02/09/23 Monday

Module-1

Data Mining

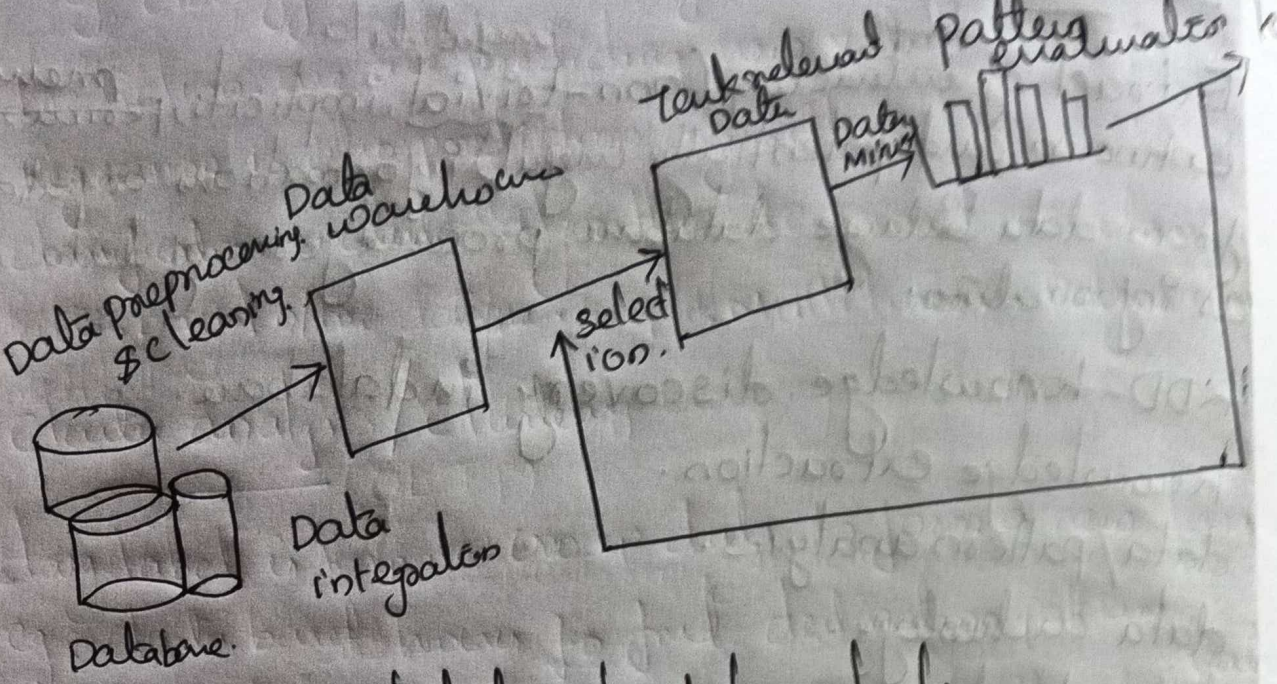
Extract of interesting (non-trivial, implicit, previously unknown and potentially useful) informations or patterns from data in large databases. processed form of data is known as information.

KDD - knowledge discovery in database.
knowledge extraction.
data/pattern analysis.
data archeology.
data dredging.
information harvesting.
business intelligence.
Web data mining.
(Deductive) query processing.
KDD is also known as data mining.

Steps in KDD

1, Data preprocessing and cleaning.
2, Data selection and Transformation.
3, Data Mining
4, Pattern evaluvation.

Data mining: the core of knowledge recovery process.



creating a target data set : data selection.

Data cleaning and preprocessing : removing unwanted data.

2) Find useful features dimensionality.

choosing functions of data mining

Summarization, classification, regression, association, clustering

choosing the mining algorithm.

