Data ~~Mining~~ Warehousing & Online Analytical processing.

Data warehousing :- decision support database — maintained separately — info processing providing a solid platform.

→ ① **subject-oriented** :-
* organized around major sub.
like → customer, product, sales.
* focuses on decision makers not on daily op or transaction
* subject issue → exclude simple and concise view that excluding not useful data.

② **Integrated** :-
* combines multiple heterogenous data sources.
[relational database, flat files, online transaction records]
* Data cleans, integrates.
* converted data when moved to warehouse.

③ **Time variant** :-
* Time horizon is longer
* provides info from a historical perspective.

④ **Non volatile** :-
* physically seperate store → tranform हुआ
* do not need transaction processing, recovery or concurrency control.
* Needs — initial loading
— access of data.

OLTP → (Database, operational) Online transaction processing.
OLAP → (" warehousing) Online Analytical processing

Difference :	OLTP (Database)	VS	OLAP (Warehouse)
① users	clerk, IT professionals (customer oriented)		knowledge workers (market oriented) data analytics
② function	day to day operations.		decision support
③ DB design	application oriented		subject oriented
④ data	current, upto date, detailed, flat relational isolated.		historical, summarized, multidimensional integrated, consolidated
⑤ usage	repetitive		ad-hoc
⑥ access.	r/w. index / hash on primary key		scans. read only

	OLTP	OLAP
⑦ unit of work	short and simple transaction	complex query
⑧ records accessed	tens 10s	millions M
⑨ users number	thousands 1000s	hundreds 100s
⑩ DB size	100 MB - GB	100 GB - TB
⑪ metric	transaction throughput	query throughput response
⑫ design	ERD (Entity Relationship model)	snowflake/star models
⑬ Access patterns	Concurrency control needed	No need, Read Only ডাটা

* Why seperate Data Warehouse?

① For long term use

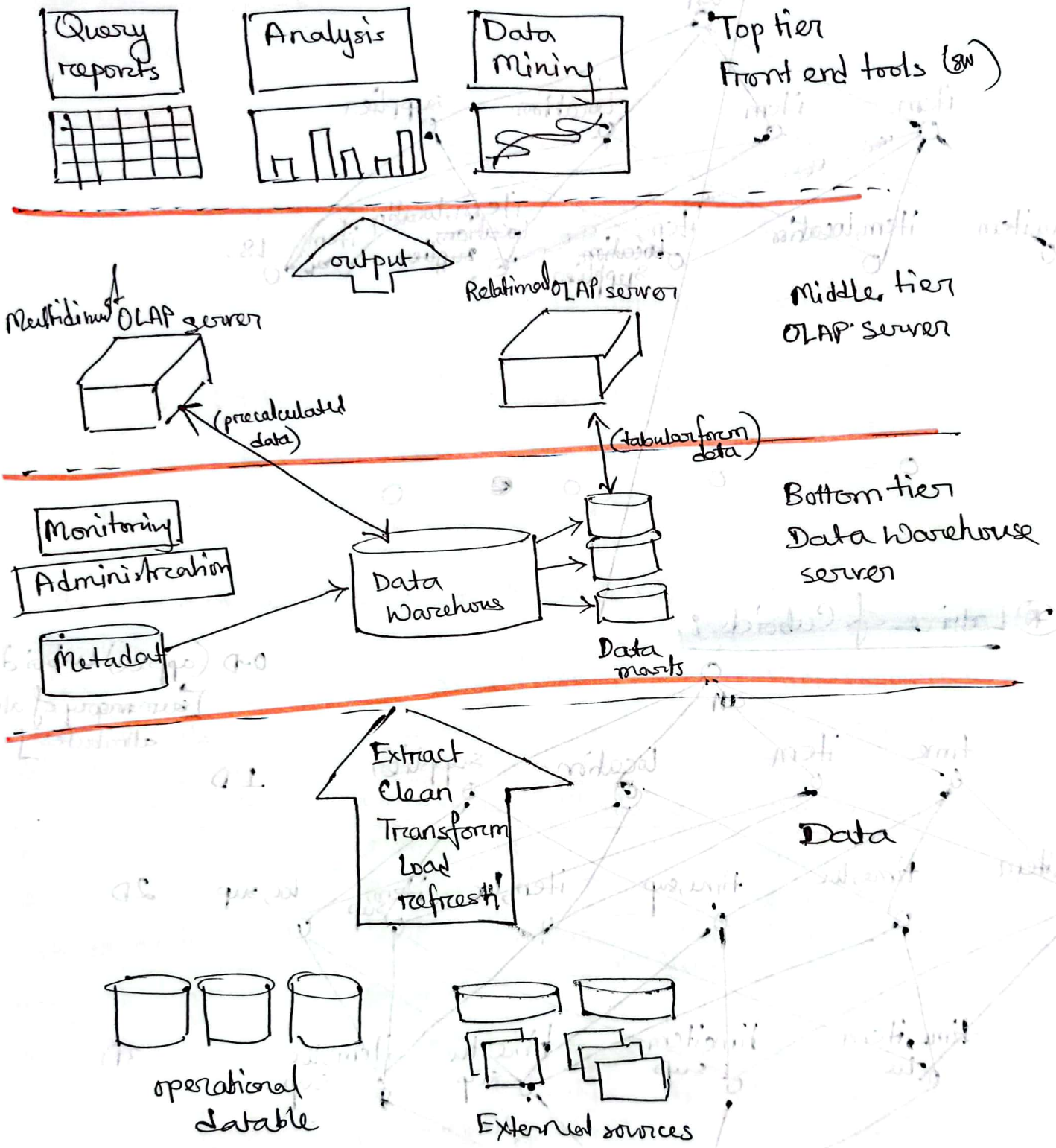
② Get historical data - no missing data.

③ DB - DB warehouse seperate না করলে data amount বেড়ে যাবে retrive করা computationally complex.

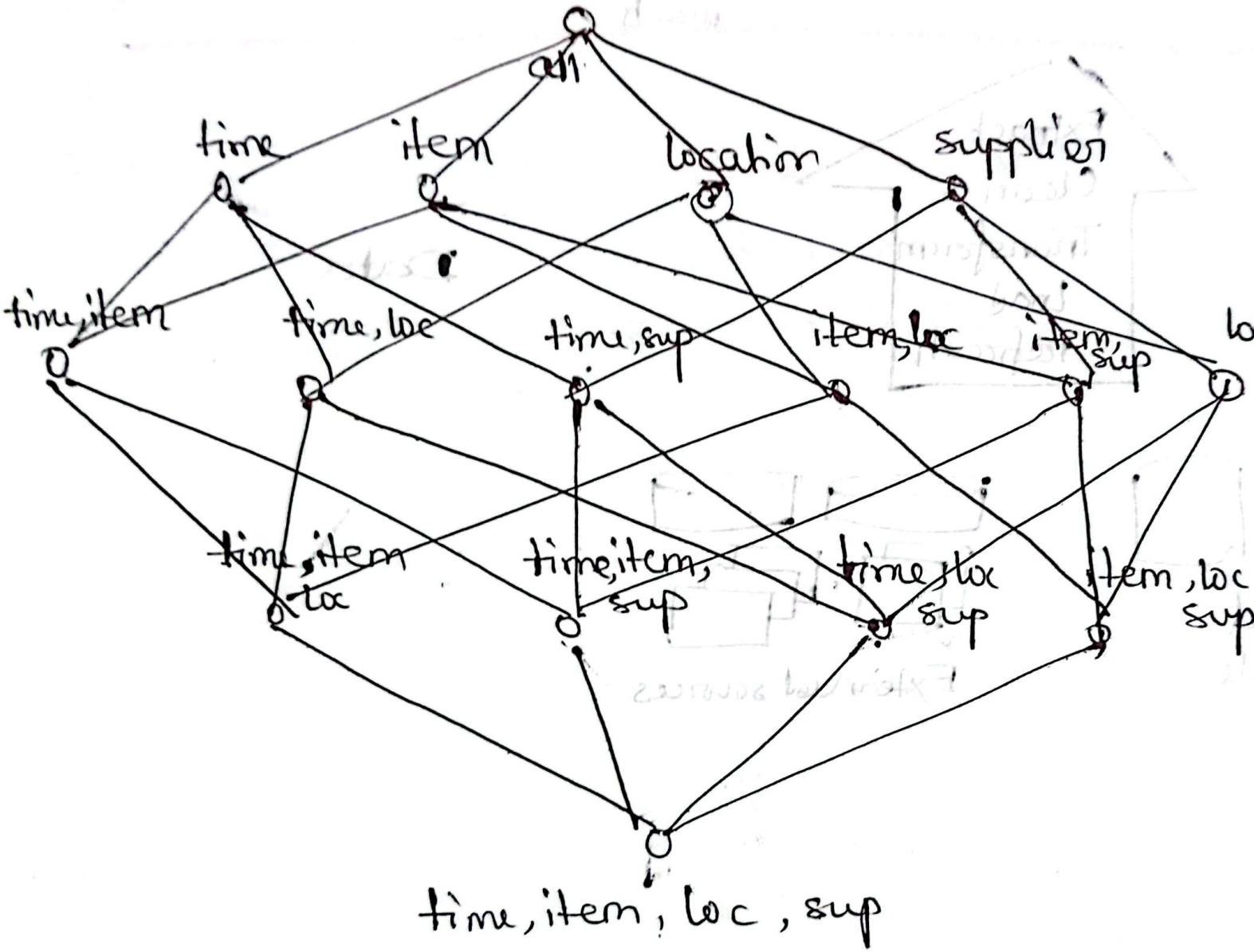
④ DB - তে (OLTP) detailed - এ data থাকিব - not necessary.

⑤ Data access - এর জন্য precalculation লাগলে বারবার calculate করে access করা লাগবে but warehouse - এ precalculated data থাকায় possible so বারবার calculation করা লাগে না

Multi Tiered Architecture: (3tier)



(*) Lattice of Cuboids :



0-D (apex) cuboid
[summary of all attributes]

1 D

2D

3D

4D (base) cuboid
[Highest Dimension]

Computing number of cuboids.

each dimension → n.

Level → L

So, total number of cuboids = $\prod_{i=1}^{n} (L_i + 1)$ → 1 যোগ করি all-এর জন্য.

