FUNDAMENTALS OF DATA MINING

Data Mining has attracted a great deal of attention in information industry in recent years due to wide availability of huge amounts of data & need for such data into useful information and knowledge. Such knowledge is used for dfy applications such as market analysis, fraud detection, production control & science exploration.

Database system is an evolutionary path in the development of following functionalities (i) Data collection and database creation (ii) Data Management (iii) Advanced data analysis.

Since 1960's Database & Information technology has been evolving systematically from primitive file processing systems to sophisticated & powerful database systems.

- Since 1970's database systems has progressed from early hierarchical and Network database systems to the development of relational database systems, Data Modelling tools, Indexing and accessing methods
- users gained flexible data access through Query languages, users interfaces, optimized Query processing & Tx Management.
- Efficient methods for OLTP has evolved for efficient storage retrieval and management of large amounts of data.

mid 1980's — Research & development activities on new & powerful database systems. This lead to development of advanced data models such as Extended relational, object oriented, object relational & deductive Models.

> Application Oriented database systems including Spatial, temporal, multimedia, active, stream, sensor, scientific engineering databases. Issues relating to distribution, diversification, sharing are studied extensively.

The progress has led to the large supplies of powerful and affordable computer, data collection equipment & storage media.

Late 1980's - present :- Data is stored in dfy repositories One such is Data Warehouse, a repository of multiple heterogeneous data sources organized under a unified schema inorder to facilitate decision making.

DW includes cleaning, Integration and OLAP — analytics technique with functionalities such as summarization, consolidation and aggregation. OLAP tools support multidimensional analysis & some tools for data classification, clustering.

1990's - present : XML - Based database systems are developed. Integration with information retrieval, Data and Information integration.

Data — Records, Webpages, documents etc

Mining — Process of extracting ore or minerals from ground

DM — Nontrivial extraction of implicit, previously unknown and potentially useful info from large amount of data.

Why DM

* Abundance of data resources: commercial database, internet, Intranet

↓

Contain large amount of valuable data

Best way to structure data depend on how one want to exploit it

Manual data organization is expensive and laborious so there is need to automate this process

DM applications

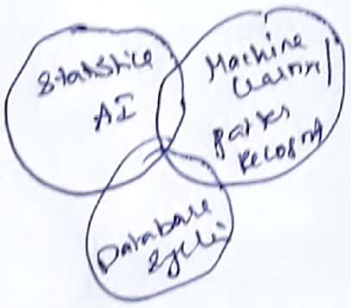
Customer analysis (impacts customer behaviour)

Medical research (impact of lifestyle / drug effect)

Insurance (risk assessment)

Stock investment (which factors impact stock performance)

fraud detection (whether is transaction okay or not)



Evolution of Database System Technology

Data collection and database creation (1960s and earlier)
- primitive file processing

↓

Database Management system (1970s - early 1980s)
- Hierarchical and network database system
- Relational database systems
- Data modeling tools: entity-relational models, etc
- Indexing and accessing methods: B-trees, hashing etc
- Query languages: SQL, etc
- user interfaces, forms and reports
- Query processing and query optimization
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

↓

Advanced database systems (mid-1980s - present)
- Advanced data models: extended relational, Object-relational, etc
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, Knowledge-based

Advanced Data Analysis: Data warehousing and Data Mining (late 1980-s Present)
- Data warehouse & OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc
- Advanced data mining app. stream data mining, biodata mining etc
- Data mining and society. Privacy-preserving data mining.

Web based databases (1990s - present)
- XML-based database systems
- Integration with information retrieval
- data and information integration

↓

New Generation of Integrated Data & Information Systems (Present-future)

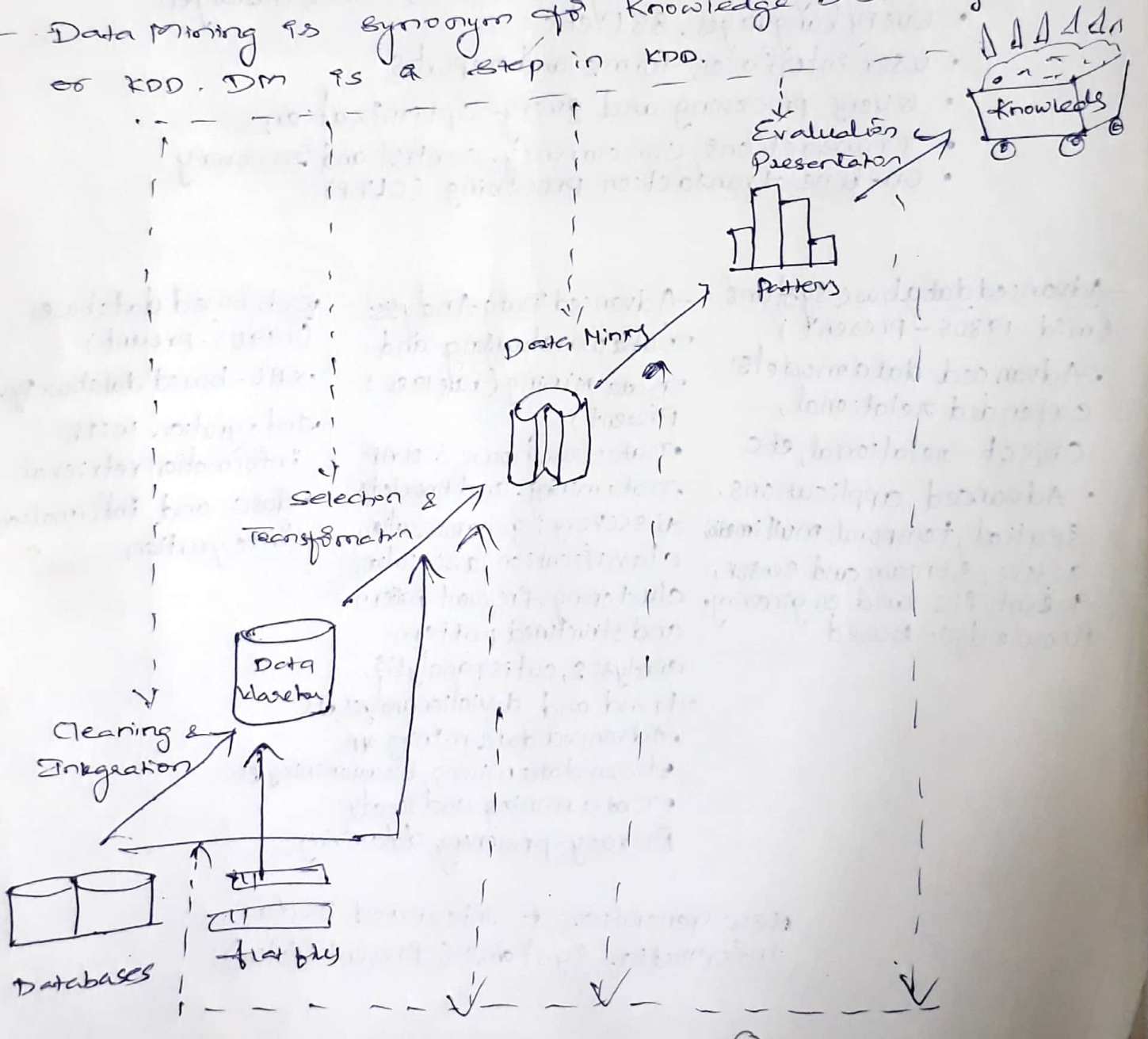
DATA MINING :-

Data mining refers to Extracting or mining knowledge from large amounts of data.

Mining of gold from rocks or sand is refereed as gold mining rather than rock or sand mining. So data mining can be appropriately named as "knowledge mining from data"

* Many other terms to Data Mining which has slightly dif meaning are knowledge mining from data, knowledge Extraction, data / pattern Analysis, data archaeology and data dredging.

* Data Mining is synonym of Knowledge Discovery from data or KDD. DM is a step in KDD.



KDD — Automatic Extraction of non-obvious, hidden knowledge from large volumes of data

Data — String of bits of Numbers and symbols of objects

Information — Data stripped of redundancy and reduced to the minimum necessary to characterize data

Knowledge — Integrated information, including facts, relations which have been perceived, discovered or learned as our "world picture"

↓

Data at right level of abstraction and generalization

How to acquire knowledge for knowledge based systems

Traditional : via knowledge Engineers

New Trend : via Automatic programs

KDD process :

Non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data

Non trivial process — multiple process

valid — Justified patterns / models

novel — previously unknown

useful — can be used

understandable — by human & machine

1, 8, 16, 22, 23
26, 33, 34, 38
36, 42, 44, 51

(5) Putting results in practice use

(4) Interpret and evaluate discovered patterns

(3) Data Mining Extract pattern/Model

(2) collect and preprocess data

(1) understand domain and define problems

KDD is inherently interactive and iterative

KDD consist of iterative seaquence of following steps

(i) Data cleaning – to remove noise and inconsistent data

(ii) Data Integration — Multiple data sources may be combined

(iii) Data Selection – Data relevant to the analysis task are retrieved from database.

(iv) Data Transformation — data are T^r or consolidated into appropriate forms for mining by performing summary & aggregation operation

(v) Data Mining — essential process where Intelligent methods are applied inorder to Extract data patterns

(vi) Pattern Evaluation — to identify truly interesting patterns representing knowledge based on some interesting measures

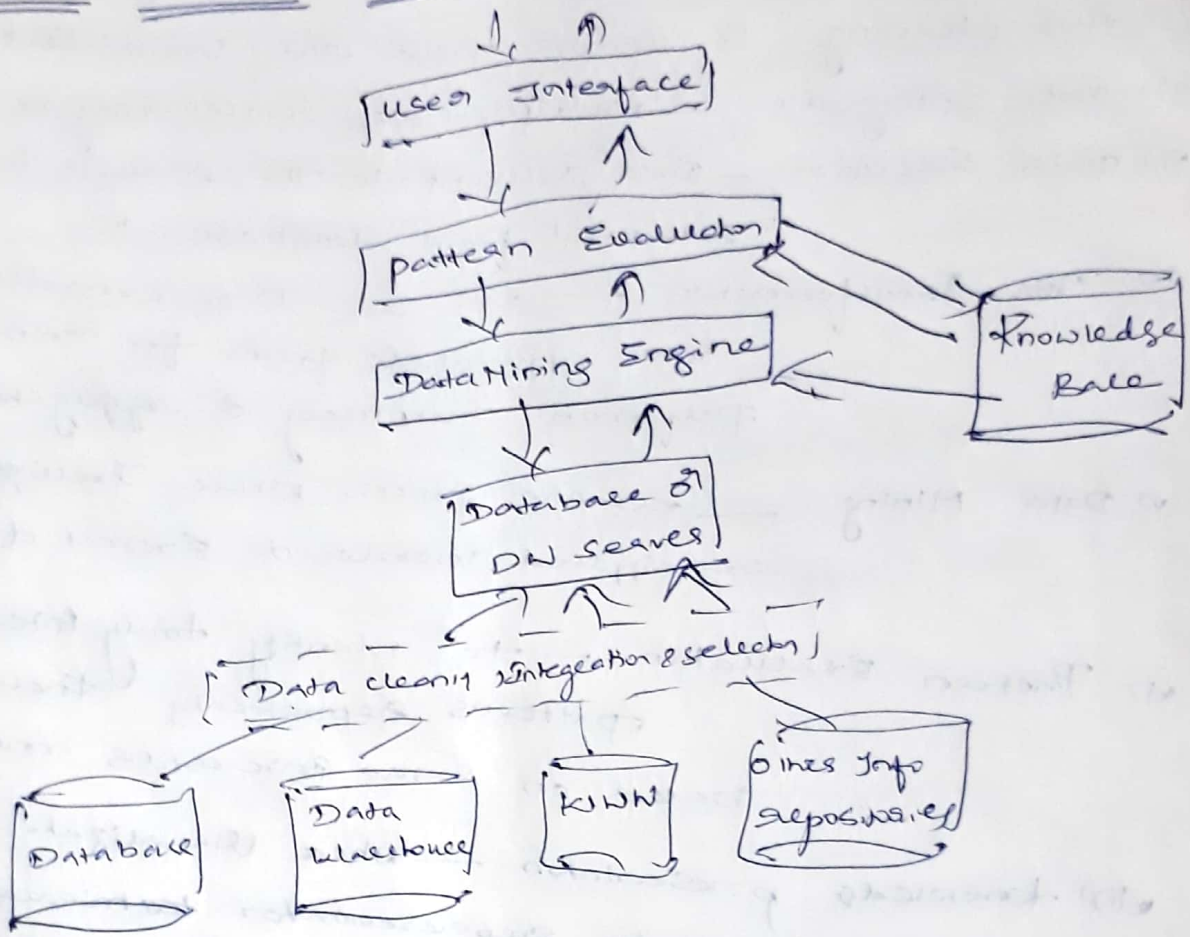
(vii) knowledge presentation — where visualization and knowledge representation techniques are used to present mined knowledge to user

(i) to (iv) are preprocessing steps. Data mining step may interact with user & knowledge base.

