17/01/23

1st class

Page No.:
Date : / /

Data Mining & Warehousing

Data Mining is classified into 3 types:-

i) Supervised → provided set of features along with associated labels
ii) Unsupervised
iii) Market - Basket Learning Analysis
↳ analyzing shopping cart of customer

Data Mining- It is the process of analysing the data from different perspective and summarizing it into useful information - Info. that can be used to increase revenue, cut-cost or both.

Data Mining Strategies-

Supervised learning	Unsupervised Clustering	Market-Basket Analysis

Supervised learning-

When we are young, we use induction to form basic concept definition, we see instances of concepts representing animals, plants, buildings etc. We here labels given to individual instances and choose what we believe to be that defining concepts, features (attribute) from our own classification models.

Later we use the model we have developed to help us identify objects of similar structures. The name of this type of learning is supervised concept learning or just supervised learning.

Page No.:
Date: / /

Supervised Learning

Classification Regression

Classification - It is a supervised learning task where output is having defined labels

There are 3 types of classification

i) Binary
ii) Multiclass
iii) Multilabel

→ In Binary classification, the model predicts either '0 or '1', 'yes' or 'no' but in case of multiclass classification, the model predicts any one class from more than two classes. In multilabel classification anyone instance can have more than 1 label or classes.

Regression - It is a supervised learning task. where O/P is having continuous value.
It is generally divided into 2 parts:-

i) Linear Reg.
ii) Logistic Reg.

Semi-supervised Learning - It is a supervised learning where the training data contains very few labelled examples and a large no. of unlabelled examples.
In this initially, similar data is clustered along with an unsupervised learning algo. & further it helps to label the unlabelled data into labelled data.

Page No.:
Date : / /

Semi-supervised learning lies b/w supervised & unsupervised learning.

Unsupervised clustering - Unlike supervised learning, unsupervised learning builds model from data without pre-defined classes. Data instances are grouped together based on the similarity scheme defined by the clustering system.

→ Clustering is the process of grouping the data into clusters, so that objects within a cluster have high similarity in comparison to one-another but are very dissimilar to objects of other clusters

18/01/23

Data mining is applied on varying no. of data:-

i) RDBMS

ii) Transaction database

Transaction ID	Items
T_1	I_1, I_2, I_3
T_2	I_7, I_{235}, I_6
T_3	I_1, I_{500}, I_2

iii) Temporal database - One attribute of time must be present

iv) Time-Series Databases -

Page No.:
Date : / /

Data Mining can be applied on many kinds of data :-

i) Transactional Databases - It consists of a file where each record represent a transaction. A transaction typically includes a unique transaction id no. and a list of items making of the transaction.

ii) Temporal Databases - It stores and manages time-varying data e.g :- financial, medical, government, etc. It maintains historical information, incorporate notion of time.

iii) ~~Incorporate Notion of Time~~

iii) Time-Series Databases - A time-series is a sequence of data points, measured typically at successive points in time spaced at uniform time-interval.
A time-series database stores sequence of values or events obtained over repeated measurement of time ex. hourly, daily, weekly

iv) Spatial Databases - It contains spatial related information ex- included geographical map database, satellite image databases.
→ A spatial databases that stores objects that change with time is called spatio temporal database.

Page No.:
Date : / /

Data Cleaning :-

1) Missing Values -

a) Ignore the tuple - This is usually done when several attributes with missing values are present or the class label is missing.

b) Fill in the missing values manually - ~~This~~ In general this approach is time-consuming & may not be feasible given a large datasets with many missing values.

c) Use a global constant to fill in the missing values - Replace all missing values by the same constant such as label like unknown, infinity, zero, etc. The problem with this is that the mining program may mistakenly conclude that they form a interesting concept since they all have a value in common.

d) Use attribute mean to fill in the missing values - Missing value is replaced by the mean value of the column.

e) Use the attribute mean for all samples belonging to the same class as the given tuple -

f_1	f_2	f_3	f_4	Income	class
v_1	$-v_4$	v_3	v_4	100	c_1
v_5	v_6	v_7	v_8	200	c_1
v_5	v_3	v_4	v_5	?	c_1 ←
v_1	v_2	$-v_3$	v_1	300	c_2
v_2	v_3	v_{19}	v_{20}	400	c_2

f) Use the most probable value to fill in the missing value - (Linear Regression)

~~g)~~

ii) Noisy Data :-

a) Binning - In this binning method smooth the sorted data values by consulting its neighbourhood that is the value around it. The two common techniques that are used are as follows :-

i) Data Smoothing by Bin means
ii) Smoothing by Bin boundaries

ex! 4, 8, 15, 21, 21, 24, 25, 28, 34

i) Data Smoothing by Bin means

Bin 1 : 4, 8, 15 → 9, 9, 9
Bin 2 : 21, 21, 24 → 22, 22, 22
Bin 3 : 25, 28, 34 → 29, 29, 29

Page No.:
Date : / /

a) Smoothing by Bin Boundaries

Bin 1: 4, 8, 15 → 4, 4, 15

Bin 2: 21, 21, 24 → 21, 21, 24

Bin 3: 25, 28, 34 → 25, 25, 34

b) Regression - Data can be smooth by fitting the data to a function such as with regression

Clustering - Outliers may be detected by clustering ~~their~~ where similar values are organised into groups or clusters.
Outliers are those value which dont belong to any clusters.

1/02/23

26.
60
40

24
21
80
135

Page No.:
Date : / /

Mean, Median, Mode

ex: 13, 18, 13, 14, 13, 16, 14, 21, 13

13, 13, 13, 13, 14, 14, 16, 18, 21

$$\text{Mean} = \frac{13\times4 + 28 + 16 + 18 + 21}{9} = \frac{135}{9} = 15$$

$$\text{Median} = \left(\frac{9+1}{2}\right)^{th} \text{ element} = 5^{th} \text{ element} = 14$$

Mode =

ex: 10, 12, 13, 16, 17, 18, 19, 21 (8)

$$\text{Mean} = \frac{126}{8} = 15.75$$

$$\text{Median} = \left(\frac{8}{2}\right)^{th} + \left(\frac{8}{2}+1\right)^{th}$$

$$= \frac{16+17}{2} = \frac{33}{2} = 16.5$$

48, 44, 48, 45, 42, 49, 48

Mode = 48.

