DOMS | Page No.
Date / /

⇒ Data mining architecture :-

The significant components of data mining systems includes :-

- Data Sources.
- Data mining engine
- Data warehouse server
- Pattern evaluation
- gui
- Knowledge base

Data Source :-

The actual source of data is database, data warehouse, world wide web (ww), text files and other document.

* Organizations typically stores data in database or data warehouse.

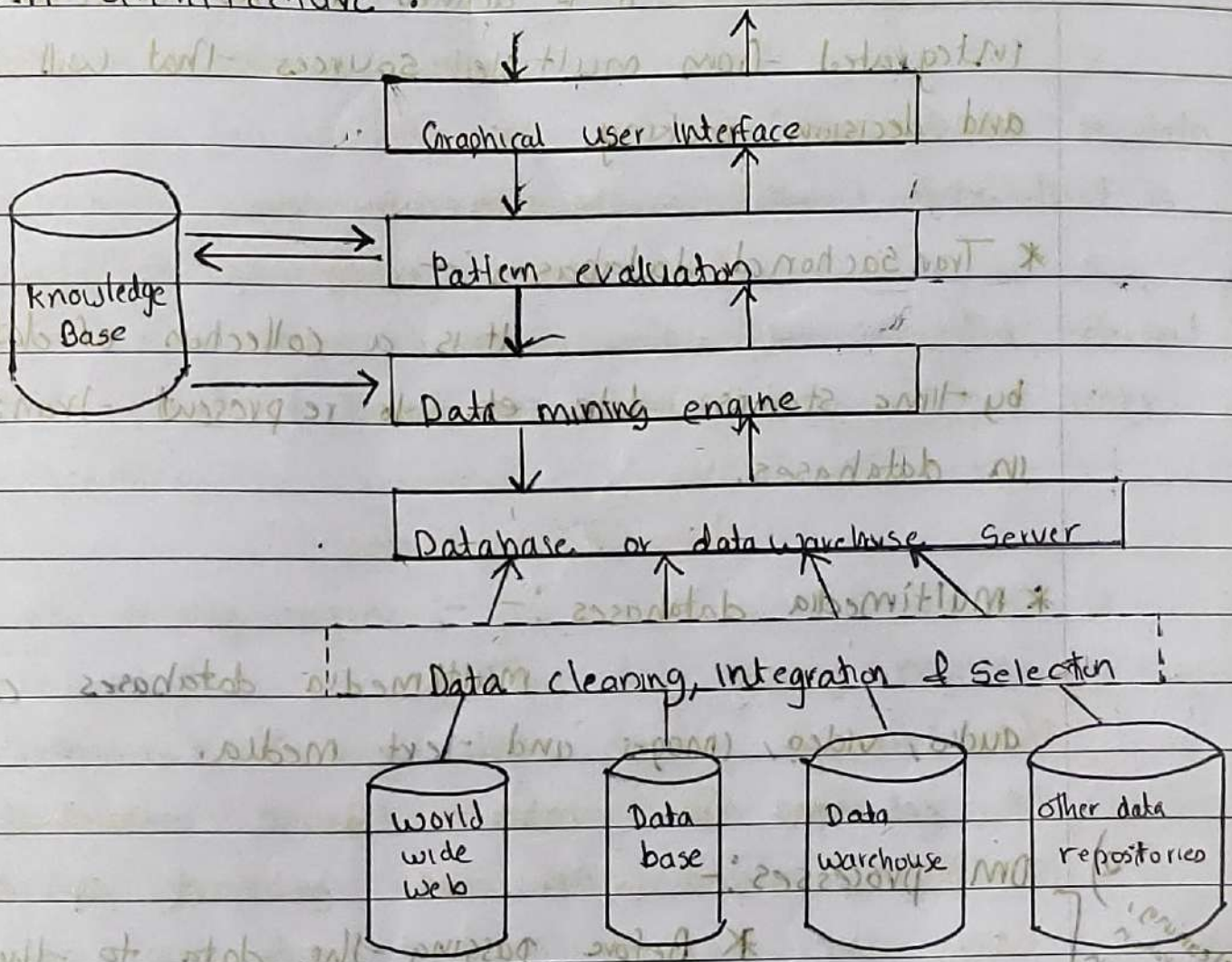
Some of the data sources may be :-

⇒ Flat files :-

Data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.

DOMS | Page No.
Date / /

⇒ DM architecture :-



⇒ Datamining components :-

1) Data Source :-

* Relational database :- A relational database is defined as the collection of data organized in tables with rows and columns.

DOMS | Page No.
Date / /

* Data Warehouse :-

It is defined as the collection of data integrated from multiple sources that will queries and decision making.

* Transactional databases :-

It is a collection of data organized by time stamps, date, etc to represent transaction in databases.

* multimedia databases :-

multimedia databases consits audio, video, images and text media.

2) DM processes :-

Cleaning, integration & selection

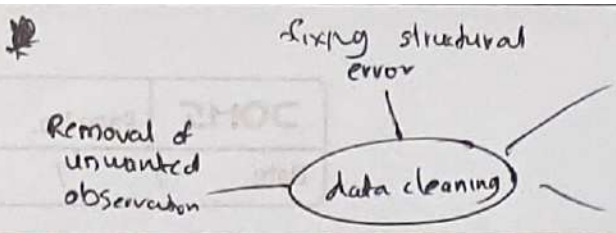
* Before passing the data to the database or data warehouse server, the data must be cleaned, integrated and selected.

Data cleaning

* As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate.
* so the first data requires to be cleaned and unified.

Data integration →

* Data integration in data mining refers to the process of combing data from multiple sources into single, unified



DOMS | Page No.
Date / /

view. after integrating the data we need to select the relevant data. that process is called data selection

3) Database or Data Warehouse server :-

* The database or data warehouse server consists of the orginal data that is ready to be processed.
* Hence the server is caused for retrieving the relevant data that is based on data mining as per user request.

