Assignment No. 1

Q.1) A) Differentiate between :

i) Datamart and metadata.

Datamart	Metadata
① Datamart can be consider as the db or collection of db's that are design to help managers in making strategic decision about buiness and organizations.	metadata is defined as data about data that describe the data warehouse the data that is used to represent other data is known as metadata.
② Data may or may not be informative	Metadata is always informative.
③ Data may or may not have been processed.	It is always a processed data.
④ data is stored as a file either navigational or hierarchical form	It is stored in data dictionary
⑤ Example : If you create a notepad file, then the content of that document is data	Example : If you create a notepad file, the name of file, storage description, type, size becomes metadata of file.

ii) OLTP and OLAP :

Category	OLAP	OLTP
Definition :	It is well known as an offline db query management system	It is well-known as an online db modifying system.
Method used :	It makes use of a data warehouse	It make use of Standard DBMS.
Application :	It is subject-oriented	It is application oriented.
Normalized :	Table are not normalized	tables are normalized (3NF)
Volume :	100GB's to TB's	100 MB's to GB's
Type of users :	This data is generally managed by CEO, MD, GM	This data is managed by clerks, managers.
Operations :	only read & rarely write operation	Both read & write operation.
productivity :	Improves the efficiency of business analysts	Enhances the user's productivity
Function :	day to day operations	decision support.

1.6) what is data model? Explain multidimensional datamodel in detail with a real time example.

→ * Data Model :→

Data model describe how a db's logical structure is represented. Data models specify how data is linked to one another, as well as how it is handled and stored within the system.

* Multidimensional data model :→

- Datawarehouse and OLAP tools are based on multi-dimensional datamodel. It is a method which is used for ordering data in the database.
- It view data in the form of data cube.
- Data cube allows the data to be modelled and view in multiple dimension. It defined by dimension and facts.
- Example:

Take the example of data of the factory which sells the products per quarter in location bangalore. The data is represent in table given below:

Time (Quater)	location: Bangalore			
	Type of Items			
	Jam	Bread	Sugar	Salt
Q1	350	389	35	50
Q2	260	528	50	90
Q3	483	256	20	60
Q4	436	396	15	40

In the above presentation, the facts for the sales manager are for the time dimension which is organised into quarter and the dimension of items which are sorted w.r.t the kind of items which is sold. The facts here are represented in Rupees (in thousands).

Time	location = "kolkata"			location = "delhi"			location = "mumbai"		
	item			item			item		
	Jam	Bread	Sugar	Jam	Bread	Sugar	Jam	Bread	Sugar
Q1	340	664	38	335	365	35	336	484	80
Q2	680	583	10	689	496	48	595	594	39
Q3	535	490	50	389	385	15	366	385	20

Fig: 3D data representation as 2D

The data can represented in the form 3D conceptually, which is shown in the image below

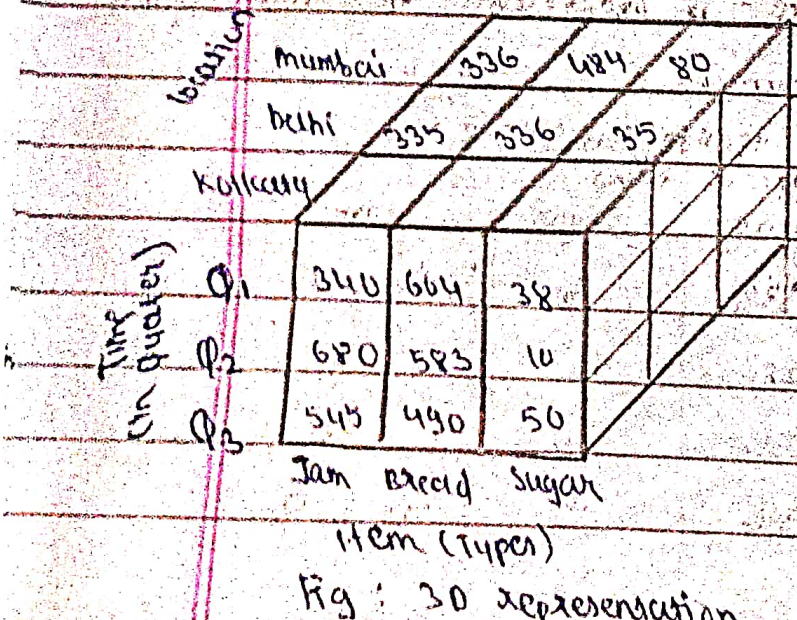


Fig: 3D representation

Q.2 Define data warehouse. Draw the architecture of data warehouse and explain the three tier in detail.