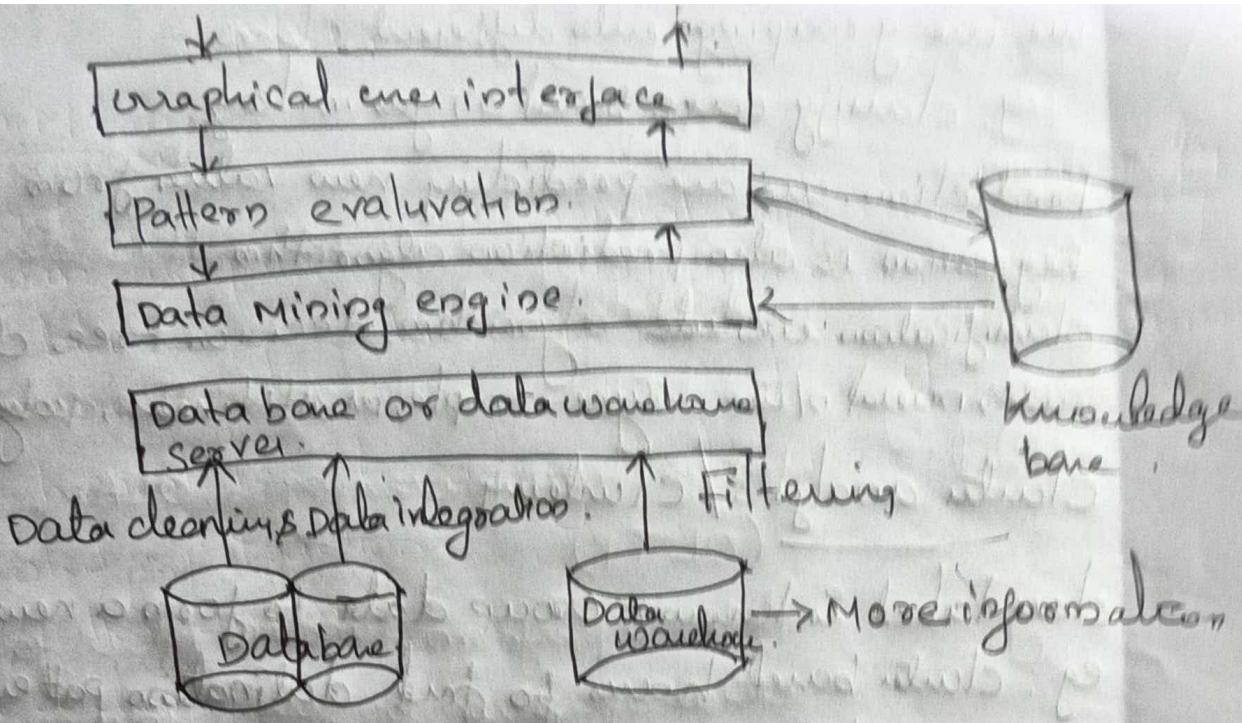
Data mining : search for patterns.

Pattern evaluation & knowledge

visualization, transformation, removing redundancy patterns.

use of discovered knowledge.

Architecture a Typical Data mining System.



6 Major Components

1. Database, data warehouse, WWW, or other information repository
2. Database or data warehouse server.
3. knowledge base.
4. Data Mining engine
5. pattern Evaluation module
6. user interface.

11/2/22
Tuesday

Data mining functionalities

- Concept description: characterization and discrimination
  - Generalize summarize and Constant data.
  - characteristic eg: dry vs wet regions.
- Association (correlation and causality)
  - Multidimensional vs single dimensional association.
- classification and prediction.
  - Finding models that distinguish class or concept for future Prediction

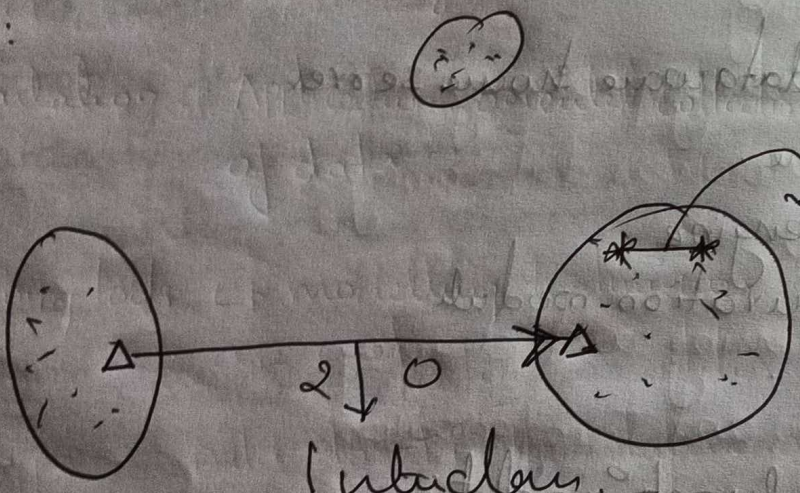
we are grouping into different classes.
eg: classify cars based on gas mileage.

Prediction: we are predicting new values from previous values
Regression is also prediction algorithm.
classification include algorithm, decision tree, classification rule, neural n/w. Type - simple, Multi linear, polynomial - curve

Cluster analysis / clustering

class label is known: group data to form a new classes
eg: cluster based hours to find destination pattern.
cluster based on the principle maximize ~~b/w the class~~ value ↑ the intra
similarity and minimizing value ↓ the interclass similiarity.

eg:



Similiarity distance measure
Inter class similiarity - b/w the class.
Intra " " - within the class.