4) Data mining engine :-

* The data mining engine is a major component of any data mining system.
* It contains several modules for operating data mining tasks, including association, classification, clustering, prediction, time-series analysis etc.
* It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

5) Pattern evaluation module :-

grouped data since one particular result

* The pattern evaluation module is primarly responsible for the measure of investigation of the pattern by using a threshold value.

DOMS | Page No.
Date / /

* It collaborates with the data mining engine to focus the search on exciting patterns.
* The pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used.
* for efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

6) Graphical user interface :-
* The graphical user interface (GUI) module communicates b/w the data mining system and the user.
* This module helps the user to easily and efficiently use the system without knowing the complexity of the process.
* This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

7) Knowledge Base :-
* The knowledge base is helpful in the entire process of data mining.
* It might be helpful to guide the search or evaluate

DOMS | Page No.
Date / /

Using threshold value wanted data is identified and used remaining data is send back to knowledge base.

the stake of the result patterns the knowledge base may even contain user views and data from user experences that might be helpful in the data mining process.
The data mining engine may recieve inputs from the knowledge base to make the result more accurate and reliable.

Data source.

⇒ spatial database:-
store geographical information store data in the form of coordinate, topology, line, polygon etc.

⇒ KDD :- (knowledge discovery in database)
KDD is a process involves the extraction of useful, previously unknown and potentially valuable information from large datasets.
It is a field of intrest to researchers in various fields, including-
- AI
- Meaching learning
- pattern recognization
- Database
- knowldge aquisation for expert systems
- data visualization.

• The KDD process is an iterative process and it

requires multiple iterations of the various steps to extract accurate knowledge from the data.

steps:-

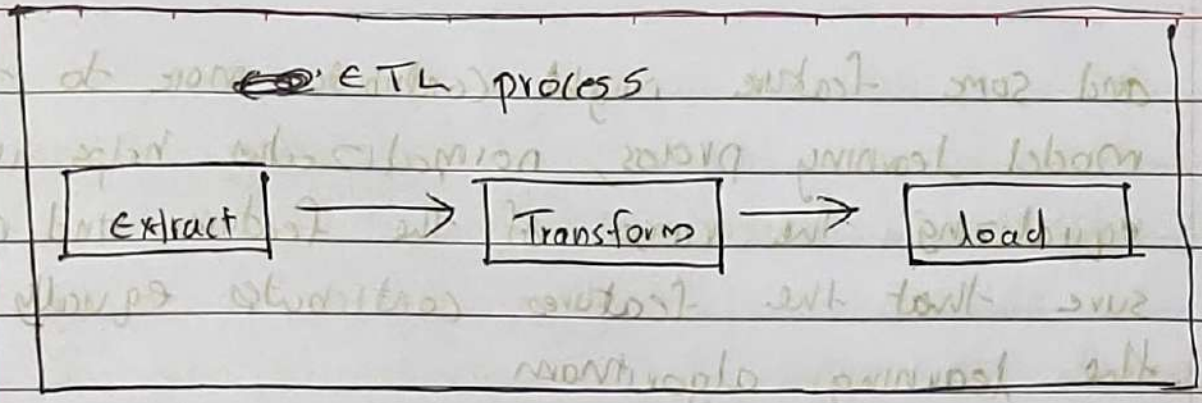
1) Data cleaning:-

* Data cleaning is also known as data cleansing or data scrubbing, is the process of identifying and correcting (or removing) errors, inconsistencies and inaccuracies within a dataset.
* It can be done by filling missing value and smoothing noisy (unwanted data) data, analyzing and removing outliers and removing inconsistencies in the data.
* The inconsistency can be recorded in various transactions during data entry or arising from integrating information from multiple database.

2) Data Integration:-

process of combing data from multiple sources to create a unified dataset.

* The hetrogeneous data from multiple sources are combined to a common source.
* It is also known as ETL (Extract - Transform - Load) process.



iii) Data selection :-

It is defined as the process where data relevant to the analysis is decided and retrived from the data collection.

iv) Data transformation :-

It is defind as the process of transforming data into appropriate form required by mining procedure

* The common method used for transformation are :-

i) discretization
ii) standardization
iii) normalization.

→ normalization :-

- It is used to rescale the features to a standard range of values which is usually 0-1.
- ~~It is used to~~ Normalization is usually used when different features have different range of values.

DOMS | Page No.
Date / /

and some feature might contribute more to the model learning process, normalization helps in equalizing the range of the features and makes sure that the features contributes equally to the learning algorithm.

* minimum and maximum value of features are used for scaling.

iv) ⇒ Standardization :-

It is used to transform the data to standardized format.

* Mean and standard deviation is used for scaling.

v) ⇒ Discretization :-

It is used to convert the a huge number of data values into smaller ones so that the evaluation and management of data become easy.

vi) ⇒ Data mining :-

It is defined as techniques that are applied to extract patterns potentially useful.

* It transforms task relevant data into patterns, and decides purpose of model using classification or characterization.

vi) ⇒ **pattern evaluation** :-

It is the process of assessing the quality of discovered patterns.

* This process is important in order to determine whether the patterns are useful and whether that can be trusted.

vii) ⇒ **knowledge representation** :-

is the presentation of knowledge to the user for visualization in terms of trees, tables, rules, graphs, charts, matrices, etc.

* It is a place where visualization and knowledge representation techniques are used to present mined knowledge to users.

⇒ KDD vs Data mining :-

* data mining is a step with in KDD process.

KDD	data mining
* refers to a process of identifying valid, novel, potentially useful and ultimately understandable patterns and relationships in data	* Data mining refers to a process of extracting useful and valuable information or patterns from large data sets.
* Technologies used: Data	* Technologies used:

DOMS | Page No.
Date / /

cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation and visualization.

