11:25 to 12:20. LH-30 MK publication

Han & Kamber 2nd edn

9/7/'18 Data Warehousing & Mining

Data Mining Concepts & Techniques

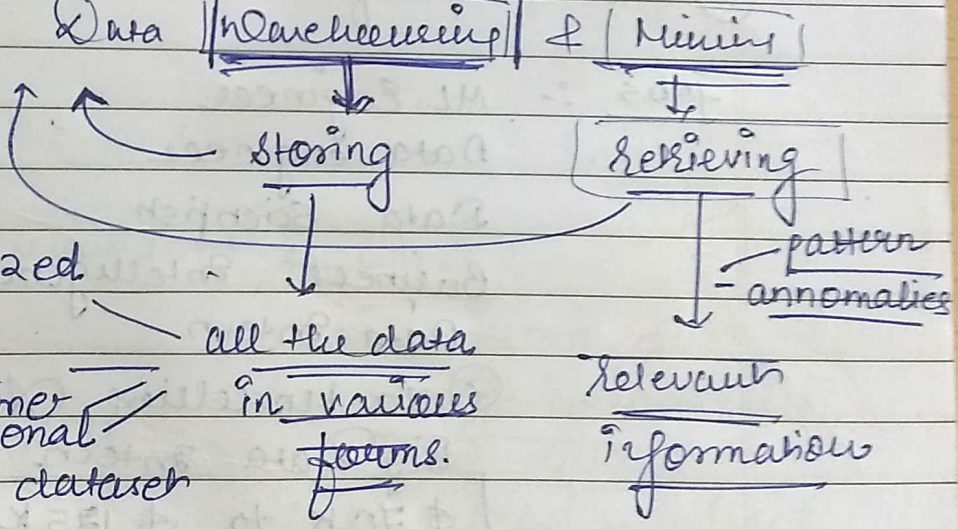
Lecture 1

- Subject Intro.
- Applications & Need
- Scope
- Importance of DM
- Architecture
- functionalities

contents :-

- Introduction to the Subj
- Importance of the
- Applicability of the
- Unit 1
  - Importance of DM
  - ~~Data Mining functionalities~~
  - ~~Classification of DM systems.~~

Introduction to subject



IoT + ML + AI → Increasing data over the cloud.

huge data collection so as to analyze and find insights. Data collected in a data warehouse, mining is cleaning - data filling - visualizing the findings

9/7/'18 Data Warehousing & Mining

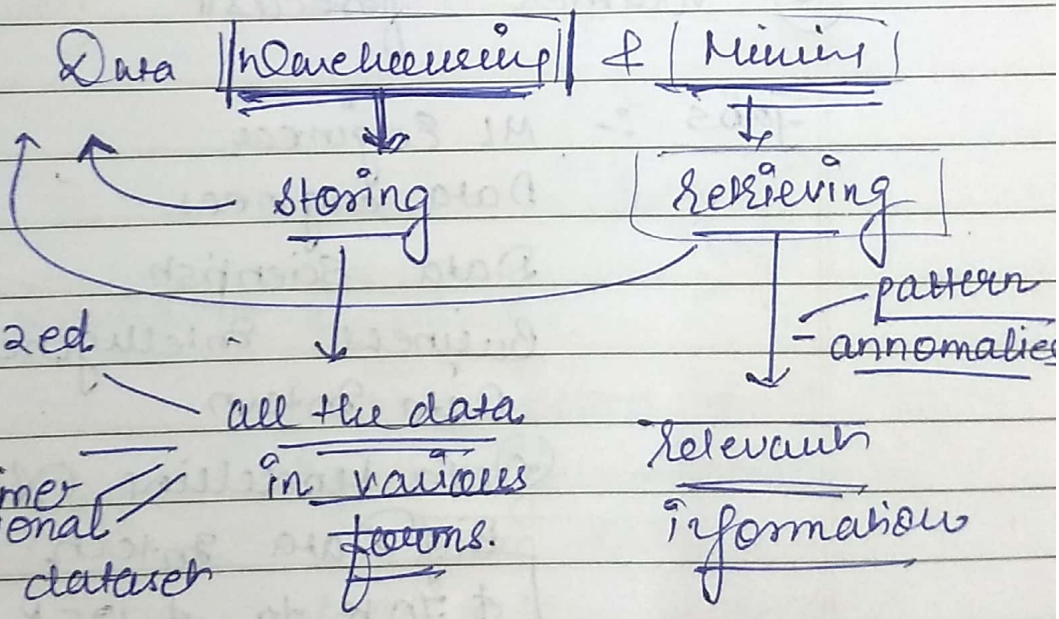
Data Mining Concepts & Techn

Lecture 1

contents :-

- Introduction to the Subject
- Importance of the Subject
- Applicability of the Subject.
- Unit 1
  - Importance of DM & DM Architectur
  - ~~Data Mining Functionalities~~
  - ~~Classification of DM Systems.~~

Introduction to subject



IoT + ML + AI → Increasing data over the cloud.

huge data collection so as to analyze and find insights. Data collected in a data warehouse, mining is cleaning - data filling - visualizing the findings

Learn the huge sets of data, looking for patterns, anomalies, associations, with the goal of extracting values.

1. market analysis
2. online transaction theft detection
3. improving user experience based on the features based on user choice.
4. Strategic decisions in companies
5. energy saving of appliances based on usage
6. weather forecast

jobs :- ML Engineer
Data engineer
Data Scientish
Business Intelligence Analyst
Data Intern
Data Modelling Analyst
Big Data Intern

$70K to $125K / year

tools :- { Spark, R, Hadoop } | Java, Python - Languages

Microsoft Big Data Certificate Program

Size of the digital universe will double every 2 years

50× growth rate of data.

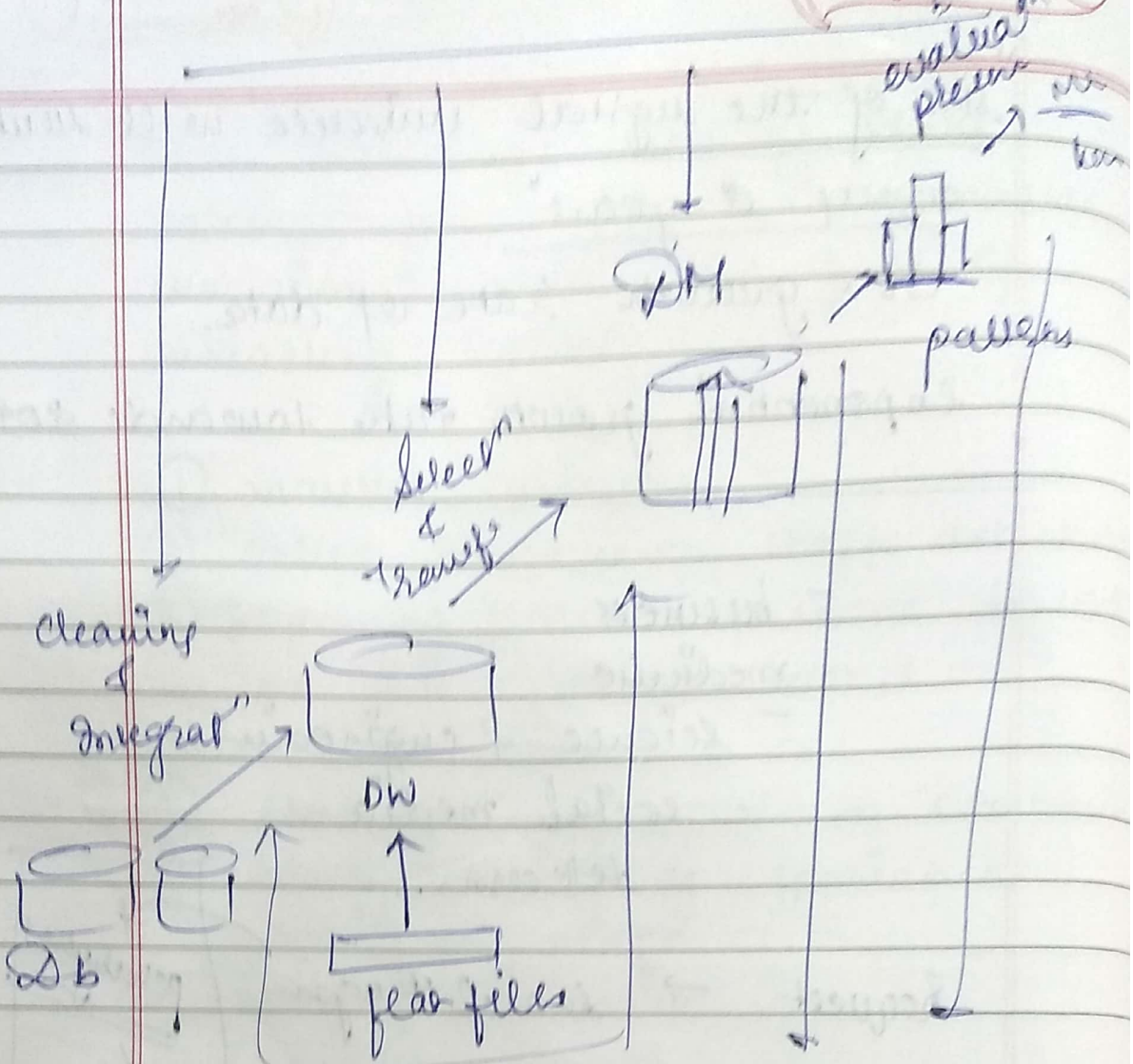
Exponential growth rate towards 2020.

- business
- medicine
- science & engineering
- social media
- telecom

Request → Search Engine → Databases

Relevant info.