Suppose a 3D cuboid. → location, item, time.

↓

location has attributes (district, division)

So, $L_L = (2+1) = 3$ [location] (dist + div + all)

$L_I = (1+1) = 2$ [item] (item + all)

$L_T = (1+1) = 2$ [time] (time + all)

So total number of cuboids = $3 \times 2 \times 2 = 12$.

Group By:-

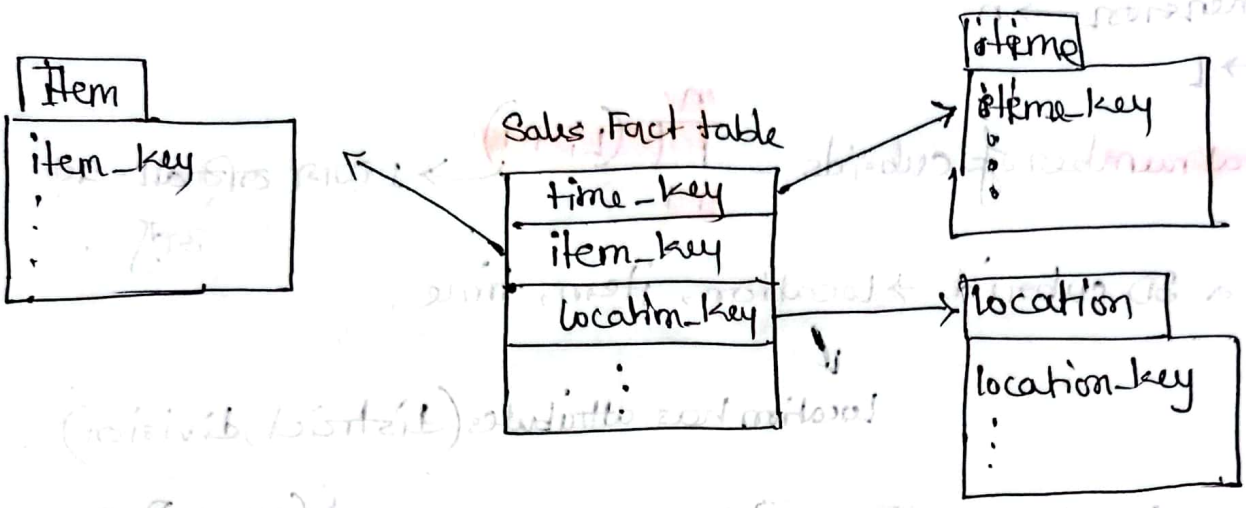
when attributes do not have sub attribut

so, 3D cuboid group by = 2^n or $2^3 = 8$.

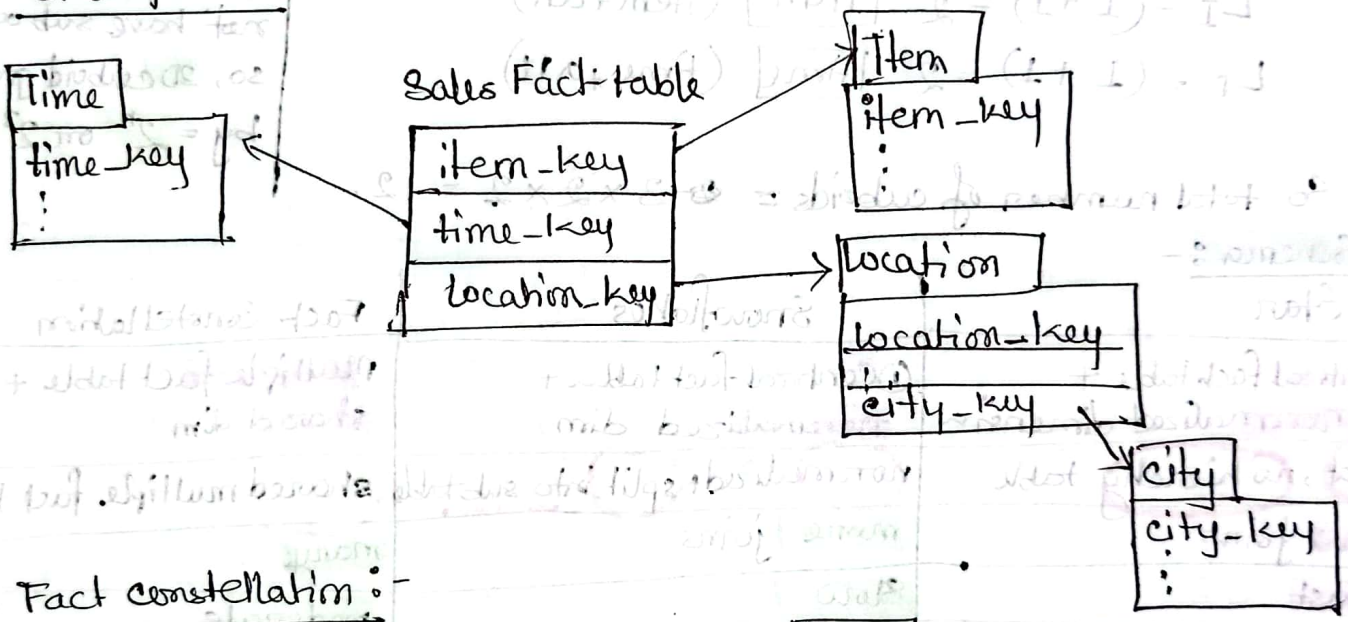
Schema:-

Star	Snowflakes	Fact Constellation
① Central fact table + denormalized dimension	① Central fact table + normalized dim	multiple fact table + shared dim
② Flat, no hierchy table	normalized + split into sub table	shared multiple fact table
③ Few joins	more joins	many
④ Fast	slow	moderate
⑤ redundant data	high (less redundancy) storage efficiency	med to high storage efficiency
⑥ easy maintanance	complex	complex
⑦ Readable	moderately readable	hard for beginners.
⑧ Dashboards, BI tools	⑧ storage optimization, integrity focus.	⑧ Complex systems with multiple fact types
usage: simple + fast.	reduce redundancy + normalize.	multiple types of facts.

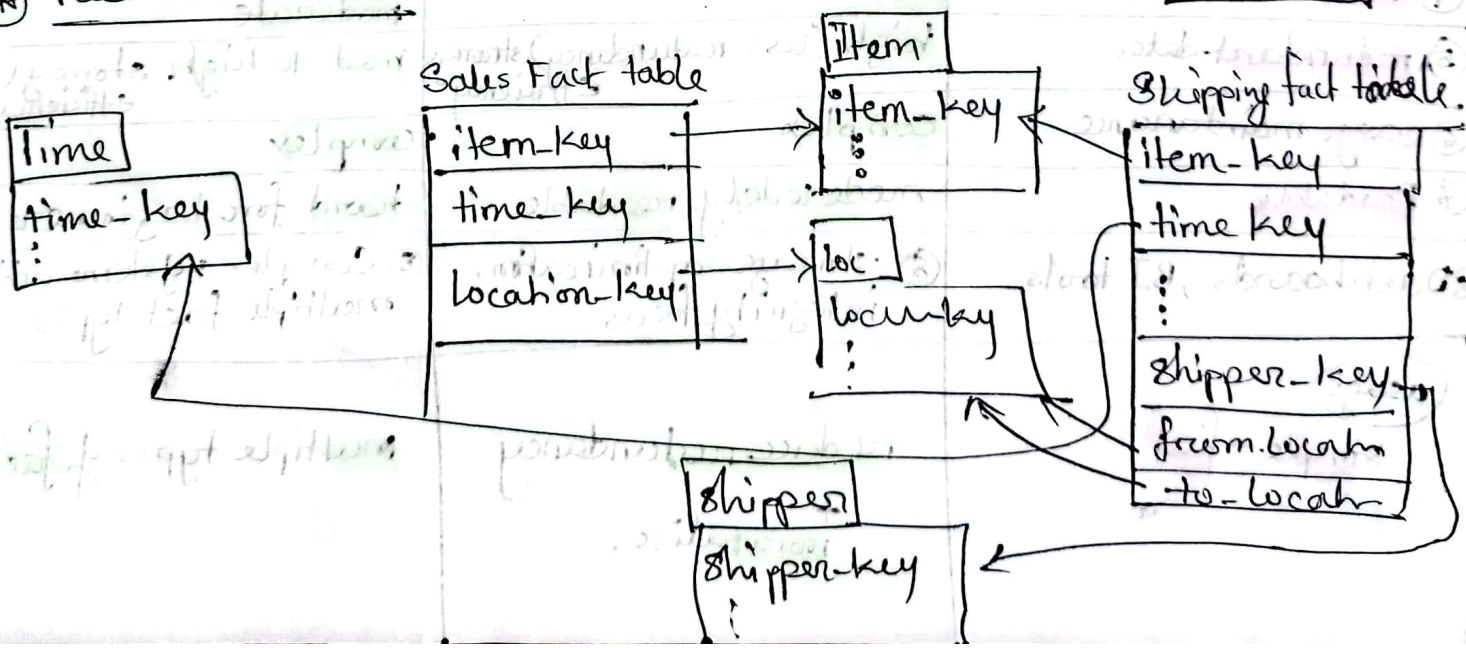
(*) Star Schema



(*) Snowflake



(*) Fact constellation



Data Cube: (Measures)

① **Distributive**: result derived by applying the fn to n aggregate values is same as that derived by applying the fn on all the data without partitioning.

Eg: count(), sum(), min(), max().

