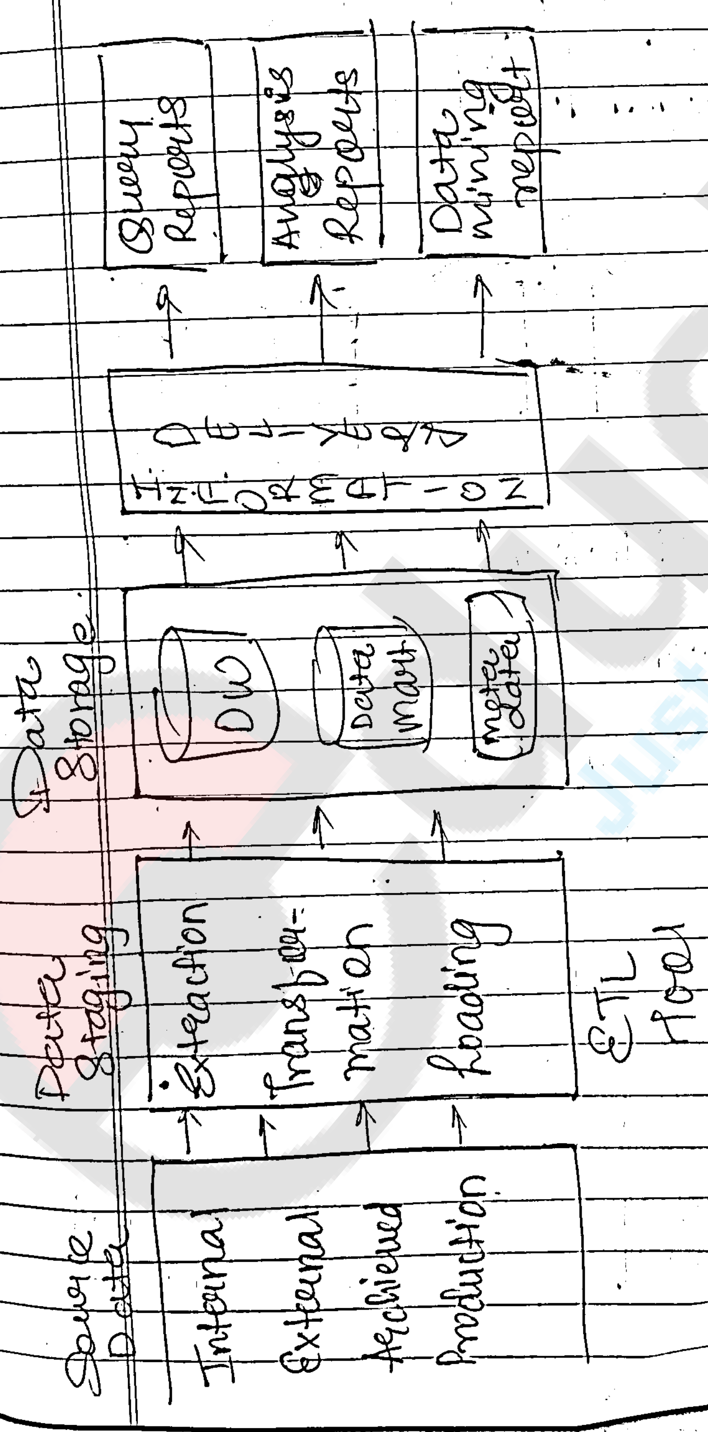
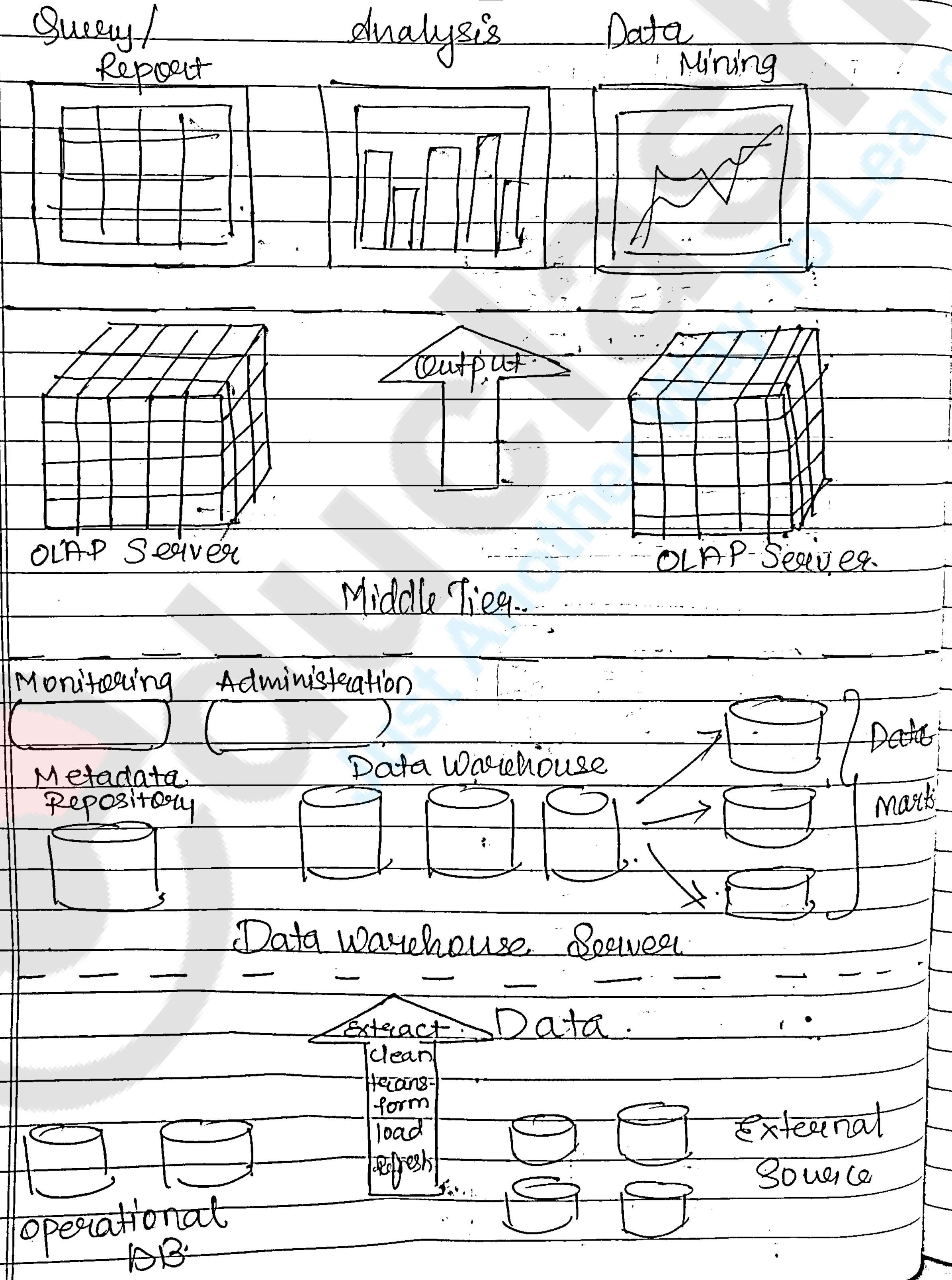
31st Jan

Data warehouse is a collection of subject oriented, integrated, non-volatile, time varient data

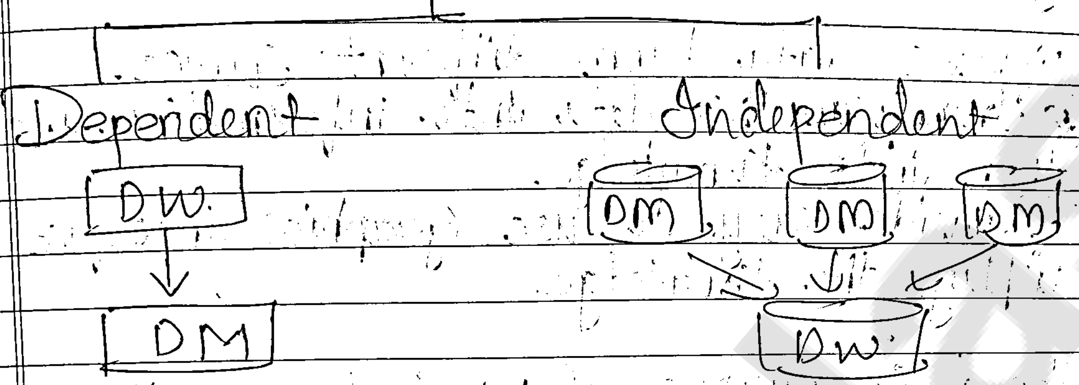
* DW Architecture:



* 3-Tier Architecture of DataWarehouse.



Data Mart



Module 1 Business Intelligence

* **Business Intelligence**

BI is defined as a set of mathematical models and analysis methodology that exploit the available data to generate information & knowledge which is useful for complex decision making process.

Q Explain Architecture of BI.

Q Difference b/w Data, Information, Knowledge.

DATA → Data Preparation → Information → Data Mining → Knowledge

classmate

Date
Page

* Goal of BI system

(1) Access data from different sources.
(2) Transform this data into information and then into knowledge.
(3) Provide easy to use Graphics interface to display the knowledge

The main purpose of BI is to provide knowledge workers with tools and methodologies that allow them to make effective and timely decision.

* Difference b/w data, information & knowledge.

Data: Data are collected on a daily basis in the form of ~~Th~~ bits, number, symbols, objects, etc. Data represents a structural codification of single primary entity.
Eg: for retail company primary entity or data are customer, point of sale (POS), salesman, product & sales receipt.

