3/7/19

PAGE No.
DATE / /

Data - Raw and unprocessed facts.
Information - Processed and structured data.

Ex- The marks of all the students can be considered as data while the average calculated is information.

Data mining is important because not all data is relevant to us and can be discarded.

Extracting or mining knowledge from a large amount of data is called data mining.

Knowledge discovery in database (KDD)
→ It is a synonym of data mining.
→ Database consists of a collection of inter-related data.
→ Data warehouse is used to store large amount of data.
⇒ It is a repository of information collected from multiple sources.
Datamart - subset of datawarehouse used at dept. le

Q. Difference between database and data warehou

KDD is a step-by-step process:
i) Cleaning - Removing inconsistent and noisy data.
Table A: AB → C Table B: AB → E
inconsistent

- Missing value : Solution:
  (1) Take average of data.
  (2) Manually insert data (ineffective)
  (3) Use global variable
      [Ex- unknown or $-\infty$]
  (4) Ignore the tuple (may lead to loosing of important info)

ii) Integration - Combining data from multiple sources.

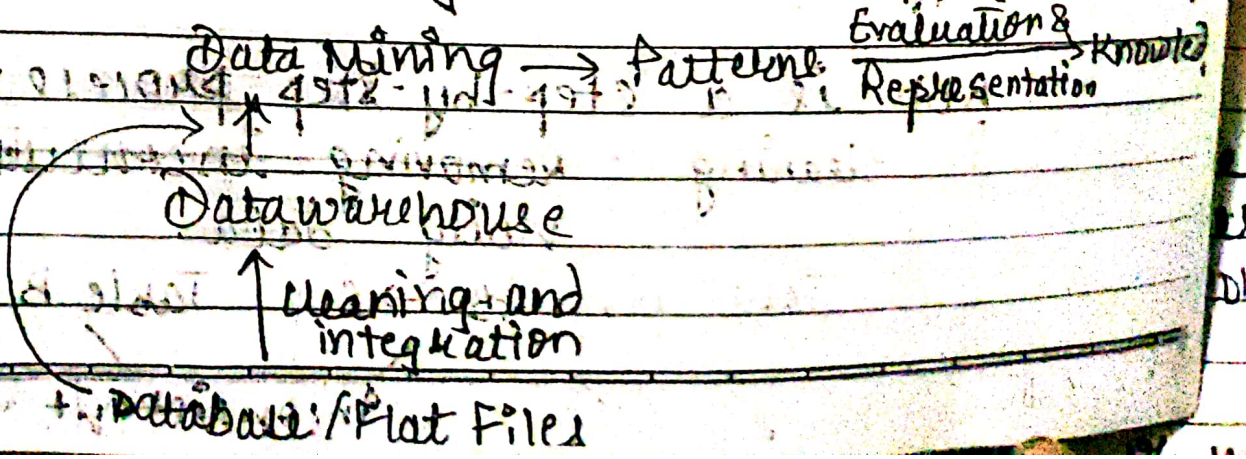
iii) Data Selection - where data relevant to the analysis task are retrieved from the database.

iv) Data Transformation - where data are transformed into appropriate forms for mining (Representing available data according to need)

v) Data mining

vi) Pattern analysis

vii) Knowledge presentation - Displaying items based on the pattern recognized in the previous step



PAGE No.
DATE / /

20/7/19

Datamining components :