Interesting patterns are presented to user and may be stored as new knowledge in knowledge base

∴ Data Mining is process of discovering interesting knowledge from large amounts of data stored in databases / Warehouses - or other information repositories.

ARCHITECTURE OF TYPICAL DATA MINING SYSTEM



Database, data Warehouse, WWW or other Info repository:-
This is set of databases, DW, Spreadsheet or other repositories. Cleaning & Integration may be performed on data.

Database or Data Warehouse Server:- It is responsible for fetching relevant data based on users data mining request.

Knowledge Base:- It is used to guide the search or Evaluate interestingness of resulting patterns. Such Knowledge includes Concept hierarchies. — organize attribute in dif levels of abstraction

Data Mining Engine :- It has set of functional modules such as characterization, association, correlation analysis, classification, prediction, cluster analysis, outlier analysis.

Pattern Evaluation module :- It employs interestingness measures and interacts with Data mining module so as to focus search towards interesting patterns. For efficient DM it is recommended to push evaluation of pattern interestingness as deep as possible into mining process to confine search to only interesting patterns.

User Interface :- This module communicates b/n user & DM Sys which allows user to interact with system by specifying DM query or task, to focus search, perform data analysis based on results.

→ DATA MINING — ON WHAT KIND OF DATA :-

Data Mining should be applicable to any kind of data repository, transient data, such as data stream. Data repositories include relational databases, Data warehouses, Transactional databases, Advanced database systems, flat files, data streams & WWW. Advanced database Sys include object relational databases, Spatial, Time Series, Text and multimedia databases.

Relational Databases :-

DBMS consist of collection of interrelated data known as database & set of pgms to manage and access data.

A relational DB is collection of tables each of which has a unique name. Each table has set of attributes & stores large set of tuples. Each tuple in table represents an object

* ER model is constructed for Relational database
  └ represent set of entities & relationships

Customers

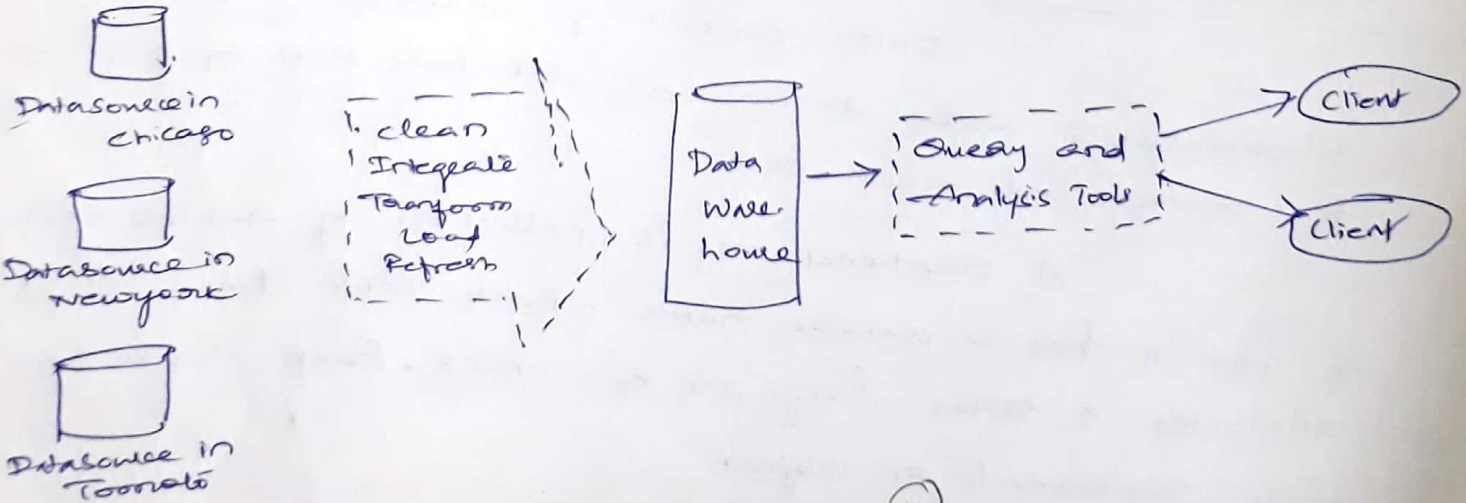
Cust_ID	name	address	age	income	Credit_info	Category
c1	Smith	--	--	--	---	--
--	--	--	--	---	--	--

Relation Customers has set of attributes, a unique Customer id, Cust_name, age etc.

* Relational data can be accessed by database Queries written in SQL or they uses Interfaces. The Query is transformed to set of relational operations such as join, selection and projection & then optimized for efficient processing. Relational language include aggregate functions.

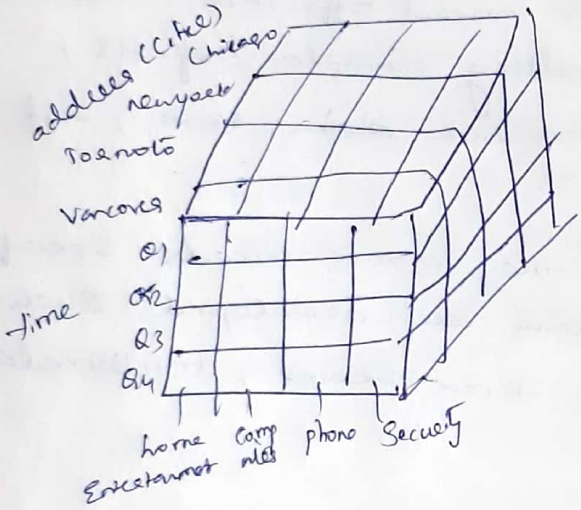
Data Warehouse :-

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema and resides at single site. Data Warehouse is constructed via cleaning, integration, transformation, loading & periodic data refreshing.



(5)

* Data are stored to provide info from historical perspective and are summarized. DW is modelled by a multidimensional database structure where each dimension corresponds to attributes, each cell stores aggregate measure.
* Physical structure — Multidimensional data cube
* Data cube — provide multidimensional view of data & allows precomputation, fast access of summarized data.



Cube has 3 dimensions address, time, item

* A Datawarehouse collects information about subjects that span an Entire organization

DataMart:- — department subset of a datawarehouse
It focuses on selected subjects

* Datawarehouses are well suited for OLAP — which use background knowledge to present data at diff levels of abstraction. OLAP operations — drill down
roll up

Transactional Database:-

It consist of a file where each record is a transaction. It has unique transaction identity number & list of items makes up a transaction

* It has additional tables

trans-ID	list of item-IDs
T100	I1, I3, I5, I6
T200	I2, I8

(11)

Advanced data & Information systems and advanced applications :-

Spatial data — maps

Engineering design data — design of building, system components, IC's

Hypertext & multimedia data — Text, image, audio, video data

Time related data — Stock Exchange data

Stream data — Sensed data, video surveillance

World wide web — available by Internet.

All above applications need efficient data storage and scalable methods for handling complex objects, variable length records, semi-structure data, text, and multimedia data

In response to the above needs specific application oriented database systems are developed such as object-relational, temporal, time series, multimedia, legacy etc

Object-relational Databases :-

* It is constructed based on object relational data model
* This model extends relational model
* This model integrates concepts of object oriented database where each entity is considered as object.

Each object is associated with following.

- Set of variables — describe object
  - corresponds to attributes in ER Model.
- Set of message — objects use it to communicate with other object.
- Set of methods — Each method holds code to implement a message.

UNIT - 2...

2.1 DATAWAREHOUSE :-

Datawarehouse refers to database ie maintained separately from an Organization's operational database. Integerates variety of application system.

By William H. Inmon

"A Data Warehouse is subject oriented, integerated, Time Variant, non voltaile collection of data in support of managements decision making process"

Subject - oriented — A DW is organized around major subject such as Customers, supplier, product. Rather than concerned on day to day and T^x processing DW focuses on modelly and analysis of data for decision makers. So DW Concerns only on concise view eliminating data that are not useful for decision support

Integrated — A DW is usually constructed by integerating multiple heterogeneous sources such as relational database, flat files and online T^x records. Data cleaning and data integration techniques are applied to ensure Consistency in naming conventions, Encoding structures, Measures

Time - Variant — Data is stored to provide info from a historical perspective of data.

Non - Voltaile — A DW is separate store of data T^x from application data found in Operational Environment. DW doesn't require T^x processing, recovery and Concurrency Control mechanisms.

Requires 2 operations for accessing { initial loading, access of data

* Construction of DW needs cleaning, Integration & Consolidation. Utilization of Datawarehouse needs Collection of decision support technologies

How Organizations are using IT from datawarehouse :

* Many of them use this IT to support decision making activities
  - > Increasing customers focus - analysis of customers buying patterns
  - > Managing product portfolios, Comparing performance of sales by Quarter, year, by geographic region to fine tune production strategies
  - > Analyzing operation & looking for sources of profit
  - > Managing customers relationship, Cost of corporate assets, environmental correction.

* On top of multiple heterogeneous database we should build Wrappers and Integrator. In traditional approach when a Query is posed to client site metadata dictionary is used to translate Queries into appropriate for individual heterogeneous site involved. Queries are mapped to local Query processors. Results are merged to form a global answer set.

  Query driven approach — Requires Complex filtering & Integration process and competes for resources with processing at local sources. It is inefficient, expensive for Queries with aggregates

  Update-Driven approach — IT is integrated in advance and stored in warehouse for querying & analysis.

  It has high performance. : data are copied, Integrated, preprocessed, annotated, summarized & restructured into one Semantic data store. It stores historical information & can support Complex multidimensional Queries.

objects that share common properties (6) are group as object class. Each object is an instance of the class

Temporal Database, Sequence & Time-Series Database:

↓

Stores relational data which include time related attributes. These attributes involve diff timestamps each has diff semantics

Sequence — stores sequence of ordered events with or w/o notion of time

Ex — weblick streams
Biological sequences

Time Series — stores sequence of value or events obtained over repeated measurement of time

Ex — hourly, daily, weekly.

DM techniques are used to find characteristics of object Evolution. Such info is useful for decision making

Spatial and Spatiotemporal Database:

↓

Spatial related information.

Ex — geographic databases, VLSI or CAD databases
Medical, Satellite image databases

> Represented in raster format — n dimensional bit map or pixmaps

Ex — 2D Satellite image in raster data where each pel registers rainfall in given area

> Maps — vector format
Roads, bridges, buildings are represented as unions or overlays of basic constructs such as point, line, polygon.

> Geographic database — used in forestry, ecology planning, Vehicle navigation, dispatching system.

Ex - Taxi would store city map with info how to move from Region A to region B during rush hour.

> Clusters and outliers are identified by spatial cluster Analysis

> Spatial classification can be performed to construct models for prediction

* Spatial databases that store spatial objects that change with time is called spatiotemporal database.

Text Databases and Multimedia databases :-

↓ contain word description for objects.

↳ long sentences, paragraphs, product specification, error or bug report, warning messages, summary reports.

> Highly unstructured but some text database may be semi structured or well structured.

> By mining text data we may uncover general & concise descriptions of text documents, clustering behaviour of text objects so we need to integrate DM techniques with IR retrieval techniques

Multimedia — store image, audio, video

> used in picture content based retrieval, voice-mail systems, WWW, speech based user IF that recognize spoken commands

> These database must support large objects

> It need specialized storage and search techniques ; there should'nt be gaps in picture or sound and also to avoid system buffer overflows.

> we need to integrate DM techniques with storage & search techniques

Heterogeneous Database :- — Consist of set of interconnected, autonomous component databases. They communicate to exchange dlf and answer queries. objects differ greatly from one component to other

Legacy Database :- - group of heterogeneous databases

> combines dlf kinds of databases
> Heterogenous in legacy are connected by inter or intra computer N/w.

