DWM :-

-

-

-

Knowledge discovery in databases.

1] Data cleaning to remove noise and inconsistent data.

2] Data Integration where multiple data sources maybe combined

3] Data selection where data delevent to analysis task are retrieved from the database

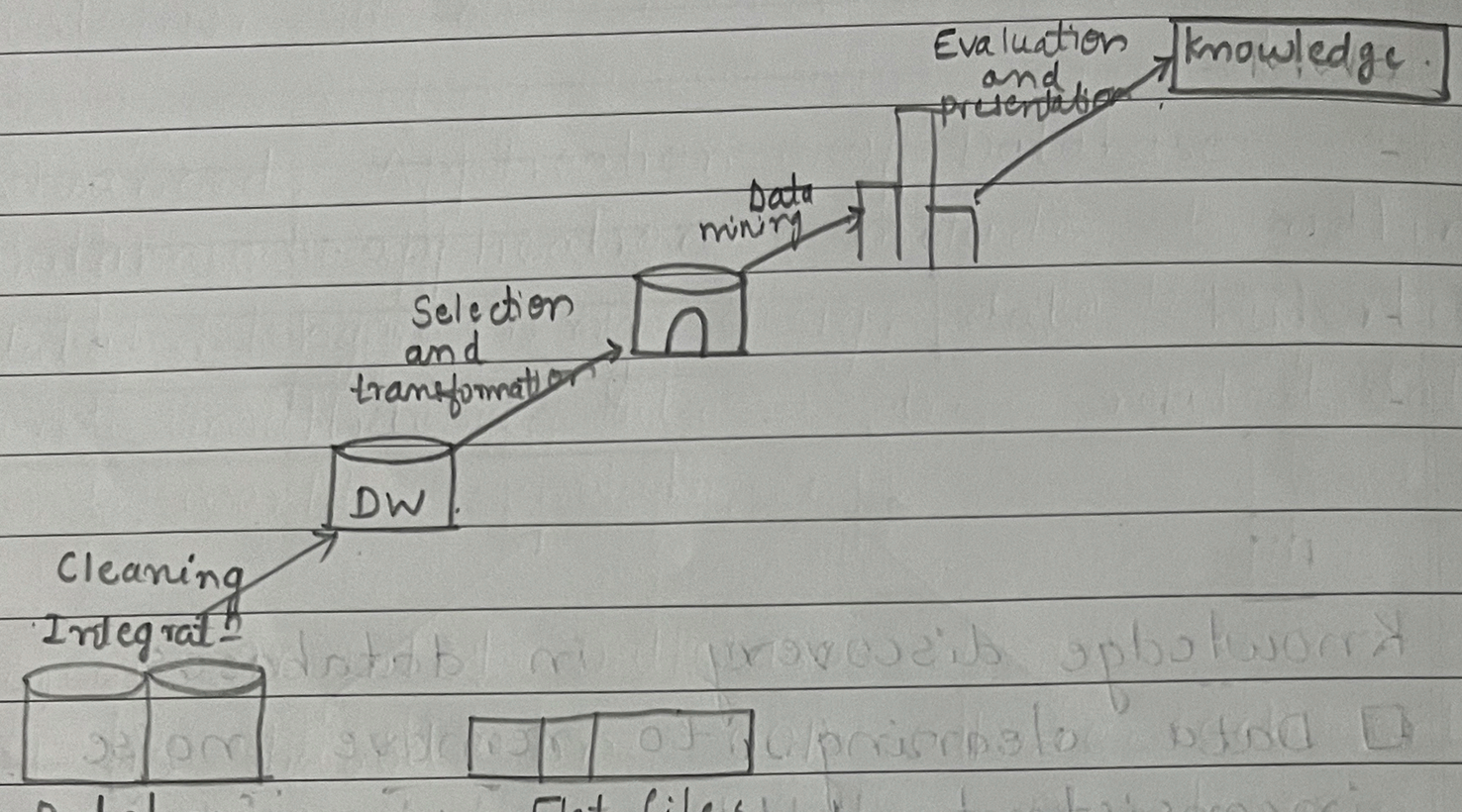
4] Data transformation where data are transformed or consolidated into forms for mining by performing memory or

5] Data mining where intelligent methods are applied to extract data patterns.

6] Pattern evaluation to identify the truly intresting patterns representing knowledge.

7] Knowledge presentation where visualization and knowledge representation techniques are used to present mined knowledge to the user.