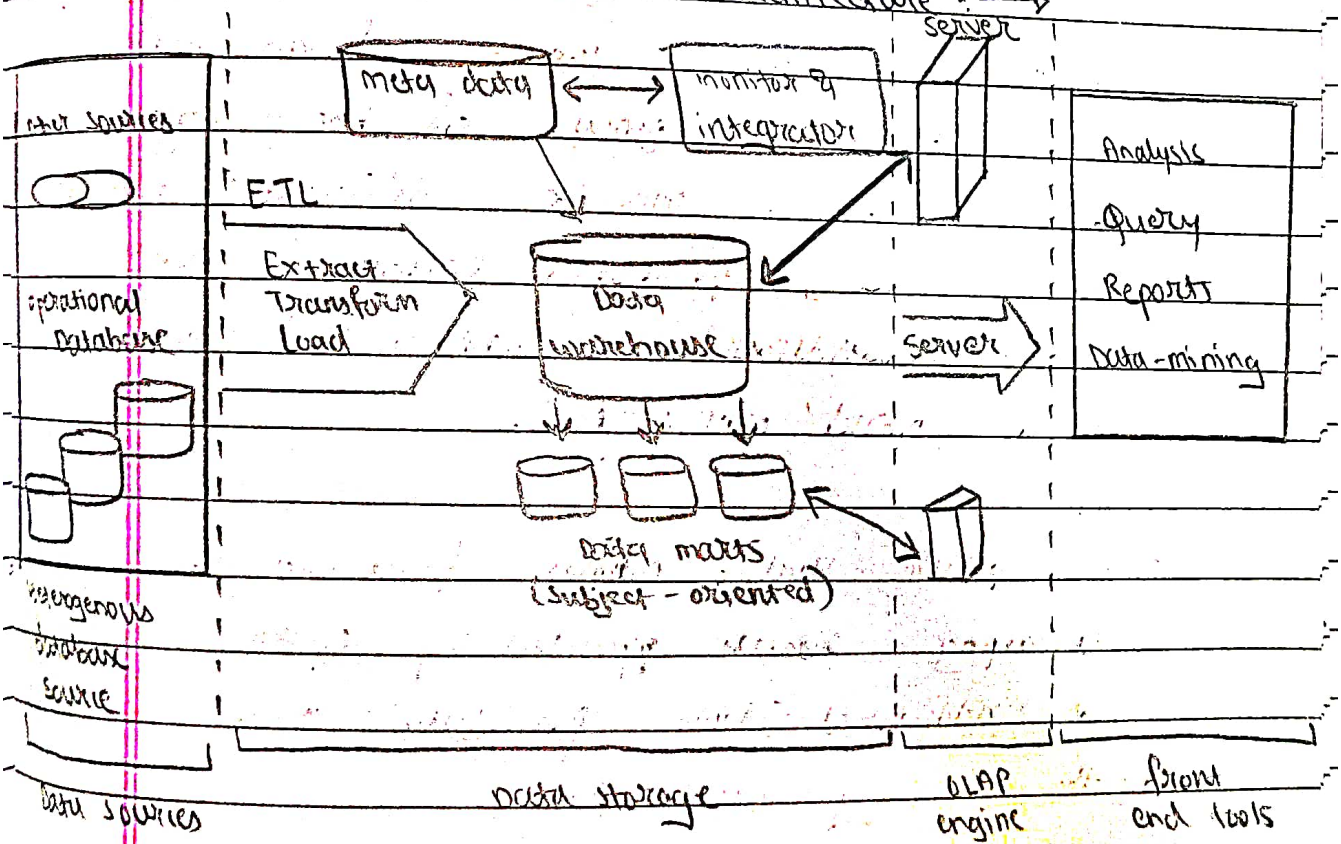
→ * Data warehouse :→

① It is a centralized storage system that allows for the storing, analyzing and interpreting of data in order to facilitate better decision-making.

② A data warehouse is large collection of business data used to help an organization make decision.

③ It is subject-oriented, integrated, non-volatile and time variant collection of data in support of management decision making process.

* 3-tier data warehouse Architecture :→



Data warehouse usually have three levels (tiers) architecture that includes :-

① Bottom tier that consist of data warehouse server which is almost a RDBMS. It may include several specialize data marts and many data repository. Data from operational databases and external sources are extracted using application program interfaces called a gateway.

A gateway is provided by the underlined DBMS and allows customer program to generate SQL code to be executed at a server example of gateways contains ODBC (open database connectivity) & JDBC (Java database connectivity).

② middle tier, which consist of an OLAP server for fast query of the data warehouse. There can be different types of OLAP servers ie. ROLAP (relational OLAP server), MOLAP (multidimensional OLAP server) and HOLAP (Hybrid OLAP server).

③ A top tier, that contains front-end tools for displaying results provided by OLAP, as well as additional tools for data mining of the OLAP generated data.

Ques: 3) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

→ The steps are :

① Data cleaning :→

To remove noise and inconsistent data.