I/f exchange is difficult so we need precise transformation rules.

Ex — Exchange of i/f ~~se~~gardy student academic performance b/n 2 universities is difficult. each university maintains its own grading system. DM techniques provide solution to I/f exchange problem by performing statistical data distribution & correlation analysis & Transforming data into generalized & conceptual levels through which information exchange is easier.

Data streams :-

↓ data flow in & out of observation platform dynamically

features — huge or infinite volume
dynamically changing flowing in & out in fixed order
allowing small no of scans, fast response time

Example — power supply, Network traffic, Stock Exchange, web click streams, Telecommunication, Video Surveillance weather or Environment monitory.

> Not stored in any kind of repository
> we use continous query model — predefined queries constantly Evaluate data streams, collect aggregate data, report current status of data streams as concept flow.

World Wide Web :- In yahoo, Google, Alta Vista etc objects are linked together to facilitate interactive access. Capturing the users access patterns in distributed information environment is called Web usage mining.

- It is difficult for computers to understand semantic meaning of web pages and structure them in an organized way for systematic IIf retrieval & DM. Web services that provide keyword based search will help offer limited help to users. DM can often provide additional help than web based search engines
- Authoritative web page analysis based on linkages among webpages can help rank webpages based on Importance, Influence and topics

⇒ 1.2 DATAMINING FUNCTIONALITIES

↓ used to specify kind of patterns to be found in the DM task.

- Descriptive ↓ Characterize general properties of data in database
- predictive ↓ perform inference on current data in order to make predictions.

users have no idea of what kind of patterns may be interesting. So Data mining system should mine multiple kinds of patterns.

Data mining functionalities ~~are~~ ki are given below

Concept / Class Description : Characterization and Discrimination:

Data is associated with classes & concepts

Ex Classes of item for sale — computers, printers
concepts of customers — bigspenders, budget spenders

we can describe classes, concepts in precise terms called as concept / class descript (16)

Mining Various Kinds of Knowledge and Association:-

Descriptions are derived via characterization & discrimination

Characterization — Summarization of general characteristics or features of target class of data. Data of user specified class are collected by Query.

There are several methods for summarization & characterization.

Ex — OLAP roll up operation for summarization

Attribute oriented Induction — for generalization w/o user interaction

o/p forms — pie charts, bar charts, curves, Multidimensional data cube, Multidimensional tables with Correlation

Characteristic rules ie rule form

Example — DM system should produce summarizing the characteristics of customers who spend more than $1000 a year. Result is customer of 40-50 years, employed, Excellent Credit ratings. System should drill down on dimension occupation to view these customers according to type of Employment.

Discrimination :- Comparison of general character of target class data object with one or set of contrasting classes. Target and Contrasting classes are specified by user & objects are retrieved through Queries.

Discrimination descriptions include comparative measures to distinguish b/n target and contrasting classes.

o/p form — Expressed in rule called discriminant rules.

Ex — Comparison b/n who shop regularly & who rarely shop. On query we get some o/p. We can drill down on occupation or add income-level which helps to find even more discriminative features b/n 2 classes.

Mining frequent patterns, Associations and correlations

Frequent patterns — patterns that occur frequently in data

Frequent itemset — set of items that frequently appear together in dataset

Frequent subsequence — pattern that customers tend to purchase sequential pattern a first PC followed by camera then memory card

Substructure — diff structural forms such as trees, lattices combined with itemsets & subsequences

↓

If occur frequently called as structured patterns.

Example of a rule is

buys (X, "computer") ⇒ buys (X, "software") [support = 1% confidence = 50%]

X → Customer

Confidence — If cust buys computer there is 50% chance that she buys s/w.

Support — 1% of transactions shows computer and software are purchased together.

Single dimension Association rule — ARs that contain single predicate

Multi dimension Associate rule — one or more predicates

Ex -

age (X, "20...29") ∧ income (X, "20K - 29K") ⇒

buys (X, "CDplayer") [support = 2% confidence = 60%]

* ARs are uninteresting if they don't satisfy both minimum support threshold and minimum confidence threshold

Classification and prediction:-

* process of finding a model that describes & distinguishes classes & use the model to predict class of objects whose class label is unknown. Model is based on analysis of training data (class label is known)
* presentation of derived Model — classification rules, decision trees, Neural N/w, Mathematical formulae.

* Decision tree — flow chart like structure
  - ↓ easily converts to classification rule
  - Node — test on attribute value
  - branch — outcome of test
  - leaves — classes & class distribution
* Neural Network — collection of neuron like processing units with weighted connection b/n units.
* Some other methods — Bayesian, SVM, k-nearest neighbour
* classification predicts categorical labels

prediction — models continuous valued fns
— used to predict missing values or unavailable data

* Method used for prediction — ~~Regression~~ analysis

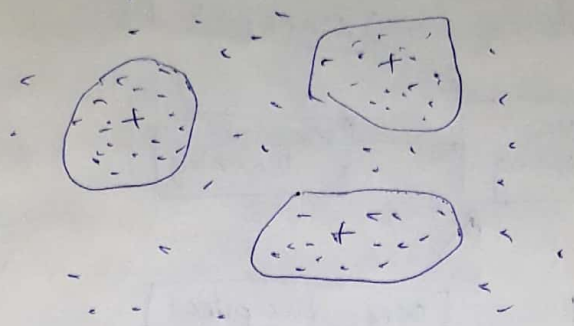
Example of classification — 3 classes — good response, avg, no response

we have to derive model for 3 classes based on feature of item i.e price, type, brand ..... The result should minimally distinguish each class. If classification is expressed in decision tree then tree may find price as the one that best distinguish classes and tree is helpful to understand impact of sales & design more effective campaign.

Rather than predicting categorical if we predict the amount of revenue that each item generate during an upcoming sale based on previous sale data, i'e nothing is prediction.

Cluster Analysis :-

↳ analyzes class objects w/o consulting known class label



* Class labels are not present in Training data. Clustering is used to generate such labels.

Principle — Maximizing intra class similarity and minimizing Inter class similarity

Objects within a cluster have high similarity to one other but are very dissimilar to objects in other clusters. From each cluster we can derive rules.

Outlier Analysis :-

Database contain data objects that do not comply with general behaviour of model of data. Such data objects are called as outliers. DM system discard outliers. Analysis of outlier data is referred as outlier mining

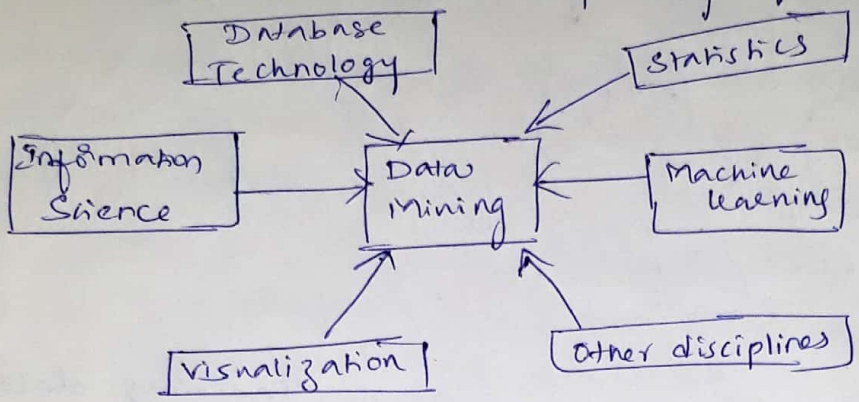
Detection of outliers → using statistical tests ie probability model
> using distance measures
> deviation based methods - by examining differences in main characteristics of objects in group

Evolution Analysis :- describes and models trends or regularity of objects whose behaviour changes over time

It includes all DM functionalities.

1.3 CLASSIFICATION OF DATA MINING SYSTEMS

DM is a interdisciplinary field ie set of disciplines.



Depending on DM approach used, techniques from other disciplines are also used such as NN, fuzzy & roughset theory DM tec may also integrate techniques from spatial data Analysis, IR, Retrieval, Image analysis, computer graphics, Business, psychology, Bioinformatics etc.

* Classification of DM systems help users distinguish b/n such systems & identify those that best match their needs DM systems can be categorized according to various criteria

Classification according to the kinds of databases mined:

A DM system can be classified based on kinds of databases mined. Different criteria are data models, types of data or application involved each requires its own DM technique.

Ex If it is data model we use relational, Transactional, or Datawarehouse mining system.

If it is special type of data we have spatial data mining & multimedia DM or world wide web mining system.

Classification according to kinds of knowledge mined:-

ie Based on characterization, discrimination, association & classification, prediction, clustering, Outlier analysis.

DM system can be distinguished based on granularity or levels of abstraction of knowledge mined. A DM system should discover knowledge at diff levels.

Differences between operational Database Systems & DW :-

The major task of on-line operational database systems is to perform on-line transaction and Query processing and these systems are called as OLTP (On-line Transaction Processing) systems .. — day-to-day Operations

OLAP Systems — DW use users as knowledge workers for analysis and decision making . These systems organize & present data in various formats

Major distinguishing features

Users and System Orientation :

OLTP — customer Oriented , IT & Query processing by clerks, clients)

OLAP — Market Oriented . Data Analysis by knowledge workers (Managers, Executives , Analysts

Data Content —

OLTP — manage current data — used for decision making

OLAP — Manage Historical data
facilitates summarization and aggregation
Manage diff at diff levels of granularity

Database design —

OLTP — E-R Model
Application oriented database design

OLAP — Star or Snowflake Model
Subject Oriented database design

View

OLTP — mainly focuses on current data within department . doesn't use historical data

OLAP — Use historical data
Spans multiple versions of database schema
data is stored on multiple storage media

Access patterns :

OLTP — has short, Atomic Transactions
Requires Concurrency Control & Recovery Mechanisms

OLAP — Access to OLAP systems are read-only.
∵ it stores historical rather than up-to-date information

Some of other features which distinguishes are database size, frequency of operations, performance metrics.

COMPARISON BETWEEN OLTP and OLAP SYSTEMS

Feature	OLTP	OLAP

Why to have separate Warehouse?

Why can't we perform OLAP directly on databases instead of constructing a separate Warehouse?

Reason — to promote performance of both the systems

* Operation database — Tuned for indexing (hashing using primary keys, searching for particular records, optimizing "canned" Queries.
* Warehouse — Queries are complex & need computation of large group of data at summarized levels & require special organization, access and implementation methods. processing OLAP Queries in operational database degrade the performance of operational Tasks
* Operational DB need concurrent processing of Tx. Concurrency Control and Recovery mechanisms are needed to ensure Consistency and Robustness. If we apply such concurrency Control and Recovery mechanisms to OLAP Queries it may the Execution of Concurrent Tx and reduce throughput of OLTP System
* Separation is based on diff structures, Contents and Uses

Decision support Queries ⟶ need historical data
> Requires consolidation of data from heterogeneous sources

Operational database ⟶ don't maintain historical data
> for form decision making
> Contain Raw data that is to Consolidated for analysis.

Since 2 systems provide diff functionalities, diff kinds of data it necessary to maintain separate databases.

2.2 Multidimensional Data Model

DW and OLAP tools based on Multidimensional data model and model view data in form of data cube.

Data cube —— allows data to be modelled and viewed in multiple dimensions.

Dimension —— entities w.r.to organization want to keep records

Each dimension has table associated with it called as Dimension table —— Specified by users or automatically generated based on data distribution

Ex Item dimension table has describe dimension for attributes — item_name, brand, type.

* Multidimensional data model is organized around central theme like "Sales" for instance. Represented by fact table

Fact —— Numerical Measures.

for "Sales" facts —— dollars_Sold
Units_Sold

fact table —— Name of fact or measure
Key to each of related dimension table

location = "vancouver"

time (Quarter)	item (type)			
	home Entertainment	Computer	phone	Security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

* Data cube is n-dimensional

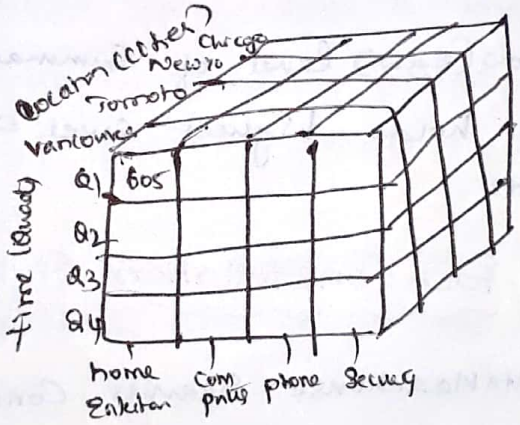
2-D representation, sales of Vancouver are shown with respect to time dimension (organized in Quarters) and item dimension (organized according to types of item sold).