Information: Information is organized data. Information is the outcome of extraction & processing activity carried out of data & it appears meaningful for those who receive it in a specific domain.
Eg: To the sales manager of a retail company

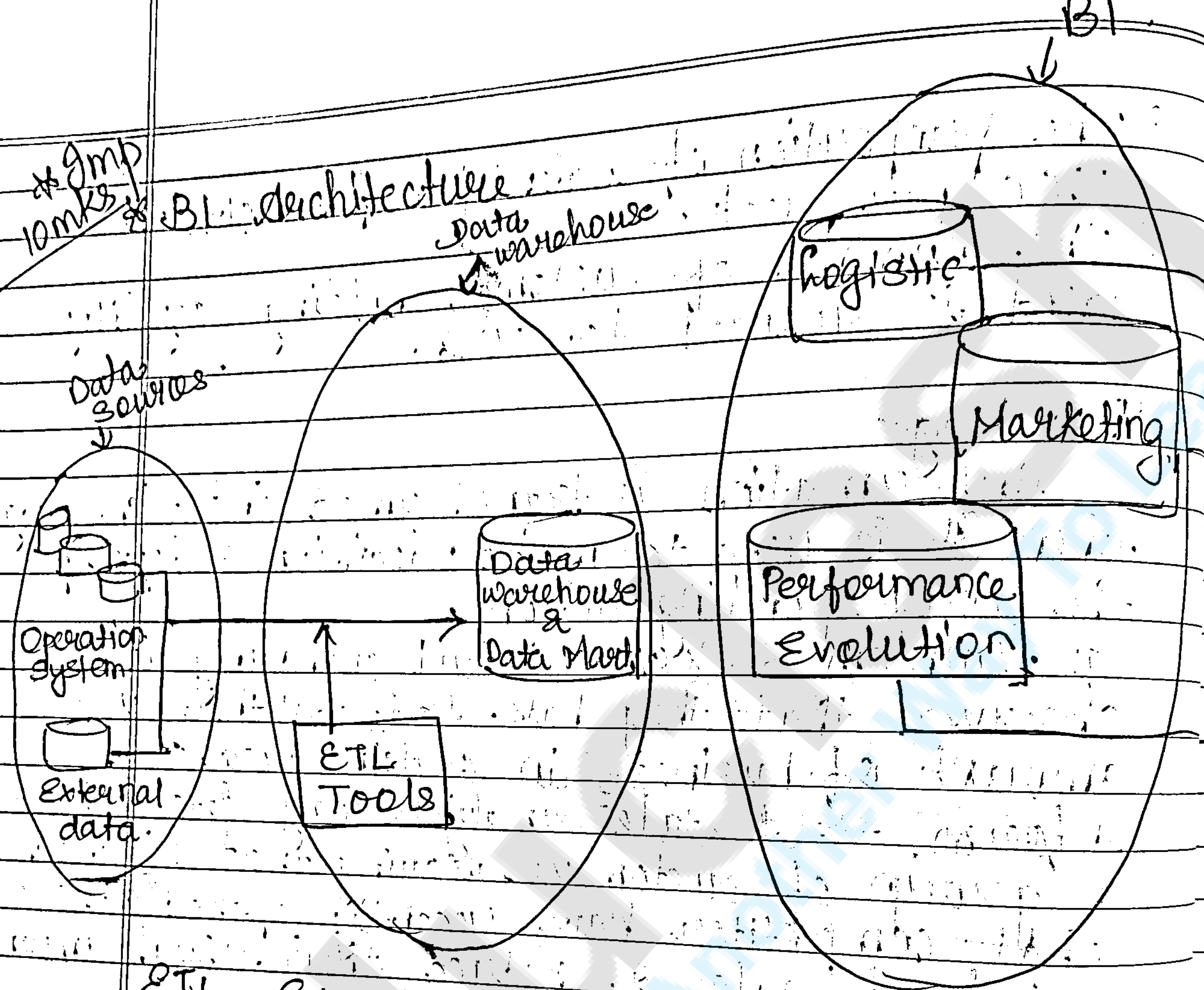
1) The proportion of sale of x product is 1000 per week.
2) The number of customer holding loyality card is reduced to 50%.

Knowledge:
Information is transformed into knowledge when it is used to make decision & develop the corresponding action.
Eg. Problem → For a Retail company, it is observed that they have reduced their usual amount of business in certain area.
Inform → A sales analysis may detect that a group of customers, leaving that area with that competitor have recently opened a new point of sale & reduce their usual amount of sale.
Knowledge → The problem can be solved by introducing new services for the customer residing in that specific area.

Passive Way → Knowledge Management
Active Way → BI

classmate
Date
Page

Imp 10mks * BI Architecture:



ETL = Extraction Transformation & Load

→ Components of BI Architecture:

1] Data Sources:

In first ~~section~~ stage it is necessary to gather & integrate the data stored in the various primary & secondary sources. Basically the data sources are operational sim, external data such as email, files, document etc.

2] Data Warehouse -

Using ETL tool, the data originating from different sources are stored in databases, Q What do you mean by BI & Architecture of BI.

1) MultiDimensional Cubes
2) Exploratory data analysis
3) Time Series analysis
4) Data Mining
5) Optimization

which is referred as datawarehouses & data marts.

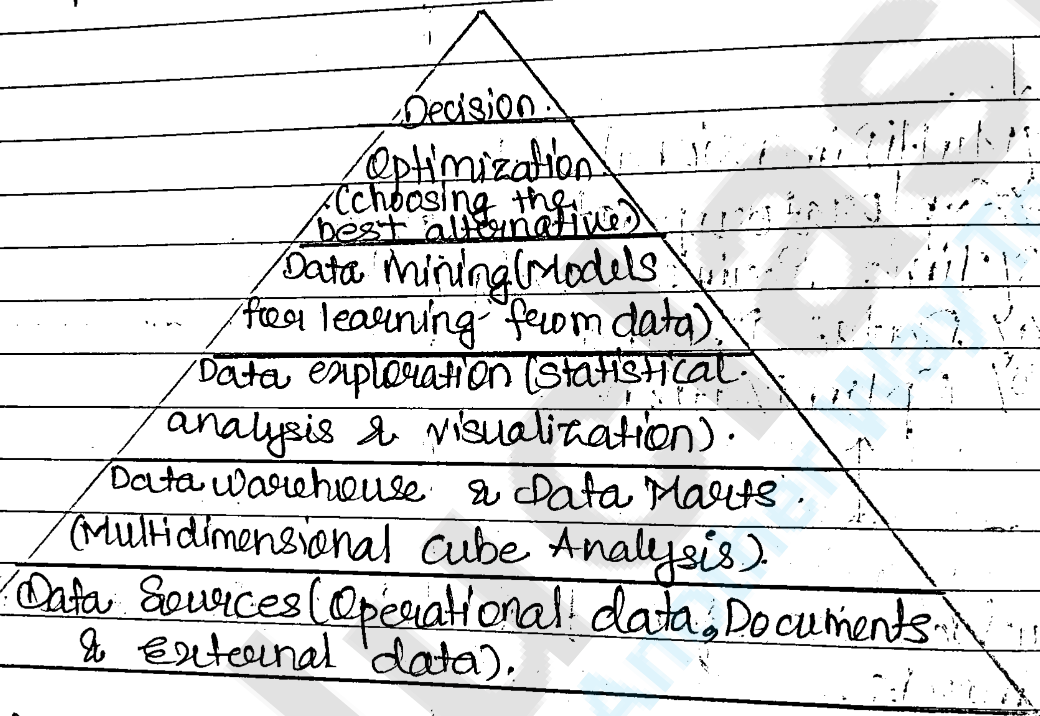
3] Business Intelligence:

Data are finally extracted and use to feed mathematical models & analysis methodologies to support decision.

Different decision support applications are:

a) Multidimensional cubes,
b) Exploratory data analysis,
c) Time series analysis
d) Inductive learning models for Data Mining.
e) Optimization models

* Components of BI



Top to Bottom:

1. Decision } Decision Makers
2. Optimization } Decision Makers
3. Data Mining - Active BI Analysis (Analyst & Experts)
4. Data Exploration - Passive BI Analysis (Analyst & Experts)
5. DW & Data Marts } - Db Administrator
6. Data Sources } - Db Administrator

Unit 3 (20m).

Introduction to DW

* Difference b/w operational system & information system.

Operation System OR OLTP (Online Transaction Processing).	Information System OR OLAP (Online Analytical Processing).
1) Contains only current data values	1) Contains archieved, derived summerized data.
2) OS is optimized for transactions.	2) optimized for complex queries
3) Access frequency is high.	3) Access frequency is medium to low.
4) The data can be read, write, delete & update.	4) The data can only be read.
5) Large no. of users.	5) Very few no. of users.

Q What do you mean by datawarehouse & architecture of DW.

Date
Page

* Define Datawarehouse.

Bill Inmon → 1990 → USA
↳ Father of DataWarehouse.

Datawarehouse is a subject oriented, integrated, non volatile, and time varient collection of data in support of management decision.

Subject oriented data: In the DW the data is stored by business objects or topics, not by operations applications like cust no, cust name address etc.
Eg: for manufacturing company, the subjects are customer, account, loan.

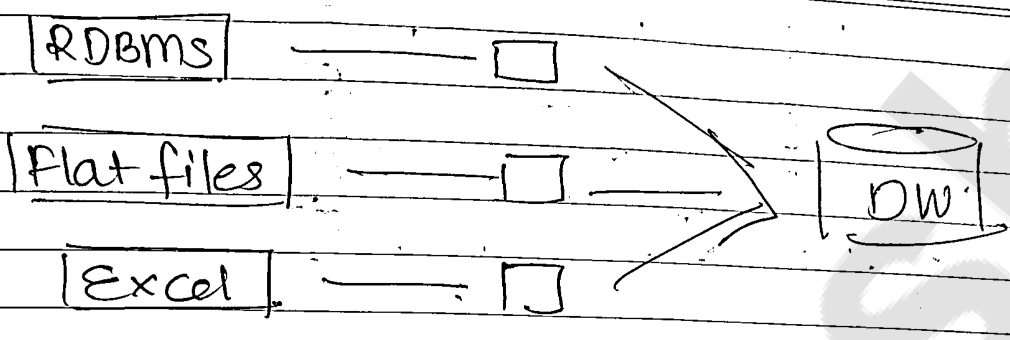
Integrated data: Integrated means the data are stored as a single unit, not as a collection of files that may have different structure.

Non-volatile: Non volatile means the data do not keep changing in DW, once it is intersected in DW, it cannot be deleted

① cust1 cust 2 → customer

Subject oriented data.

(2)



Integrated data

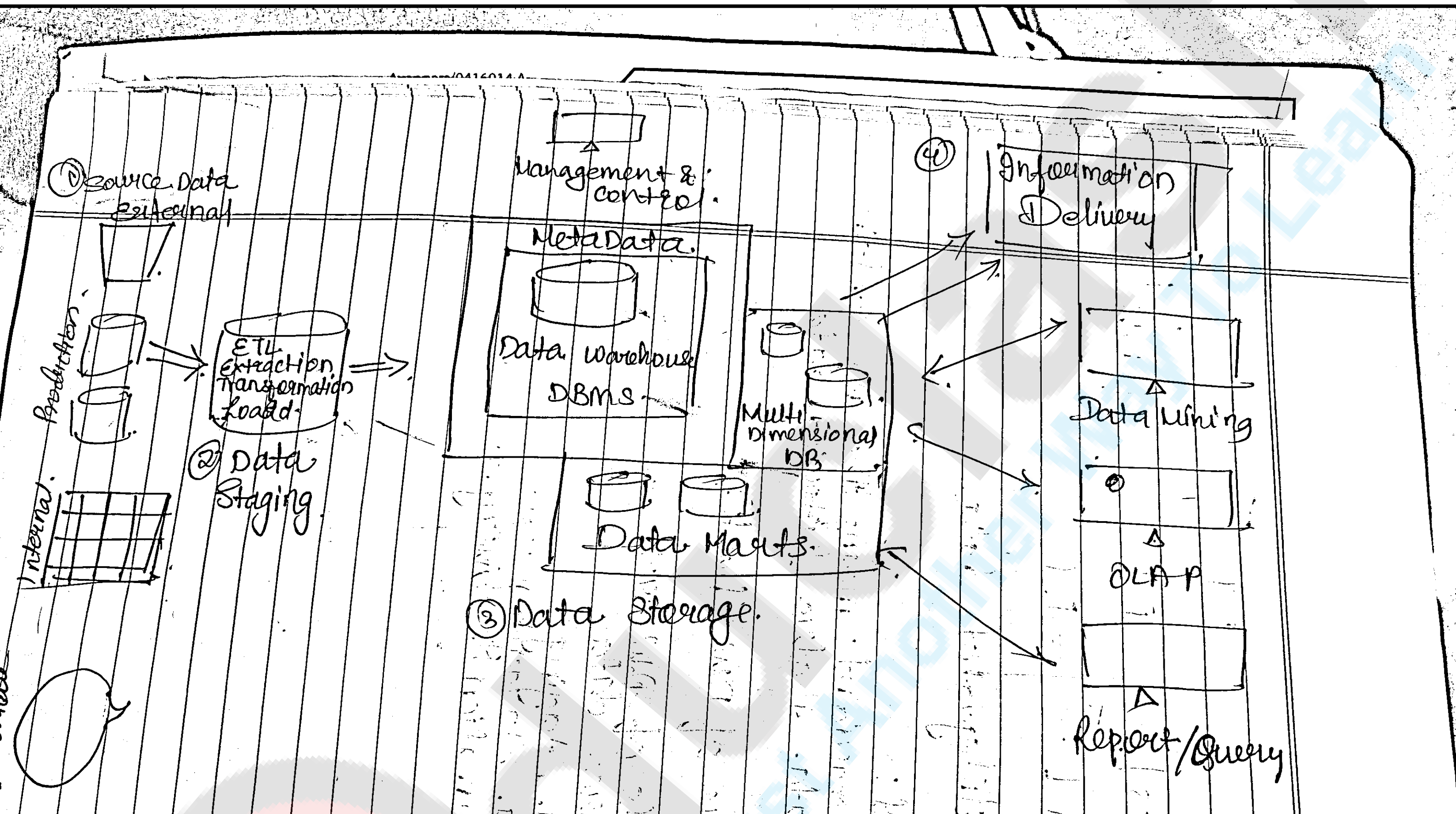
Time Variant:

Time variant means the DW contains current data & also historical data.

* Architecture of DW (back):

1) Components of DW are:

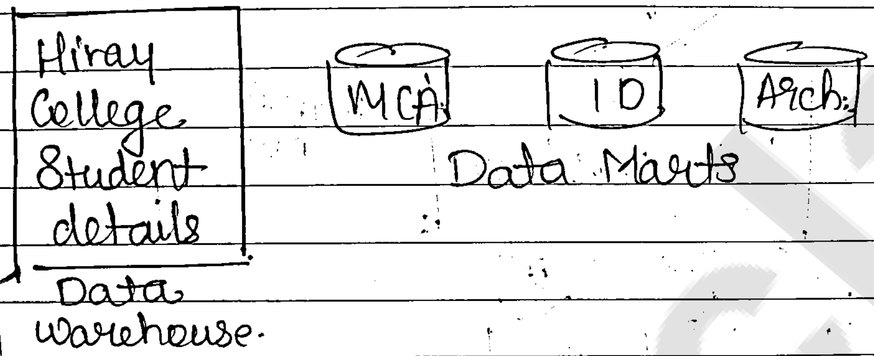
1. Source data
2. Data Staging
3. Data Storage
4. Information Delivery



DW Components or Building Stage

Q What is DataWarehouse? Explain the architecture.
Q what is ETL? Explain it in detail.
Q what is Metadata? Explain its types.
Q what is dataMart? Explain top down approach and Bottom up approach.

SN. * DataMart (departmental wise)

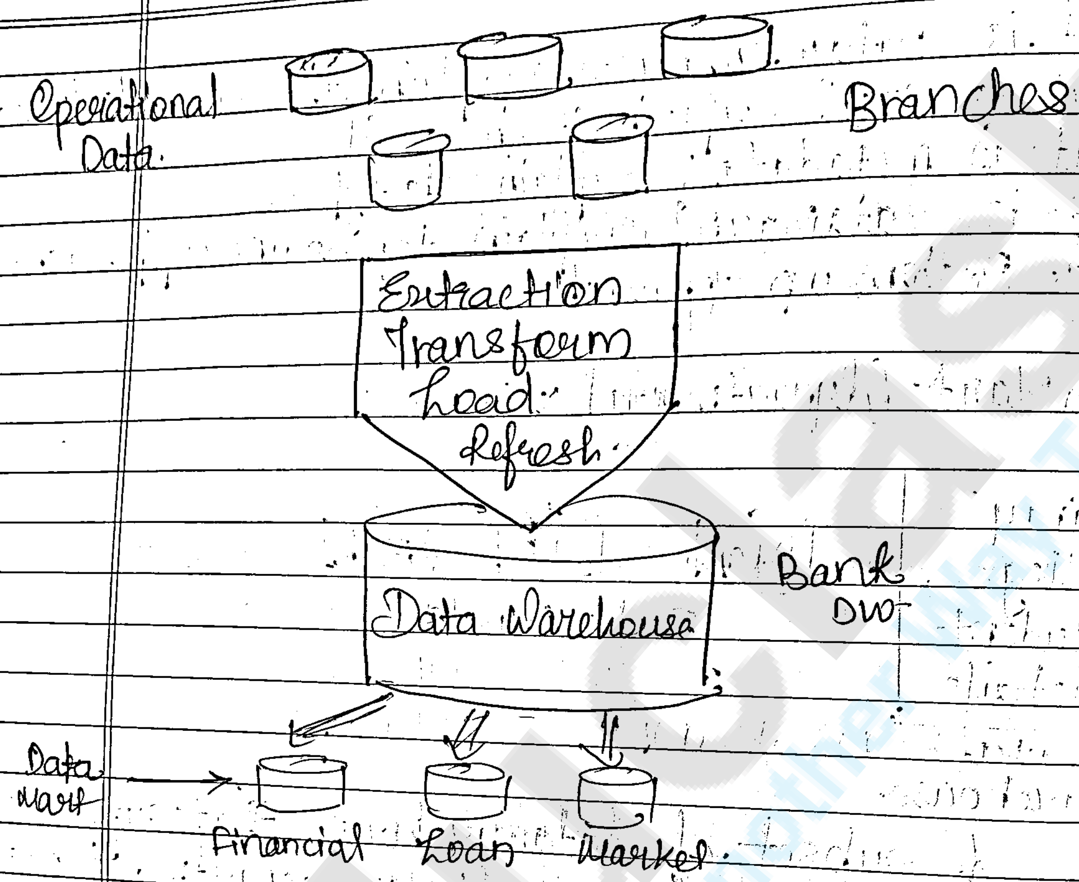


A subset of datawarehouse that supports the requirements of a particular department or DW.

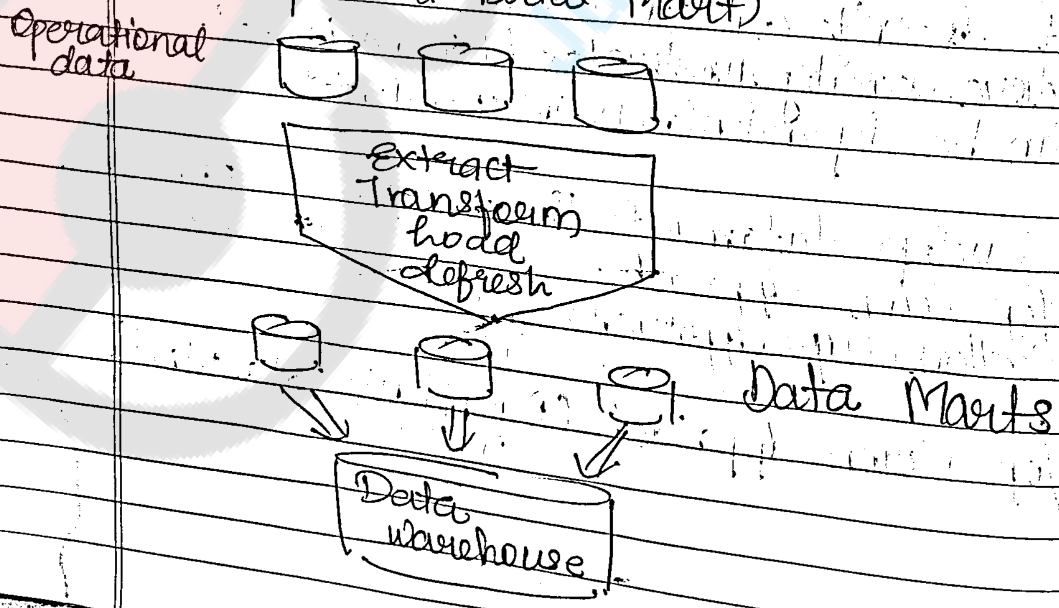
Characteristics of DataMart:
① It doesn't contain detail data like DW.
② More easily understood & navigated.
③ Can be dependent or independent

To develop datamart there are 2 approaches:
① Top Down Approach
② Bottom up approach.

i) Top Down approach (Dependent Data Mart)



2] Bottom Up Approach
(Independent Data Mart)



Date
Page

8N 4 Metadata.

MetaData is a DW contains the answer to the questions about the data in the DW. The answer will be stored in a place which is called as meta data repository

Definitions of Metadata.

1. Metadata is data about data.
2. Metadata is table of contents for data
3. Metadata is the DW road map.
4. Metadata is the glue that holds the DW content together.

Eg: Product Metadata. //Metadata format.

Source: Finished Goods orders,
System Maintenance contracts,
Online Sales.
Create date: Jan 15 2016
Last update date: weekly.
Last full refresh data: Dec 29 2016.

Full refresh cycle: Every year.
Data quality Reviewed: Jan 25 2017.
Last Duplication: Jan 10 2018
Planned Archival: Every six months.
Responsible user: Jane Brown

Date
Page

* Types of Metadata.

(1) Human Metadata
(2) Computer Based Metadata for user to use.
(3) Computer Based Metadata for computer to use.

A metadata repository should contain the following information:

- A description of datawarehouse structure.
  > Schema view
  > dimension view
  > data mart location & contents.

- Operational Metadata
  > history of migrated data
  > Currency of data.
  > Monitoring Information.

- Algorithm used for summarization
  > Data granularity
  > Partitions
  > Aggregation.

- Mapping from operational environment to DW
  > Source Database
  > gateway
  > extraction.
  > cleaning.

- Data related to system performance.
  > How to improve data access.
  > Rules for timing and scheduling to refresh
  > updates.

- Business Metadata.
  > Data ownership
  > Business terms & definition.

* ETL Process.