Outlier analysis (noise)

outlier: A data object that doesnot couply with general behaviour of the data.
outlier Analysis: It can be considered as noise or exception but it is useful in fraud detection and even analysis

Trend and evaluvation analysis

Trend and deviation: regression analysis similiarity based analysis

Seaquential pattern mining, periodicity analysis

other pattern: directed or statistical analysis

~~it~~. Major issues in data mining

- Mining methodology and user interaction
  Mining different kinds of knowledge in database.
- performance and scalability.
- Issues relating to the diversity of the datatypes.
- Issues related to application and social impacts.

Human interaction.

overfitting

changing data.

9/2/23 Thursday Data warehousing

A datawarehouse is subject-oriented, integrated, time-variant non-volative collection of data in support of management's decision making proces.

Decision support database.

Datawarehousing: The proces of conducting and using data warehouse.

Time variant: entirely different from database.

non-volatile: operational update of data doesnot occur in the data ware house environment.

Initial loading of data access of data

Database V/s Datawarehouse.

Feature	Database	Datawarehouse.
Purpose	It is designed to record data	It is designed to analyze
Processing method.	The database uses the online transaction processing (OLTP)	The datawarehouse uses OLAP
Usage	Helps to perform fundamental operations for your business.	Allows you to analyze your business.
Tables & joins	complex as they are normalized	simple in because they are denormalized.
orientation	Application-oriented collection of data	A subject-oriented collection of data.
Designing tools	ER modelling techniques are used.	Data modelling techniques are used.
Data type	Data stored in the database is up to date.	Current and historical Data is stored in datawarehouse. May not be up to date.
query type	simple hierarchical queries are used.	complex queries are used
Data summary.	Detailed data stored in the database.	It stores highly summarized data

OLTP - Online Transaction processing

	OLTP	OLAP
ues	clerk, IT professional	knowledge workes.
function	day-to-day operation.	decision support
DB design.	application oriented.	subject oriented.
Data	current up to date detailed, flat relational isolated	historical, summarized multidimensional, integrated consolidated
usage	repeletive	ad-hoc
acces	read/write/index/hash on primary key.	lots of scans.
unit of work	short, simple transaction	complex query.
#records of accessed	tens	millions.
#users	1000s	100s.
DB size	100MB-GB	100GB-TB.
metric	transaction throughput	query, throughput response

3/2/23 Monday

A multidimensional data model

A data warehouse is based on a multidimensional which view data in the form of a data cube. Data is represented in the form of data.

location="chicago" loc="Newyork" loc="Toronto" loc="vancou

item item item item

home	home	home	home
time ent. comp. phone sec.	ent. comp. phone sec	ent. comp. phone sec	ent. comp. Phone s
Q1			
Q2			
Q3			
Q4			

Data Cube

Data cube allows data to be modelled and view in multiple dimen.

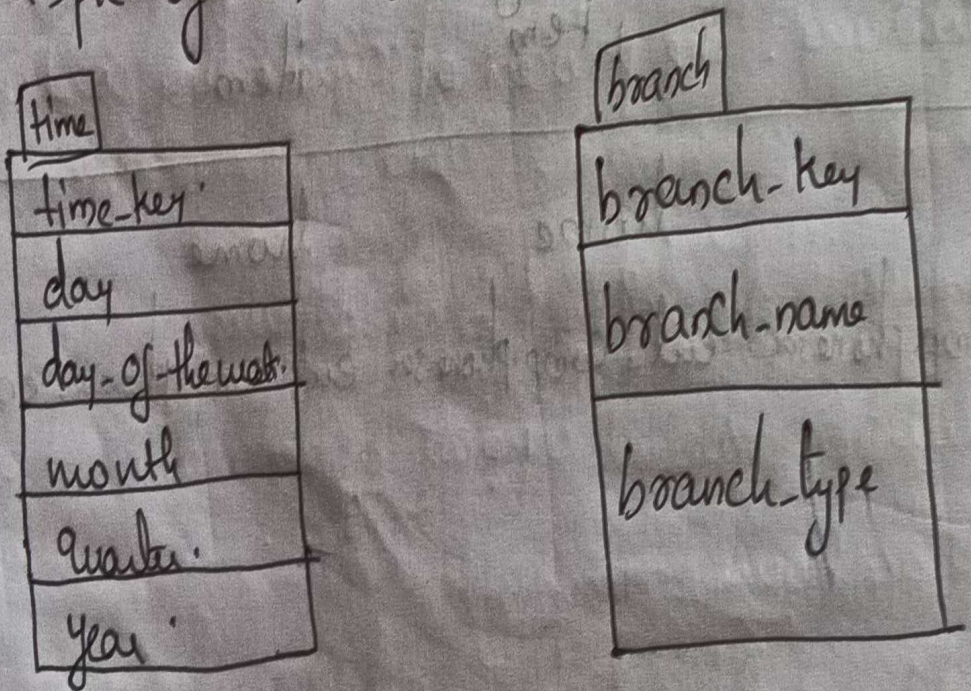
It can be defined by dimensions and facts.

