23/01/23

Vidyalekhan

DATE

PAGE

Course Outcome

UNIT - I

→ Syllabus

→ Introduction :-

→ Data Warehouse And Mining

- A data warehouse is a subject Oriented, Integrated, Non-Volatile and time-variant collection of data in support of managements decision making process.

→ Textbooks

1. Jiawei Han and Micheline Kamber
"Data mining concepts and techniques"
Second edition Elsevier

Containing Def"

- The four keywords, Subject oriented, Integrated, Time-variant and non-volatile distinguish Data warehouse from other data repository system such as Rational DB system, transactional processing system and file system

1. Subject Oriented

A data warehouse is organised around major subjects such as customers, suppliers, product and sales rather than concentrating on the day to day operations and transaction processing of an organization

2. Integrated

A data warehouse is usually constructed by integrating multiple heterogeneous sources such as rational database (DBS) flat file and online transactional record

3. Time - Variant

Data are stored to provide information from a historical perspective (the past 5-10 years)

4 Non - Volatile.

A datawarehouse is always a physically seperate store of data transform from the application data found in the operational environment. Due to this seperation the data warehouse does not require transaction processing recovery and concurrancy control mechanism

Data warehouse

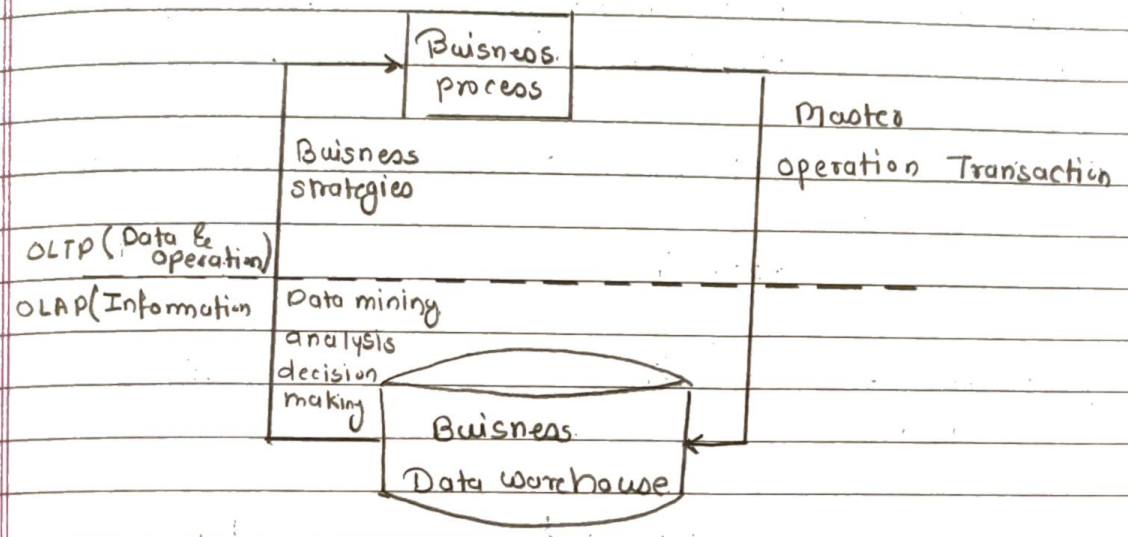
It is usually requires only two operation in data accessing i.e. initial loading of data and access of data.

* OLTP - Online Transaction processing
* OLAP - Online Analitic processing

25/01/23

Vidyalekhan
DATE
PAGE

OLAP Vs OLTP



	Features	OLTP	OLAP
1)	characteristic	operational processing	Information processing
2)	Orientation	Transcation	Analysis
3)	User	clerk, DBA and any database professional	Knowledgeble worker. eg manager, executive & analyst
4)	Function	day to day transaction/operation	Long term information requirements, decision suppor
5)	PB design	ER diagrams / App oriented	Star / Snow flake, Subject oriented

6)	Data	Current data : guaranteed, upto date data	Historical data
7)	Access	Read/write access	mostly read.
8)	Focus	data in	information.
9)	Number of Records Access	Tens	millions
10)	No. of users	Thousands	Hundreds
11)	DB size	100 MB to GB	100 GB to TB
12)	priority	High performance, High availability	High flexibility, end user autonomy
13)	Metric	Transaction throughput	query throughput, response time

Multidimentional Data Model

- Data Warehouse and OLAP tools are based on multidimensional data model.
- The multi-dimensional data model is a method which is used for ordering data in the database.
- Multidimensional data models views data in the form of data cube.
- Data Cube - A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions & facts.
- Dimensions are the prospectives or the entities with respect to which and organisation want to keep records

eg Take the example of data of a factory which sells products per quarter in location banglore. The data is represented in the table given below.

	Location = Banglore			
Time Quarter	Type of item			
	Jam	Bread	Sugar	Salt
Q_1	350	389	35	50
Q_2	260	528	50	90
Q_3	483	256	20	60
Q_4	436	396	15	40

2D - Factory Data.

In the above given presentation, the factories sells for banglore are, for the time dimension, which is organised into quarters and the dimension of items, which is sorted according to the kind of item which is sold.

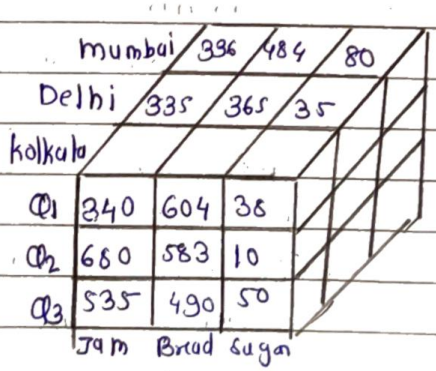
The facts here are represented in Rs. (1000)

Now, if we desire to view the data of the sells in a 3 dimensional table, then it is represented in the diagram given below.

Here the data of the sells is represented as a 2 dimensional table. Let us consider the data according to item, time and location.

Time	Location = "kolkata"			Location = "Delhi"			Location = "mumbai"		
	item			item			item		
	Jam	Bread	Sugar	Jam	Bread	Sugar	Jam	Bread	Sugar
Q1	340	604	38	335	365	35	336	484	80
Q2	680	583	10	684	490	48	595	504	39
Q3	535	490	50	389	385	15	366	385	20

3D - data representation in 2D.



The data can be represented in the form 3D conceptually, which is shown in the image below.