10 MKS. VVImp.

Extraction:

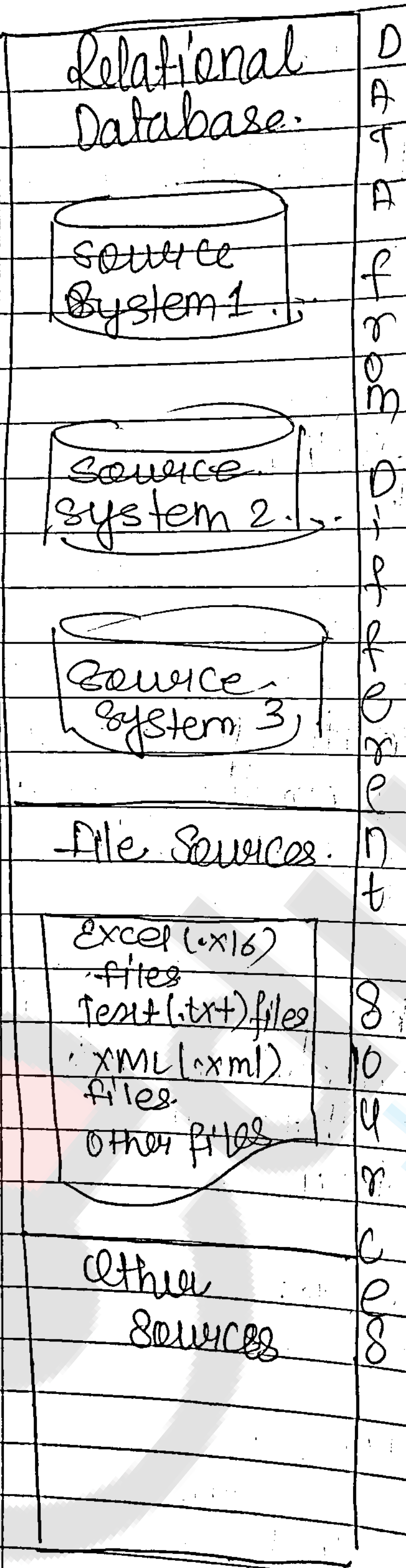
Extraction refers to the process of retrieving the required data from the operational system tables.

Transformation:

Transformation is the process which transforms the extracted data in accordance with the business rules & standards that have been established for the datawarehouse.

Transformation techniques are:

1) Format changes.
   DD mm YY.
   YY mm DD.
2) Duplications.
3) Splitting up fields.
4) Integrating fields.
5) Replacement of values.


Relational Database.
source System 1.
source system 2.
Source system 3
File Sources.
Excel (.xls) files
Text (.txt) files
XML (.xml) files
other files
Other Sources
DATA from Different sources

Loading:

* Dimension Modeling

Q Difference between ER modeling & dimensional modeling.

ER Modeling	Dimensional Modeling
① Remove data redundancy	① Captures critical measures
② Ensures data consistency	② Views data along dimensions
③ Express microscopic relationship	③ Intuitive to bussiness users.

Q What do you mean by dimensional modelling with eg: [with 3 schemas]

A

Requirement Gathering → Requirements Definition Document → Data Design → Dimensional Modelling.

Dimensional Modelling →
- Star schema.
- Snowflake schema.
- fact constellation schema.

- Dimensional modeling is a designing concept used by many datawarehouse designers to build DW.
- In dimensional modeling two types of tables are used:
  - > Fact table
  - > Dimension table.

Fact table:

- Fact tables are used to record actual facts or measurement units in the business with no redundancy.
- facts are numeric data item that are of interest to the business.
- Eg: In retail: No. of units sold, sales amount

Telecommunication: Length of call in minutes, average no. of calls.

Banking: average daily balance, transaction amount.

Dimension Table:

1. Dimension table establish the context of the fact.
2. It stores the fields that describe the facts.

Types of facts (Measure)

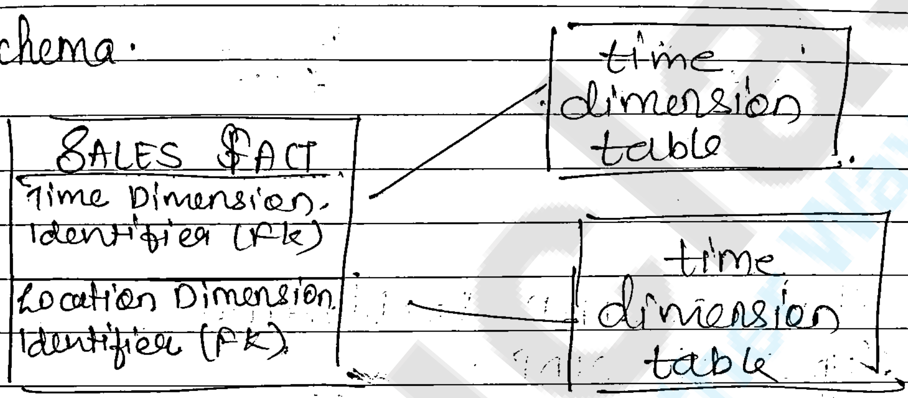
1. Additive.
2. Semi additive.
3. Non additive.

Star schema design → $\underbrace{\text{FK + measurement}}_{\text{Should be normalized}}$ } Fact table connected with multiple dimension.

Q Telecommunication system:
1. fact table.
2. Dimension table

* Star Schema.

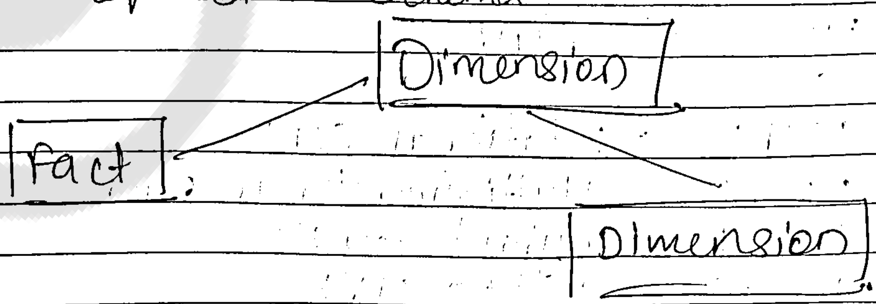
5mks



1) It is one type of dimensional model.
2) Used to view cube data.
3) One fact table connected with multiple dimension.
4) Dimension table may not be normalized.

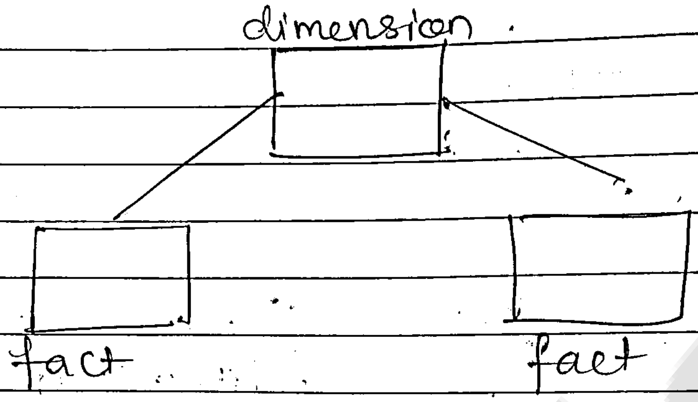
* Snowflake Schema:

— Extension of stale Schema.



classmate
Date
Page

4. Fact Constellation Schema.



Q. What is OLAP Architecture.
Q. Types of OLAP.

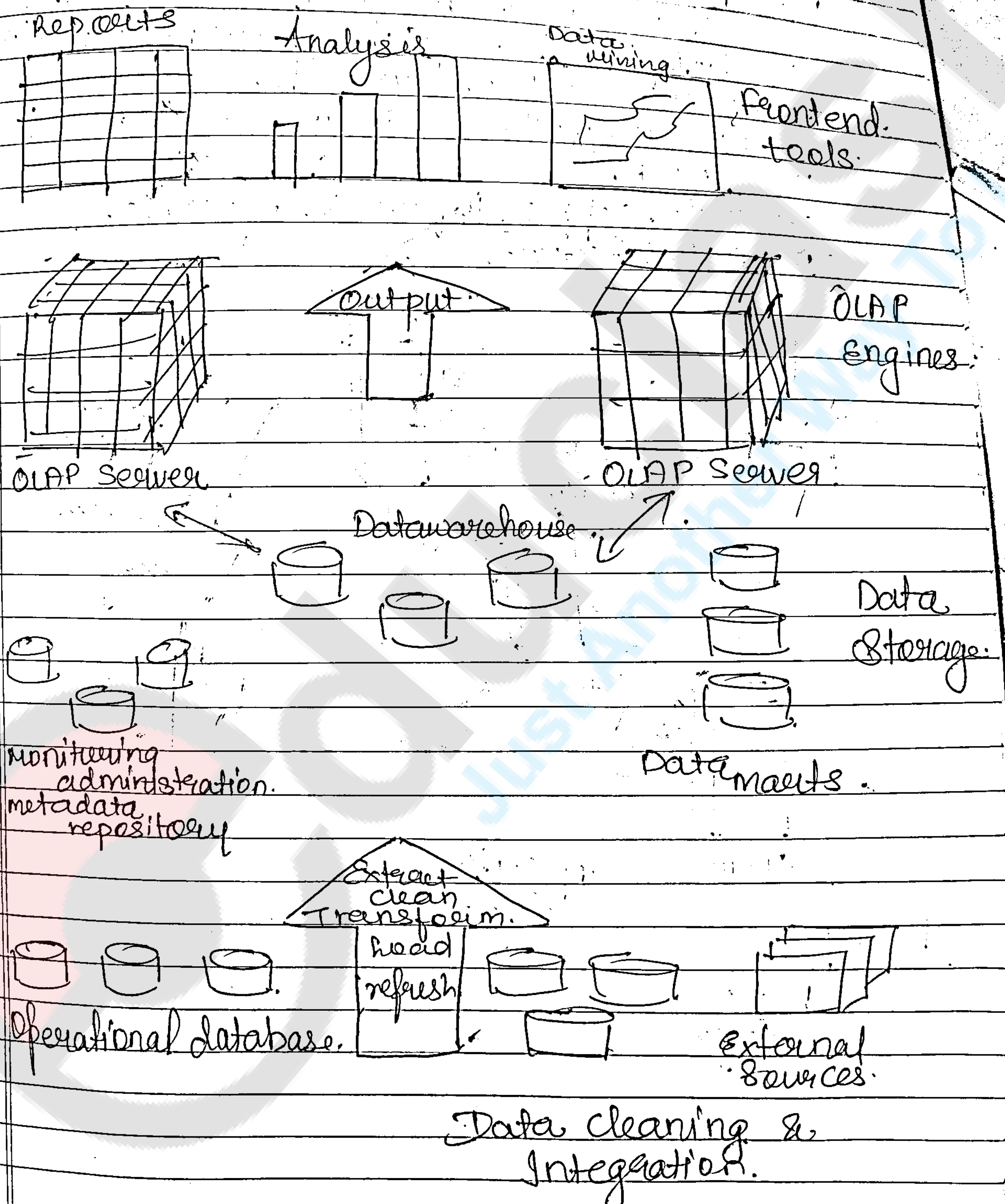
* OLAP - Online Analytical Processing.

- OLAP is a s/w frontend tool for datawarehouse.
- It is an information delivery system for dw.
- OLAP s/w provides the ability to analyze the large volume of information to improve decision making at all level of an organization.