Dimension

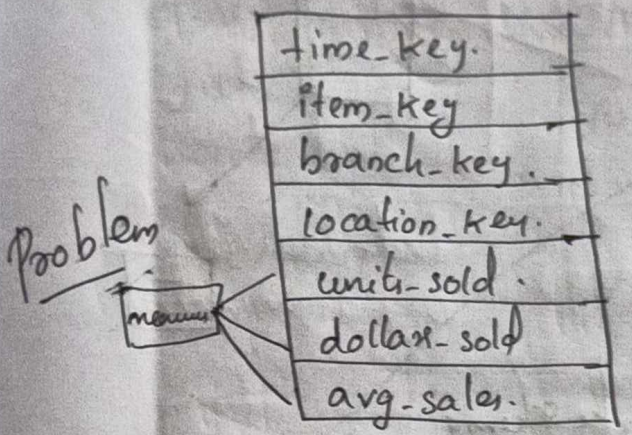
Dimensions are the perspectives or entities with which is table, of an entity.

fact - numerical tables, relationship b/w dimensions tables.

Example of dimension table:



eg of fact table

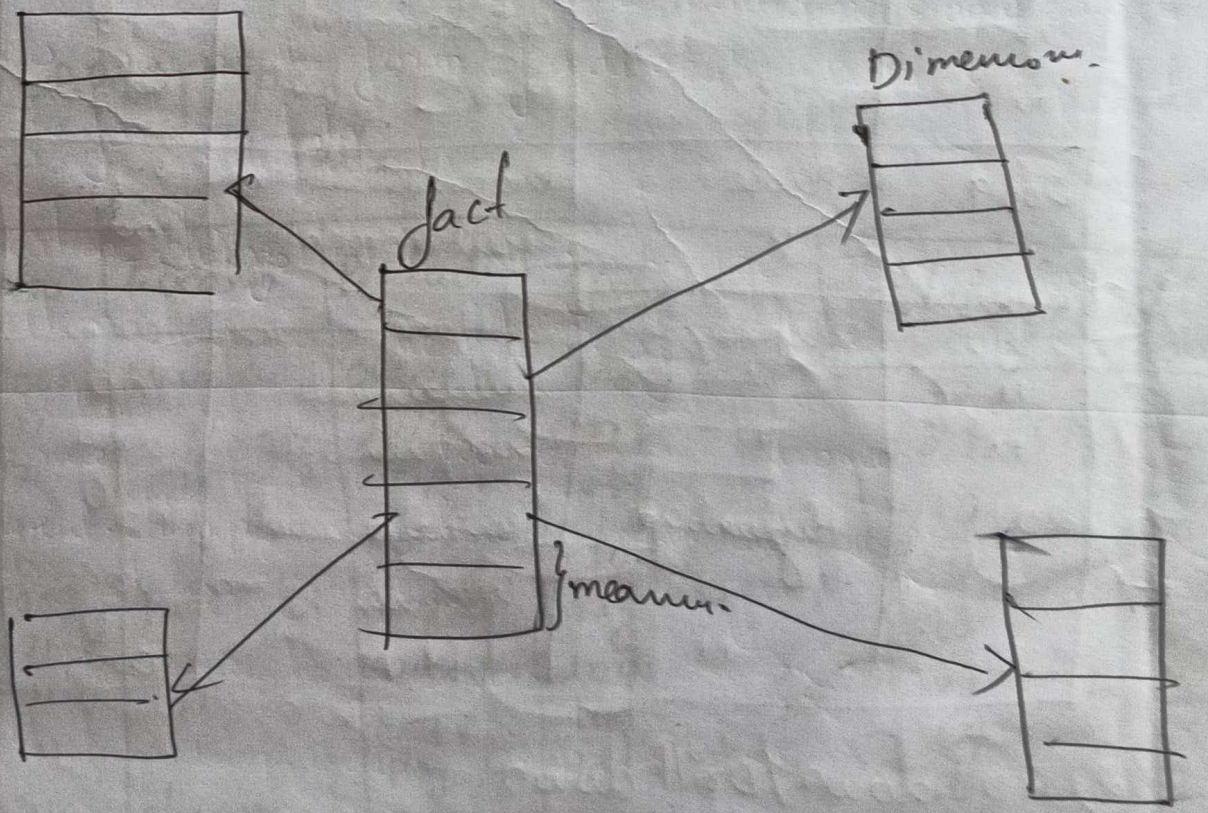


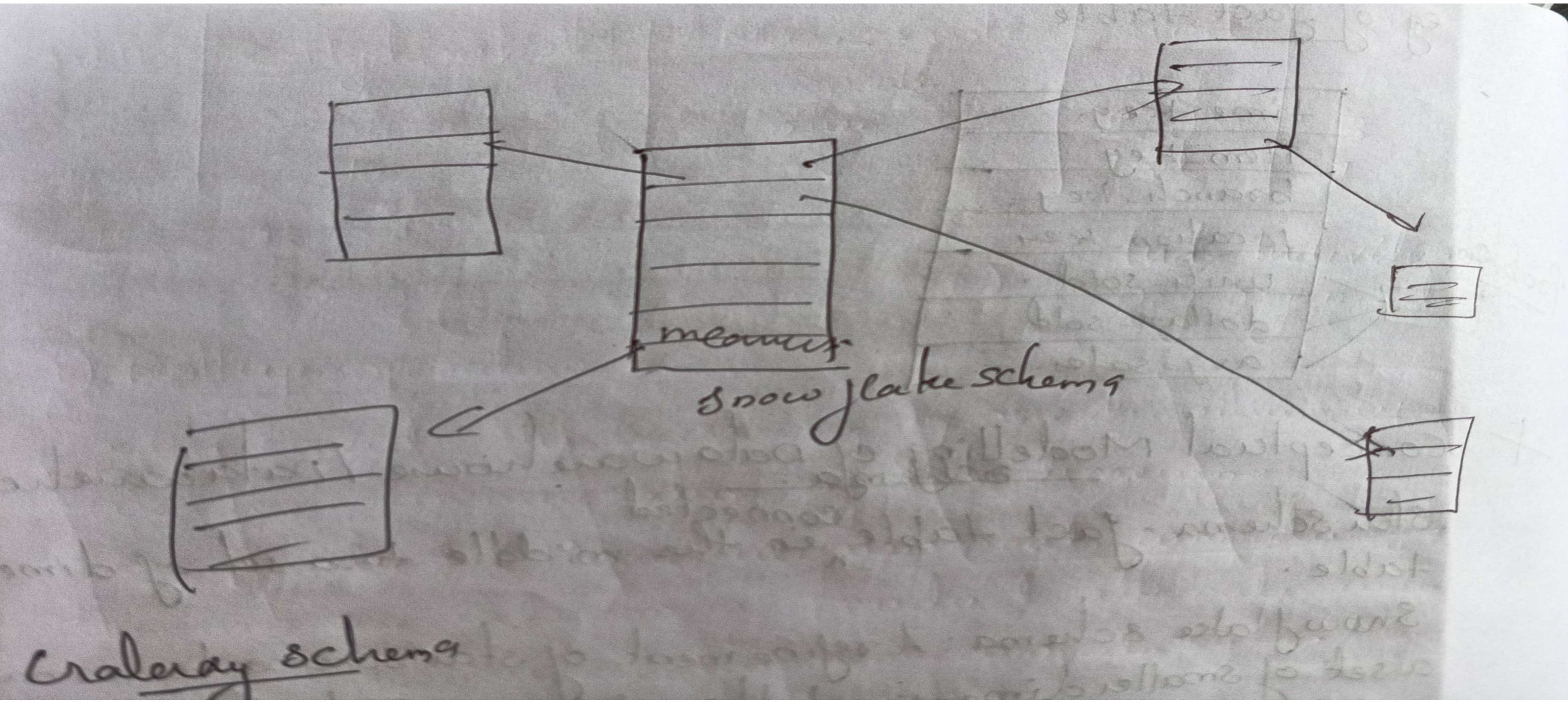
:+ Conceptual Modelling of Data warehouse / Data warehouse schema.

star schema - fact table connected to the middle to a set of dimension table.