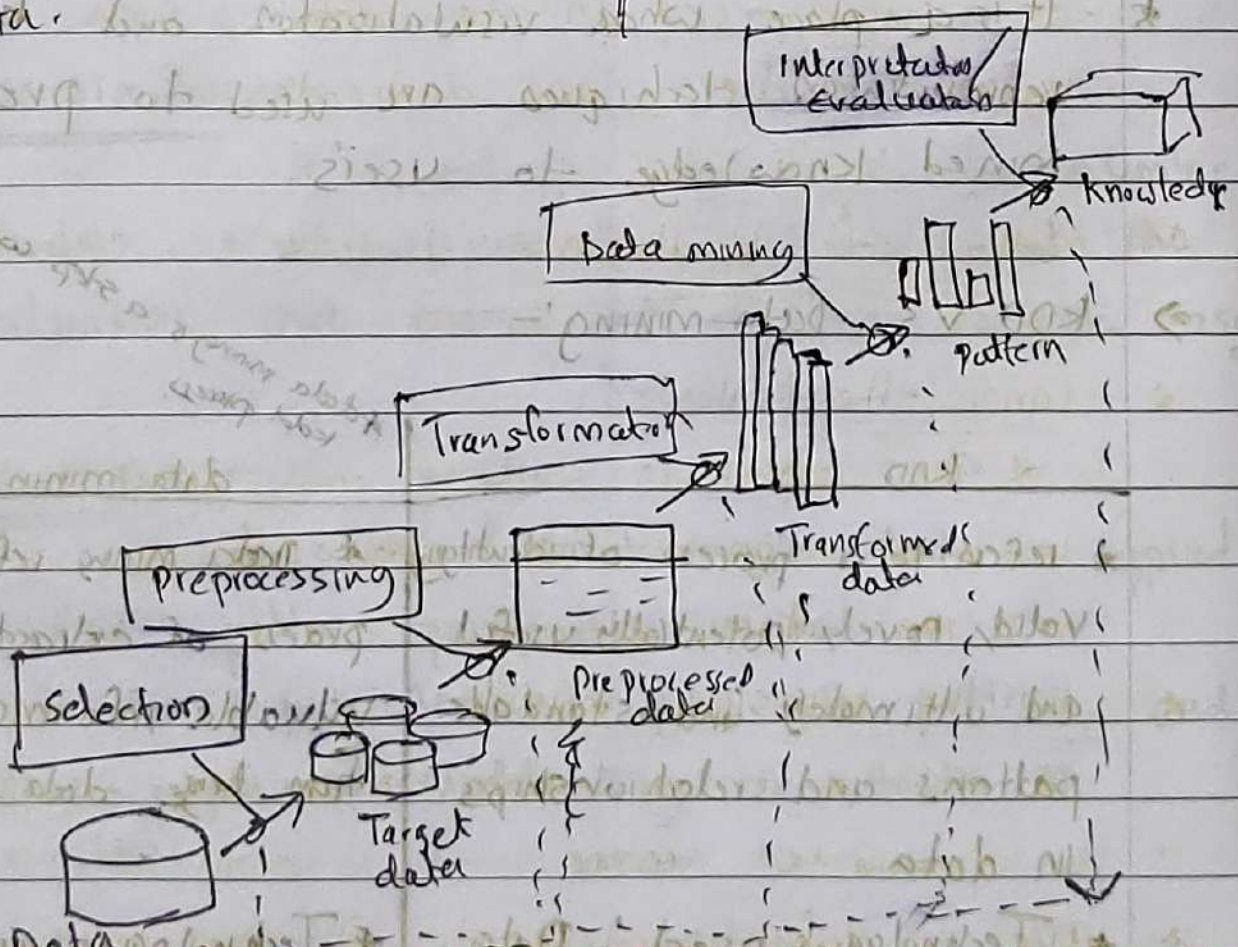
* Output :- structured information such as rules and models, that can be used to make decisions or predictions

* Focus is on the discovery of useful knowledge, rather than simply finding patterns in data.

Association rules, classification, clustering, regression, decision tree, neural networks, and dimensionality reduction.

* output :- patterns, associations, or insights that can be used to improve decision-making or understanding.

* Data mining focus is on the discovery of pattern or relationship in data.



DOMS | Page No.
Date / /

⇒ Data preprocessing :-

essay

It is the process of converting raw data into an understandable format.

Steps :-

1) Data cleaning
2) Data integration
3) Data transformation
4) Data reduction (selection)

⇒ Data cleaning :-

This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates.

⇒ Methods for handling missing data :-

* Ignore the tuples :-
  * This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
* Fill the missing value :-
  * Fill the missing values manually or by the most probable value.
* Use a global constant to fill the missing value :-

DOMS | Page No.
Date / /

* In this replace all the missing attribute by a same constant.
* use a measure of central tendency for the attribute to fill the missing value.
* In this replace all the missing attribute by central tendency values such as mean or median.

⇒ Methods for handling Noisy data:-

* Noisy data is a meaningless data that can't be interpreted by machines.
* It can be generated due to faulty data collection, data entry error etc.
* Steps —
  - Binning
  - Regression
  - clustering.

⇒ Binning:
The data is divided into several segment of equal size. After that, the different methods are executed to complete the task.

⇒ regression:
Data can be made smooth by fitting it to a regression function. The regression used may be

linear (having one independent variable) or multiple (having multiple independent variables).