* Types of OLAP

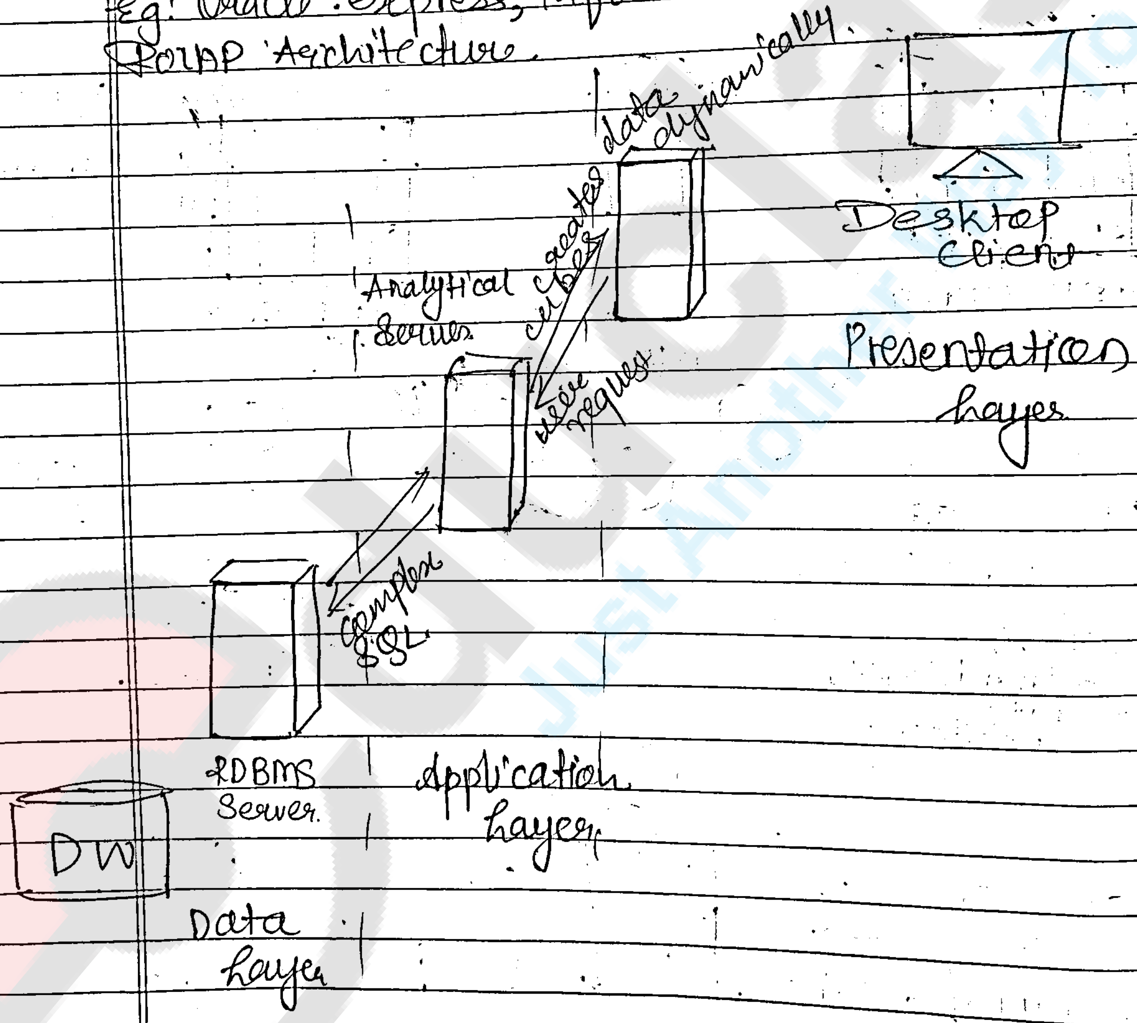
ROLAP → Relational OLAP.
MOLAP → Multidimensional OLAP.
HOLAP → Hybrid OLAP.
DOLAP → Desktop OLAP
WOLAP → Webbased OLAP.

* OLAP Architecture.



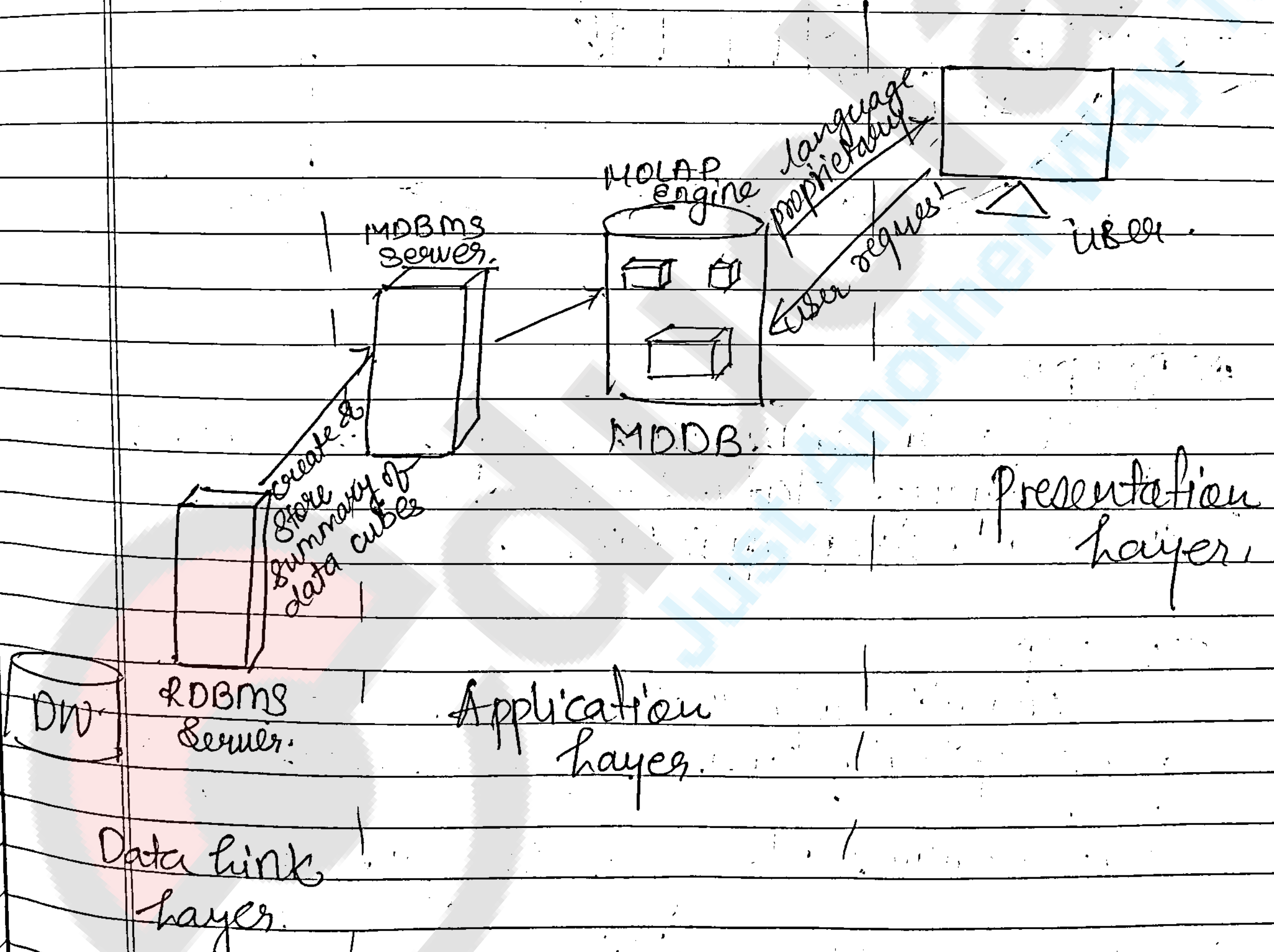
1) ROLAP

ROLAP is a S/w which is used to retrieve the data from a database oriented relationship.
It is suitable for large db.
ROLAP has ability to analyse a business better then MOLAP.
Eg: Oracle express, Informix metacube.
ROLAP Architecture.

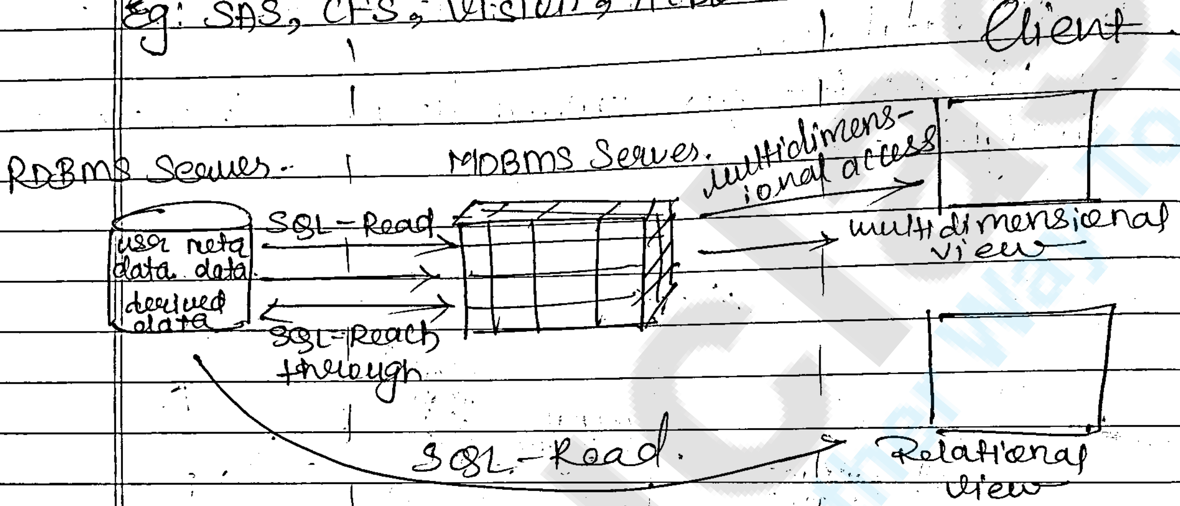


2) MOLAP.

It is a s/w tool which is used to retrieve multidimensional datacubes from database.
-It's faster then ROLAP but it is suitable for small database.
Eg: Microsoft s/w commander, FDC, Dimensional insight etc.



3) HOLAP
- Used to retrieve both relational & multi-dimensional relation data from the db.
Eg: SAS, CFS, Vision, Arbor Essbone.



4) DOLAP
Used in only one single standalone PC to retrieve data from db.
It is a desktop application.

5) WOLAP
- Analytic information is embedded into html pages, specially written for the needs of special users.
Eg: Executives, Managers, Analysts, Specialists.

* Types of OLAP Operations.
- Roll-up (drill-up)

* - Drill-down (roll down)
- Slice & Dice.
- Pivot (rotate).

7th Feb

* Data Mining:

- DM is a technique to find the hidden data from dw.
- DM is a collection of techniques for efficient automated discovery of previously unknown valid and understandable patterns in large databases

- The patterns must be actionable so that, they can be used for decision making process.

- DM is the generic term used to look for hidden patterns in db.

Data + [Interesting criteria] = Hidden Patterns.

↓

different data mining algorithms

It has been also called exploratory data analyses, data driven discovery & deductive learning process.

Data Mining Model

- Predictive
  - Classification
  - Regression
  - Time Session
  - Prediction
- Descriptive.
  - Clustering
  - Summarization.
  - Association Rules
  - Sequence Discovery

Predictive model makes a prediction about values of data using known results found from different data predicting modelling or maybe from the other historical data.

Descriptive model identifies patterns or relationship in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined.

* Association Rule Data Mining.
  An association is a rule when x then y
  It is denoted as $x \rightarrow y$
  It is used for finding the relationship among the data.
  Eg: If India wins in cricket world cup then sales of sweets rises.