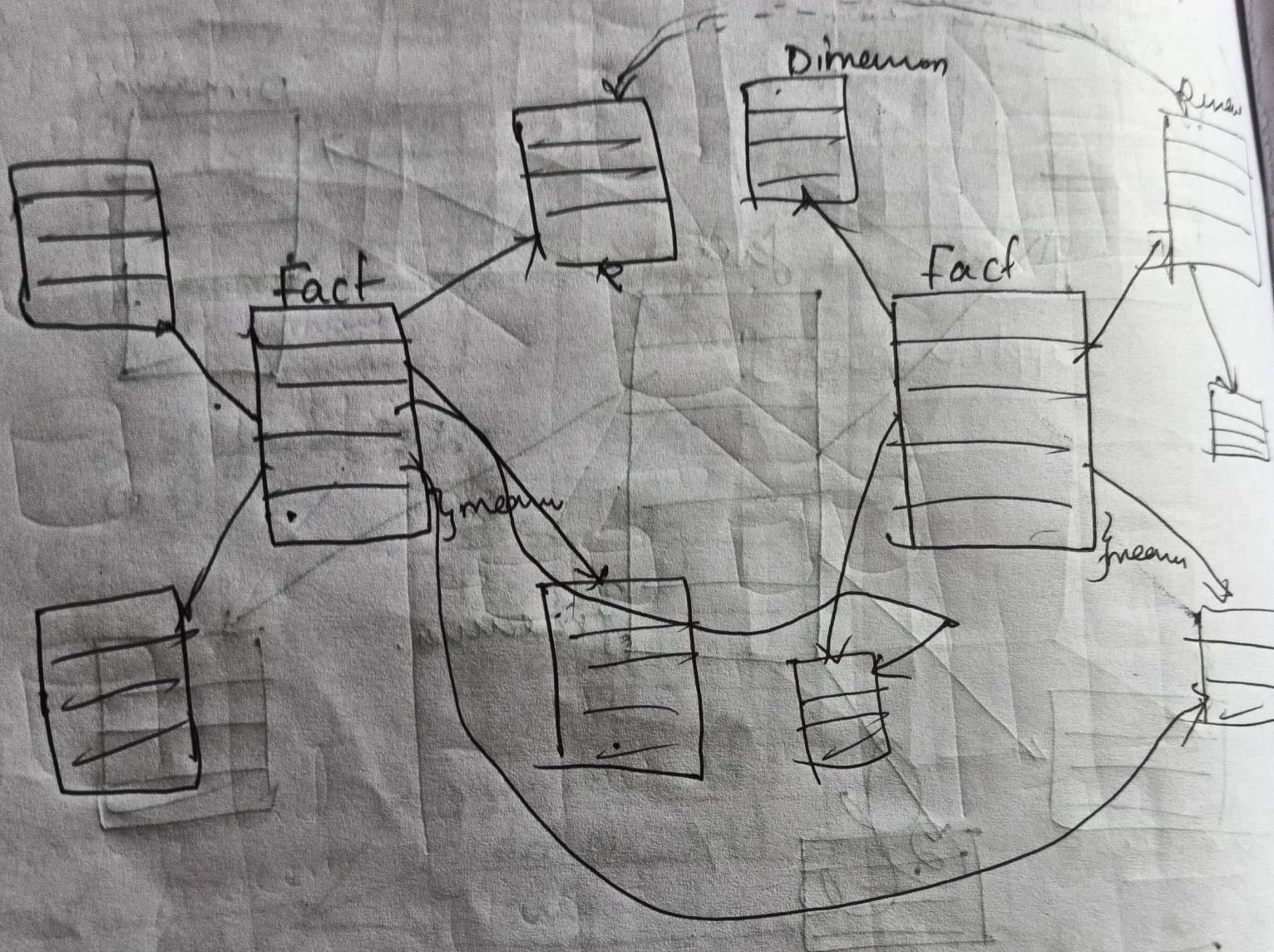
Snowflake schema: A refinement of star schema ~~to~~ normalized into a set of smaller dimension tables, forming a shape similar to.

Galaxy schema: Multiple fact tables share dimension tables, viewed as a collection of stars therefore called galaxy schema.




memory
snowflake schema
Galaxy schema

Galaxy schema.