② **Algebraic**: if it ~~give~~ can be ~~appl~~ computed by an algebraic fn with M arguments (M is a bounded int), each of which is obtained by applying a distributive aggregate fn

Eg: avg(), min_N(), standard_deviation()

③ **Holistic**:

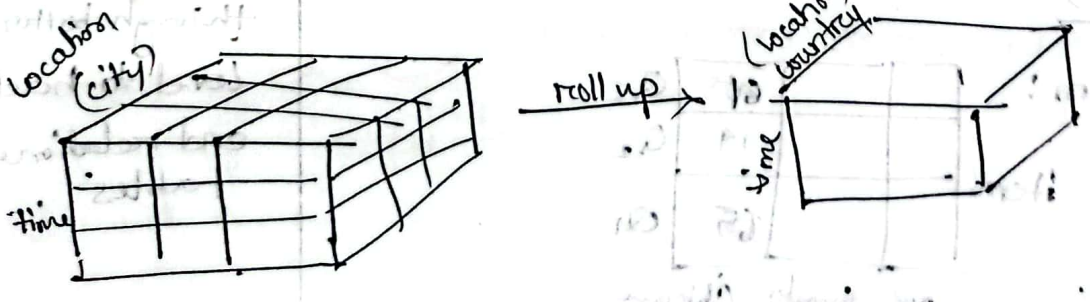
i) pending memory size: no constant bound on the storage size.

ii) Non distributive: partition wise median বের করে, join করলে correct result পাব না – ~~dat~~ full dataset -ই median করা possible।

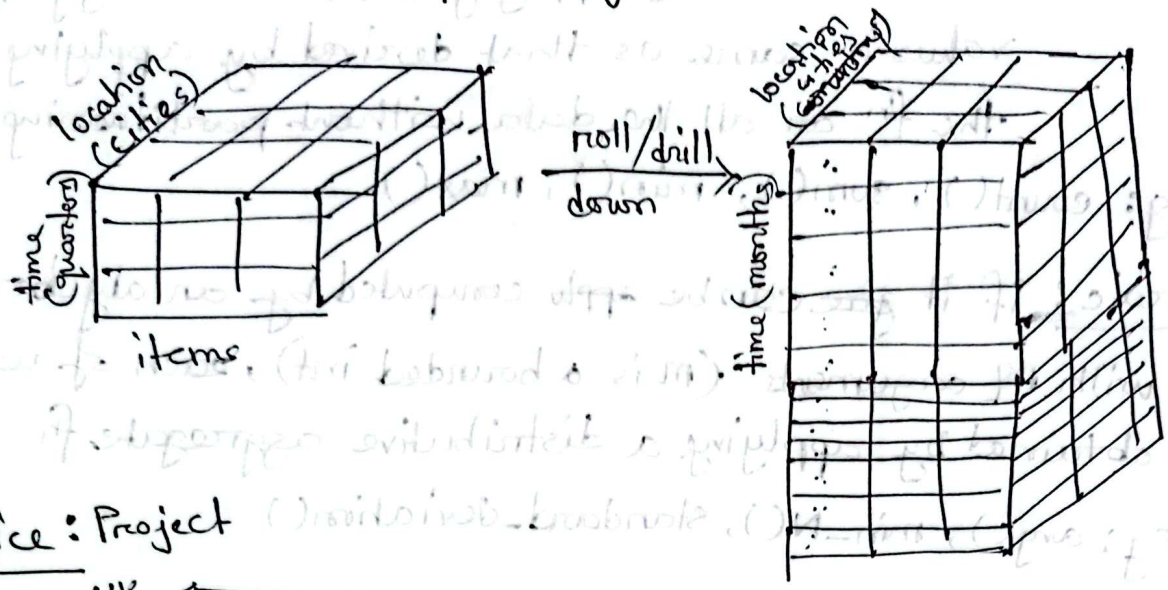
Eg: median(), mode(), rank().

OLAP operations:-

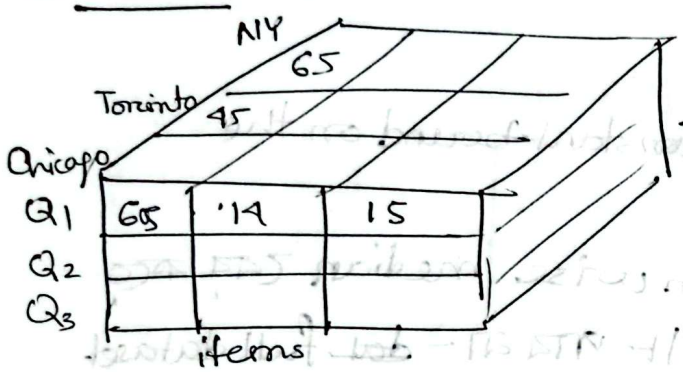
① Roll-up / drill-up: climb to higher hierarchy or dim reduction.



② Roll down / Drill down : introduce new dim, from high to lower level summary.



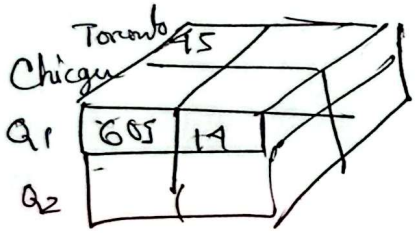
③ Slice : Project



Slice (Q1) :

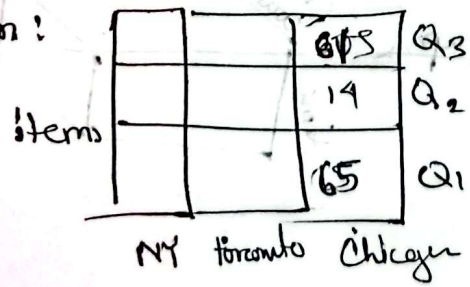
NY			
Toronto			
Chicago	605	14	15

④ Dice : Select



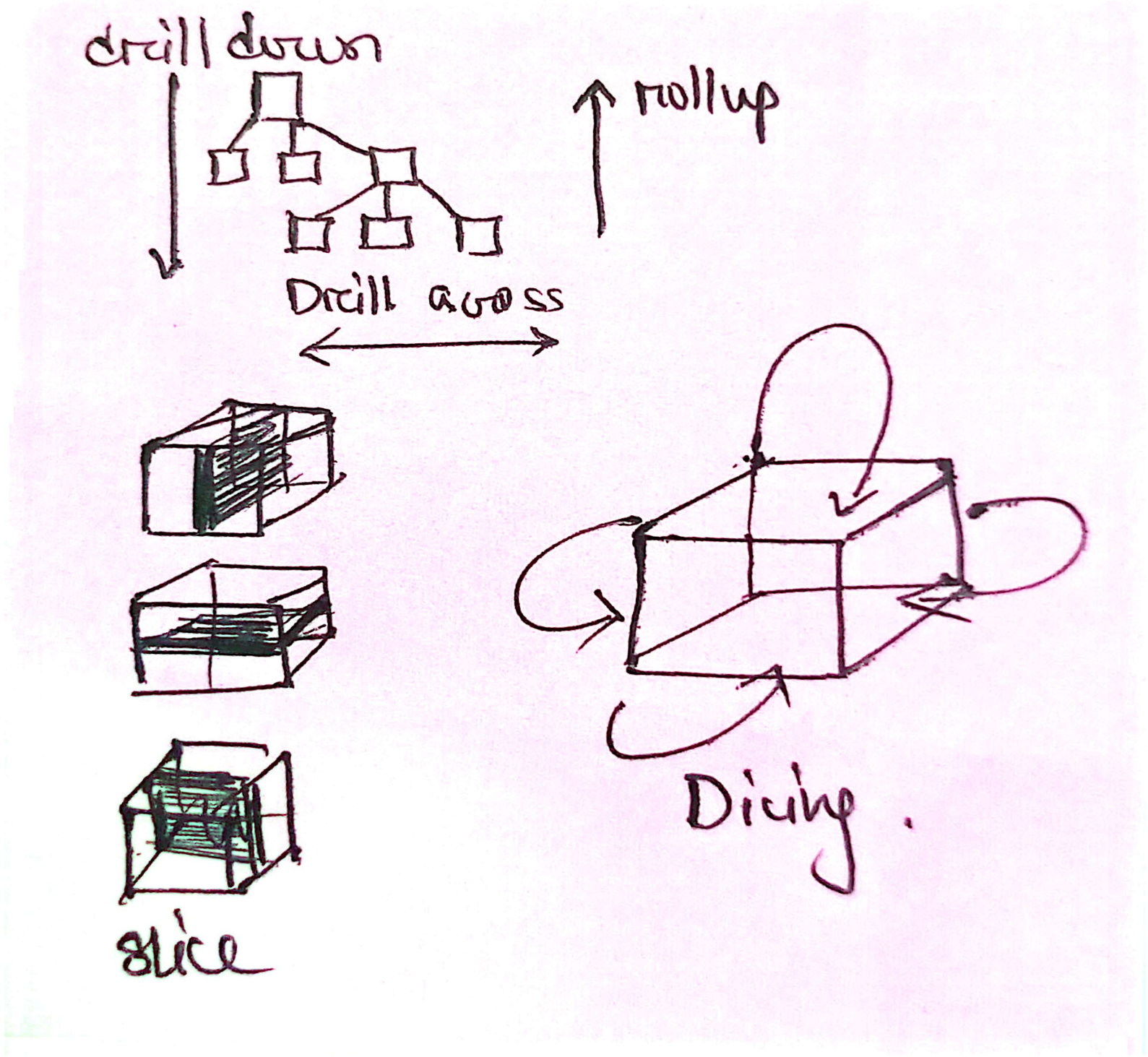
⑤ Pivot (rotate) : reorient, 3D to 2D planes

pivot to item :



⑥ Drill across : more than 1 fact table

⑦ Drill through : through bottom level to back end relational tables


drill down
roll up
Drill across
Slice
Dicing.

Data Warehouse Design Process

2 Approaches: ① Top-Down :- starts with overall design + planning (mature)

② Bottom-Up :- " " experiments + prototypes (rapid)

Methods :- (swe -এর pov).

① **Waterfall** :- এক process -কে stage -এ ভাগ করে - each stage 100% complete না করে পরের stage -এ যাব না।

② **Spiral** :- short + quick turn around, interval time between stages is short so 100% stage complete না করেই jump করা যাবে

- **Steps :-** ① Choose a business process to model (কোন type business করব — orders, invoices etc).

- ② Choose grain of the bp (data warehouse / fact table-এ কি for রাখব)

③ Choose dimension (attribute) for each fact table

④ " measure that will populate each fact table.