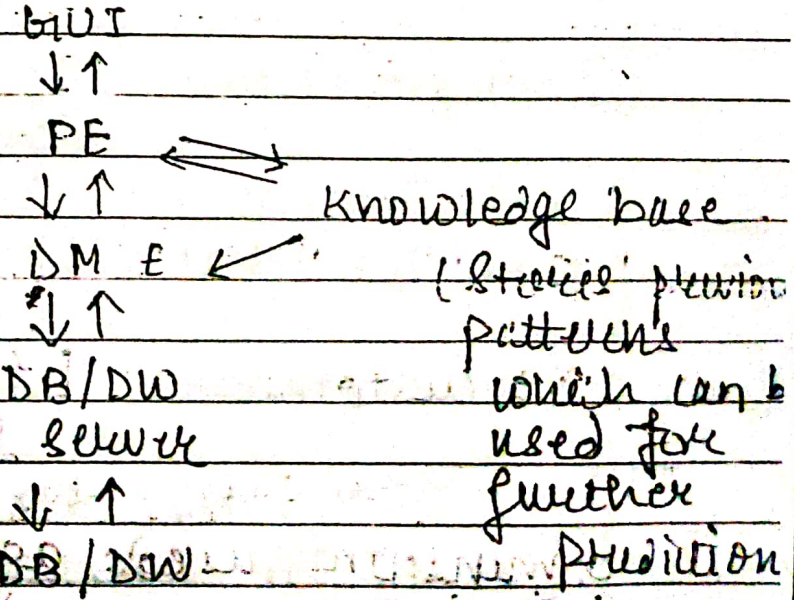
i) DB / DW (Database or Datawarehouse)
Data can be mined from either of these.

ii) DB / DW Server
When user fires a query, this server is responsible for fetching the data from DB or DW

iii) Datamining Engine
It determines the technique by which data is to be mined.

iv) Pattern Evaluation
Ex- To find the probability of an user buying certain items after buying an item.

v) GUI



Datamining is of two kind of data :
i) Persistent (DB) ii) Datawarehouse

iii) Transactional DB.
iv) Advanced DB system. (Object Oriented DB)
v) Advanced DB applications. (Spatial DB) stores info. relating to geographical resources.

vi) Temporal and time-series database (ex- Transactions in a bank/ stock) To store data which changes ~~based~~ with respect to time.

vii) Multi-media DB / Text based DB
viii) World wide Web database. (Search history)

Data Mining functionalities:

i) Association:

age(X, "20...29") ^ income(X, "20k...30k") ⇒ buys(X, "CD Player")

Interpretation - If a customer within the range of 20 to 29 years and income of 20k to 30k, it is probable that he/she will buy a CD Player

X → customer [confidence = 50%
support = 2%]

dimensions used: age, income, buys (attributes)

∴ Multi-dimensional association rule

50% means that CD Player is bought

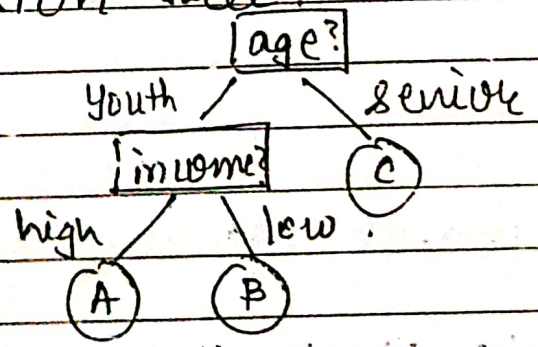
2% entries in the database which abide by this rule].

⇒ Contains (T, "computer") ⇒ contains (T, "S/w [confidence = 60%. S = 1%]
Single dimensional association rule.

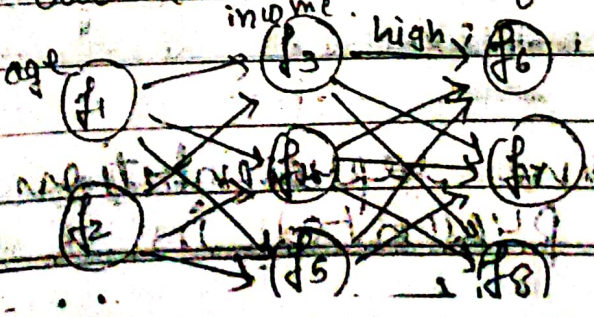
If transaction T contains comp, then is a 60% chance that it contains software.

ii) Classification and Prediction
- IF-THEN rules
Ex - age (X, "youth") ∧ income (X, "high") → class (X, "A").
If a person can be classified under age group of 'youth' and has a high income, then the response is put under class A.

- Decision Tree
Ex:



iii) Neural network



iii) Cluster Analysis (involves finding similarities & then grouping them)
Intra-class similarity higher.
Inter-class similarity lower.