5/02/23

Data Warehouse Schema :

1. Star schema.
2. Snowflake Schema.
3. Fact Constellation schema.

1. Star Schema.

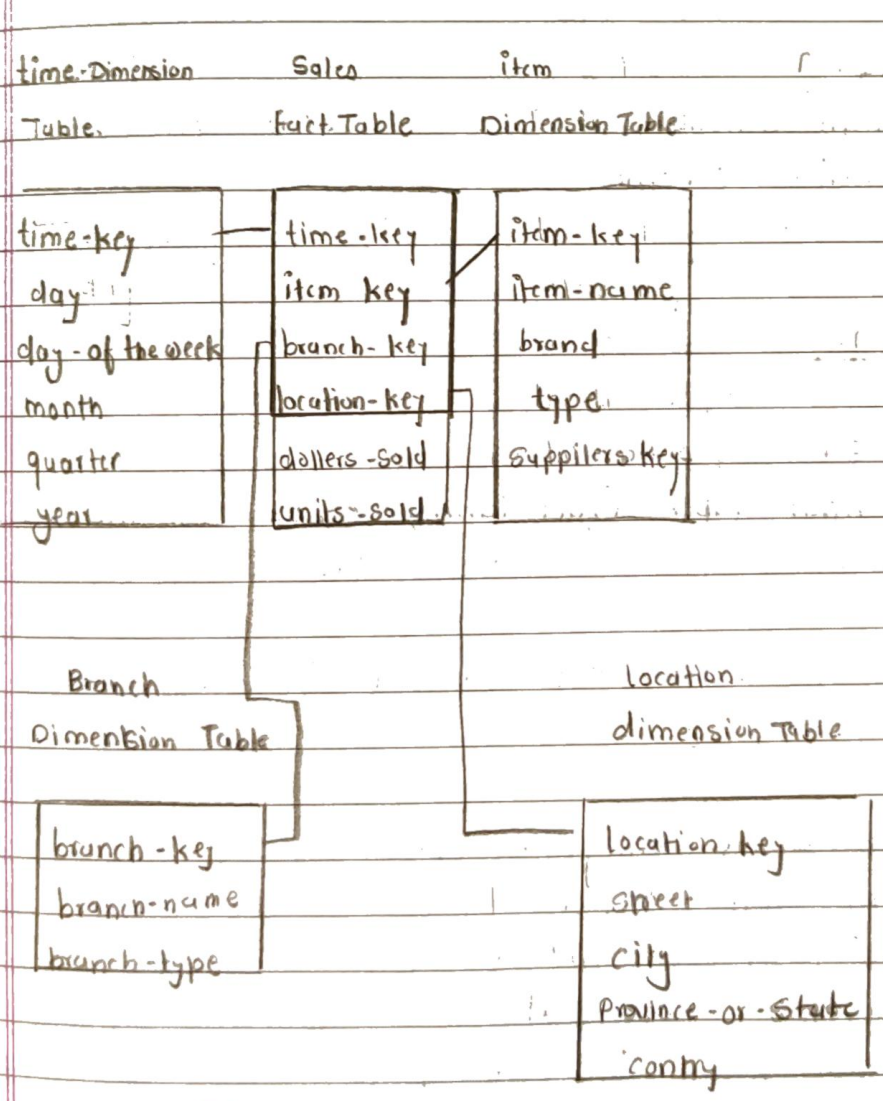


Fig : star schemas

Datawarehouse Schema

Schema is a logical description of the entire database. It includes the name and discription of records of all record types including all associated data items and aggregates like entity - relationship data model which is commonly used in relational databases in the design of where the databases schema consists of a set of entities relationship betn them.

The most popular data model for data warehouse is a multi-dimensional mode. Such a model can be exist in the form of the following types of schemas:

1. Star Schema
2. ~~Sta~~ Snowflake schema.
3. Fact Constellation Schema.

1. Star Schema

In star schema data warehouse contains

1) A large central table (fact table) containing the bulk of data, foregein keys with no redundancies.

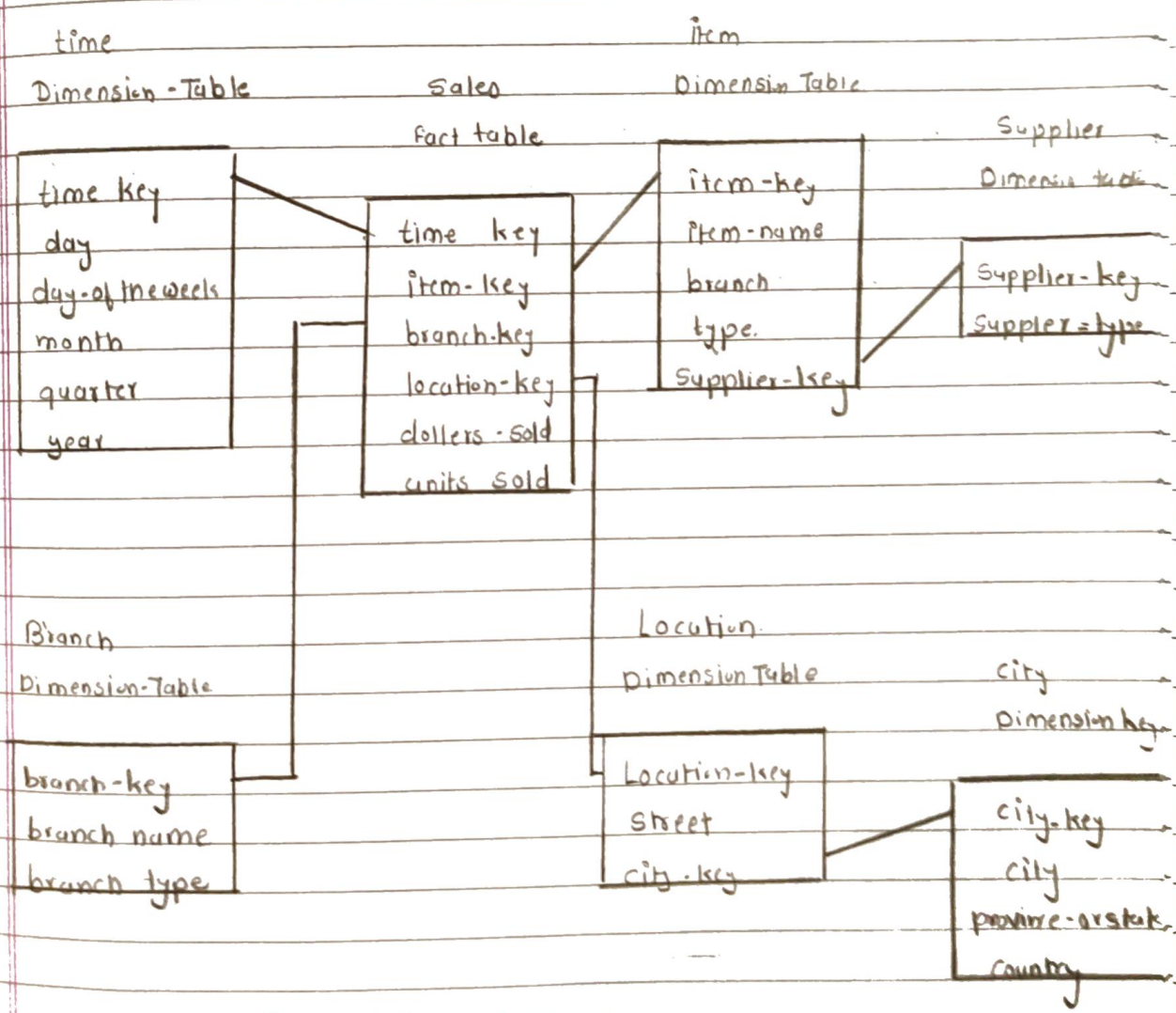
2) A set of smaller attendent tables (dimension table), one for each dimension.

