	2	6	5.38
7	4	3	8	5.65
7	4	4	7	4.24
7	4	6	2	2.23
7	4	6	4	1.0
7	4	7	3	1.0
7	4	8	5	1.41



<u>find the cost</u>

$$\text{cost}(x,c) = \sum_{i=1}^{d} |x-c|$$

where $x \rightarrow$ data object
$c \rightarrow$ midoid
$d \rightarrow$ dimension

<u>cluster-1</u>

(3,4) (2,6) (3,8) (4,7)

<u>cluster-2</u>

(7,4) (6,2) (6,4) (8,5) (7,6) (7,3)

$\text{cost}\{(3,4)(2,6)\} + \text{cost}\{(3,4)(3,8)\}$
$+ \text{cos}\{(3,4)(4,7)\} + \text{cost}\{(7,4)(6,2)\}$
$+ \text{cost}\{(7,4)(6,4)\} + \text{cost}\{(7,4)(8,5)\}$
$+ \text{cost}\{(7,4)(7,6)\} + \text{cost}\{(7,4)(7,3)\}$

[x − c means
(3−2) + (4−6)

$= (3+4+4) + (3+1+2+2+1)$

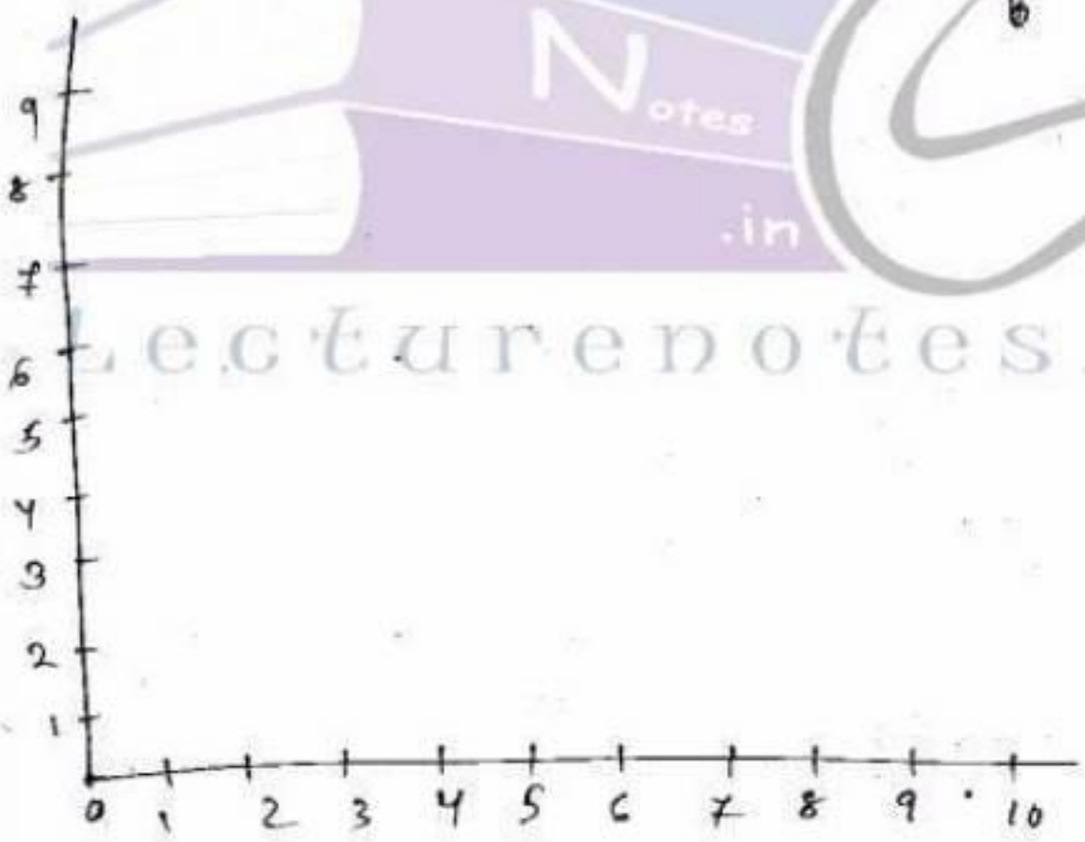
$= 20$

$c_1 = (3,4)$

orandom $= (7,3)$

7	3	2	6	5.83
7	3	3	4	4.12
7	3	3	8	6.40
7	3	4	7	5.0
7	3	6	2	1.41
7	3	6	4	1.41
7	3	7	4	1
7	3	8	5	2.23
7	3	7	6	3.0