If we want to view data with 3rd dimension. If we want to view data with: time, item, location — for cities. It is represented as series of 2D tables.

location = "Chicago"			location = "New York"			location = "Toronto"			location = "Vancouver"		
item			item			item			item		
time	home ent	comp	phone	sec	home ent	comp	phone	sec	home ent	comp	phone	sec	home ent	comp	phone	sec
Q1	854	882	...													
Q2	943	890														
Q3	1032	924														
Q4	1129	992														

3D view of sales data

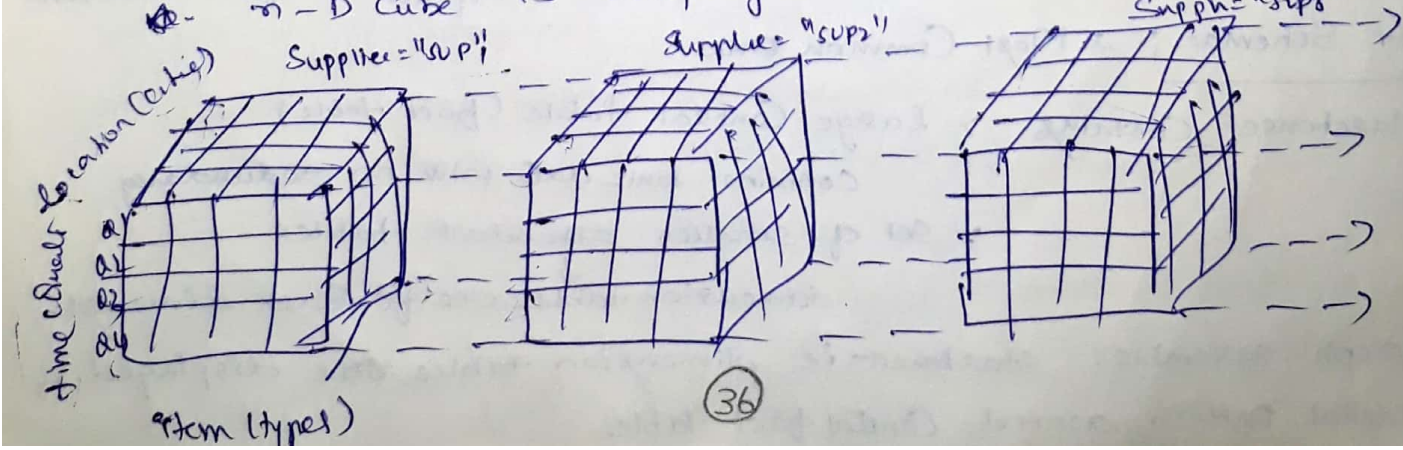
3D-Data Cube



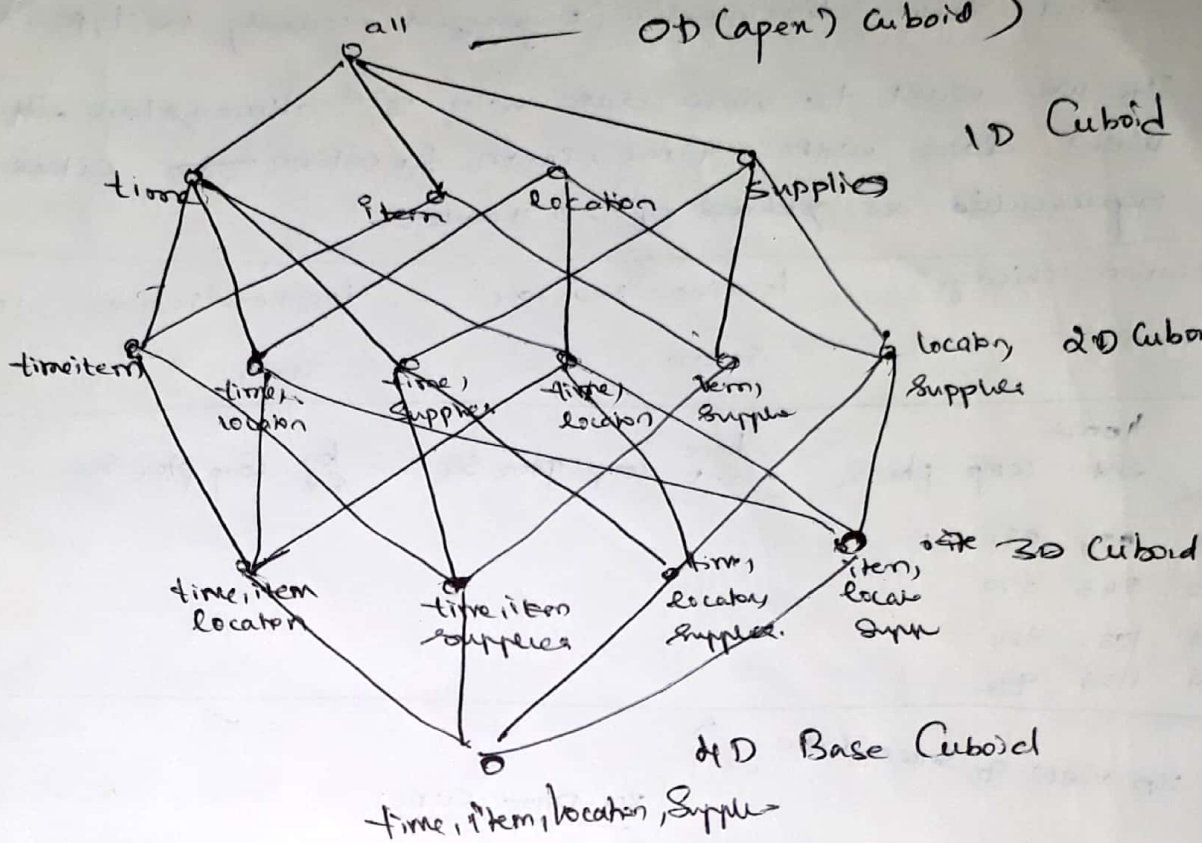
If we want to view data with additional dimension such as "Supplier"

4D Cube is conceptualized as series of 3D Cubes.

n-D Cube is displayed as series of (n-1) D Cubes.



* Data Cube is often refered to as Cuboid. For given set of dimensions we can generate possible subsets forming lattice of cuboids and data is at diff levels of summarization



Cuboid that holds low (raw) level of summarization is called Base cuboid. which holds highest level of summarization is called Apex Cuboid

Star, Snowflake, Fact Constellation Schemas

A Datawarehouse requires concise, subject oriented Schema that facilitates on-line data Analysis. The most popular data model for Warehouse is multidimensional model

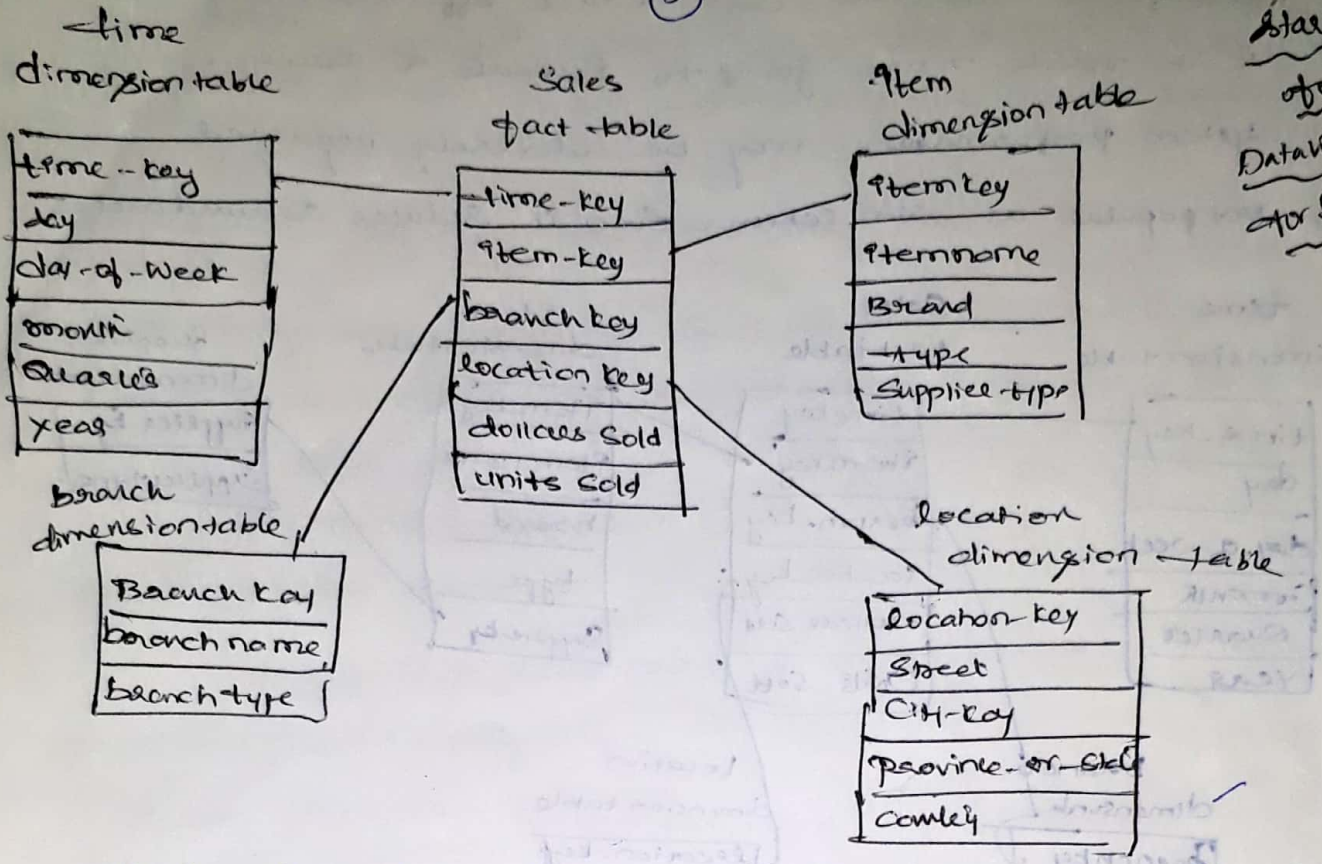
Star, snowflake, fact constellation

Star Schema :- * Most Common One

Warehouse Contains → large central table (fact table)
contains bulk data with no redundancy
→ set of smaller attendant tables
dimension table one for each dimension.

Graph resembles starburst - ie dimension tables are displayed in radial pattern around central fact table.

33 Star Schema of a Data Warehouse for Sales



Each dimension is represented by only one table. Each table contains set of attributes. The above schema contains central fact table for Sales that contains keys to each of four dimensions along with 2 measures.