Data mining → knowledge mining from data.
knowledge discovery from data
KDD
Alternative pattern analysis
data dragging
data archaeology



preprocessing

cleaning - Remove noise
fill incomplete
Remove inconsistency

integrat - multiple sources combined

Select - Relevant to analysis
data retrieved

Transformat - data transformed &
consolidated into forms
appropriate for mining.

pattern evaluation - To identify interesting patterns rep: knowledge based on ~~data~~

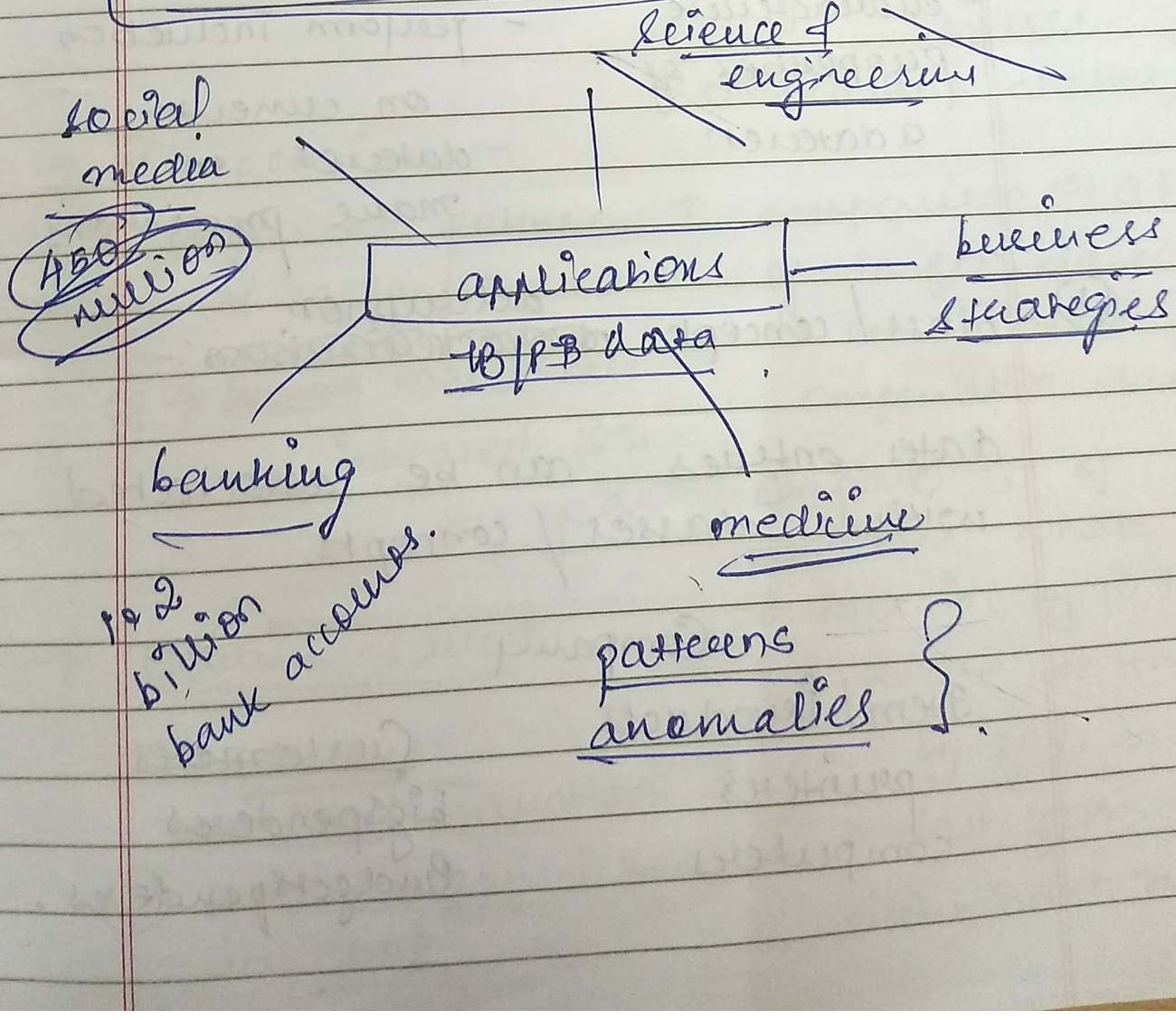
knowledge representation - visualization.

Whatsapp

- 450 million users
- 55 billion msg/day
- 45 million photos
- 1 million videos.

2018 stats

1.02 billion - bank accounts.



Data Mining Functionalities

- characterization } class/concept description
- discrimination
- mining frequent patterns
- associations & correlations
- classification
- Regression
- cluster analysis
- outlier analysis

task

descriptive
- characterize properties of a dataset

predictive
- perform induction on current dataset to make predictions

① class/concept description ~~characterization~~:-

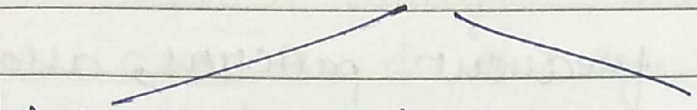
data entries can be associated with classes / concepts

Company

Item / Products
printers
computers

Customers
bigspenders
Budgetspenders.

this helps define a group in the form of a concept in a summarized, precise form.
& this description is called class/ concept description.



summarizing the data following a class in general terms (target class)

by comparison of our target class with a set of comparative classes. (contrasting classes)

→ characterisation of general features of a target class

product

eg → ~~people~~ whose sale increased 10% this year

→ comparison of GF of target class to GF of other comparative class.

compare GF of products whose sale inc by 10% & whose decreased by 30% in last year

→ techniques used

① attribute oriented induction

② roll up operation in OLAP

discriminant rules

Output - pie chart,
bar chart, MD cubes,
MD tables,
generalised
relations.

(2) Mining frequent patterns, associations & ~~path~~ correlations

→ Frequent pattern - appears frequently in your dataset.

frequent pattern :- ~~set of i~~

frequent itemset - set of items that appear together in a transaction

frequent subsequence - an item purchased (sequential) followed by another item,

frequent substructure - in the form of a graph/tree is substructure

↳ associations & correlations

(3)

association analysis -

buys (X, "computer") → buys (X, "s/w")

[support = 1%. confidence = 50%]

confidence 50% means if customer buys a laptop there are 50% chances that he will buy S/w as well.

support 1% means 1% transactions under analysis show that computer & S/w are purchased together.

by single dimensional Rules

dependent on 1 attribute

multidimensional Rules

dependent on multiple attributes

age (X, "29") ∧ income (X, 25K) →

buys (X, laptop)

minimum support & confidence threshold must be satisfied.

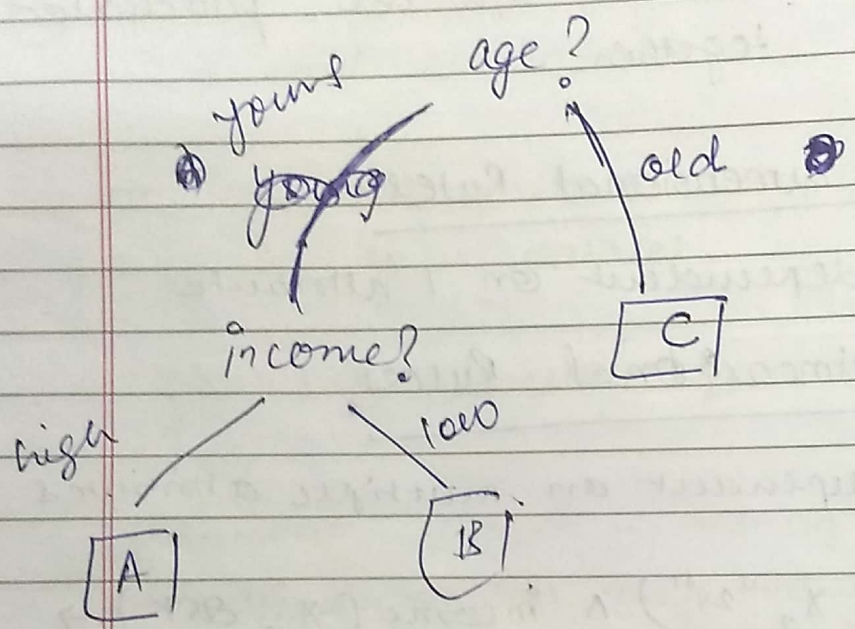
(2) classification & Regression.

classification - finding a model that describes & distinguishes data classes / concepts.

Model derived from analysis of the training dataset.

50%]

This model is used to then predict class labels of objects whose labels are unknown
- classification rules, decision trees, neural n/w, mathematical formulae



age (X, "young") → ∧ incom(X, "high")
↓ ↓
income (X, "low") cla (X, "A")
↓
clas ("B")

age (X, "middle") → C

age (X, "senior") → C.

classification

categorical (unlabelled data, discrete)

Regression - models continuous values to predict missing / unavailable numerical data values

Regression analysis is used for numeric prediction.

It encompasses idea of distribution trends over data

Both preceded by Relevance Analysis

identifies attributes relevant to classification / regression process.

4) Cluster Analysis

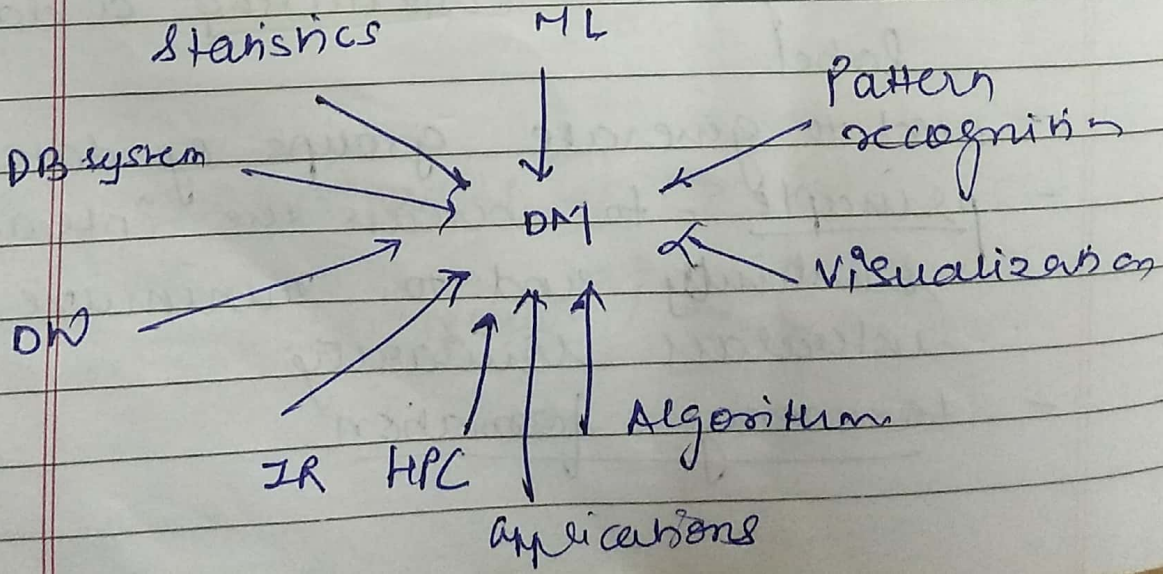
- classification & regression analyze labelled data whereas cluster analysis is constructing a class label.
- used to generate groups of data
- principle - to increase the intraclass similarity and to minimize the interclass similarity
- taxonomy formation

(5) Outlier Analysis

- data set may contain data entries that do not comply with the general behaviour / modelling of data. These data objects are called outliers. / noise / exception
- In fraud detection, outlier is an interesting pattern than the regularly occuring pattern.
- anomaly mining
- cluster analysis can be used for the purpose wherein an object distant from any cluster is an outlier

Classification of DM systems

- Domains that influence the development of DM methods.
- application-driven field



(1) Statistics

- study collection, analysis, interpretation & explanation of presence data
- statistical model - mathematical func^n describing behaviour of objects in a target class
  eg: classification & data characterization
  statistical model can be built

Also, a DM task can be built on top of a statistical model.

