14/3/23

What is data Mining ?

→ Data mining is one of the most useful techniques that help entrepreneurs, researchers and individuals to extract valuable information from large set of data. Data mining is also called Knowledge Discovery in Database (KDD).

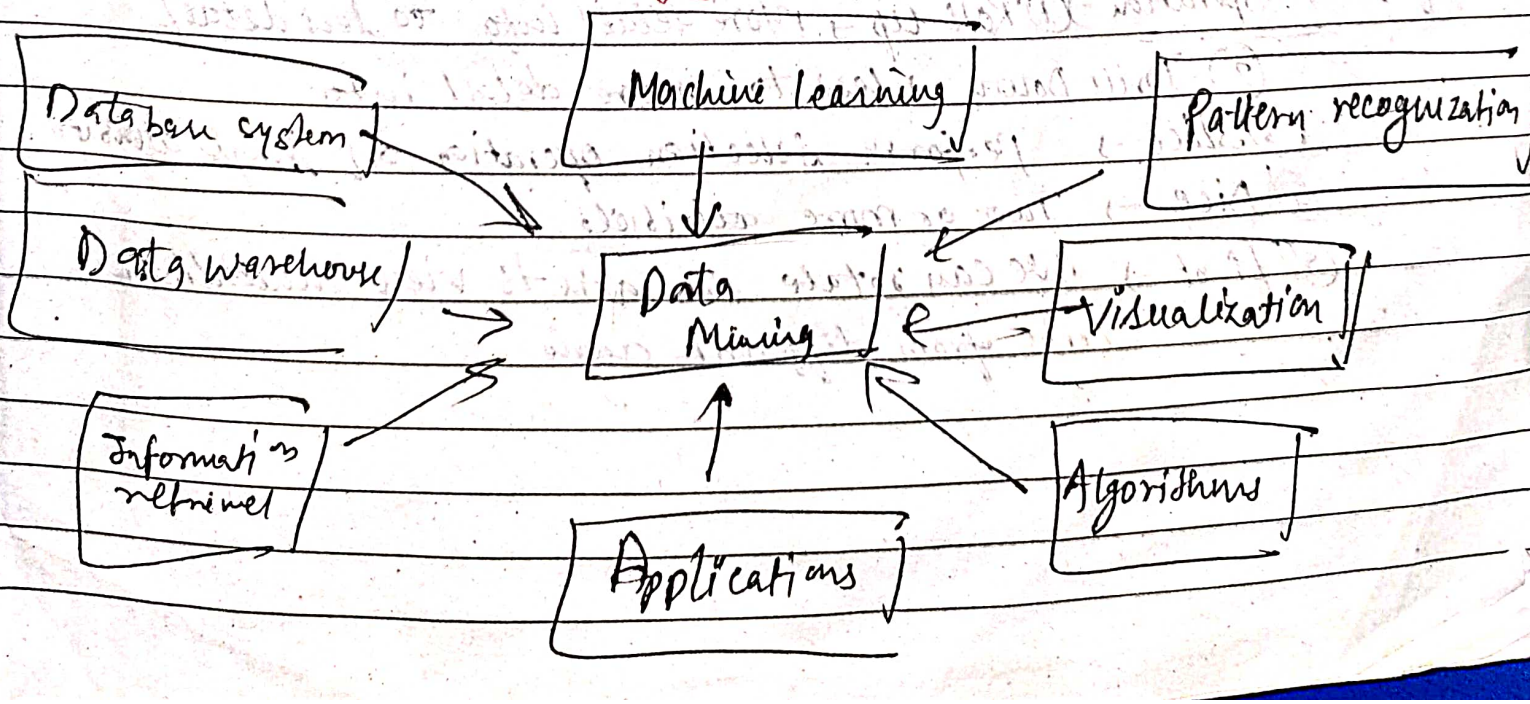
Or

→ Process of extracting useful information from large sets of data.

→ Also known as knowledge discovery / knowledge extraction / pattern analysis

1960's → Storing of data in form of hardcopy.
1970's → RDB on database. → Storing of data in the form of many table, in which row & column are there.