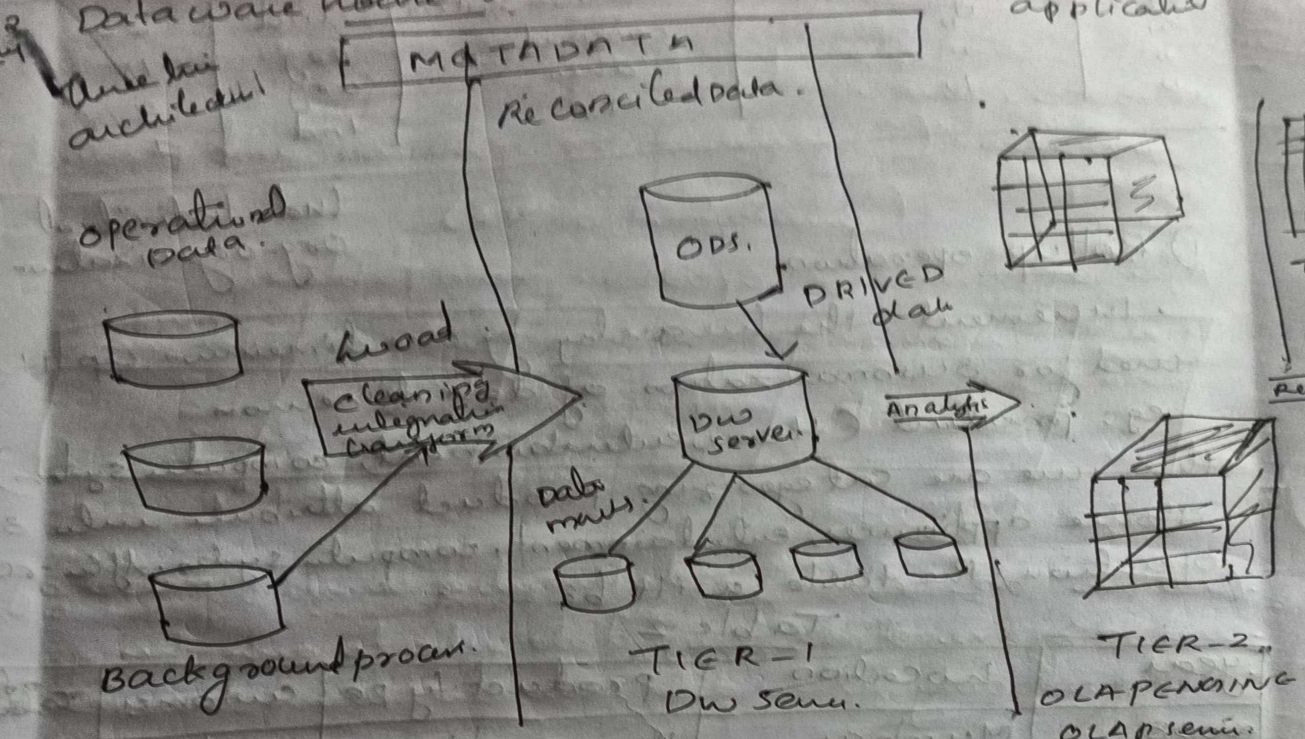


14/2/22
Tuesday Datawarehouse architecture

14/2/23 Tuesday

Data ware house architecture

Data Mining applicati



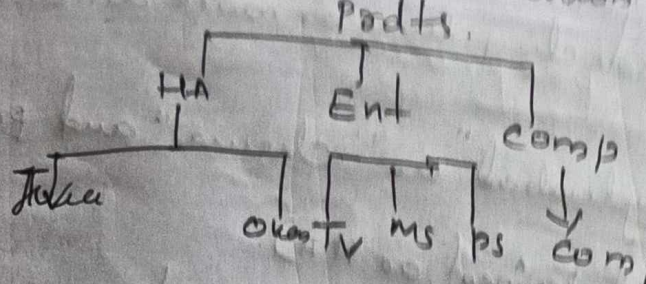
15/2/23 Wednesday

OLAP Operation

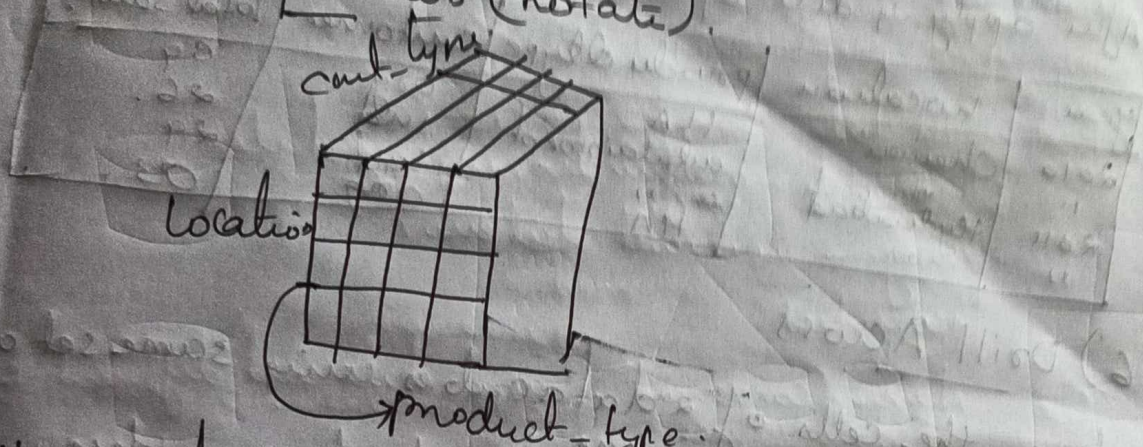
Table :