Consider the above diagram which shows the sales data of a company with respect to the 4 dimensions namely time, item, branch and location.
There is a fact table at the center. It contains the keys (foregein keys) to each of four dimensions. The

fact table also contains the attributes namely dollers sold and units-sold.

In dimension table there can be data redundancy of data.

2. Snow flakes Schema.

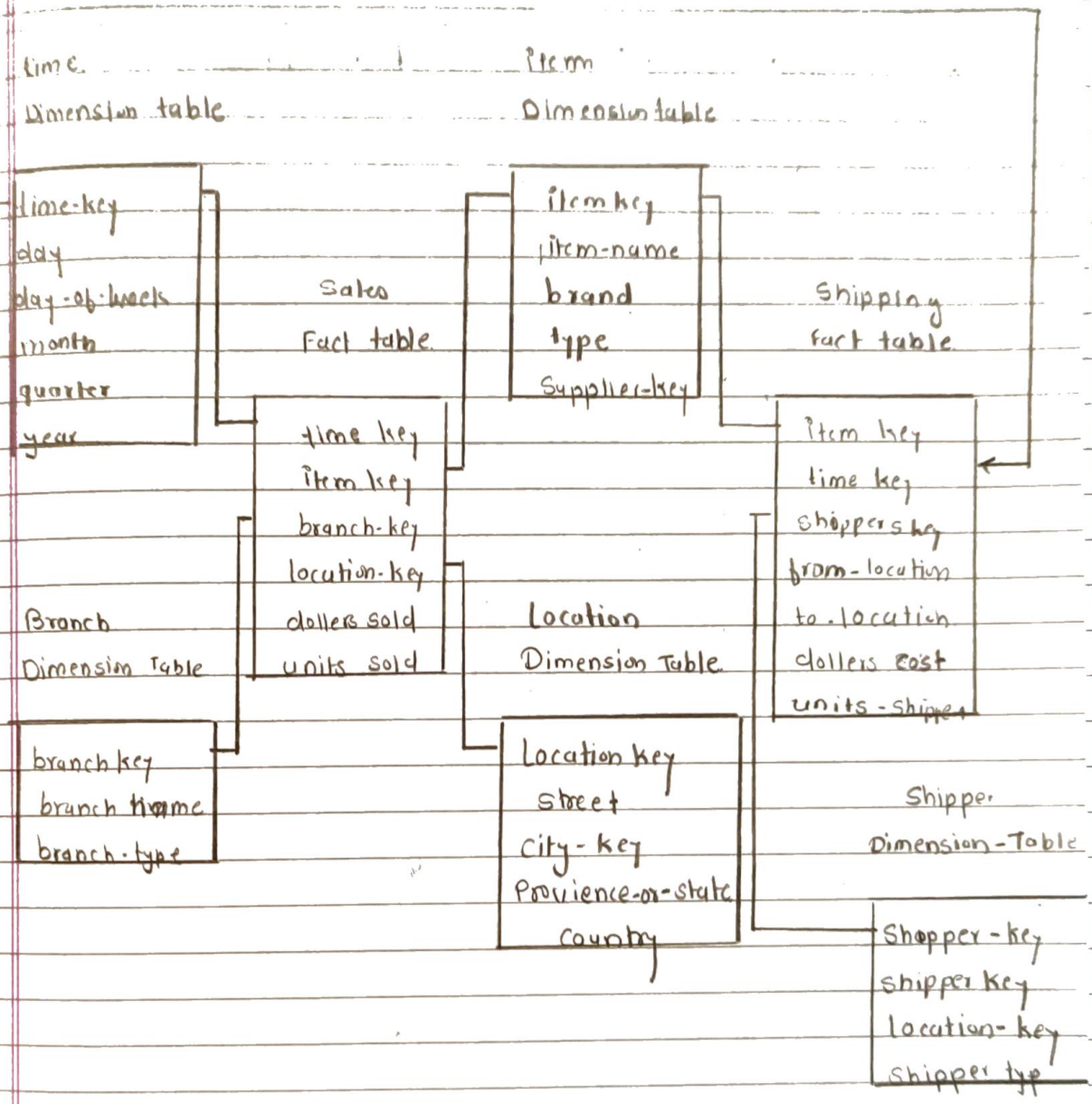


Snowflake schema

In snowflakes schema, The main difference betn the star and snowflakes schema is in the defination of dimension table, the single dimension table for item in the star schema is normalized in the snowflakes schema, resulting in new item and supplier tables. The normalisation splits up the data into additional table.

