ClassRoom
Page No.
Date / / /

SE-417

Data Warehousing & Data Mining

Unit 1 → warehousing
Unit 2-3-4 → 3 different techniques of Data Mining

Data Warehousing

DBMS
SQL
SYSTEM & SOFTWARE

↳ stores the functional data from DBMS which is not operational anymore.
↳ Historic Data
⇒ strategic information

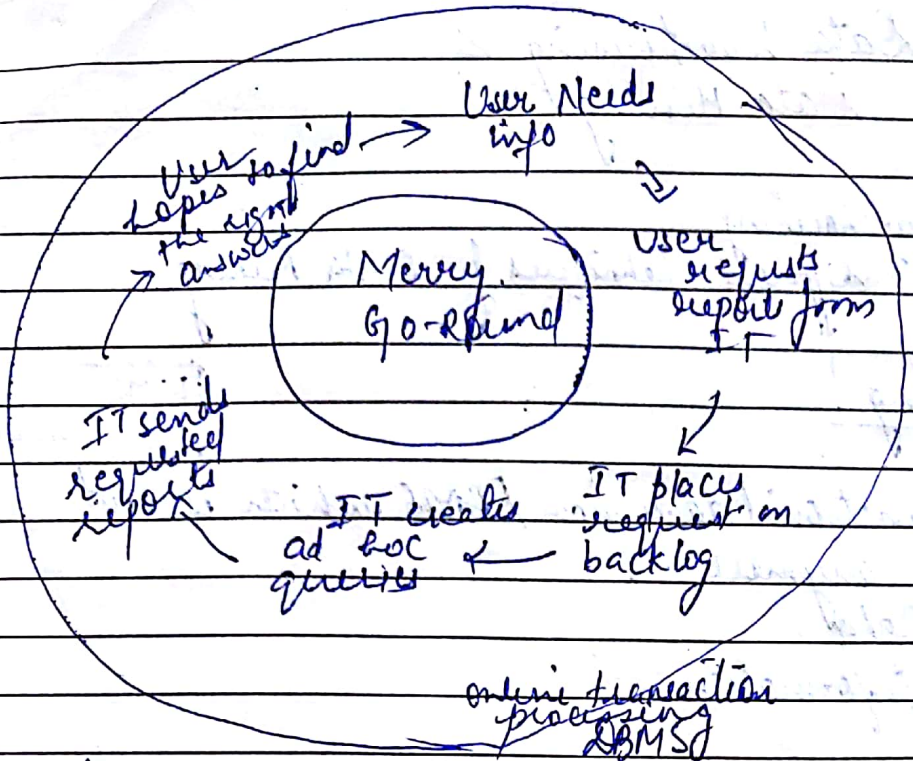
Data Warehouse

It is an informational environment that → provides
(i) an integrated and total view of an Enterprise
(ii) makes the interprisis current and historic information easily available for strategic decision making
(iii) makes decision support transaction possible without hindering the operational systems
(iv) renders the organizations information consistent
(v) presents a flexible and interactive source of strategic information.

Earlier proposed Decision Support Systems.

1) Ad hoc Reports
2) special Extract reports
3) Information Centres
4) small applications
5) Decision Suppod System → Digitized
6) Executive Informatio System

