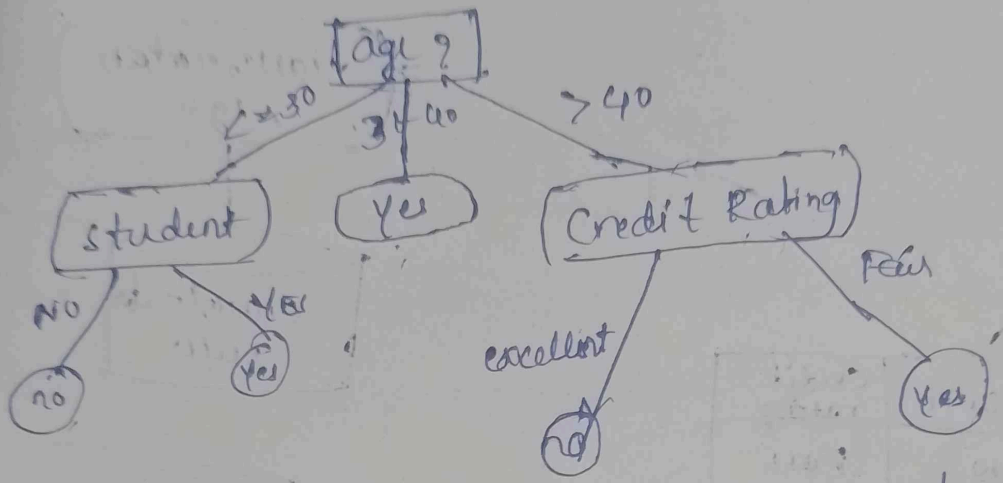


Decision Tree : is a flow chart like tree structure where each internal node denotes the test on an attribute each branch represents an outcome of the test and leaf node represents classes or class distribution. Top most node in a tree is a root node.

04/01/2024

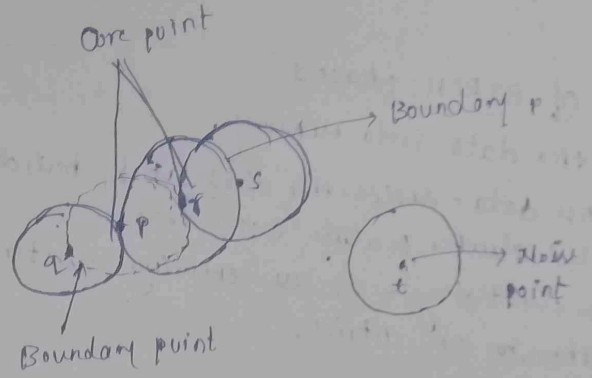
DBSCAN (*** 7-8 marks)

(Density Based Spatial clustering of Application with Noise)

→ Forms clustering based on density.

→ Eps
Minpts = 3] → 2 I/p parameters.

Minpts → Minimum pts.



→ Directly density reachable must be neighbour & core pt so q is directly density reachable from pt p.

→ p & r are core pts → satisfies Minpts = 3

→ s & q → Boundary pt → should be neighbour of any core point. t is noise point.

Classification and Prediction

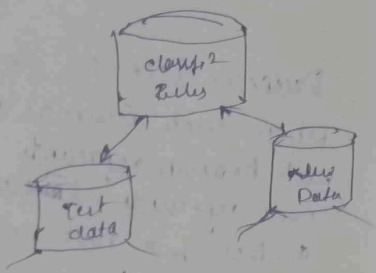
→ Data classification is a 2 step process.
In the 1st step a model is built describing a predetermined data classes or concepts.

→ Test data are used to estimate the accuracy of the classification rules.

2 steps
→ Training data set
→ Testing data set
→ supervised learning
→ Unsupervised learning.

name	age	income	credit rating
AAA	<=30	Low	Fair
Bill	31-40	HIGH	Excellent

Fig: Data classification Process



Issues regarding classification and prediction.

1) Data cleaning: Preprocessing of data to remove or reduce noise.

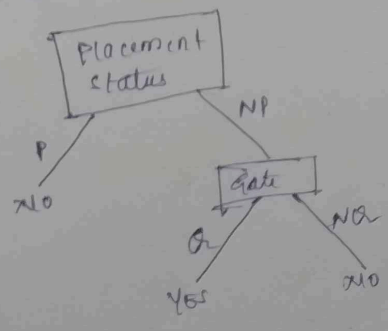
→ Relevance analysis.