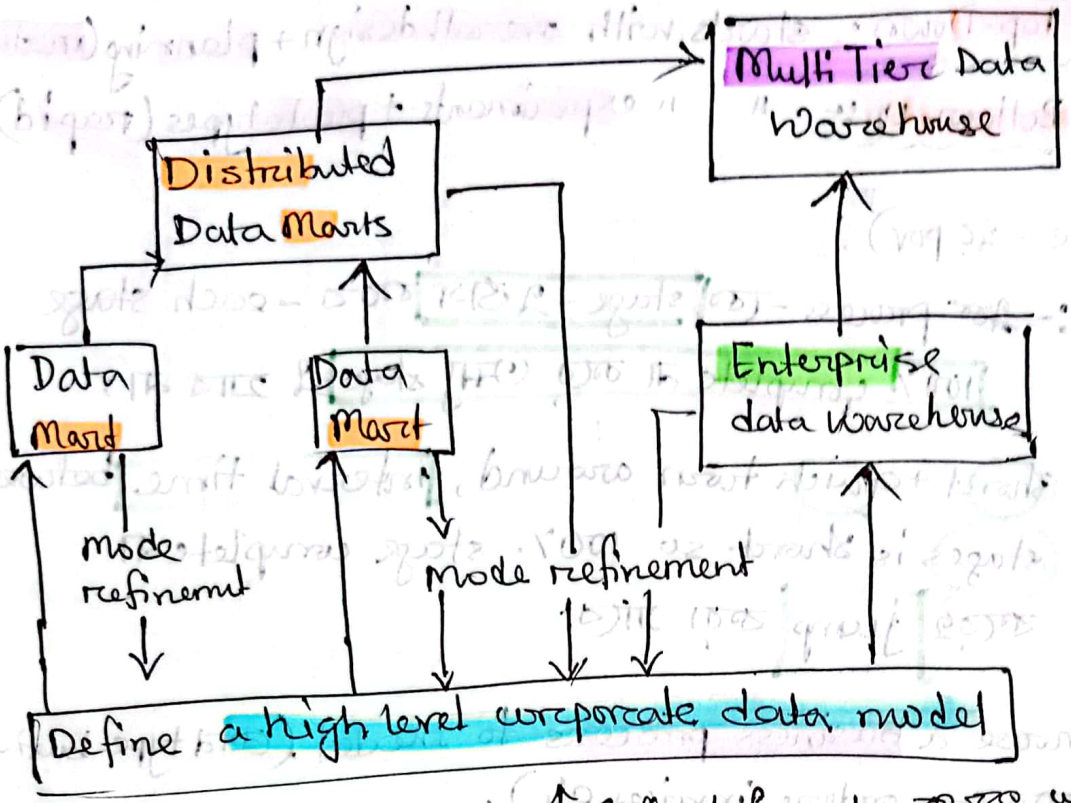
Applications :-

① Info processing - graph, table, chart.

② Analytical processing - Cube, ROLAP, MOLAP etc.

③ Data mining - new কোনো info বের করব, new pattern.

Datawarehouse Development: A Recommended Approach :-



↑ (Different format - থেকে unified model -এ convert করব - does precalculated aggregated value - for data cube)

Partial Materialization কেন use করি?

① Full Materialization :- সব cuboid আগে বানায় রাখব
② Partial " :- কিছু " বানায় রাখলে not all
③ No " :- only main cuboid টা থাকবে

** সব আগে বানায় রাখলে — operation-এ data পাওয়া easy but storage waste — this computing time faster.

Compute Cube Operator: The operators used to ~~op~~ in the mathematical operation of data cuboids.

→ mean(), median() etc.

OLAP Indexing:

① **Bitmap indexing**: It is difficult to search a string amongst so much data. So we can represent a value using a bit creating seperate columns for each value and assigning existence of the value row wise.

Example:

Base table

Cust	Region	Type
1	Asia	Retail
2	Europe	Dealer
3	Asia	Dealer
4	Russia	Retail

bitmap index according to Region:

RecID	Asia	Europe	Russia
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1

1 = present
0 = absent

according to type

	Retail	Dealer
1	1	0
2	0	1
3	0	1
4	1	0

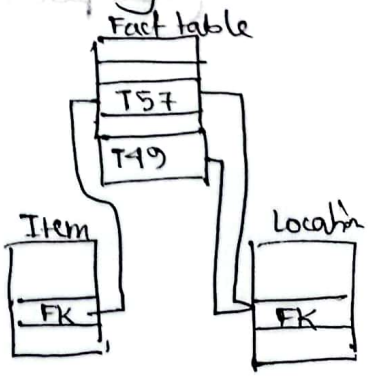
~~bitmap~~

* fast search
* complex architecture.

② **Join indexing**: can be done in fact tables only.

* Join the primary key of fact table with foreign keys of other tables.
* Fact table do not have details, so searching is easy.

Steps:
** input a key into the system
** system searches the key in fact table
** that key is joint with the foreign keys of other table, thus system goes to the location.



✱ Conversion from Data Cube to Table :

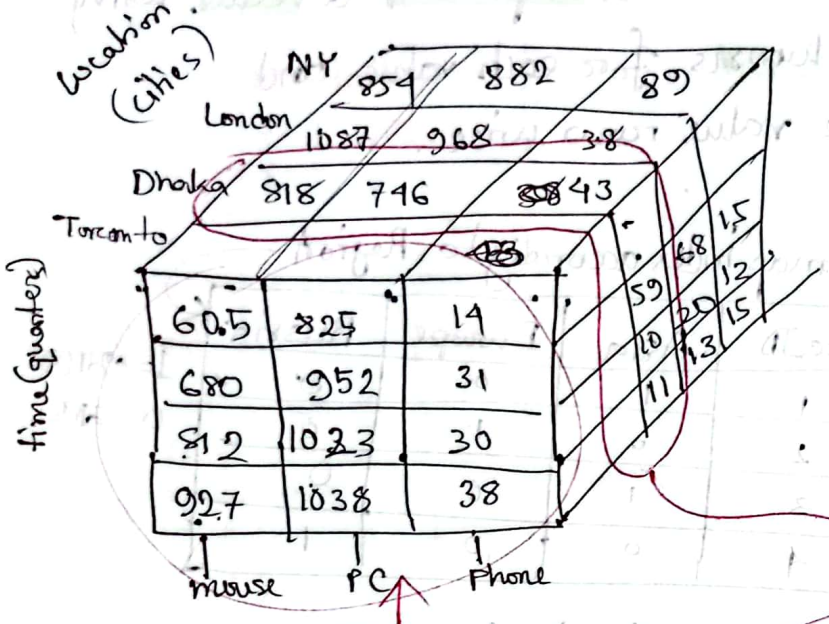


Table:

Time	location NY			London			Dhaka			Toronto		
	Item			item			item			item		
	mouse	PC	Phone	mouse	PC	phn	mou	PC	phn	mou	PC	phn
Q1	854	882	89	1087	968	38	818	746	43	605	825	14
Q2	634	712	15	1033	321	68	352	402	59	680	952	31
Q3	512	892	12	113	481	20	415	411	10	812	1023	30
Q4	321	1012	15	135	702	13	113	1413	11	927	1038	38

✱✱✱ (Red pen-এ যে cube not-visible.)

Association Rule for Data Mining:

Support → x %, there are x% case in the training data of and condition where (A∩B) (A ∪ B)

Confidence → y%, there is a probability of y%. that a new case meeting the condition A will cause the condition B which means (A → B).

⊛ association rule:

$$\text{Support } (A \to B) = P(A \cup B) = \frac{\text{count}(A \cup B)}{\text{Total no. of transactions}}$$

$$\text{confidence } (A \to B) = P(B|A) = \frac{\text{count}(A \cup B)}{\text{count}(A)}$$

suppose, ~~support = 2%~~
~~confidence = 60%~~ for {milk, bread} → table-1 দুইবার আছে

Here, support = $\frac{2}{10}$

ID	Item (frequent)
1	{milk, bread}
2	{~~bread~~, milk, butter}
3	: , , ,
4	{bread, milk}
:	:
10	{milk, ~~bread~~ banana}

$$\text{confidence} = \frac{2}{4} \to (\text{milk, 4বার আছে})$$

Binning: Partitioning large number of data points into seperate bins.

3 techniques:

① Equal Frequency : [10, 11, 14] [16, 17, 18] (3, 3)

" width : [10-14] , [15-19]

② By median : median বসায় দিব সবগুলোকে [10, 12, 14] → [12, 12, 12]

③ By boundary : দুইটা boundary-র কাছে যেটা বেশি কাছাকাছি [10, 12, 15] → [10, 10, 15] (nearer to 10.)

Correlation for Nominal Data : (Chi square).

step1 : Get expected frequency

step2 : $\chi^2 = \sum \frac{(\text{Actual} - \text{Expected})^2}{\text{Expected}}$.

Degree of freedom = (row − 1) × (column − 1).

Significance level → 0.001 or 0.01 (or given)

if χ^2 (chi-square) > table of chi square (given) [rejected]

Example:

	male	female	total
fiction	250	200	450
nonfiction	50	1000	1050
total	300	1200	1500

Expected value :

$e_{male,fiction} = \frac{\text{Total}_{male} \times \text{T}_{fiction}}{T_{m;f}} = \frac{300 \times 450}{1500} = 90$

$e_{male,nonfic} = \frac{300 \times 1050}{1500} = 210.$

$e_{female,fic} = 360$; $e_{female,nonfic} = 840.$

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

$$= 507.93.$$

will compare χ^2 with given chi square (যদি হলে rejected) no correlation.

Correlation for Numeric Data:

① covariance কাকে বলে → $Cov(A,B) = \frac{\sum(A_i \times B_i)}{\text{Total}} - (\bar{A} \times \bar{B})$

② correlation " " = $\frac{Cov(A,B)}{\sigma_A \cdot \sigma_B}$

Example:-

ID	A($)	B($)
1	$ 6	20
2	5	10
3	4	14
4	3	5
5	2	5

covariance

$$E_A = \frac{6+5+4+3+2}{5} = \$4$$

$$E_B = \frac{20+10+14+5+5}{5} = \$10.80.$$

$$Cov(A,B) = \frac{(6\times 20) + (5\times 10) + (4\times 14) + (3\times 5) + (2\times 5)}{5} - (4 \times 10.80)$$

$$= 7.$$

$\sigma_A = \sqrt{\frac{\sum(A-\bar{A})^2}{n}} = 2$. and $\sigma_B = 32.56$.

$$Corr(A,B) = \frac{Cov(A,B)}{\sigma_A \cdot \sigma_B} = \frac{7}{2 \times 32.56} = 0.107.$$

$r_{AB} > 0$; positively correlated (higher the stronger); $r_{AB} < 0$; negatively correlated.