→ not required



1 1 1 1 0

0 0 0 0 1

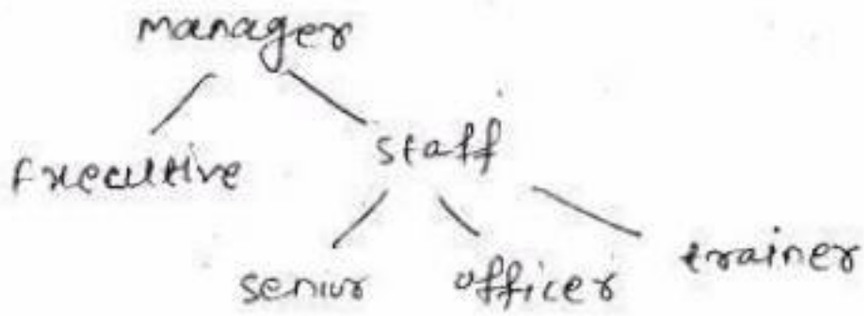
1 1 1 0 0 0 0 0

0 0 0 1 1 1 1 1

Lecture Notes .in
Lecturenotes.in

It works by grouping data object into hierarchy or tree of cluster. Representing data object in the form of a hierarchy is useful for data summerization & visualisation.

Ex:-



There are several ways to catagorise hierarchical clustering method :-

① Agglomerative method
② Divisive method

Agglomerative Method:-

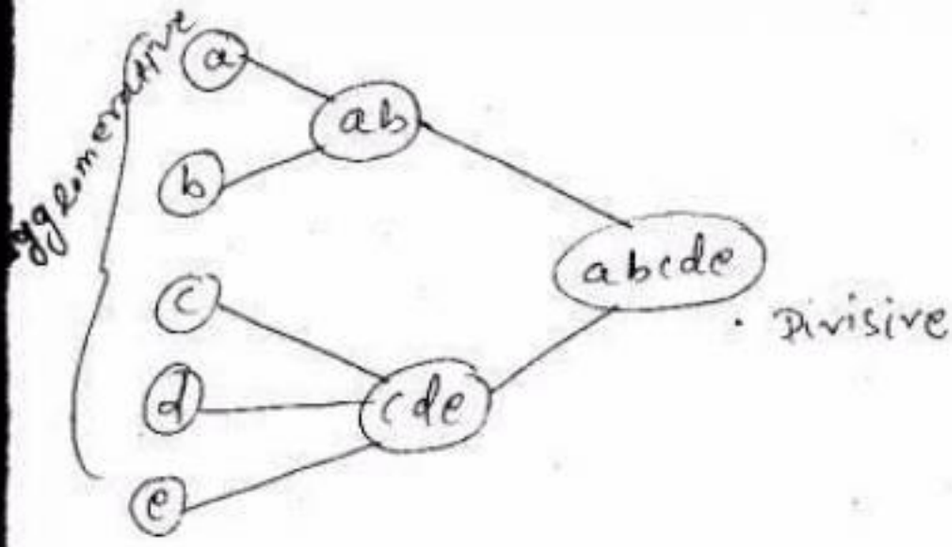
→ It is a bottom-up approach where we start from the bottom in which all the objects are ~~group~~ going up through merging of object.

→ We begin with each individual object & merge two closest object.

This process is iterated until all the objects are aggregated into a single group.

Divisive Method:-

We start with assumption that all objects are group into a single group, then we ~~split~~ split the group into two recursively until each group consists of single object.



Algorithm (Agglomerative)

Step-1

convert object feature to distance matrix.

Step-2

set ~~distance matrix~~ each object as a cluster.

Step-3

Iterate until the no. of cluster is '1'.

(i) Merge two closest cluster, ~~update the~~

(ii) update the distance matrix

Ex:-

	x_1	x_2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

$d_{ij} = (x_{ik} - x_{jk})^2$

Solution

	A	B	C	D	E	F
A	0.00	0.71	5.66	3.60	~~3.20~~ 4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.0	0.5
E	4.24	3.54	1.41	1.00	0.00	1.12

0.5 is min^m

Lecturenotes.in

	A	B	C	(D,F)	E
A	0	0.71	5.66	?	4.24
B	0.71	0	4.95	?	3.54
C	5.66	4.95	0	?	1.41
D,F	?	?	?	0	?
E	4.24	3.54	1.41	?	0

How objects should be grouped into cluster?