→

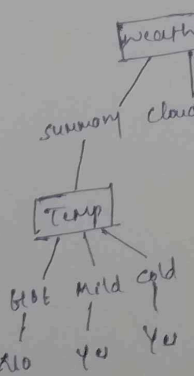
3) Data Transformation: Data can be generalised to higher level concepts.

Decision Tree Induction

→ To take a decision



playing football



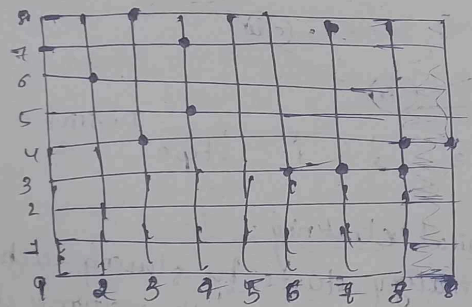
BIRCH

(Balanced Iterative Reducing & Clustering using Hierarchies)

→ It is scalable clustering.
→ Works for very large dataset.
→ Only one scan of data is necessary.
→ Clustering is based on CF (Clustering Feature).
→ CF Tree → stores the cluster features.
→ cluster of data pts is represented by triple of numbers (N, LS, SS)

N = No. of items
LS = Linear sum of pts
SS = Sum of squard of pts.

eg:



e.g.; CF

CF = (N, LS, SS)

N = No. of data pts

$LS = \sum_{i=1}^{N} X_i \quad SS = \sum_{i=1}^{N} X_i^2$

(3,4) (2,6) (4,5) (4,7) (3,8)

N = 5

LS = 3+2+4+4+3 = 16
= 4+6+5+7+8 = 30

SS = $3^2+2^2+4^2+4^2+3^2 = 54$
$4^2+6^2+5^2+7^2+8^2 = 190$

(6,2) (7,3) (8,3) (8,4)

N = 4

CF = (N, LS, SS)

LS = 6+7+8+8 = 29
2+3+3+4 = 12

SS = $6^2+7^2+8^2+8^2 = 213$
$2^2+3^2+3^2+4^2 = 38$

CF = (4,)

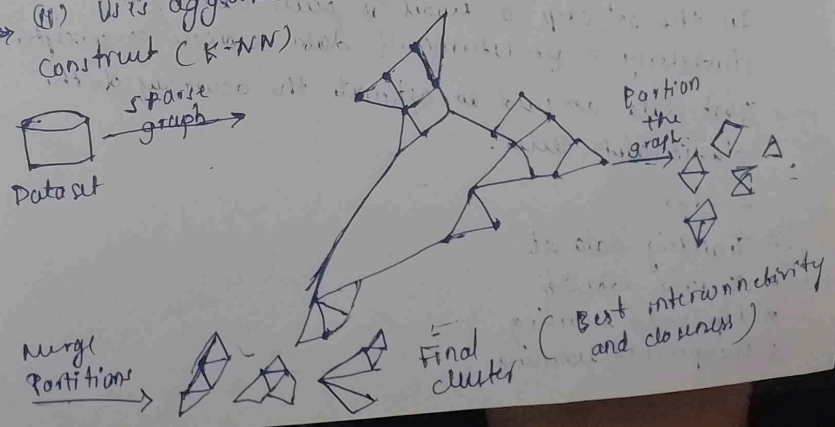
Basic algorithm of BIRCH: ~~Phase 1~~

Phase 1: Load the data into memory.
Phase 2: Condense data - resize the data set by building smaller cluster feature (CF) tree.
Phase 3: Global clustering - It uses existing clustering algorithm on CF entries.
Phase 4: Cluster refining.

Chameleon: Hierarchical clustering using dynamic modelling

→ Two clusters are merged → if interconnecting & closeness b/w 2 clusters are high.
→ 2 parameters → Interconnectivity & closeness.
→ Graph based & a two phase algo. (i) uses graph partitioning algorithm.
→ (ii) Uses agglomerative hierarchical clustering algo.

Construct (K-NN)



→ then compute the new mean for each cluster. This process iterates until all the objects are in one of the k cluster.

K-medoid Algorithm

I/P : No. of k clusters, n objects

O/P : Set of k cluster.

Method:

1) Arbitrarily choose k objects as initial medoids.
2) Repeat
3) Assign each remaining object to cluster with nearest medoid.
4) Randomly select non medoid object O_{random}
5) Compute total cost, s of swapping O_j with O_{random}
6) If $s < 0$ then swap to form new set of k medoid
7) Until no change.

k-Medoid Algorithm : (8M) ***

		x	y	k_1	k_2	cost
$k_1 = m_1$	①	8	2	0	6	0
	②	3	5	8	2	9
	③	4	7	9	3	3
	④	8	4	2	4	2
$k_2 = m_2$	⑤	5	5	6	0	0 / 7

Manhattan distance

→ $|x_2 - x_1| + |y_2 - y_1|$

$= |5-8| + |5-2| = 3+3 = 6$.

→ $|3-8| + |5-2| = 5+3 = 8$

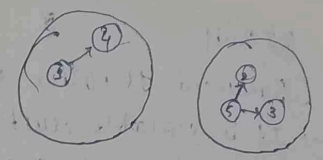
→ $|3-5| + |5-5| = 2$

→ $|4-8| + |7-2| = 4+5 = 9$

→ $|4-5| + |7-5| = 1+2 = 3$

→ $|8-8| + |4-2| = 2$

→ $|8-5| + |4-5| = 3+1 = 4$

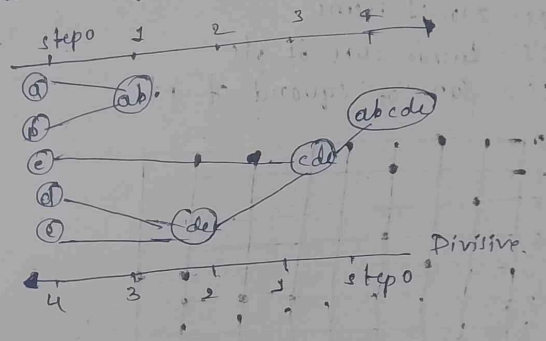


04/01/2024

Hierarchical Method (***Imp)

→ Works by grouping data objects into tree of clusters.

→ Divided into → Agglomerative & Divisive Hierarchical clustering.



Agglomerative Hierarchical clustering:
This is bottom-up strategy starts by placing each objects in it's own cluster and then merge these atomic clusters into larger & larger clusters until all the objects are in single cluster.

Divisive Hierarchical clustering:
This is top-down approach it starts with all objects in one cluster. It subdivides the cluster into smaller & smaller pieces until each object forms a cluster on it's own.

03/01/2024

Categories of clustering methods

1) Partitioning method
→ DB of n objects, k partitions of data
→ $k \leq n$
→ classify data into k groups
1) Each group must contain atleast one object
2) Each object must belong to exactly one group.
→ Uses iterative relocation technique methods
1) k-means
2) k-medoid

2) Hierarchical Method
→ Hierarchical decomposition of given set of data objects
→ Methods
1) Agglomerative → Bottom-up
2) Divisive → Top-down

3) Density Based Method
→ It is based on density, it continuous growing the given cluster as long as the density exceeds some threshold.

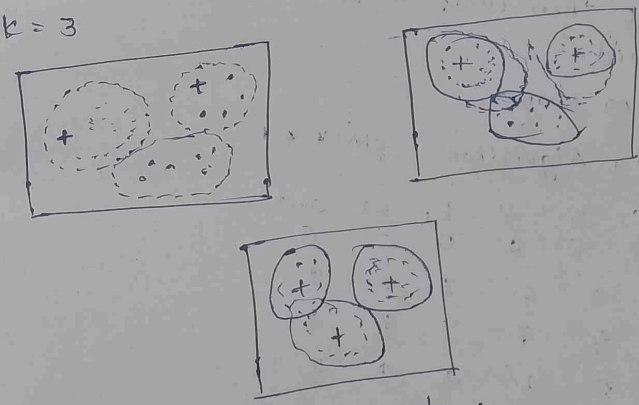
4) Grid Based Method
→ It quantise the object space into finite no. of cells that form a grid structure. All the clustering operations are performed on grid structure.

Partitioning Method (10 M) ***

→ $K \leq n$
→ K-means Algorithm
I/P : no. of cluster k, n objects.
O/P : set of k clusters.

Method
1) Arbitrarily choose k objects as initial cluster centre.
2) Repeat
3) (re) assign each object to cluster to which the object is most similar based on mean value of the object in the cluster.
4) Update the cluster mean.
5) Until no change.

k = 3



How k-means algorithm works :
→ First it randomly selects k of the objects, each of which initially represents cluster means or cluster centre.
→ For each of the remaining objects an object is assigned to a cluster to which it is most similar based on the distance blw object and cluster mean.

Datawarehouse Implementation

21/12/2023

Compute cube operator and it's implementation

→ Compute cube → Aggregate over all subsets of dimensions.

e.g., 3 dimensions - city, item, year and Sales-in-dollars as measures

$2^3 = 8$ Possible groupings {city, item, year}, {city, item}, {city, year}, {item, year}, {city}, {item}, {year}, {}

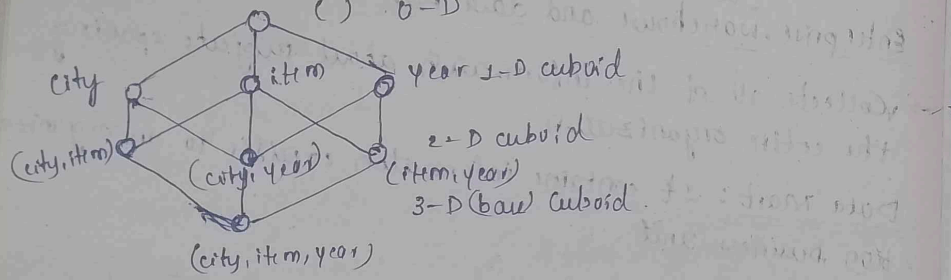


Fig. 3 D Database

Datawarehouse Backend Tools and Utilities.

1. Data extraction
2. Data cleaning
3. Data Transformation
4. Load
5. Refresh. → All the updation to the data has to be done

UNIT-04

CLUSTERING

Imp *** (8-10M)

→ Cluster is a collection of data objects that are similar to one another within the same cluster and dissimilar in other cluster.

→ It is widely used in numerous applications like image processing, pattern recognition, data analysis, market research

→ Clustering is an example of unsupervised learning.

* Unsupervised learning do not rely on predefined classes and class labeled training examples.
* Supervised → Predefined sets or classes.

Requirements of clustering

1. Scalability
2. Ability to deal with different types of attributes.
3. Minimal requirements for domain knowledge to determine input parameters.
4. Ability to deal with noisy data.
5. High dimensionality.
6. Constraints based clustering.
7. Interpretability and usability.

***

Types of data in cluster analysis

1. Interval scaled variable → 200-300
2. Binary variable → 0 or 1
3. Nominal variable → map-clear colour (Representing more than one state)
4. Ordinal → Assistant, Associate, professor. (order/sequence)

Database Design Process steps

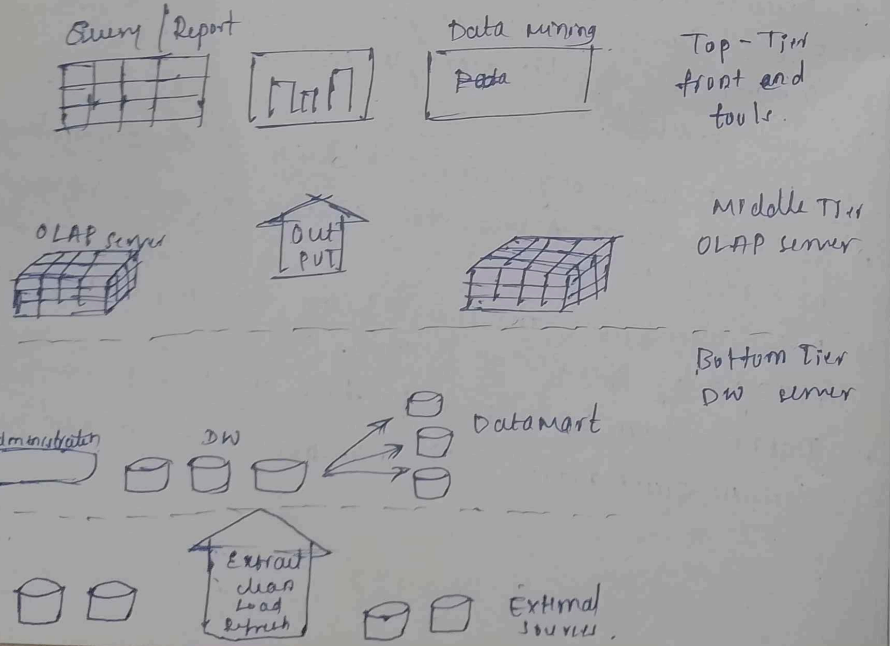
1) Choose a business process to be modeled e.g., orders, shipments, sales.
2) Choose grain of the ~~result~~ business process (key) e.g., Individual transaction
3) Choose the dimensions.
4) Choose the measures for each fact table record.

20/12/2023

Process of Database Design

→ Top down approach - design of planning.
→ Botton-up approach - experiment and protypes.
→ Combined approach.
→ Software engg → planning, requirement study, problem analysis, warehouse design, Data integration and testing, deployment.

Three-Tier DW Architecture (8-10 M) ***



* Bottom tier is almost a relational DB where the data extracted from this tier in order to create a data ware-house.
* Middle tier is implemented using relational OLAP and multidimensional OLAP.
* Top tier contains reporting tools, analysis tools and data mining tools.

Enterprise Data warehouse and data mart

→ Collects all of the information about subjects spanning the entire organization;

Data Mart: It contains information specific to an organization business unit.

Types of OLAP servers

1) Relational OLAP server.
2) Multidimensional OLAP server.
3) Hybrid OLAP server.

1) Relational OLAP server

→ Application based on relational DBMS. It can handle large amount of info., it has greater scalability; and it's performance is slow or low.

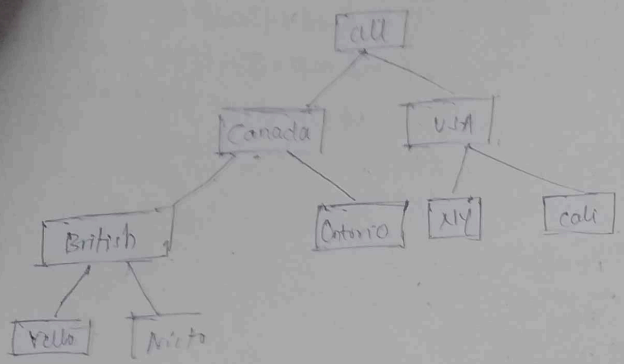
2) Multidimensional OLAP ser

Application based on multidimensional DBMS. It has fast info. retrival, performance complex calculation and limited info. it can handle.
→ Through indexing info. can be retrieved.

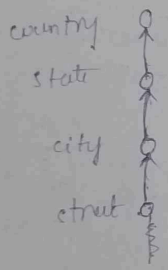
3) Hybrid OLAP server

→ It is combination of both relational and multidimensional OLAP server.

Concept hierarchy.

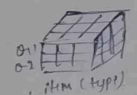


def: Concept hierarchy defines a sequence of mappings from set of low level concepts to higher level. The mappings form a concept hierarchy for the dimension location mapping a set of low level concepts (cities) to higher level concepts (countries)

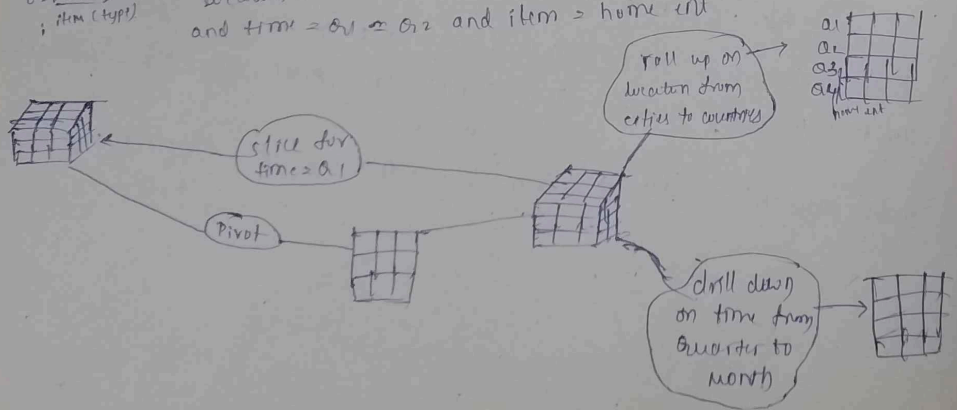


07/12/2023

OLAP operation in Multidimensional data model



dice for location = "Toronto" or "Vancouver" and time = Q1 or Q2 and item = "home ent"