year	location	Product	Customer	Total sale
2008	Mumbai	oven	ind	32
"	"	Juicer	org	34
"	"	TV	org	12
"	"	Mixie system	ind	11
2009	Pune	Computer	org	43
"	"	Mixie system	ind	21
"	"	TV	ind	14
"	"	play_st	ind	10
2010	chennai	oven	org	9
"	"	Mixie sy	ind	8
"	"	computer	org	13
				24

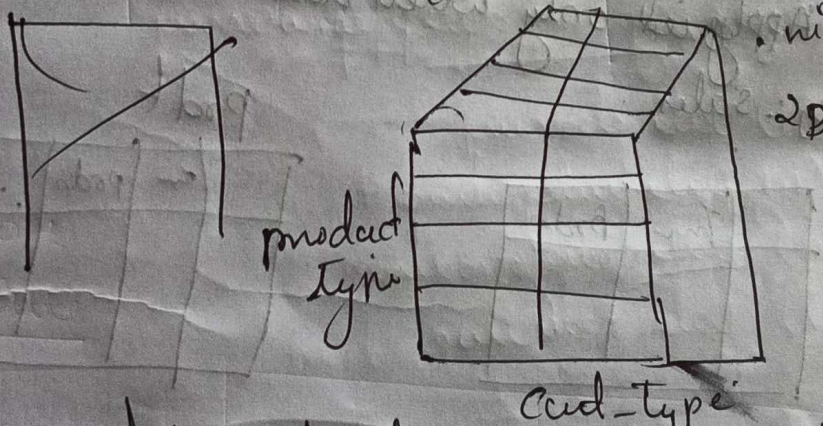
used to document the dimension hierarchy by using



3) Pivot operation (Rotate).



It is used to view data cubes in different form.



4) Slice: It is used to extract a slice of the original cube corresponding to a single value of a given dimension (similar to select operation RDBMS) for eg. Slice Table 2 to on location = "Hyderabad".

year	location	product type	Custom	Total sales
2011	Hyderabad	entertainment	ind	29
"	"	computer	org	15
"	"	HA	org.	03

16/2/23
Thursday

5) Dice operation

It is used for selecting two or more dimension and the or more value.

eg: location ("hyderabad" or "chennai" and product_ty... (entertainment or HA).

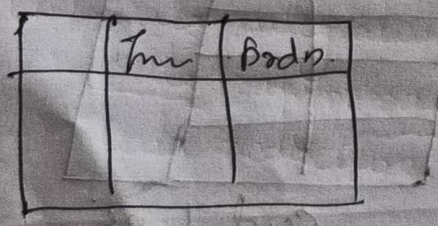
After applying dice operation from Table 2.

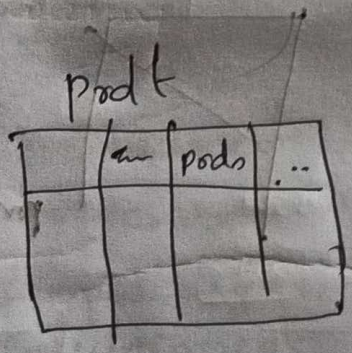
year	location	productType	customer	total sales.
2010	chennai	HA	org	89
"	"	entertainment	ind	26.
2011	Hyderabad	"	ind.	29
"	"	HA	org	03.

6) Drill Across

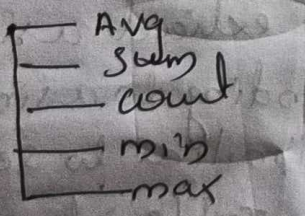
. Analyse the cells of 2 data cubes using same set of dime
. used to move data from org data cube to another.
. Applied only when data cubes are having same/equal dime

eg: Sales.





7) Range query



It applies a given aggregation operation over a set of select cells.

For eg: find the total no: of sales for a period 2008-2010.

Then the range query will return the value by aggregating the values total sales from 2008 - 2010. It can be applied on dimensions such

numerical values.

Range sum query [sum].

year	location	product	customer	Avg-sales	sold available
2008	Mumbai	Over seicer.	Ind	12	03.
"	"		org	14	04.
"	"	TV	org	12	01.