⇒ If we have data with same no. of frequency, then no mode exists

2000
1150
44
3194

Age	Frequency	
1-5	200	200
5-15	450	650
15-20	300	950
20-50 (median interval)	1500	2450
50-80	700	3150
80-110	44	3194

find mean, median and mode.

Formula

$$\text{Median} = L_1 + \left(\frac{N/2 - (\Sigma \text{freq})_L}{\text{freq}_{\text{median}}} \right) \text{width}$$

where L_1 is the lower boundary of the median interval, and N is the no. of value in the entire dataset

$(\Sigma \text{freq})_L$ is the sum of the frequencies of all the intervals that are lower than median interval

$\text{freq}_{\text{med.}}$ - frequency of the median interval

width - width of median interval

How to find median interval ?

Soln :

$N = 3194$ $\quad \text{freq}_{\text{median}} = 1500$

$L_1 = 20$

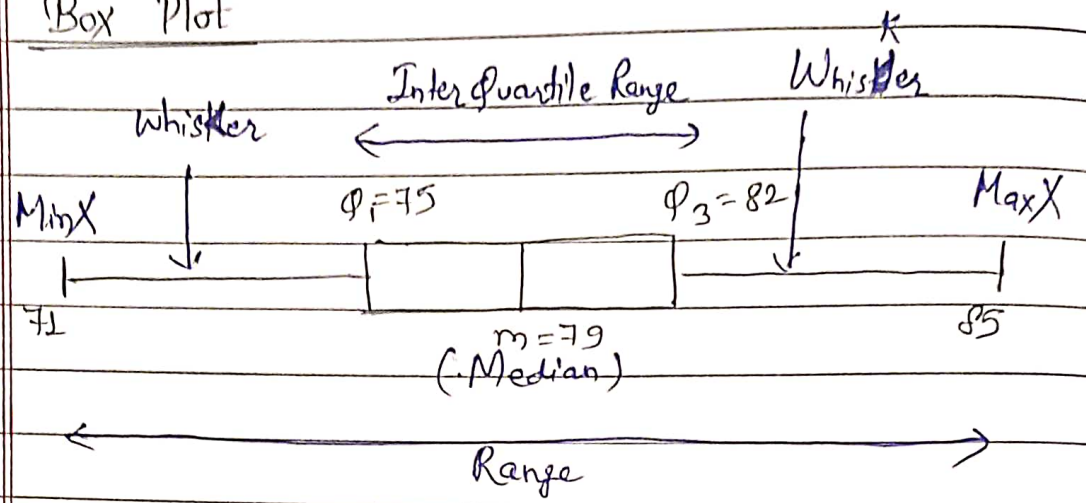
$(\Sigma \text{freq})_L = 950$ $\quad$ width $= 30$

$\frac{3194}{2} = 1597$

Page No.:

Date : / /

Box Plot



Data:

76, 79, 76, 74, 75, 71, 85, 82, 82, 79, 81

71, 74, 75, 76, 76, (79), 79, 81, 82, 82, 85

N = 11

Median = $\left(\frac{11+1}{2}\right)^{th}$ element = 79

Median of left half = 75

" " right " = 82

Min element = 71

Max element = 85

ex: Calculate the standard deviation,

4, 2, 5, 8, 6

formula

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$$

$$\bar{x} = \frac{25}{5} = 12.5$$

$$\Sigma (x - \bar{x})^2 = (4-12.5)^2 + (2-12.5)^2 + (5-12.5)^2 + (8-12.5)^2 + (6-12.5)^2$$

$$=$$

$$\Sigma (x - \bar{x})^2 = 1^2 + 3^2 + 0^2 + (-3)^2 + (-1)^2$$

$$= 1 + 9 + 9 + 1 = 20$$

$$\sigma = \sqrt{\frac{20}{5}} = \sqrt{4} = 2$$

Correlation Analysis

Co-relation

Pearson ~ +ve co-related

-ve co-related

⇒ Pearson Co-relation values lies btw -1 to 1

~~If the value are~~

⇒ Gives two attributes, such analysis can measure how strongly one attribute implies the other, based on available data. The value of pearson co-relation lies between -1 to +1.

Pearson - $(-1 \leq r_{A,B} \leq +1)$

If $r_{A,B} > 0$ then A and B are +ve ly co-related. (ie. when the value of A increases, the value of B also increases). The Higher the value, the stronger the co-relation.

If the resulting value is equal to zero then A and B are independent and there is no co-relation btw them.

If the resulting values is less than zero, then A & B are -vely co-related.

$$r_{A,B} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^{N} (a_i - \bar{A})^2 \sum_{i=1}^{N} (b_i - \bar{B})^2}}$$

ex:

Tree Height (y)	Trank Diameter (x)
35	8
49	9
27	7
33	6
60	13
21	7
45	11
51	12

$r = 0.886$ Ans

190
80
51
322

$\bar{y} = \frac{321}{8}$

$= 40.125$

$\bar{x} = \frac{73}{8}$

$= 9.125$

Page No.:
Date : / /

X^2 (chi-square) Test :-

for categorical data, a relationship btw two attributes A and B can be discovered by chi-square Test.

Suppose A has 'c' distinct values $a_1, a_2, a_3, \ldots, a_c$ and B has 'r' distinct values $b_1, b_2, b_3, \ldots, b_r$ let (A_i, B_i) denotes the event that attribute A takes the value A_i and B takes the value B_i. This is represented by a distinct slot in the table. This table is known as **Contingency Table**.

The chi-square value is computed as

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

Page No.:
Date: / /

where O_{ij} is the observed frequence (ie the actual count) of the joint event.