Roll up Operation : Climbing up a concept hierarchy for a dimension.
e.g., street, city, state, country

Drill down : stepping down a concept hierarchy for a dimension.
e.g., month, week, day

slice operation : Performs selection on 1 D of the given cube.
e.g., time = Q1

Dice operation : defines subcube by performing a selection on two or more dimensions.

Pivot : It provides an alternative presentation of data.
e.g., Rows → columns or columns → rows.

Datawarehouse Architecture

→ steps for design and construction of DW
What does DW provide for business analysts
1) Presents relevant ~~information~~ information.
2) Enhances business productivity.
3) Customer relationship management
4) Cost Reduction

4 different views regarding the design of data warehouse
1) Top-down views: Allows the selection of relevant information necessary for the DW.
2) Data source view: The information being captured, stored and managed by operational systems.
3) Data warehouse view: includes fact and dimension tables
4) Business query view: The viewpoint of the end user.

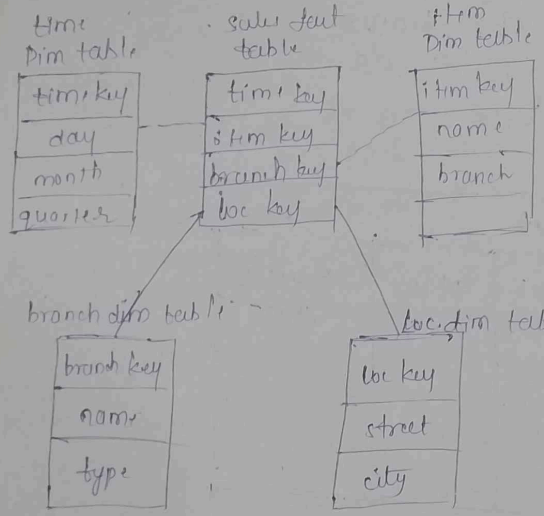
Star, snowflake and fact constellations:

Schema for multidimensional DB (CO-1 S M **)

1) Star schema of Datawarehouse for sales

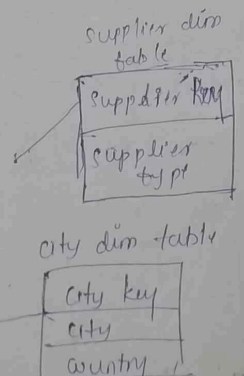
→ Dy?: The Data warehouse contains a large centr table (fact table) containing the bulk of data with no redundancy.

→ set of smaller attendance table (Dimension table) one for each dimension. (Many dim table one fact table)



2) Snowflake Schema

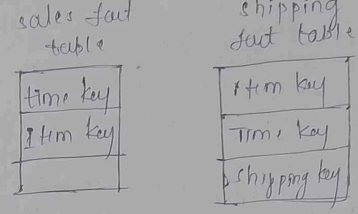
It is the variant of the star schema table where some dim tables are normalized them by further splitting the data into additional tables.



3) Fact constellation schema

It requires multiple fact tables to share dimension tables. This kind of schema can be viewed as collection of star hence it is called as galaxy schema or fact constellation schema.

(Many fact table one dim table)



24/10/2023

Examples for defining star, snowflake and fact constellation schema

→ SQL based DM query language - DMQL

→ DW defined using 2 language primitives

- one for cube dy?
- one for dimension dy?

syntax → define cube <cube name> <dimension list>: <measure list>

define dimension <dimension name> as (<attribute_list>)

e.g., star schema

define cube sales_star [time, item, branch, location]:
dollars_sold = sum(sales_in_dollar), units_sold = count(*)
define dimension time as (time_key, day, month, year)
" " item as (item_key, name, brand)
" " branch as (branch_key, item)

e.g., snowflake schema

define cube sales_star [time, item, branch, location, supplycity]
define dimension location as (location_key, street, city(city_key), state, country)
define dimension item as (item_key, name, type, supplier(supplier_key, type))