(2) ML

- these systems can learn on the basis of data.
- automatically recognize patterns & make intelligent decisions based on it

→ Supervised learning :- synonym for classification.
- learning from the labelled example of the training dataset.

→ Unsupervised learning - clustering
- training data / ip examples are not class labelled.
- grouping on general feature but semantic meaning not drawn.

Semi-supervised - Both labelled & unlabelled.
labelled - labels used to classify
unlabelled - refine boundaries.

Active learning :- users are asked to label the examples so as to keep users active in the process

(3) DB Systems & DW

DB systems - creation, maintenance & use of db for org. & end users
- data storage
- indexing
- accessing
- query proc.

DM can be used to extend the capability of conventional DB.
- analysis

DW & DM - DW integrates data originating from multiple sources & various time frames & consolidates in a multidimensional space.

④ IR

→ Science of searching document/info. in a document.

doc → text / multimedia & may over ~~internet~~ web.

IR assumes ① unstructured data

② queries formed from keyword

Text Document / bag of words

Language Model – function that generate these bag of words

Similarity b/w 2 documents is similarity b/w corresponding lang. model.

Topic | text doc may have 1 / more topics. IR & DM help find major topic in a collection of document & for each doc, major topic is involved.

Applications ① Business Intelligence
② Web Search Engines

ates ① BI :- helps acquire better understanding of customers, the market, supply & resources & the competitors.

BI due to DM provides historic, current & predictive views of business operation

customer feedback on similar products, effective market analysis, strength & weakness of competitors, smart decision

OLAP (online Analytical Processing) in BI relies on DW & multidimensional data mining

CRM - customer relat° mgmt
clustering
↑
characterization - general feature
↑
Predictive analysis
↑
classification & prediction

(2) Web Search Engines
↳ specialized computer server that searches for info on the web.

- search results returned as hits (list)
- web page, images
- Search engines, working algorithmically or times search & return results from data available public db/ directories.

- crawling → spider bot / crawler
- indexing
- context awareness
- ranking

} pages.

robot.txt → no crawling allowed

- Scaling DM algos to cloud.
- context awareness.
- previous user query based search, few time - queries.

* Major Issues in DM

① Mining Methodology

(1) - Mining various new kinds of knowledge - tasks
(2) - Mining in Multidimensional dataspace.
(3) - DM - interdisciplinary.
IR + NLP
Big Mining + SE
(4) - Boosting power of discovery in networked environment

- Handling uncertainty, noise or inconsistency of data
- Pattern evaluation / pattern constraint guided mining

(2) User Interaction
- Interactive mining
- Incorporation of background knowledge. constraints, rules
- ad hoc DM & query languages
- presentation & visualization

(3) efficiency & scalability-
① algorithms.
② parallel, dis, incremental
③

(4) Diverse db

(5) DM & Soc

Multidimensional Data Modelling

Data cube – to be viewed & analysed in various/ dimensions.
multiple.

dimension – perspective / entity with which an org. wants to keep records.

eg Company → item, time, location → branch } sales.

each dimension has a table associated to it, dimension table.

eg item :- item_id
item_name
item_type
item_brand.

These multidimensional data is based around a single theme
i.e sales. Theme is represented by fact table. — key of dimension
measures.

data

Quantity to be analyzed through which, we analyse relationship bet the dimensions.

eg → item_key, time_key, loc_key,
branch_key, units_sold, dollar_sa

time & item & ~~location~~ view for cities

NY, Toronto, Vancouver.

Ahmedabad.

~~Vancouver~~

2D location =

Sales data:

item	home	comp	phone	sec
Q1	6	10	14	18
Q2	7	11	15	9
Q3	8	12	16	20
Q4	9	13	17	21

3D view Ahm, Vad, Surat, Ray

Ahm. Vad. Surat Ray

item. item item item

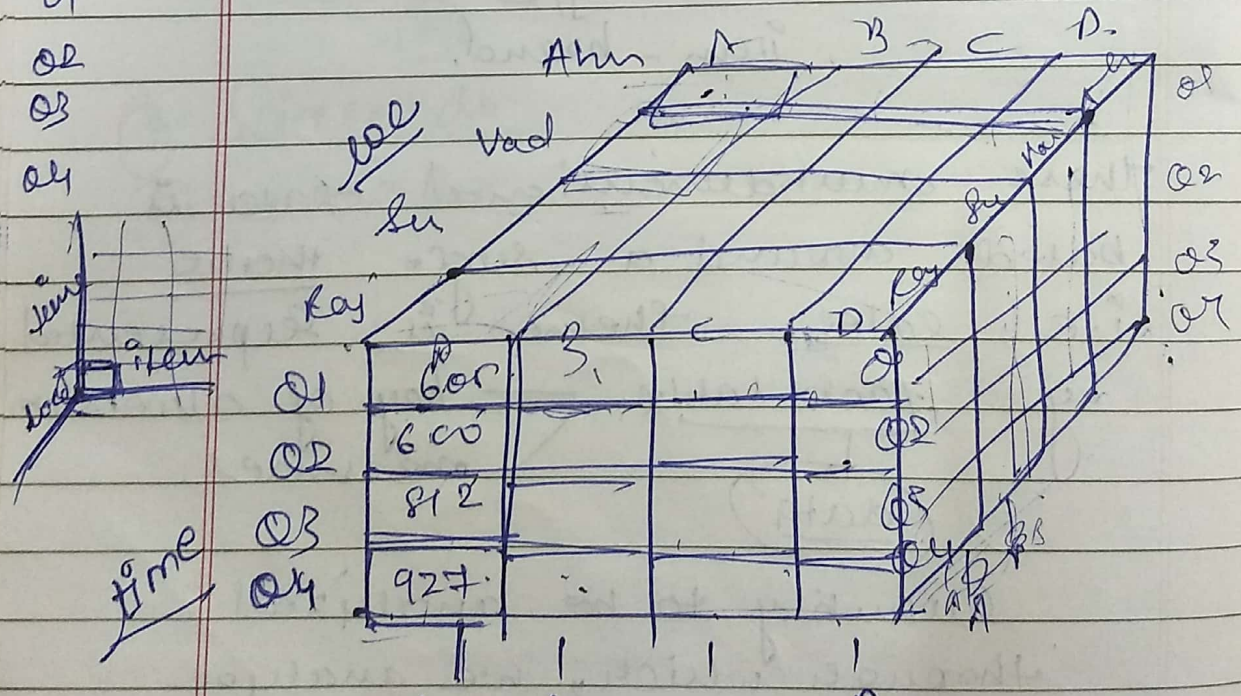
item A B C D item A B C D item A B C D item A B C D

Q1

Q2

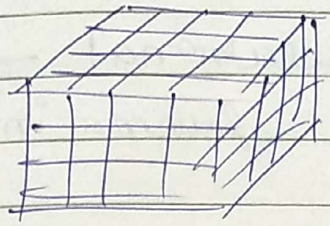
Q3

Q4



item.

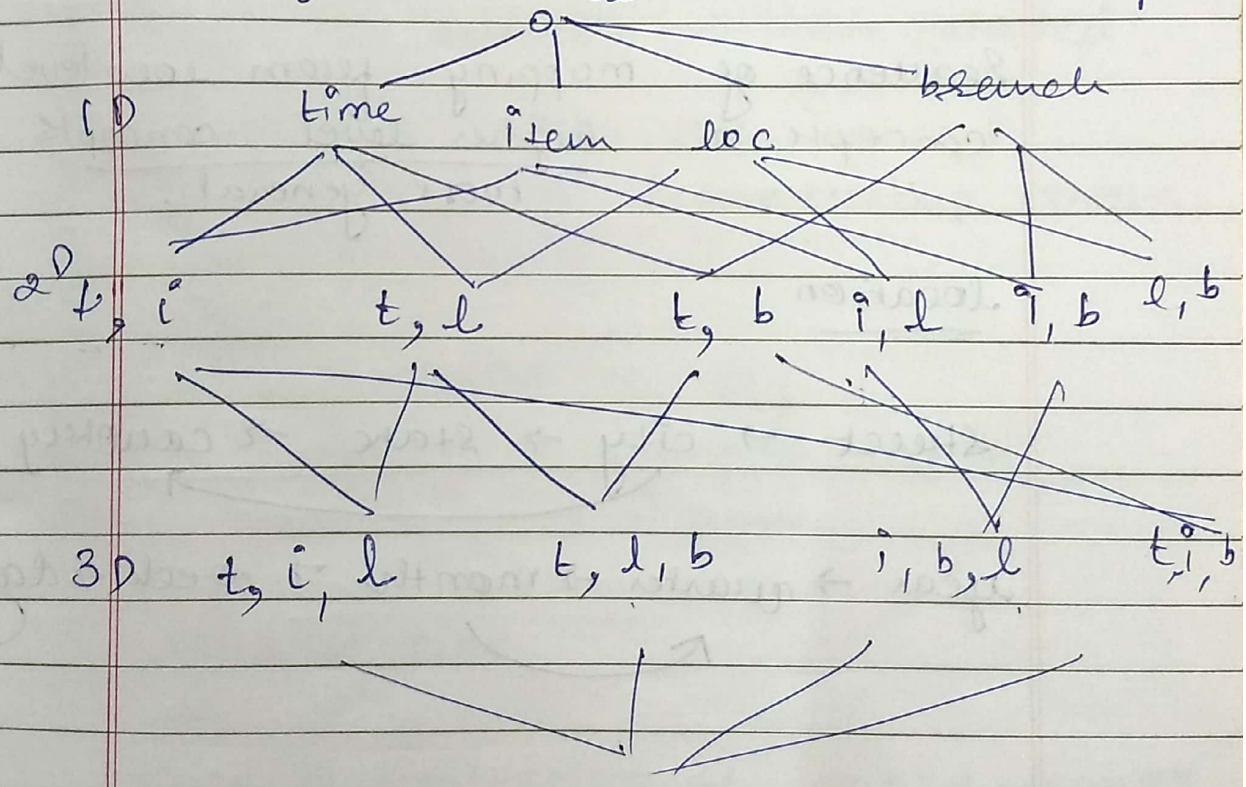
4D



least summarized (most detailed) is called base cuboid. 4D

most summarized (~~least~~ detailed) is called apex cuboid.

0-D all (total sales) apex.



Star Schema

ER diagram in relational → Star schema in DW:

- Star
- snowflake
- constellation, galaxy.