ClassRoom
Page No.
Date



Challenges in all earlier decision Support systems

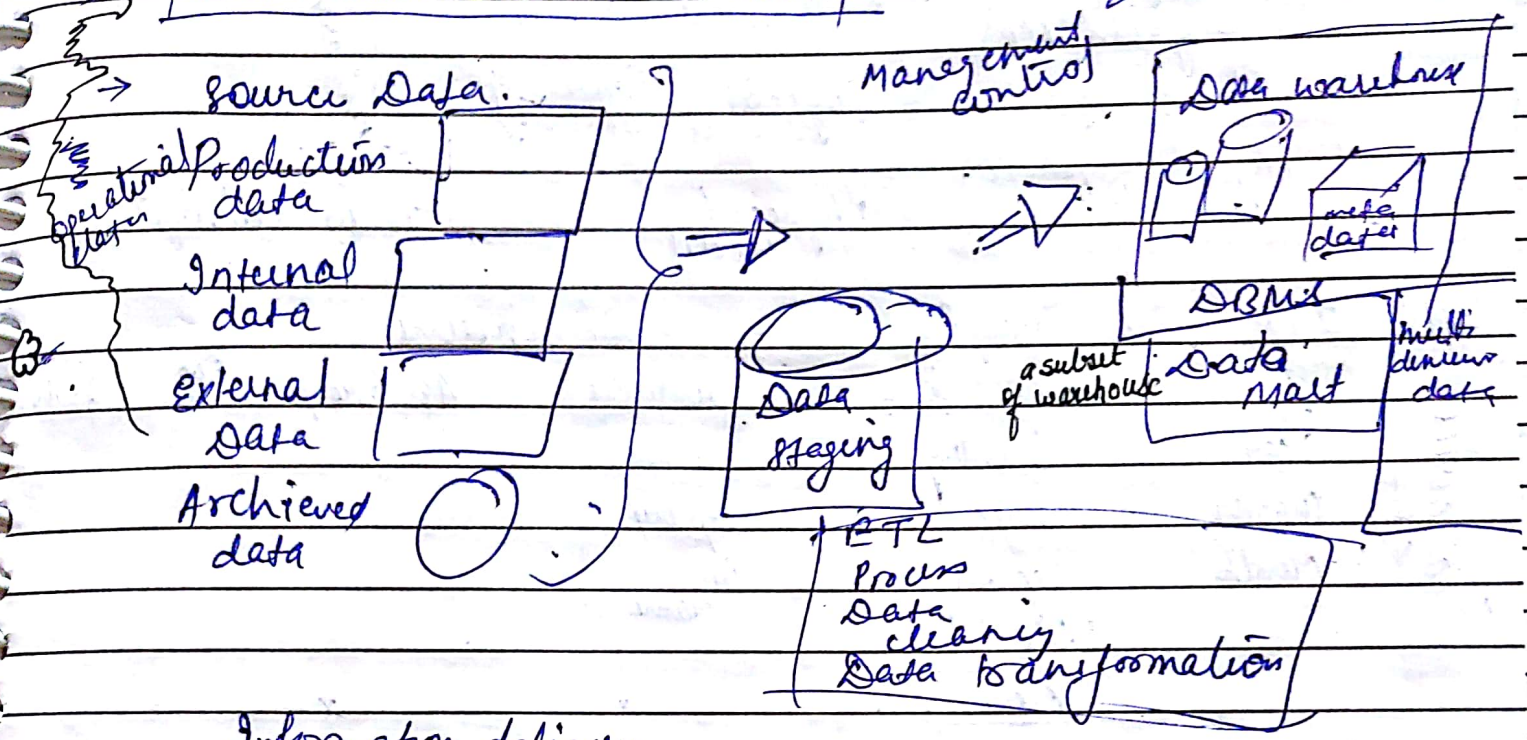
Power

online transaction processing

#	OTLP systems DBMS	DW Informational System
	Operation System	Decision Support System
Data Content	Current Values	Historic, archived & aggregated
Access frequency	Very high	Very low.
no of users	Very large Number	Small numbers (Executives only)
type of access	Read, write & update	only read.
usage	Predictable, Repetitive	Random, Ad-hoc Heuristic
Response time	Subseconds	several seconds to few minutes
Data structure	optimized for transaction	optimized for complex Queries
Design Goal	designed to put the data into the database making the wheels of Business turn	designed to get strategic information out of database watching the wheels of Business turn

Data Warehouse Architecture

=> Data Warehouse Architecture



Information delivery
Data mining
OLAP
Query/Report

Purpose

Features of Data in Data Warehouse.

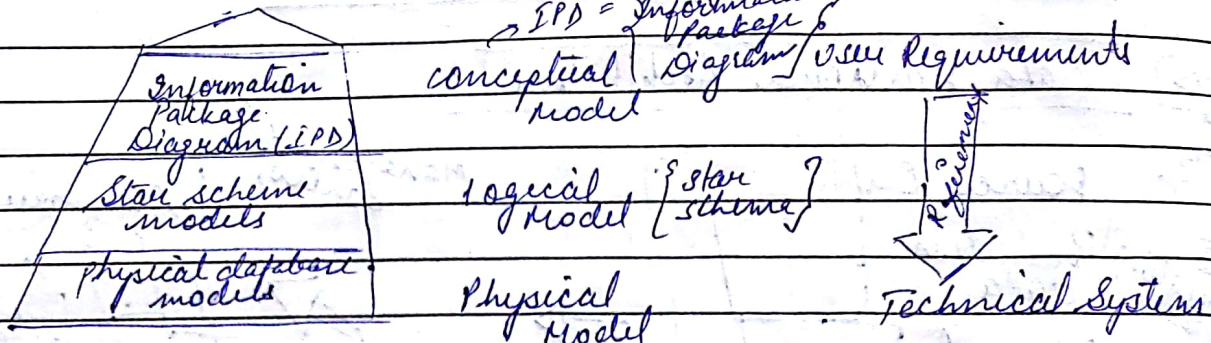
Accounts →
Academic →
Department →
Student

1. Integrated
2. Subject oriented
3. Non Volatile
4. Time variant

Best tool for Data Warehousing & OLAP is Oracle used DBS.