6-8 marks. Process of KDD :-



* Architecture of typical DM system :-

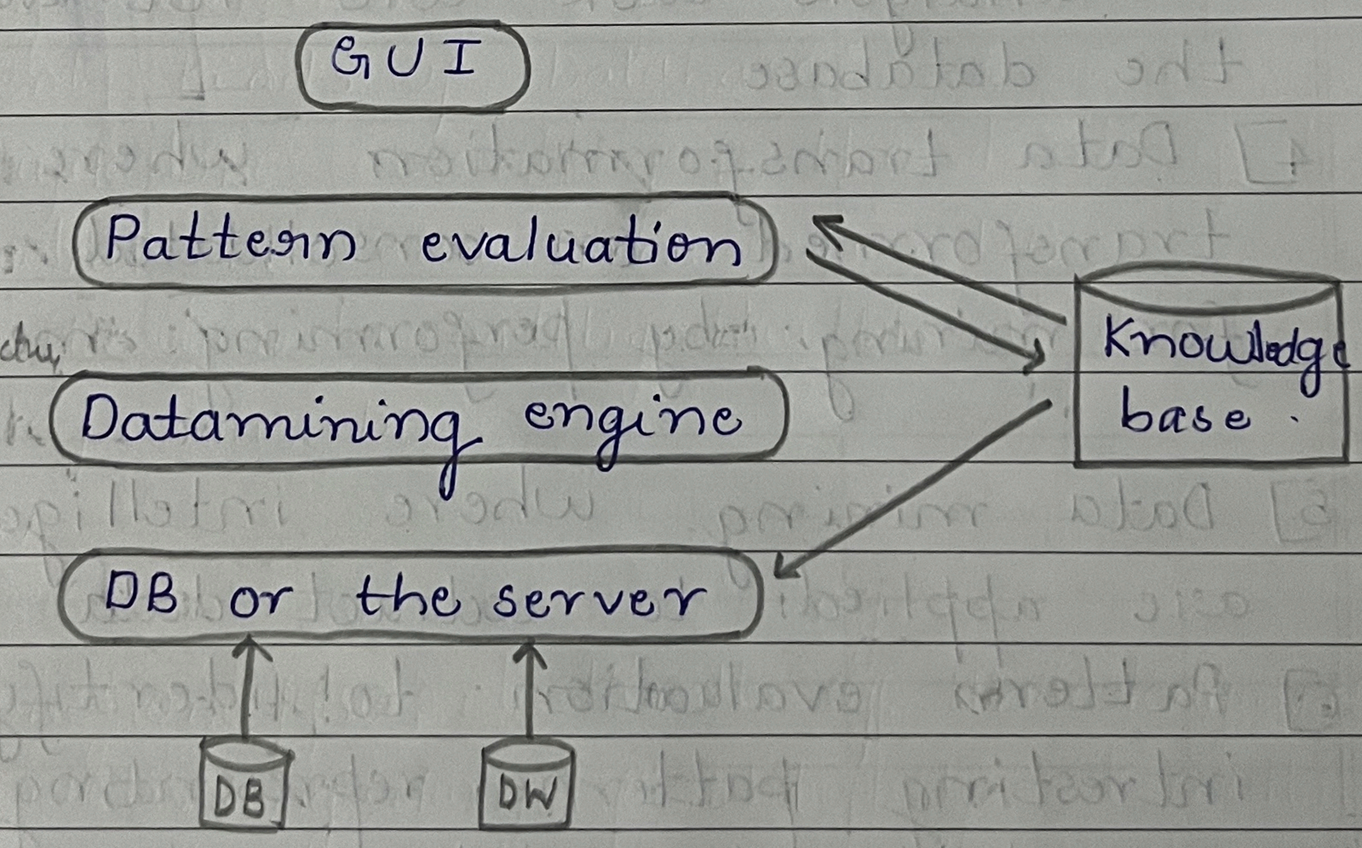


fig. Architecture of typical DM system

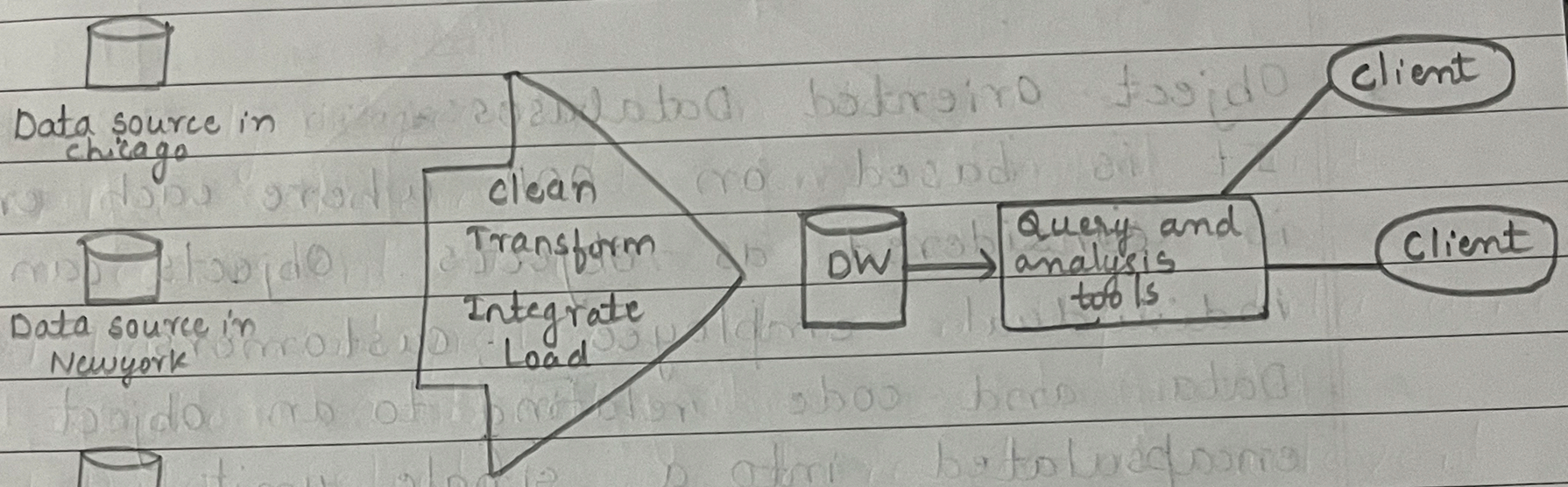
* Types of Data :-

1] Relational Databases :-

Database Management System collection of interrelated data called as database. A relational Database is a collection of tables each of which is assigned a unique name each table consist of set of

attributes and stores large set of tuples.