⇒ clustering (outlier analysis):-

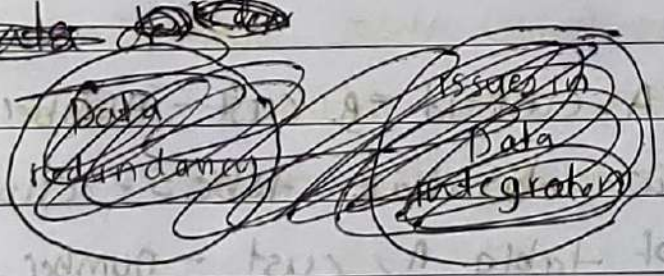
clustering group the data in a cluster. Then the outliers are detected with the help of clustering

⇒ Data Integration:-

It is data preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of the data

* The process involves identifying and accessing the different data sources, Mapping the data to a common format.

⇒ Issues in data integration:-



* Data redundancy:-

redundant data occurs while we merge data from multiple databases

DOMS | Page No.
Date / /

* If the redundant data is not removed, incorrect results will be obtained during data analysis.

⇒ Duplicate data attributes:- (multiple columns same type)
Identify the duplicate data attributes and remove those attributes.

⇒ Irrelevant attributes:-
Some attributes in the data are not important, and they are not considered while performing the data mining task.

⇒ Entity Identification problem:-
Integration of data from multiple resources, some data resources match each other and they will become redundant if they are integrated.

* for example - A.cust-id = B.cust-number. Here A, B are two different database tables. cust-id is the attribute of table A, cust-number is the attributes of table B.

* Here cust-id and cust-number are attributes of different tables and there is no relationship

b/w these tablets but the cust-id attributes and cust-number attribute are taking the same values.

* It helps in detecting and resolving data value conflicts.

⇒ Data transformation:-

• This involves converting the data into a suitable format for analysis.

• Common techniques used in data transformation include. The different method involves:-

→ Normalization

→ Standardization

→ Discretization

⇒ methods for data normalisation:-

• Decimal scaling

• Min-max Normalization

• z-Score Normalization

• (zero-mean normalization)

⇒ Data reduction:-

data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information

DOMS | Page No.
Date / /

- This is done to improve the efficency of data analysis and to avoid over fitting of the model.
- Data cube aggregation
- Attributes subset selection
- Dimensionality reduction.

⇒ Data cube aggregation :-

The technique is used to aggregate data in a simpler form.

- Data cube aggegation is a multi-dimensional ~~aggregat~~ aggregation that uses aggregation at various levels of a data cube to represent the orginal data set, thus achieving data reduction

⇒ Attribute Subset Selection :-

- The large data set has many attributes, some of which are irrelevant to data mining.

* The core attribute subset selection reduces the data volume by eliminating redundant and irrelevant attributes
* The attribute subset selector ensures that we get a good subset of orginal attributes even after eliminating the unwanted attributes.

DOMS | Page No.
Date / /

⇒ Attrib

* Ste

set

(Cha

the

of observations. Whenever we
median value. There are two

)/2 -th observation
= th observation

$n + (n/2) + 1)^{th}$ obsrvat

added to a reduced set.

* ~~Step wise~~ backward elimination :-
  - Here all the attributes are considered in the initial set of attributes
  - In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than significance level

* Combination of forward selection and backward Elimination :-
  - The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently.
  - This is the most common technique which is generally used for attribute selection.

DOMS | Page No.
Date / /

⇒ Attributes subset selection methods :-

* Stepwise forward selection :-

• This procedure start with an empty set of attributes as the minimal set.

• The most relevant attributes are choosen (having minimum p-value) and are added to the minimal set.

• In each iteration, one attribute is added to a reduced set.

* ~~Step wise~~ backward elimination :-

• Here all the attributes are considered in the initial set of attributes

• In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than significance level.

* Combination of forward selection and backward Elimination :-

• The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently.

• This is the most common technique which is generally used for attribute selection.

DOMS | Page No.
Date / /

⇒ Dimensionality reduction :-

The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction. It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information.

⇒ Measures of central tendency :-

* mean
* median
* mode
* mid range
* range

⇒ mean :-

mean of data = sum of observations / no. of observations

⇒ median :-

median means to identify the middle value

DOMS | Page No.
Date / /

from a set of observations. Whenever we need to calculate median value there are two conditions,

when n is odd,

median = $(n+1)/2$ th observation

eg- $9 // = \frac{9+1}{2} = \frac{10}{2} = 5^{th}$ observation //

when n is even,

median = $\frac{[(n/2)\text{th observation} + (n/2)+1)\text{th observation}]}{2}$

⇒ Mode:-

most frequently occuring value in the data set.

eg- 2 3 3 3 7 7 9 14 15

here mode is 3//

Q Find the mean of first 10 odd integers.

odd = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19

mean = $\frac{\text{sum of observation}}{\text{no. of observation}}$

$= \frac{100}{10} = 10//$

⇒ **mid-range**:-

It is defined as the sum of maximum and minimum value divided by 2.

$$\text{mid range} = \frac{\max + \min}{2}$$

⇒ **range**:-

It is the difference b/w the maximum and minimum values of observation in the data.

range = maximum value in data − min value in data

⇒ **Variance**:-

variance of data is given by measuring the distance of observed values from the mean of distribution.

$$\sigma^2 = \sum_{i=0}^{n} \frac{(x_i - \bar{x})^2}{n}$$

⇒ **standard deviation**:-

Square root of variance is called standard deviation.

$$\sigma = \sqrt{\frac{\sum_{i=0}^{n} (x_i - \bar{x})^2}{n}}$$

DOMS | Page No.
Date / /

Q What is the median of following data set?
32, 6, 21, 10, 8, 11, 12, 36, 17, 16, 15, 18, 40, 24, 21, 23, 24, 24, 29, 16, 32, 31, 10, 30, 35, 32, 18, 39, 12, 20

Find the mid-range and range of the data set.

Step 1 first to ascending order.
6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40

no. of values in data set $\Rightarrow n = 30$

$$\text{median} = \frac{(n/2)^{th} + ((n/2)+1)^{th}}{2}$$

$$= \frac{21 + 21}{2} = 21$$

$$\text{mid range} = \frac{\max + \min}{2} = \frac{40 + 6}{2} = \frac{46}{2} = 23$$

$$\text{range} = \max - \min = 40 - 6 = 34$$

DOMS | Page No.
Date / /

Module - 2

⇒ **Database:-**

It is a collection of data and stores data in tables. It deals with operational or transactional data It can store MB and GB of data. A student db can store details of students and retrive information based on user query. used for OLTP (online Transcation process).

⇒ **Data warehouse:-**

Store huge amounts of data. Data collected from multiple hetrogeneous sources like files, DBMS etc. It is mainly use to store historical data. Can store TB's of data. eg- how the placement of cs students has improved over the last 10 years. It is used for OLAP (online analytical processing). A data warehouse is a system that stores data from a company's operational databases as well as external sources.

- Data warehouse platforms are diffrent from operational databases because they store historical information, making it easier for business leaders to analyze data over a specific period of time.

DOMS | Page No.
Date / /

Module - 2

⇒ **Database:-**

It is a collection of data and stores data in tables. It deals with operational or transactional data. It can stores MB and GB of data. A student db can store details of students and retrive information based on user query. used for OLTP (online Transcation process).

⇒ **Data warehouse:-**

Store huge amounts of data. Data collected from multiple hetrogeneous sources like files, DBMS etc. It is mainly use to store historical data. Can store TB's of data eg- how the placement of cs students has improved over the last 10 years. It is used for OLAP (online analytical processing). A data warehouse is a system that stores data from a company's operational databases as well as external sources.

- Data warehouse platforms are differnt from operational databases because they store historical information, making it easier for business leaders to analyze data over a specific period of time.

DOMS | Page No.
Date / /

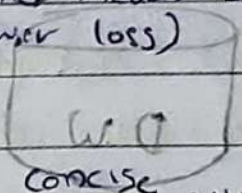
father of data ~~science~~ warehouse

- According to bill Inmon, "A warehouse is a subject oriented, integrated, time, variant and non-volatile collection of data in support of managements decision making process.

⇒ Characteristics of data warehouse :-

- Subject oriented
- Integrated
- non volatile
- Time variant.

(volatile means data lose during power loss)



⇒ Subject oriented :-

DW provide a ~~concise~~ view around a particular subject such as customer, product or sales, instead of the global organization's ongoing operations.

- DW is always a subject-oriented one, as it provides information about a specific theme.

⇒ Integrated :-

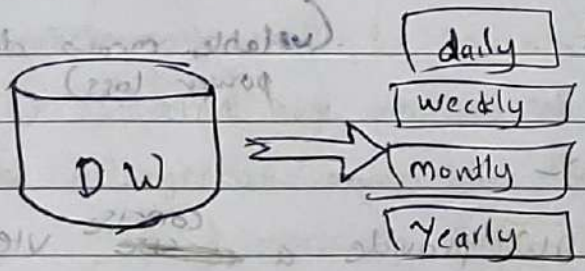
DW integrates data from various heterogeneous data sources like RDBMS, flat files, and online transaction records, and combines it in a relational database.

DOMS | Page No.
Date / /

- It requires performing data cleaning and integration during data warehousing to ensure consistency among different ~~dat~~ different data sources.
- It must be consistent, readable and coded.

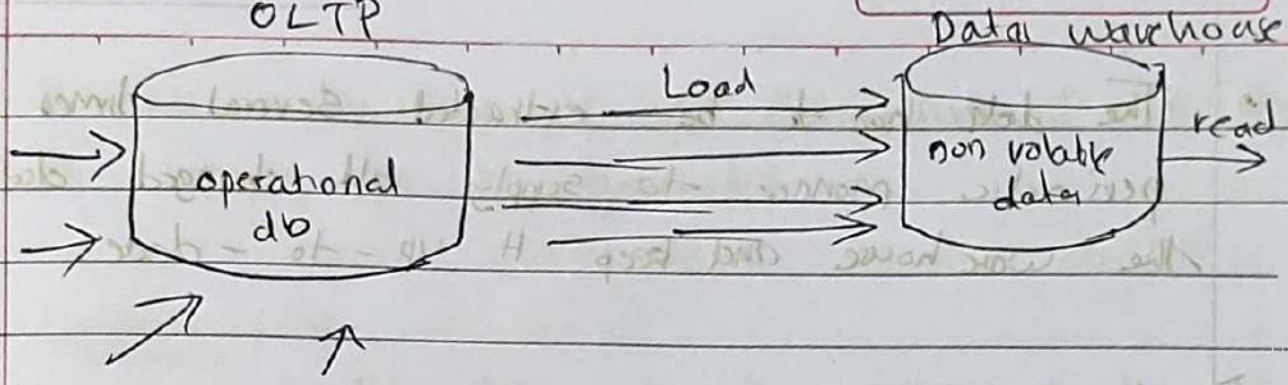
⇒ Time variant :-

- Historical information is kept in a data warehouse.
- For example, one can retrieve files from 3 months, 6 months, 12 months or even previous data from a data warehouse.