② Data integration :→

where multiple data sources may be combined.

③ Data selection :→

where data relevant to the analysis task are retrieved from the database.

④ Data transformation :→

where data are transformed or consolidated into form appropriate for mining by performing summary or aggregation operations, for instance.

⑤ Data mining :→

an essential process where proces intelligent methods are applied in order to extract data patten.

⑥ pattern evaluation :→

To identify the truly interesting patten representing knowledge based on the some interestingness measures.

⑦ knowledge presentation :→

where visualizations and knowledge representation techniques are used to present the mined knowledge to the user.

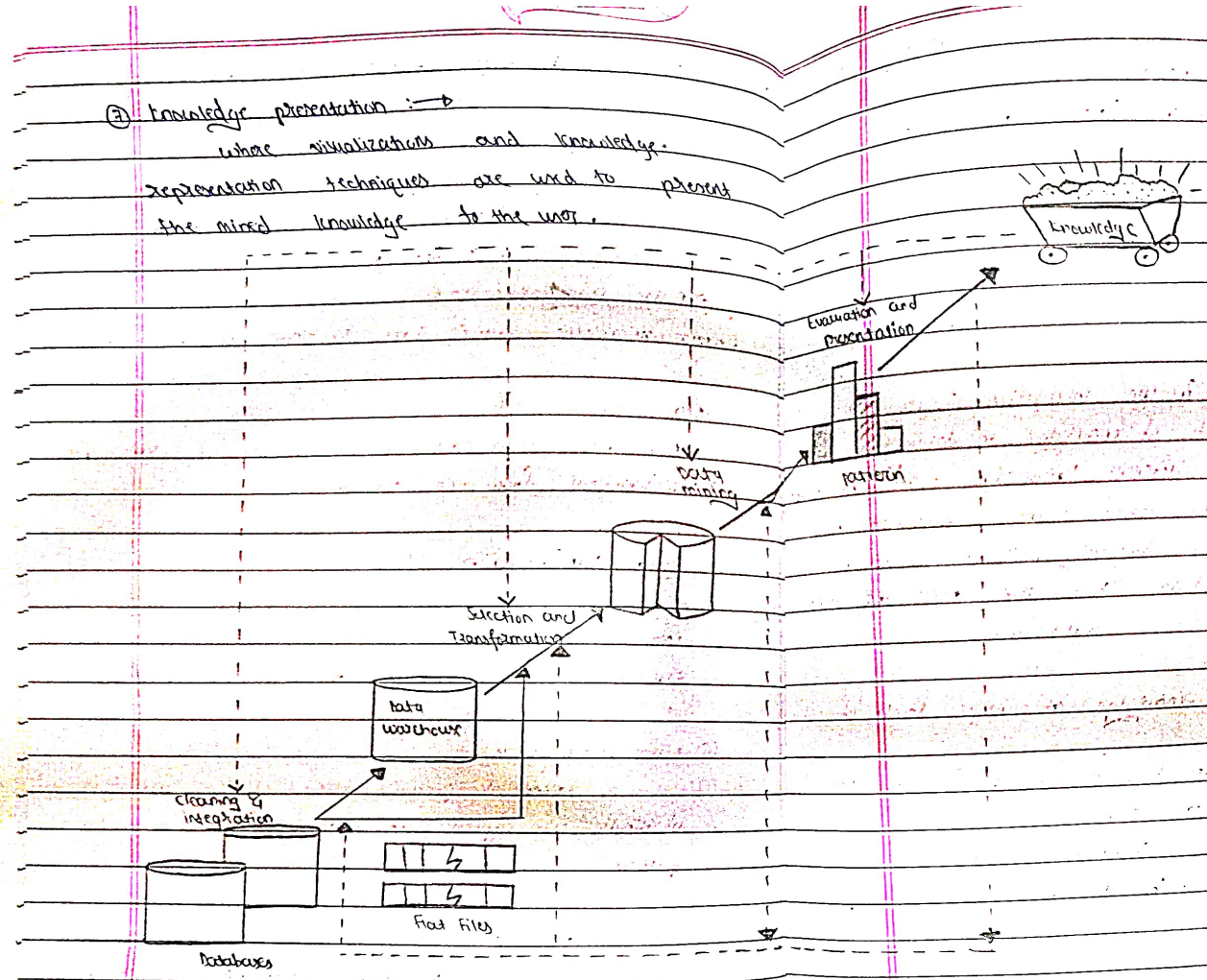


Fig: Steps in data mining when viewed as a process of knowledge discovery

Date
Page

Ques 4) Explain various major issues and challenges in data mining in detail.

→ Data mining system face a lot of data mining challenges and issues in today's world. Some of them are:

1. Mining methodology and user interaction issues.
2. performance issues.
3. Issues relating to the diversity of db types.