Main purpose of Data Mining :-



☆ Data Warehouse:-

A data Warehouse is a repository of heterogeneous database collected from multiple sources over a history of time, stored under a unified schema and used for analysis and decision making.

→ Used for OLAP → Online analytical processing

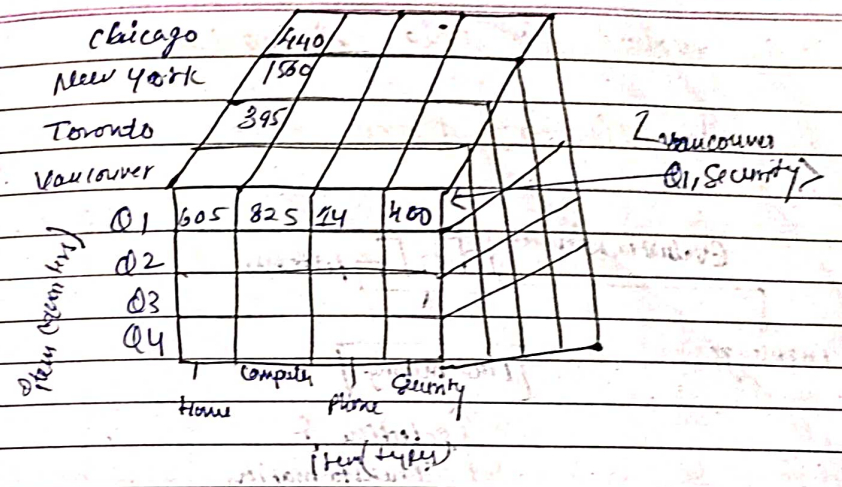
→ Data warehouse are constructed via a process of data cleaning, data integration, data integration, data transformation, data loading &.

→ Usually modeled by a multidimensional data structure called a data cube.

→ Each dimension corresponds to an attribute or a set of attributes

→ advantage → 1) Performs various types of OLAP operation
2 View multiple data base at once

Ex of OLAP operation → (1) Roll up → More detail/info to less detail
(2) Drill Down → less to more detail info
(3) Slice → Perform selection operation on single attribute
(4) Dice → Two or more attributes
(5) Pivot → We can rotate the axis to view multidimensional cube from different angle



☆ Data base →

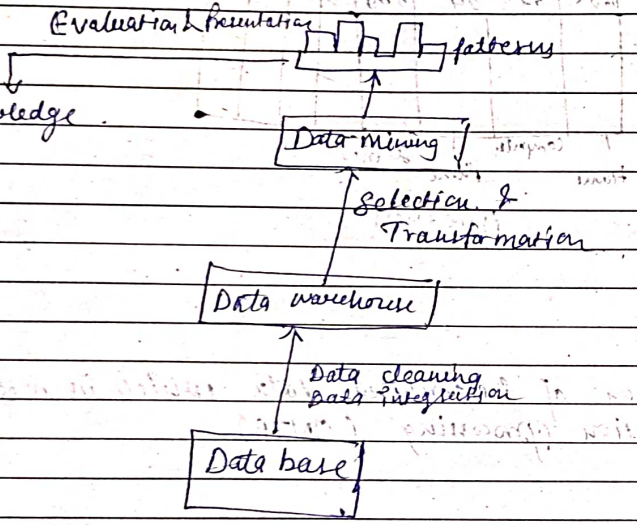
collection of interrelated data which is used for Online Transaction Processing (ONTP).

→ Similarities between them is both string data
→ Huge amount of data

→ Different is one is (ONTP) & another (OLAP).

☆ Steps involved in data mining :-

The Knowledge discovery process :-



(1) Data cleaning → To Remove noise and inconsistent data.
→ In real life we never get the cleared data have to check it first

(2) Data Integration → Where multiple data sources may be combined.

(3) Data selection → Where data relevant to the analysis task are retrieved from the database.
→ Selection of attributes only the relivent features out of all the data.

(4) Data transformation → Where data are transformed and consolidated into forms appropriate for mining by performing summary.

(5) Data Mining → An essential process where intelligent methods are applied to extract data patterns.