Concept hierarchy

Sequence of mapping from low level concepts to higher level concepts. more general.

location

street → city → state → country

year → quarter → month → week → day

OLAP operations

roll up → drill up.
aggregation on data cube
climbing up the concept hierarchy
attribute reduction.
city
~~state~~ state.
~~city~~ → ~~country~~

one/more dim removed.

drill down → roll ~~up~~ down.

less detailed → more detailed

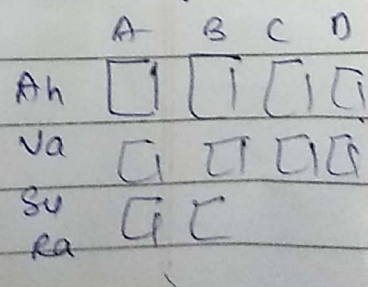
stepping down the concept hierarchy. or introducing more attributes.

country → city.
month → ~~quarter~~ week.
quarter → month.

slice & dice.

slice → selection of one dimension of the give cube resulting in a sub cube.

time = Q1 Q1



dice - selection of 2/more dimen

loc = ahm
quarter : Q1

Ahm

	A	B	C	D
Q1	—	—	—	—

pivot - Rotate

alternative data representation.

item

loc

others - drill across more than 1 fact table
drill through relational sql

- each dimension has a table associated with it - dimension table

eg item
item_key
item_brand
item_type
item_name

eg location
location_key
city
province/state
country

- central theme - fact is represented by fact table

↓

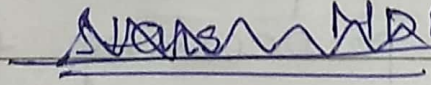
consists of keys of all dimensions & the measures over the relationship amongst these dimensions needs to be analyzed.

fact table

sales
item_key
loc_key
time_key
branch_key
sold_units
earned_profit

ch table

sol



Data Warehouse

DW facilitates higher level executives to effectively analyze, organize & understand data to make strategic decisions.

DW is a huge data repository maintained separately from the org. operational database.

It is a subject-oriented, integrated, time-variant, non volatile collection of data in support of mgmt's decision making process.

Subject oriented - DW is organized around major subjects like customers, sales, products, suppliers - rather than processing the data on a day to day basis the data is analyzed to be used for decision making.

Integrated - usually constructed merging multiple heterogeneous data sources eg Relational db, flat files, online transaction logs etc. Data cleaning & integration done before data merger into the DW.

es

Time-variant - past 5-10 years of data (historic) perespective. Every key structure consists of an implicit/ explicit time key element.

Nonvolatile - Always physically separate from app data in operational env. Dw doesn't require transaction processing, recovery, concurrency control, it only requires loading & access.

Managers / Executives / Analysts - use warehouse to quickly obtain an overview of data & to make sound decision based on info in warehouse.

Eg: ① Inc customer focus (analyzing buying patterns)
② Repositioning products to inc sales
③. Analyzing op & looking for profit
④. CRM.

Heterogeneous data integration-

wrapper/ integrators built on top of multiple/ het. databases. When query posted to a client site,

a meta dictionary used to translate into query appropriate for each individual heterogenous db involved. These queries are mapped & sent to local query processors. Results obtained are integrated into a global answer set.

query-driven approach

update driven → integrated in advance & stored in DW.

OLAP VS OLTP

online transaction processing - day to day activities of an org like inventory, purchase, manufacture, payroll, banking, registrar, & accounting

online analytical processing -

data analysis & decision making. over persistent data

Why DW?

(1) Major reason for separate DB & DW is performance

Operation db is made/tuned on basis of known tasks and workload like indexing, hashing using p-key, searching for particular records, & to optimize "canned" queries

↓

common & stored in

a way that they can easily be executed without to re-enter details of query each time.

DW queries are complex. Involve computation of large data groups at summarized level & may require use of special data organizat°, access, index methods based on MD view.

(2) Operational db has multiple transactions. Concurrency control & recovery mechanism required

OLAP query has read only access of data for sum, aggregation

(B) Diff structure, content, uses

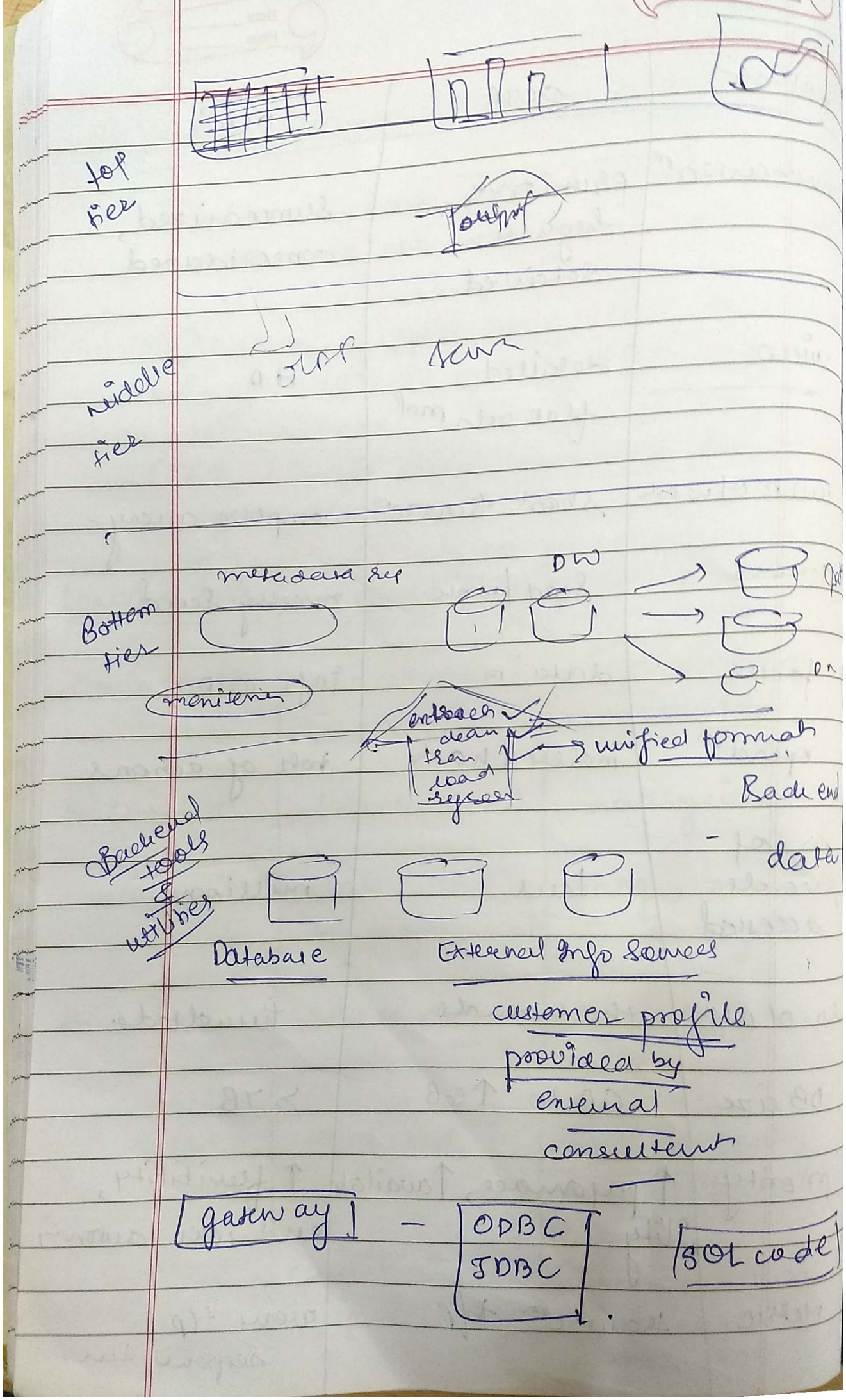
Decision support needs historic data, ODB do not maintain historic data

Decision support requires consolidated (aggregated & consolidated) data from her sources.

ODB contain only detailed recent data of transaction.

feature	OLTP	OLAP
characteristic	operational proc.	informational proce.
execution	transactional	analysis
user	clerk, DBA, DB professional	mng, execute analyst
function	day to day OP.	long term info.
DB design	ER based	star, snowflake, constellation
data	current, up to date	historic, accuracy maintained over time

feature	OLTP	OLAP
summarization	primitive, highly detailed	summarized, consolidated
view	detailed, flat relational	MD
unit of work	short transaction	complex query
access	read/write	mostly read
focus	data in	info out
operation	index/hash	lots of scans
no. of records accessed	tens	millions
no. of users	thousands	hundreds.
DB size	GB to ↑TB	>, TB
priority	↑performance, ↑availability	↑ flexibility, end user autonomy
metric	transaction th/p	query th/p response time.



Metadata repository - contains information about the data warehouse.

① DW structure - schema, view, dimension, hierarchy & derived data defn, data mart location & content

② operational metadata - data lineage (history of migrated data & sequence of transf. applied to it)
- (DW statistics, error report)
- data - archived/active

mark ☆

DW server

③ Algorithms used for data summarization

④ Mapping from operational env to DW - source db, gateway, e, c, t rules

⑤ Data related to system performance - indices to improve data access & retrieval. Refresh/update schedule, replication cycles

⑥ Business ~~mechanism~~ metadata - Business terms & defn, data ownership info & charging policy.

① Extraction - gathers information from multiple heterogeneous sources.