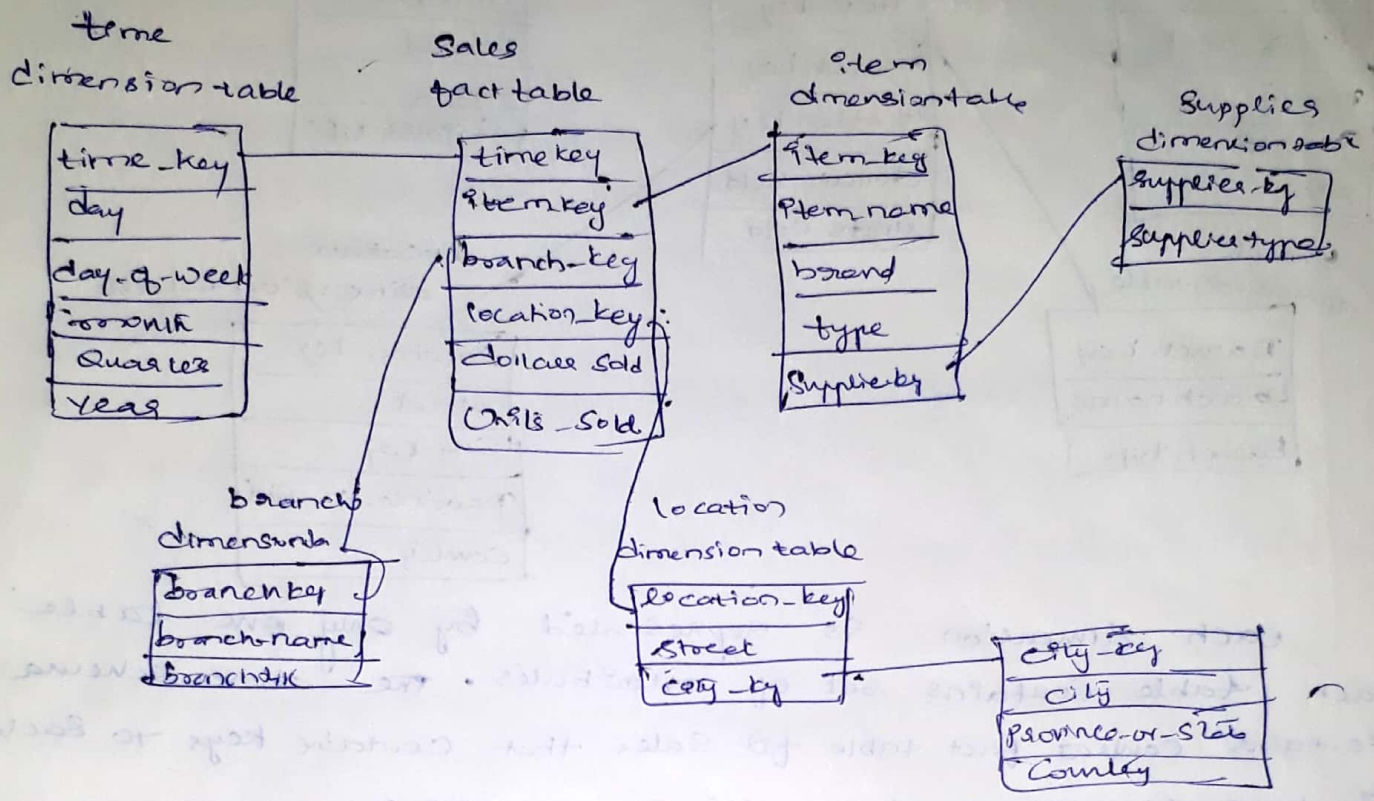
* location dimension contains attribute set {location_key, street, city, province-or-state, country}. It has some redundancy

ie Vancouver and Victoria are both cities in Canadian province of British Columbia. Such Entries causes redundancy among attributes ie (---, Vancouver, British Columbia, Canada---) and (--, Victoria, BritishColu, Canada, ---)

Snowflake Schema:-

* Variant of Starflake
* Some dimension tables are Normalized
  ie spitting data into additional tables
* Shape is similar to Snowflake
* Dimension tables are normalized to reduce redundancies
  ↓ reduces space, easy to maintain

* Snowflake Structure can reduce effectiveness of browsing
  - it needs more joins to Execute a Query.
  System performance may be adversely impacted
* Not popular as Star schema Even it reduces redundancies.

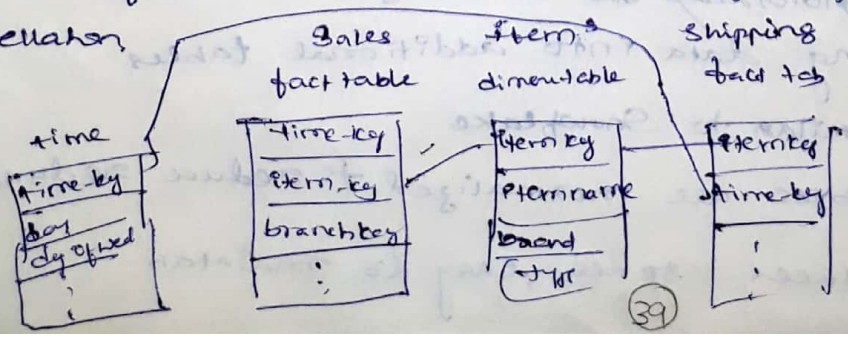


→ Here item table is normalized in Snowflake resulting in new item and Supplies tables. Here in item dimension table Supplies_key is linked to Supplies dimension table which has Supplier key and supplier type information.

Location table is normalized to location, city

city_key of new location table points to city dimension further normalization can be performed on state and country in Snowflake

Fact Constellation - Sophisticated applications require multiple fact tables to share dimension tables. It is viewed as collection of stars and is called as galaxy schema or fact constellation



In the above Examples Schema (B) has 2 fact tables — sales, — shipping. Fact Constellation Schema allows dimension tables to be shared between fact tables.

Examples for defining star, snowflake & fact constellation schema:

* Data mining query language is used to specify data mining tasks. DM and data marts are defined using a language primitives.
  - > Cube definition
  - > dimension definition

Cube definition Syntax:

define cube <cube-name> [<dimension-list>]: <measure-list>

dimension definition Syntax

define dimension <dimension_name> as ((attribute & dimension list))

Star Schema definition

define cube sales-star [time, item, branch, location]:

dollars-sold = sum (sales-in-dollars), units_sold = count(*)

define dimension time as (time-key, day-day-of-week, month, quarter, year)
define dimension item as (item-key, item-name, brand, type, supplier-type)
define dimension branch as (branch-key, branch-name, branch-type ----)
define dimension location as (location-key, street, city ----)

define cube define a data cube called sales star which contains. Corresponds to fact table. It has 2 measures dollars-sold, units-sold.

Snowflake definition.

define cube sales-snowflake [time, item, branch, location]:
dollars-sold = sum (sales-in-dollars), units_sold = count(*).

define dimension time as
define dimension item as (item-key, item-name, brand, type, supplier
(supplier-key, supplier-type))

define dimension branch as (

define dimension location as (location-key, street, city (city-key)
city, province or state, country)

Fact Constellation schema definition:

define cube sales [time, item, branch, location]:
dollars_sold = sum (sales_in_dollars), units_sold = count(*).
define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier_type)
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key, street, city, state, country)

define cube shipping [time, item, shipper, from_location, to_location]:
dollars_cost = sum (cost_in_dollars), units_shipped = count(*).
define dimension time as time in cube sales
define dimension item as item in cube sales
define dimension shipper as (shipper_key, shipper_name, location as location in cube sales, shipper_type)
define dimension from_location as location in cube sales
define dimension to_location as location in cube sales

Measures: Their Categorization and Computation

A data cube measure is a numerical function that can be evaluated at each point in data cube space.

Measures are categorized into 3 categories
- Distributive
- Algebraic
- Holistic

Distributive:-

An aggregate fn is distributive if it can be computed in a distributed manner. Data is partioned into n sets. We apply fn to each partition resulting in 'n' aggregate values. If result derived by applying fn to n aggregate values is same as that derived by applying fn to entire data set the 'fn' can be computed in distributed manner.

A measure is distributive if it can obtained by applying a distributive aggregate function.

Ex: count(), sum()
max()
min()

In the above Examples schema has 2 fact tables — sales, shipping. fact constellation schema allows dimension tables to be shared between fact tables.

Examples for defining star, snowflake & fact constellation schema:

* Data mining query language is used to specify data mining tasks. DM and data marts are defined using 2 language primitives.
  - > Cube definition
  - > dimension definition

Cube definition Syntax:

define cube <cube-name> [<dimension-list>]: <measure-list>

dimension definition Syntax:

define dimension <dimension_name> as ((attribute & dimension list))

Star Schema definition

define cube sales-star [time, item, branch, location]:

dollars_sold = sum (sales_in_dollars), units_sold = count(*)

define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier_type)
define dimension branch as (branch_key, branch_name, branch_type ---)
define dimension location as (location_key, street, city ---)

define cube define a data cube called sales star which corresponds to fact table. It has 2 measures dollars_sold, units_sold.

Snowflake definition.

define cube sales_snowflake [time, item, branch, location]:

dollars_sold = sum (sales_in_dollars), units_sold = count(*).

define dimension time as
define dimension item as (item_key, item_name, brand, type, supplier (supplier_key, supplier_type))

define dimension branch as (
define dimension location as (location_key, street, city (city_key, city, province_or_state, country)

Algebraic — An aggregate fn is algebraic if it can be computed by an algebraic fn of M arguments Each of which is obtained by applying distributive aggregative function. A measure is algebraic if it is obtained by applying aggregate fn

Ex — avg() → sum() / count()
min_N()
max_N()
standard_deviation()

Holistic — An aggregate fn is holistic if there is no constant bound on storage size needed to describe subaggregate. ie there doesn't exist an algebraic fn with M arguments.

A measure is holistic if it is obtained by applying holistic aggregate fn.

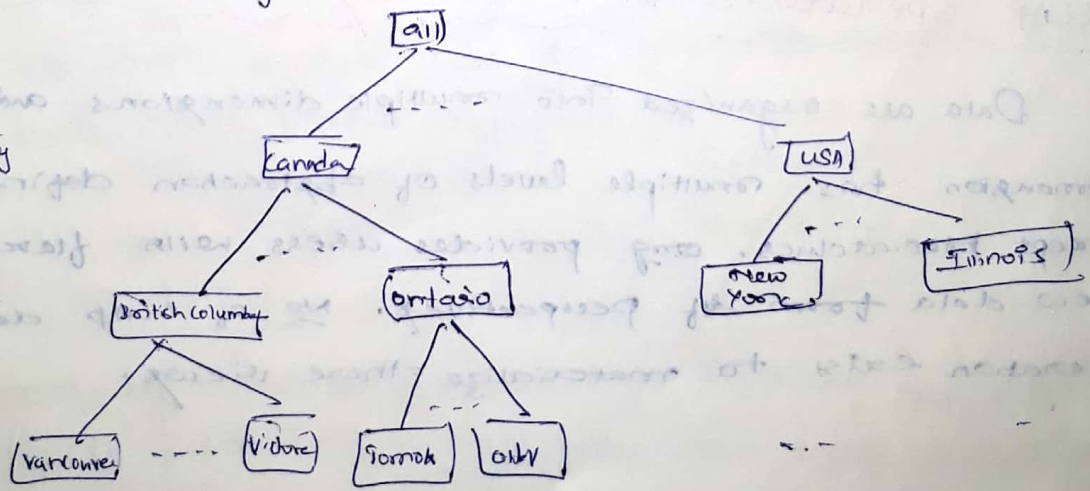
Ex — median()
mode()
rank()

Concept Hierarchies:-

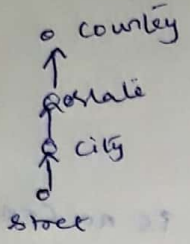
A Concept hierarchy defines a sequence of mapping from a set of low level concepts to higher level concepts

Ex - location dimension

Each city mapped to state to which it belong. province are mapped to country to which they belong. These mapping form a concept hierarchy for dimension location mapping a set of low level concepts to higher level concepts



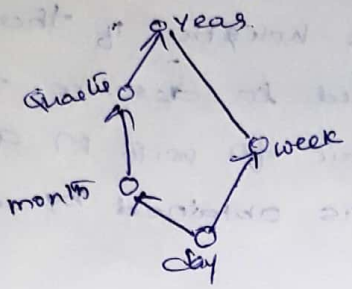
* Many concept hierarchies are implicit within database attributes of location are related by total order forming a concept hierarchy ie "street < city < state < country"



Attributes may also be organized in partial order forming a lattice.