(6) Pattern evaluation :- To identify the truly interesting patterns representing knowledge based on interestingness measures.

(7) Knowledge presentation → Where visualization & knowledge representation techniques are used to present mined knowledge to user.

☆ Data Mining Engine :-

(1) Characterization → Summarization of the general features of a target class of data.

(2) Association Rule Mining →

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

(3) Correlation analysis :-

(4) Classification → The process of finding a model that distinguishes data classes for the purpose of being able to use the model to predict the class of object whose class level is unknown.

(4) Prediction → To predict missing or unavailable numeric data.

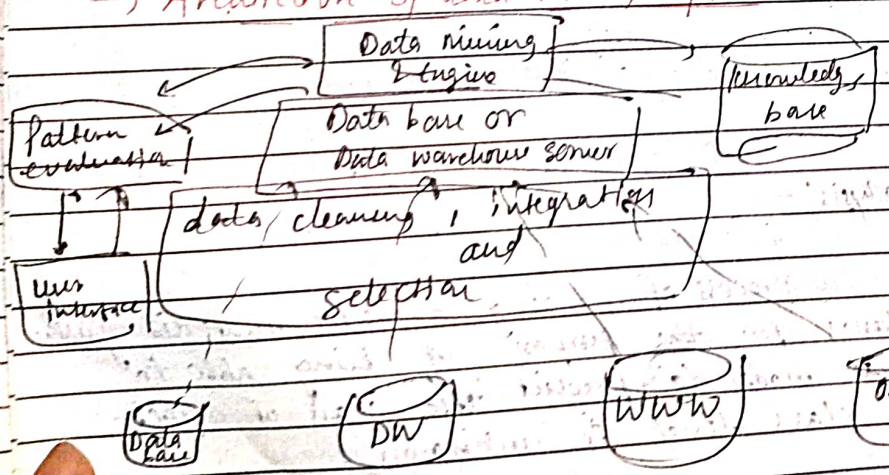
(5) Clustering → Unlike classification & regression, which analyze class labeled data sets, clustering analyzes data objects without consulting the class labels.

(6) Outlier Analysis → Analysis of noisy data.

(7) Evaluation analysis → Trying to develop the model which describes and models trends for objects whose behaviour changes over time.

15/03/23

→ Architecture of data Mining System:



① collect the data
② Then do processing

→ knowledge base → Basic information which will help or guide the search process

☆ Types of databases:-

(1) Relation Database:-
A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores large set of tuples (record or row).
Each tuple in a relational table represents an object identified by a unique key & described by a set of attribute values.

(2) Data warehouse → already written

(3) Transactional database → Each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user clicks on a web page.
A transaction typically includes a unique transaction identity number and a list of the items making up the transaction.
→ Nested relationship not allowed.

(4) Object relational database → All OOPs concepts

(5) Temporal database → (Relational database which includes time related attributes)
Ex → Booking (meeting Room, Time)

16/03/23

(8) Sequence database → Stores sequence of ordered events with or without concrete notion of time.

Ex → Formation of DNA
→ Customer shopping sequences

(9) Time-series database → Stores sequences of values or events obtained over repeated measurement of time.
(Hourly, daily, weekly, monthly)

Ex → Stock exchange data.

(10) Spatial database →
→ In this
- Geographic database
→ medical images
→ Satellite images.

→ VLSI → designers the ability to design utilize less space.

→ It saves a huge amount of space-related data, including maps, preprocessed remote sensing or medical imaging records, and VLSI chip design data.

Application of geographical database →

① Forestry and Ecology planning.

② Providing public service, information regarding the location of telephone and electric cable, water pipes & sewage system.

③ Vehical navigation and dispatching system.

(11) Text Database :- These contain word description for Not simple keywords but objects
→ (Contains long sentences & paragraphs)

Ex → Product specification
→ Error report
→ warning message.

→ Highly unstructured - Ex - Web page
→ Semi structured - Ex - Email
→ structured - Ex - library catalogue

(12) Multimedia database :- Stores images audio, videos.