PCA (Principal Component Analysis)

Given, $[x_1 \dots x_i]$ and $[y_1 \dots y_i]$

Step 1: Normalize (Zscore).

① real value-র mean $= \frac{\Sigma x}{n}$

② " " SD $= \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

③ Normalize $= \frac{x-\bar{x}}{SD}$ or $\frac{x-\bar{x}}{\sigma}$

④ Normalized value-র mean, $\mu = \frac{\Sigma(N)}{n}$

Step 2: Covariance Matrix

$$\begin{bmatrix} \sigma^2(x) & cov(x,y) \\ cov(x,y) & \sigma^2(y) \end{bmatrix}$$

$$cov(x,y) = \frac{\Sigma(xy)}{n-1} - (\bar{x}.\bar{y})$$

$$\sigma^2(x) = \frac{x-\bar{x}}{n-1}$$

$$\sigma^2(y) = \frac{y-\bar{y}}{n-1}$$

→ normalized data-র

Step 3: PC1 বের করব

$$C - \lambda \vec{I} = 0.$$

$$\text{or, } \begin{bmatrix} \sigma^2(x) & cov(x,y) \\ cov(x,y) & \sigma^2(y) \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0.$$

$\lambda = ?$ (Highest value of λ is PC1).

Step 4: Projections.

$(C - \lambda \vec{I})\vec{V} = 0$

ultimately.

$$\begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} v_1 & v_1 \\ v_2 & v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$v_1 = ?$

$v_2 = ?$

if $v_1 v_2 = -v_1 v_2$ → $\begin{bmatrix} t \\ -t_2 \end{bmatrix}$ → let $v_1 = t$, $v_2 = -t_2$

if $v_1 v_2 = v_1 v_2$ → $\begin{bmatrix} t \\ t \end{bmatrix}$ → let $v_1 = v_2 = t$

So, $\vec{V} = ?$

$\hat{V} = \frac{\vec{V}}{|V|} = ? \rightarrow \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$

$$P_i = V_1 x_i + V_2 y_i$$

DWT (Wavelet Transformation)

Dataset → subset- এরকম দুইটা করে set → High frequency set (core feature)
Low " "

Given Dataset, D = {x, y, z, m}.

① Length = 2^n (padding করব if need)

② High fq set = difference of adjacent pairs → $\{x-y, z-m, \dots\}$
Low fq set = avg " " " → $\{\frac{x+y}{2}, \frac{z+m}{2}, \dots\}$

③ will consider low fqset again to seperate:

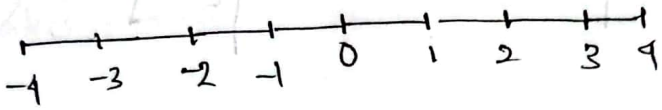
H. fq set = $\{\frac{x+y}{2} - \frac{z+m}{2}, \dots\}$.

L. fqset = $\{\frac{(\frac{x+y}{2}) + (\frac{z+m}{2})}{2}, \dots\}$.

④ Result = { L1, L2, High frequency set ....}.

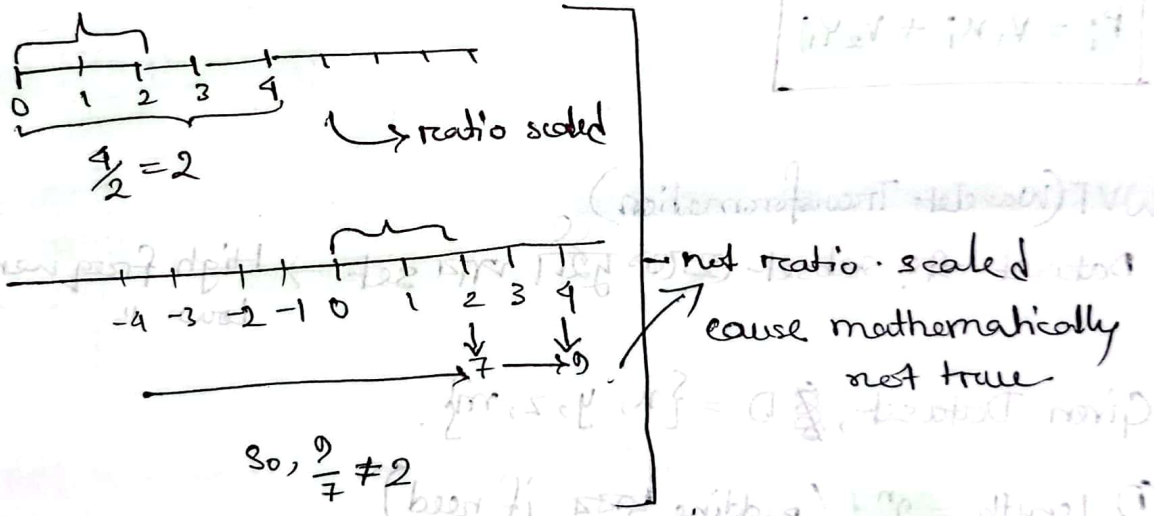
= [~~Low~~ দুইটি low fq set -এর first এর value +
দুইটি high fq set -এর শেষ value]

Interval Scale :- no true zero point in the difference of an attribute



যেমন, 0°C ≠ ~~0~~ no heat
0K = no heat

Ratio scale :- ratio calculation must be meaningful.



Mean = $\frac{\sum x_i}{n}$

weighted mean = $\frac{\sum (w_i x_i)}{\sum w}$

(Sensitive to outliers)

Mode : value frequent.

1 2 3 3 3 4 5
uni ~~root~~ modal

1 2 3 3 4 4 4 5
Bimodal.

Mode- Media :

Index | 0 | 1 | 2 | 3 | 4 |

Dataset length odd = $\frac{N-1}{2}$ → 1 2 (3) 4 5 . $\left(\frac{5-1}{2}=2\right)$

" " even = $\frac{N-2}{2}$, $\frac{N-2}{2}+1$ → 1, 2, (3), (4), 5, 6 $\left(\frac{6-2}{2}=2,\ 3\right)$

$\frac{3+1}{2}$ = 3.5 median.

For table, median = $L_1 + \left(\frac{\frac{N}{2} - (\Sigma fq)_{low}}{fq_{median}}\right) \times W$

$\frac{N}{2}$ = cumulative fq → যে রো! Row is median row.

L_1 = median row-এর min range.

$(\Sigma fq)_{low}$ = median-এর আগের cum. fq.

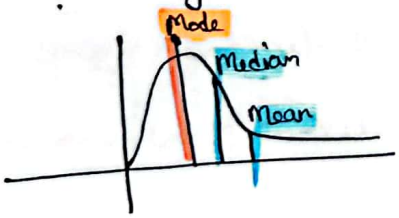
fq_{med} = median-এর fq

width = interval.

Symmetric : Central point with both side equal distribution

Asymmetric : Due to outliers

Positively skewed



Negatively skewed

Moderately skewed : outlier বেশি not high বা low (unimodal).

mean − mode ≈ 3 (mean − median)

Range : Substraction of min-max of an dataset.

Quantiles : points that partitions dataset equally

2Quantile → 1 2 3 | 4 5 6
(1) (2)

3Quantile → 1 2 | 3 4 | 5 6
(1) (2) (3)

Quartile → 1 2 3 | 4 6 5 | 6 7 8 | 10 11 12
(1) (2) (3) (4)

percentile → 99 points that seperates datasets into 100 equal portions

~~1 percen~~ 1 quartile = 25 ~~percentile~~ টি ডাটা
4 " = ~~99~~ 100 টি ডাটা = 1 percentile.

Inter Quartile Range : IQR = $Q_3 - Q_1$; outliers < 1.5IQR < outliers

(Even length) Example: 40, 36, 47 | 50 52 52 | 56 60 63 | 70 70 1000
Q_1 median Q_3

Which datapoints are outliers?

(** sorted না থাকলে sort করে ডাটা বসাব)

Here, $Q_1 = \frac{47+50}{2} = 48.5$

$Q_3 = \frac{63+70}{2} = 66.5$.

~~1.5 IQR = 1.5 × 48.5 =~~

~~1.5 IQ3R = 1.5 × 66.5 = 99.75~~

$IQR = Q_3 - Q_1 = 66.5 - 48.5 = 18$

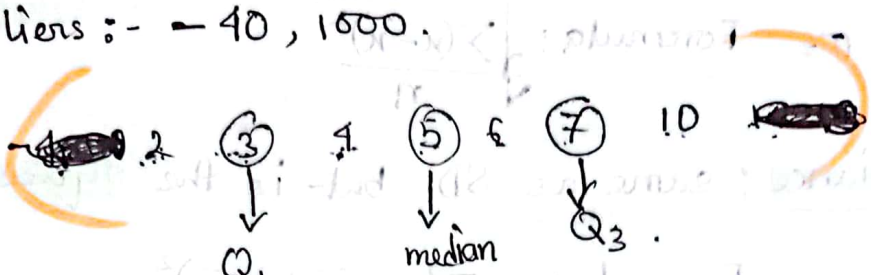
So, $1.5 \times IQR = 1.5 \times 18 = 27$.

$Q_3 = 27 + 66.5 = 93.5$ (more than 93.5)

$Q_1 = 48.5 - 27 = 21.5$. (less than 21.5)

Ans: outliers :- $-40, 1000$.

(odd length) Example:



$Q_1 = 3$

$Q_3 = 7$

So, $IQR = 7 - 3 = 4$.

$1.5\ IQR = 6$.

Now, $Q_1 = 3 - 6 = -3$.

$Q_3 = 7 + 6 = 13$.

So, value less than -3 and more than 13 are outliers.

~~Here, outliers 2~~. (Ans).

step-1 mark Q_1, Q_3, median on sorted data.