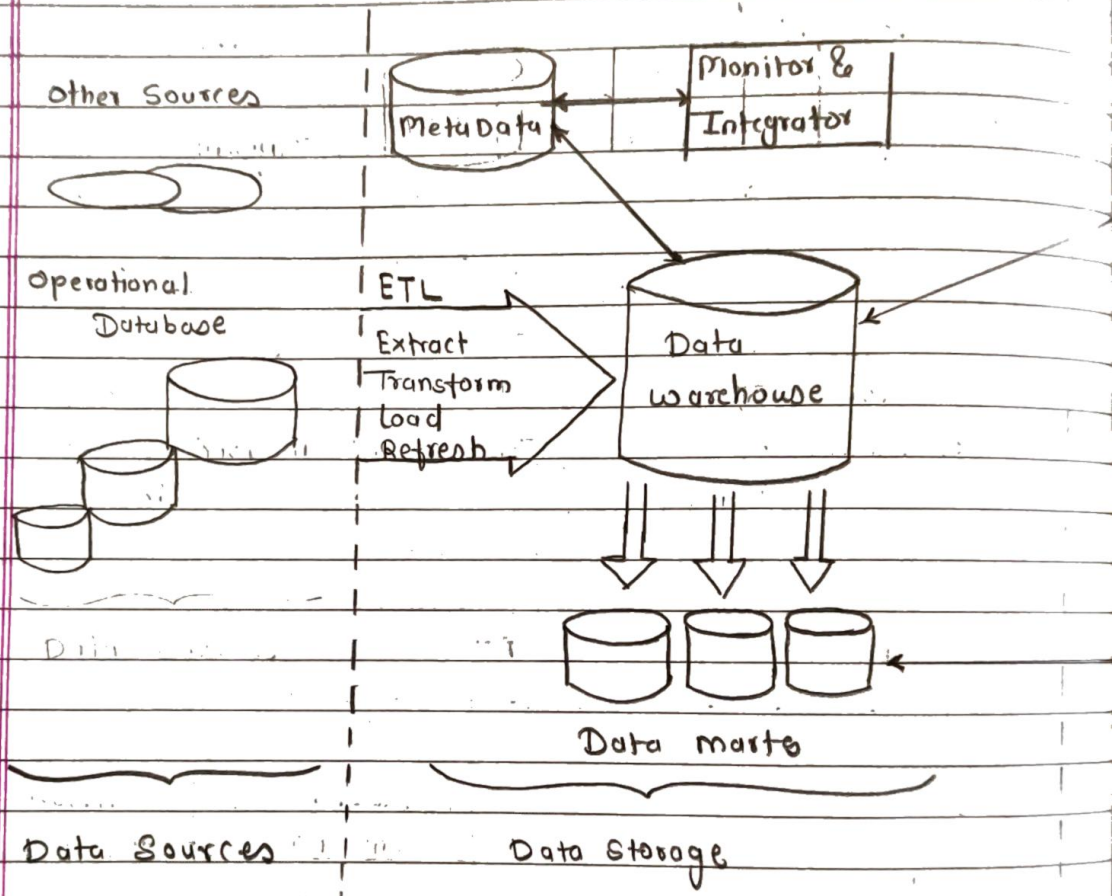
3. Fact - Constellation Table

In fact - constellation table there can be multiple fact tables This ~~kun be~~ kind of schema can be viewed of a collection of stars and hence it is called as "Galaxy Schema"



7/02/23

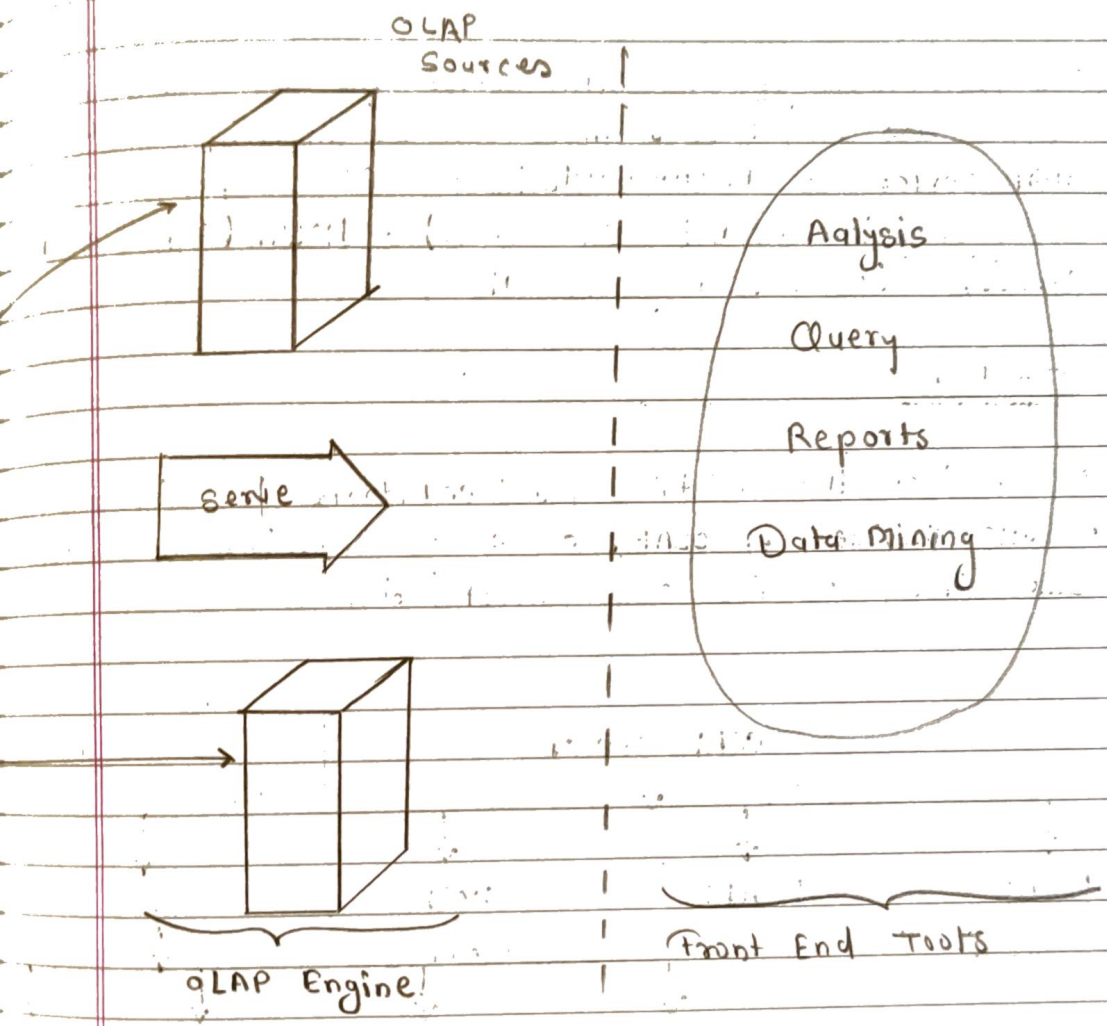
3-tier Data Warehouse Architecture



Data warehouse usually have 3 level (tier) Architecture that includes

1) Bottom Tier

That consists of data warehouse server which is almost an RDBMS. It may include several specialised data marts and a metadata repository.
Data from operational databases and external sources (such as user profile data provided by external consultant)



are extracted using application program interfaces called a gateway.

A gateway is provided by the underline DBMS and allows costomer programs to generate SQL code to be executed at a server.