2.] Data Warehouse:-
Modelled by multidimensional DB structure where each dimension corresponds to an attribute in the schema.



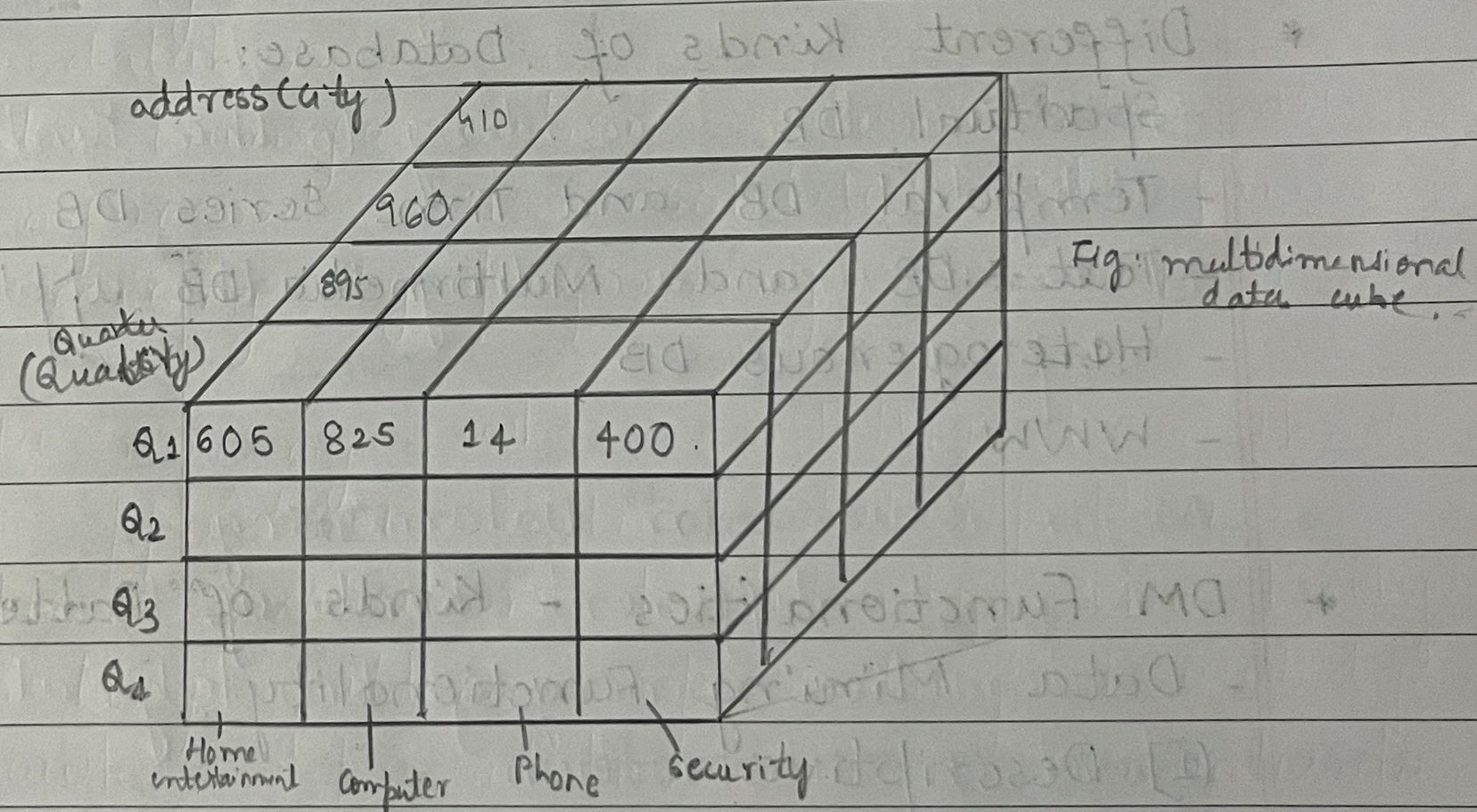


Fig: multidimensional data cube.

3. Transational Databases:-
TDB consist of a file where each record represents a transaction. A transaction includes a unique transaction identity number (transaction), and list of items

making up the tranaction, it contains other information such as date of transaction, customer id no., branch, id. no. of sales.

* Advanced Application of Database:-
Database applications include handling special data, engineering design data, hypertext and multidesign data, time related data and worldwide web.

*. Object Oriented Databases:-
It is based on OOP where each entity is considered as objects. Objects can be individual employee, customers.
Data and code relating to an object are encapsulated into a single unit.

* Different Kinds of Database:-
- Spatial DB.
- Temporal DB and Time Series DB.
- Text DB and Multimedia DB.
- Heterogenous DB.
- WWW.