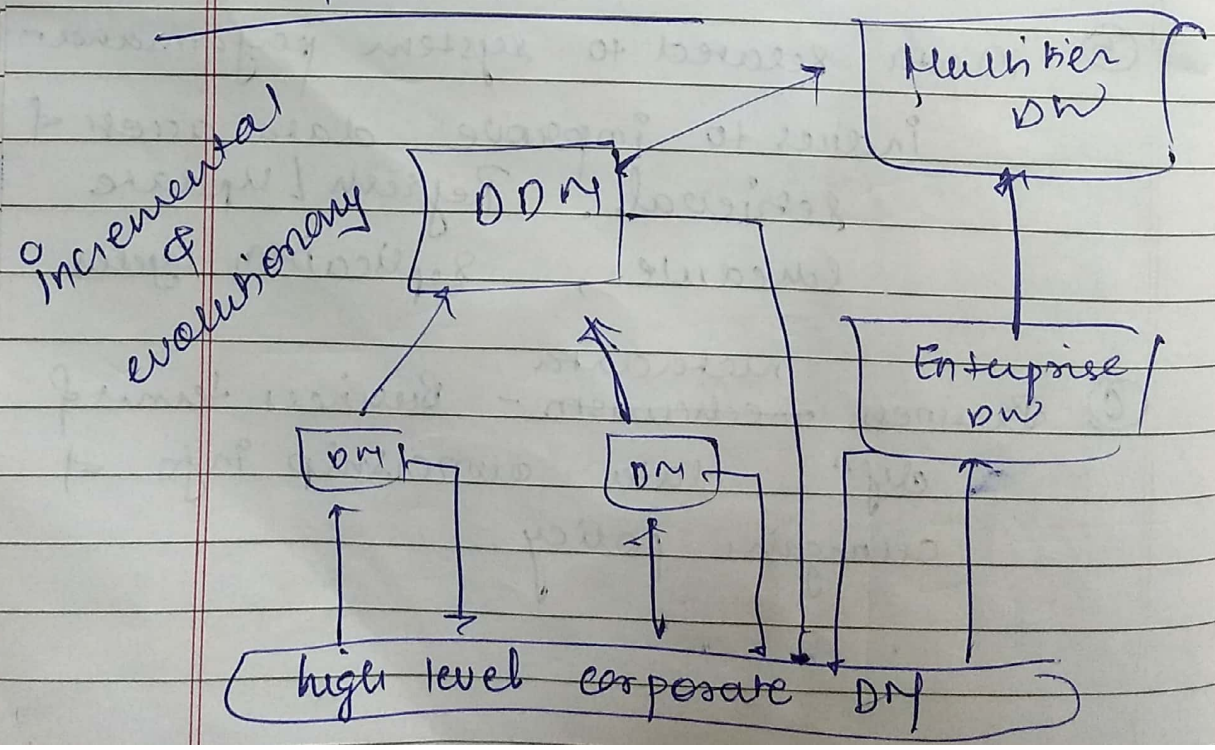
② Cleaning - detects errors & rectifies them.

③ Transformation - converts data from host format to warehouse format

④ Load - sort/summarize/consolidate computes views, checks integrity & partitions.

⑤ Refresh - propagates updates from data sources to DW.

DW, DM & VDW



DW development approach

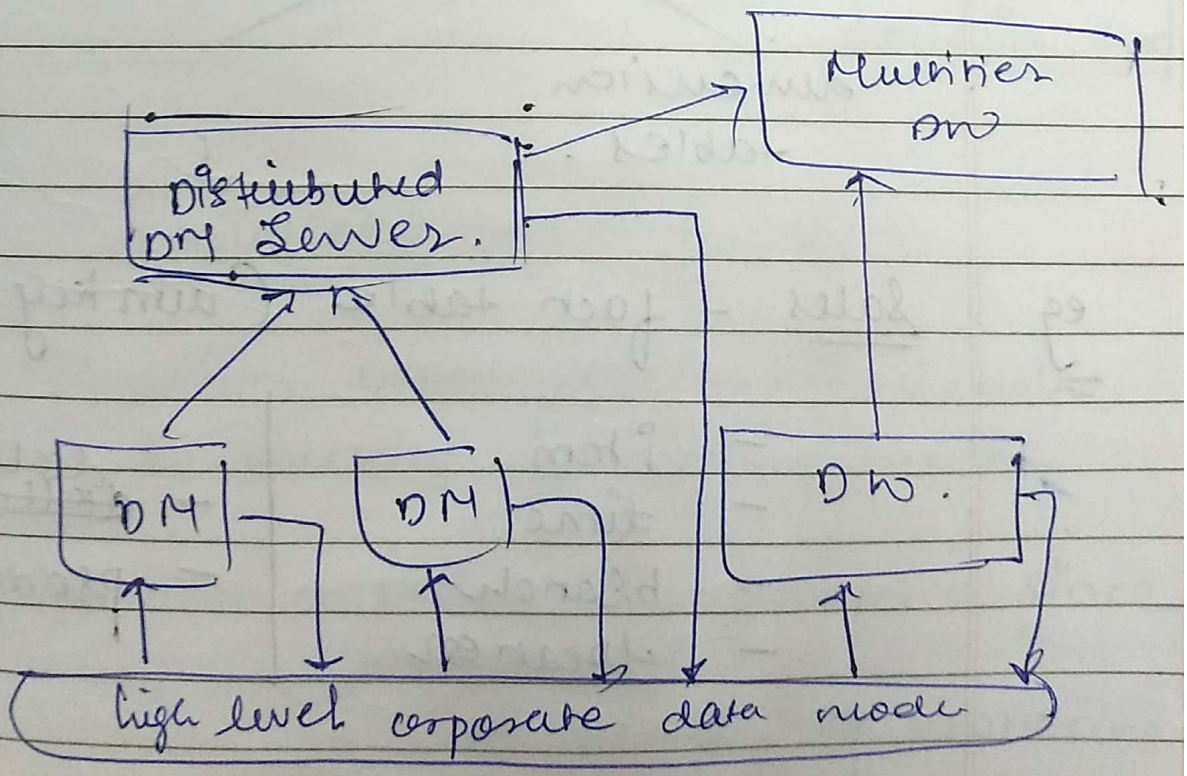
- Incremental & evolutionary.

first - high level corporate data model is defined in a month or two.
- provides integrated view of data among diff subjects & usages. (Chaudhuri model defined later)

Second - Independent Data Marts are prepared in parallel to datawarehouse on the basis of model.

Third - Distributed DM is constructed to integrate diff DM via hub servers.

Finally - Multitier data warehouse is constructed

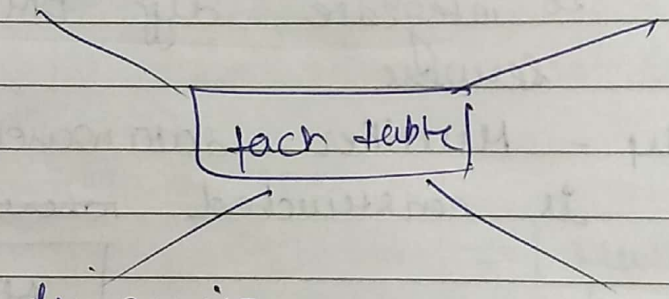


Star schema

① large central table - fact table
bulk of data
with no redundancy
↓
central theme table

② set of smaller attendant table (dimension table) one for each dimension.

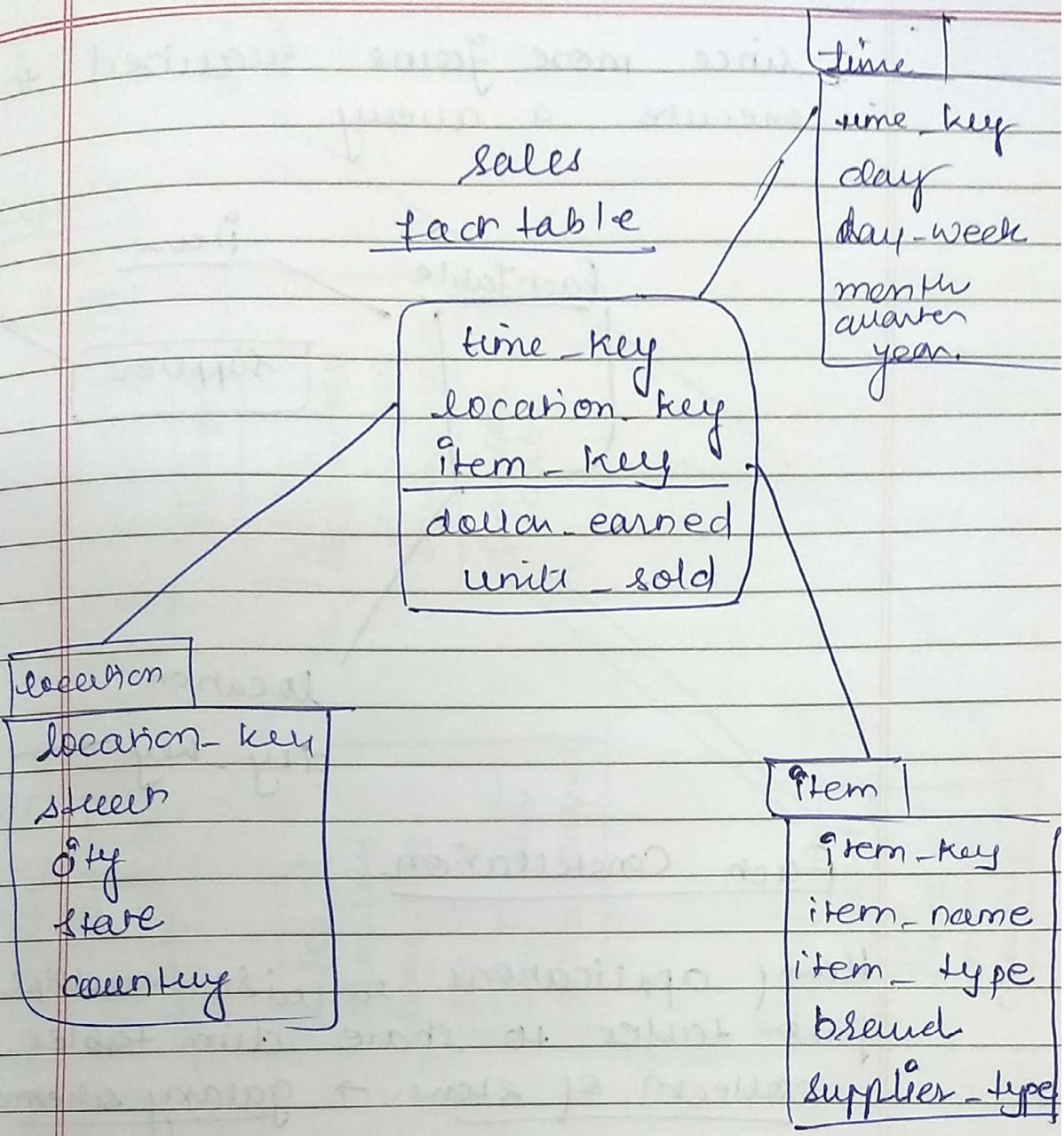
③ Graph resembles a star burst pattern radiates from central object



eg Sales - fact table (dim key + measures)

- item
- time
- branch
- location

- dollar_earned
- ~~dollar_sold~~
- product_sold



earned)