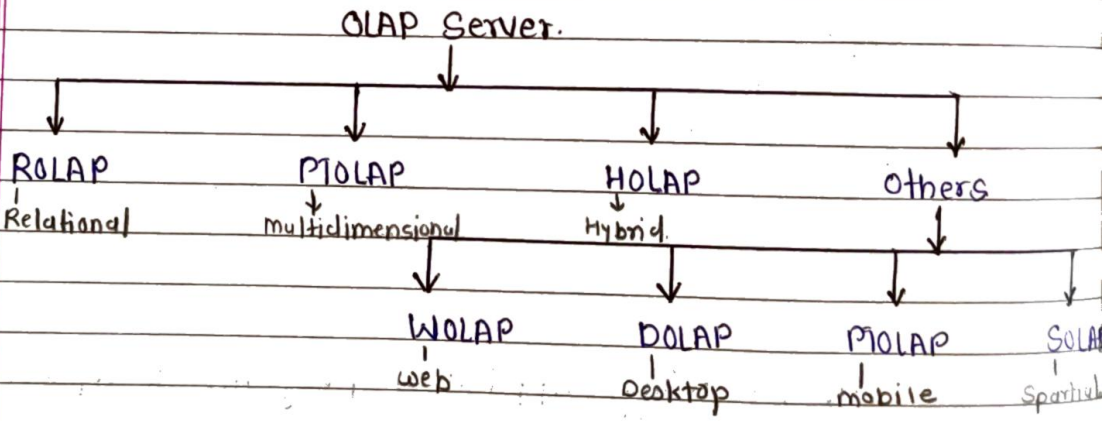
Examples of gateways contains ~~OD~~ ODBC (open database connectivity) and JDBC (Java database connectivity).

2) Middle tier.

It is consists of an OLAP server for fast querring of the datawarehouse. Their can be different types of OLAP server i.e. ROLAP (Relational OLAP server), MOLAP (Multidimensional OLAP server), HOLAP (Hybrid OLAP server). (you have to write discription)

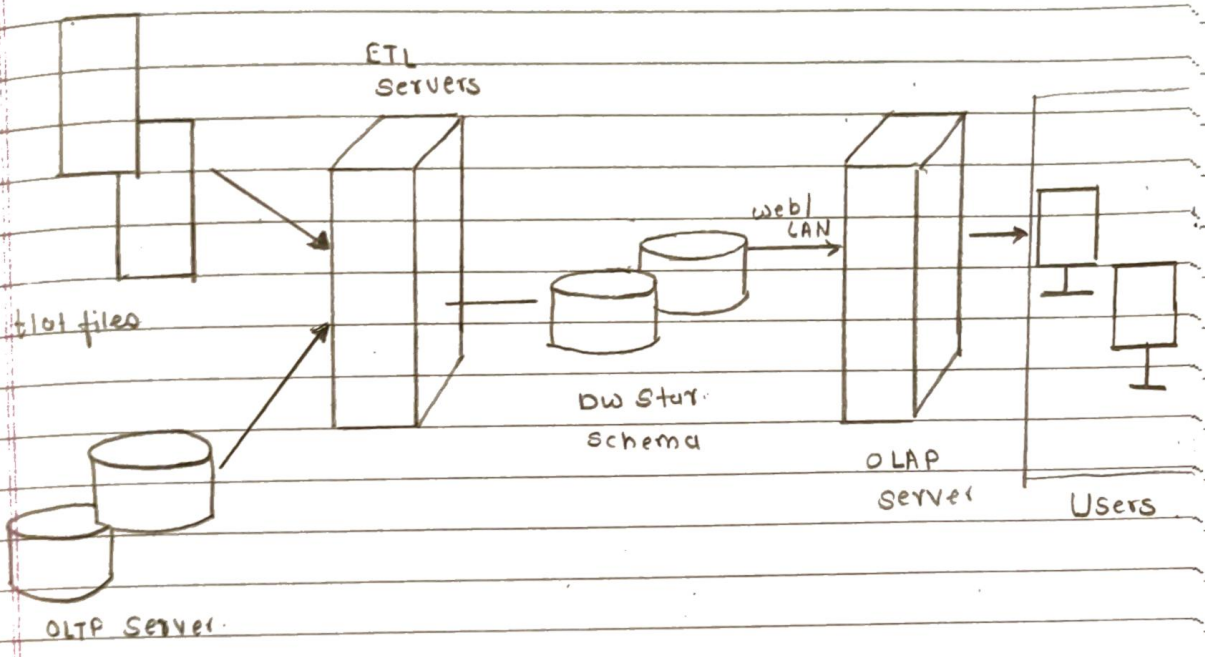
3) Top tier.

A Top tier that contains front end tools for displaying results provided by OLAP, as well as additional tools for data mining of the OLAP generated data.