* DM Functionalities - Kinds of Patterns canbeMined
- Data Mining Functionality.
(i) Descriptive
(ii). Predictive.

1. concept / class Description : characterizations and Discrimination.

2. Association Analysis.

Association Analysis is the discovery of association rules showing attribute value

conditions that occur frequently together in a given set of data.

1. age (X, "20 29") ∧ income (X, 20K..29K") ⇒ buys (X, "comp
(support = 2%, confidence = 60%.)
2. contains (T, computer") ⇒ contains (T, "software")
(support = 1%, confidence = 50%).

3. Classification and Prediction
- Distinguish data classes / concepts.
- Training data.
- (IF - THEN) rules.

Eg - All electronics company
good response, mild response and no response.

4. Cluster Analysis
- Analyse data objects without consulting a known class label
- Class labels are not present in training data.

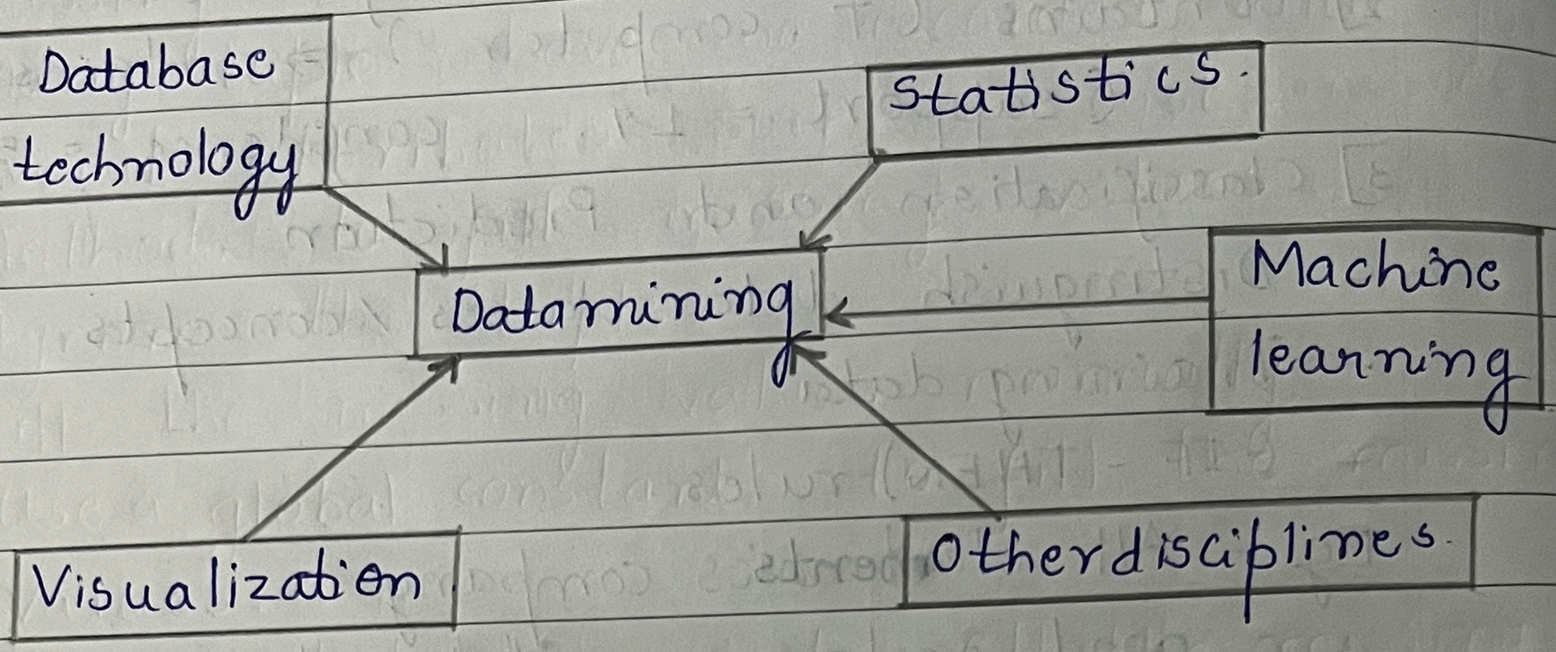
Eg:-

S No.	Question	Label
1.	What is cluster analysis	Descriptive
2.	Water content in banner.	Numeric
3.	Where GIT is located	Location.

The objects are clustered or grouped based on the principle of maximising the intra-class similarity and minimizing the inter-class similarity.
The cluster of objects are formed so that objects within a cluster have high similarity in comparison to one-another and very dissimilar to objects in other clusters.

5] Outer Analysis.
- Do not comply with the general behaviour

* Classification of Datamining systems:-



1] Classification according to the kinds of databases mined.
2] Classification according to the kinds of knowledge mined.
3] Classification according to the kinds of techniques utilized
4] Classification according to the applications adapted.

