Data Mining — An Introduction.

- we live in a world where vast amounts of data are collected daily. Analysing such data is an important need.

What is Data Mining?

- Data Mining refers to extracting or mining knowledge from large amounts of data. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

KDD — Knowledge Discovery from data.

generally the terms KDD and data mining are used interchangeably.

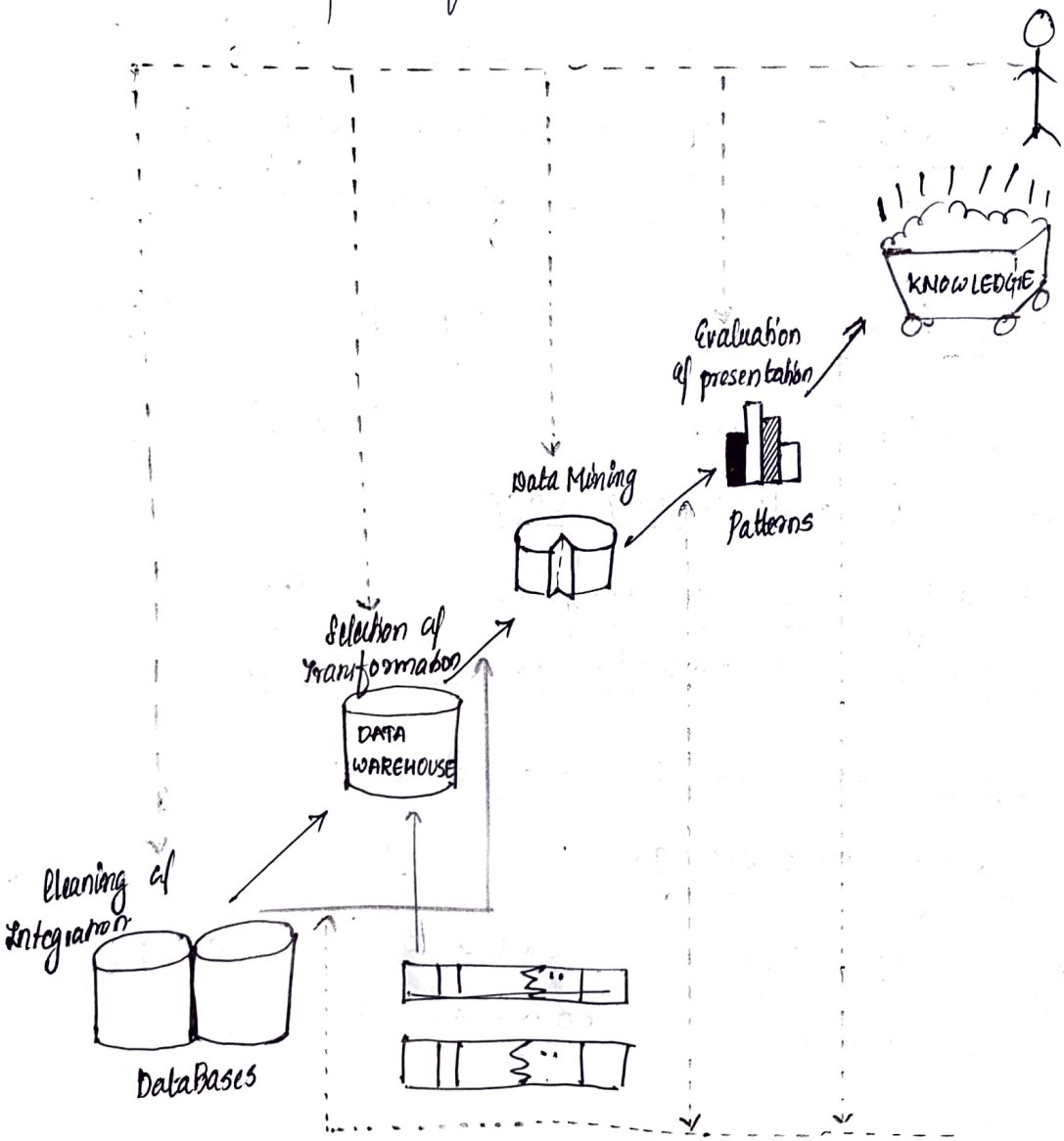
Actually data mining is one step in the process of knowledge discovery.