<u>Different methods</u> :-

① single linkage (min^m distance critiera)

$$\text{dist}_{min}(c_i, c_j) = \min(|p - p'|) \quad p \in c_i, p' \in c_j$$

② complete linkage (max^m distance critiera)

$$\text{dist}_{max}(c_i, c_j) = \max(|p - p'|) \quad p \in c_i, p' \in$$

③ mean Distance

$$\text{dist}_{mean}(c_i, c_j) = |m_i - m_j|$$

④ Average Distance

$$\text{dist}_{avg}(c_i, c_j) = \frac{1}{n_i n_j} \sum_{p \in c_i, p \in c_j} |p - p'|$$

$$d_{(D,F) \to A} = \min(d_{DA}, d_{FA}) = \min(3.61, 3.20) = 3.20$$

$$d_{(D,F) \to B} = 2.5$$

$$d_{(D,F) \to C} = 2.24$$

which min^m we take i

	A	B	C	D,F	E
A	0	6.71	5.66	3.20	4.24
B	6.71	0	4.95	2.5	3.54
C	5.66	4.95	0	2.24	1.41
D,F	3.20	2.5	2.24	0	1.00
E	4.24	3.54	1.41	$\boxed{1.00}$	0

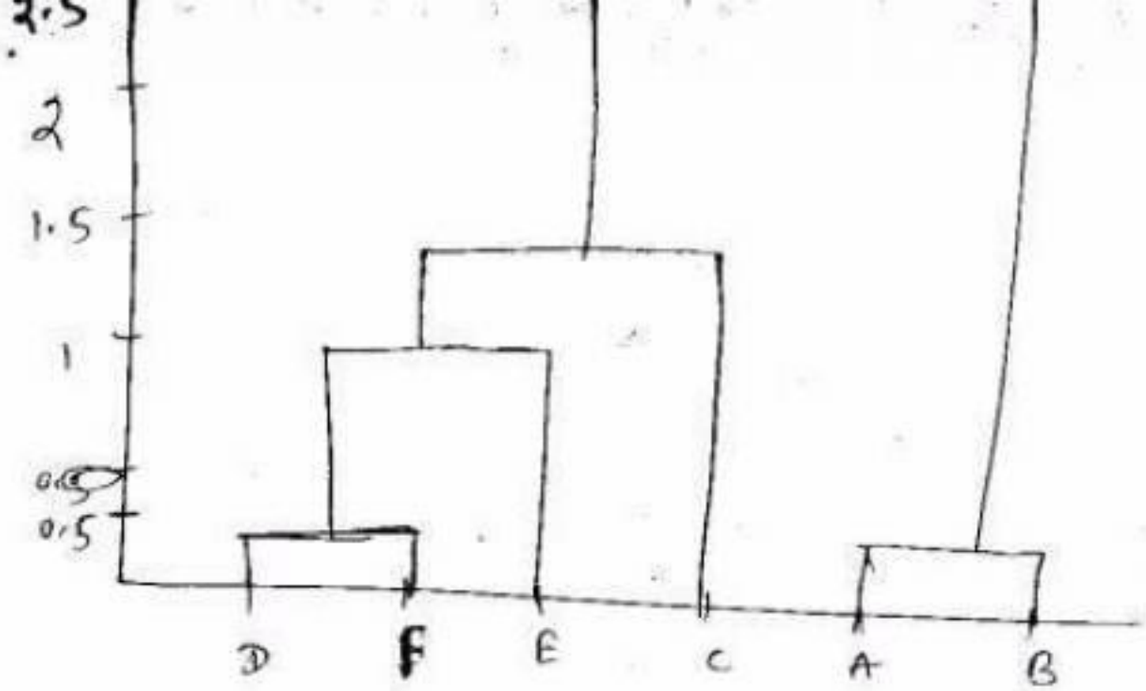
	(A,B)	C	D,F	E
A,B	0	4.95	2.5	3.54
C	4.95	0	2.24	1.41
D,F	2.50	2.24	0	1.00
E	3.54	1.41	1.00	0