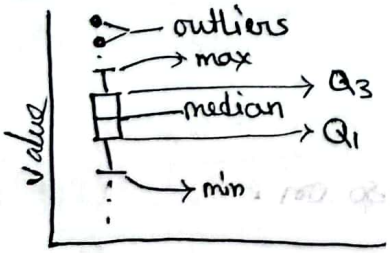
step-2 $IQR = Q_3 - Q_1$.

$Q_1' = Q_1 - 1.5IQR$

$Q_3' = Q_3 + 1.5IQR$

step-3 find out less than Q_1' and more than Q_3' → outliers.

Boxplot : To visualize dataset-ly value



Standard deviation: Measures how spread out the ~~number~~ datapoints are from the mean value

σ = Formula: $\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$

Variance: same as SD but is the square of SD.

Formula: $\sigma^2 = \frac{\sum(x-\bar{x})^2}{n}$

Proximity Measures: similarity/dissimilarity in the data.

Data Matrix = $\begin{bmatrix} L_1 & L_2 & L_3 & \cdots \\ S_1 & S_2 & S_3 & \cdots \end{bmatrix}$

Dissimilarity Matrix = $\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$

(*) For Nominal attributes (colors, genders, bg, city, marital status etc).

$d(i,j) = \frac{p-m}{p}$ | p = total attribute considered. | m = no. of matched attribute

Example:

ID	test 1	test 2
1	codeA	A
2	codeB	A
3	codeC	B
4	codeA	C

$d(2,1) = \frac{2-1}{2}$ (codeA, codeB; test2 भिन्नहुन्छ)

$= \frac{1}{2} = 50\%$.

$d(4,3) = \frac{2-0}{2} = 1$

$= 100\%$

	1	2	3	4
1	0			
2	d(2,1)	0		
3	d(3,1)	d(3,2)	0	
4	d(4,1)	d(4,2)	d(4,3)	0

Thus $d(2,1) \Rightarrow$ 50% dissimilar

$d(4,3) \Rightarrow$ 100% dissimilar. and so on.

(*) For Ordinal (Size, ratings, levels, ~~etc~~ status, etc) numbering করা যায়

Step-1: Numeric representation.

Object	satisfaction level (test2)	
1	Excellent (3)	1
2	Fain (1)	0
3	Good (2)	0.5
4	Excellent (3)	1

Let, fair = 1
good = 2
excellent = 3.

Step-2: Normalize $\Rightarrow Z_{if\,feature} = \frac{Value_{rf} - 1}{Max_{rf} - 1}$ (row, test2)

Here, Maximum feature, $Max_f = 3$.

So, $Z_{12} = \frac{3-1}{3-1} = \frac{2}{2} = 1$

$Z_{22} = \frac{1-1}{3-1} = \frac{0}{2} = 0$

$Z_{32} = \frac{2-1}{3-1} = \frac{1}{2} = 0.5$

$Z_{42} = \frac{3-1}{3-1} = 1$

Step-3: Calculate dissimilarity. $\longrightarrow d(i,j) = \sqrt{(i-j)^2 + (i-j)^2 - - -}$ (test1, test2)

$d(2,1) = \sqrt{(1-0)^2} = 1 \longrightarrow$ 100% dissimilar

$d(3,1) = \sqrt{(0.5-1)^2} = 0.5 \longrightarrow$ 50% dissimilar

$d(4,1) = \sqrt{(1-1)^2} = 0 \longrightarrow$ 0% dissimilar

$d(3,2) = \sqrt{(0.5-0)^2} = 0.5 \longrightarrow$ 50% "

$d(4,2) = \sqrt{(1-0)^2} = 1 \longrightarrow$ 100% "

$d(4,3) = \sqrt{(1-0.5)^2} = 0.5 \longrightarrow$ 50% "

Min Max Normalization

$\frac{v - min}{max - min}$

(*) For numeric : Each datapoint- rowwise distance বের করব,

3 approaches:

(i) Euclidean : $d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (\cdots)}$

(ii) Mahattan : absolute value usage

$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{ip}-x_{jp}|$

(iii) Minkowski : generalized one.

$d_{i,j} = (x_{i1}-x_{j1})^h + (x_{i2}-x_{j2})^h + \cdots + (x_{ip}-x_{jp})^h$

if $h=1$, then Manhattan

" $h=2$ " Euclidean

(*) For binary :

Step-1 : Contingency table বানাব.

Feature 1 → / Feature 2 ↓	1	0	
1	q (11)	r (10)	q+r
0	s (01)	t (00)	s+t
	q+s	r+t	

For symmetric attribute (দুটো state-ই important):

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

For assymmetric attribute (যেকোনো একটা state not imp):

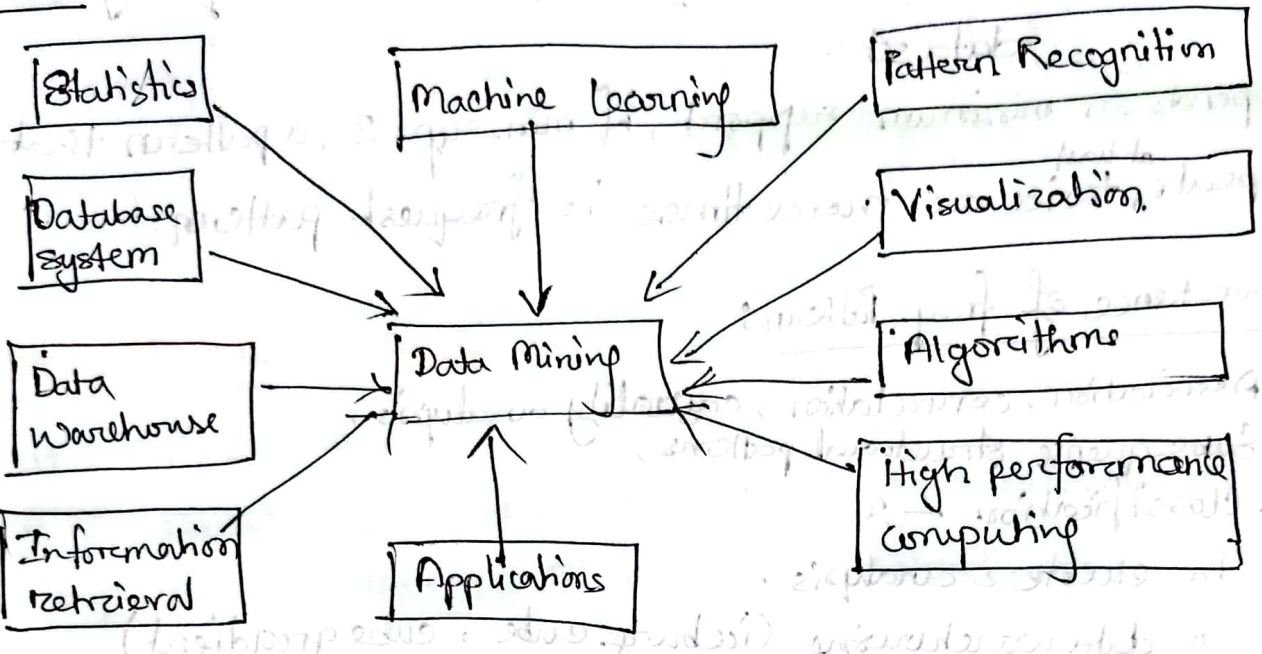
$$d(i,j) = \frac{r+s}{q+r+s}$$

[t = (0,0) state নাও করতে পা]

Theory:

(*) "Data Mining adopts techniques from various domains".

Ans:



Describe:

Statistics: Hypothesis testing, regression.
ML: Decision tree, clustering, neural nets.
Database Systems: Efficient data retrieval, storage.
Info retrieval: handle unstructure data
Pattern: complex data trends
visualize: meaningfully representation.

Stats ML PR
Datas DM VIS
DW Algo
InfoR APP Hperf computing

(**) Support → 5%. $[(A \wedge B) \rightarrow C.]$ if A∧B, then C
Conf → 85%.

Support কম হলে conf কেমন হবে why?

* support: 1000-এ 50 জন (A∧B) condition true $\left[\frac{5 \times 1000}{100} = 50\right]$
* conf: তাই 50 জনে $\left[\frac{85}{100} \times 50 \approx 43 \text{ জন}\right]$ → actually does C while (A∧B) true.

(বাকি 7 জন (A∧B) true but C true না)

Chap-6

* **Frequent Pattern:** A pattern (a set of items, subsequences, substructures etc) that occurs frequently in a data set.

Depends on minimum support, if min_sup = 2, a pattern that repeats at least twice or more times is frequest pattern

⇒ **Importance of freq. Pattern:**

1. Association, correlation, causality analysis.
2. Subsequence structural patterns,
3. Classification →
4. In cluster analysis.
5. " data warehousing (iceburg cube, cube gradient)
6. Semantic data compression.
7. Broad applications.

Frequent itemset Formulae:

$$\text{Support } (A \to B) = P(A \cup B) = \frac{\text{count} - (A \cup B)}{\Sigma \text{transactions}}.$$

$$\text{Confidence } (A \to B) = P(B|A) = \frac{\text{count } (A \cup B)}{\text{count } (A)}.$$

Association Rule:

Support: Probability that a transaction contains $A \cup B$.

Confidence: Conditional probability that a transaction having A also contains B.

Example:

TID	Itemset
1	{bread, butter, banana}
2	{Mango, berry, banana}
3	{bread, butter, banana}
4	{mango, banana}
5	{bread, butter, oil}

ধরি,
A = {bread, butter}
B = {banana}.

$$support = \frac{c(A \cup B)}{\Sigma trans} = \frac{2}{5}.$$

$$conf = \frac{c(A \cup B)}{c(A)} = \frac{2}{3}$$

✱ Closed Frequent Itemset :

* will consider support count
* কোনো subset -এর sup-count যদি তার কোনো superset -এর sup-count -এর similar হয়, then সেই subset বাদ

Ex Suppose,
① {milk, bread, butter, oil} — 2 time (sup-count)
② {milk, bread, butter} — 3 times
③ {milk, bread} — 3 times.