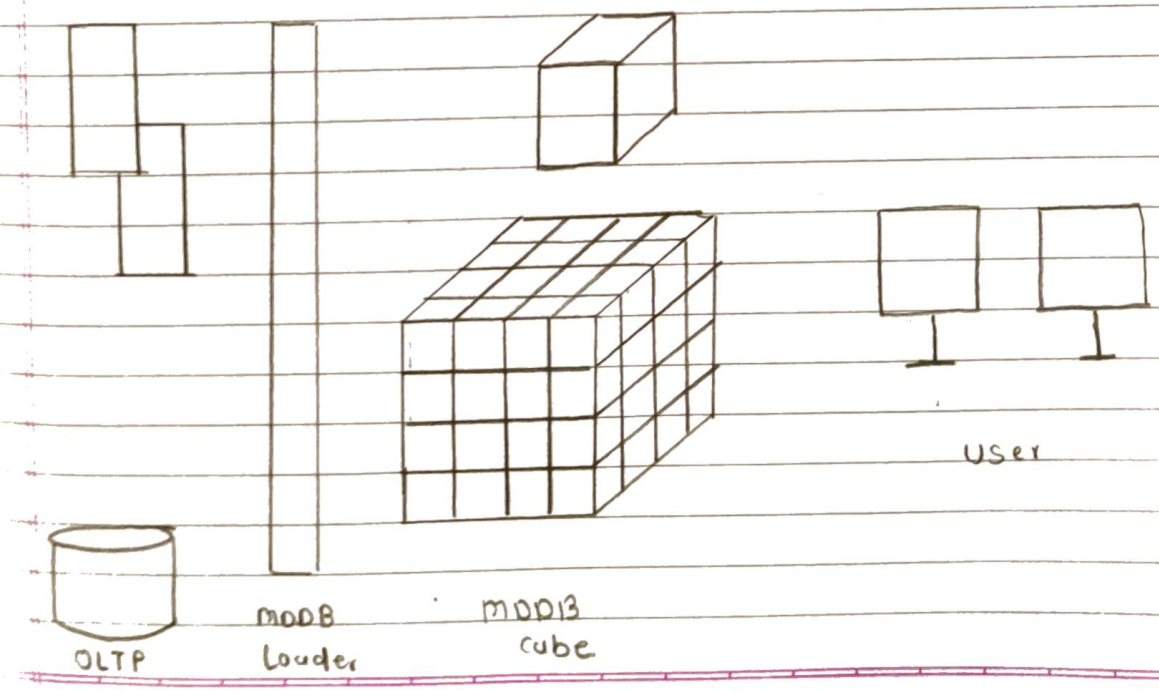


Types of OLAP server

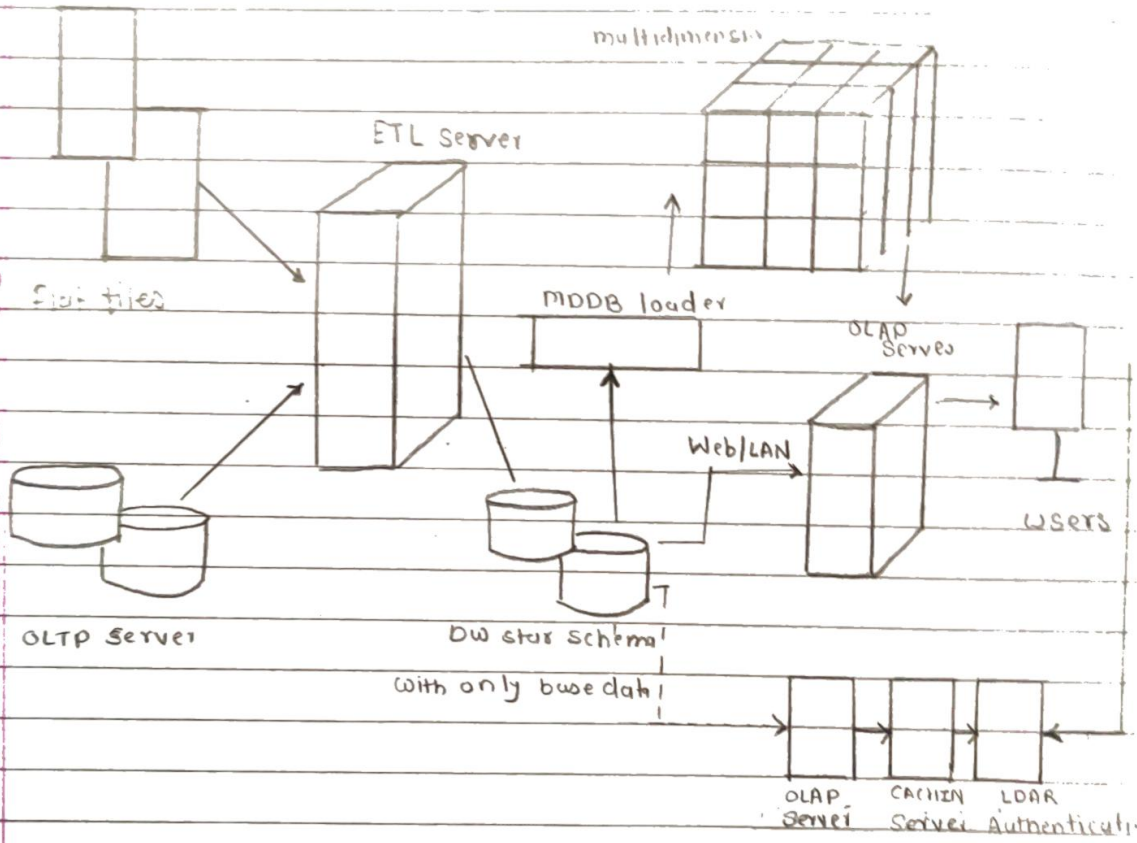
1) ROLAP Server



2) MOLAP Server



3) HOLAP Server



16/2/23

Vidyalekhan

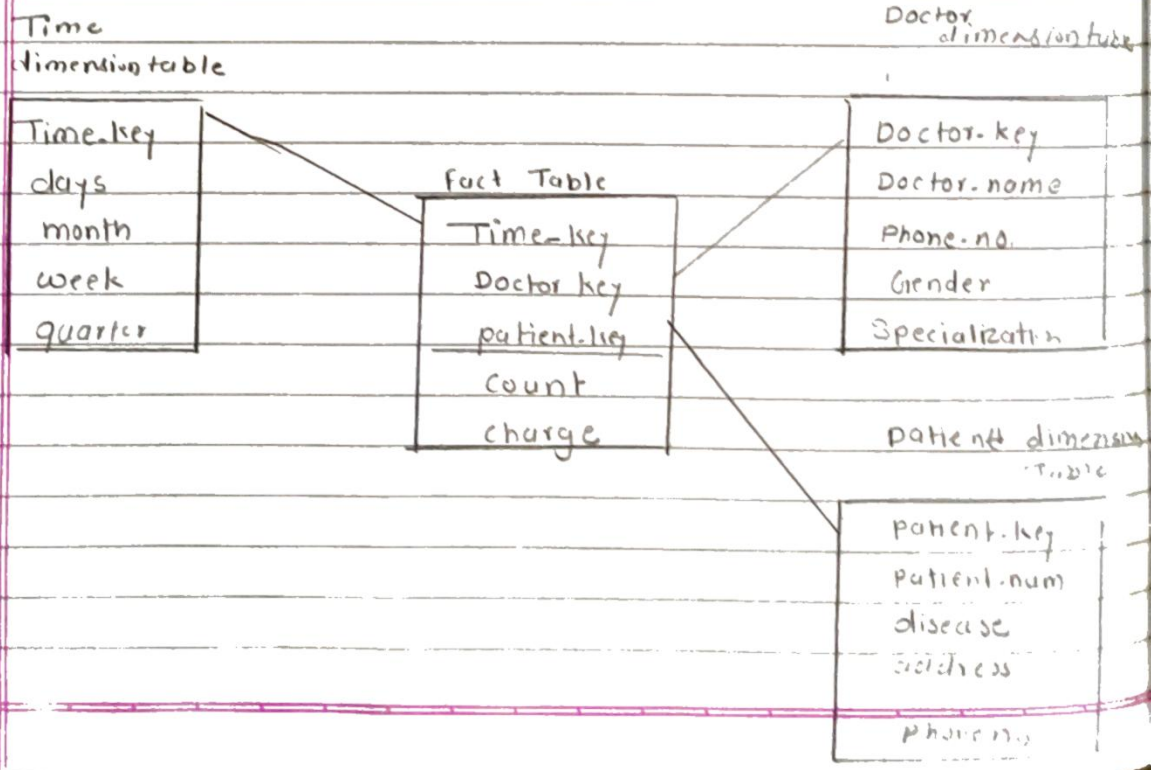
DATE
PAGE

Design & Construction of datawarehouse

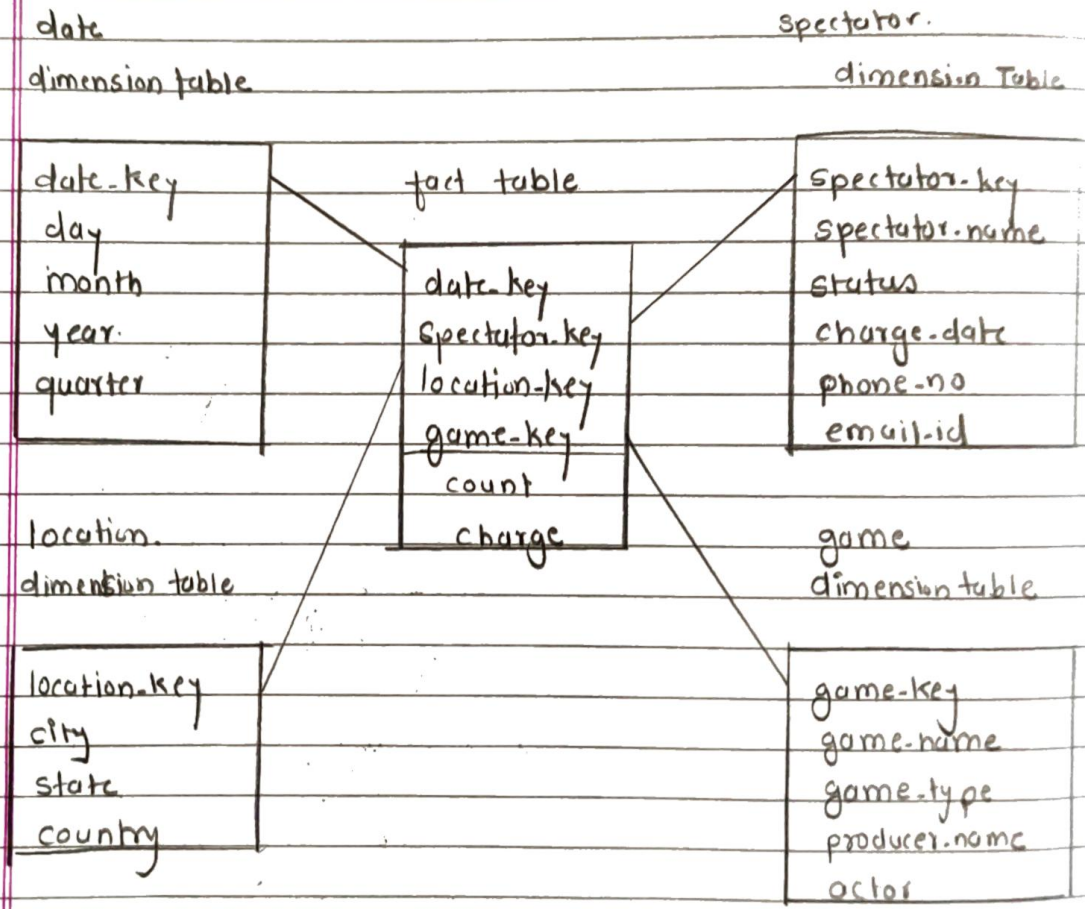
In general warehouse design consists of following steps:

1) Choose a buisness process model.
2) choose the grain of the buisness process (fact table)
3) choose the dimension
4) choose the measures

Suppose that datawarehouse consist of 3 dimensions Time, doctor, patient and 2 majors i.e. count & charge where charge is the fee that a doctor charges to the patient for a visit. Draw star schema diagram for the above datawarehouse



Suppose the datawarehouse consist of the 4 dimensions date, spectator, location and game and 2 major count and charge where charge is the fair that a spectator ~~phase~~ pays for watching a game on a given date. Spectator may be student, adults or seniors with each spectator having a its own charge rent. Draw the star schema



KDP

Unit 2 — Knowledg discovery from data

Vidyalekhan

Syllabus

Fundament data mining

1) Data mining functionalities, classification of data mining system, data mining task premitives, megor issues and challange in data mining, data preprocessing / need for data preprocessing, data cleaning, integration, transformation data reduction, data mining application areas.

* Data Mining

Data mining refers to extracting or mining knowledge from large amount of data.