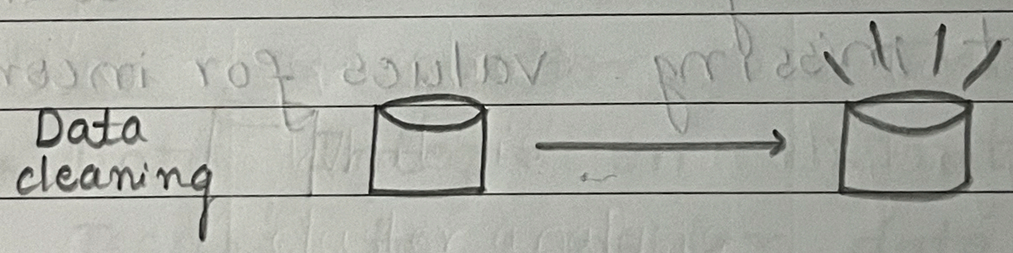
* Major issues in Datamining:-
- Mining different kinds of knowledge in DB
- Interactive mining of knowledge at multiple level of abstraction.
- Data mining query languages.
- Presentation and visualisation of DM results.
- Handing noisy or incomplete data.

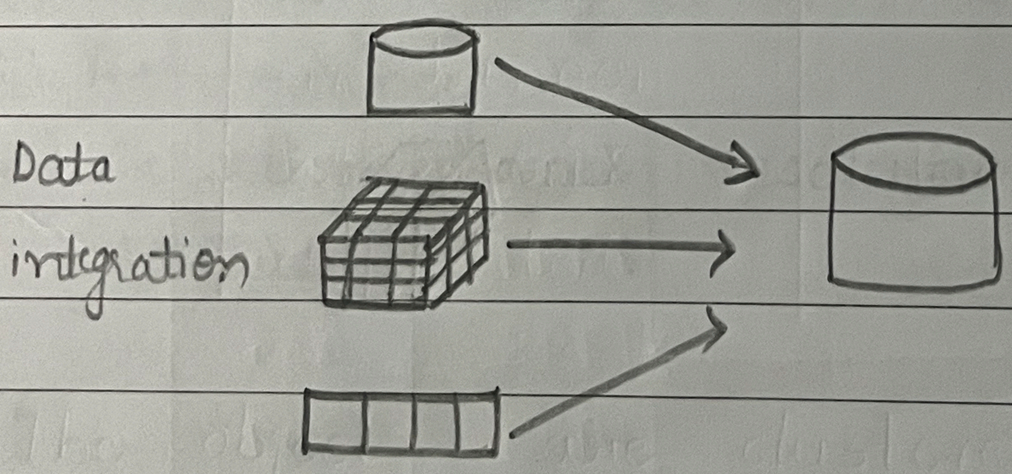
* Data Pre Processing:-
- Need of Data Preprocessing.
- What is Data Preprocessing.

- To improve quality of data.
- Data preprocessing techniques.

Data preprocessing techniques :-
Data cleaning, data integration, data transformation, Data reduction.

* Data cleaning can be applied to remove noise and correct inconsistency in the data.
* Data integration merges data from multiple sources into a coherent datasource such as data warehouse.
* Data transformation normalization maybe applied to improve efficiency and accuracy of mining algorithm.
* Data reduction can reduce the datasize, elemination and aggregating redunt features or clustering.





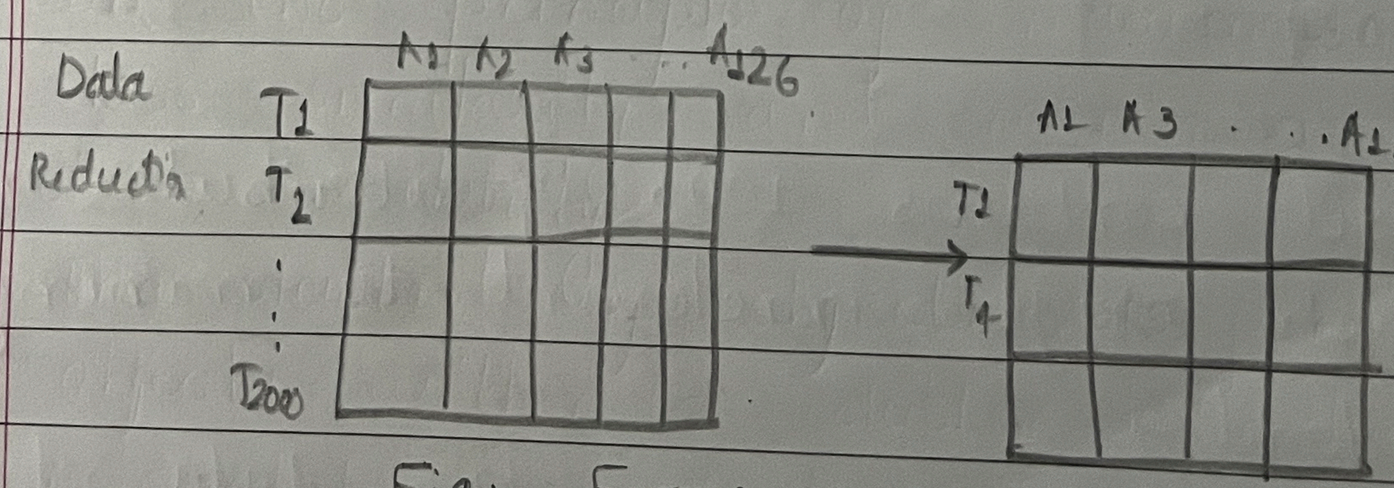


Fig:- Forms of Data preprocessing

-2, 32, 100, 59 ⟶ 002, 0.32, 1.00, 0.59

Data cleaning:-

* It attempts to fill the missing values, smooth out noise, identifying the outliers and correcting inconsistency in the data

(i). missing values

- Many tuples have no recorded value for several attributes such as cust_income.
- Methods to fill missing values.

a. Ignore the tuple.
b. Fill the missing values manually.
c. Use a global constant to fill the missing values.
d. Use attribute mean to fill the missing values.
e. Use the most probable value to fill the missing values.
f. Using other customer attributes in a dataset you may construct a decision tree to predict missing values for income.