এখানে ② is subset of superset ①
again ③ is subset of supersets ① and ②.
but ③ ②-এর sup-count same, thus we will remove ③

Closed → sub = sup (support count)
Maximal → sup supcount >= min_supcount

→ subset বাদ দাও

* Maximum Frequent Itemset

* will only consider ~~too~~ minimum sup-count.
* If superset not frequent than subset, then keep subset
* " " is frequent than subset, remove subset.

☐ Super subset → sup-count >= min sup count; subset ✓

* Challenges of Apriori Method:

① Multiple scans.
② Huge # of candidates.
③ Tedious workload of support counting for candidates.

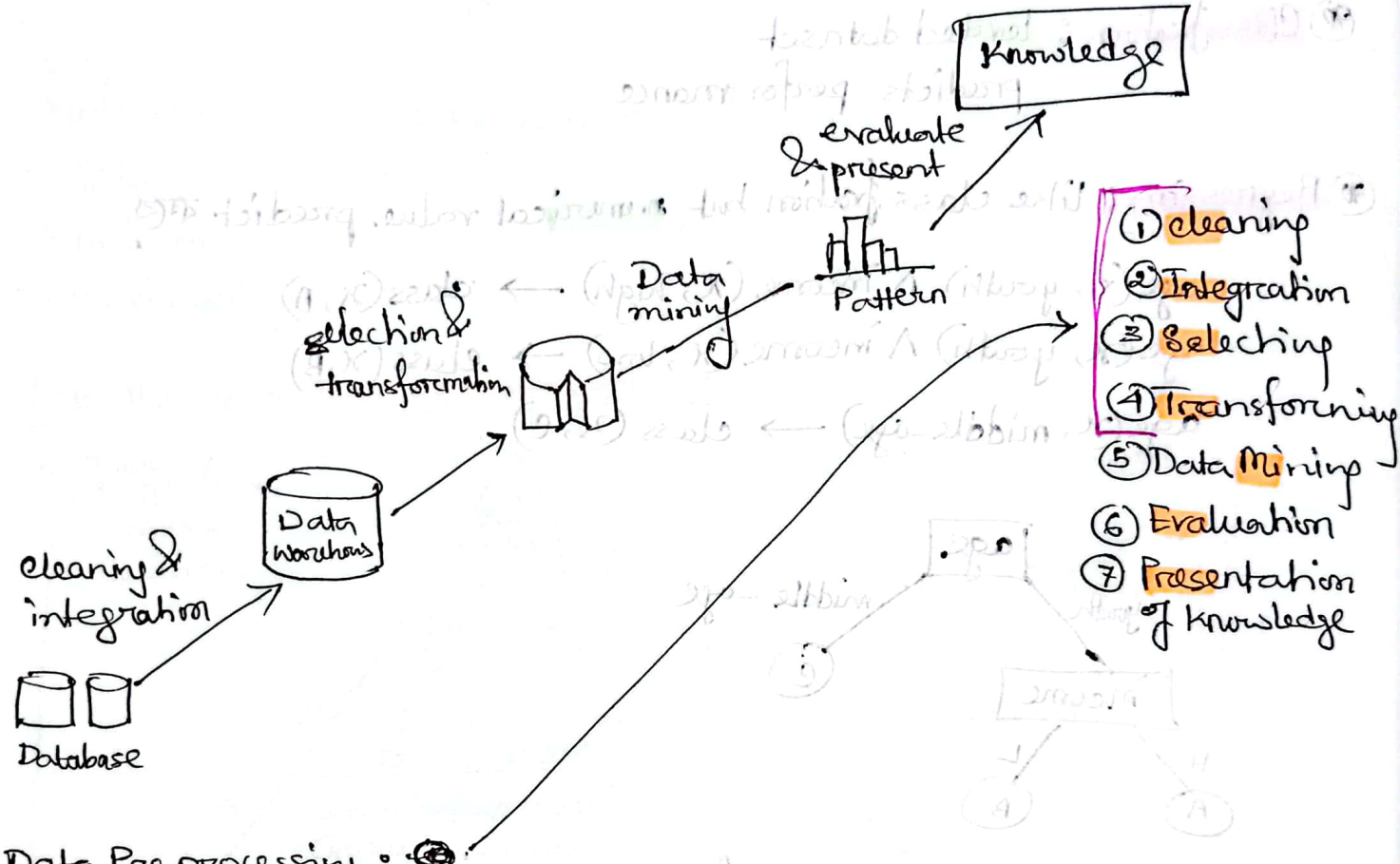
* Improve Apriori

① Reduce passes of transaction database scans (twice - local, global)
② Shrink no. of candidates.
③ Facilitate support counting of candidates.

Chapter - 1

Data Mining: Extracting interesting patterns or knowledge from huge amount of data.

Knowledge Discovery from Data (KDD)



(A) Data Pre processing : (B)

(B) What kind of data can be mined ?

1. Data Streams / Sensor data
2. Time series
3. Structure, graphs, networks
4. Obj-relational databases
5. Heterogeneous data.
6. Spatial - spatiotemporal data
7. Text
8. Audio, video, www.

(M) Patterns that can be mined :

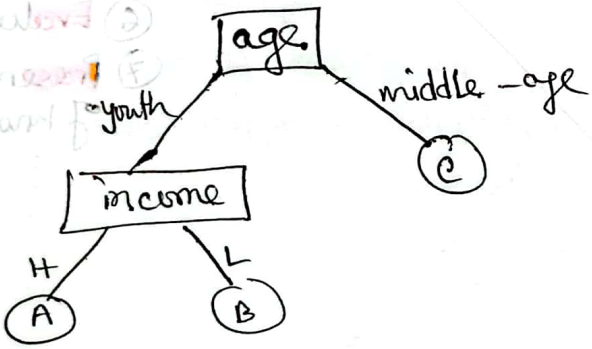
1. Generalization
2. Characterization, Discrimination
3. Frequent Pattern
4. Association & correlation
5. Classification & regression,
6. clustering analysis
7. Outlier analysis.

(*) Frequent patterns : ① Itemset → { bread কিনলে সাথে butter কিনে }
② Subsequence → { pc কিনলে, পরে গিয়ে OS লাগে }
③ Substructures → { chemistry, geometrics etc, 2H থাকলে O থাকে → H_2O } (mostly)

(*) Classification : labeled dataset
predicts performance

(*) Regression : like classfation but numerical value predict করে,

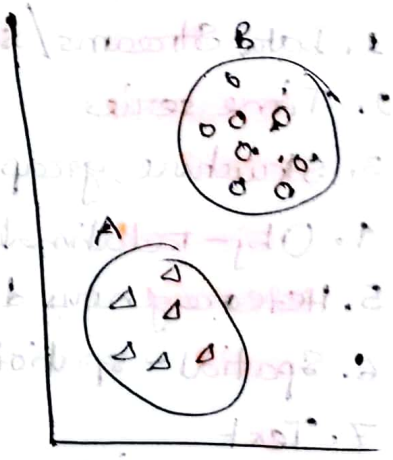
age(x, youth) ∧ income(x, high) → class(x, A)
age(x, youth) ∧ income(x, low) → class(x, B)
age(x, middle-age) → class(x, C)



** too much branch = bad performance.

(*) Cluster Analysis :
1. Unsupervised learning.
2. Groups, plots in graph to find pattern.
3. Principle : maximize intra class similarity
minimize inter class

4 Common character → cluster
datapoints.

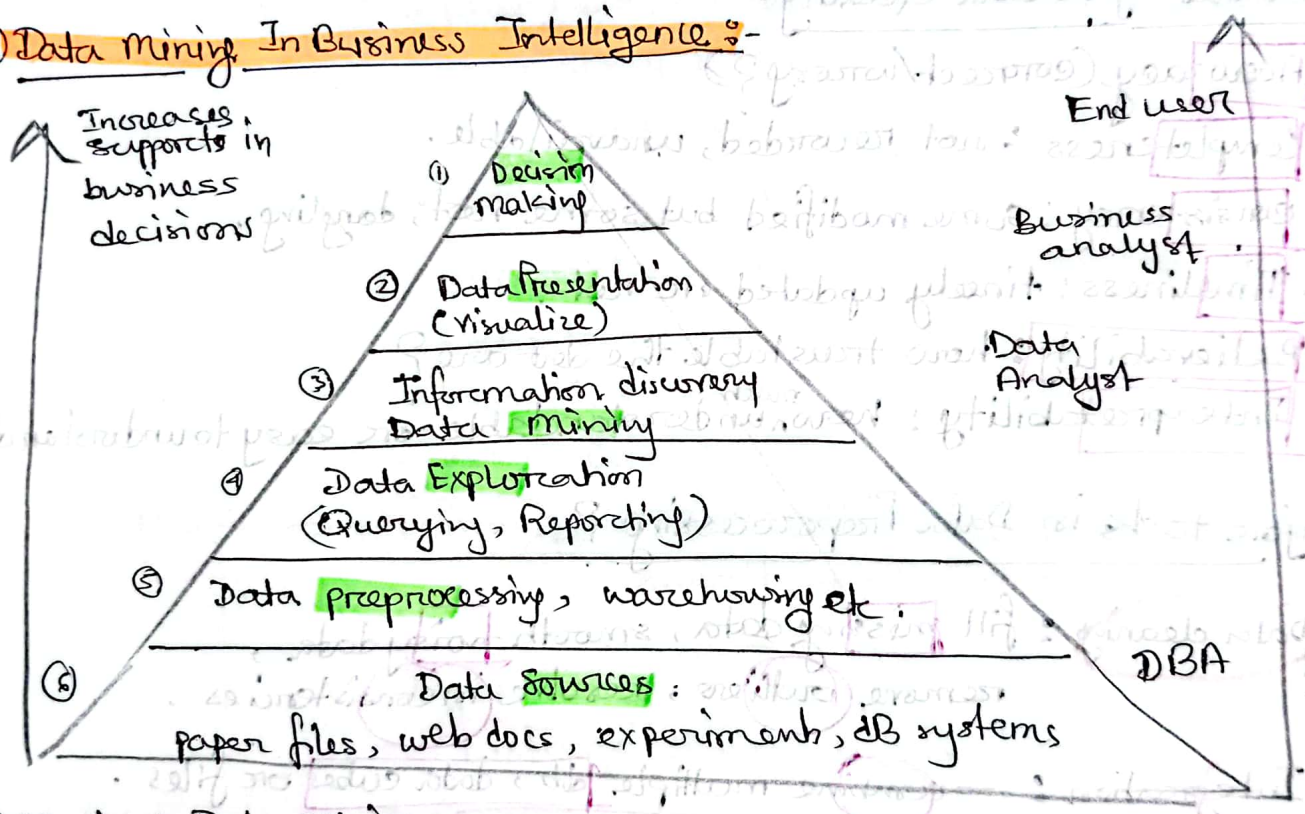


* **Outlier Analysis** : 1. data not complying with general behaviour
2. Noise
3. Exception.
4. Methods : By product of clustering / regression analysis.
5. Fraud detection, rare events analysis.