Similar to classification but, classification has certain pre-defined labels. Class is which are not available for clustering.

iv) Outlier Analysis
Used to find certain anomaly in the pattern Ex- fraud.

v) Evolution Analysis
Useful for analyzing data which keeps changing with respect to time (undergoes continuous evolution)

12/7/19 Parameters to check whether pattern is interesting
i) understandable
ii) Valid
iii)
iv) Novel

Issues of DM
i) Based on human interaction and data mining techniques
It is possible that the results differ when different visualization and classification techniques are used.

ii) Visualization and representation
Data can be presented in a graphical

format or tabular format.
Ex- Most sold item - table more helpful
Profit/loss - Graph more helpful

iii) Pattern detection (evaluation)
Threshold value needs to be correctly determined for detecting outliers and minimize the variation to evaluate a pattern.

iv) Scalability and efficiency of DM techniq
The size of the dataset shouldn't has an effect on the results.
The algorithm should produce correct results in optimal time. (efficient).

v) Parallel processing & distributed env.
Given the vastness of the dataset, the DM technique should be able to perfo parallel processing in a distributed env.

vi) Incremental approach.
When new data flows in, the DM technique should be able to pick up the process where it left or paused. Should start from scratch.

(vii) Variety of database
Ex- DM tech should be global and be able to mine data based on the parameters of a particular database Ex- Co-ordinates for long lat dataset

geographical data and pixel info for images, etc

17/2/19 characteristics of Data warehouse

i) Subject - Oriented
It is not transaction oriented. Rather, it is based on the specific topic for which we intend to find patterns

ii) Integrated
It ensures that the naming / dimensions, structural conventions are same (or taken care of) in all the databases used to create the warehouse.

iii) Time variant
Databases may have a wide range of time. While they are combined to form a warehouse, a specific range is selected.

iv) Non-volatile
It deals with historic data, mostly and thus, does not need any concurrency or reconciling measures.

OLAP & OLTP:
OLAP - Online Analytical Processing - DW
OLTP - Online transaction processing - DB

OLTP	OLAP
Operational	Informational

	Transaction	Analysis
Orientation:		
Users :	DBA / Clerk / DB Professional	Knowledge worker / analysts
Storage:	Day-to-day	Long-term informational req.
Design:	ER-based designing schemes	Star or Snow flod Schema
Data :	up-to-date (current)	Historical
Dimensionality:	2-D	Multi-Dimensional
Operations:	Read / Write	Read
No. of users:	In terms of thousands	In terms of hundreds
Memory size:	100 MB to GB	100 GB to TB

⇒ When the requirements are clear, top-down approach can be used to design the structure of data warehouse. (waterfall model)
Bottom-up approach is preferred when the req. are unclear or may change. (Iterative model)

Q. Advantage & disadvantage of top-down and bottom-up approach.

3-tier architecture of DW. (diagram in TB)
External data source is not a 'tier' in the model but it served as a base over which the entire architecture depends.

Meta-data acts as an index for the actual data stored in DW.

Virtual Warehouse
When we represent the information residing in a DW, it can be said as a view which is termed as virtual warehouse.

Types of OLAP Server
i) Relational OLAP (ROLAP)
ii) Multi-dimensional OLAP (MOLAP)
iii) Hybrid OLAP (HOLAP)

Q. Difference between ROLAP, MOLAP, HOLAP

ROLAP: Has large memory but, limited speed. Data stored in RDBMS.

HOLAP = (ROLAP + MOLAP)

MOLAP: Stores summarized data and has comparitively higher speed than ROLAP

19/7/19

location "Chicago"

item	time	home entertainment	computer	phone	security
Q_1	854	882	89	623
Q_2	943	890	64	698
Q_3	1032	924	59	789
Q_4	1129	992	63	870

location "New York"

item

time	HE	Comp.	Ph.	Sec
Q_1	1087	968	38	872
Q_2	1130	1024	41	925
Q_3	1034	1048	45	1002
Q_4	1142	1091	54	989

Location "Toronto"