Data Preprocessing:-

1. Noisy Data:-
   * It is a random error or variance in a measured variable.
2. Clustering:-
   * Outliers may be detected by constraints where similar values are organised into group clusters. values that fall outside of set of clusters considered as outliers.
3. Regression:-
   * Data can be smoothed by fitting data to a function.
   * Linear regression involves finding two lines such that to fit two variables one variable can be used to predict other. Multiple regression where more than two variables are involved.

Data Integration and Transformation:-

* Data Integration Definition:-

  Issues to consider data Integration:-

(i) Schema Integration → Entity identification problem

(ii) Redundancy.

(iii) Detection and resolution of data value conflicts.

Data Trans

* Data are transformed or consolidated into forms appropriate for mining. It involves.

(i) Smoothing - which works to remove the noise from data.

(ii) Aggregration - where summary or aggregation operations applications are applied to the data.

(iii) Generalization where low level data are replaced by higher level concepts through higher level hierarchy.
street → city or country.
age → young, middleage or senior.

(iv) Normalization where attribute data are scaled to fall within specified rage (0.2 - 1.0)

(v). Attribute construction:- where new attributes are constructed and added from given set of attributes.

Data Reduction:-

* To obtain a reduced representation of the data set.
* Maintain integrity of original data.

(i). Data cube aggregation.
(ii). Dimension Reduction.
(iii). Data compression.

Data cube Aggregation:

Year = 1999
Year = 1998
Year = 1999

Quarter	Sales.
Q1	224
Q2	408
Q3	350
Q4	58

→

Year	Sales.
1997	156
1998	235
1999	359

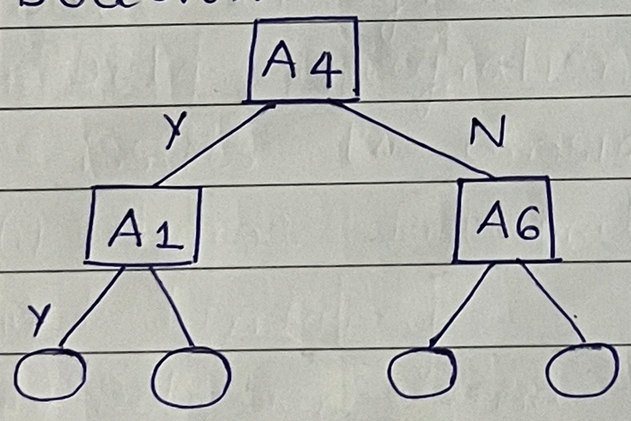
Data set contains 100's of attributes many of which maybe irrelevant to mining tasks redundant.
Dimensionality reduction reduces the data set size by removing such attributes or dimensions

from it.

* Forward selection
Initial attribute set
[A1, A2, A3, A4, A5, A6]
Initial reduced set
{ }.
→ {A1}
→ {A1, A4}.
Reduced → {A1, A4, A6}

* Backward elemination
Initial attribute set
{A1, A2, A3, A4, A5, A6}.
→ {A1, A3, A4, A5, A6}.
→ {A1, A4, A5, A6}.
Reduced → {A1, A4, A6}.

* Decision Tree Induction.



A1, A4, A6.

Data compression :-
Data encoding or transformations are applied to reduce or obtain compressed representation if the original data can be reconstructed from compressed data without any loss of information is called lossless.

22/11/2023.

UNIT-II.

DATA WAREHOUSE

Provides tools for business executives to systematically organize, understand and use their data to make strategic decisions

6 marks **

Key features of Data Warehouse:

1. Subject oriented:-
   It is organised around major subjects such as customer, supplier, products and sales
2. Data warehouse is construced by integrating multiple heterogeneous sources.
3. Time varient:-
   Data are stored to provide information from historical perspective [past 5-10 years]
4. Non-volatile:-
   Data warehouse is a phyisically separate store of data transformed from application data found in the operational environment.

Difference between operational DB systems and Datawarehouse:-

- OLTP → cover day to day operations.
  Eg:- Banking, inventory, purchasing.
- DW → Data analysis and Decision making.

Features	
Characteristics	O
Orientation	T
User	C
	Ex
Function	D
DB Design	
Access	
No. of records accessed	
No. of users	
DB size	

Features	OLTP	OLAP
Characteristics	operation Processing	Informational Processing
Orientation	Transaction.	Analysis.
User	Clerk, DBA, DB professional. Ex:- DB Administrator.	Knowledge worker Ex:- Manager, executive, analyst
Function	Day-to-day operations	Long term informational requirements.
DB Design	ER Based.	Star / Snowflake, subject oriented.
Access	Read and write	Mostly read
No. of records accessed	Tens	Millions.
No. of users.	Thousands	Hundreds.
DB size	100 mb to gb.	100 gb to tb.