Ex time dimension

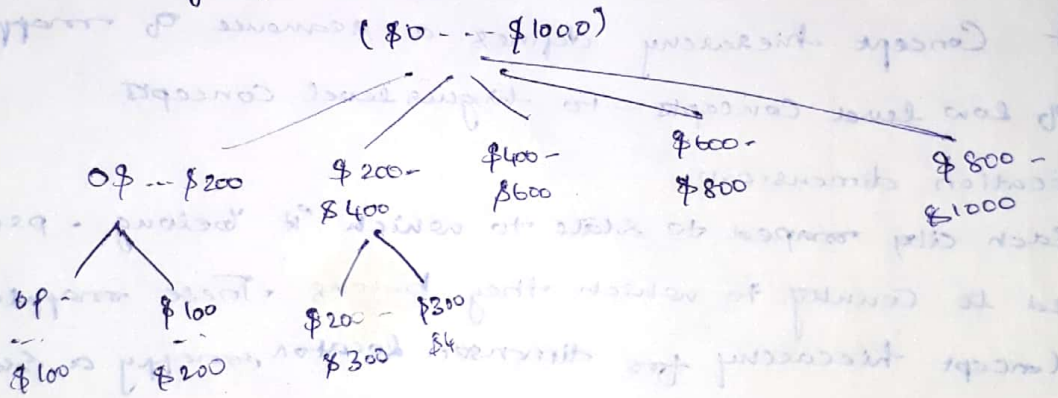
"day < {month < quarter, week} < year"



A concept hierarchy is total or partial order among attributes in a database schema is called "schema hierarchy"

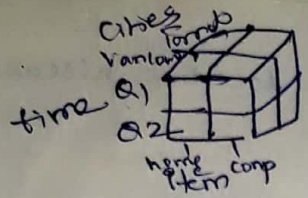
* Concept hierarchies is also defined by discretizing or grouping values for given dimension or attribute, resulting in a set-grouping hierarchy.

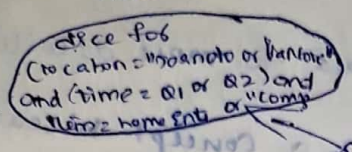
Ex of Set grouping hierarchy

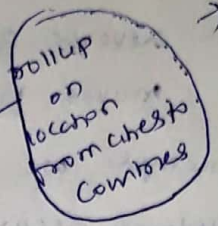


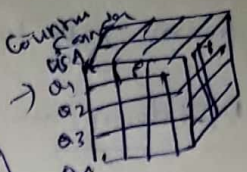
OLAP operations in Multidimensional Data Model

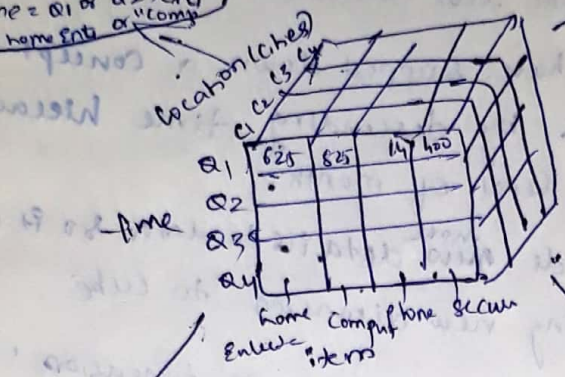
Data are organized into multiple dimensions and each dimension has multiple levels of abstraction defined by the concept hierarchies. org provides users with flexibility to view data from different perspectives. No of OLAP data cube operation exist to materialize these views.

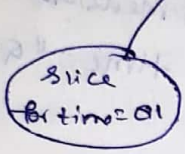


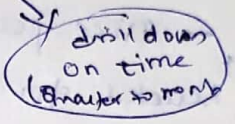


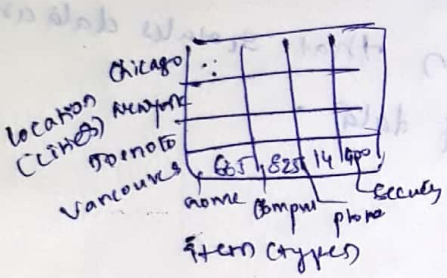


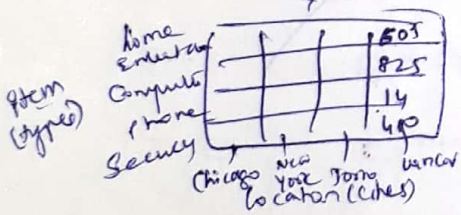


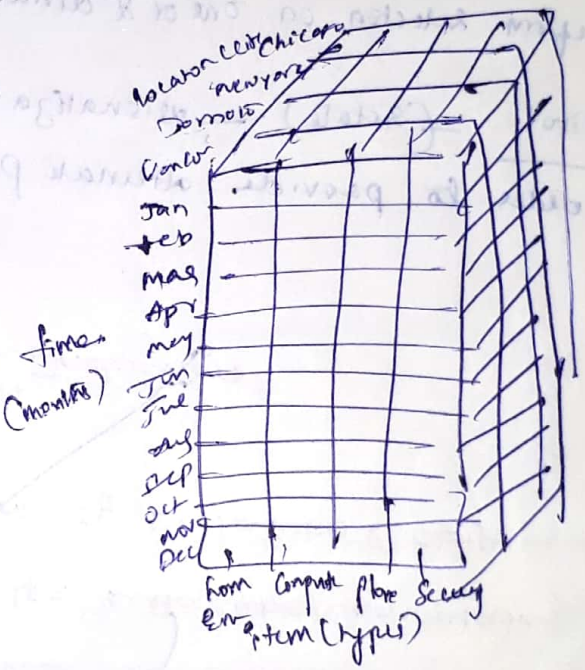












At the centre we have cube for sales. It contains dimensions location, time, item. location aggregated to cities
time " quarters
item " item types

Rollup :- (or) Drill up

Performs by climbing up a concept hierarchy for dimension. hierarchy is defined with total order street < city < state < country

Roll up operation aggregates data by ascending location hierarchy from city to country

Drill down — reverse of roll up
moves from less detailed data to more detailed data
In the example we have stepped down a concept hierarchy for time. Drilldown occurs by descending time hierarchy from quarter to more detailed level of month.

Drill down adds more details to data so it can also be performed by adding new dimension to cube

Slice & dice — slice performs selection on one dimension. In a sales data is selected for time dimension using time = "Q1". Dice operation perform selection on one or more dimension

pivot (rotate) — visualization operation that rotates data axes in order to provide alternate presentation of data.

OLAP Systems Vs Statistical Databases :-

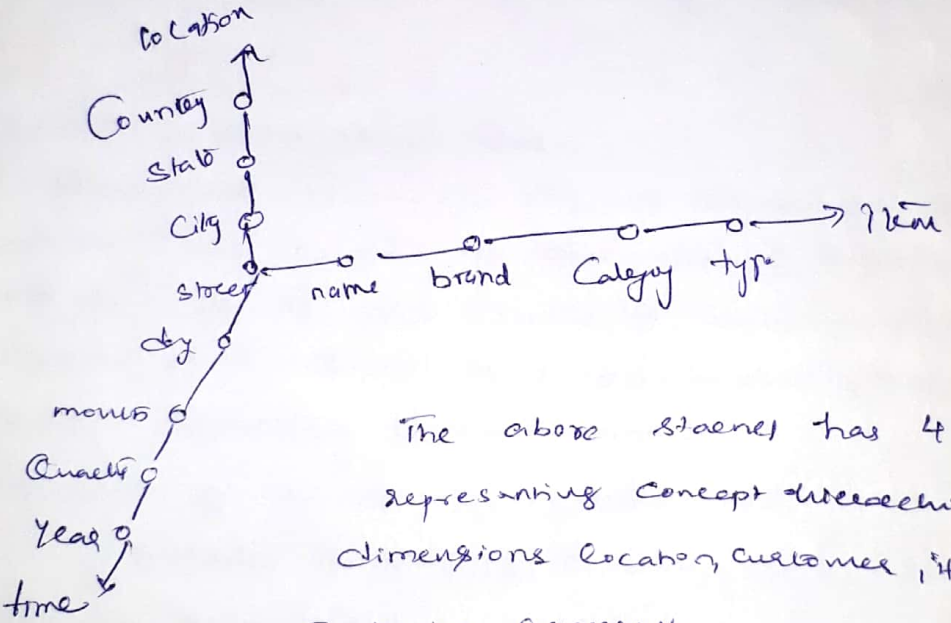
A statistical database is a database system that is designed to support statistical information

SDB — focuses on socioeconomic applications, privacy issues of concept hierarchy

OLAP — Business applications

Starnet Query Model for Querying MDB :-

Querying of MDBs can be based on starnet model. It consist of radial lines emanating from central point where each line represents concept hierarchy for dimension. Each abstraction level on hierarchy is called a footprint.



The above starnet has 4 radial lines. representing concept hierarchies for 4 dimensions location, customer, item, time. Footprint represent — levels of dimension

* Concept hierarchies are used to "generalize" data by replacing low level values by higher level abstractions. (or) to specialize data by replacing higher level abstractions with lower level values.

Data Warehouse Back-End Tools and Utilities

- DW systems use back end tools and utilities. Tools include following fn

Data Extraction — (generates data tool) gathers data from multiple heterogeneous and External sources

Data cleaning — detect Errors in data & rectify them

Data Transformation — Convert data from legacy or host format to warehouse format

Load — sorts, Summarizes, Consolidates, Computes views, Check Integrity, build indices and partitions.

Refresh — propogates updates from data source to warehouse

In addition to these metadata definition tools & warehouse usually provide a set of warehouse management tools

Metadata Repository

Metadata — data about data

In warehouse metadata defines warehouse objects.

- Metadata are created for datanames & definitions of given warehouse. It is also created for timestamping any Extracted data, Source of Extracted data, Missing field

Metadata Repository Contain following

> Description of structure of datawarehouse
includes warehouse schema, view, dimensions, hierarchy, derived data definition

> Operational metadata
includes data lineage (history of migrated data & transformats applied to it)
Currency of data (active, archived, purged)
Monitoring info

> alg used for Summarization (47)
include measure & dimension definition alg, data on granularity, Partitions, aggregation, Summarization predefined queries & reports

> Mapping from Operational Environment to datawarehouse
include source databases & its contents
gateway descriptions
Data Extraction, cleaning, T^r rules
data Refresh and purging rules
Security (user authorization and access control)

> Data related to system performance
include indices, profiles that ↑ data access & retrieval performance
Rules for timing & scheduling
Replication cycles

> Business metadata
include business terms, definitions
Data Ownership I/f
Charging policies

Types of OLAP Servers

For OLAP processing Implementation of Warehouse servers include following

Relational OLAP (ROLAP) servers:

> It lies b/n Relational back End server and client front end tools
> It uses relational or Extended relational DBMS to store and manage Warehouse data
> These servers include optimization for Each DBMS back End, Implementation of aggregation Navigation logic, additional tools
> Data is stored in form of table about not multidimensionally

Advantage — high speed of accessing data
large storage space
greater Scalability than MOLAP

Characteristics — All OLAP features & fn like cleaning, Extraction, are supported by ROLAP servers
- data is stored in relational format
- have capability of applying some form of aggregation

Multidimensional OLAP (MOLAP) Servers :-

- Support multidimensional View of data thru array based multidimensional storage engines.
- They map multidimensional views directly to data cube array structure.
- Cube allows fast indexing to precomputed summarized data
- Storage utilization is low if data set is sparse and so they use sparse matrix compression techniques
- MOLAP servers handle both sparse and dense data and adopt a 2 level storage representation to handle such data.