And e_{ij} is the expected frequency of (A_i, B_j). e_{ij} is computed as

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

where N is the no. of data tuples

$$\text{Degree of freedom} = (r-1) \times (c-1)$$

ex:-

	male	female	Total
fiction ~~(90)~~	250 (90)	200 (360)	450
Non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

$$e_{11} = \frac{count(male) \times count(fiction)}{N}$$

$$= \frac{300 \times 450}{1500} = 90$$

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

$$= 507$$

Page No.:
Date : / /

Degree of freedom = $(r-1) \times (c-1)$
$= (2-1) \times (2-1) = 1$

Significance level → 5% = 3.84

If $\chi^2 > 5\%$ → Co-related else Independant

7/02/22nd Data Information

i) Smoothing – It works to remove noise from the data, such techniques include, binning, regression or clustering.

ii) Aggregation – where summary or aggregation generations are applied to the data.
ex – Daily sales may be aggregated to compute monthly or annual sales.

iii) Generalization – where low level primitive data are replaced by higher level concepts through the use of concept hierarchies.
ex – city can be generalized to higher level concept like state or country.

iv) Normalization – where the attribute data are scaled so as to fall within a specified range.
ex – -1 to +1, 0 to +1

v) Attribute construction- where new attributes are constructed from a given set of attributes

Min-Max Normalization :-

$$V' = \frac{V - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

where max_A and min_A are maximum and minimum values of an attribute A.

Min-Max Normalization maps a value V of an attribute A to a value V' in the range new_min_A to new_max_A.

Q. Suppose the minimum and maximum value of attribute income are Rs 12,000 and Rs 98,000 respectively. Map the income in the range 0.02 to 1.0 by using min-max normalization. And the normalized value of Rs 73,600 of attribute income.

Ans- 0.716

$$V' = \frac{73600 - 12000}{98000 - 12000} (1.0 - 0.02) + 0.02$$

$$= \frac{61600}{86000} (0.08) + 0.02$$

Page No.:
Date : / /

z-score Normalization - The values of an attribute A, are normalized based on the mean and standard deviation of A.

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

Q. The mean and the standard deviation of the values for the attribute income are Rs 54000 and Rs 16000 respectively. With z-score normalization transform the value of income Rs 73600.

$$v' = \frac{73600 - 54000}{16000} = \frac{1966\phi}{160\phi\phi} = \frac{196}{160} = 1.225$$

Data Warehousing :- (→ storage of data)

definition given by Kliliam H. Inmon

→ According to William H. Inmon, a data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of mangement decision-making process.

i) Subject-oriented- A data warehouse is organized around major subject areas such as customer, supplier, etc

ii) Integrated- A data warehouse is usely constructed by integrating multiple heterogeneous sources.

iii) Time-Variant - Data are stored to provide information from a historical perspective. Every key structure in the data warehouse contains either implicitly or explicitly an element of time.

iv) Non-Volatile - It usely requires only two operations in data accessing : intial loading of data and accessing of data.

Location = "Buxar"

item (type)

Time (Quarter)	Home Entertainment	Computer	Phone
Q_1	605	365	12
Q_2	625	205	13
Q_3	425	195	19
Q_4	365	315	29

Location = "Patna"

item (type)

Time Quarter	~~Time~~ Home ~~Entertainment~~ (Quarter)	~~Home Computer~~ Computer ~~Entertainment~~	Phone
Q_1	395	1600	
Q_2	200	3000	
Q_3	900	200	
Q_4	400	400	

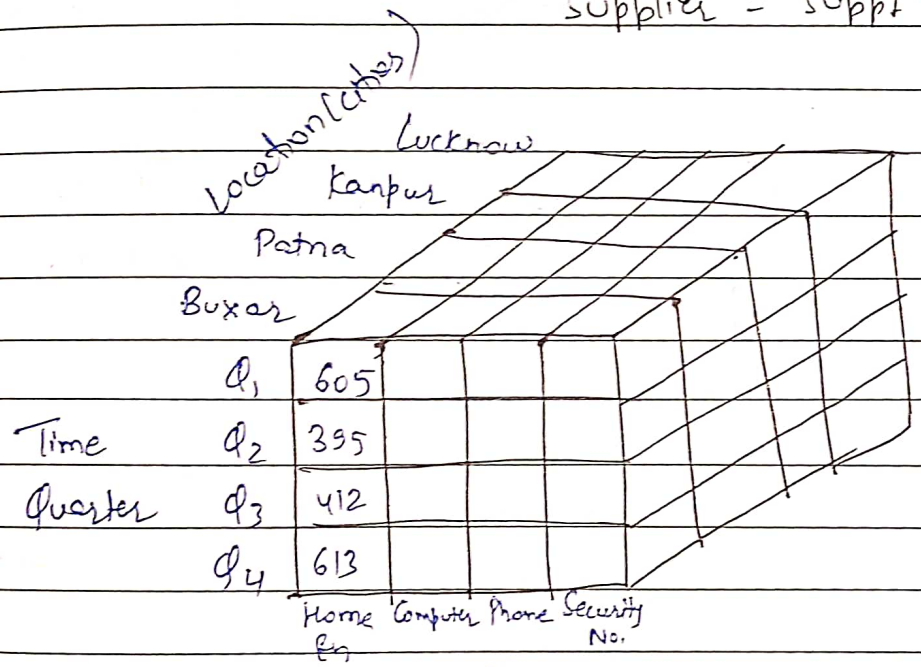
Page No.:
Date: / /

Location = "Kanpur"

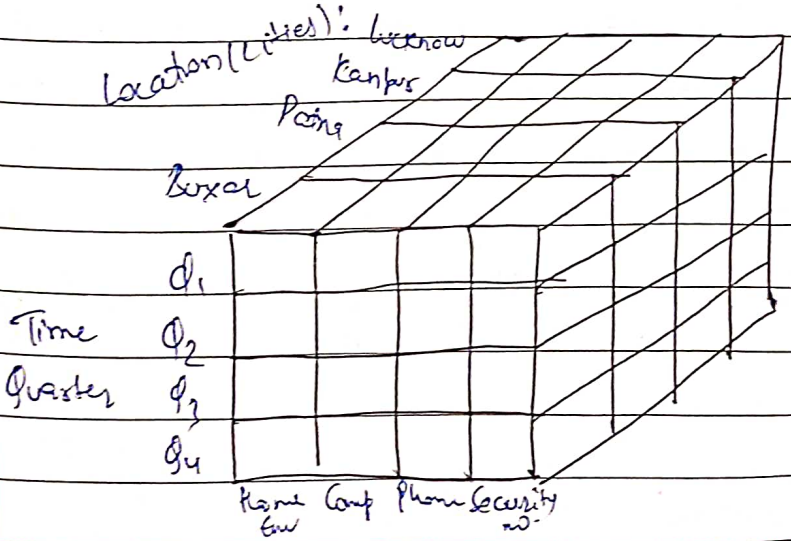
item (type)

Time	Home	Computer	Phone
Quarter	Entertainment		
Q_1			
Q_2			
Q_3			
Q_4			

Supplier = 'supp1'



Supplier = 'supp2'



Page No.:
Date : / /

If we continue in this way, so n-dimensional data can be viewed as a series of (n-1) dimensional queves.