Suppose x = India wins cricket.
y = Sales of sweets rises.
$x \rightarrow y$.

In association rule mining we use two terms:
1) Support 2) Confidence.

1) Support: probability that a transaction contains $x \cup y$ i.e it is the percentage of transaction in which item x & y occurred together.

2) Confidence: it is a conditional probability that transaction having x also contains y.
Confidence is a rule for finding the ratio of the no. of occurance of y given x among all other.

Q) Find the support & confidence for the rule Bag → Uniform by using the foll database.

t1 → Bag, Uniform, Crayons
t2 → Pencil, Uniform, ~~Bag~~ Books
t3 → Book, Bag, Uniform.
t4 → Bag, Pencil, Book.
t5 → Bag, uniform, pencil.
t6 → Uniform, crayons, Bag.
t7 → Bag, Pencil, Book.
t8 → Crayons, Uniform, Bag.
t9 → Books, crayons, Bag.
t10 → Uniform, crayons, pencil

$$\text{Support} = \frac{\text{No. of transactions containing Bag \& uniform}}{\text{Total no. of transaction}} \times 100$$

$$= \frac{5}{10} \times 100 = 50\%$$

$$\text{Confidence} = \frac{\text{No. of transactions both bag \& uniform}}{\text{No. of transactions containing bag}} \times 100$$

$$= \frac{5}{8} \times 100 = 62.5\%$$

Q find support & confidence for the rule Milk → Butter.

t1 → Milk, Bread, Butter
t2 → Milk, Butter
t3 → Jam, Butter
t4 → Bread, Egg, Fruit

$\text{Support} = \frac{2}{4} \times 100 = 50\%$ $\text{Confidence} = \frac{2}{2} \times 100 = 100\%$

Q ~~Confiden~~ $t_1 = \{F, A, D, B\}$
$t_2 = \{D, A, C, E, B\}$
$t_3 = \{C, A, B, E\}$
$t_4 = \{B, A, D\}$

$\{B, D\} \rightarrow A$

$\text{Support} = \frac{3}{4} \times 100 = 75\%$ $\text{Confidence} = \frac{3}{3} \times 100 = 100\%$

Q What is the need of data preprocessing? Techniques of preprocessing.

28/02

Data Quality:

• Measures for data quality: A multidimensional view.
- Accuracy.
- completeness.
- Consistency.
- Timeliness
- Believability
- Interpretability.

Techniques of Preprocessing.

• Data Cleaning
• Data integration
• Data reduction
• Data transformation and data discretization.

(1) Data cleaning.
Data in real world is dirty: lots of potentially incorrect data eg: instrument faulty, human or computer error.
- incomplete
- noisy.
- inconsistent
- intentional.

* How to handle missing (incomplete) data
- Data is not always available.
- Missing data may need to be inferred.
- Missing data may be due to:

> requipment malfunction.
> data not entered due to misunderstanding
> inconsistent with other recorded data
& thus deleted.
> not register history or changes of the data.

- Ignore the tuple.
- fill in the missing value manually.
- fill in it automatically with
  > global constant.
  > attribute mean.
  > the most probable value: reference-based
Beyesian formula or decision tree.
  > attribute mean for all samples belonging to
Same class.

Noisy Data.

Noise: random error or variance in a
measured variable.
Incorrect attribute values.
  > faulty data collection instruments
  > data entry
  > data transmission
  > technology limitation
  > inconsistency in naming convention.
Other data problems -> requires data cleaning

How to handle Noisy data.
- Binning
- Regression
- Clustering
- Combined computer & human inspection.

Binning

- Binning method smooth sorted data values by consulting its neighbourhood i.e the values around it.

Method:

① Assort the data in descending or ascending order.
② Partition the sorted data into equal frequency bins
③ apply any binning method such as -Smooth by bin means,
-Smooth by bin median
-Smooth by bin boundaries, etc.

Q Suppose a group of 9 sales price records has been sorted as follows.
4, 8, 15, 21, 21, 24, 25, 28, 34.
By considering bin depth of 3 use smooth by bin means and smooth by boundaries.

→ Partition the given data in equal freq bins i.e 3.
Bin 1 : 4, 8, 15
Bin 2 : 21, 21, 24
Bin 3 : 25, 28, 34.

Date
Page

i) Smooth by bin mean
- Find mean of each bin ~~mean~~ & replace the values of each bin.

Bin 1: $\frac{4+8+15}{3} = \frac{27}{3} = 9$

Bin 2: $\frac{21+21+24}{3} = 22$

Bin 3 = $\frac{25+28+34}{3} = 29$

After smoothening the bin values are:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

ii) Smooth by bin boundaries
find the max & mean of each bin.

		min	max
Bin 1:	4, 8, 15	4	15
2:	21, 21, 24	21	24
3:	25, 28, 34	25	34

Replace the data in bin by boundary values

Bin 1: 4, 4, 15

2: 21, 21, 24

3: 25, 25, 34

Q. Apply Smooth by bin means & Smooth by bin boundaries, Smooth the foll data.
Bin depth = 4.
10, 4, 3, 10, 11, 6, 7, 5, 6, 12, 15, 16

→ (1) Sort the data:
3, 4, 5, 6, 6, 7, 10, 10, 11, 12, 15, 16.

(2) Partition
Bin 1: 3, 4, 5, 6
Bin 2: 6, 7, 10, 10
Bin 3: 11, 12, 15, 16.

(3) Smooth by bin mean
Bin 1: $\frac{3+4+5+6}{4} = 4.5$

Bin: $\frac{6+7+10+10}{4} = 8.25$

Bin: $\frac{11+12+15+16}{4} = 13.5$

After smoothening the bin values are:
Bin 1: 4.5, 4.5, 4.5, 4.5
2: 8.25, 8.25, 8.25, 8.25
3: 13.5, 13.5, 13.5, 13.5

(4) Smooth by bin boundary
Bin 1: 3, 4, 5, 6 min = 3 max 6.
2: 6, 7, 10, 10 min = 6 max 10
3: 11, 12, 15, 16 min = 11 max = 16

Scan

Date
Page

After Smoothening:

Bin 1: 3, 3, 6, 6.
2: 6, 6, 10, 10
3: 11, 11, 16, 16.

— Regression method is used to smooth by fitting the data in regression function.
> Linear regression used to find the best line to fit two attributes so that one attribute can be used to predict the other.

> multiple linear regression extension of linear regression where more than two attributes are involved & the data are fit to multidimensional surface

— Clustering:
~~detect~~ By using ~~Euclidean~~ Eculidian distance formula the distance between outliers & clusters will be found out then the nearest cluster will be consider for the outlier

9.

* Data Integration

Combines data from multiple source into a coherent store.

Schema Integration

Integrate metadata from different source.

- Entity Identification problem:

  - Identify real world entities from multiple data sources.

Detecting and resolving data value conflicts

For the same real world entity, attribute values from different sources are different.

Possible reasons: different representations.

- Are any attributes corelated
- Tuple duplication.
- Value confliction.

Date
Page

Apriori Algorithm.
is used to discover all the frequent item-sets from a large db in an efficient manner.
The idea behind Apriori algorithm is:- if all item set is frequent then all subsets of this item set should also be frequent.

Q Find frequent item sets or association rule with 30% support & 70% confidence of foll db:

Transaction Id	Item Set
t1	Bag, uniform, crayons
t2	Books, Bag, Uniform
t3	Bag, Uniform, Pencil
t4	Bag, pencil, Books
t5	Uniform, Crayons, Bag
t6	Bag, Pencil, Books
t7	Crayons, Uniform, Bag
t8	Books, crayoons, Bag
t9	Uniform, Crayons, Pencil
t10	Pencil, Uniform, Books

Soln:
Given support = 30% i.e out of 10 transaction atleast $\frac{10}{100} \times 30 = 3$ or more than 3 frequency should be considered.

1st Item Set.

Items	Frequency
Bag	8
Uniform	7
Crayons	5
Books	5
Pencil	5

Since all the items have more than 3 frequency they will be considered for the 2nd frequency item set.

2nd frequency Item Set.

Items	Frequency
Bags, uniform	5
Bags, Crayon	4
Bags, Books	4
Bags, Pencil	3
Uniform, Crayons	4
Uniform, Books	2
Uniform, Pencil	3
Crayons, Books	1
Crayons, Pencil	1
Books, Pencil	3

Since {uniform, Books}, {crayon, book} & {crayon, pencil} has frequency < 3 $\therefore$ it would be discarded and others will proceed to the next level.

3rd frequency Item Set

Items	Frequency
Bag, Uniform, Crayons	3
Bag, Uniform, Books	1
Bag, Uniform, pencil	1
Bag, crayon, books	1
Bag, crayons, pencil	0
Bag, books, pencil	2
Uniform, crayons, pencil	1
Uniform, pencil, books	1

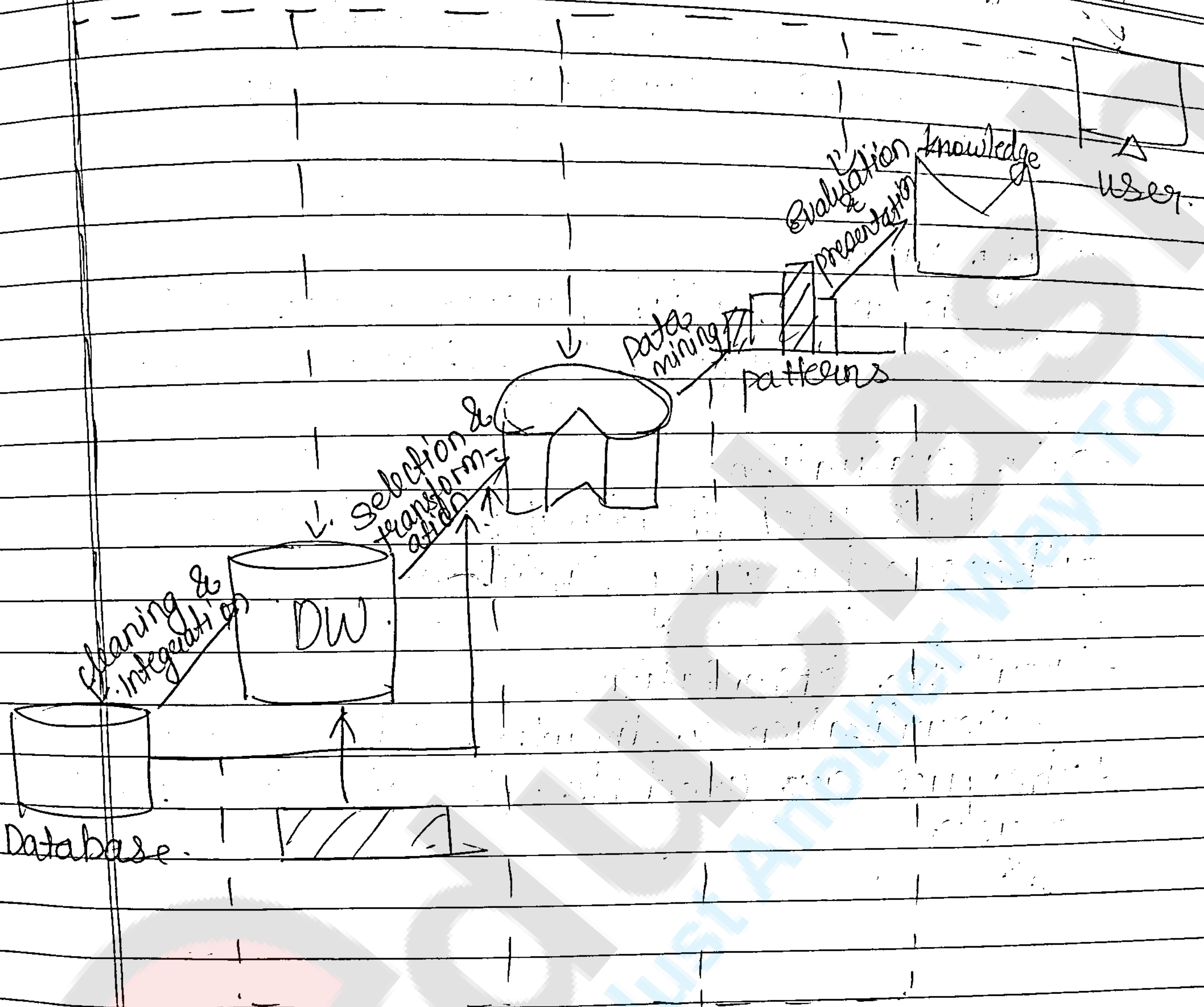
From the item set only Bag, uniform, crayons satisfy the frequency 3 i.e support 30%. Other item sets will be discarded.
Since there is only one frequency item set left, there will be no further combination.