KDD — refers to the broad process of finding knowledge in data.

goal of KDD is to extract knowledge from data in the context of large DBs.

It does this by data mining meth
(algorithms) to extract (identify) wh
is deemed knowledge according to t
specifications of measures and thresh
using a DB along with any require
preprocessing, subsampling and
transformations of that DB.

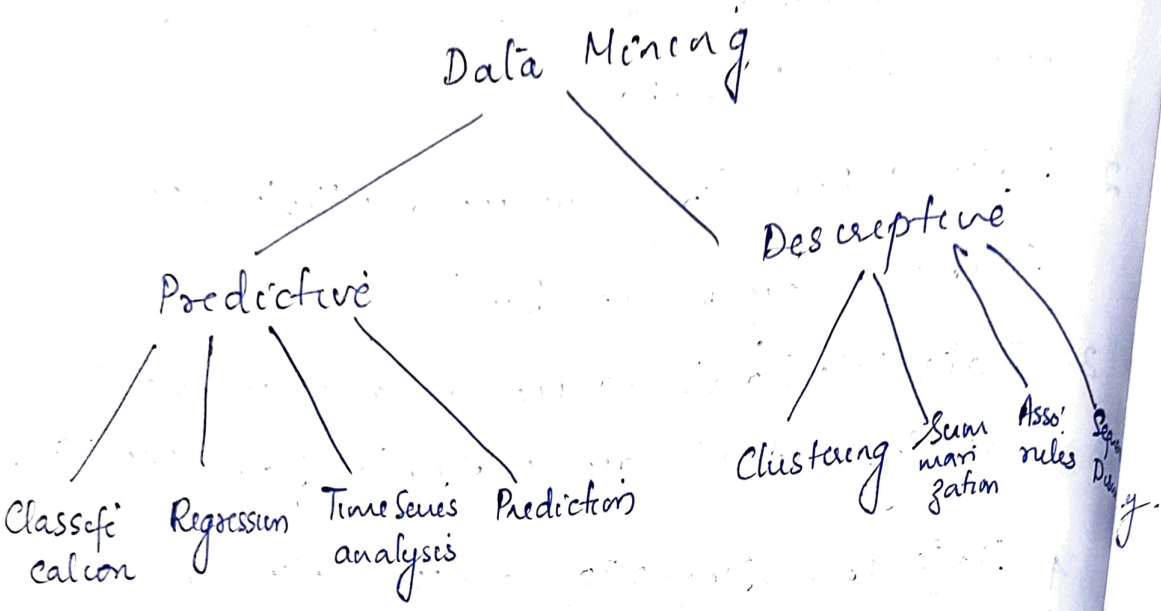
Steps of the KDD Process



Data Mining Stages.

1. Data Cleaning: To remove noise and inconsistent data.
2. Data Integration: where multiple data sources may be combined
3. Data Selection: where data relevant to analysis task are retrieved from the DB.
4. Data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations
5. Data Mining: an essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation: To identify the truly interesting patterns representing knowledge based on some interestingness measures
7. Knowledge Presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data Mining Models and Tasks



Predictive Model.

- makes a prediction about values of data using known results from different data.
- may be based on the use of other historical data.
- Predictive model tasks include classification, regression, time series analysis and prediction.

Descriptive Model.

- identifies patterns or relationships in data.
- Descriptive model serves as a way to explore the properties of the data

examined, not to predict new properties.

- Clustering, Summarization, association rules and sequence discovery are usually viewed as descriptive in nature.

Predictive Models.