8/02/23

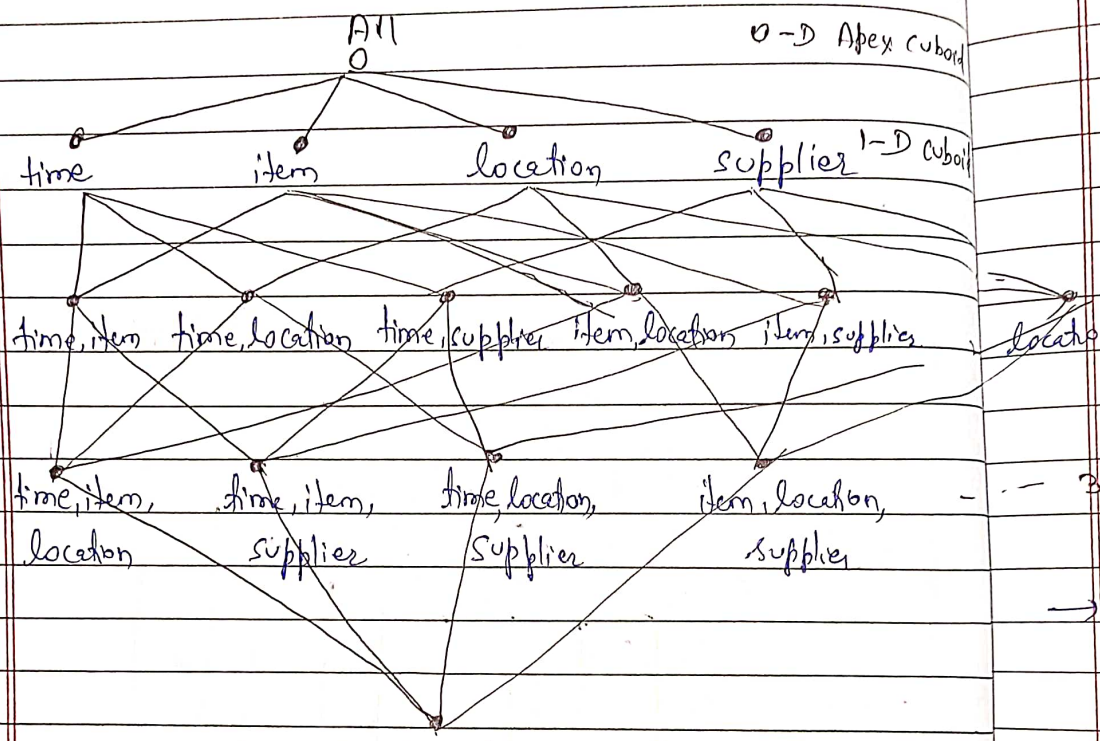


fig: lattice cuboid

→ Given a set of dimensions, we can generate a cuboid for each of the possible subsets of given dimensions. The result would form a lattice of cuboids, each showing a data at different level of summarization. The fig. shows the lattice of cuboids forming a data cube for the dimension time, item, location and supplier.

→ The cuboid t
summarization

→ The zero-dim
level of sum

Schemas for
most common
i) Star sche
ii) Snowflake
iii) Fact-cons

→ The E-R mod
relational da
consist of
between them.

$^4C_2$

$\frac{4 \times 3}{2}$

Page No.:
Date : / /

2-D cuboid

location, supplier

3-D cuboid

→ The cuboid that holds the lower level of summarization is called the base cuboid.

→ The zero-dimension cuboid which holds the highest level of summarization is called the apex cuboid.

Schemas for Multidimensional Databases :-

most commonly used

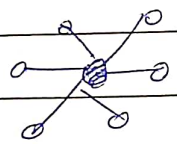
i) star schema
ii) snowflake "
iii) fact-constellation "

→ The E-R model is commonly used in the design of relational databases, where a database schema consist of a set of entities and the relationship between them.

→ The most common data models for data warehouse are:

i) star schema
ii) snowflake "
iii) fact-constellation "

Star-schema:



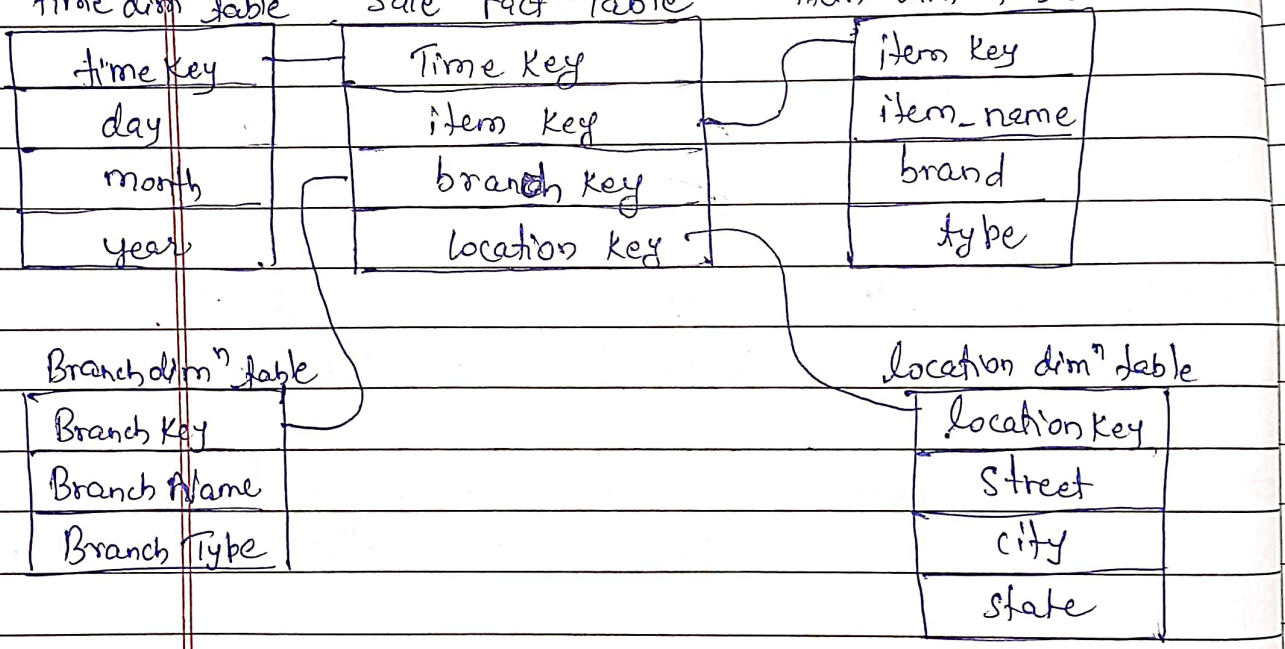


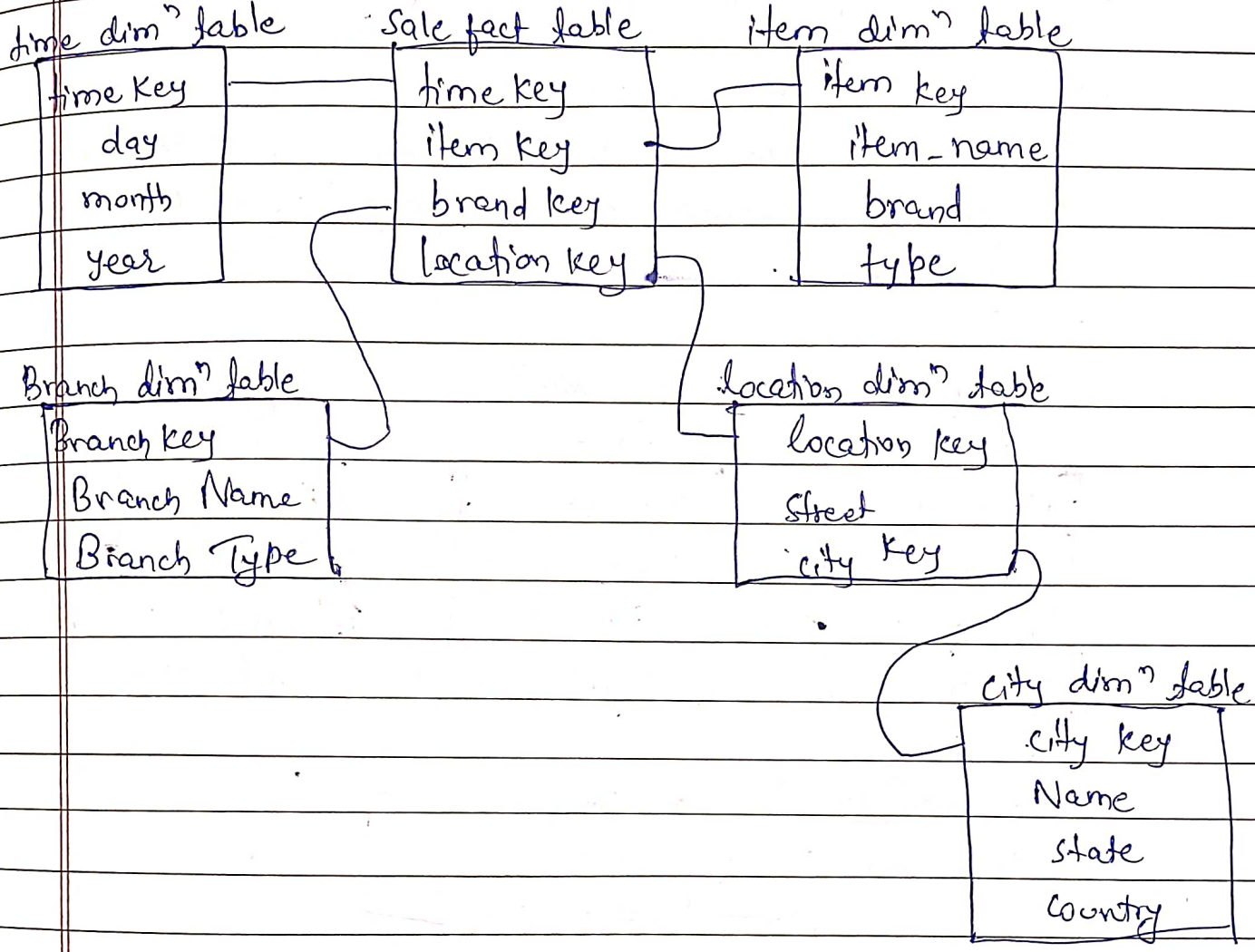
Fig: Star schema for sales

→ The most common modelling is star schema, in which data warehouse contains ~~the~~ m.

i) a large central fact table containing the bulk of data with usually no redundancy.
ii) a set of smaller attendent tables (dimension tables), one for each dimension.
iii) The schema graph resembles a star bus.

Page No.:
Date : / /

Snowflake schema :-



→ The snowflake schema is a variant of star schema where some dimension tables are normalized, thereby further splitting of data into additional table.

Page No.:
Date : / /

Fact - constellation

→ Sophisticated application may require multiple fact tables to share dimension table.

→ This kind of schema can be viewed as a collection of stars. At hence is called as galaxy schema or a fact - constellation schema.