22/02

Vimp KDD Process (Knowledge discovery from database)

- KDD is the process of finding useful information & patterns in data

- Data Cleaning & Integration:
  Noisy data & inconsistent date will be removed from the db.



- Data Integration:
    Multiple data sources are combined into a single form.

- Data Selection:
    In this step the relevant data to the analysis task are retrieve from the db.

- Data transformation:
    In this step data are transformed & consolidated

into forms which are appropriate for mining while performing summary & aggregation appreciation.

- Data Mining:
It is an essential process where intelligent methods are applied to extract data patterns.

- Data evaluation:
It is used to identify the truly interesting patterns representing knowledge.

- Knowledge presentation:
Visualization & knowledge representation techniques are used to present mind knowledge to use.

Data Integration:

~~Problems:~~

* ① Handling redundancy in Data Integration.

- Correlation analysis is used for nominal data.

$$x^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \text{ — (one dimensional)}$$

correlation relationship between two attributes A & B can be discovered by x^2 test.
Suppose A has c number of distinct values i.e. $a_1, a_2 \ldots a_c$.
and B has r distinct values represented as $b_1, b_2 \ldots b_r$
To perform x^2 test, a contingency table will be organised with c values of A making columns & r values of B making of the rows.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \text{ — (two dimensional)}$$

$$E_{ij} = \frac{\text{count}(A=a_i) * \text{count}(B=b_j)}{n}$$

where n = no. of tuples
count($A=a_i$) ⇒ no. of tuples having value a_i
count($B=b_j$) ⇒ no. of tuples having value b_j.

The test is base on ~~an~~ significant level with. $(r-1)*(c-1)$ degrees of freedom.
If the hypothesis is rejected then we say that A & B statiscally correlated.

Q Correlation analysis of nominal attribute using χ^2:

Suppose a group of 1500 people ~~were~~ are surveyed. The gender of each person was noted. Each person was polled as to wether his or her preferred type of reading material was fiction or non-fiction.
The foll table gives the contingency table data for the surveyed data. Find are gender & preferred reading correlated?

	male	female	
fiction	250 (90)	200 (360)	450
non-fic	50 (210)	1000 (840)	1050
	300	1200	1500

→ $C_{11} = \frac{\text{count(male)} * \text{count(fiction)}}{}$

$= \frac{300 * 450}{1500} = \frac{135000}{1500} = 90$

Since our calculated value (507.92) > 10.828
∴ the hypothesis is rejected.
It will be concluded the 2 attributes are strongly correlated.

$$e_{12} = \frac{\text{count(female)} * \text{count(fiction)}}{n}$$

$$= \frac{1200 \times 450}{1500} = 360$$

$$e_{21} = \frac{\text{count (male)} * \text{count (non-fi'c)}}{n}$$

$$= \frac{300 \times 1050}{1500} = 210$$

$$e_{22} = \frac{\text{count (female)} * \text{count (non-fi'c)}}{n}$$

$$= \frac{1200 \times 1050}{1500} = 840$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= \frac{25600}{90} + \frac{25,600}{210} + \frac{25600}{360} + \frac{25600}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.47$$

$$= 507.92$$

Degree of freedom: $(r-1) * (c-1)$

$= (2-1) * (2-1)$

$= 1 * 1 = 1.$

For 1 degree of freedom the χ^2 value needed to reject the hypothesis at 0.001 significance level is 10.828

class fruit & achieve polymorphism into it.

Q 13/02

DMBI

Q Given a simple transactional database. Find FP tree for this database if minimum support threshold is 3.

TID	~~Itemset~~ Itemset
01	f, a, c, d, g, i, m, p
02	a, b, c, f, l, m, o
03	b, f, h, j, o
04	b, c, k, s, p
05	a, f, c, e, l, p, m, n

Soln:
Scan the database & find frequent item having frequency 3.

Items	frequency
f	4
a	3
c	4
d	1
g	1

i	1
m	3
p	3
b	3
d	2
o	2
h	1
j	1
k	1
s	1
e	1
n	1

Qualified items are
{(f,4), (a,3), (c,4), (m,3), (p,3), (b,3)}.

- FP growth model

Arrange the items in descending order of frequency.

L = {(f,4), (c,4), (a,3), (b,3), (m,3), (p,3)}

L → List of frequent items

* Developing FP tree

The root of the tree will be labelled as ROOT.

Scan the db second time.

Scan the first transaction which leads to the construction of first branch of FP tree.

Scan of first transaction. (f, a, c, d, g, i, m, p)
Only those items that are in the list of frequent items ~~only those are items~~ are selected.
i.e. {(f,1), (c,1), (a,1), (m,1) (p,1)}

[NOTE: sequence as L is compulsory]

The FP tree will be.

ROOT
↓
f,1
↓
c,1
↓
a,1
↓
m,1
↓
p,1

Scan of second transaction.
(a, b, c, f, l, m, o).

{(f,1) (c,1) (a,1) (b,1)(m,1).}

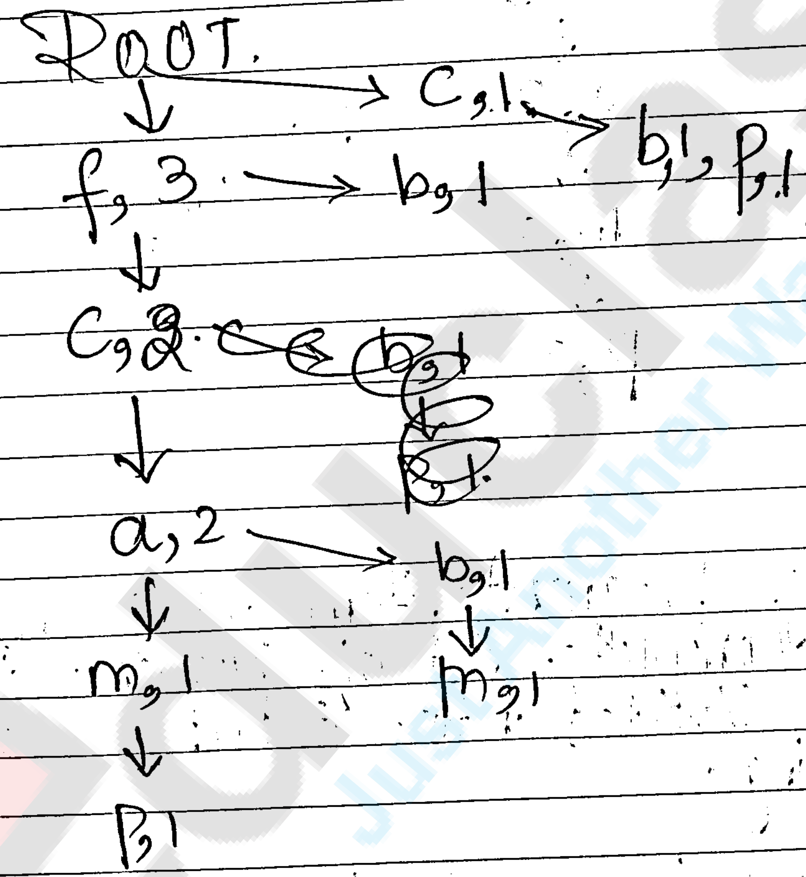
ROOT
↓
f, 2
↓
c, 2
↓
a, 2 → b, 1
↓ ↓
m, 1 m, 1
↓
p, 1

Scan third transaction
(b, f, h, j, o)
{(f, 1), (b, 1)}

ROOT
↓
f, 3 → b, 1
↓
c, 2
↓
a, 2 → b, 1
↓ ↓
m, 1 m, 1
↓
p, 1

Scan 4^{th} transaction
(b, c, k, s, p)

{(c, 1), (b, 1), (p, 1)}



Scan 5^{th} transaction
(a, f, c, e, l, p, m, n)

{(f, 1) (c, 1) (a, 1) (m, 1) (p, 1)}

ROOT → c, 1 → b, 1 → p, 1

↓

f, 4 → b, 1

↓

c, 3

↓

a, 3 → b, 1 → m, 1

↓

m, 2

↓

p, 2

Finding frequent item set from FP tree.

The complete set of frequent item set can be divided into 6 subsets without overlapping.

1. frequent item sets having item p.
2. frequent item sets having item m but not p.
3. frequent item sets having b but not m & p
4. f.i.s having a without b, m, p.

⑤ fis having c without a, b, m & p.

⑥ fis having f.

① freq item sets having p

{(f,4),(c,3),(a,3),(m,2),(p,2)} and

{(c,1),(b,1),(p,1)}

two parts will be selected from the FP tree i.e

Samples with a freq item PRule

{(f,2) (c,2) (a,2) (m,2) (p,2)} and

{(c,1), (b,1), (p,1)}

The given threshold value 3 satisfies only the frequent items are.

{(c,3) (p,3)} or {c, p}.

② freq item set having m but not p.

{~~(f,4) (c,3) (a,3) (m,2)~~} and

{(f,1) (c,1) (a,3) (b,1) (m,1)}

Sample with freq item m are

{(f,3) (c,3) (a,3) (m,3)} and

{(a,1) (b,1) (m,1)}

The given treshold 3 satisfies

{(a,3) (m,3)} or {a,m}

{(f,4) (c,3) (a,3) (m,2)} and

{(f,4) (c,3) (a,3) (b,1) (m,1)}

{(f,2) (c,2) (a,2) (m,2)} and

{(f,1) (c,1) (a,1) (b,1) (m,1)}

threshold ⇒ 3

{(f,3) (c,3) (a,3) (m,3)} or {f, c, a, m}

③ freq item set having b.