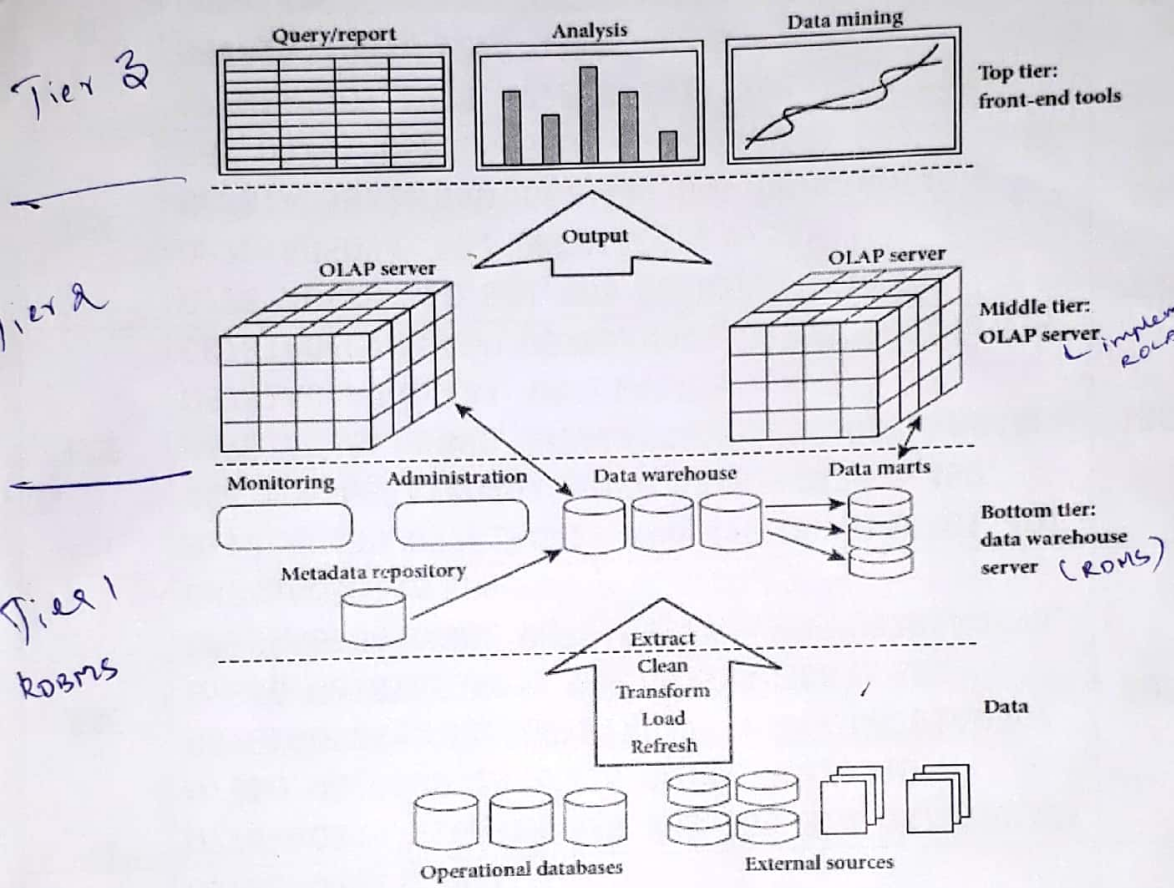
Hybrid OLAP (HOLAP) Servers :- Combines ROLAP and MOLAP ie benefiting from greater scalability of ROLAP and faster computation of MOLAP.

Specialized SQL Servers

* To meet growing demand of OLAP processing :- provides advanced query language and query processing support for SQL queries

2.4 DATAWAREHOUSE IMPLEMENTATION

1.9.2 A Three Tier Data Warehouse Architecture:



Tier-1:

The bottom tier is a warehouse database server that is almost always a relationaldatabase system. Back-end tools and utilities are used to feed data into the bottomtier from operational databases or other external sources (such as customer profileinformation provided by external consultants). These tools and utilities performdataextraction, cleaning, and transformation (e.g., to merge similar data from differentsources into a unified format), as well as load and refresh functions to update thedata warehouse . The data are extracted using application programinterfaces known as gateways. A gateway is

supported by the underlying DBMS andallows client programs to generate SQL code to be executed at a server.

Examplesof gateways include ODBC (Open Database Connection) and OLEDB (Open Linkingand Embedding for Databases) by Microsoft and JDBC (Java Database Connection).

This tier also contains a metadata repository, which stores information aboutthe data warehouse and its contents.

Tier-2:

The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model or a multidimensional OLAP.

- OLAP model is an extended relational DBMS thatmaps operations on multidimensional data to standard relational operations.
- A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

Tier-3:

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

P. Jyotheeswari
Assoc Prof
SVCET.

1.9.3 Data Warehouse Models:

There are three data warehouse models.

1. Enterprise warehouse:

- An enterprise warehouse collects all of the information about subjects spanning the entire organization.
- It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
- It typically contains detailed data aswell as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.
- An enterprise data warehouse may be implemented on traditional mainframes, computer superservers, or parallel architecture platforms. It requires extensive business modeling and may take years to design and build.

2. Data mart:

- A data mart contains a subset of corporate-wide data that is of value to aspecific group of users. The scope is confined to specific selected subjects. For example,a marketing data mart may confine its subjects to customer, item, and sales. Thedata contained in data marts tend to be summarized.

- Data marts are usually implemented on low-cost departmental servers that areUNIX/LINUX- or Windows-based. The implementation cycle of a data mart ismore likely to be measured in weeks rather than months or years. However, itmay involve complex integration in the long run if its design and planning werenot enterprise-wide.

- Depending on the source of data, data marts can be categorized as independent ordependent. Independent data marts are sourced from data captured from one or moreoperational systems or external information providers, or from data generated locally within a particular department or geographic area. Dependent data marts are sourced directly from enterprise data warehouses.

3. Virtual warehouse:

- A virtual warehouse is a set of views over operational databases. Forefficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.

1.9.4 Meta Data Repository:

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured fortimestamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

A metadata repository should contain the following:

- A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.

- Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).

- The algorithms used for summarization, which include measure and dimension definitionalgorithms, data on granularity, partitions, subject areas, aggregation, summarization,and predefined queries and reports.

- The mapping from the operational environment to the data warehouse, which includessource databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, andsecurity (user authorization and access control).

- Data related to system performance, which include indices and profiles that improvedata access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

- Business metadata, which include business terms and definitions, data ownershipinformation, and charging policies.

1.10 OLAP(Online analytical Processing):

- OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly.
- OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining.
- OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives.

OLAP consists of three basic analytical operations:

- Consolidation (Roll-Up)
- Drill-Down

➢ Slicing And Dicing

- Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions. For example, all sales offices are rolled up to the sales department or sales division to anticipate sales trends.

- The drill-down is a technique that allows users to navigate through the details. For instance, users can view the sales by individual products that make up a region's sales.

- Slicing and dicing is a feature whereby users can take out (slicing) a specific set of data of the OLAP cube and view (dicing) the slices from different viewpoints.

1.10.1 Types of OLAP:

1. Relational OLAP (ROLAP):

Base data & Relational Dimen table & tables new tables – aggregated info

- ROLAP works directly with relational databases. The base data and the dimension tables are stored as relational tables and new tables are created to hold the aggregated information. It depends on a specialized schema design.
  - This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.
  - ROLAP tools do not use pre-calculated data cubes but instead pose the query to the standard relational database and its tables in order to bring back the data required to answer the question.
  - ROLAP tools feature the ability to ask any question because the methodology does not limit to the contents of a cube. ROLAP also has the ability to drill down to the lowest level of detail in the database.

2. Multidimensional OLAP (MOLAP):

- MOLAP is the 'classic' form of OLAP and is sometimes referred to as just OLAP.
- MOLAP stores this data in an optimized multi-dimensional array storage, rather than in a relational database. Therefore it requires the pre-computation and storage of information in the cube - the operation known as processing.
- MOLAP tools generally utilize a pre-calculated data set referred to as a data cube. The data cube contains all the possible answers to a given range of questions.
- MOLAP tools have a very fast response time and the ability to quickly write back data into the data set.

3. Hybrid OLAP (HOLAP):

- There is no clear agreement across the industry as to what constitutes Hybrid OLAP, except that a database will divide data between relational and specialized storage.
- For example, for some vendors, a HOLAP database will use relational tables to hold the larger quantities of detailed data, and use specialized storage for at least some aspects of the smaller quantities of more-aggregate or less-detailed data.
- HOLAP addresses the shortcomings of MOLAP and ROLAP by combining the capabilities of both approaches.
- HOLAP tools can utilize both pre-calculated cubes and relational data sources.

DATA WAREHOUSE IMPLEMENTATION

* OLAP servers requires query results to be answered in seconds which means that DW should support efficient cube computation techniques, access methods and query processing techniques.

Efficient Computation of Data Cubes :-

In SQL Aggregation → group-by's
↓
represented by cuboid
Set of cuboids — lattice of cuboids — data cube

The Compute Cube Operator & Curse of Dimensionality

Compute Cube operator — computes aggregates over all subsets of dimensions specified in operation which requires more storage space

Example :-

AllElectronics Sales
↓ has
City, item, year, sales-in-dollars

We would like to analyze data with following queries

> "Compute sum of sales, grouping by city & item"
> "Compute sum of sales, grouping by city"
> "Compute sum of sales, grouping by item"

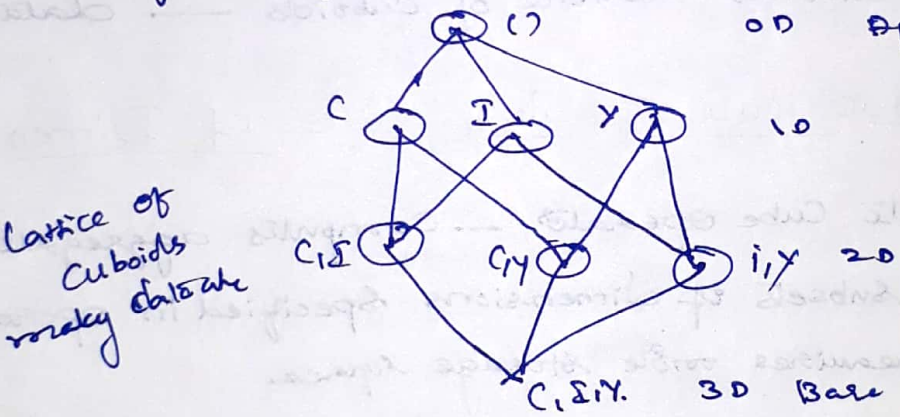
* Total no of cuboids for this data cube is $2^3 = 8$
possible group by s are {(city, item, year), (city, item),
(city, year), (item, year), (city), (item), (year), ()}
All these group by form lattice of cuboids.

Ex 0D Cuboid → ()
(Apex)
3D Cuboid → (city, item, year)

* If there are n dimensions, then there are total of 2^n cuboids

* Compute Sum of total sales → 0D operation
↓
No group by operations

one group by ie "compute Sum of Sales by city"



Lattice of Cuboids making data cube

Cube operator → Generalization of group by operators

* OLAP needs to access dif cuboids for dif queries
So all can some of cuboids should be precomputed
in advance which leads to faster response time
and avoids redundant computation — But it
requires more storage space

Curse of Dimensionality:-

If Cube has many dimensions which is associated with concept hierarchies each with multiple levels, and if all cuboids in cube are precomputed then storage requirements are Excessive and it is known as Curse of Dimensionality

No of cuboids in n D Cube

* If no hierarchies with each Dimension then the total no of cuboids for n D cube is 2^n
* If Concept hierarchies are associated

Total no of cuboids = $\prod_{i=1}^{n} (L_i + 1)$

$L_i \rightarrow$ no of levels associated with each dimension

$+1 \rightarrow$ to include virtual top level all

* If Cube has 10 Dimensions
  each dim has 5 levels.
  Total no of cuboids $\rightarrow 5^{10} \approx 9.8 \times 10^6$
* It is unrealistic to precompute and materialize all cuboids for a data cube. so Hence we go for partial Materialization — ie only some of possible cuboids are precomputed

Partial Materialization:-

* 3 choices for materialization is available

No Materialization

- precompute only base cuboid & no other cuboids
- slow computation

Full Materialization:

- precompute all cuboids
- Requires huge space

Partial Materialization selectively compute proper subset of whole set of cuboids.

- Consider 3 factors
  - > Identify cuboids to materialize — based on workload, frequency, accessing cost, storage
  - > Exploit materialized cuboids during query processing
  - > update materialized cuboids during load and refresh.

* Now we can compute iceberg cube — which stores only cube cells whose aggregate value (count) is above some minimum support threshold.

* Shell Cube — precomputing cuboids for only small no of dimensions (3 to 5)

Once selected cuboids are materialized then

* . Selection of cuboids involves how to determine relevant cuboids from among the candidate materialized cuboids.
  - how to use available index structures on the materialized cuboids
  - how to transform OLAP operations onto the selected cuboids.

* During load and Refresh

Materialized cuboids are updated efficiently. parallelism and incremental update techniques are used for this purpose.

Indexing OLAP Data:-

* Two types are indexing methods are used to index OLAP data
  - Bitmap Indexing
  - Join Indexing

Bitmap Indexing

- Most popular one
- Allows quick searching in datacubes.
- Alternative representation to RID list
- Index on a particular column
  Each distinict value in a column has bit vector / Bit
- i^{th} bit is set if i^{th} row of basetable has value for Indexed column.
- This approach is not suitable for high cardinality domains.