Are all pattern interesting ?

→ No
→ interesting – when human understand it
→ Valid on new data.
→ potential.
→ novel, unique.

Data Mining In Business Intelligence :-



Medical Data Mining :

→ Statistics, ML
→ Preprocess data (feature extraction, dimension reduction).
→ Classification, clustering.
→ post processing for presentation.

Mining Methodology:

1. Mining various, new kinds of knowledge
2. Mining in multidimensional space (sub-attribute-এ হাতে ধরা)
3. An interdisciplinary effort
4. Boosting the power of discovery in a networked environment
5. Handling noise, uncertainty, incompleteness of data
6. Pattern evaluation
7. Pattern/Constraint guided mining.

Chapter-3

Measures for data Quality:

1. Accuracy (correct/wrong??)
2. Completeness: not recorded, unavailable
3. Consistency: some modified but some not, dangling.
4. Timeliness: timely updated or not.
5. Believability: how trustable the dat are?
6. Interpretability: how much understandable or easy to understand?

Major tasks in Data Preprocessing?

1. Data cleaning: fill missing data, smooth noisy data, remove outliers, resolve inconsistencies.
2. Integration: — combine multiple db, data cubes or files.
3. Data reduction: Dimensionality reduction
   Numerosity "
   Data compression
4. Data Transformation: — Normalization
   — Concept hierarchy generation.

Data Cleaning:

MISSING DATA : 1. Ignorance the tuple (not good)
2. Fill in manually (not possible for huge data)
3. Use global constant
4. " central tendency (mean/median/mode).
attribute wise
5. Central tendency for same attribute and same class.
6. Most probable value (using regression model).

NOISY DATA : (Random error or variance in a measured variable is called Noise.)

→ due to - faulty data collection instruments
- data entry problems.
- data transmission "
- technology limitation
- inconsistency in naming convention.
- duplicate records.
- incomplete "
- inconsistent "

Handle → 1. Binning - partition into equal frequency.
smooth by bin means/median/boundaries.

2. Regression
3. Clustering - detect + remove outliers
4. Computer ft. human inspection
(detect করতে check করবে).

Data Integration:

Handle redundancy: correlation analysis
covariance "

(1) Nominal → χ^2 - chi square value $= \sum \left\{ \frac{(\text{actual} - \text{expected})^2}{\text{expected}} \right\}$

(large = related)

(2) Numeric → Pearson's product-moment coefficient.

$$r_{A,B} = \frac{\sum (a - \bar{a})(b - \bar{b})}{\sigma_a \sigma_b}$$

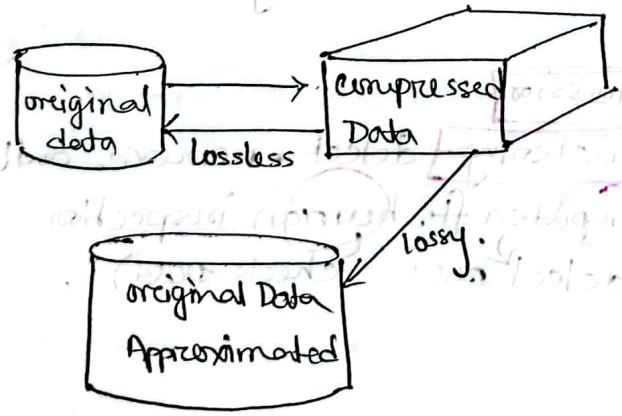
$$r_{AB} = \frac{\text{cov}(A,B)}{\sigma_A \cdot \sigma_B} = \frac{\frac{\sum (A.B)}{n-1} - \bar{A} \cdot \bar{B}}{\sigma_A \ \sigma_B}$$

$r_{AB} > 0$, positive correlation.
$r_{A,B} < 0$, negative "
$r_{A,B} = 0$, independent.

Data reduction:

(1) Dimensionality reduction:

1. Data compression:
   a) Principle Component Analysis.
   b) Wavelet Transformation.



2. Attribute subset selection. (Forward + Backward selection)
3. Attribute construction.

② Numerosity reduction :

a) Regression : — linear regression $[Y = wX + b]$

— multiple " $[Y = b_0 + b_1 X_1 + b_2 X_2]$

— log-linear models.

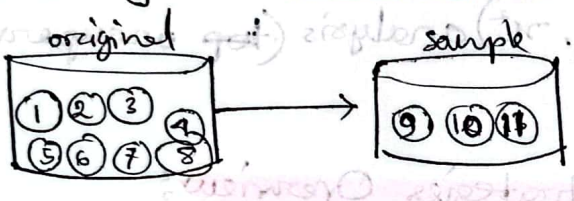
b) Histogram Analysis :

c) clustering

d) Sampling :- taking a random sample from dataset of N.

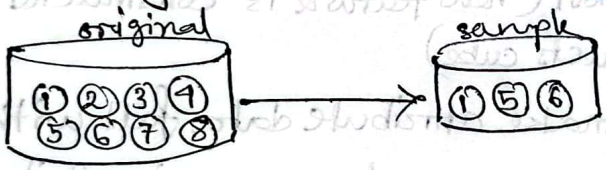
(*) SRSWOR (Simple Random Sample without replacement)

→ selected object is removed from the population.

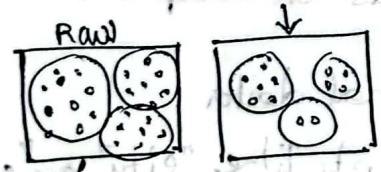


(*) SRSWR (Simple Random Sample with Replacement).

→ selected object not removed from the population.



(*) Cluster : — main memory-তে process-এর জন্য part saved

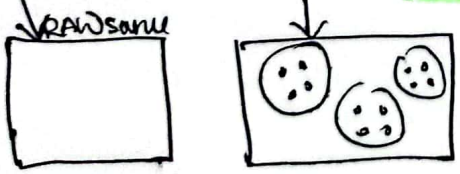


— Full dataset is divided into k fold and randomly sampled out.

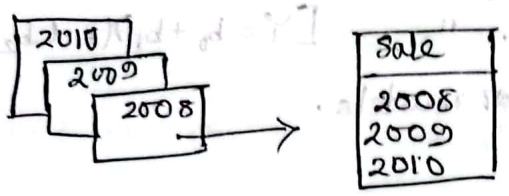
— SRSWOR or SRSWR — যেকোনো could be used.

— imbalanced হয়ে যায়।

(*) Stratified : balanced sample — each class same instance.



e) Data Cube Aggregate: Combine data into cube but may loss data.



✱ Data Discretization Methods:

1. Binning
2. Histogram Analysis (top down split, Unsupervised)
3. Clustering Analysis (" " / bottom up merge split, Unsupervised)
4. Decision tree analysis (" " split, supervised)
5. Correlation ($e.g.\ \chi^2$) analysis (~~top~~ unsupervised, bottom-up merge)

✱ Data Transformation Strategies Overview:

1. Smoothing (remove noise: Binning, regression, clustering).
2. Attribute construction (new feature is constructed + added)
3. Aggregation (constructs cube)
4. Normalization (to make attribute data fall within small range)
5. Discretization (raw values of a numeric attribute are replaced by interval labels or conceptual labels.
6. Concept ✱ hierarchy generation for nominal data.
   ("streets" generalized to higher level concepts like "city" or "country")

Normalization:

* min-max → $\frac{\text{actual value} - \text{min}}{\text{max} - \text{min}}$
* Z-score → $\frac{x - \bar{x}}{\sigma}$
* decimal scaling.

(*) Ass. Rule:

$$\text{conf}(A \to B) = P(B|A) = \frac{\text{sup\_cnt}(A \cup B)}{\text{supcnt}(A)}$$

[y% chance of a new data of $A \to B$]

$$\text{support}(A \to B) = P(A \cup B) = \frac{\text{count}(A \cup B)}{\text{Total} \ldots}$$

[In trainy data, x% case among 100 sample of $A \to B$]

(*) Apriori:

Base table

$I_1 I_2 I_5$
$I_2 I_4$
$I_2 I_3$
$I_1 I_2 I_4$
$I_1 I_3$
$I_2 I_3$
$I_1 I_3$
$I_1 I_2 I_3 I_5$
$I_1 I_2 I_3$

I_1 I_2 I_3 I_4 I_5

Step-1

C_1 (1 item) (each count):

I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

→ L_1 সব নাও

s = 20%
c = 80%
n = 9

so, min-support $\left\lceil \frac{20}{100} \times 9 \right\rceil = 2$.

(2 <= count মানে বাদ্যের)

Step-2 (2 items) (Pas eligible : I_1 I_2 I_3 I_4 I_5)

C_1

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

L_2

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

Step-3 (3 items) (Eligible: I_1, I_2, I_3, I_4, I_5)

C_3:

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

L_3:

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

$I_1 I_2 I_3$
I_n
I_5

$I_2 I_3 . I_4$
I_5

$I_3 I_4 I_5$

Step 4 - (4 items) (Eligible: $I_1 I_2 I_3 I_5$),

So. | $I_1 I_2 I_3 I_5$ | 1 | not 2 or more.

Thus fq set = $I_1 I_2 I_3$
$I_1 I_2 I_5$.

Math आएगा ?

(*) Bin.

Equal fq ⟶

Bin1 : 4 8 15
Bin2 : 21 21 24
Bin3 : 25 28 34

(*) Bin mean:

Bin1 : 9 9 9
22 22 22
29 29 29

Boundaries:

Bin1: 4 4 15
Bin2 : 21 21 24
Bin3 : 25 25 34: