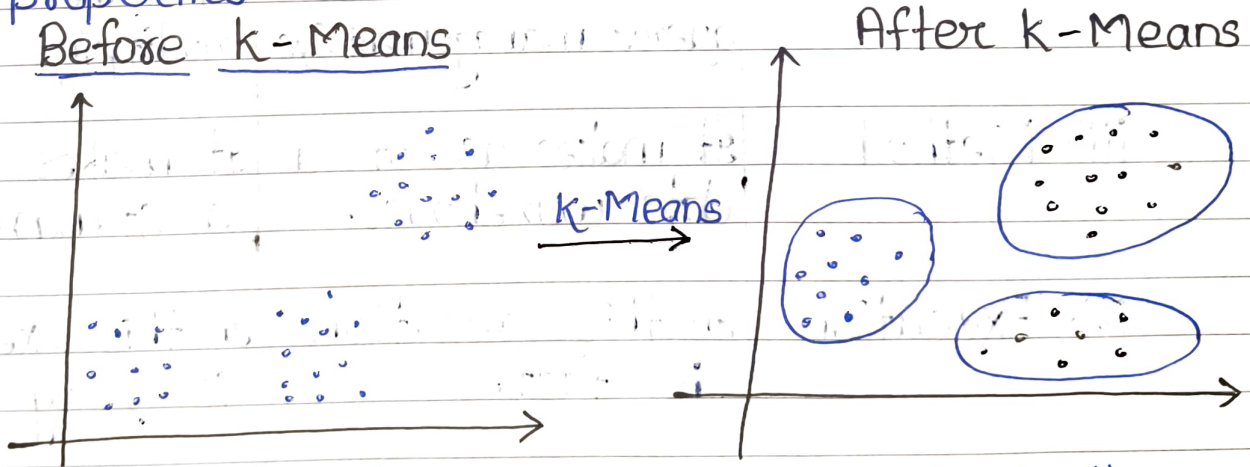
Management Support System

1. K-means clustering algorithm
K-means clustering algorithm groups similar objects or data points into clusters. This approach is popular for multivarite numeric data.

- objective :- i) Grouping similar Data points
K-Means is designed to cluster data points that share common traits, allowing patterns or trends to emerge.

ii) Minimizing within-cluster Distance
Another objective is to keep data points in each group as close to the cluster's centroid as Possible.

Algorithm :- It is an iterative algorithm that divides that unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.



Step 1 :- Select the number k to decide the number of clusters.

step 2 :- Select random k points or Centroid.

step 3 :- Assign each data point to their closest Centroid, which will form the predefined k clusters.

step 4 :- Calculate the variance and place a new Centroid of each cluster.

step 5 :- The model is ready.

2. OLAP and OLTP :-
   ↓ Online analytical Processing
   ↳ Online transaction Processing

Differences :-

	~~Data Source~~	OLAP	OLTP
i)	Data Source	Transcation data based.	Data warehouse Or data mart.
ii)	Definition	It is well-known as a online database query management Sys.	It is well known as an online database modifying System.
iii)	Method	It makes use of a data warehouse.	It makes use of a standard.
iv)	Execution	In this execution is fast.	In this execution is very slow.
v)	Database design	design with a focus on the Subject.	designed that is focused on the application.

3. What is a Cube and what to do drill up roll down and slice and dice mean ?

Ans :- The main operational structure of OLAP is based On a Concept Called Cube. A Cube in OLAP. is multi dimensional data structure is might be actual or virtual that allows fast analysis of data. It can also we define as the Capability of efficiently and analysis data from multiple propectives.

- Slice :- A slice is a Subset of the Cubes Corresponding to a Single value for one or more members of the dimensions.
- Dice :- The dice operation describes a Subcube by operating a Selection on two or more dimension.
- Drill down :- Drilled down or up is a specific OLAP technique where by the user nevigates among levels of data record from rest Summarise to the most detail.
- Roll-up :- A roll-up involves Computing all of the data relationship for One or more ~~arranges~~ dimensions to do this a Configurational relationship or a formula might be defined.
- Pivot :- A Pivot is means of changing the dimensional Orientation report or ad-hoc Query page display.

4. ✡ Apriori Algorithm :- The most usually Commonly to discovered association rule by recurrisely identify frequent items sets.

Tran no.	SK Us
1	1,2,3,4
1	2,3,4
1	2,3
1	1,2,4
1	1,2,3,4
1	2,4

⟹

item sets	Support
1	3
2	6
3	4
4	5

⟹

item sets	Support
1,2	3
1,3	2
1,4	3
2,3	4
2,4	5
3,4	3

⟹

item sets	Support
1,2,4	3
2,3,4	3

→ draw this table

→ describes this tables

↳ given a set of example of items to above tables the algorithms attempts to find Subsets that are common to atleast a minimum no. of the item set.

- Apriori uses a bottom up approach where frequent Subsets are extended one item at a time and groups of items at each level are tested against the data for minimum Support the algorithm terminates number of further Successful extension are found.

* Components of apriori algorithms -

i) Support
ii) Confidence
iii) Lift

__/__/____

Advantage :- It is used to calculate large itemsets.
Simple to understand and apply.

- Disadvantage :- It is an expensive method.

5. Differences – cluster analysis and classification

Ans :-

Parameter	classification	clustering
Type	used for supervised learning.	used for unsupervised learning.
Basic	Process of classifying the input instances based on their class labels.	grouping the instances based on their similarity without the help of class.
Complexity	more complex as Compared to clustering.	less complex to compared to classification.
Methods	decision tree, logistic regression	K-mean clustering, hierarchical clustering.
Data Requirement	works with unlabeled data.	Requires labeled data for training.
Output	Produces clusters without specific labels.	Produces discrete known labels for each data point.

6. General algorithm of decision tree with example.

Ans: A decision tree is a type of Supervised Learning algorithm that is Commonly used in machine learning to model and predict Outcomes based on input data. It is a tree-like structure where as internal node test on attribute, each branch Correspond to attribute value and each leaf node represents the final decision or prediction. The decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification.

- Root node
- internal nodes
- Leaf nodes
- Branches
- spilitting
- Parent node
- child node

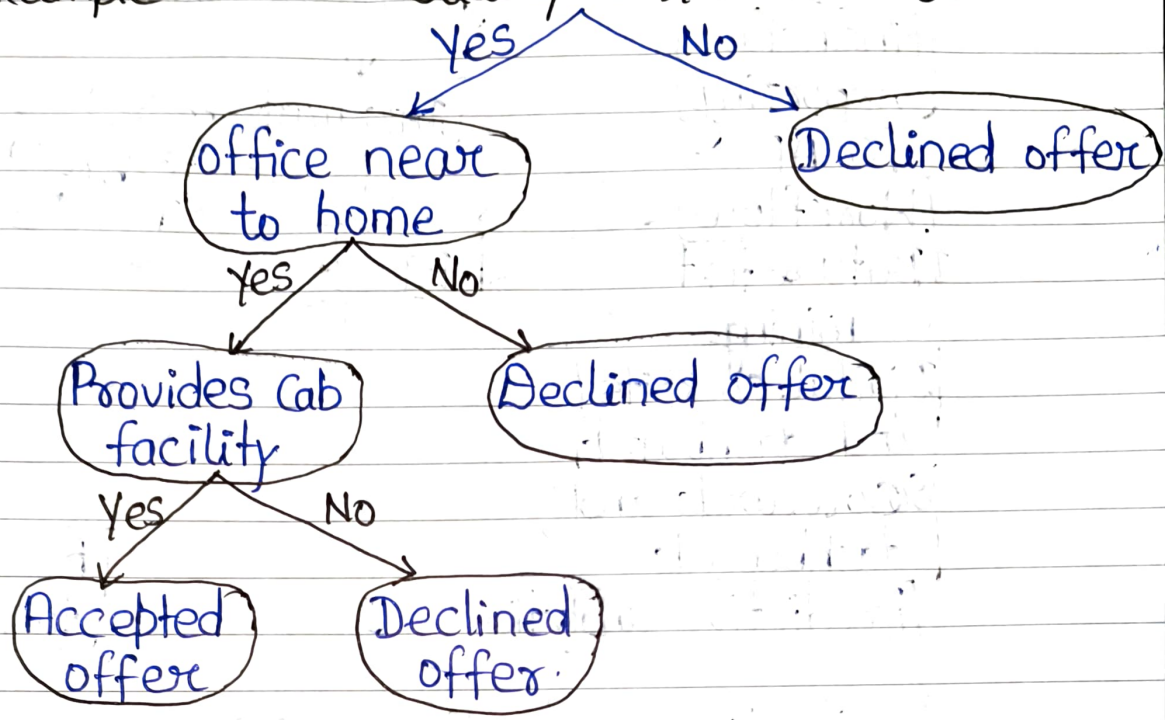
Algorithm :- (i) Begin the tree with the root node → Says X which Contains the Complete datasets.

(ii) find the best attribute in the dataset using attribute selection.

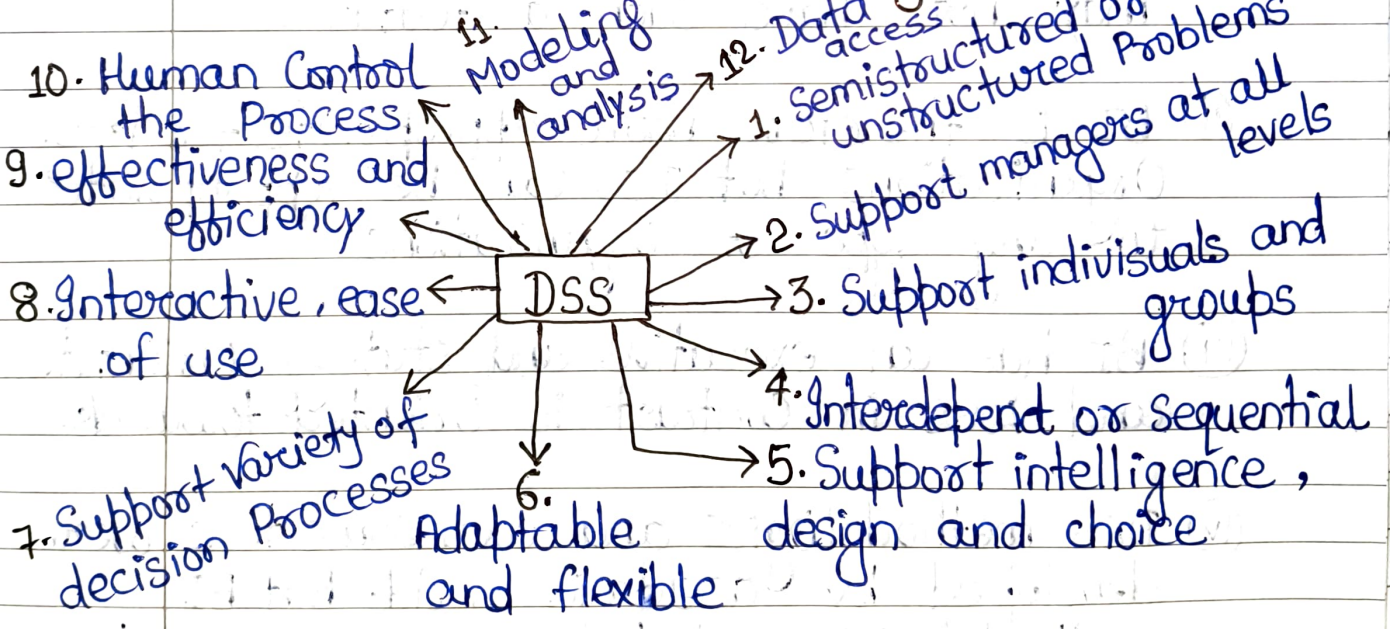
(iii) Divide the S into subsets that Contains possible values for the best attributes.

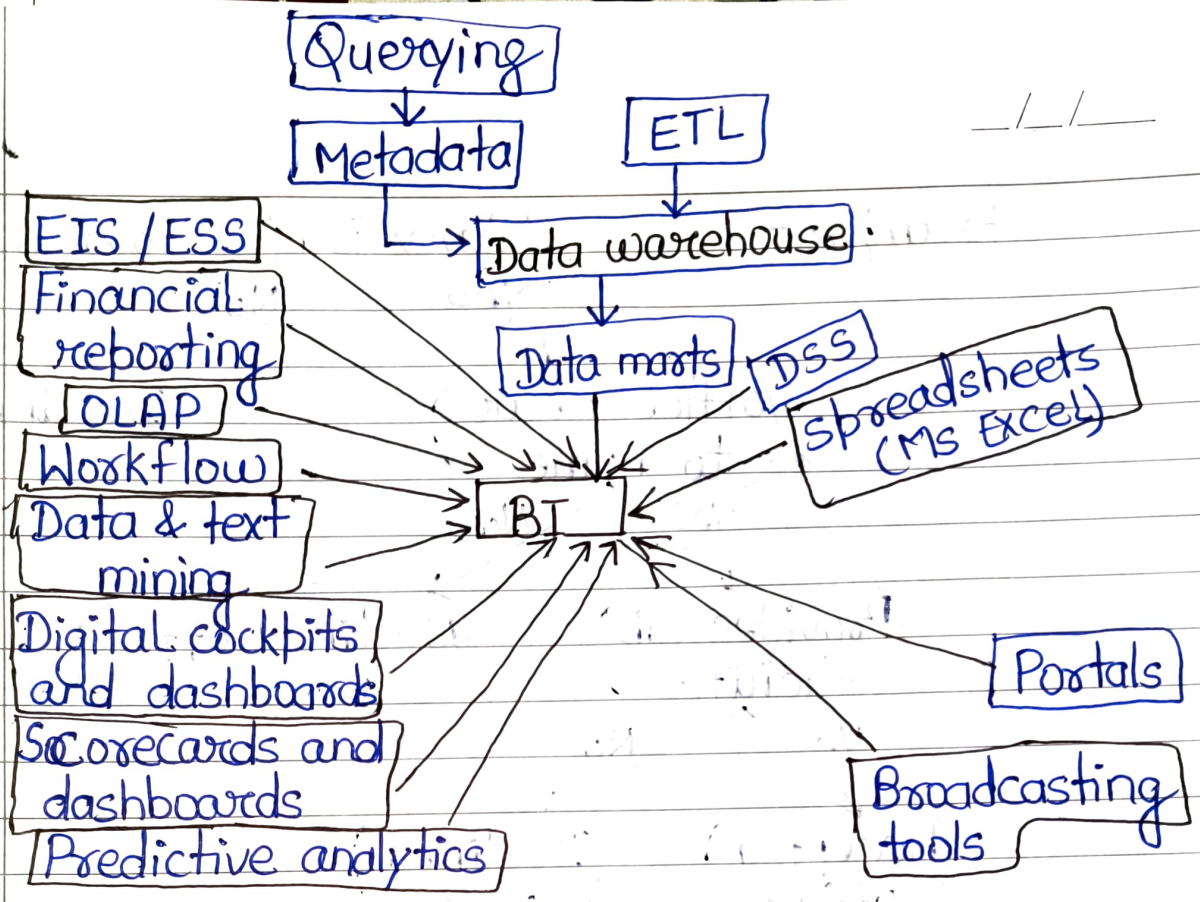
(iv) Generate the decision tree node, which Contains the best attribute.

Example –



7. Difference and Similarity b/w DSS & BI.





Similarities and differences -

(i) Their architecture are very similar because BI envolved from DSS. but DSS may or may not have such a feature.

(ii) BI is therefore more appropriate for large Organizations but DSS can be appropriate to any type of Organizations.

(iii) BI has an executive and stategy Orientation and DSS in Contrast is Oriented towards.

(iv) DSS methodologies and even some tools were developed by academic world. But BI methodologies and tools were developed by S/w Companies.

8. Gini index and how does it measure?

Ans:- Gini index is a metric to measure how often a randomly choosen element would be incorrectly identified.

- It means an attribute with a lower gini index should be preferred.
- Sklearn supports "Gini" criteria for Gini index and by default it takes "gini" value.

$$Gini(S) = 1 - \sum_{i=1}^{c} P_i^2$$

The Gini index is a measure of the inequality or impurity of a distribution, Commonly used in decision trees and other machine learning algorithms. It ranges from 0 to 0.5 where 0 indicates a pure sets and 0.5 indicates a maximally impure set.

- In decision trees, the Gini index is used to evaluate the quality of a split by measuring the difference b/w the impurity of the Parent node and the weighted impurity of child node

9. What an Organizations should Consider before making a decision do purchase data mining Software?

Ans: Before Purchasing data mining Software, an Organizations should Consider things like :-

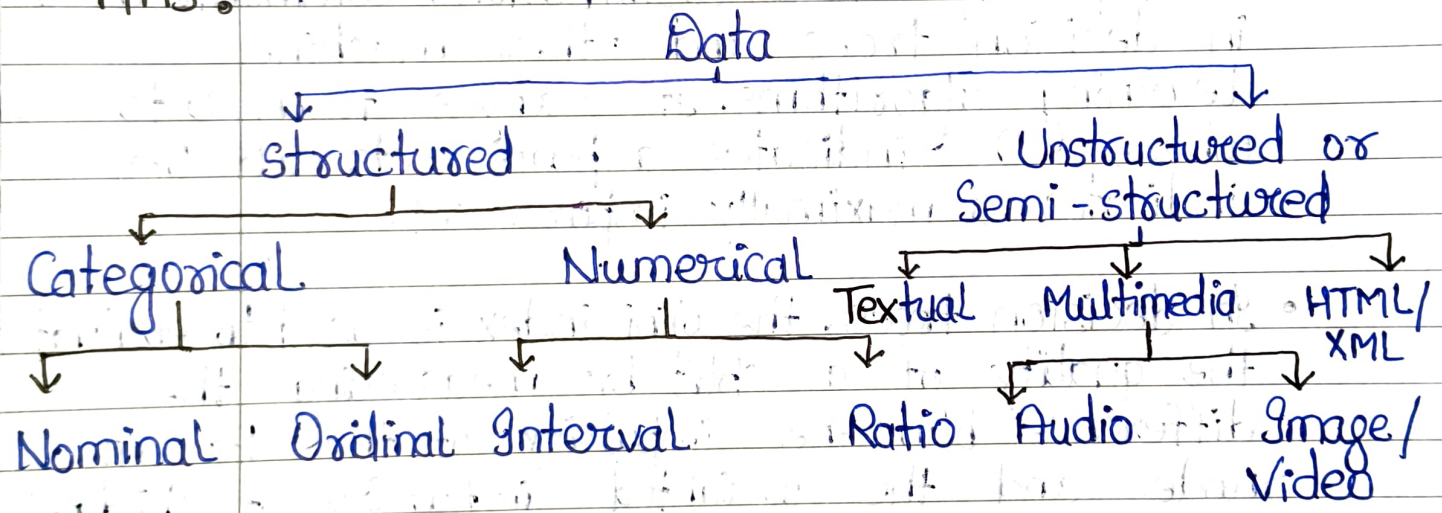
- Data quality and accuracy :- How well the S/w Can handle the quality and Size of the data.

__/__/____

- Usability :- How easy the s/w is to use.
- Visualization :- what visualization features the s/w offers.
- Cost :- How costly the s/w will be to invest
- Effciency and Scalability :- How efficient and Scalable the s/w is.
- Support Services :- what Support Services are available for the Software.
- Vendor Reputation :- The reputation of the s/w vendor should also be Considered.

10. Taxonomy of Data mining

Ans :-



- Categorical data represent the labels of multiple classes used to divide a variable into specific groups.
- Nominal data Contain measurements of simple codes assigned to objects as labels, which are not measurements.

__/__/____

- Oridinal data Contains codes assigned to objects or events as labels that also represent the rank order among them.
- Numeric data represent the numeric values of specific variables. e.g - numeric value
- Interval data are variables that can be measured on interval Scales.
- Ratio data includes measurement variables Commonly found in the physical sciences and ~~engen~~ engineering. e.g - Mass, length, time

other datatypes including textual, spatial, imagery, and Voice, need to be Converted into some form of Categorical or numeric representation before they Can be processed by data mining algorithms. Data Can also be classified as static or dynamic.

The nominal or ordinal Variables are Converted into numeric representations using Some type of 1 of -N Pseudo variables.

10. How the DSS are Supported ?

Ans :- A Decision Support System is an interactive Computer-Based System that assists individuals or Organizations in making informed decisions by analyzing large volumes of data, Provides insights and helping predict Outcomes.

Supported ? :- Here we relate specific technologies to the decision making Process. Databases, data marts, and especially data warehouses are important technologies in Supporting all phases of decision making.

① • Support for the intelligence Phase :- The Primary requirement of decisions Support for the intelligence Phase is the ability to Scan external and internal information Sources for opportunities and Problems and to interpret what the Scanning discovers.

• Decision Support / BI technologies Can be very helpful. Graphical Systems information System Can be utilized either as stand-alone Systems or integrated with these Systems so that a decision maker Can determine opportunities and Problems in a spatial Sense. Another aspect of identifying internal Problems and Capabilities involves monitoring the Current status of operations.

Expert Systems in Contrast Can render advice regarding the nature of a Problem, its classification, its seriousness, and the like.

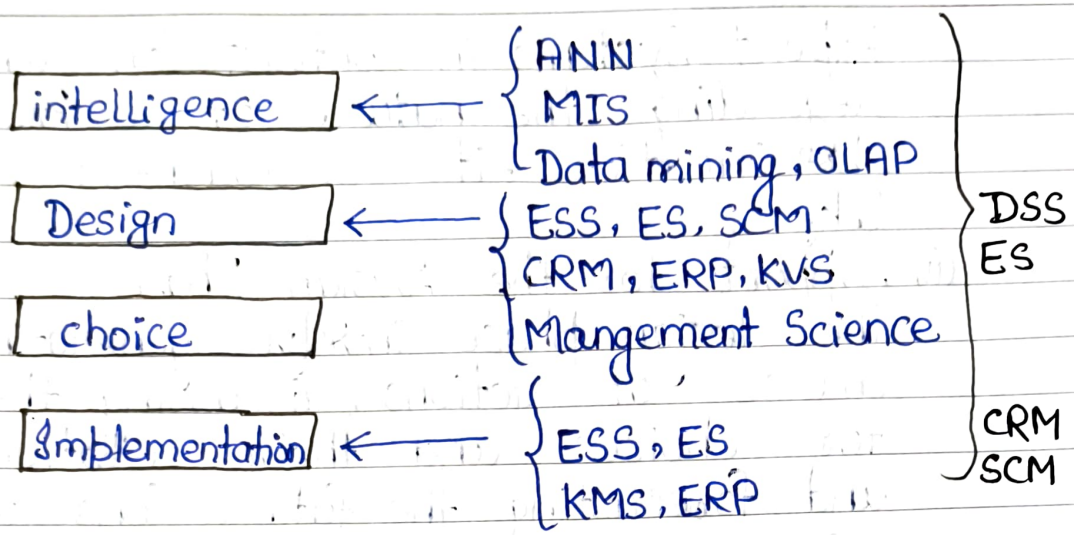
(ii) Support for the Design Phase
The design phase involves generating alternatives Course of actions, discussing the Criteria for choices and their relative importance and forecasting the future Consequences of using Various alternatives. Several of these activities Can use standard models provided by a DSS.

(iii) Choice Phase :- In addition to Providing models that rapidly identify a best or good enough alternatives, a DSS Can Support the choice Phase through what-if and goal-seeking analyses.

(iv) implementation Phase :- DSS Can be used in implementation activities Such as decision Communication, explanation, and Justification. Implementation Phase DSS benefits are partly due to the vividness and detail of analyses and reports. Reporting Systems and others tool variously labeled as BAM, BPM, KMS, EIS, ERP, CRM and SCM are all useful in tracking how well an implementation is working.

- and draw the diagram.

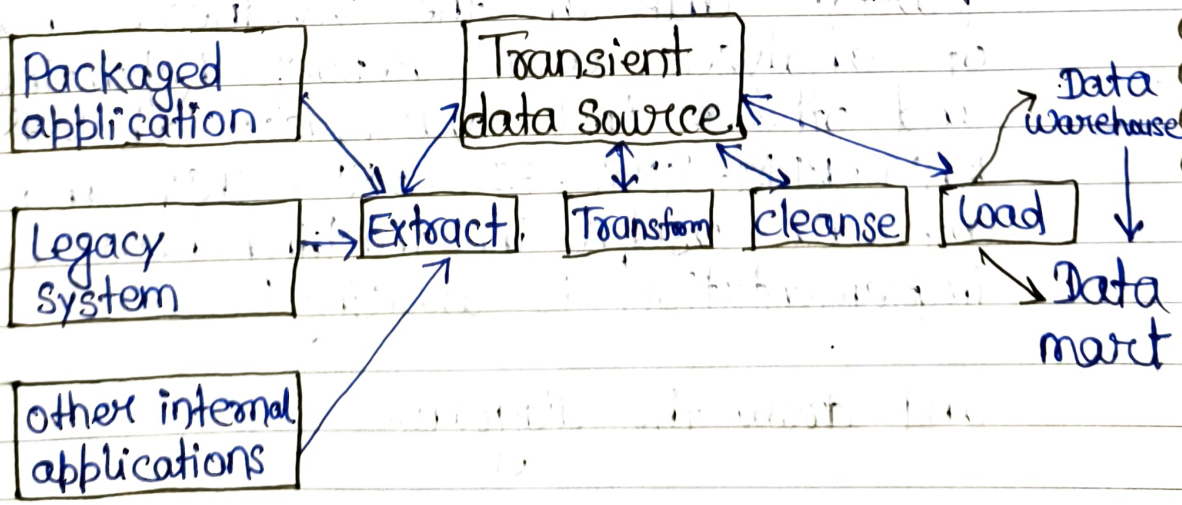
10.



11. Data Integration and the Extraction, Transformation and Load (ETL) Process

Ans :- Data integration comprises three major processes that, when correctly implemented, permit data to be accessed and made accessible to an array of ETL and analysis tools and the data warehousing environment: data access, data federation and change capture.

- Enterprise application integration (EAI)
- Service-Oriented architecture (SOA)
- Enterprise information integration (EII)
- Extraction, transformation, and load (ETL)



At the heart of the technical side of the data warehousing Process is extraction, transformation and load (ETL)
~~Extraction~~ ETL technologies, which have existed for some time, are instrumental in the Process and use of data warehouses. The ETL Processes is an integral Component in any data-centric Project. IT managers are often faced with challenges because the ETL Process typically Consumes 70% of the time in a data-centric Project. ~~IT managers are often faced with challenges because the ETL Process typically Consumes~~ Transformation Occurs by using rules or lookup tables or by Combining the data with other data.

- Data transformation tools are expensive.
- Data transformation tools may have a long learning Curve.

** Data mining
- Prediction
  - classification
  - Regression
- Association
  - Link analysis
  - Sequence analysis
- clustering
  - Outlier analysis

__/__/____

Learning method ⟶ Popular algorithm

① • Supervised ⟶ classification and regression trees, ANN, SVM, GA

② • Supervised → Decision trees, ANN/MLP, SVM, Rough sets, GA

Supervised → Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM

③ Unsupervised → Apriory, OneR, ZeroR, Eclat

→ Expectation maximization, Apriory algorithm

④ Unsupervised → K-Means ANN/SOM

→ K-Means, Exceptation maximization.

Fill in the blanks –

(i) _____ is an important Component of text mining and is soft field of AI and Computational Linguistic it study of Problem of understanding natural human language.

Ans:- NLP

(ii) _____ is One or more web pages that provides a Collection of links authority Pages.

Ans:- Hub Page

(iii) _____ mining is the Process of extracting useful information from the links ambedded in web documents.

Ans:- Web structure

(iv) _____ Segmentation is one of the unsupervised machine learning methods where in data is put into several groups based on their similarity.

Ans:- clustering

(v) _____ deals with the kind of uncertainity and partial information that is human in nature.

Ans:- fuzzy logic

(vi) _____ is a relational approach to design and developed of a model based System.

Ans – Relational model based System

vii) if the simulation results do not match the equitance and Judgement of decision maker than a Confidence Can occur in the result.

viii) Continuous distributions are situations with limited no. of possible events that follow density function Such as the normal distributions.