1. Mining methodology & user interaction issues: →

i) Mining different kinds of knowledge in databases:

Different user - different knowledge - different way. That means different client want a different kind of information so it become difficult to cover vast range of data that can meet the client requirement.

ii) Interactive mining of knowledge at multiple levels of abstraction:

It allow user to focus the search for patterns from different angles. The data mining process should be interactive because it is difficult to know what can be discovered within a db.

iii) Incorporation of background knowledge:

It is used to guide discovery process and to express the discovered patterns.

iv) Query language and ad hoc mining :→
It allow users to pose ad-hoc queries for data retrieval.

v) Handling noisy or incomplete data :
In a large database, many of the attribute values will be incorrect. This is may be due to human error or because of any instruments fail. Data cleaning methods and data analysis methods are used to handled noise data.

② Performance issue :→
i) Efficiency and scalability of data mining algorithms :
To effectively extract information from a huge amount of data in databases, data mining algorithm must be efficient and scalable.

ii) Parallel, distributed and incremental mining algo :
The huge size of many databases, the wide distribution of data and complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algo.

③ Issue relating to the diversity of database types :→
i) Handling of relational and complex types of data :
There are many kinds of data stored in db's

and data warehouses. It is not possible for one system to mine all these kind of data. So different data mining system should be constructed for different kinds of data.

2) mining info from heterogeneous databases and global information systems:

Since data is fetched from different data sources on LAN and WAN. The discovery of knowledge from different sources of structured is a great challenge to data mining.

Q.4 B) Discuss

i) Data cleaning :→

① Data cleaning is a crucial process in data mining.

② Data cleaning is fixing or removing incorrect, corrupted, incorrect formatted, duplicate or incomplete data within a dataset.

③ When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

④ Generally, data cleaning reduces errors and improves data quality.

⑤ It is not simply about erasing info. to make space for new data, but rather finding a way to maximize a data set's accuracy

without necessarily deleting information.

(1) Steps of data cleaning :

a) Remove duplicate or irrelevant observations

b) Fix structural error

c) Filter unwanted outliers

d) Handle missing data

e) validate and QA.

ii) Data integration :→

(1) Data integration is a data preprocessing technique that combines the data from multiple heterogenous data sources into a coherent data store and provides a unified view of the data.

(2) The goal of data integration is to make it easier to access and analyze data that is spread across multiple systems or platforms, in order to gain a more complete and accurate understanding of the data.

(3) There are mainly 2 type or approach for data integration :

a) Tight coupling :→

This approach involves creating a centralized respository or data warehouse to store the integrated data. This approach is also known as data warehousing, and it enables data consistency and integrity, but it can be inflexible and difficult to change or update.

b) Loose coupling :→

This approach involves integrating data at the lowest data level, such as level of individual level data element or record. This approach is also known as data federation, and it enables data flexibility and easy updates, but it can be difficult to maintain consistency and integrity across multiply data sources.

Q.5 A) Explain tree induction algorithm for building decision tree.

→ Input :

- data partition, D, which is a set of training tuples and their associated class labels;
- attribute-list, the set of candidate attributes;
- Attribute-selection-method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of splitting-attribute and possibly either a split point or splitting subset.

Output : A decision tree.

Method :

1) create a node N;
2) if tuples in D are all of the same class, C then
3) return N as a leaf node labeled with the class C;

4) if attribute_list is empty then
5) return N as a leaf node labeled with the majority class in D; // majority voting
6) apply Attribute_selection_method (D, attribute_list) to find the 'best' splitting_criterion;
7) label node N with splitting_criterion;
8) if splitting is discrete-valued and
multiway splits allowed then
// not restricted to binary trees
9) attribute_list ← attribute_list - splitting_attribute
// remove splitting_attribute
10) for each outcome j of splitting_criterion
// partition the tuples and grow subtrees for
each partition
11) let Dj be the set of data tuples in D
satisfying outcome j; // a partition
12) if Dj is empty then
13) attach a leaf labeled with the
majority class in D to node N;
14) else attach the node returned by
Generate_decision_tree (Dj, attribute_list) to
node N;
end for
15) return N;

Q.3 a) Differentiate betn hierarchical and non-hierarchical clustering.

→

Hierarchical clustering	Non-hierarchical clustering
① It involves creating clusters in a predefined order from top to bottom	It involve formation of new clusters by merging or splitting the clusters instead of following a hierarchical order
② less reliable than non-hierarchical clustering	more reliable than hierarchical clustering
③ slower than non-hierarchical clustering	faster than hierarchical clustering
④ It is relative unstable	It is relative stable
⑤ It is comparatively easier to read & understand	The cluster are difficult to read and understand as compared to hierarchical clustering
⑥ It is very problematic to apply this technique when we have data with high level of errors	It work better than even when error is there