5th Sep 2018

Data Warehouse Modeling



#. IPD / Sales Analysis → Dimensions

	Time	Locations	Products	Age Groups	Edu class	Gender
categories	Year	Country	class			
	Quarter	Area	Group			
	Month	Region	Product Name			
		District				
		Store				

Measures : Forecast Sales, Budget Sales, Actual Sales, Forecast Variance, Budget variance.

Warehouse is based on business organisation. Based on that an IPD is created. For example → Sales Analysis

categories are hierarchial
Time will always be there

Q) Make an IPD for Employee Performance Analysis

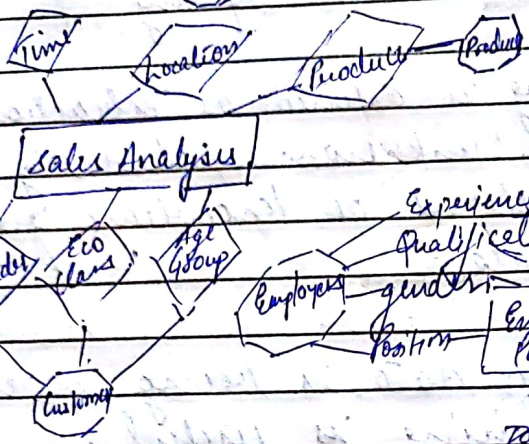
Time	Position	Achieved Targets	Gender	Location	Experience	Qualification
Year	Manager	Projects	Male	Country		
Month	Team Leader	Modules	Female	State		
Week	Employee			District		
				Branch		

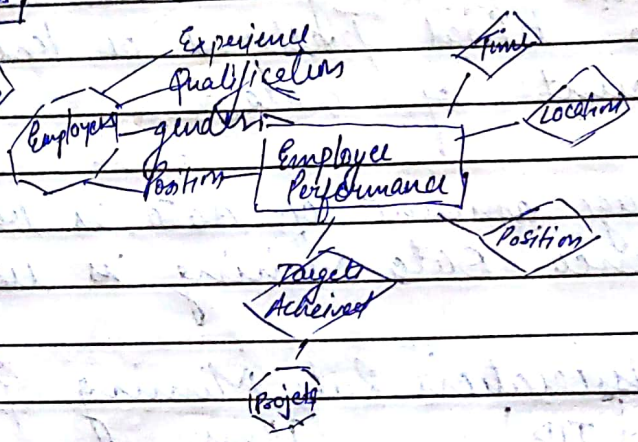
Measures : Daily Working hours, Avg working hours, Expected working hours

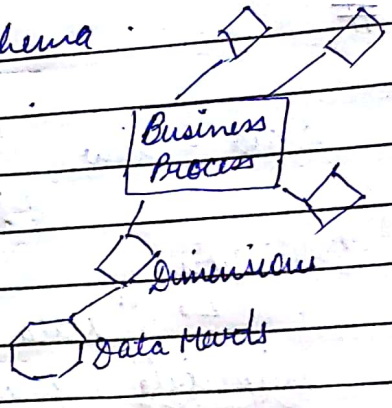
Page No.
Date

Star Schema => Convert IPD to star schema.

- ▭ → business process.
- ◇ → dimensions
- ⬡ → Data Mart







7th Sep 2018

Data

Two theoretical Data Mining topics

Class Room

Page No.
Date

7 Sep 2018

DATA MINING

what, why

KDD = knowledge discovery data

Data = Raw $\xrightarrow{\text{processed form}}$ Information $\xrightarrow{\text{further refining}}$ Patterns Knowledge

Definition

↳ "Data Mining is defined as extraction of interesting, implicit, previously unknown, potentially useful and non trivial patterns or knowledge from huge amount of data"

Huge amount of data is being generated everyday. Hence Data Mining is useful.