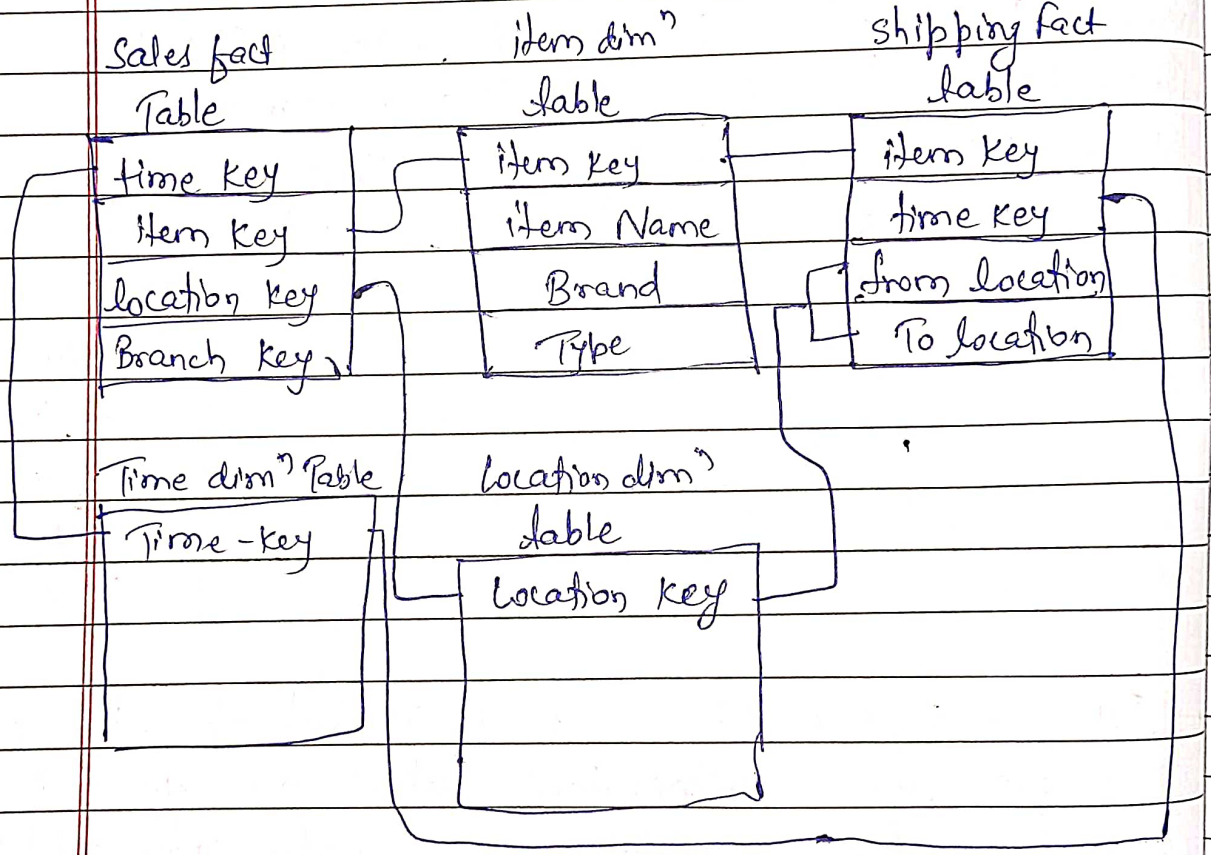


fig - Fact-constellation schema for sales & shipping

14/02/23

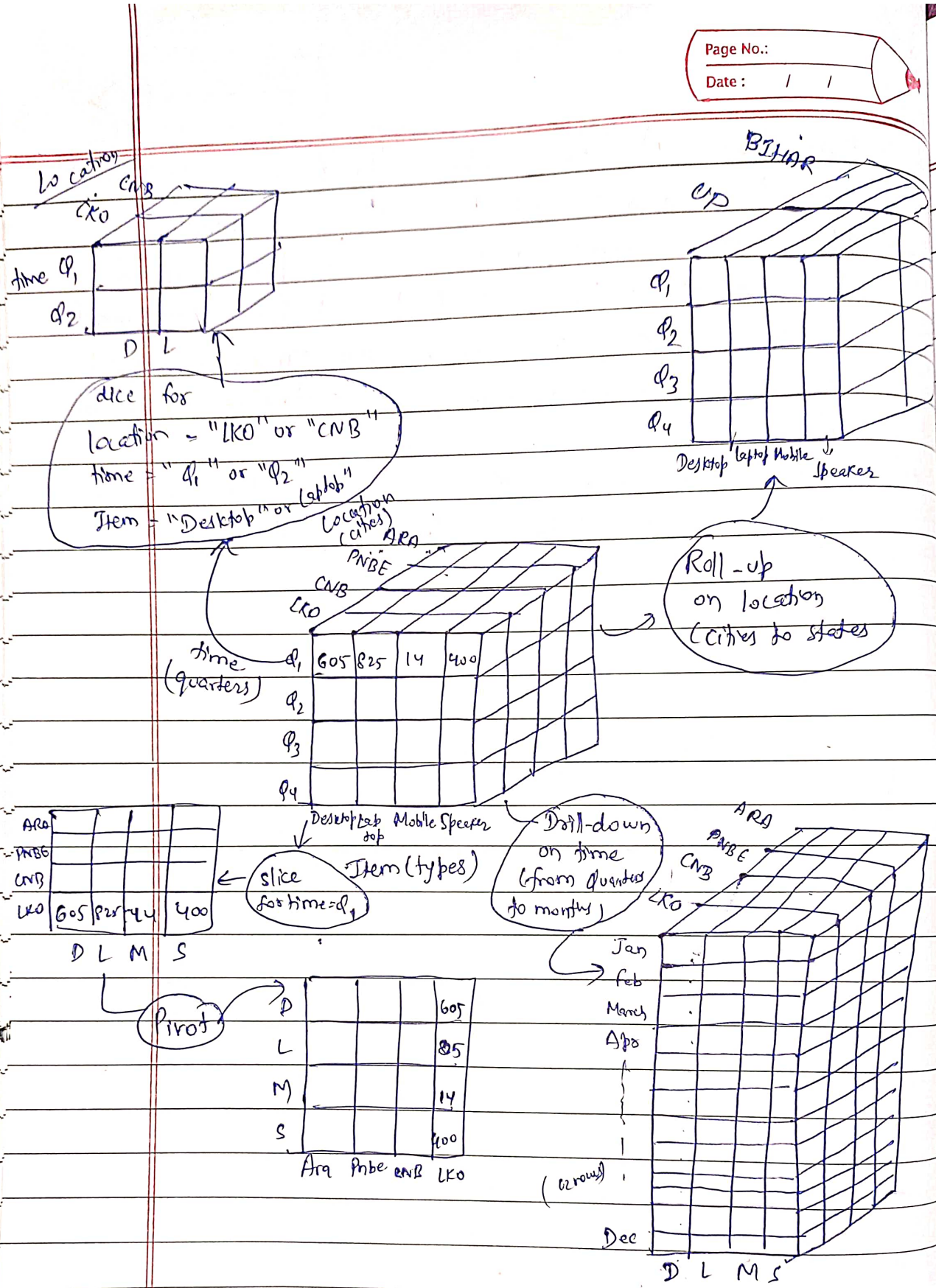
Page No.:
Date: / /

Data Cube - A data cube allows data to be modelled and viewed in multiple dimension.

→ In general terms, dimensions are the perspectives or entities with respect to which an organisation wants to keep records. Each dimension may have a table associated with it called a dimension table.

→ Fact table contains the name of the fact or keys of each related dimension tables.

OLAP Operation -


Page No.:
Date : / /
Location
CNB
LKO
time Q1
Q2
D L
slice for
location = "LKO" or "CNB"
time = "Q1" or "Q2"
Item = "Desktop" or "Laptop"
BIHAR
UP
Q1
Q2
Q3
Q4
Desktop Laptop Mobile Speaker
Roll-up
on location
(cities to states
location (cities)
ARA
PNBE
CNB
LKO
time (quarters)
Q1 605 825 14 400
Q2
Q3
Q4
Desktop Laptop Mobile Speaker
Item (types)
slice for time=Q1
Drill-down on time (from quarter to months)
ARA
PNBE
CNB
LKO
605 825 14 400
D L M S
Pivot
Jan
Feb
March
Apr
(12 rows)
Dec
D L M S
Ara Pnbe CNB LKO

Page No.:
Date : / /

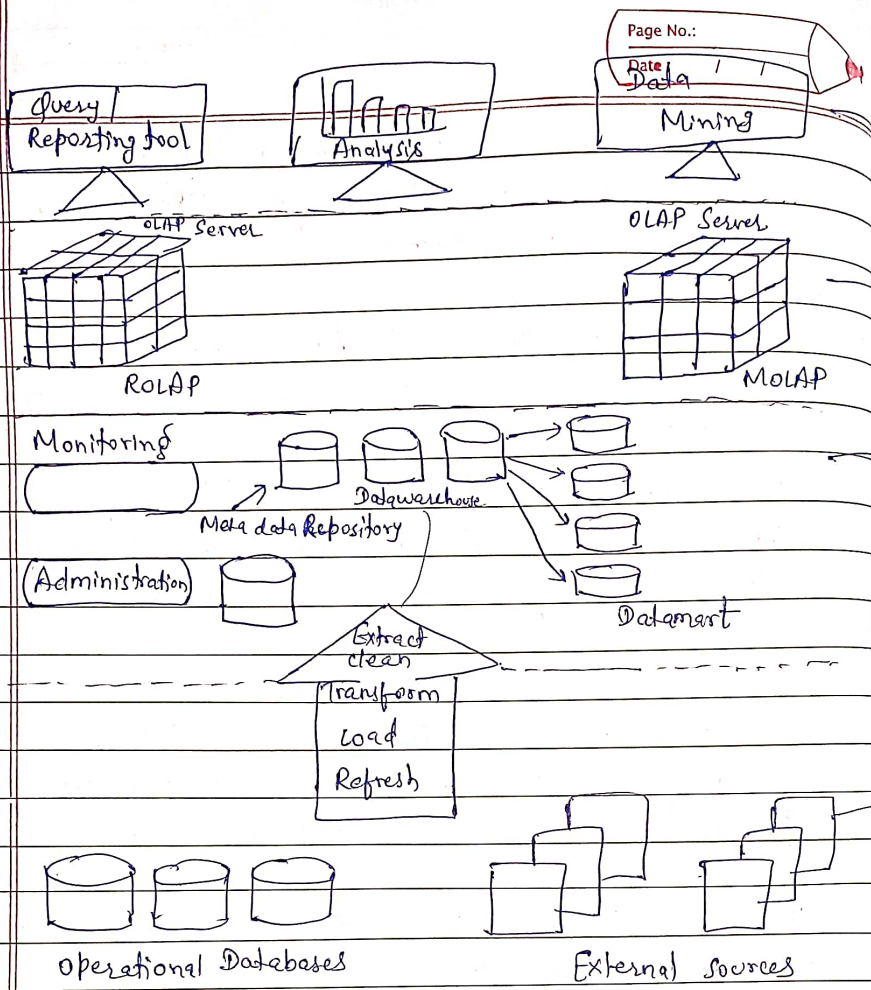
Roll-up → The roll-up operation performs the aggregation on a data cube either by climbing up or dimensionality reduction.

Drill-down → The drill-down is reverse of roll-up. It navigates from less-detailed data to more-detailed data.
→ Drill-down can be ~~realized~~ released either by stepping down a concept hierarchy for a dimension or introducing additional dimensions.

Slice and Dice- The slice operation performs a selection on one-dimension of the given cube resulting in a sub-cube. The dice operation performs a selection on two or more dimensions.

Pivot- It is a visualisation operation that rotates the data access in view in order to provide an alternate presentation of the data.

Datawarehouse Architecture:- (Three-Tier)



→ A three-tier datawarehouse structure:-

i) The bottom tier is the warehouse database server that is almost always a relational database system. Back end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources.

ii) The middle tier is a OLAP server ie typically implemented using either ROLAP model or MOLAP model.
ROLAP – It is a extended relational dbms that maps operations on multi-dimensional data.