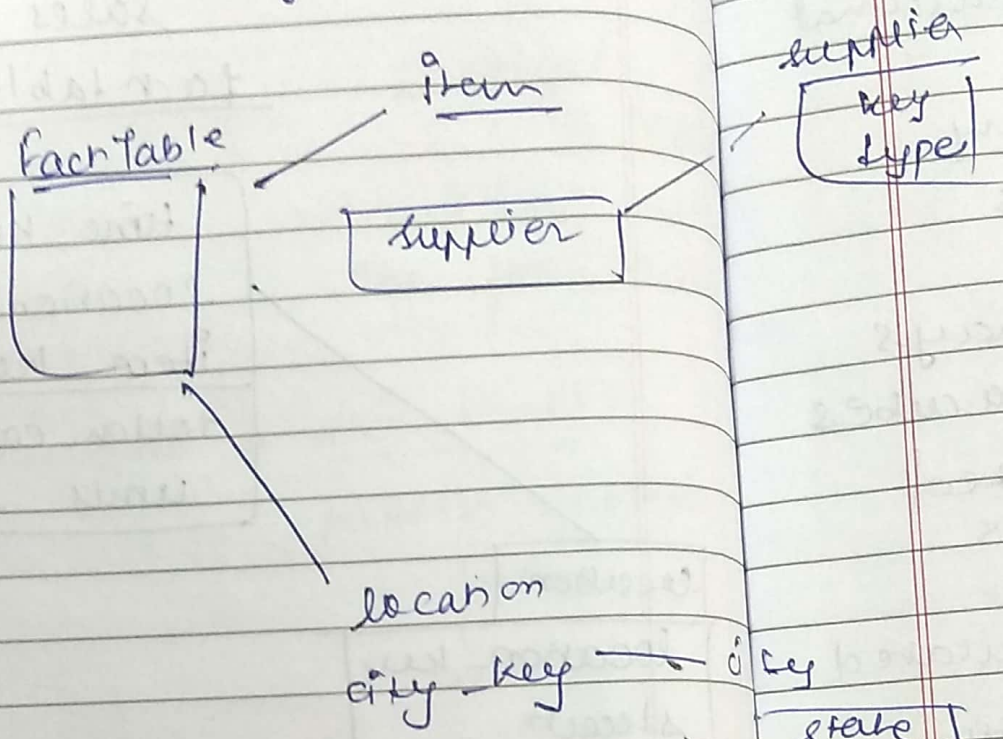
Snowflake Schema

Dimension tables are normalized.
by splitting into more tables.

arned.
d
old

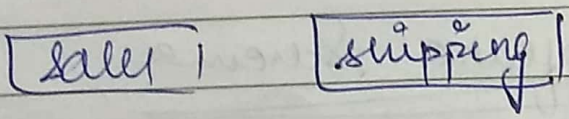
difference - saves storage space since normalized.
reduces effectiveness of browsing

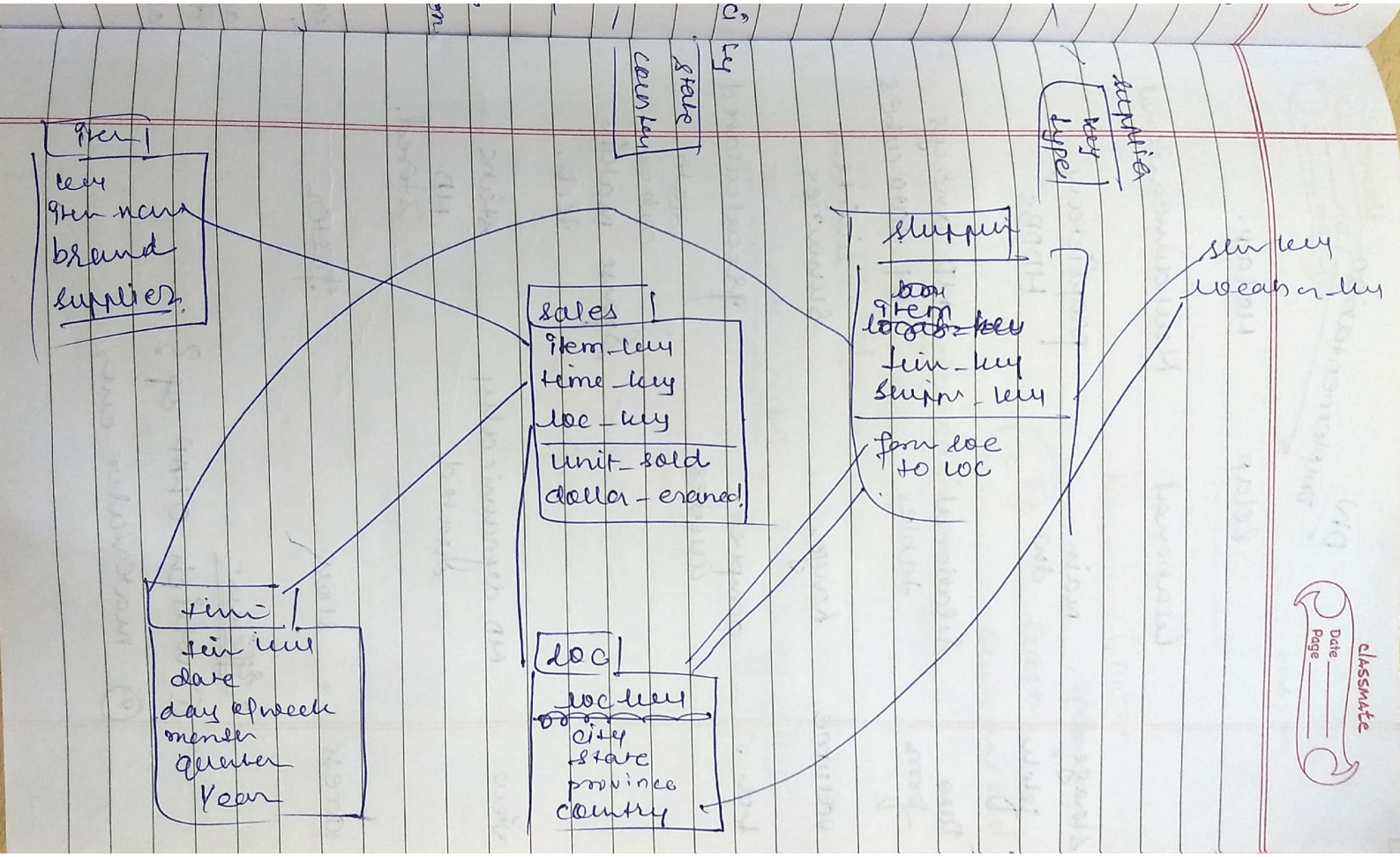
since more joins required to execute a query.



Fact Constellation

Many applications require multiple fact tables to share dim tables.
collection of stars → galaxy schema
fact constellation





DW Implementation

	Rolap	Molap.
	Relational	Multidimensional
Storage & fetch	main dw	proprietary MDDBs
Data form	Relational tables	MD arrays of data cubes
volume	large.	limited summaries
tech.	compiler queries	precalculated data cubes Sparse matrix tech.
View	MD dynamically shared	static MD stored.
access	slow	faster

Efficient

1. which OLAP op?
2. materialize em

apex mesh

nto

full

partial.

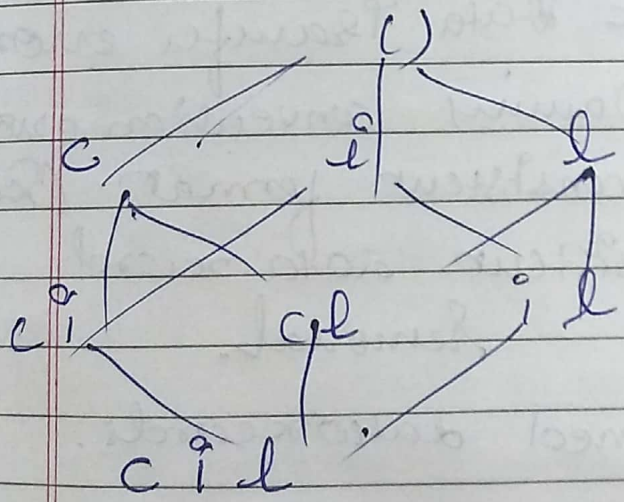
base (most specific, least summarized)

2^n

curse of dimensionality.

group by

compute cube, group by city, item.



Preprocessing

(1) Data Quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability.

Inaccuracy - ✓- machine / system faults
✓- human entry faults

Incompleteness - ✓- Required missing values

Inconsistency - ✓- Data Transfer errors.
✓- Naming convention error.
✓- Inconsistent format (Date)
✓- Inconsistent data record Removal.
✓- Expired date records.

- Address not present in db.
  - ↳ db accurate for SEO (IT) department.
  - ↳ db not accurate for Sales/ marketing dept.

Timeliness - Yearly report

If data not submitted in time is false timeliness.

Believability - If errors in data have earlier resulted in failures, the next time using the same dataset is not trusted by user.

Interpretability - how easy it is to understand data.
Accounting codes in a db are not understood by the Sales Department

Major Task

- Data Cleaning - filling missing values
  - smoothing noisy data
  - removing outliers
  - resolving inconsistencies.

- Data Integration -
  - eid, customer id
  - naming convention to be resolved
  - remove redundancies.

- Data Reduction -

Reduction in volume of data producing same analytical results.

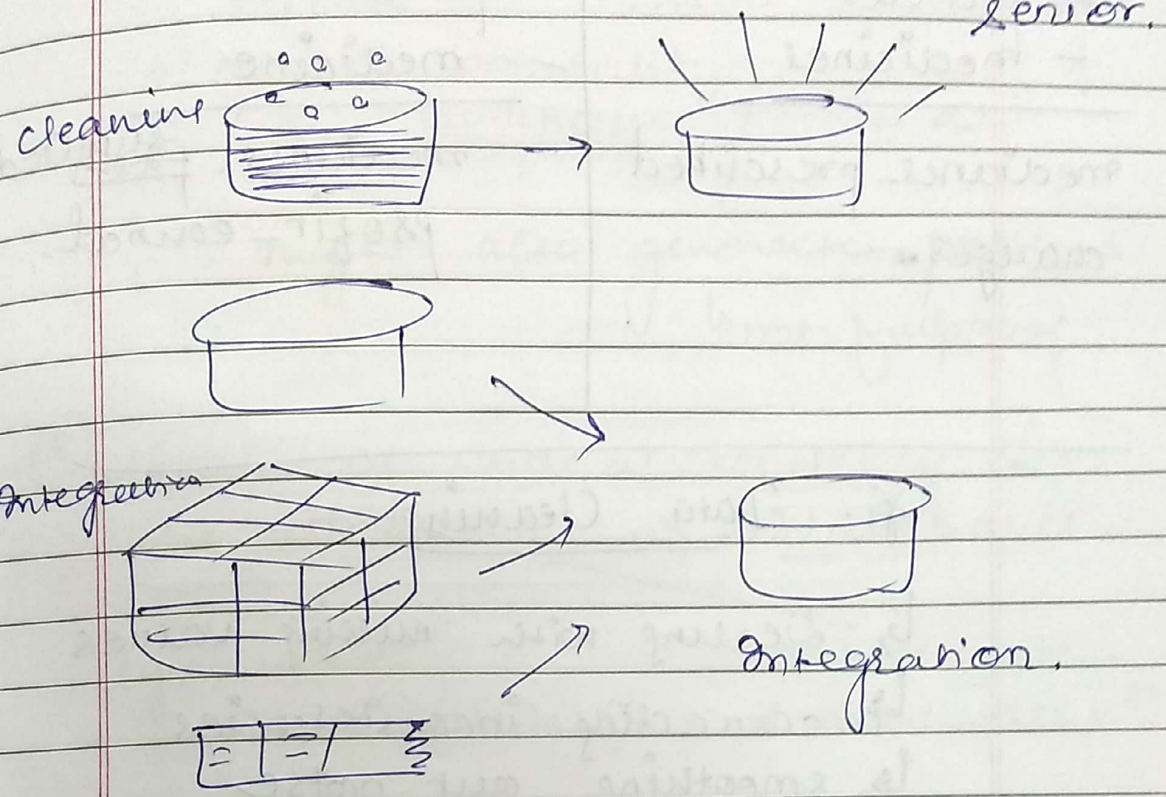
- Dimensionality Reduction -
  "compressed" representation
  - Principal component analysis
  - attribute subset selection
- Numerosity Reduction -
  - Data replaced by alternative, smaller rep. parametric model.
    eg. Regression / log-linear models