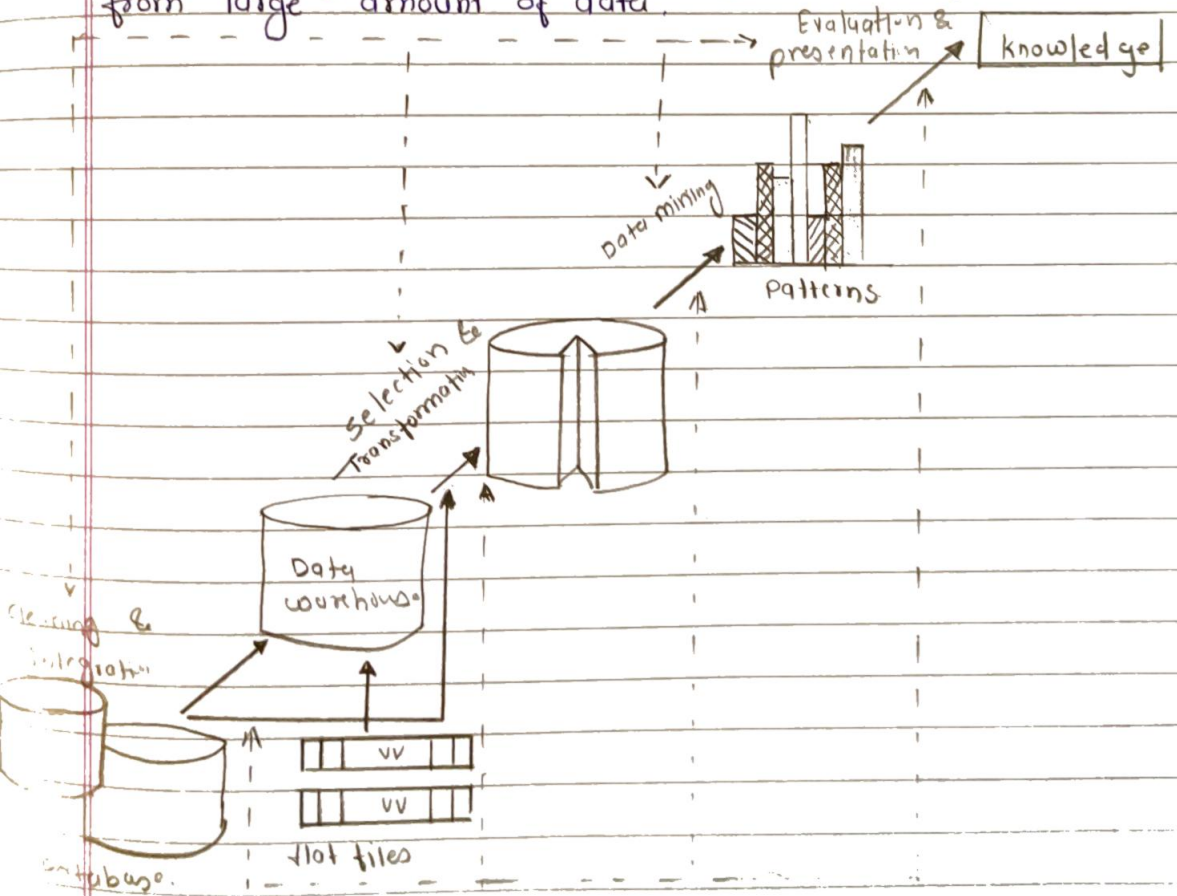


fig. Data mining as a step in the process of knowledge discovery

20/02/23

Vidyalekhan
DATE
PAGE

★ Architecture of Data mining : Major Components in Data mining

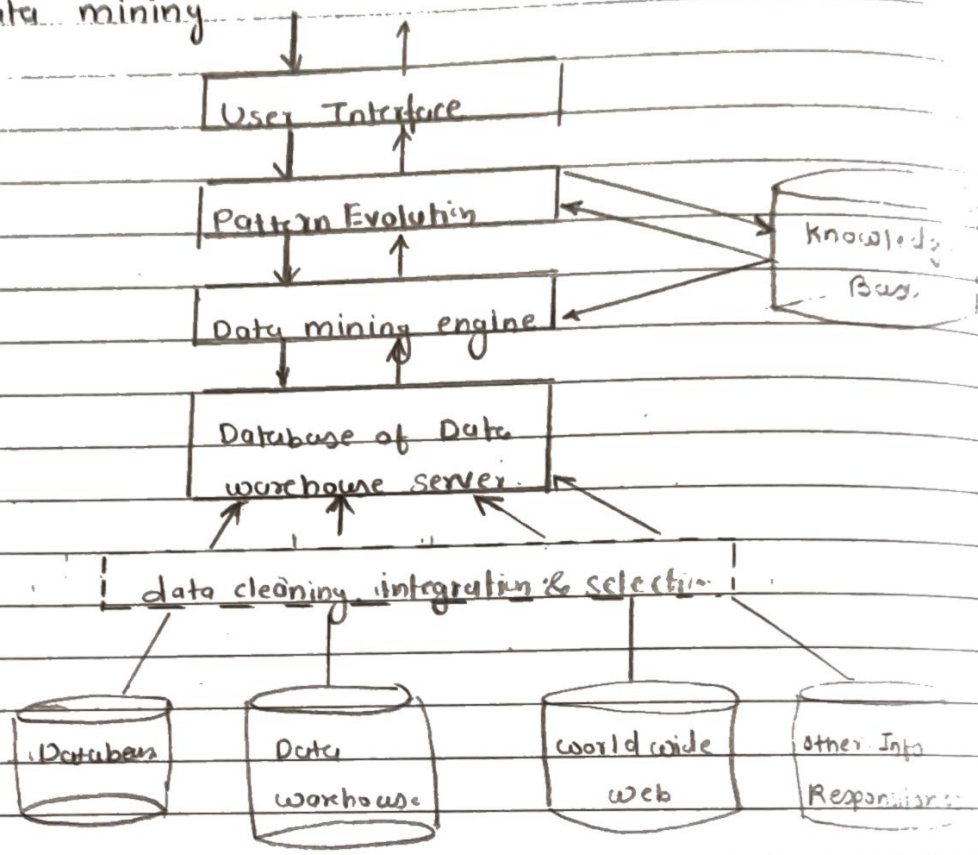


Fig Architecture of Typical Data mining System

★ Data Mining Functionalities

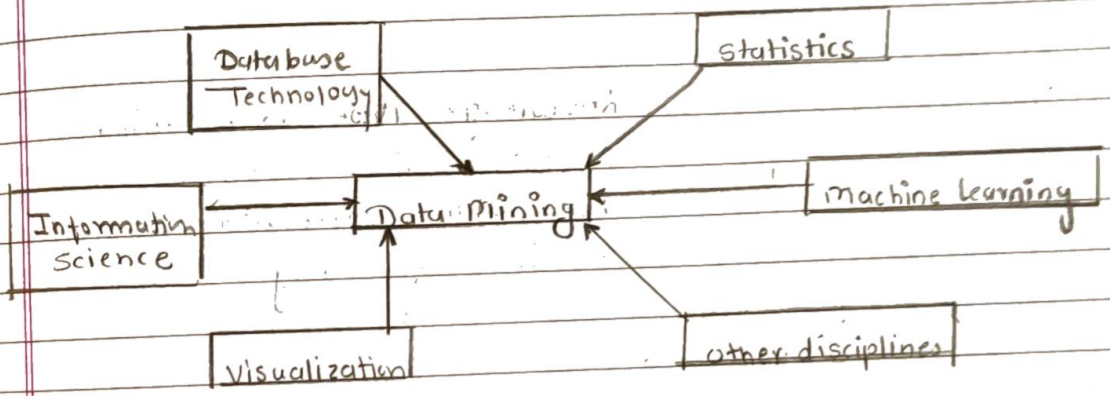
1) Concept / class Discription.
2) Mining Frequent Patterns, association, correlation
3) classification and prediction
4) Cluster Analysis

5) Outlier Analysis

6) Evolution Analysis

18/02/23

* Classification of Data Mining system



Data mining as confluence of multiple disciplines

1) classification according to kinds of database mined

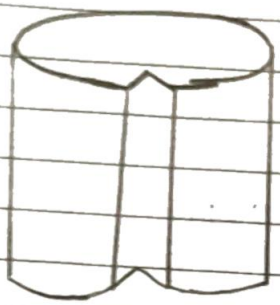
2) classification according to the kinds of knowledge mines

3) classification according to the kinds of technique utilized

4) classification according to the applications adopted.

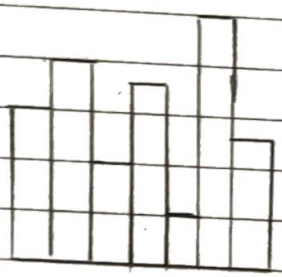
★ Data Mining Task Premitives

1)



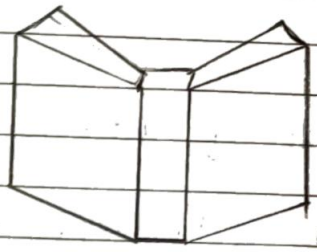
1) Task relevant Data
2) Database or Data warehouse name
3) Database Table or Data cubes
4) Condition for Data Selection
5) Relevant attributes or dimensions.
6) Data Grouping criteria

2)



knowledge type to be mind
characterization, Discrimination,
Association, correlation, classification,
prediction, clustering

3)

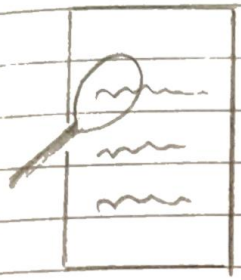


Background knowledge
concept Hierarchies user
Beleifs about relationships
in the data

4)

Pattern interestingness measures
simplicity
certainty (e.g. confidence)

5)



Visualization of discoured patterns
Rules, Tables, reports, charts, graphs, decision, trees & cubes
Drill down and rollup.