8

Association Rule Mining { Transaction Data Set }

TID	Items
T0	Beer, Nuts, Diaper
T1	Beer, Coffee, Diaper
T2	Nuts, Eggs, Milk
T3	Beer, diaper, Eggs.
T40	Nuts, Coffee, Diaper, Eggs, Milk

3

I = {Beer, Nuts, Diapers, Coffee, Eggs, Milk}

D : Transaction database

I : Itemset

T : event $T \subseteq I$

T contains itemset X iff $X \subseteq T$

Association Rule :- $X \rightarrow Y$

$X \subset I$, $Y \subset I$ and $X \cap Y = \phi$

$X \rightarrow Y$: support $= \frac{|X \cup Y|}{|D|}$ can be calculated for individual items & itemsets as well.

confidence $= \frac{|X \cup Y|}{|X|}$

Apriori Algorithm => gives frequent itemsets

Support ≥ 50% (0.5) Confidence (0.5)

Iteration 1 :- Candidates	Frequent
F1	Beer, Nuts, Diaper, Eggs
1-itemset Beer, Nuts, diaper	
Coffee, Eggs, Milk	

Page No.
Date

Iteration 2. : Candidates

F2
2-itemsets. [Beer, Nuts], (Beer, Diaper)
~~(Beer, Coffee)~~, (Beer, Eggs) ~~(Beer, Milk)~~
(Nuts, Diaper), ~~(Nuts, Coffee)~~, (Nuts, Eggs) ~~(Nuts, Milk)~~
~~(Diaper, Coffee)~~ (Diaper, Eggs), ~~(Diaper, Milk)~~
~~(Coffee, Eggs)~~, ~~(Coffee, Milk)~~, ~~(Eggs, Milk)~~

Frequent :- (Beer, Diaper) // (Nuts, Eggs)

F3 = ϕ. Association Rule :- Beer → Diaper
Diaper → Beer

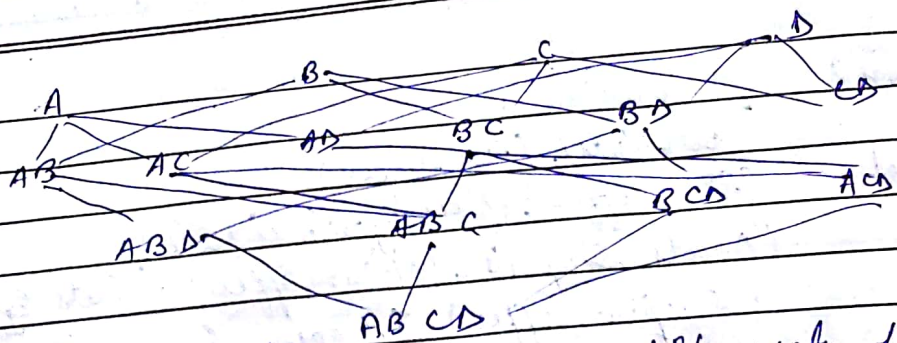
Making Association Rules = RHS contains only 1 item

Confidence (Beer → Diaper) = 3/3 = 1
(Diaper → Beer) = 3/4 = 0.75
} Both are valid

APRIORI ALGORITHM

C_K : Candidate itemsets of size K
L_K : frequent itemsets of size K
L_1 : { frequent items }

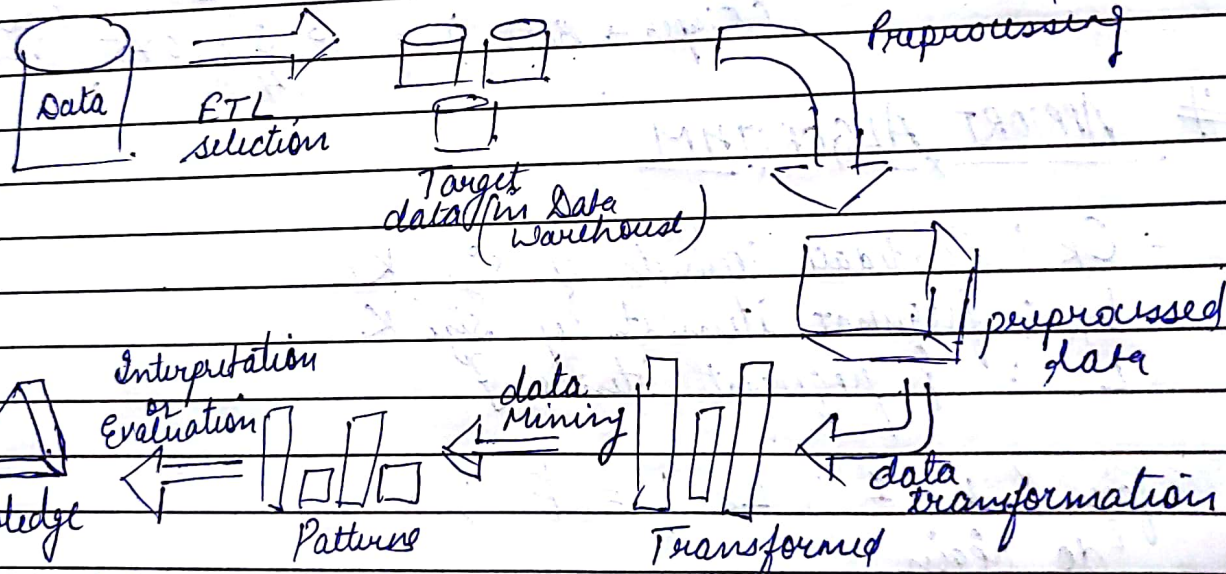
for (K = 1 ; L_K ! = ϕ ; K++)
{ do begin
C_{K+1} = Candidates generated from L_K (cartesian product)
for each transaction t in D.
{ increment the count of all
candidates in C_{K+1} that are contained in t
L_{K+1} = candidates in C_{K+1} with min-support
}
return L_K
}