⇒ Non volatile :-

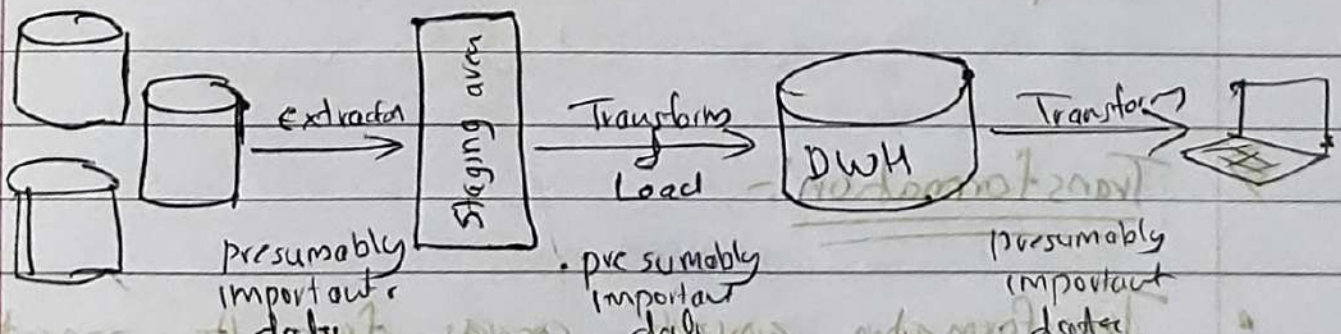
- The data residing in the data warehouse is ~~volatile~~ permanent
- It ensures that when new data is added, data is not erased or removed.
- A data warehouse is kept seprate from the operational db and thus the data warehouse does not represent regular changes in the operational database.



~~Select~~ select / insert / delete / update

⇒ ETL process (extract, transform and load):-

- The mechanism of extracting information from source systems and bringing it into the data warehouse is commonly called ETL.



⇒ Extraction:-

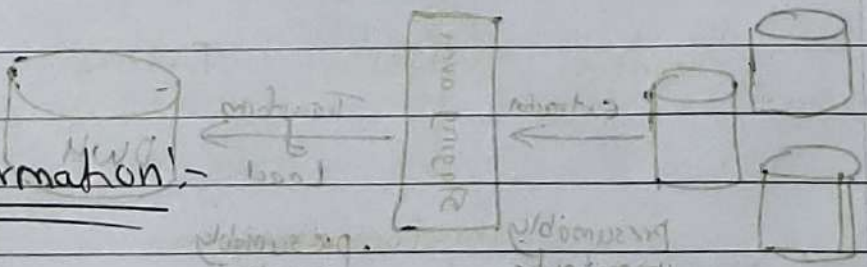
- Extraction is the operation of extracting information from a source system for further use in a data warehouse environment.
- This is the first stage of the ETL process.

DOMS | Page No.
Date / /

- The data has to be extracted several times in a periodic manner to supply all changed data to the warehouse and keep it up-to-date.

⇒ **Cleansing:-**

- The cleaning stage is crucial in a data warehouse technique because it is supposed to improve data quality.
- The primary data cleansing features found in ETL tools are rectification and homogenization.
- They use specific dictionaries to rectify typing mistakes and to recognize synonyms and defines appropriate associations b/w values.



⇒ **Transformation:-**

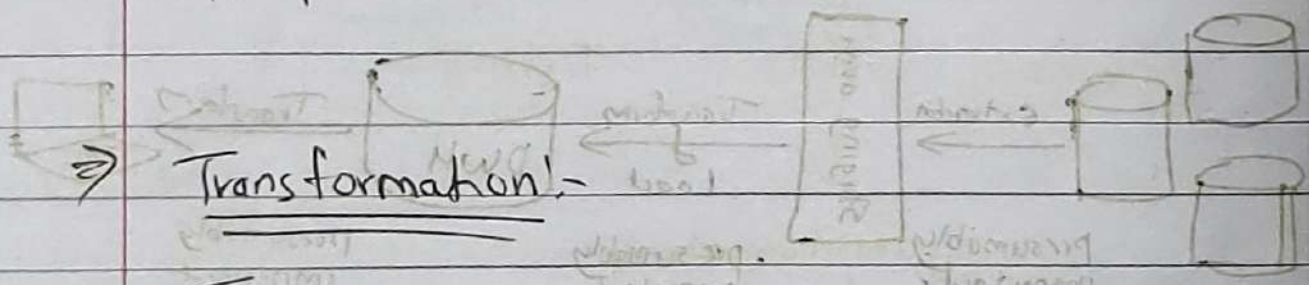
- Transformation converts records from its operational source format into a particular data warehouse format.
- ie, data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed.

DOMS | Page No.
Date / /

- The data has to be extracted several times in a periodic manner to supply all changed data to the warehouse and keep it up-to-date.

⇒ **Cleansing:-**

- The cleaning stage is crucial in a data warehouse technique because it is supposed to improve data quality.
- The primary data cleansing features found in ETL tools are rectification and homogenization.
- They use specific dictionaries to rectify typing mistakes and to recognize synonyms and defines appropriate associations b/w values.



⇒ **Transformation:-**

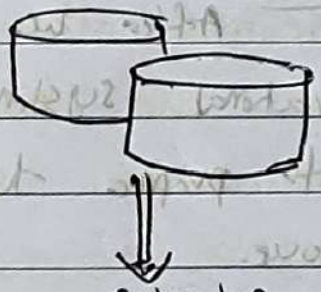
- Transformation converts records from its operational source format into a particular data warehouse format.
- ie, data extracted from source server is raw and not usable in its original form, therefore it needs to be cleansed, mapped and transformed.

- It adds value and changes data such that insightful BI reports can be generated.

⇒ **Loading :-**

Loading data into the ~~target~~ target data warehouse is the last step of the ETL process.

- Huge volume of data needs to be loaded in a relatively short period.
- Hence, load process should be optimized for performance



DOMS | Page No.
Date / /

⇒ Components or building blocks of data warehouse system

i) ⇒ Data source components :-

i) Internal data
ii) Archived data
iii) External data

ii) ⇒ data staging components :-

After we have been extracted data from various operational systems and external sources, we have to prepare the files for storing in the data warehouse.

i) Data extraction
ii) Data transformation
iii) Data Loading

iii) ⇒ data storage component :-

Data Storage is the component of the data warehouse that stores the data. Advantages of data storage in a data warehouse include the ability to store large amounts of data in a single location, fast and efficient data retrieval, and improved data quality due to data cleansing and standardization.