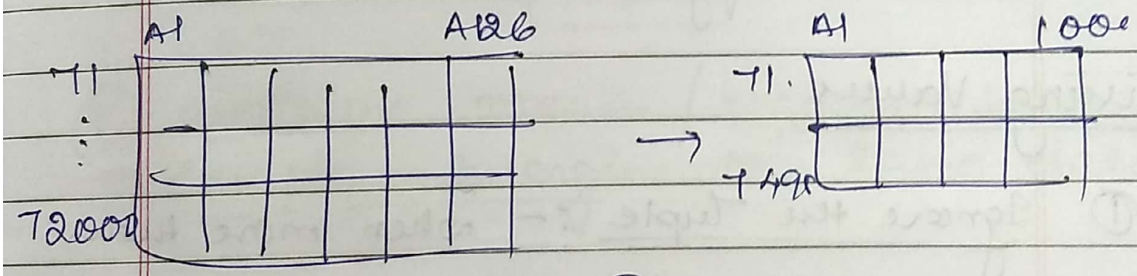
nonparametric
- histograms,
- clustering,
- sampling,
- data aggregation.

Normalization,
Discretization,
concept hierarchy generation -

~~xxxx~~ transforming data to a format that is such that all attributes fall in a range such that they don't outweigh each other

age → raw values → adult, young, senior.





Transformation

-2, 32, 100, 59, 48 → -0.2, 0.3
1.00, 0.5,
-0.3

hospital	chemist shop
- location - doctor - patient - medicines	- location. - chemist - patient - medicine
medicines_prescribed charges.	medicines. ~~supplied~~ profit_earned.

Data Cleaning

↳ Dealing with missing values
↳ correcting inconsistencies
↳ smoothing our noise
↳ identifying outliers.

Missing Values

① Ignore the Tuple :- when more than one attri. missing. Otherwise, we lose our on other attri that were contributes to patterns.

② Fill values manually :- time consuming & not feasible for large dataset.

(4) Using global constant to fill values:-

Replace all missing values by a global constant.

eg "Unknown" / "-∞"

might also generate patterns
& not foolproof.

(5) Measure of central tendency for the attribute → middle value.

symmetric data distribution — mean. eg income

asymmetric data distribution — median.

(6) attribute mean / median for a sample belonging to same class.

eg: customers in same credit risk → same mean

(8) Most probable value
- regression
- inference based
- decision tree based

null values

Noisy Data

Random error / variance in the measured variable.

numeric attribute - noisy data smoothing

① Binning - sorted data values.
- neighbourhood
- boundaries
- central tendency

(bucketing / bins)

local smoothing

equal frequency - bin size.

① Smoothing by mean.

② Smoothing by bin medians

③ Smoothing by bin boundaries

4 8 15 21 21 24 25 28 34

equal freq.

4 8 15
21 21 24
25 28 34.

mean - 9 9 9
22 22 22
29 29 29

11 19.
34 38
3 3
8 8
12

2/ (equal width).

0 - 12	(4) (8)
13 - 26	15 21 24
27 - 34	28 34

median - 8 8 8
21 21 21
25 28 34

boundaries - 4 8
15 24 24 24
28 34

(Regression:)

Linear Regression involves finding the "best line" that fits two attributes so that one attribute can be used to predict the other.

Logistic / multiple Reg → to find more attributes.

$y = B0 + B1x$ ← slope.

intercept

line

$B0 = \text{mean}(y) - B1 \cdot \text{mean}(x)$

$b0 = mean(x) \cdot B1 + mean(y)$

B1 $B1 = \frac{\sum (x - mean(x)) \cdot (y - mean(y))}{\sum}$

x · mean(x) x

y · (mean(y)

2

8 8

10

10

Outlier

Clustering

Cleaning Process

① Identify Discrepecency detection

- poor design forms
- data decay
- ~~cuberian~~ error.
- human entry
- inconsistency
- instrument -

How to identy ?

use data about data - metadata

what data type ?

what are acceptable values ?

mean ?

median ?

mode ?

length

classmate
Date
Page

Field Overloading - 31 out of 32 bits used for storage

1 bit secured.

3 rules for analysis

1. unique
2. null
3. consecutive.

Data Scrubbing / Data Auditing tools

↓ domain knowledge. (parsing) (fuzzy)

↓ Rules of relationship
correlation
outliers

Data Transform

Data Migration Tools

Inconsistency Removal

ETL

eid cusid

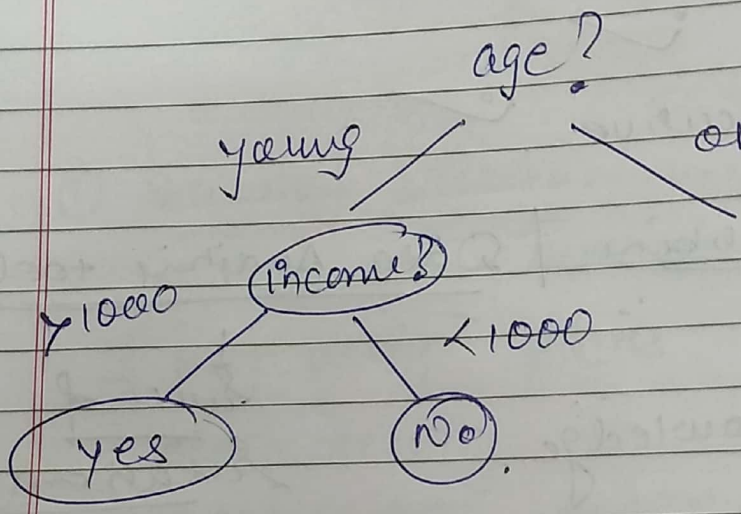
Potter's Wheel

Date
Page

label

(1)

A	1500	young	✓
B	1600	old	X
C	~~2~~2000	young	✓



income(X, "1000") ∧ age(X, "young")

(2) → ~~credit risk(X,~~ loan(X, "granted")

x - mean x
y - mean y

$y = (B0) + (B) \cdot x$

$B1 \Rightarrow \frac{\sum (\text{mean}x \cdot x \times \text{mean}y \cdot y)}{\sum (x - \text{mean}x)^2}$

Scanned by CamScanner

new y =

$B0 = y - (B1 \cdot \text{mean } x)$

$$B1 = \frac{\sum (x - \text{mean } x) \cdot (y - \text{mean } y)}{\sum (x - \text{mean } x)^2}$$

$B0 = \text{mean } y - B1 \cdot \text{mean } x$

$y = B0 + B1 \cdot x$

Correlation

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

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

-1 to +1

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

$$e_{mf} = \frac{(300) \times 450}{1500} = 90$$

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

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

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

Correlations

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

$$\sum_{j=1}^{n} \frac{(a_i - \bar{A})(b - \bar{B})}{n \sigma_A \sigma_B}$$

Test

1. Which one of them provides better/ efficient query processing?
   (a) star (b) schema
2. Name 5 OLAP operations
3. ~~Write a~~ Which ~~form~~ of data does DB & DW store?
4. Name 5 DM functionalities.
5. Diff bet supervised & unsupervised learn?
6. What is freq. itemset mining?
7. Diff bet single dim. associan rule & MD asso. rule with eg.
8. Which is most summarized? apex cube / base cube

$(840)^2 + (200 - 360)^2$

840.

9. 4 properties of DW

(10) Define DW, DM & VDW.

integrar.

* ~~Tuple Duplica~~ DV conflicts

1. weigar
2. hotels
3. gradings.

Tuple duplicar.

* ~~Data value conflict~~

denormalized.

* Redundancy & correlation

covariance

derived attribute

Data integration

Nominal Data

① X^2 correlation test for Nominal Data :

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

CW, Answer.

Are gender and preferred_reading correlated?

$$X^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij} → observed frequency of count

e_{ij} → expected frequency.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{total\ (n)}$$

A has c values $a_1 \dots a_c$ (cols)
B has r values $b_1 \dots b_r$ (rows)

contingency table.

count $(A = a_i)$ no of tuples having value a_i for A

$(r-1) \times (c-1)$ degree of freedom

1500 people

- gender
- preferred - reading fiction/non fiction

observed count summaries in table

expected :-

$$e_{11} = \frac{\text{count}(M) \times \text{count}(F)}{n}$$

$$= \frac{300 \times 450}{1500}$$

$$= 90$$

(data distribution)

$$\chi^2 = \frac{(450-90)^2}{90} + \frac{(60-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 507.93$$

Degree of freedom = $(C-1) \times (R-1)$

$= 1 \times 1$

$= 1 \rightarrow 0.001$ significance level

10.828

upper percentage point of χ^2.

Correlation coefficient for Numeric

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n \sigma_A \sigma_B}$$

$$= \frac{\sum_{i=1}^{n} (a_i b_i) - n(\bar{A}\bar{B})}{n \sigma_A \sigma_B}$$

$$-1 < r_{A,B} < 1$$

$r_{AB} > 0 \Rightarrow$ +vely correlated.
A ↑ with B
& can prove that there is Redundancy

$r_{AB} = 0 \Rightarrow$ A & B independent

$r_{AB} < 0 \Rightarrow$ A & B are negatively correlated. They discourage each other.