(13) Heterogeneous database → Consist of a set of interconnected autonomous components of database.

(14) legacy database → group of heterogeneous databases that combines different kinds of datasystem.
→ Exchange of info root possible.

(15) Data Streams → Here data flow in and out of a window dynamically or continuously Ex - all time series, Stock exchange, weather

(16) WWW → Capturing user access patterns in www is known as weblog mining.

21/03/23

Data warehouse → It is a subject oriented, integrated time variant and non-volatile collection of data in support of management decision making process.

Difference between OLTP v/s OLAP :-

OLTP	OLAP
① These are the online operational processing database systems	Online information data warehouse system
② online transaction and query processing	② Data analysis and decision making.
③ customer oriented	③ Market oriented
④ Manages current data	④ Manages Historical data
⑤ ER datamodel	⑤ Star schema or snowflake schema or Galaxy schema
⑥ 100 TPS - some GB	⑥ 100 GB - TB
⑦ Users → DBA, clerks, database professors	⑦ Users → manager executive and data analysts.

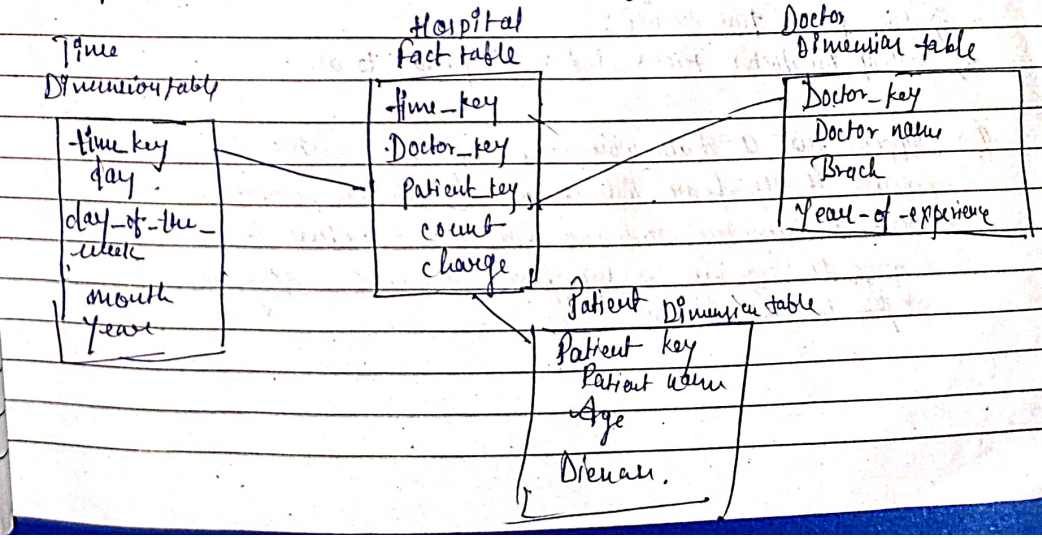
☆ Different Operations we can perform on multidimensional data cube :-

① Roll up → Performes aggregation on a data cube.
② Drill down → It navigates from less detailed to more detailed data.
③ Slice → Performs a selection on one dimension.
④ Dice → Selection on two or more dimension.
⑤ Pivot → Rotates the data axis in view.

☆ Star schema :-

The schema contains a central fact table with a bulk of data with no repetation at the centre.
→ Set of smaller dimension tables one for each dimension.

① → There is a datawarehouse consist of 3 dimension, time, doctor, patient & two measures → count, charge. draw star schema for this

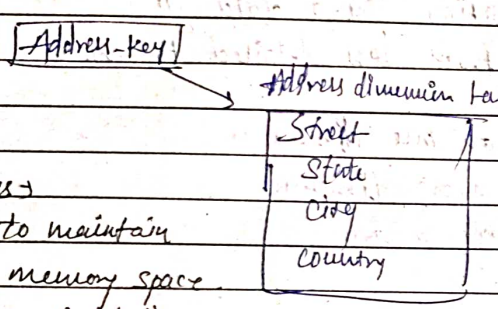


22/08/23

Date
Page No.