{(f,4) (c,3) (a,3) (b,1)} and

{(f,3),(b,1)} and

{(c,1) (b,1)}.

Sample with freq item b.

{(f,1) (b,1)} and

{(f,1) (b,1)} and

{(c,1) (b,1)}

threshold ⇒ 3 {b,3}.

Since only one item is obtained, it is not considered.

④ f is having a.

{(f,4) (c,3) (a,3)}.

{(f,3) (c,3) (a,3)}

threshhold ⇒ 3 {f,c,a}

⑤ freq item set having c.
{(f,4), (c,3)} and {c,1}

Sample of c.
{(f,3) (c,3)} {c,1}

thresheld ⇒ 3 {(f,3) (c,4)} or
{f,c}.

⑥ freq item set of f

The frequent item sets we get are
{c,p} {f,c,a,m} {f,c,a} {f,c}

→ {f,c,a} and {f,c} are subsets of frequent item set {f,c,a,m}.

∴ the final solution of in FP growth model is: {c,p} and {f,c,a,m}

20/02. DMBI

Q What is the difference between Association and classification?
Explain Associative classification method with suitable example.

* Associative Classification Method (ACM)

**
CMAR Classification based on Multiple Association Rule.
- It is a classification association rule mining method which is used for associative classification method.

Method: The association classification method consists of two phases.
1) Rule Generation / Training
2) Classification / Testing.

1) In rule generation method (phase), CMAR computes the complete set of rules in the form of $R : P \rightarrow C$ such that sup(R) and conf (R) pass the given thresholds where $P = \{a_1, a_2, \ldots a_k\}$ is a pattern with a set of attribute value.
c = class level.

For a given support threshold and confidence threshold the ACM finds the complete set of class association rules passing the thresholds.

② In classification phase, when a new sample ~~comple~~ comes the classifier represented by a set of association rules, selects the rule that match with the sample and has the highest confidence and is used to predict the classification of the new sample

Q Apply ~~assoc~~ ACM to the following training data set T with support threshold 2 and confidence threshold 70%

Tid.	Item set	Class.
01	a1, b1, c1, d1	A
02	a1, b2, c1, d2	B
03	a2, b3, c2, d3	A
04	a1, b2, c2, d3	C
05	a1, b2, c1, d3	C

Solⁿ:

Scan the training data set and find the set of item set occuring >= 2 freq.

Itemset	frequency
a_1	4 ✓
b_1	1
c_1	3 ✓
d_1	1
b_2	3 ✓
d_2	1
a_2	1
b_3	1
c_2	1
d_3	3 ✓
c_3	1

The qualified itemsets are:

$\{(a_1, 4)\ (c_1, 3)\ (b_2, 3)\ (d_3, 3)\}$

Sort the items in descending order:

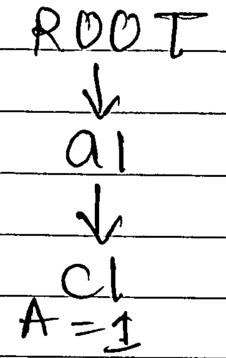
$F = \{(a_1, 4)\ (b_2, 3)\ (c_1, 3)\ (d_3, 3)\}$

$\{a_1, c_1, b_2, d_3\} \Rightarrow \{a_1, b_2, c_1, d_3\}$.

Step 2: Scan the training dataset to construct FP tree.

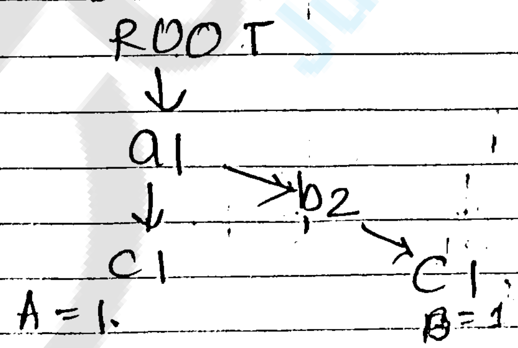
F (a1, b2, c1, d3)

Scan of 1st transaction (a1, b1, c1, d1).
The extracted items are: (a1, c1)
The path is a1 → c1.
FP tree will be.

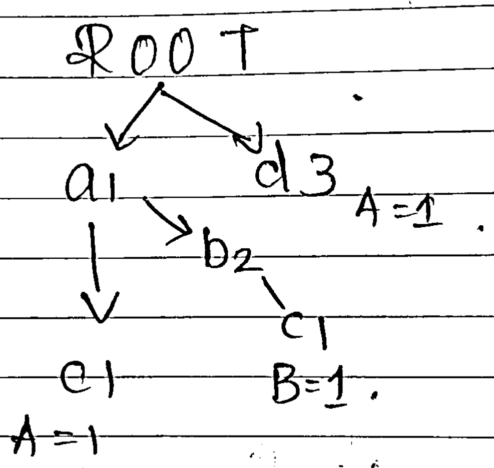


The class label of the sample & the corresponding counter are attached to the last node of the path.
A = 1.

Scan of 2nd transaction (a1 b2, c1, d2)
(a1, b2, c1). a1 → b2 → c1.
Fp tree will be

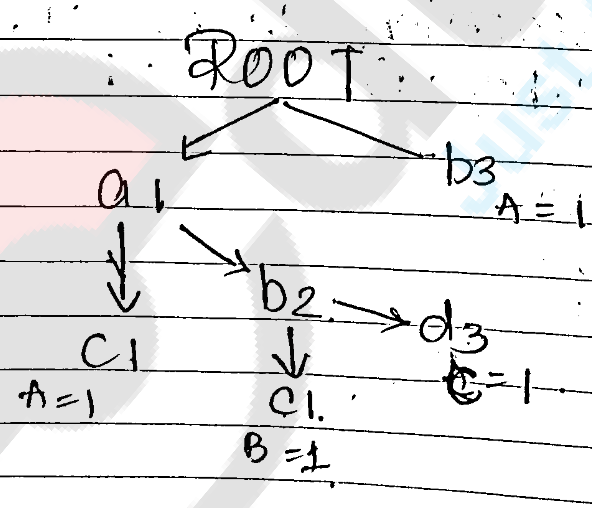


Scan of 3rd transaction (a2, b3, c2, d3)
Extracted (d3)



Scan of 4th transaction (a1, b2, c3, d3)
Extracted (a1, b2, d3)

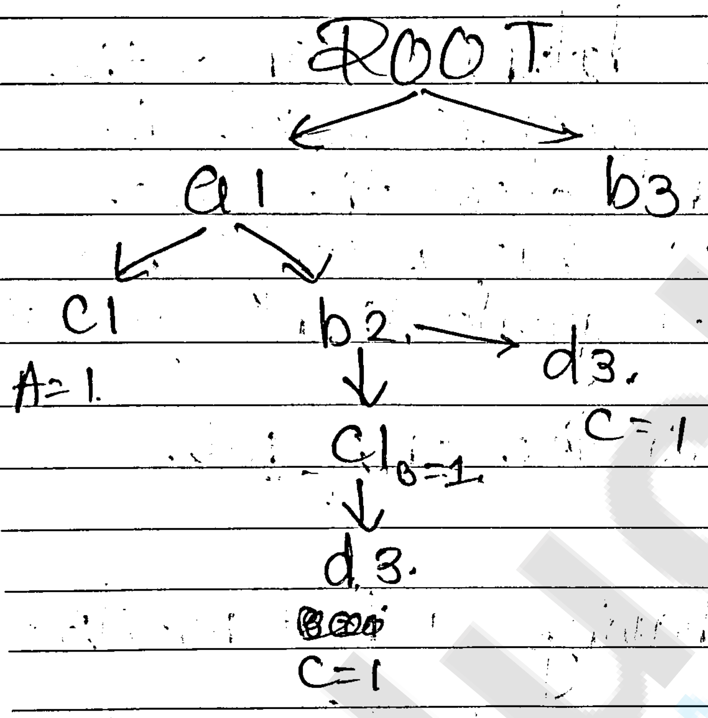
$a1 \rightarrow b2 \rightarrow d3$



Scan of 5^{th} transaction. (a1, b2, c1, d3).

Extracted (a1, b2, c1, d3)

a1 → b2 → c1 → d3



Classification Association Rule: can be generated by dividing all rules into subsets without overlap.

1. The rules having d3 value.
2. The rules having c1 but not d3.
3. The rules having b2 without d3 & c1.
4. The rules having only a1.

Subset of rules having d3 value.

There are 3 samples represented in the FP tree with d3.

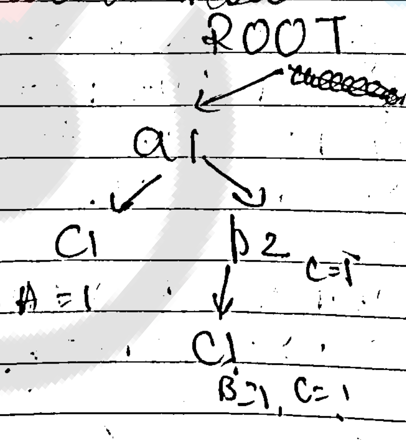
They are a1, b2, c1, d3 : C. | 1.

	support	confidence
$a_1 \rightarrow b_2 \rightarrow c_3$: C	2	$2/2 \times 100 = 100\%$
d_3 : A	1	

In the d3 projected database since the pattern a1, b2, c3 occurs twice i.e support = 2 and also the rule $(a_1, b_2, d_3) \rightarrow C$ has a confidence 100% which is greater than equals 70%

So $(a_1, b_2, d_3) \rightarrow C$ is the only rule generated

(2) Subset of rule having c1 but no d3.
~~There are 2 sa~~
After a search for rules having d3 value all the nodes of d3 and their correspon ding class labels are merged into their parent nodes at the FP tree.



There are 3 samples represented in FP tree with c1

	Support	Confidence
a1 c1 : A	1	
a1 b2 c1 : B	1	
a1 b2 c1 C	1	

all the rules are with < 2 value. so no rule will be generated.

③ Subset of rule having b2 but not d3, c1.

ROOT

a1 A=1.

↓

b2. C=2, B=1

There are 2 samples with b1

a1, b2 : C	2	$\frac{2}{3} \times 100 = 66.66\%$
a1, b2 B	1	

rule (a1, b2) → C has support value 2 but confidence < 70%
So no rule will be generated

④ Subset with a3.

$$\text{ROOT} \downarrow a1$$

A=1 B=1 C=2.

↺ sample

$a1 \rightarrow A$	1		
$a1 \rightarrow B$	1		
$a1 \rightarrow C$	2.	$\frac{2}{4} \times 100$	50%

$(a_1) \rightarrow C$ has support value 2 but confidence < 70%
So this rule will not be considered.

∴ the only association rule generated from the trending process with the above given db is $(a1, b2, d3) \rightarrow C$ with support value 2 & confidence 100%.

cont... ~~out~~ back.

DMBI

phase 2:
- Classification or testing:
for a new sample CMR collects the subset of rules matching the samples from the total set of rules
- If all the rules have same class CMAR simply assigns that level to the new sample
- If the rules are not consistent in the class level CMAR divides the rules into groups according to the class level & find the strongest ROOT.
- CMAR uses the strongest rule in the group as its representative i.e the rule with highest χ^2