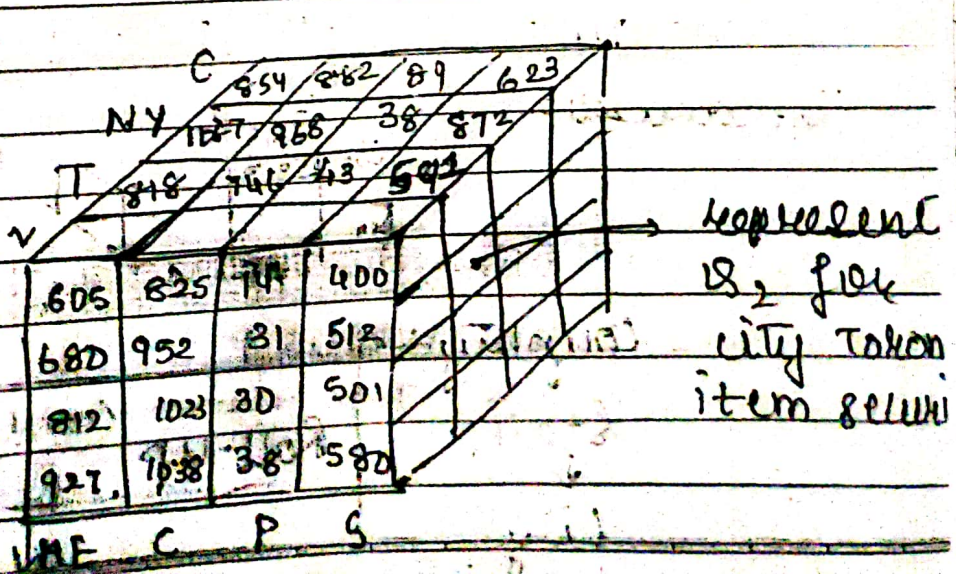
item

time	HE	Comp.	Ph.	Sec.
Q_1	818	746	43	591
Q_2				682
Q_3				728
Q_4				784

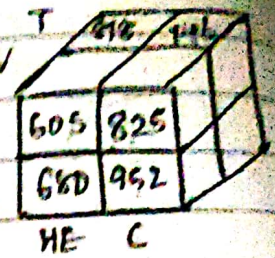
location "Vancouver"

item

time	HE	Comp.	Ph	Sec.
Q_1	605	825	14	400
Q_2				512
Q_3				501
Q_4				580



Dice:
location = "V" or "T"
time = "Q_1" or "Q_2"
item = "HE" or "COMP"



Slice: Perform select operation for single attribute.

Ex: time = "Q_1"

	C	P	S	
C	854	882	89	623
NY	1087	968	38	872
T	818	746	43	591
V	605	825	14	400

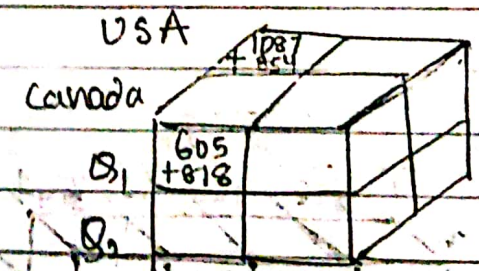
HE

Pivot: Rotating the axie.

HE	605	818	1087	854
C				
P				
S				

V T NY C

Roll up: (Generalization)



country
↓
state
↓
city

Roll-up

Drill-down

PAGE No.
DATE / /

Drill down: Q_1 : Jan
(Specification) Feb
Mar
Apr.

Schemas for Multi-dimensional DB:

i) Star schema (faster than snowflake)
(generally used to obtain facts)
Terms:
- Fact Table (larger)
The analysis carried out is stored.
- Dimension Table.
Attributes used for analysis are stored.

Ex: Dimension Table
{item, time, branch, location}

Time
(Dimension Table)
- time key
- day
- day of the week
- month
- quarter
- year