Snowflake :-

The snowflake schema is a variant of the star schema model, where some dimension tables are further normalized.



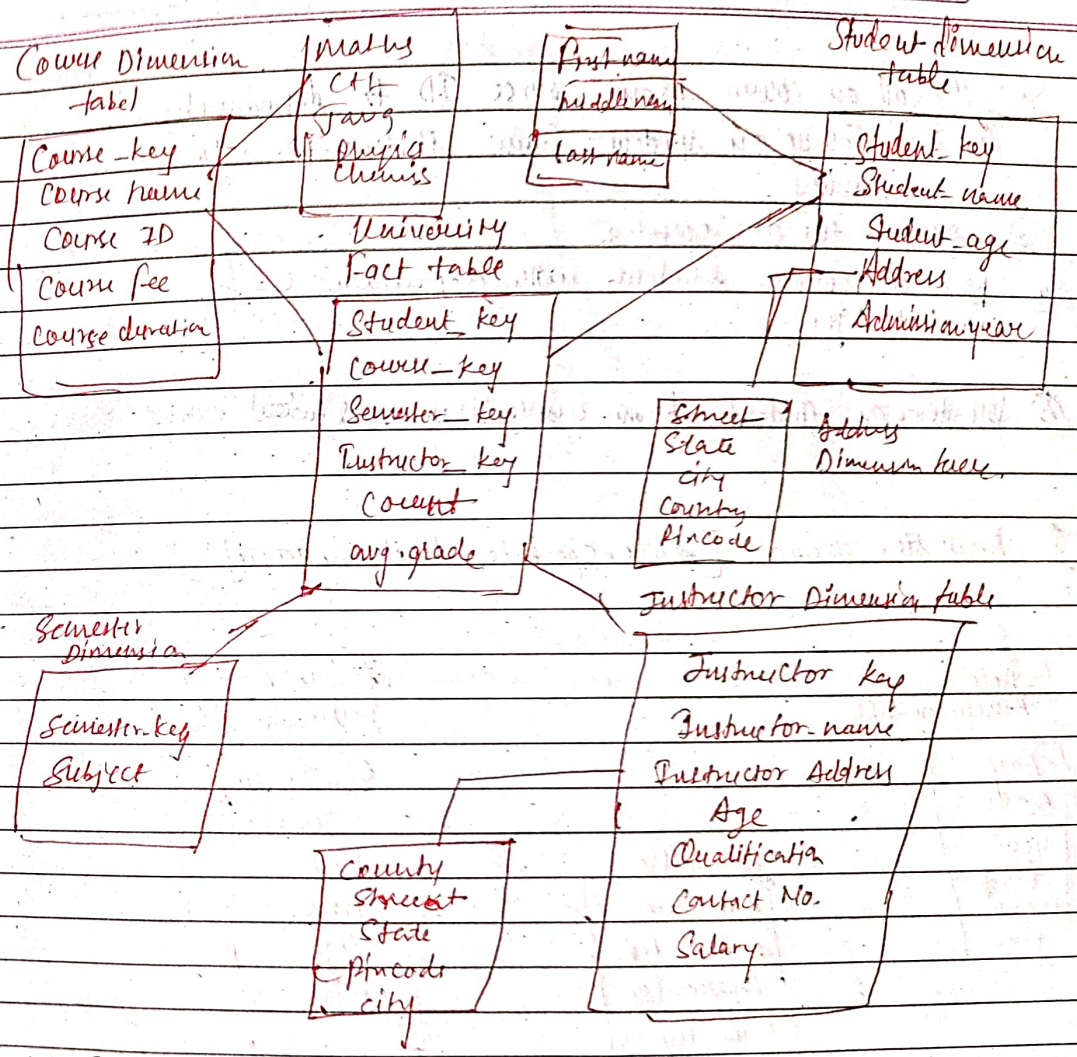
Advantages →
1. Easy to maintain
2. Saves memory space
3. Reduces Repetation

Q→ Starting with the base cuboid [day, doctor, patient] what specific OLAP operation should be performed in order to list the total fee collected by each doctor in 2010?