→ Top Tier
front end Tools

→ Middle Tier
OLAP Servers

→ Bottom tier
Datawarehouse
server

→ Data

MOLAP – It is a speci
implements multidimen

iii) The top-tier is fro
contains querying a
tool, data mining t

Datamarts – A datam
corporate wide data
group of users. Its so
selected subjects.
→ Depending on the
can be categorised a
→ Independent datamart
captured from one or
external information

→ Top Tier
front end Tools

→ Middle Tier
OLAP Server

→ Bottom tier
Datawarehouse
server

→ Data

MOLAP - It is a special purpose server that directly implements multidimensional data and operations.

(1) The top-tier is frontend client layer which contains querying and reporting tools, analysis tool, data mining tools, etc.

Datamarts - A datamart contains a subset of a corporate wide data that is of value to a specific group of users. Its scope is confined to specific selected subjects.

→ Depending on the source of data, datamarts can be categorised as independent or dependent.

→ Independent datamarts are sourced from data captured from one or more operational systems or external information providers.

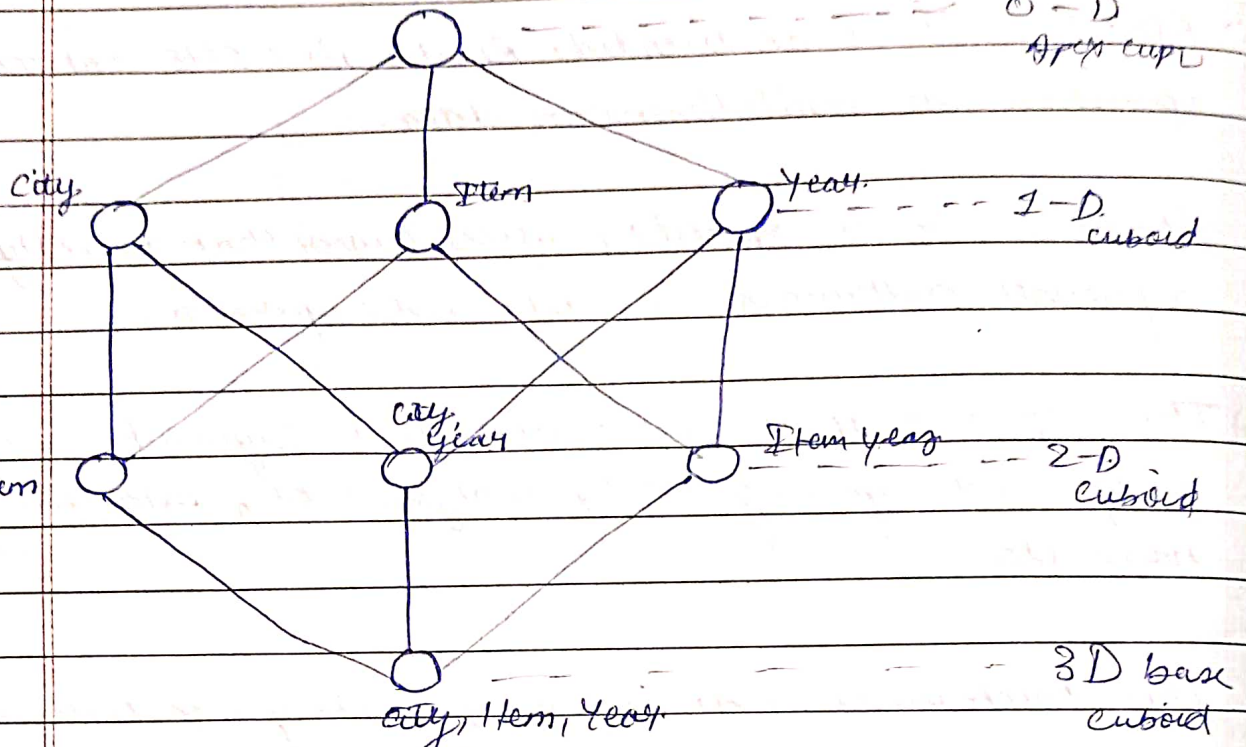
Page No.:
Date : / /

→ Dependent datamarts are sourced directly from enterprise data warehouse.

15/02/25

Date : / /
Page:

Curse of Dimensionality



Latex of cuboid

0-D ⇒ Total Sales

3-D ⇒ All dimension

→ The total number of cuboids or group by that can be computed for this data cube is 2^3. These group by forms a lattice of cuboid for the data cube

→ The base cuboid contains all 3 dimensions city, item & year. It can return the total sale for any combⁿ of these three dimⁿ.

→ The apex cuboid / 0-D cuboid refers to the case

Date : / /
Page:

where group by is empty. It contains the total sum of all sales.

→ This precomputation requires large amount of storage and it may explode when many of the dimension have associated concept heirarchies.

For n dimensional data cube:

Total no. of cuboids that can be generated if the dimen@ have hierarchies associated with it.

$$\text{Total no. of cuboid} = \prod_{i=1}^{n} (L_i + 1)$$

where L_i → no. of levels associated with dim@ i

i → no. of dimension

Q If the data cube has 10 dim@ and each dimension has 5 levels associated with it then find the total no. of cuboids generated.

$$= 4 \times 2 \times 5$$

$$= \frac{10!}{5!}$$

$$= 5^{10}$$

Date : __/__/__
Page: ____

Materialization of cuboids

There are 3 choices for datacube materialization given above cuboid:

→ No materialization
→ Full materialization
→ Partial Materialization

No Materialization: Do not pre compute any of the non-base cuboid but this may lead to extremely slow operation.

Full Materialization: Precompute all of the cuboids It typically requires huge amount of memory space in order to store all of the precomputed cuboids.

Partial Materialization:- Selectively compute a proper subset of the whole set of the possible cuboids. It considers 3 factors

1) Identify the subset of cuboids / subcubes to materialize.

2) Use the materialized cuboids during query processing

3) Efficiently update the materialized cuboid during load and refresh

Date : ___/___/___

Page: _____

Association Rule / Frequent Pattern / Market Basket Analysis