Branch
(DT)
- b_key
- b_name
- b_type

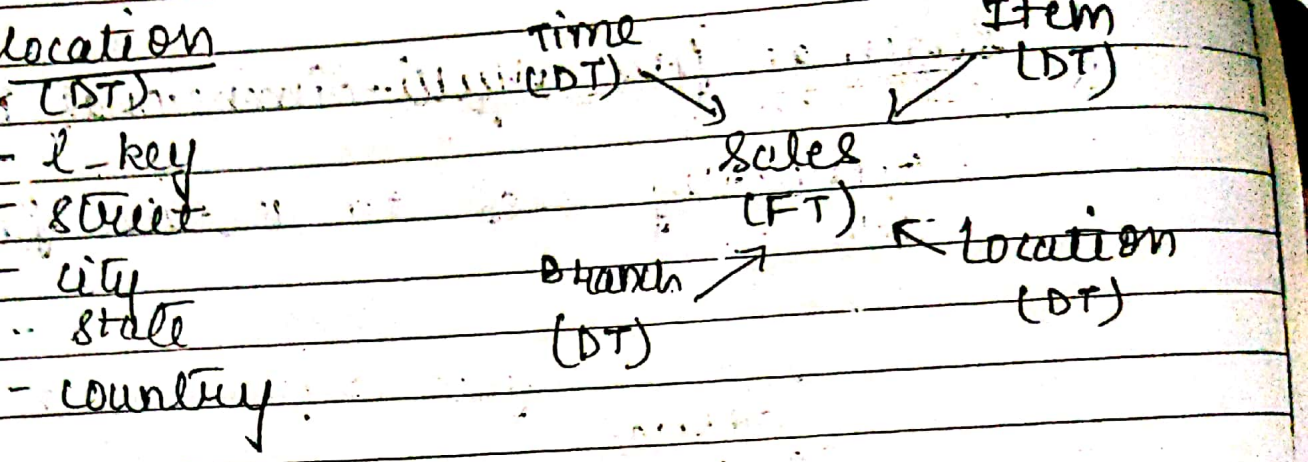
Item
(DT)
- i_key
- i_name
- brand
- type
- Supplier ty

Sales (Every primary key sto
(Fact Table) as foreign key)
- time_key
- b_key
- i_key

DATE / /

- dollar_sold
- sold_unit

} generated term with
Stores the analysis
of the entire DW



24/7/19

Dimension Table

Dimension Table — fact Table — Dimension Table

Dimension Table

D.T are not normalized in star schema
Redundancy present

To remove this redundancy, we use the snow-flake schema. It has more than one fact table