Sol⇒ 1) Roll up in time from day to year.
2) Slice for the time 2004.
3) Roll up on doctor from individual point to all.

Q→ Suppose that a data warehouse for Big university consists of the four dimension [Student, Course, Semester & Instructor and two measure count & avg_grade. Draw a snowflake schema diagram for the datawarehouse.

Date
Page No.

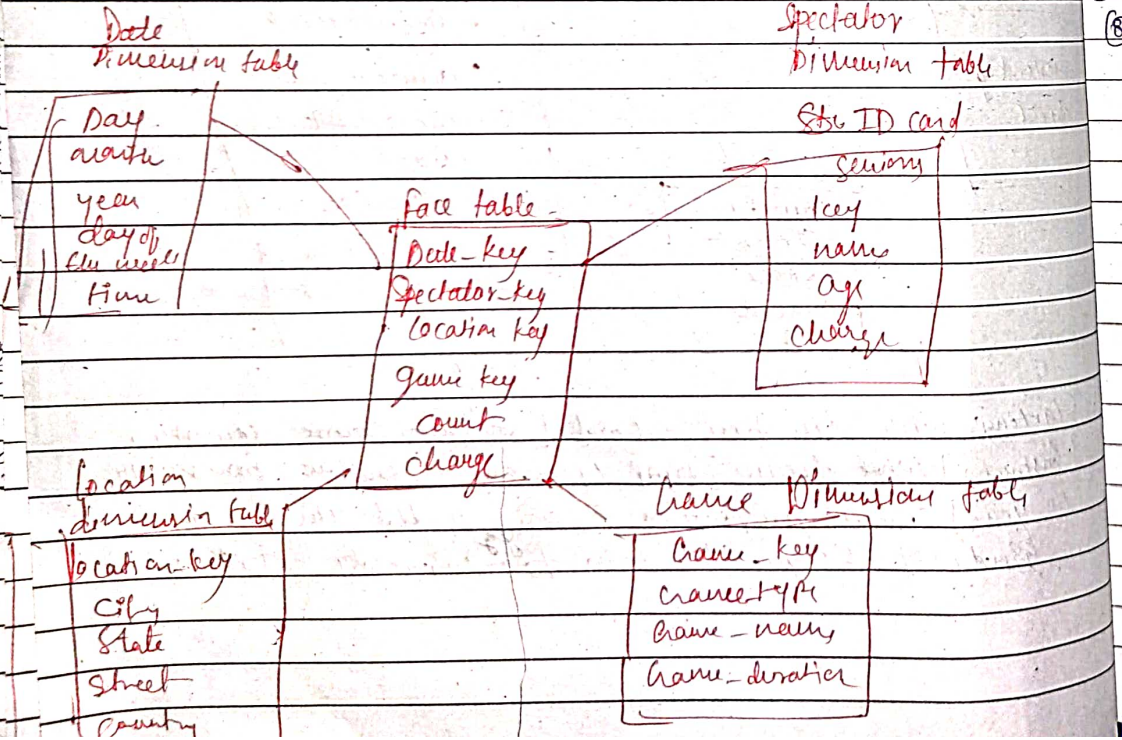


Q→ Starting with the base cuboid [student, course, semester, instructor] what specific OLAP operations eg roll-up from semester to year] should you perform in order to list the average grade of CS course for each ~~Big~~ University Student.

Sol:- (1) Roll on course from course ID to department
(2) Then Roll up on student from student-ID to university
(3). perform the dice operation
(4) Dice on course, student with department CS & University "gju"

(5) Drill-down on student from university to student name

Q. Draw star schema. [date, spectator, location, game]



Difference b/w Database and Datawarehouse

Database System	Data
(1) It supports operational process.	It supports analysis and performance reporting.
(2) Capture and maintain the data.	Explore the data
(3) Current data.	Multiple years of history.
(4) 100 MB to GB	100 GB to TB
(5) Application oriented	Subject oriented.
(6) Flat ~~rel~~ relational	Multidimensional
(7) ER based	snowflake/star
(8) Data is updated when transaction occurs.	Data is updates of scheduled processes.