DOMS | Page No.
Date / /

⇒ Data warehouse architecture :-

Data warehouse architecture defines the arrangement of the data in different databases. The different architecture includes

⇒ Single tier architecture
⇒ Two tier architecture
⇒ Three tier architecture.

⇒ Single tier architecture :-

An operation system is a method used in data warehousing to process the day to day transcations of an organizations.

- A flat file system is a system of files in which transcational data is stored, and every file in the system must have a different name.
- Meta data summarizes necessary information about data which can be used to acces data more easily
- end-user access tools provides information to the business managers for strategic decision-making.
- The various end user tools includes :-
  * reporting and query tools
  * Application development tools.
  * Executive information system tools.
  * online analytical processing tools.
  * Data mining tools.

DOMS | Page No.
Date / /

- A single tier architecture help to minimize the amount of data stored to reach the goal, ie, it removes data redundencies.

⇒ Two tier architecture:-

⇒ Source layer:-

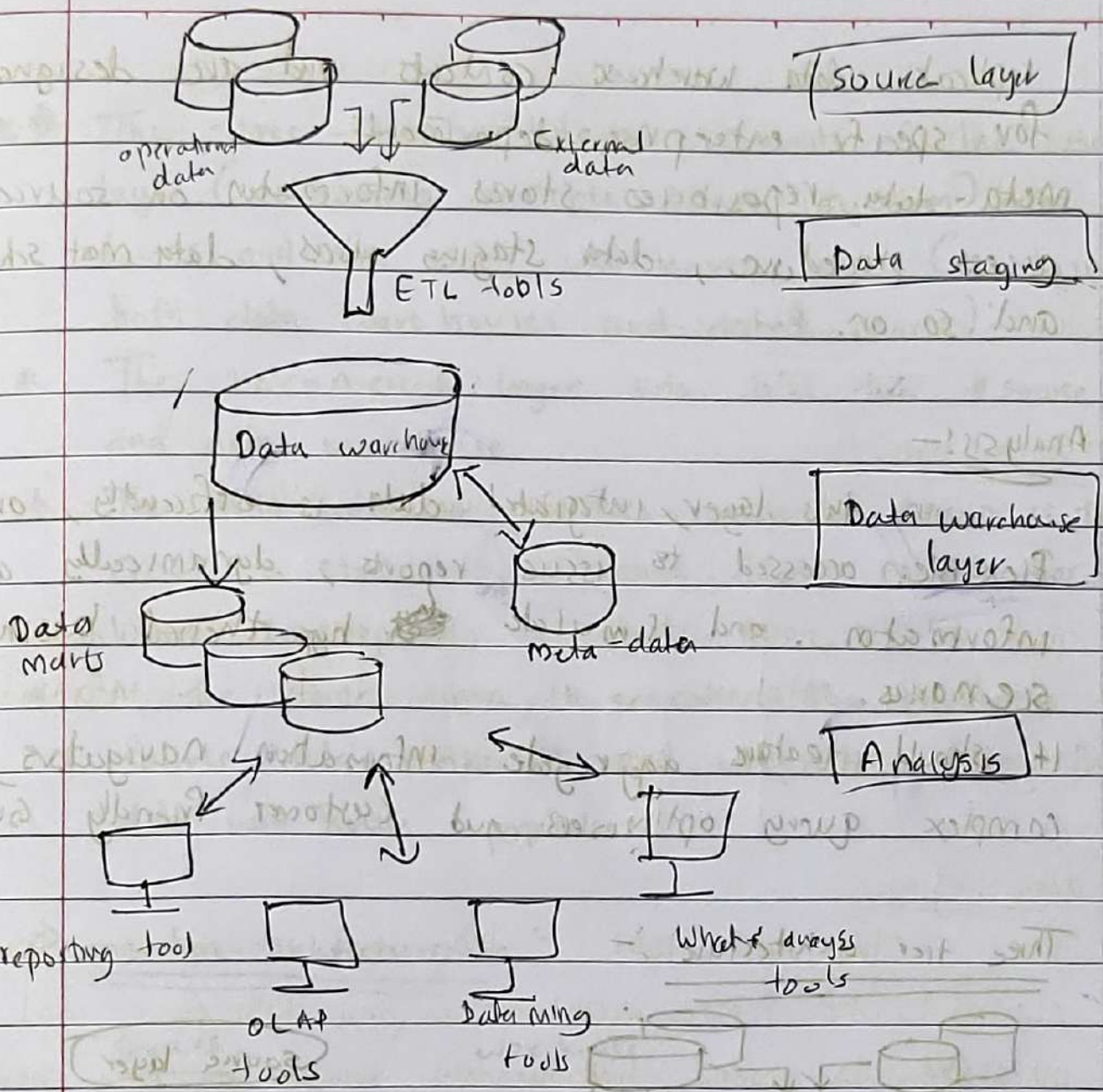
A data warhouse system uses a hchrogeneous source of data

- That data is stored initially to corporate relational databases or legacy databases, or it may come from an information system outside the corporate walls.

⇒ data staging:-

The data stored to the source should be extracted, cleansed to remove inconsistencies and fills gaps, and integrated to merge heterogeneous sources into one standard schema.

- Extraction, Transformation and loading Tools (ETL) can combine heterogeneous schemata, extract, transform, cleanse, validate, filter and load source data into a data warhouse.