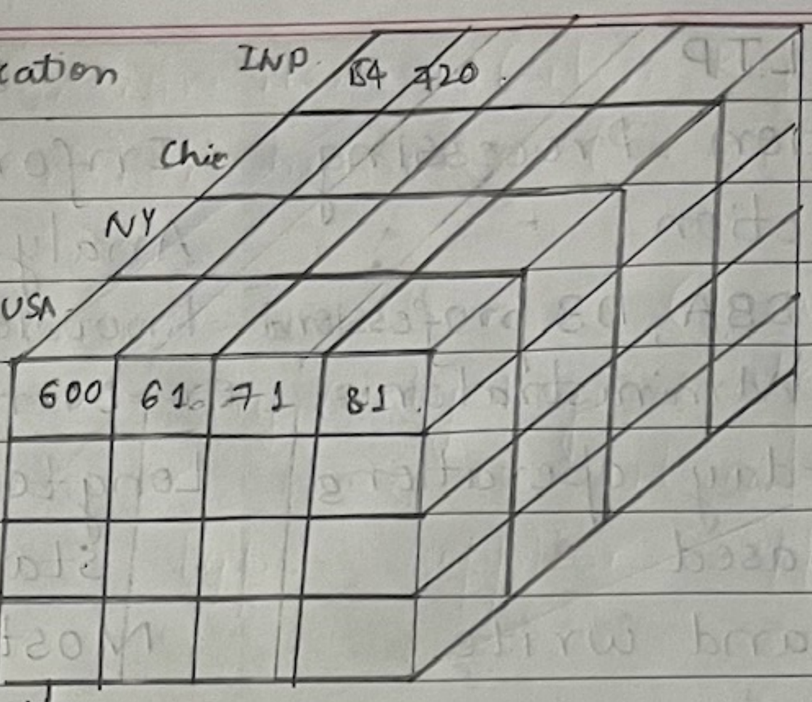
Multidimensional Data model:.

* Views data in form of datacube
* Datacube - Data to be modelled and viewed in multiple dimensions.
* Defined by dimensions and facts.
  Ex:- Dimensions → time, item, branch.
  Facts → Dollars sold, units sold.

9-D view of sales data:.

location = USA.

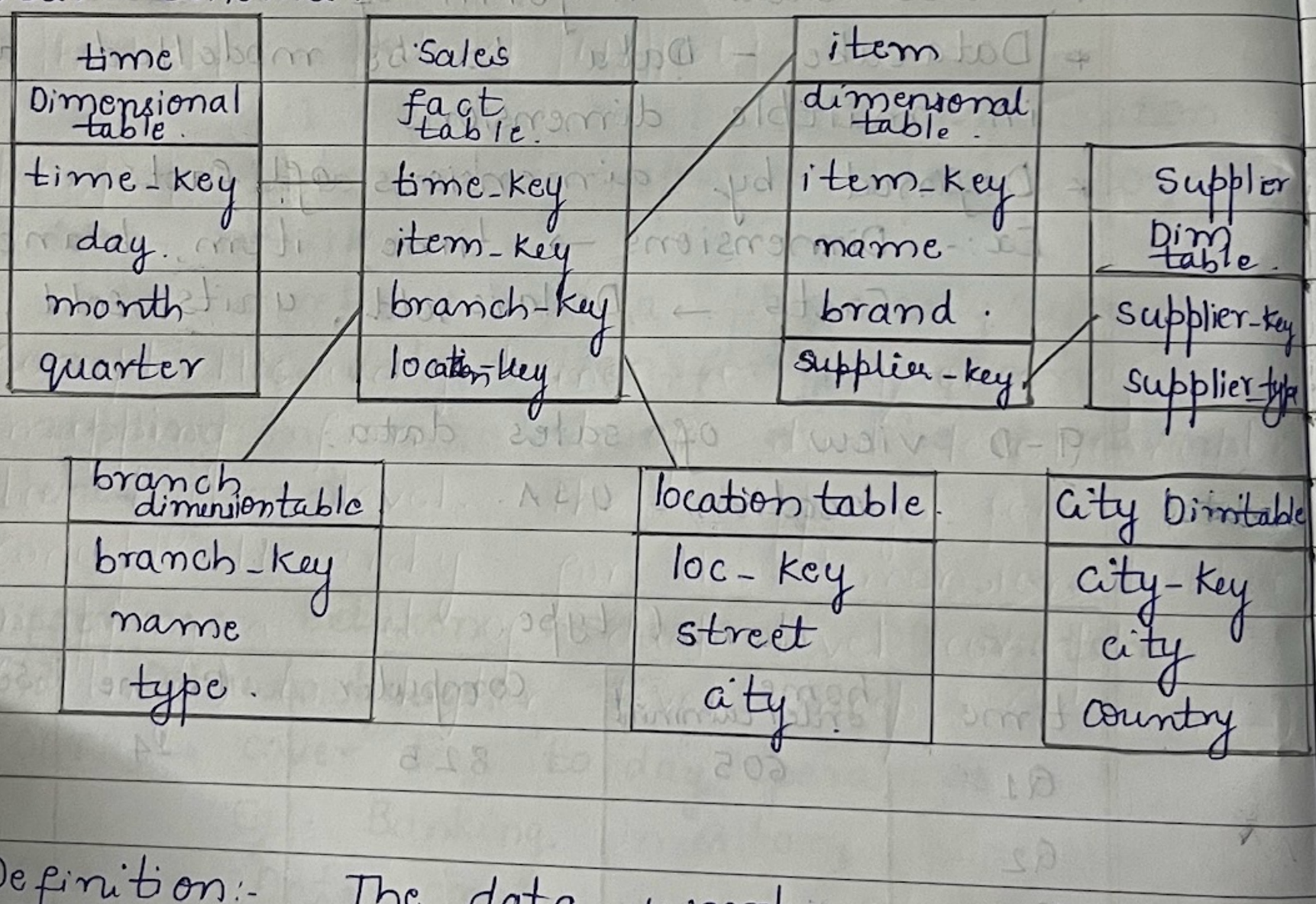
item (type.				
time.	home entertainment	Computer	Phone	Security.
Q1	605	825	14	400
Q2				
Q3				
Q4				