Covariance

mean - expected.

$$E(A) = \frac{\sum_{i=1}^{n} a_i}{n}$$

$$E(B) = \frac{\sum_{i=1}^{n} b_i}{n}$$

Date
Page

covariance $= E(A - \bar{A})(B - \bar{B})$

$$= \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$

$$\rho_{ab} = \frac{COV\,ab}{\sigma_A \sigma_B}$$

$$\boxed{COV\,ab = E(A \cdot B) - \bar{A}\bar{B}}$$

If one of A is larger than $\bar{A}$ then one B ↑ $\bar{B}$. positive

One larger & one below → negative

Independent, $E(A \cdot B) = E(A) \cdot E(B)$

time	A allelec	B nightech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(A) = \frac{6+5+4+3+2}{5}$$

$$= 20/5 = 4$$

$$E(B) = \frac{20+10+14+5+5}{5} = 10.8$$

$$COV = \frac{120 + 80 + 56 + 15 + 10 - 4 \times 10.8}{5}$$

$$= 47.$$

positive covariance

1. Tuple Dupication

- Dupication arises due to denormalized tables (done to improve performance)
- Inaccurate data / appdates / Deletion of records

2. Data Value Conflicts & Rep.

- Metric | British Imperial
- A to F+
- Hotel prices

Entity Identification Problem

Schema integration
Object Matching

format
ETP
attribute

customer_id & cus_id.

- structure ✓
- metadata ✓

E(A.B) - Add

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

Covariance

t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$E(A) = \frac{6+5+4+3+2}{5} = 4$

$E(B) = \frac{20+10+14+5+5}{5} = 10.80.$

covariance $= \frac{6\times20 + 5\times10 + 4\times14 + 3\times5 + 2\times5}{5} -$

$- (4)(10.8)$

$= 50.2 - 43.2$

$= 7.$

+ve covariance → related.

-ve covariance → negatively related

0 covariance → not related.

Data Reduction

- to obtain reduced representation of data set that is much smaller in volume but more efficient to have mining performed over it.

(maintain integrity of original dataset)

(1) Sampling

(1) Simple Random Sample without Replacement (SRSWOR) of size S

S < (N) total (D)

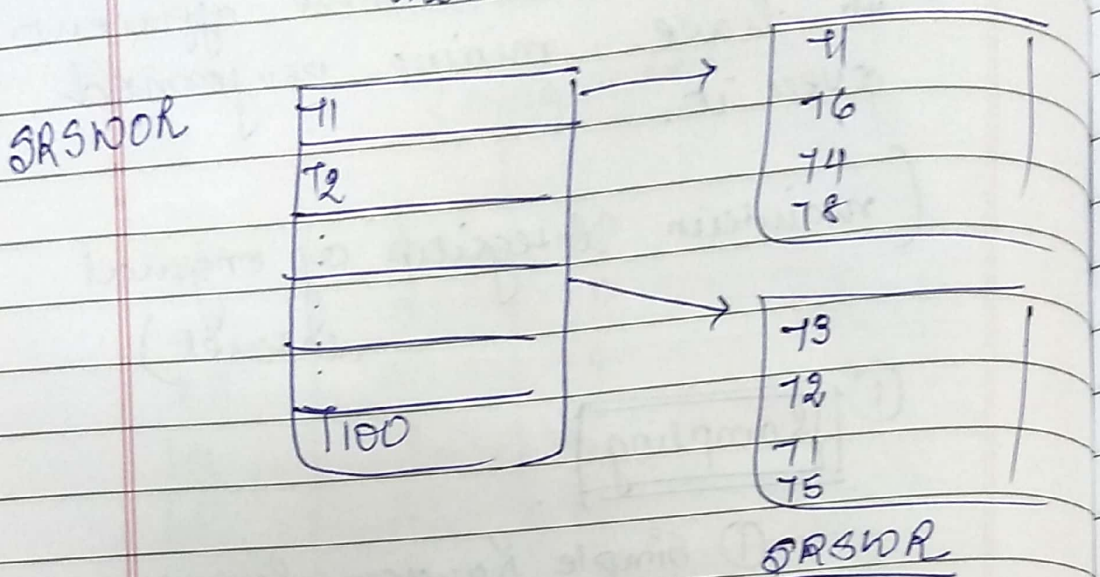
each tuple = 1/N

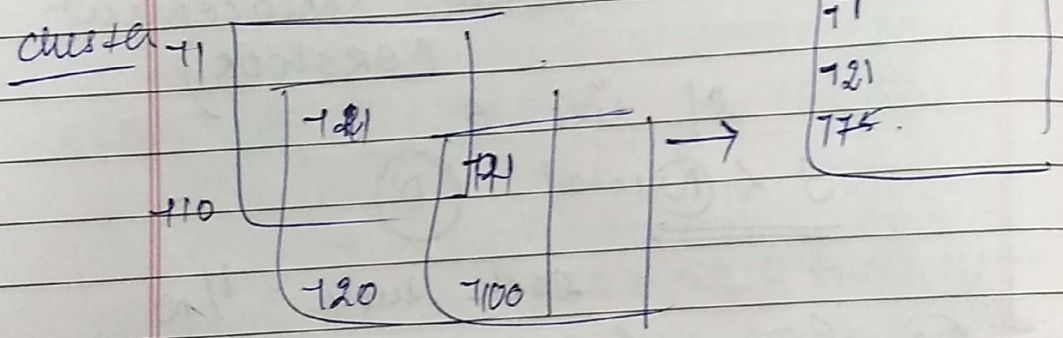
(2) Simple Random Sample with Replacement of size S (SRSWR)

each time a tuple is drawn from D, recorded & replaced. to be drawn again.

(3) Cluster Sample → If tuples in D are grouped into M

disjoint clusters then an SRS of 5 clusters can be obtained when SRSWOR





Stratified sampling | according to age | as per attribute

T1 young
T20 young
T60 young

T12 middle-aged
T16 middle-aged
T88 middle-aged

T11 Senior
T78 Senior

T38 youth
T91 youth
T101 middle aged
T21 Senior

Advantages:-

efficiency

Disadvantages:-

Sampling errors

bias / discrimination

↳ Iterative Procedure

Sample size ↑ or ↓

Size of Sample depends upon the Population.

(2)

Histograms

- uses Binning to approximate a data distribution.

Histogram for attribute A, partitions the data distribution of A into disjoint subsets → buckets / bins.

one element in bucket → singleton bucket.

single attribute

(52)/(17)

ed

Q 1 1 5 5 5 5 5 8 8 10 10 10
10 10 10 10 15 15 17 17 17 20 21
21 21 22 25 26 31 32 32.

Size = 5 frequency 5 element list

width. range

(B)

Attribute Subset Selection

- there might be some attributes in the dataset that might be irrelevant to the mining task

eg - to classify customer on the basis of their liking / not liking to purchase a popular new CD or A company when notified about sale

age / music-taste are relevant

address / phone no are all irrelevant

When data behaviour is not well known, it is difficult to pick / time consuming to pick the relevant attributes.

Attribute subset selection reduces the dataset size by removing the irrelevant / redundant attributes.

goal: to find minimum set of attributes such that the distribution of data classes is as close as possible to the original distribution obtained with all attributes

for n attributes, 2^n subsets.

If n increases & no. of classes increase, search for optimal subset becomes exhaustive.

∴ Greedy methods are used

↓

while looking through attributes they choose the one that appears to be the best choice at the moment

↳ make locally optimal decision to reach globally optimal sol'.

Test - statistical test used for finding if attr' dependent on one another.

① Stepwise forward selection:-

Initial set
{A1, A2 .... A6}
empty

proc starts with an empty set

Initial reduced set
{A1}
{A1, A3}
{A1, A3, A6}

Best original attribute is defined and added

Best from remaining is kept on being added

② Stepwise backward elimination

A1 ... A6
A1 A4 A5 A6
A1 A4 A6

Starts with fullset & removes attributes one by one

③ Combination of each step select best & worst both

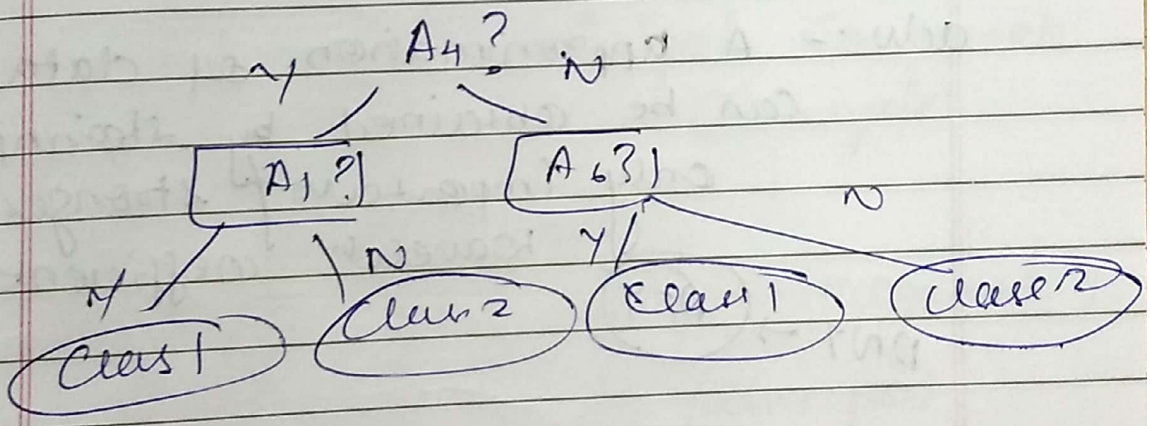
(4) Decision Tree Induction

ID3, CART, C4.5

a flowchart where internal nodes are test on attribute, each brand is outcome of test, external leaves are classes

at each stage, best attribute is chosen for partitioning.

Stopping criteria - threshold



A1 A4 A6

1. Information gain
2. gini index
3. gain ratio

} max

(2) Wavelet Transform

DWT Discrete Wavelet Transform

linear signal processing technique that when applied to X (data vector) transforms it to a numerically different (X'), of wavelet coefficients

2 vectors are same length

Here, n-dimensional data

$X = (x_1, \ldots, x_n)$

adv:- A approximation of data can be obtained by storing only important/strongest wavelet coefficients.

DWT → (DFT)

→ Length L of i/p data vector must be 2^n.

pad necessary as (L Y, n)

→ each transform — data smoothing / wt/avg

— weighted diff to bring out detailed features

→ These 2 functions are applied to pair of data points in X that is to all pair x_{2i}, x_{2i+1}

Results in 2 datasets of $L/2, L/2$ (lower freq. version of i/p)

→ Two func^n recursively applied to data set in previous loop.

→ Selected values in previous iter^n are designated as wavelet transforms.

(2) PCA Principal Component Analysis

① normalization

② eigenvector (orthogonal)

③ the eigen value associated with these vectors ↳ principal components.

magnitude higher significance kept

wq