classmate
Date
Page

DMBL

Subject Notes...

By :-

1) Hardin Shingala
2) Karan Raiyani

DOMS Page No.
Date / /

GTU - 5 TIMES

⇒ Apriori Algorithm.

Purpose :- It is an influential algorithm for mining frequent itemsets for boolean assocation rules.

Key concepts :- 1) Frequent Itemsets
2) Apriori Property
3) Join Operation

Example :-

TID	List of Items
T100	L1, L2, L5
T200	L2, L4
T300	L2, L3
T400	L1, L2, L4
T500	L1, L3
T600	L2, L3
T700	L1, L3
T800	L1, L2, L3, L5
T900	L1, L2, L3

Now, → Database D = 9
→ Min. support count required is 2
(∴ min_sup = 2/9 = 22%)
→ Let minimum confidence required is 70%

DOMS | Page No.
Date: / /

step 1:- For C1 For L1

Itemset	sup.count
{L1}	6
{L2}	7
{L3}	6
{L4}	2
{L5}	2

Itemset	sup.count
{L1}	6
{L2}	7
{L3}	6
{L4}	2
{L5}	2

step 2:- For C2 For L2

Itemset	sup.count
{L1, L2}	4
{L1, L3}	4
{L1, L4}	1
{L1, L5}	2
{L2, L3}	4
{L2, L4}	2
{L2, L5}	2
{L3, L4}	0
{L3, L5}	1
{L4, L5}	0

Itemset	sup.count
{L1, L2}	4
{L1, L3}	4
{L1, L5}	2
{L2, L3}	4
{L2, L4}	2
{L2, L5}	2

step 3:- For C3 For L3

Itemset	sup. count
{L1, L2, L3}	2
{L1, L2, L5}	2

Itemset	sup. count
{L1, L2, L3}	2
{L1, L2, L5}	2

step 4:- For C4, C4 = ϕ

step 5 :- Generating association rules.

Here, min. confidence threshold is say 70%.

so, from the formula.

R1 : $L1 \wedge L2 \rightarrow L5$ = $sc\{L1, L2, L5\} / sc\{L1, L2\} = 2/4 = 50\%$ X
R2 : $L1 \wedge L5 \rightarrow L2$ = $sc\{L1, L2, L5\} / sc\{L1, L5\} = 2/2 = 100\%$ ✓
R3 : $L2 \wedge L5 \rightarrow L1$ = $sc\{L2, L5, L1\} / sc\{L2, L5\} = 2/2 = 100\%$ ✓
R4 : $L1 \rightarrow L2 \wedge L5$ = $sc\{L1, L2, L5\} / sc\{L1\} = 2/6 = 33\%$ X
R5 : $L2 \rightarrow L1 \wedge L5$ = $sc\{L1, L2, L5\} / sc\{L2\} = 2/7 = 29\%$ X
R6 : $L5 \rightarrow L1 \wedge L2$ = $sc\{L1, L2, L5\} / sc\{L5\} = 2/7 = 100\%$ ✓

In this way we have three strong association rules, which are R2, R3 and R6.

Methods to improve Apriori's Efficiency :-

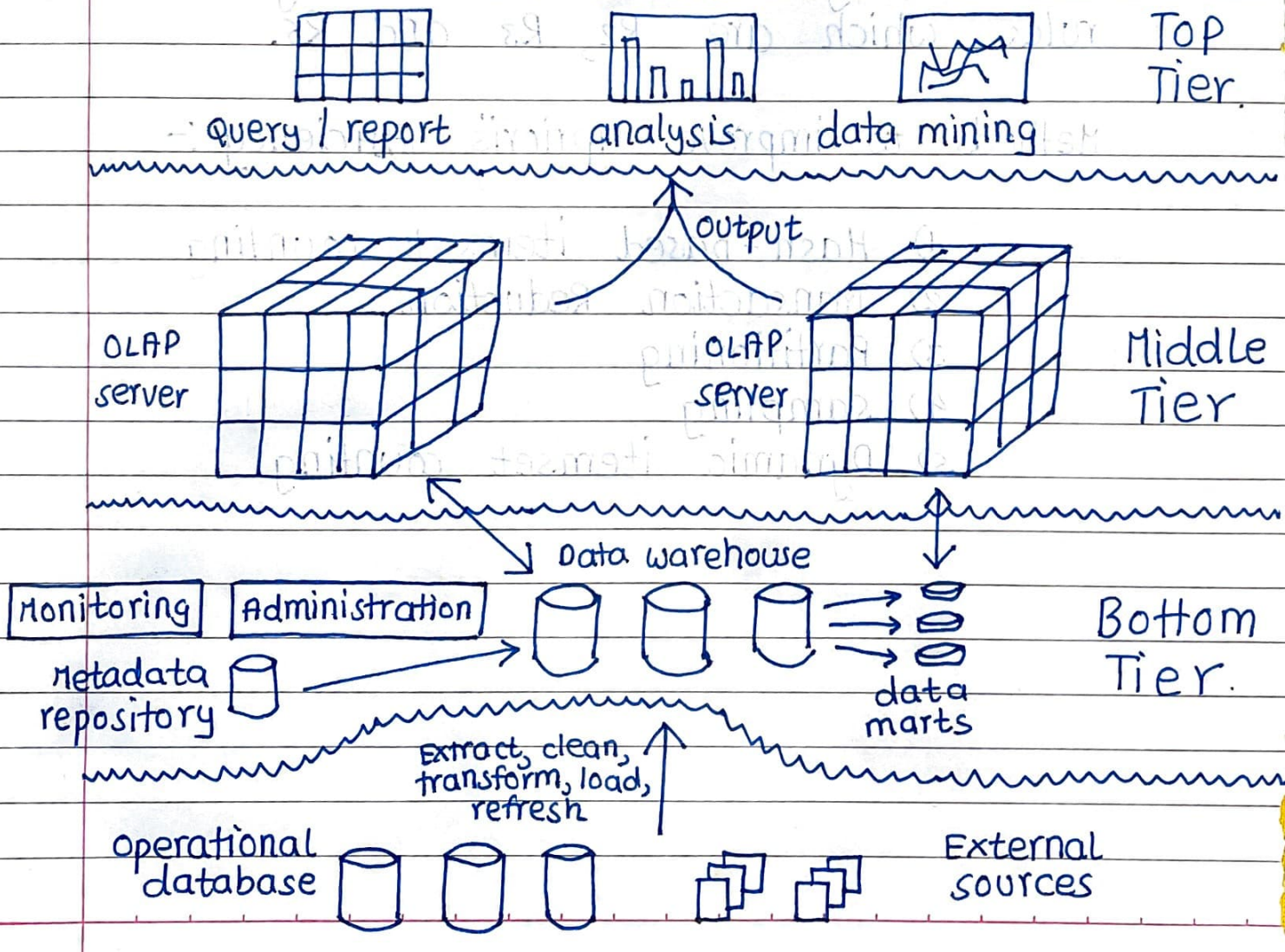
1) Hash-based itemset counting
2) Transaction Reduction
3) Partitioning
4) sampling
5) Dynamic itemset counting.

⇒ Data Warehouse architecture and its features.

Data warehouse :- it provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.

It refers to a database that is maintained separately from an organization's operational database.

Data warehouse architecture.



Bottom tier :- Data Warehouse server

It is a warehouse database server that is almost always a rational database system.

Back-end tools and utilities are used to feed data, the data are extracted using application program interface known as gateways.

Examples of gateway are ODBC, OLEDB and JDBC.

Middle tier :- OLAP servers

It is an olap server that is typically implemented using either :-

1) A relational OLAP (ROLAP) model, which maps operations on multidimensional data.

2) A multidimensional OLAP (MOLAP) model, which is a special-purpose server that directly implements multidimensional data.

Top tier :- Front-end tools

It is a front-end client layer, which contains query and reporting tools, analysis tools and/or data mining tools.

Features of data warehouse

1) Subject oriented - it is organized around major subjects such as customer, supplier, product or sales. It provides simple and consise view of a particular subject by excluding non-useful data.

2) Integrated - It is usually constructed by integrating multiple heterogeneous sources such as flat-files, etc. This technique is applied to ensure consistency.

3) Time variant - Data are stored to provide information from a historical perspective (eg. past 5-10 years).

4) Nonvolatile - It is always a physically separated store of data transformed from the application data found in the operational environment.

⇒ Define clustering, cluster analysis along with importance and application of clustering.

Clustering - cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

A cluster of data objects can be treated as one group.

clustering is a process of partitioning a set of data into a set of meaningful sub-classes called clusters.

Cluster Analysis :- It is a process in which, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.

Note :- the main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Importance of clustering in Data Mining.

1) Scalability - We need highly scalable clustering algorithms to deal with large databases.

2) Ability to deal with different kinds of attributes - algorithms should be capable to be applied on any kind of data.

3) Discovery of clusters with attribute shape - the clustering algorithm should be capable of detecting clusters of arbitary shape.

4) High dimensionality - they should not only be able to handle low-dimensional data but also the high-dimensional space as well.

5) Ability to deal with noisy data - database contains noisy, missing or erroneous data. some algorithms are sensitive to such data and may lead to poor quality clusters.

6) Interpretability - the results should be interpretable, comprehensible and usable.

- Applications of Clustering in data mining

1. Economic Science
2. WWW
3. Pattern Recognition
4. Spatial Data Analysis
5. Image Processing

GTU - 4 TIMES

Q Explain data mining with its features and importance.

→ Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems

Data mining is also known as knowledge discovery in data (KDD)

Features of data mining

- Prediction of likely outcomes
- Focus on large data sets and database
- Automatic pattern prediction based on behavio analysis
- Calculation - To calculate feature from other features

Importance of data mining

Data mining helps develop smart market decision, make predictions, analyze customer behavior and their insights which leads to great success and data driven business.

Main purpose of data mining is to manipulate huge amount of data existance or stored in databases by determing suitable variables that contribute to quality of prediction.

Q Differenciate between OLTP and OLAP

→

OLTP	OLAP
• Operational processing	Informational processing
• Transaction (occurs)	Analysis (occurs)
• Day-to-Day operations	Long term informational requirements, decision support.
• Read / write access	mostly read access
• No. of records accessed are in tens	No. of records accessed are in millions
• Database size 100 MB - GB	Data base size 100 GB - TB
• short, simple transactions	complex query
• High performance, High availability	High flexibility End-use autonomy
• Number of users are in thousands	Number of users are in hundreds
• Focuses on data in	Focuses on data out
Flat relational view	Multidimensional

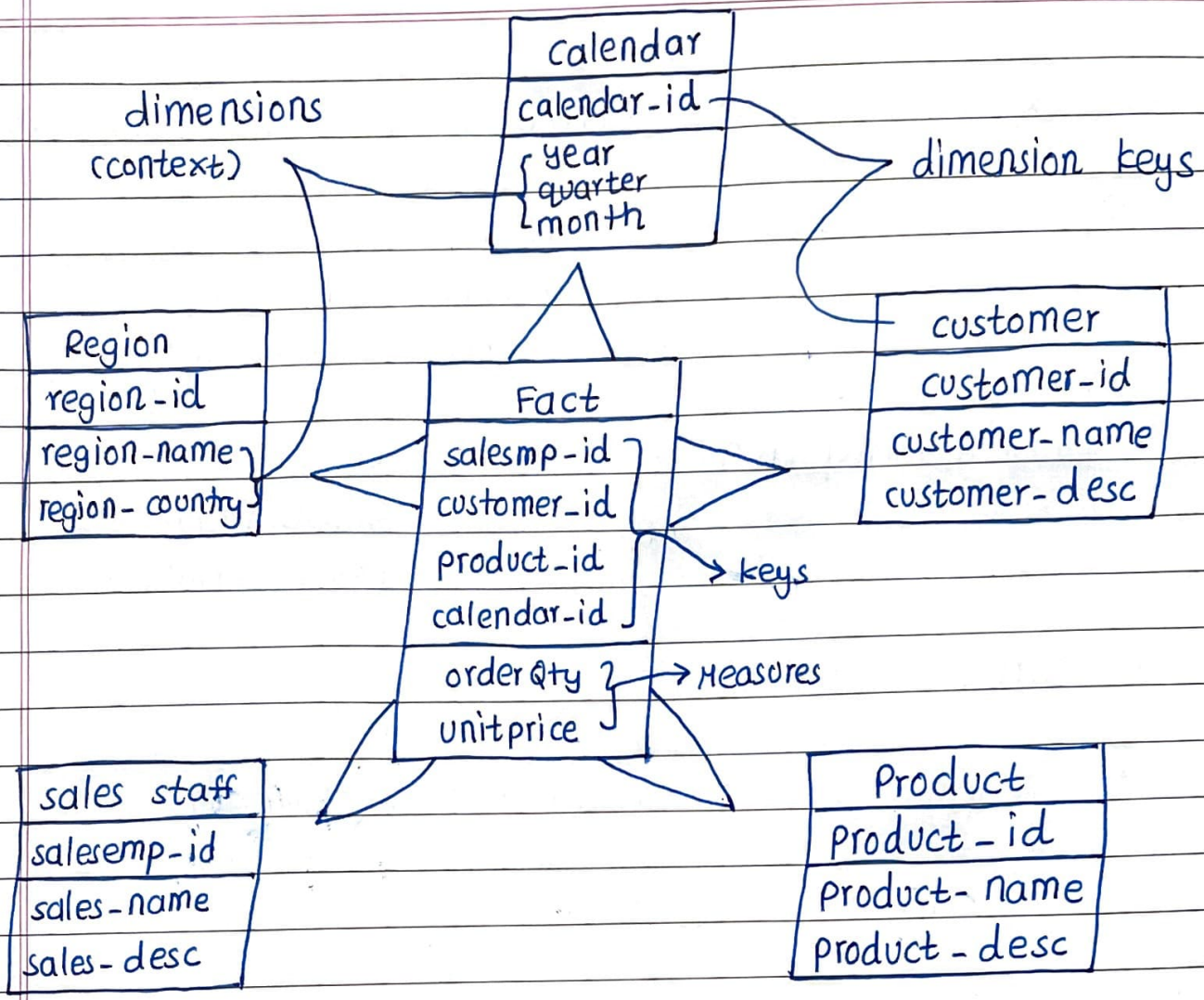
Q Explain various Data warehouse Schemas.

→ Data warehouse enviornment usually transforms the relational data model into some special architectures.

Commonly used schemas : STAR
SNOWFLAKE
FACT CONSTELLATION

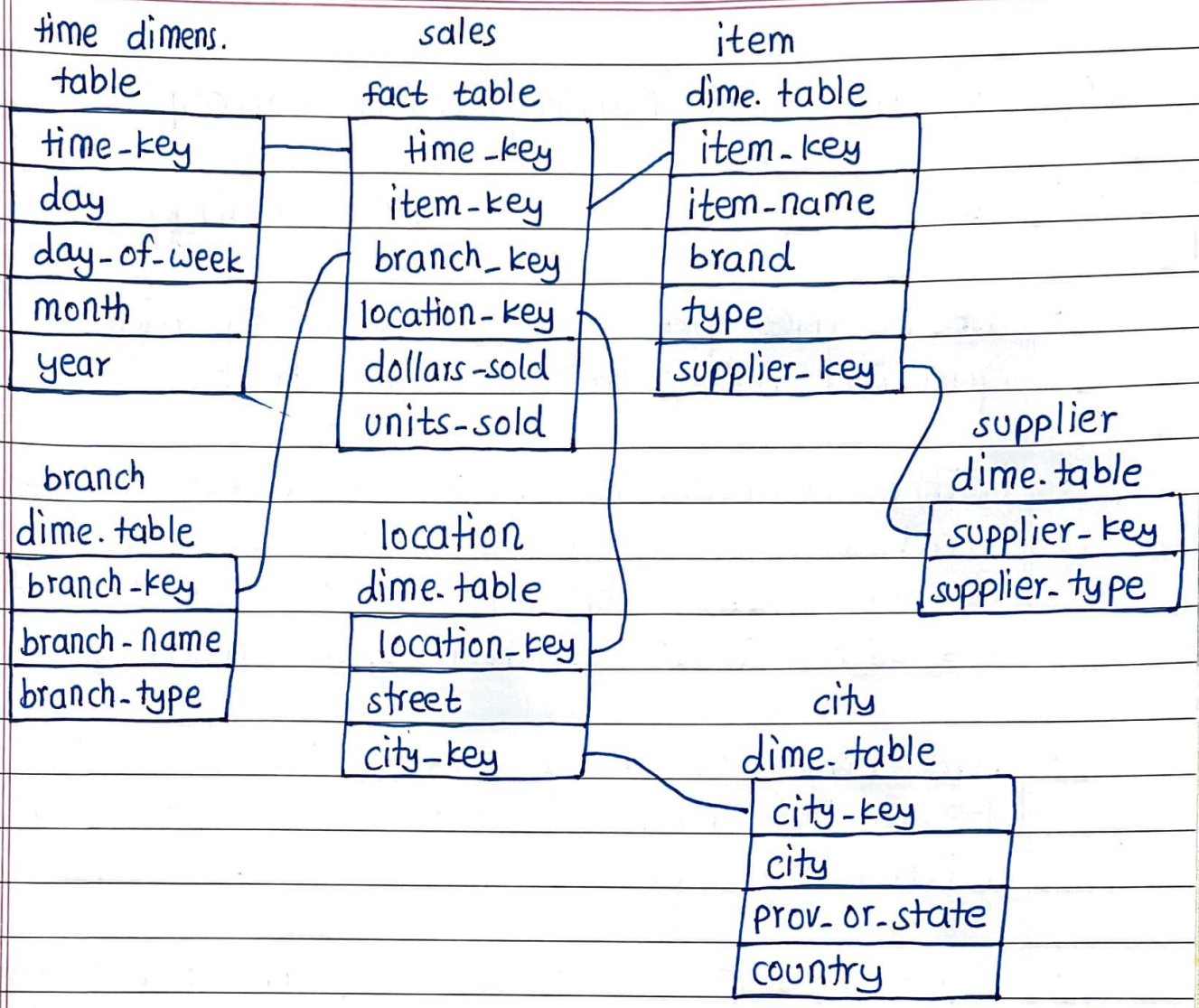
1. STAR SCHEMA

- Simplest data warehouse schema
- resembels a star
- Most commonly used now a days and recommended by oracle
- Centre of star consists - FACT TABLE
  Points of star are - DIMENSION TABLE
- FACT TABLES - 3NF
  DIMENSION TABLES - De-normalized form



2. SNOW FLAKE SCHEMA

- More complex variation of star schema
- Dimension tables are **normalized** here and hence tables are easy to maintain and saves storage space
- Reduces effectiveness of browsing as **more joins** are needed to execute a query
- snowflake schema reduces REDUNDANCY.



3. FACT CONSTELLATION SCHEMA

- Multiple fact tables
- Multiple fact tables share dimension tables
- Also called galaxy schema
  Most complicated schema design

- Ex - Dimension tables for time, item, location are shared between both sales and shipping

- Aggregation for variants is given importance and must be considered and selected

Q Topology of multilayer, feed-forward neural network.

→ Neural Networks

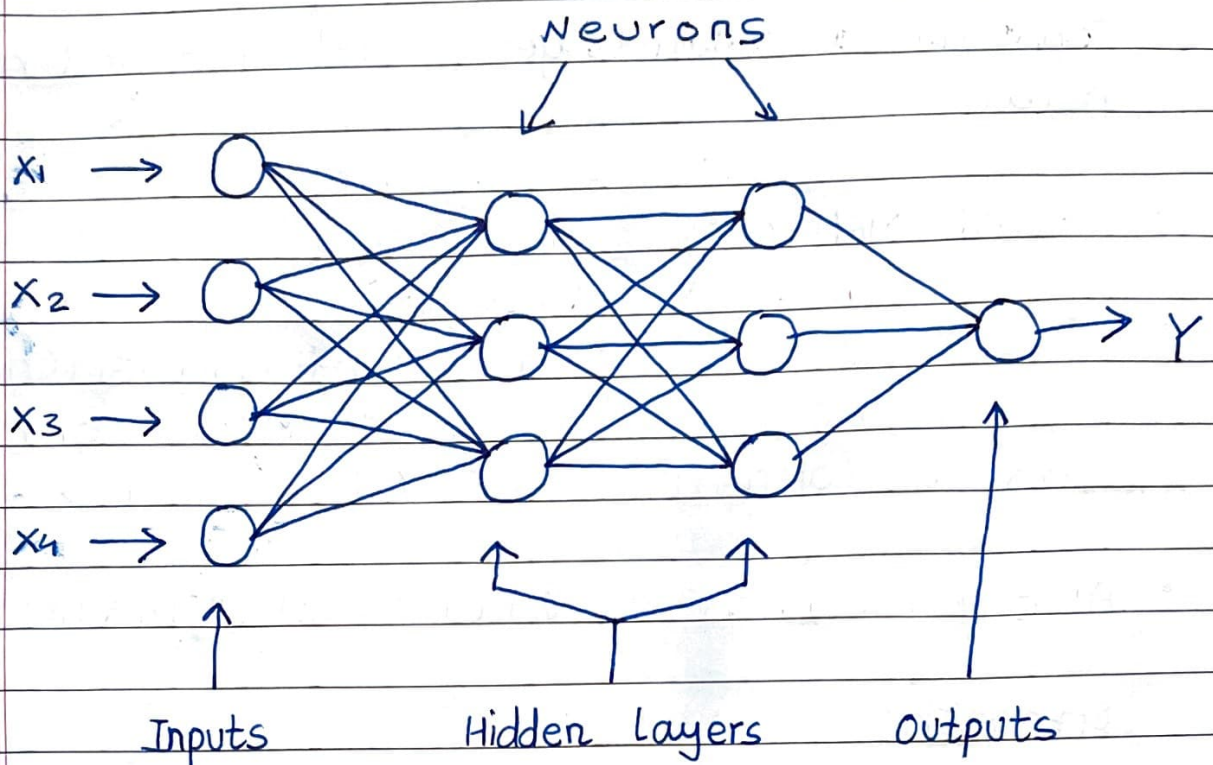
- Neural network is a set of connected input/output units where each connection has a WEIGHT associated with it
- Also known as CONNECTIONIST LEARNING

Strengths

- Can handle problems against complex data
- Can handle noise in data training
- Prediction accuracy is high
- Neural networks are robust
- Neural networks can handle missing data well

Weakness

- Neural network implementations are slower in training phase
- Acquired knowledge is difficult for humans to interpret
- Problem of overfitting of data
- Weak knowledge representation
- Construction of neural network can be a time consuming process



Note :- Hidden layers may vary according to the requirement. It may be 1, 2 or more then two.

Q Explain Text Mining with example.

→ Also refered as text data mining.

- It is the process of deriving high quality information from text.
- With advancement in technology more and more data is available in digital form, among them, most of the data (≒ 85%) is in unstructured textual form
- The kind of data stored in Data Bases, text is unstructured, ambiguous and difficult to process
- Text is the most communal way for the formal exchange of information.
- Hence TEXT mining helps to extract high quality or useful and intresting information from this large unstructured textual data.

Areas of TEXT MINING

1. Information Retrieval - Ex: Google web search Engine
2. Natural language Processing

Ex: AI, Study of human language, words etc

- Area's of application

1. Risk management
2. Knowledge management
3. Cybercrime prevention
4. Fraud detection
5. Business Intelligence

Q What is BIG DATA and BIG DATA ANALYTICS?

→ Big data is the term used for any data that is large in quantity.

- Refers to: any kind of data that is difficult to represent using conventional methods like DBMS etc.

- Big data describes any large amount of structured, semi-structured and unstructured data.

- Big data challanges: Capturing data
  data storage
  data analysis
  search
  sharing
  updating, querying
  vizualization
  information privacy.

classmate
Date
Page

- The 3 V's of BIG DATA

1. VOLUME (large amount of data produced

- Ex- FB ingests 700 TB of new data everyday
  - Boeing 737 will generate 240 TB of new data during a single long flight.

2. VELOCITY (Data moves at high speed)

- Ex- High frequency stock trading algorithms reflect market changes within micro seconds.
  - Clickstreams and Ad impressions capture user behavior at millions of events per second.

3. VARIETY (Different types of data)

- Geospatial Data
  3D data
  Audio & video
  Unstructure text
  Log Files
- Big data analysis includes different types of data.

→ BIG DATA ANALYTICS

- Big data analytics is the process of examining big data to uncover information such as : Hidden patterns
  Correlations
  Market trends
  Customer preferences

- Big data analytics help organization make informed business decisions.

- Big data analytics is a form of advanced analytics which involve complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by analytics systems.

- Importance

New revenue oppurtunities
Effective marketing
Better customer service
Improved operational efficiency
Competitive advantages over rivals

- Applications

Telecom
Traffic control
Trading Analytics
Health Care

- Tools for advanced analytics process

Data mining
↓
Predictive Analysis
↓
Machine learning
↓
Deep learning

GTU - 3 TIMES

Q Explain web Mining.

→ web mining is the use of data mining techniques to automatically discover and extract information from web documents and services.

General classes of information that can be discovered in web mining :

web activity
From server logs
web browser activity tracking

Main categories of web Mining:

web content Mining
web structure Mining
web Usage Mining

1. web content Mining

- Identify information within within given web pages. ex: text, image, records etc

2. web structure Mining

- use interconnections between web pages to give weight to the pages.

Ex: hyperlinks, tags etc

3. Web Usage Mining

- Understand access patterns and the trends to improve structure.

Eg: http logs
app servers logs

Q what is data smoothing and explain data smoothing by binning.

→ Data smoothing is the removal of noice from a data or eliminating outliers from data sets to make important patterns more noticable

Binning Method

In this method, the data is first sorted and then the sorted values are distributed into a number of bins or buckets.

As binning methods consult the neighborhood of values, they perform local smoothning

Types:

Smoothing by bin means
Smoothing by bin median
Smoothing by bin boundary

* STEPS for performing Binning

- Sort the array of given data set
- Divide the range into N intervals, each containing same number of samples
- Store mean / median / boundaries in each row

* Example

Sorted data set : 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

1. Smoothing by bin means

Bin 1	Bin 2	Bin 3
4, 8, 9, 15	15, 21, 21, 24	26, 28, 29, 34
↓	↓	↓
Mean = 9	Mean = 23	Mean = 29

Bin 1 : 9, 9, 9, 9
Bin 2 : 23, 23, 23, 23
Bin 3 : 29, 29, 29, 29

(2) Smoothing by bin boundaries

Bin 1 : Lower boundary - 4
Upper boundary - 15

4 is closer to 4 than 15
8 is closer to 4 than 15
9 is closer to 4 than 15
15 is closer to 15 than 4

Hence Bin 1 : 4, 4, 4, 15

REPEAT FOR BIN 2 & 3

Bin 1 : 4, 4, 4, 15
Bin 2 : 21, 21, 25, 25
Bin 3 : 26, 26, 26, 34

(3) Smoothing by Bin Median

Bin 1	Bin 2	Bin 3
4, 8, 9, 15	21, 21, 24, 25	26, 28, 29, 34
↓	↓	↓
median = 9	Median = ~~24~~ 24 ~~23~~	median = 29

median = middle value

Bin 1 : 9, 9, 9, 9
Bin 2 : ~~24, 24, 24, 24~~ ~~23, 23, 23~~ 24, 24, 24, 24
Bin 3 : ~~26~~ 29, 29, 29, 29

Q List various way to handle missing values in data mining.

- During data preprocessing due to noise, inconsistency and incompleteness of data leads to missing values.
- Various ways to handle missing values:

1. Ignore the data row

It is usually done when the class label is missing or many attributes are missing from the row.

2. Use a global constant to fill in for missing value.

Decide on a new global constant value, like "Unknown", "N/A" or minus infinity that will be used to fill all the missing values.

This method is used as a try and predict the missing value technique

Example: State of Residance can be filled with N/A
as can be derived from Pin / Postal / zip code

3. use attribute mean

Replace missing values of an attribute with the mean value for that attribute in the database

Use median for discrete value

Ex: For database of school average class result is X then it can be used to replace missing values for other classes of same grade.

4. Use attribute mean for all samples belonging to the same class.

Ex: In a database of luxury and low budget cars, replacing missing value of cost of luxury car with average cost of all luxury cars is more accurate

5. Use a data mining algorithm to predict the most probable value.

Use of regression
Bayesian formalism
decision trees
clustering algorithms

Q Explain Hadoop distributed file system

→ HDFS is the storage system of hadoop framework. It is a distributed file system that can conviениently run on commodity hardware for processing unstructured data.

HDFS is highly fault tolerant and designed using low-cost hardware.

HDFS holds a large amount of data and provides easier access.

HDFS makes applications available for ito parallel processing.

Features of HDFS

Suitable for distributed storage and processing

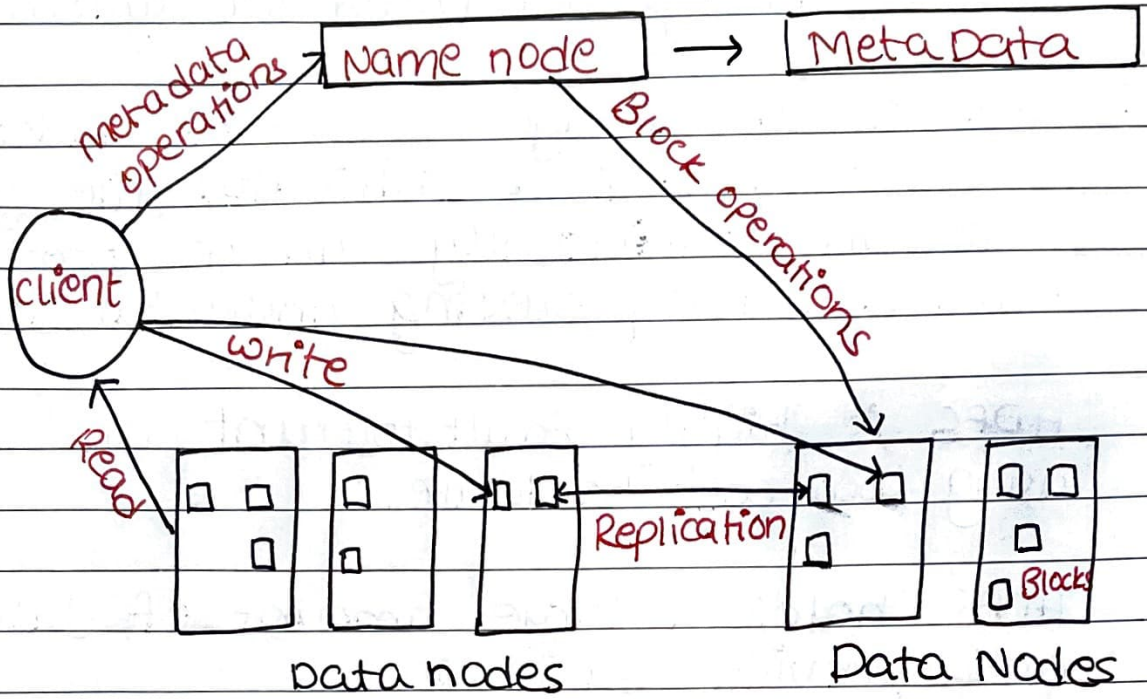
Hadoop provides a command interface to interact with HDFS

built in servers data node & Name node helps users check the status of cluster easily.

Streaming access to file system data.

HDFS provides file permissions and authentication.

HDFS architecture



The HDFS follows the master-slave architecture and has the following elements:

NAME NODE

commodity hardware that contains GNU/LINUX OS and name node software

System having name node acts as MASTER server.

tasks: Manages file system namespace

Regulates client's access to files

Executes file system operations like opening, closing files & directories

classmate
Date
Page

DATA NODE

commodity hardware having GNU/LINUX OS and data node software.

For every node (system/commodity hardware) in a cluster, there will be a data node.

These nodes manage the data storage of their system.

Tasks: Perform read/write operations on file systems, as per clients request.

Perform operations such as block creation, deletion, replication acc. to instructions of name node.

BLOCK

User data is stored in the files of HDFS.

The file in a file system will be divided into one or more segments and/or stored in individual data nodes.

These file segments are called blocks.

Minimum amount of data that HDFS can read/write is called a block default block size is 64 MB.

Goals of HDFS :

Fault detection & Recovery
Huge datasets
Hardware at data

Q Explain datamining application for fraud detection

Fraud :- In the technological systems, fraudulent activities have occurred in many areas of daily life such as telecommunication networks, mobile communications, online banking and E-commerce.

⇒ Fraud detection involves identifying fraud as quickly as possible once it has been perpetrated. Fraud detection methods are continuously developed to defend criminals in adapting to their strategies.

Types of fraud :-

1) Credit Card Fraud

2) Computer Intrusion

3) Telecommunication Fraud

Data Mining Techniques in Fraud Detection:-

1) Credit Card Fraud Detection:-

- → outlier detection.
- → Neural Networks

2) Computer Intrusion Detection:-

- → Expert Systems
- → Neural Network Intrusion Detector
- → Model-based Reasoning

3) Telecommunication Fraud Detection

- → Rule-based Approach
- → Neural Networks
- → Visualization Methods
- → Location awareness detector

GTU - 2 TIMES

⇒ Issues in Data Mining

Data mining is a dynamic and fast-expanding field with great strengths. Major issues in data mining research are listed below :-

1) Mining Methodology :-

- Mining various and new kind of knowledge.
- Mining knowledge in multidimensional space.
- Data mining - an interdisciplinary effort.
- Handling uncertainty, noise or incompleteness of data.

2) User Interaction :-

- Interactive mining.
- Incorporation of background knowledge.
- Presentation and visualization of data mining results.

3) Efficiency and Scalability :-

- Efficiency and scalability of data mining algorithms.
- Parallel, distributed and incremental mining algorithms.

4) Diversity of Database Types :-

- Handling complex types of data.
- Mining dynamic, networked and global data repositories.

5) Data Mining and Society :-

- Social impacts of data mining.
- Privacy-preserving data mining.
- Invisible data mining.

Many of these issues have been addressed in recent data mining research and development to a certain extend

These are now considered data mining requirements and others are still at research stage.

These issues continue to stimulate further investigation and improvement in data mining.

⇒ What is Regression. Explain linear regression and non-linear regression techniques of prediction.

Regression - It is a data mining function that predicts a number.

Age, weight, distance, temperature, income or sales could all be predicted using regression techniques.

→ Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values.

→ The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide.

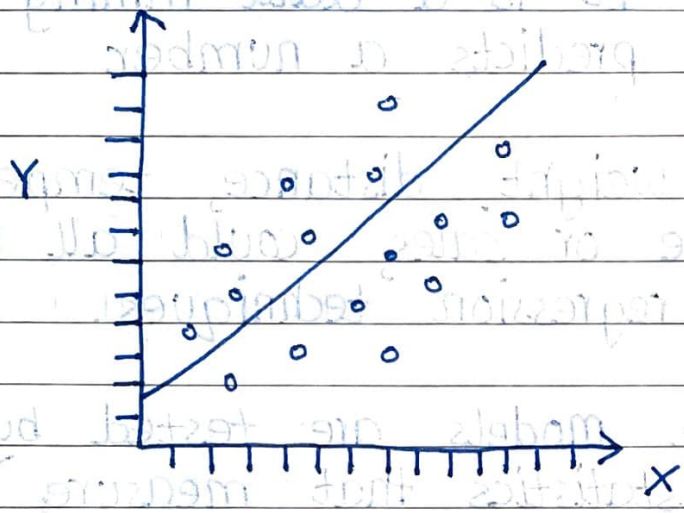
1) Linear Regression

It is the simplest form of regression to visualize with a single predictor.

This technique can be used if the relationship between x and y can be approximated with a straight line.

linear regression with a single predictor can be expressed with the given equation:-

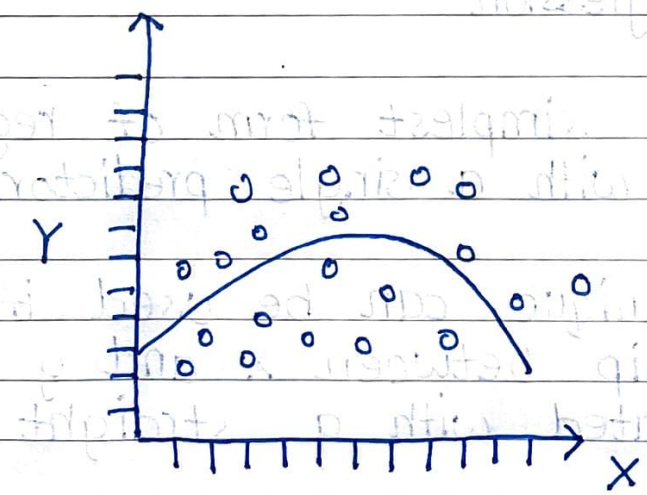
$$y = \theta 2x + \theta 1 + e$$



2) Nonlinear Regression

often the relationship between x and Y cannot be approximated with a straight line.

In this case, a nonlinear regression technique may be used. Alternatively, the data could be preprocessed to make the relationship linear.



⇒ Data Normalization in Data Mining

Normalization - It is used to scale the data of an attribute so that it falls in a smaller range.

It is generally useful for classification algorithms.

Need of normalization :-

It is generally required when we are dealing with attributes on a different scale, otherwise, it may lead to a dilution in effectiveness of an equally important attribute (on lower scale) because of other attributes having values on larger scale. So, they are normalized to bring all the attributes on the same scale.

Methods of Data Normalization :-

1) Decimal Scaling
2) Min - Max Normalization.
3) z-score or zero-mean normalization.

1) Decimal scaling method for normalization.

It normalizes by moving the decimal point of values of the data. To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data.

The data value v_i, of data is normalized to v_i' by using the formula below:-

$$v_i' = \frac{v_i}{10^j}$$

2) Min-max normalization.

Here, linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced according to the formula:-

$$v' = \frac{v - min(A)}{max(A) - min(A)} (new\_max(A) - new\_min(A)) + new\_min(A)$$

Here, A → attribute data
Min(A) → min absolute value of A
Max(A) → max absolute value of A
v' → new value of each entry in data
v → old value of each entry in data
new_max(A) → max value of the range
new_min(A) → min value of the range

3) Z-score normalization

Here, values are normalized based on mean and standard deviation of the data A.

The formula used is :-

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Here, v' → new value of each entry in data
v → old value of each entry in data
σ_A → standard deviation of A
$\bar{A}$ → Mean of A.

⇒ Naïve Bayesian Classification

Bayesian classifiers can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.

Bayesian classification is based on Bayes' theorem.

⇒ Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Naive Bayesian classifiers assume that the effect of an attribute value ~~of~~ on a given class is independent of the values of the other attributes.

This assumption is called class conditional independence. It is made to simplify the computations involved and in this sense it is considered "naive".

There are two types of probabilities :-

1) Posterior Probability

2) Prior Probability.

1) Posterior probability :-

$P(H|X)$ is the posterior probability of H conditioned on X.

2) Prior probability :-

$P(H)$ is the prior probability of H.

According to Bayes' Theorem

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)}$$

→ When it comes to effectiveness of Bayesian classifiers, they have the minimum error rate in comparison to all other classifiers.

⇒ Classification vs. Prediction

Classification	Prediction
• It is a process to identify which category a new observation belongs to.	It is the process of identifying the missing or unavailable numerical data.
• The accuracy depends on finding the class label correctly.	The accuracy depends on how well a given predictor can guess the value.
• A model or the classifier is constructed to find the categorical labels.	A model or a predictor will be constructed that predicts a continuous-valued function.
• The model can be known as the classifier.	The model can be known as the predictor.

Classification and Prediction Issues:-

1) Data Cleaning
2) Relevance Analysis
3) Data transformation and reduction
4) Normalization
5) Generalization.

⇒ Define the term "Data Mining". Explain KDD process with the help of suitable diagram.

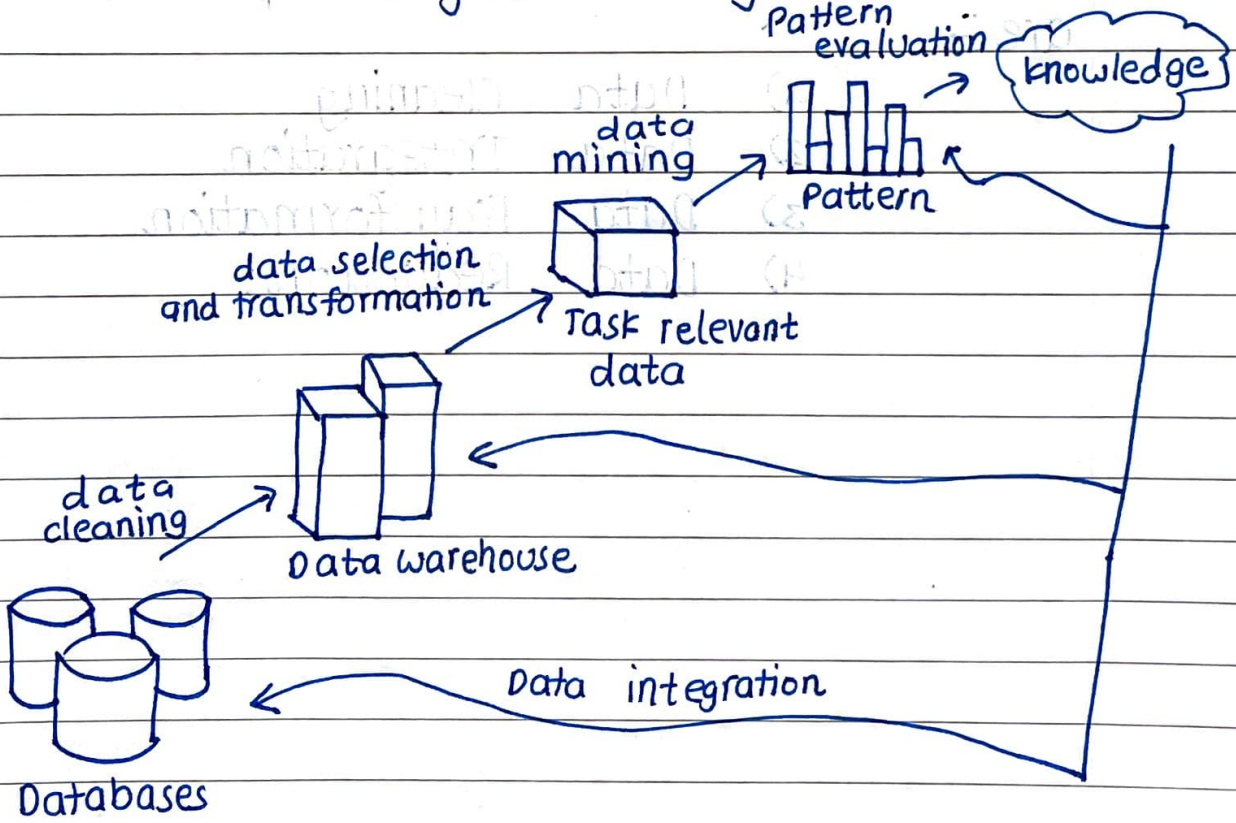
Data Mining :- it refers to extracting or mining knowledge from large amounts of data. It is also referred as knowledge mining from data.

Data mining system consist of :-

1) knowledge Base
2) Data mining engine
3) Pattern evaluation module

KDD Process

KDD → knowledge Discovery from Data



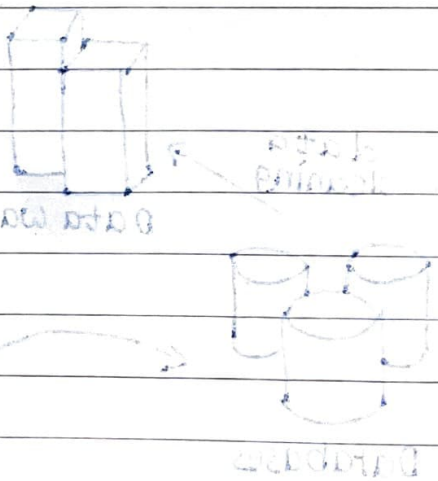
There are some pre-processing operations which are required to make pure data in data warehouse before using that data for Data Mining process.

KDD consists of an iterative sequence of steps:-

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

There are mainly 4 types of Pre-processing Activities included in KDD process which are:-

1) Data Cleaning
2) Data Integration
3) Data Transformation
4) Data Reduction.



⇒ Explain concept of Market Basket Analysis

MBA - It is a modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

The set of items a customer buys is referred to as an itemset, and MBA seeks to find relationships between purchases.

Typically, the relationship will be in the form of a rule.

Example :- IF {milk, eggs} THEN {bread}.

Types of MBA :-

1) Predictive MBA
2) Differential MBA

⇒ The main benefit of using MBA is that it can increase sales along with customer satisfaction.

DOMS Page No.
Date / /

MBA Application Areas :-

It can be also applied in areas such as :-

- Analysis of credit card purchase
- Analysis of telephone calling patterns
- Identification of fraudulent medical insurance claims.
- Analysis of telecom service purchases.

DOMS Page No.
Date / /

⇒ Fact Table Vs Dimension Table

Fact Table	Dimension Table
• It contains the measuring on the attributes of a dimension table	It contains the attributes on that truth table that calculates the metric.
• It has less attributes	It has more attributes
• It has more records	It has less records
• It forms a verticle table	It forms a horizontal table.
• The attribute format is in numerical or text format	The attribute is in only text format.
• It comes after dimension table.	It comes before fact table.
• Less in schema	More in schema.
• Used for analysis and decision making purpose.	It stores information about a business and its process.

⇒ Mean :- It is average of all the numbers in the data set which is calculated as :-

$$\text{Mean} = \frac{\text{sum of all data values}}{\text{no. of data values}}$$

Median :- Here we have to arrange the data in the increasing order and then find the mid value.

Mode :- It is the number which occur most often in the data set.

Variance :- It is the numerical values that describe the variability of the observations from its arithmetic mean.

Standard Deviation :- It is a measure of dispersion of observation within dataset relative to their mean.

Five-Numbes Summery :- It involves calculation of five quantities.

1) Median
2) 1st Quartile (25th percentile)
3) 3rd Quartile (75th percentile)
4) Minimum
5) Maximum.

Q what is the need of data warehouse and role of data marts.

→ Data warehouse

- A data warehouse is a type of data management system that is designed to enable and support business intelegence activities and analytics.

- A data warehouse centralizes and consolidates large amounts of data.

Need of data warehouse

- Ad-hoc reporting and analysis
- Dynamic presentation
- Security
- Increase Integrity
- Drill down capability
- Data mining through tools gets easier
- Highly flexible
- Increases data quality
- Automatically updates on regular basis
- Uses denormalized schemas.

DATA MART

A data mart is a subset of directorial information store, generally oriented to a specific purpose or primary data subject which may be distributed to provide business needs.

Fundamental use of data mart is Business Intelligence applications (BI).

Q Various Data Reduction Strategies.

→ Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but maintains the integrity of the original data.

- Data Reduction Techniques.

1. Data cube Aggregation

where aggregation operations are applied to data in the contruction of a data cube.

2. Attribute Subset Selection

where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

3 Dimensionality Reduction

where encoding mechanisms are used to reduce data set size

4. Numerosity Reduction

where data are replaced or estimated by alternative, smaller data representations such

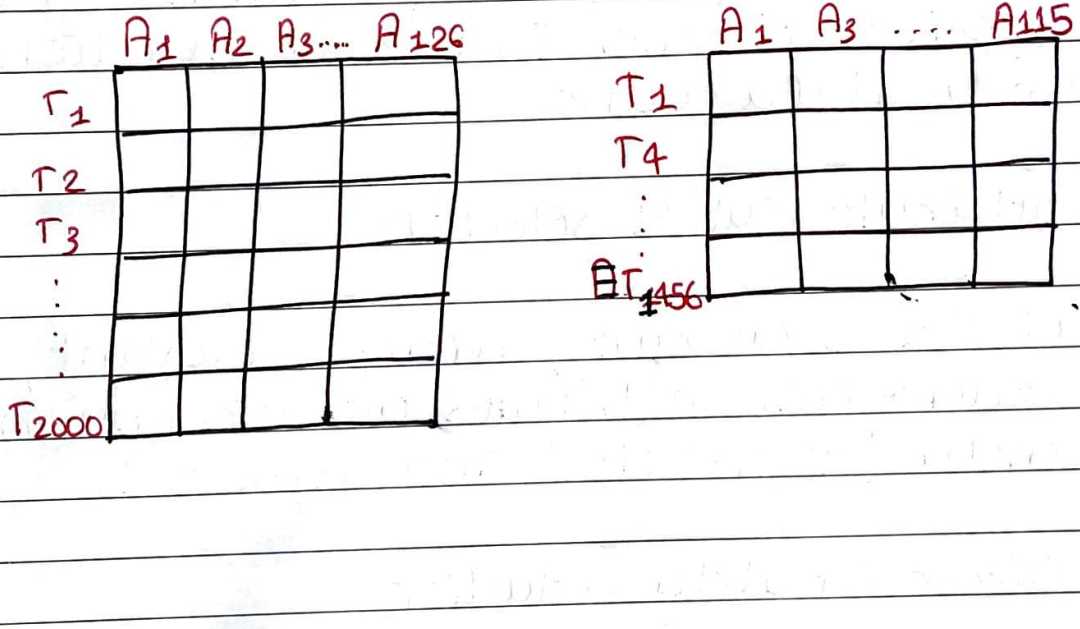
as parametric models or non-parametric methods such as clustering, sampling etc

5. Discretization and concept hierarchy generation

where raw data values for attributes are replaced by ranges or higher conceptual levels.

It is a form of Numerosity Reduction.

Eg :



Q What is concept Hierarchy? What are types of concept Hierarchy?

→ A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher level, more general concepts.

Different types of concept hierarchy

Schema hierarchy
Set grouping hierarchy
Operation derived hierarchy
Rule based hierarchy.

1. Schema hierarchy

→ Formally expresses semantic relation between attributes

Eg: house_no. < street < city < state < country

lower level ⟶ Higher level
Conceptually.

2. Set grouping hierarchy

→ Organizes values for a given attribute into groups or sets or range of values

Total or partial order can be defined among groups

Used to refine or enrich schema-defined hierarchies.

Eg: {young, middle-aged, old} all (age)

{20 ..... 30} ⊂ young

{40 ..... 60} ⊂ middle-aged

{65 ..... 90} ⊂ old

3. Operation derived hierarchy

→ Based on operation specified by users, experts or by mining systems.

Operations can include: decoding
encoding
extraction

of information from complex data clusters.

Eg: Karan Hardin@ce.mu.gtu.edu

initiates the hierarchy

user_name < dept < university < board

4. Rule Based Hierarchy

→ whole concept hierarchy or a portion of it defined by a set of rules and is evaluated dynamically based on current data & defination.

level_1 : low_profit_margin < level_0 : Price - cost < Rs50

level_1 : ~~low~~ medium _profit_margin < level_0 : Price - cost < Rs100

level_1 : high_profit_margin < level_0 : Price - cost < Rs 500