⇒

Course

Student — Student Enroll — Time

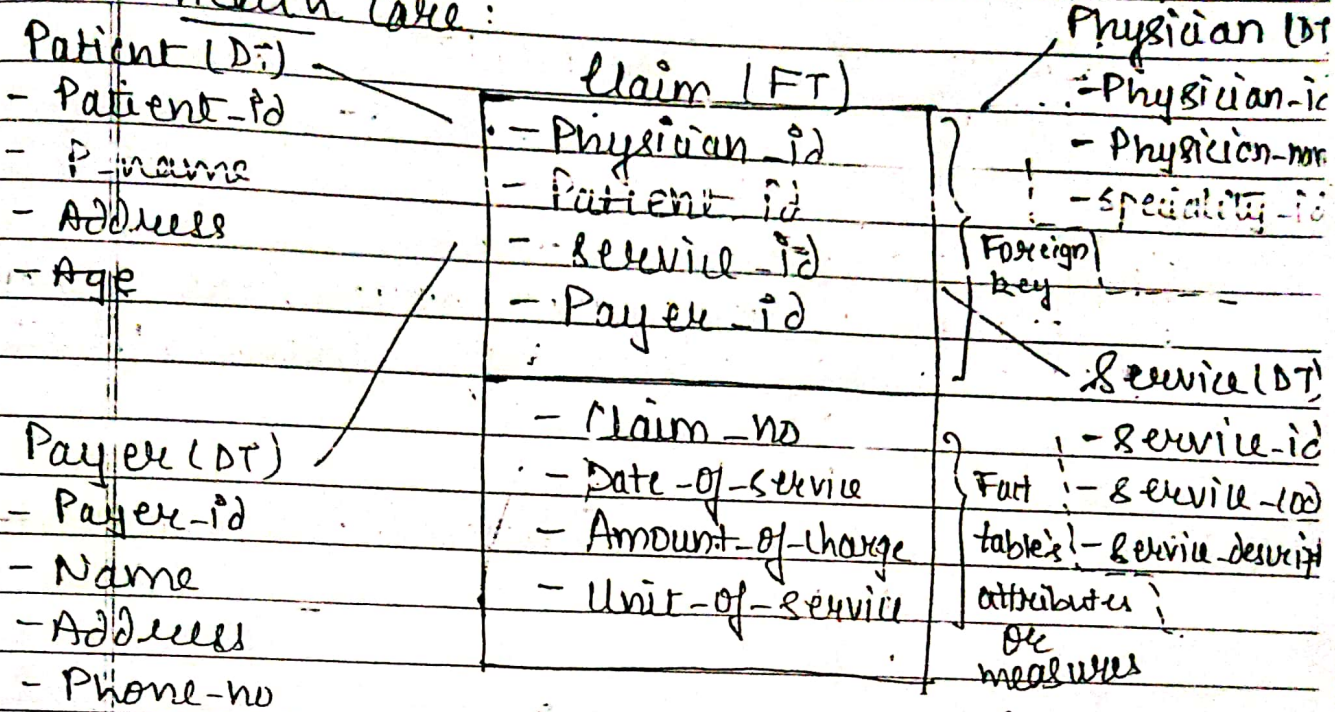
data

Course Sections

We need to find how many students

are enrolled in a particular course.

Health care:



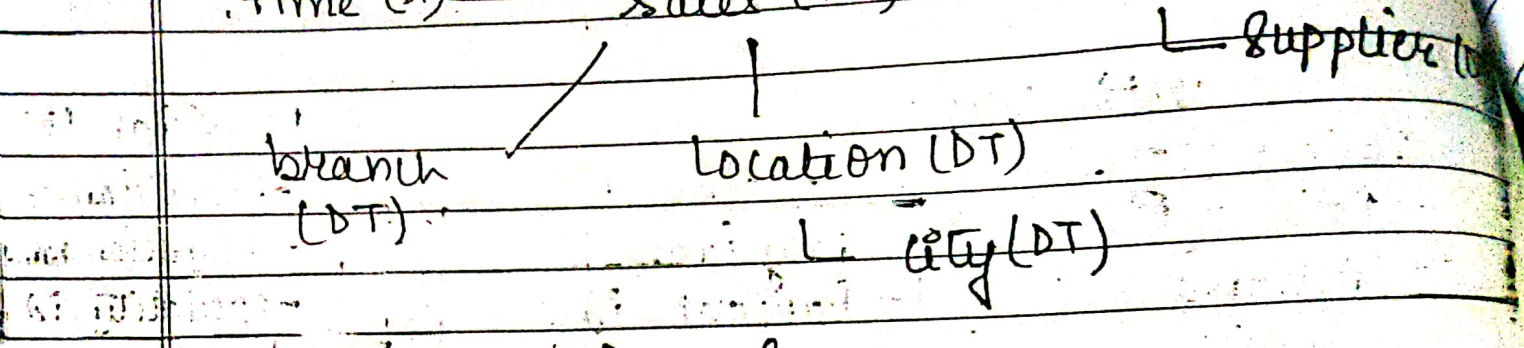
Disadvantage of Star Schema:

i) Data integrity
Tables are denormalized. Hence, redundant data persist

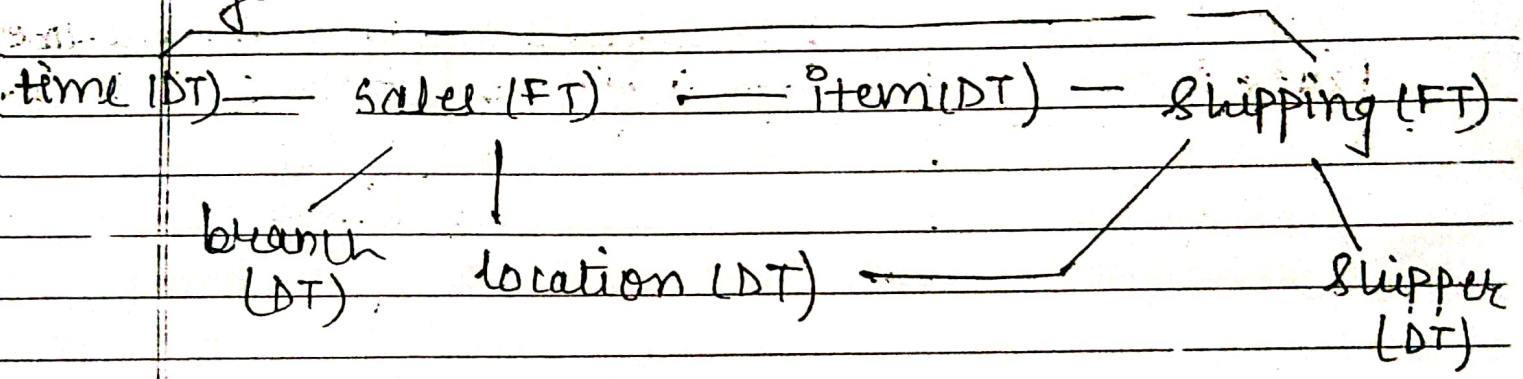
ii) Not flexible in easy way
We cannot determine the attribute to be considered in case of multi-valued attribute.

Snowflake Schema

- More than one fact table possible
- Every dimension table is normalized
- More no. of joins need to be performed
- Takes more query response time
- More reliable.

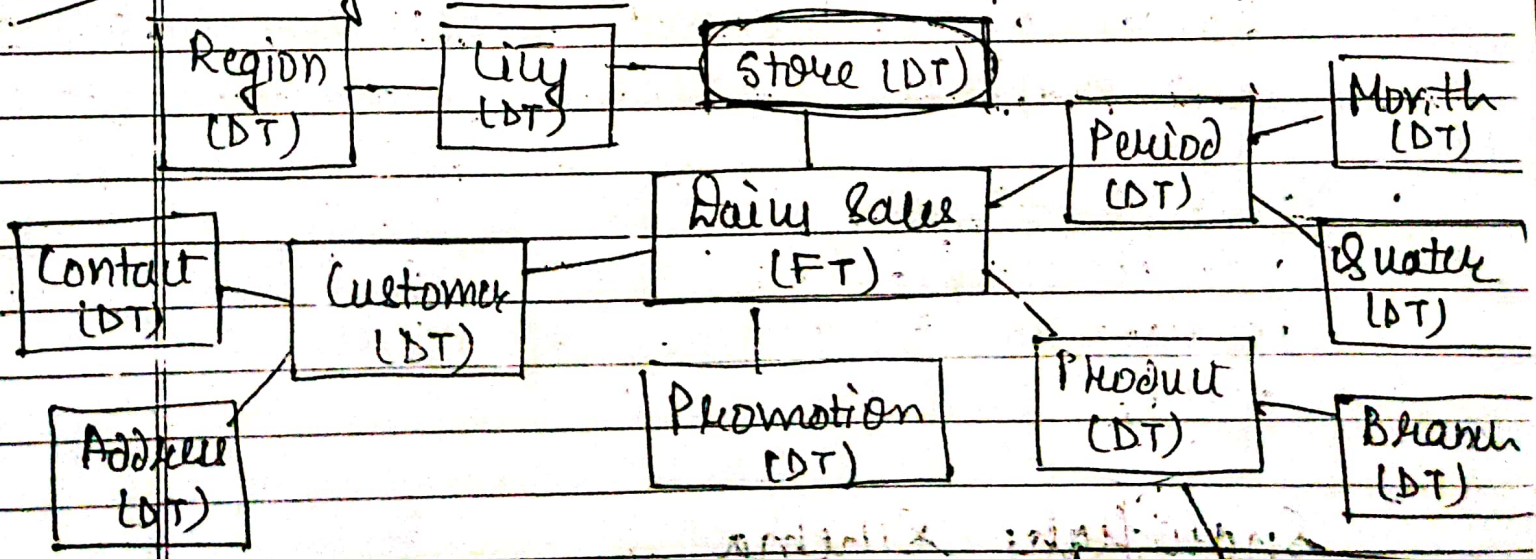


Fact constellation schema:
- collection of two or more star schema
- Tables may exist in denormalized form.