Base Table

Cust	Region	Type
C1	Asia	Retail
C2	Europe	Dealer
C3	Asia	Dealer
C4	America	Retail
C5	Europe	Dealer

Index on Region

RecID	Asia	Europe	America
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0

Index on Type

RecID	Retail	Dealer
1	1	0
2	0	1
3	0	1
4	1	0
5	0	1

Join Indexing

- Registers joinable rows of 2 relations
- Consider 2 relations R & S

  Let R (RID, A) RID } record identifiers
  S (SID, B) SID } from R, S relations

- For joining attributes A & B join index record contains pair (RID, SID)

In Traditional Databases

- Join Index maps Attribute Values ——to list of record ids

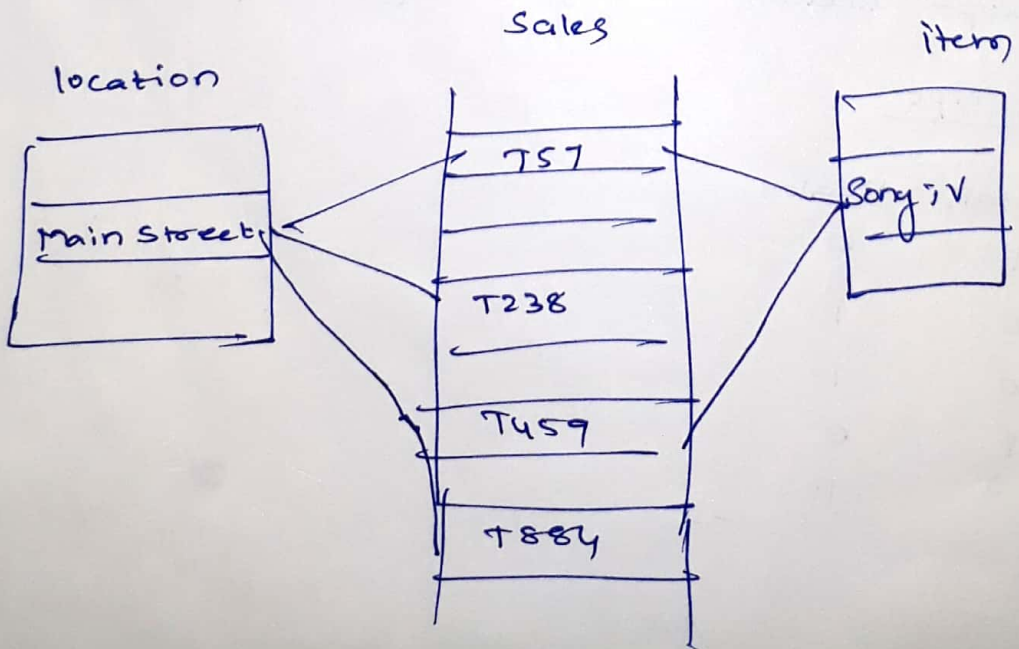
In DataWarehouse

- JoinIndex relates Values of dimensions of star schema to rows in fact table

* Join indices can span across multiple dimensions — Composite join indices

* To speedup Query processing joinIndexing and Bitmap Indexing can be integrated to form Bitmapped join indices.

Example



* Main Street value in location dimension table joins with tuples T57, T238, T884 of Sales fact table (4)
My Sony TV value in item dimension table joins with tuples T57, T459 of Sales fact table.

* Join Index Tables are shown (below) in ~~above~~ figure below

for location/sales

location	Sales-key
....	....
Main Street	T57
Main Street	T238
Main Street	T884
..	...

item/sales

item	Sales-key
...	..
Sony TV	T57
Sony TV	T459
..	...

Join Index Table linking 2 dimensions
location | item | sales

location	item	Sales
...	..	..
Main Street	Sony TV	T57
..	..	..

* If Join Indices are not used, additional I/Os have to be performed to bring joining portions of fact table and dimension table together

Efficient processing of OLAP Queries

Steps for Efficient OLAP Query processing :-

① Determine which OLAP operations should be performed on available cuboids

- Transform OLAP operations drilldown, Rollup ... to corresponding SQL and/or OLAP operations

  Ex Dicing = selection + projection

② Determine to which Materialized cuboids relevant OLAP operations should be applied.

- Involves identifying cuboids which can answer Query
- pruning of cuboids using knowledge of "dominance"
- Estimate cost of remaining materialized cuboids and Select Cuboid with least cost.

Example :-

Sales_cube [time, item, location] : sum (sales_in_dollars)

Dimension hierarchies used

day < month < Quarter < year — time

Itemname < brand < type — Item

Street < city < state < Country — location

If Query is to be processed on {brand, state} with year = 2004.

4 Materialized Cuboids

Cuboid 1 : {year, itemname, city}

Cuboid 2 : {year, brand, country}

Cuboid 3 : {year brand, state}

Cuboid 4 : {itemname, state} where year = 2004

* Which Cuboid should be selected for Query processing

Step 1 — Pruning of Cuboids

Cuboid 2 is at higher level of concept "country" and can't answer Query at lower granularity "state".

Step 2 — Estimate Cuboid cost.

Cuboid 1 cost more than 3 Cuboids..

∵ item name and city are at finer granular level than brand and state as in Query.

Step 3

If there are less no. of years and more no of itemnames under each brand then Cuboid 3 has least cost.

But if otherwise and there are efficient indexes on itemname, then Cuboid 4 have least cost.

Hence (3) or (4) can be selected accordingly

⇒ FROM DATA WAREHOUSING TO DATA MINING

How do data warehouse and OLAP relate to data mining

Data Warehouse Usage :-

* Data in DW and data marts are used to perform analysis and make strategic decisions.
DW are Extensively used in banking & financial Services, Consumer goods and retail distribution Sectors and Controlled manufacturing such as demand based production

* Longer DW in use more it is Evolved which takes place in number of phases.
first Step — generating reports & answering predefined Queries
Next Initially — DW is used to analyze Summarized & detailed data
Next — and present result in form of reports & charts.

later — used for making strategic purposes
performing multidimensional Analyses
performing slice, dice Operations.

finally — used for knowledge discovery & strategic Decision making using DM

* There are three kinds of DataWarehouse Applications
— Information processing
— Analytical processing
— Data Mining

Information processing —

- Supports Querying, Basic statistical Analysis
- Reporting using crosstab, charts & graphs
- To construct low cost web based accessing tools that are then integrated with Web browsers.

Analytical processing

- Supports basic OLAP operations
- operates on historical data in Summarized & detailed form
- Can have multidimensional analysis of DW data

Data Mining

- Supports knowledge discovery
- finds hidden patterns, association, constructing analytical models, performing classification & prediction, presenting mining results using visualization tools

Information processing is not DataMining

- If processing gives result to Queries but that is taken directly from database
- They don't have any sophisticated patterns.

OLAP is closer to DM

- Derive if ~~the~~ Summarized at multiple granularities from subsets of DW. Such descriptions are Equivalent to class/concept descriptions.

* OLAP is data summarization/aggregation tool that helps in simplifying data analysis

DM tools —— Automated discovery of implicit patterns and knowledge

* OLAP tools ⟶ simplifying & targeting data analysis

DM tools ⟶ Automate process

* DM analyze data at more detailed granularity than the summarized data in DW. It also analyze spatial, multimedia, textual and transactional data.

From On-line Analytical processing to On-line Analytical Mining

On-Line Analytical Mining (OLAM) — Integrates OLAP with data mining and mining knowledge in the multidimensional database.

* OLAM is important due to following reasons

High Quality of data in Data Warehouse:-

- DM tools need to work on Integrated, Consistent, Cleaned data which requires costly preprocessing steps.
* DW constructed through this process has high quality of data.

Available information processing Infrastructure Surrounding DW

- If processing & Analysis Infrastructures are constructed surrounding DW which include accessing Integration, Consolidation and Transformation of multiple heterogenous databases, ODBC/OLE DB Connections, Reporting and OLAP analysis tools
- So best to use available infrastructure than constructing from scratch.

OLAP-based exploratory data Analysis:-

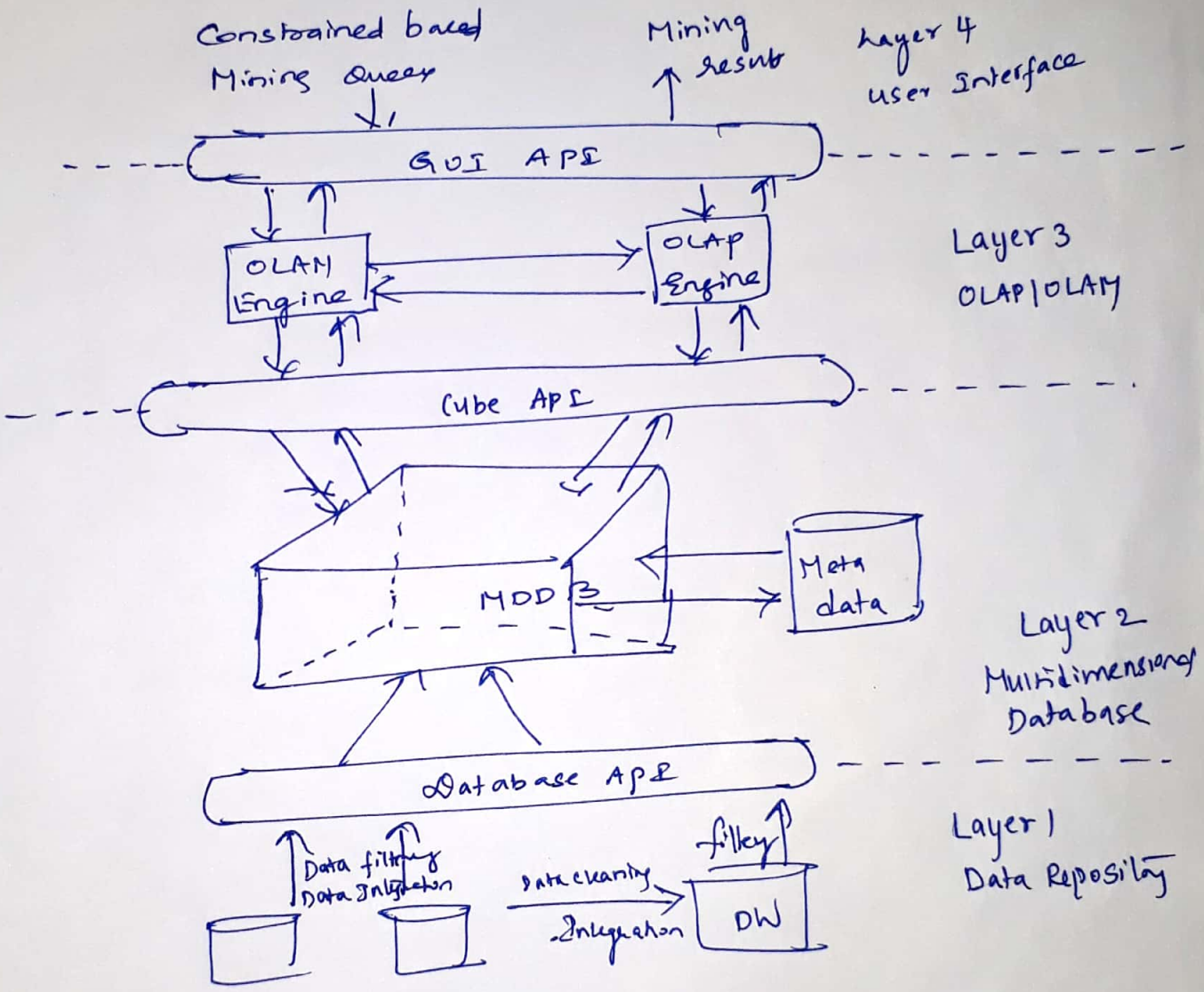
- Effective DM needs Exploratory data Analysis i.e traversing through database,
  Select portion of relevant data
  Analyze them at diff granularities
  present knowledge in diff forms

OLAM provides facility for DM on diff subset & at diff levels

On-line selection of DM functions

Users don't know what kind of knowledge to use. Integrating OLAP with DM functions OLAM provides Users with flexibility to select desired DM fn & swap DM ~~fn~~ tasks dynamically.

Architecture of OLAM:



* OLAM, OLAP Server — accepts users Queries via GUI API & work with data cube in data analysis via Cube API

* Metadata directory — used for helping in accessing data cube

Data cube constructed — by accessing integrating multiple database via MDDB API

* Filtering of database — through Database API that supports OLE DB & ODBC command