	A,B	C	(D,F)E
(A,B)	0	4.95	2.5
C	4.95	0	1.41
(D,F)E	2.5	$\boxed{1.41}$	0

	AB	C	((D,F),E)C
A,B	0	4.95	2.5
C	~~4.95~~	0	0
(D,E,f)C			

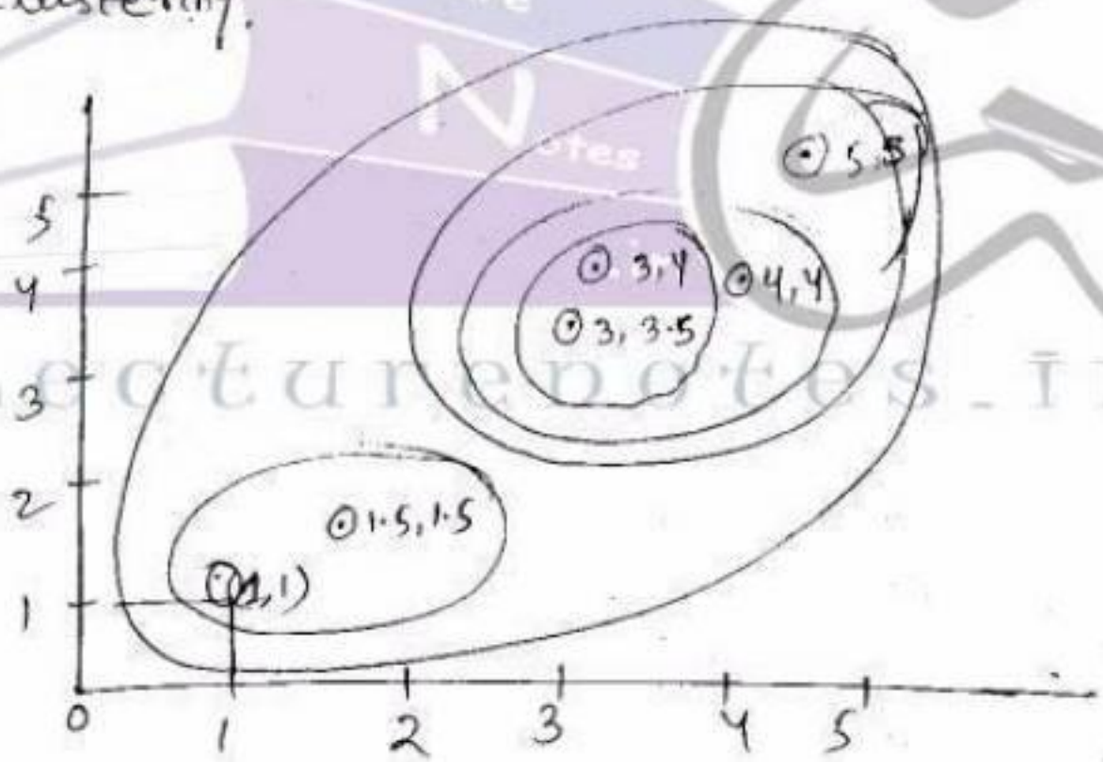
	AB	((D,F),E)C
(A,B)	0	2.5

Lecture Notes .in
Lecturenotes.in



(Dendogram)

A tree structure called a dendogram is commonly used to represent the process of hierarchical clustering.



New chapter

Dt. 1.12.11

Outlier Detection :-

Def :-

Also called anamaly detection.

This is a process of finding data objects with behaviour that are diff. from expection, such objects are called outliers or anamolie.

Application

Medical care.

Fraud detection.

Public safty & security

What are outliers ?

It is a data object that deviates significantly from rest of the object as if it were generated by diff. mechanism.

Types of outlier :-

Global outlier

contextual (conditional) outlier

collective outlier

Global outlier :-

In a given dataset a data object is called global outlier if deviates significantly from the rest of the dataset.

Contextual outlier

In a given dataset a dataobject is a contextual outlayer if it deviates significantly w.r.t specific context of the object.

Ex:- Temperature 45°, is it esceptional but if it is summer, it is normal.

In contextual outlier detection the attributes of the dataobjects are divided into two groups :-

contextual Attributes

Behavioural Attributes

The contextual attribute of a data object defines the object context.

Ex:- In the temp., the contextual attributes may date & location.

Behavioural Attribute :-

These defines the object characterstics & are used to evaluate wheather the object is an outlier in the context to which it belongs to.

Ex:- In temp., the behavioural attributes may be hummidity, Pressure, temp

Collective outlier :-

Given dataset, a subset of data objects from a collective outlier if the object as a whole, deviate significantly from the entire dataset. Importantly the individual data object may not be an outlier.

Outlier Detection Method :-

There are two ways to catagorise outlier detection method :-

① we catagorise outlier method according to whether sample of data for analysis is given with domain expert provided that can be used to built an outlier detection model.