=> Data warehouse layer:-

- Information is saved to centralized individual repository i.e, the data warehouse.
- The data warehouse can be directly accessed, but it can also be used as a source for creating data marts, ~~which~~ which, partially ~~dep~~

DOMS | Page No.
Date / /

replicate data warehouse contents and are designed for specific enterprise departments

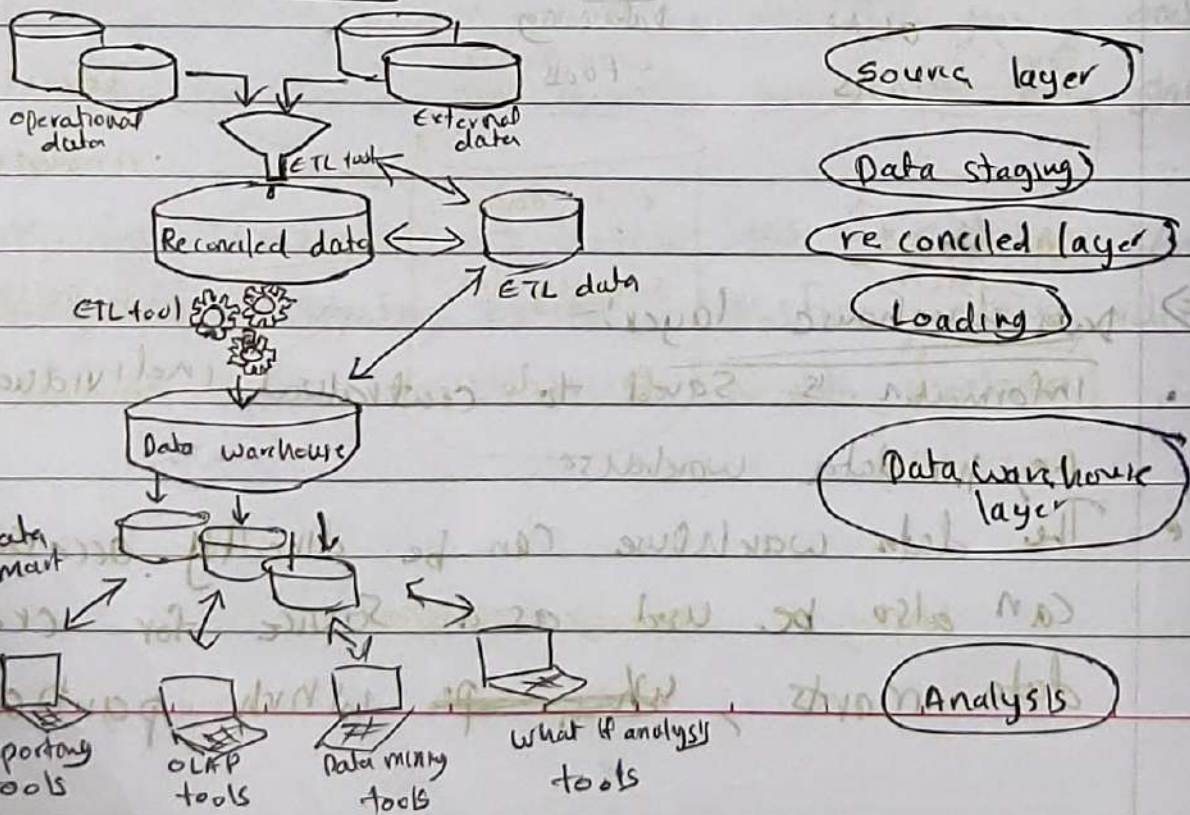
- Meta-data repositories stores information on sources, access procedure, data staging users, data mart schema and so on.

⇒ Analysis :-

In this layer, integrated data is efficiently, and flexible accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios.

- It should feature aggregate information navigators, complex query optimizers and customer friendly GUI's

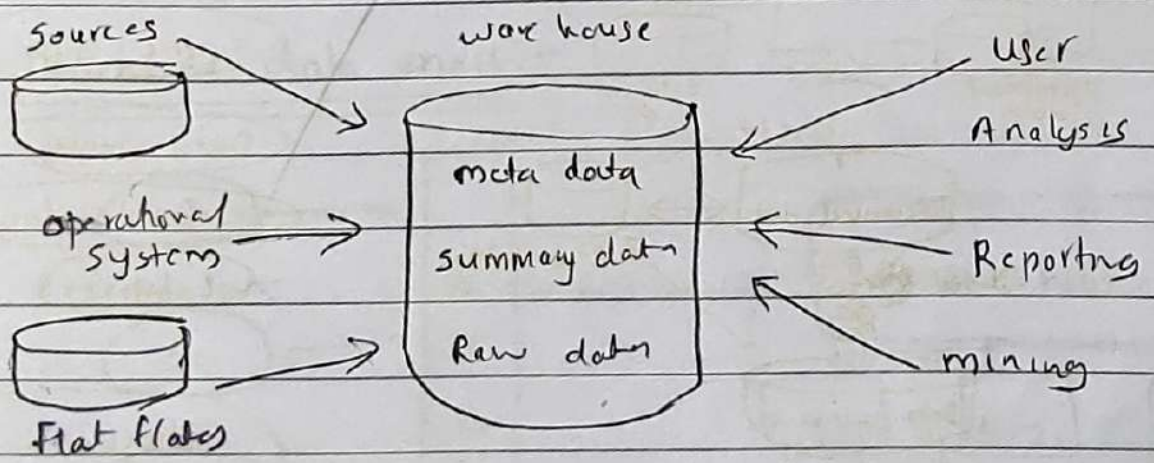
Three tier architecture :-



DOMS | Page No.
Date / /

* The three-tier architecture consits of the source layer (containing multiple source system), the reconciled layer and the data warehouse layer (containing both data warehouses and data marts):
* The reconciled layer sits b/w the source data and data warehouse.
* The main advantage of the reconciled layer is that it create a standard reference data model for a whole enterprise.
* At the same time, it seperates the problems of source data extraction and integration from those of data warehouse population

=> one tyre architecture :-



Input → Storage → o/p