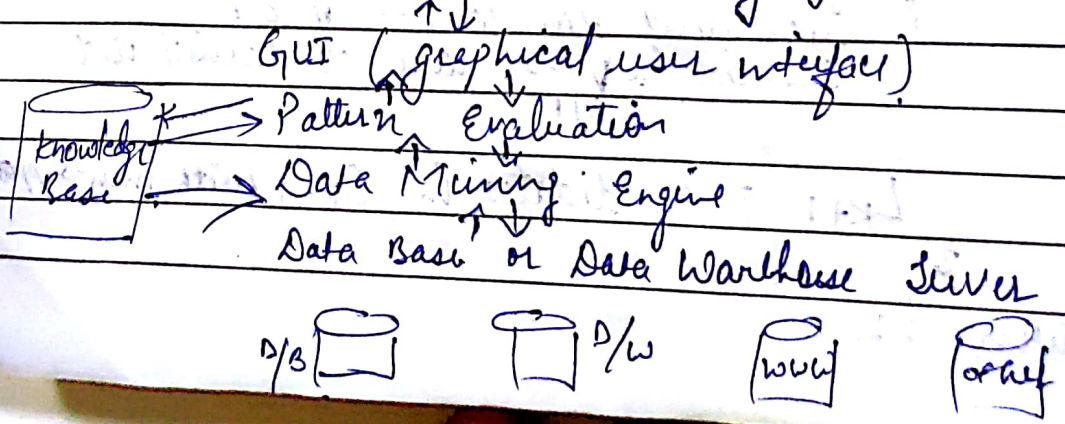
Apriori Principle : combination of itemsets forms a lattice structure.

If an itemset don't satisfy minimum support threshold, then its superset will also not satisfy ⇒ they can be discarded.
≡ PRUNING

KDD knowledge discovery Data
or knowledge discovery in Data Base



Architecture : Data Mining System.



Page No. classroom
Date / / /

12/9/18

Types of Attributes

		Example	operations
1.	Nominal:	Id, number, names. (fall, short, red)	mode, entropy, correlation, χ^2 test, =, ≠
2.	Ordinal:	Ranking, grade, Height	median, percentiles, <, >, rank correlation, run tests, sign tests
3.	Interval:	Calander data, temperature in °C or °F	=, ≠, mean, <, >, t & F tests
4.	Ratio	marks in (0 to 100) temperature in Kelvin, time length	=, ≠, GM, HM, <, >, +, -, ×, /, percent variation

Why Data Preprocessing :→

Data Quality problems → ① Noise & outliers
② missing values
③ Duplicate values.

Data Preprocessing :- to improve data mining analysis with respect to time, cost & quality.

① Aggregation ② Sampling ③ Dimensionality Reduction.

Why data Preprocessing ? ⇒ redundant data, Missing data, ① to improve data quality.

② after preprocessing data may become more valuable

③ reduce computational load as it may load unnecessary things thus we remove it

Outliers # elements do not belonging to that data.

Aggregation → n way aggregation {along n dimension}

disadvantages: information may be lost

Sales Analysis

Time	Location	Product
year (1)	Country (10)	Category (50)
Quater (4)	State (80)	Sub Category (500)
month (12)	District (360)	Product (1-5000)
Day (365)	Town (3000)	

→ compilation cost reduces.
→ higher level view of data
→ gives a better "...."
→ time save