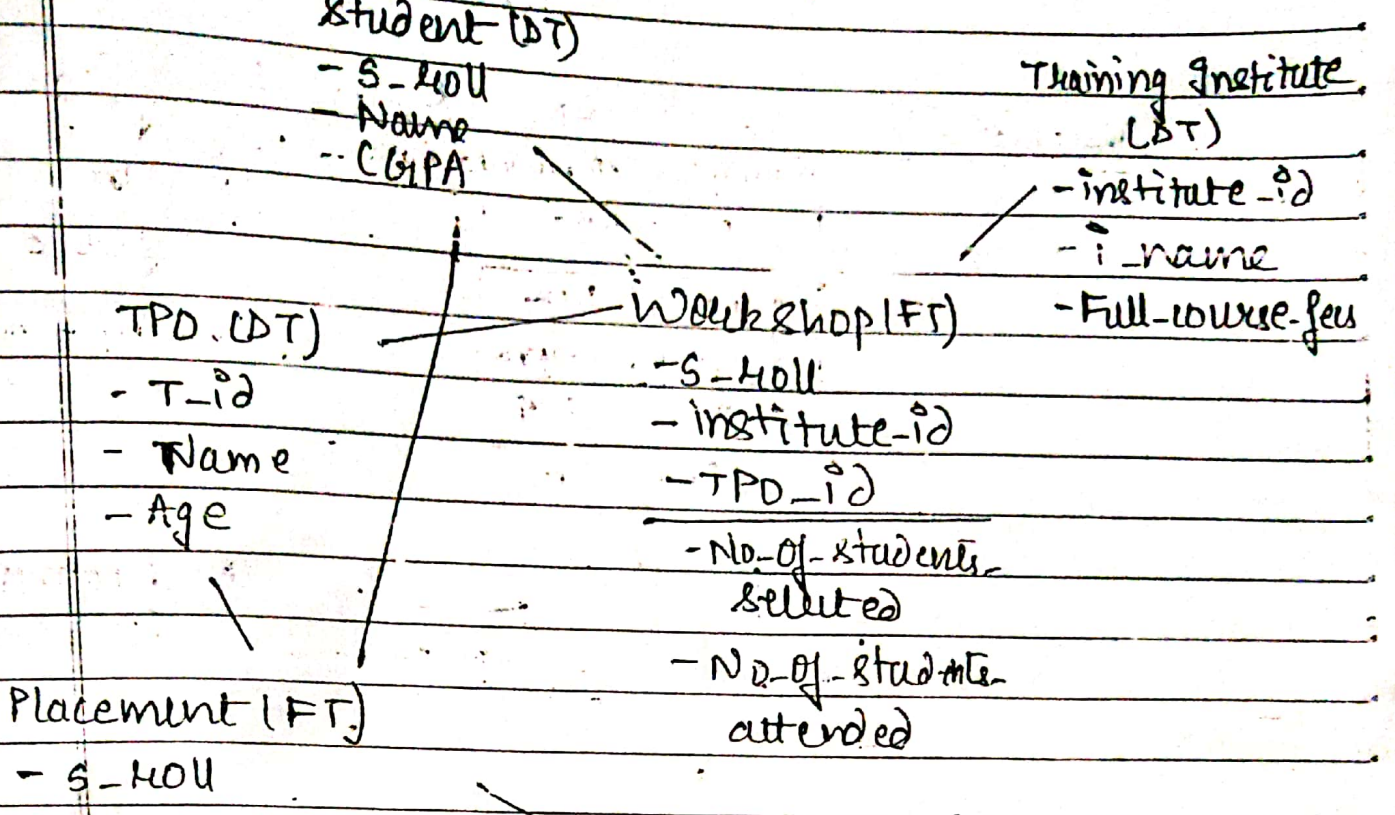
26/7/19

Snowflake schema example:



PAGE No.
DATE / /

Fact Constellation example:



From OLAP to OLAM.

OLAP: just summarizing top. over available data. Not extracting any new info.)

OLAM: finding out patterns and predicting further results.

Applications of Data warehouse:

i) Info. processing
ii) Analytical processing (Slicing / Dicing)

iii) Data Mining (Classification / Clustering / Predicting)

Architecture of OLAM (Online Analytical Mining)