② we divide method into groups according to their assumption regarding normal objects vs outlier.

In first method we use:-

(i) Supervised Learning (classification) → class level is known

(ii) Unsupervised Learning (clustering)

(iii) Semi-supervised Learning

① Supervised Learning :-

→ Supervised method's model used data normality & also abnormality.

→ outlier detection can then be model as a classification model.
The sample is used for training & testing.

→ In some application, the experts may label just the normal objects & any other objects not matching the model of normal object are reported as ~~outlayer~~ outlier.

other method models the outliers & object not matching the model of outliers are called normal objects.

② Unsupervised Learning :-

→ In some application cenaria the object ~~label~~ lebel ~~are~~ as normal or outlier or not available.

→ So unsupervised learning method has to be used. Many clustering method can be adopted to act as ~~an~~ unsupervised outlier detection method.

→ The idea is to find the cluster first, then the ~~The central idea is to find th~~ data object not belongs any cluster are detected outlier.

Dis Advantage :-

① ~~An~~ Data object not belongs to any cluster

find outliers.

③ Semi-supervised method

→ In many application although obtaining some label example is fisible.

→ The no. of such lebel example is often very sma we may encounter cases where only small sets of normal ~~or outlier~~ AND or OR outlier objects are labeled but most of the data are unlabe

→ When some labeled normal objects are availab we can use them together with unlabel obje that are closely by to train a model for normal objects.

→ The model of normal object then can be used to detect outlier.

→ These objects are not fitting the model of norm object are classified as outlier.

Dt. 2.12.12

Other methods are also there:-

① Clustering Based method
② Proximity Based Method
③ Stastical Method

Clustering Based Method :-

→ Clustering Based approach detect outliers by examining the relationship between objects & clust

→ An outlier is an object than belongs to a small & remote cluster or does not belongs to any cluster.

~~The~~ Three general approaches to Clustering Base outlier detection.

① Does the object belongs to any cluster. If not then it is identified as an outlier.

② Is there a large distance betˆ the object & the cluster to which it is closest, if yes it is a outlier

③ Is the object part of small or sparce cluster, if yes then all the object in that cluster are outlier.

①

→ Detecting outlier as object that donot belong to any cluster.

→ The clustering based outlier detection consider the distance betˆ an object & cluster, to which it is closest. If the distance is larger then object is likely an outlier w.r.t the cluster.

② Clustering based outlier detection using distance to the closest cluster.

→ Using K-means clustering method we can partation the data points into the cluster, the center of each cluster is marked as +ve.

→ For each object 'O' we can assign an outlier score to the object according to the distance betˆ the object & center.

→ Suppose the closest center to object 'O' is C_o, then the distance betˆ O & C_o is $\boxed{dist(O, C_o)}$.

→ The average distance betˆ the C_o & the object assign to O is $\boxed{L_{C_o}}$.

The ratio is $\frac{dist(O, C_o)}{L_{C_o}}$ measures how ~~distance from~~ $dis(O, C_o)$ stands out from the average. If the ratio is larger & farther away from object O is relative from the center & more likely object O is an outlier.

small or sparce cluster & declares the object in those clusters to be outlier.

Algorithm

① Find the cluster in a dataset and sort them accordingly to decreasing size.

② It uses a parameter α to distinguish larger from small cluster

Any cluster that contains at percentage ($\alpha = 90'$) of the dataset as considered a large cluster. The remaining cluster are reffered to as small cluste

③ To each data~~set~~ point assign a cluster based local outlier factor (CBLOF). For a point belonging to large cluster, its CBLOF is the product of the cluster size and similarity betˆ the point & the cluster.

For the point belonging to a small cluster its CBLOF is calculated as the product of the size of the cluster and the similarity betˆ the point & closest large cluster.

How to find the similarity betˆ point & closest large clust

→ CBLOF defines the similarity betˆ a point & cluster is a stastical way that represents the probabili that the point belongs to that cluster.

→ Larger the value more similar point to the cluster.

→ The small cluster that far away from larger cluster are considered to be outlier.

→ The points with lowest CBLOF scores are suspected to be outlier.

Advantages

① First they can detect outliers without requiring any label data ie unsupervised way.

② once the clusters are obtain clustering based methods need only to compare any object against to determine wheather the object is outlier or not.

Disadvantage

① It is very costly on very large dataset

② It heighly depends on clustering based method.

Lecture Notes .in
Lecturenotes.in