3D Datacube Representation

* Star, Snowflake and fact constellations:- Schemas for multidimensional DB.

1. Star Schema:



Definition:- The data warehouse contains
* A large central table containing bulk data with no redundancy.
* A set of smaller attendant tables (dimension table) one for each dimension.

2. Snow
It
mod
mo
int

3. Fact
I
dim
can
it
con

Ex
fa
*
*
S

2. Snowflake schema :-
It is the variant of the star schema model where some dimension tables are modified. Thereby further splitting the data into additional tables.

3. Fact constellations schema :-
It requires multiple fact tables to share dimension tables. This kind of schema can be viewed as collection of stars hence it is called as galaxy schema or fact constallation schema.

Sales
fact table
time_key
item_key

shipping
fact table
item_key
time_Key
Shipper_key

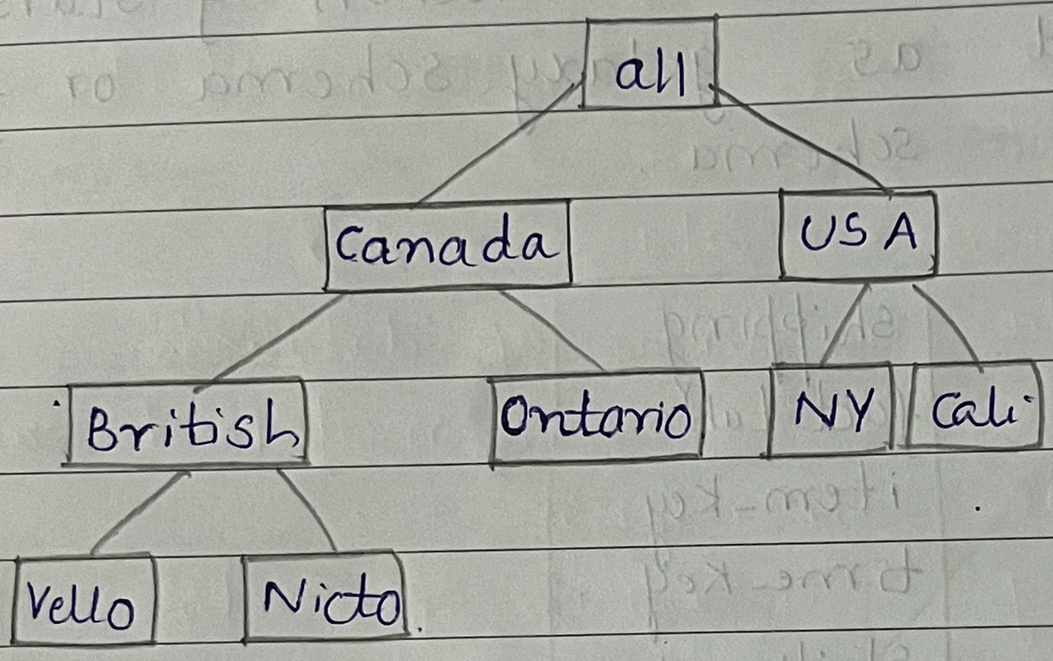
Examples for defining star, snowflake and fact constellation schema :-
* SQL based DM Query language - DMQL.
* DW defined using 2 language primitives.
  - One for cube definition.
  - One for dimension definition.

Syntax :-
define cube <cubename> <dimensionlist> : (measure list)
define dimension <dimensionname> as (<attribute_list

Eg :- Star Schema.
define cube sales_star [time, item, branch, location]
dollars_sold = sum(sales_in_dollars), units_sold
= count(*)
define dimension time as (time_key, day, month, year)

define dimension item as (item_key, name, brand)
define dimension branch as (branch_key, name)
define dimension location as (location_key, street, city (city_key, state, country).
define dimension item as (item_key, name, type, supplier (supplier_key, type)

* Concept Hierarchy:-



Concept Hierarchy defines a sequence of mapping from set of dual level concepts to higher level. The mappings form a concept Hierarchy for the dimension location mapping a set of dual level concepts (~~settings~~ cities) to higher level concepts (~~contents~~ countries).