1) classification:

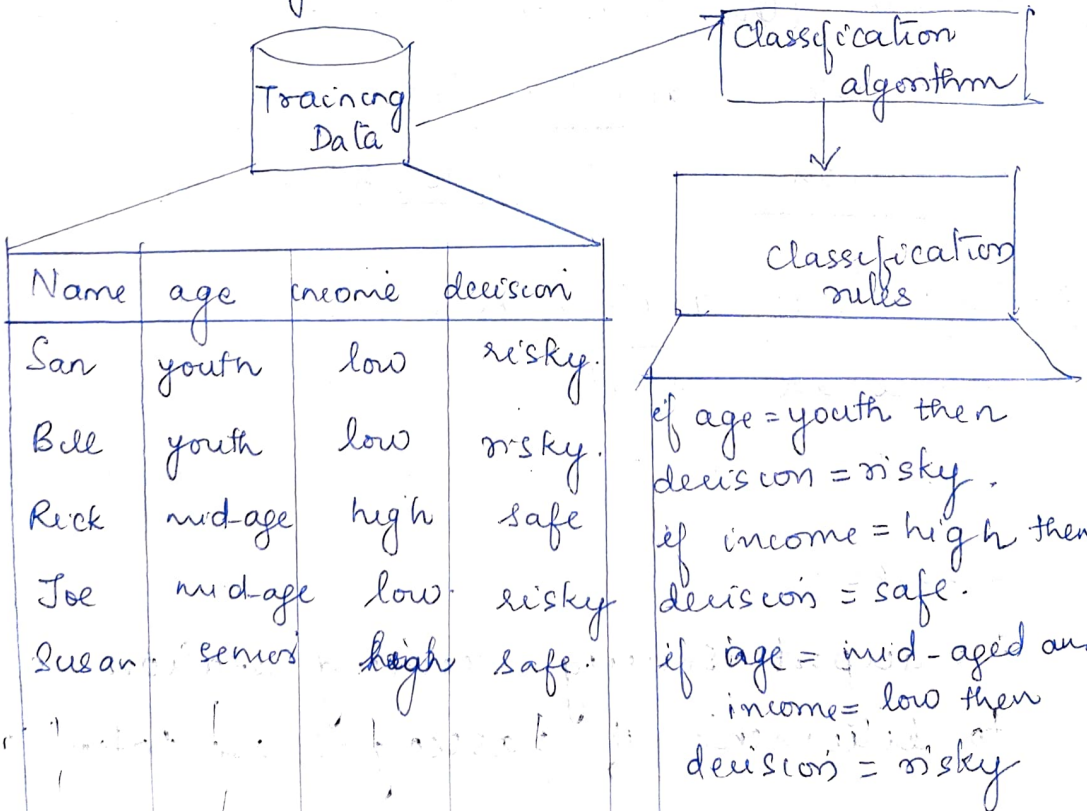
- maps data into predefined groups or classes.
- often referred to as supervised learning because the classes are determined before examining the data.
- classification algorithms require that the classes be defined based on data attribute values.
- They often describe these classes by looking at the characteristics of data already known to belong to the classes.

eg: Pattern recognition is a type of classification where an input pattern is classified based on its similarity to these predefined classes.

An airport security screening station is used to determine if passengers are potential

terrorists or criminals. (To do this, the face of each passenger is scanned and its basic pattern (distance b/w eyes, size and shape of mouth, shape of head etc) is identified. This pattern is compared with entries in a DB to see if it matches any patterns that are associated with known offenders.

Q. A bank loan officer wants to analyze the data inorder to know which customer (loan applicant) are risky or which are safe.



Name	age	income	decision
San	youth	low	risky.
Bill	youth	low	risky.
Rick	mid-age	high	safe
Joe	mid-age	low	risky
Susan	senior	high	safe

Test data: John, middled, low income.

decision → risky.

② Regression.

- Regression is used to map a data item to real-valued prediction variable.

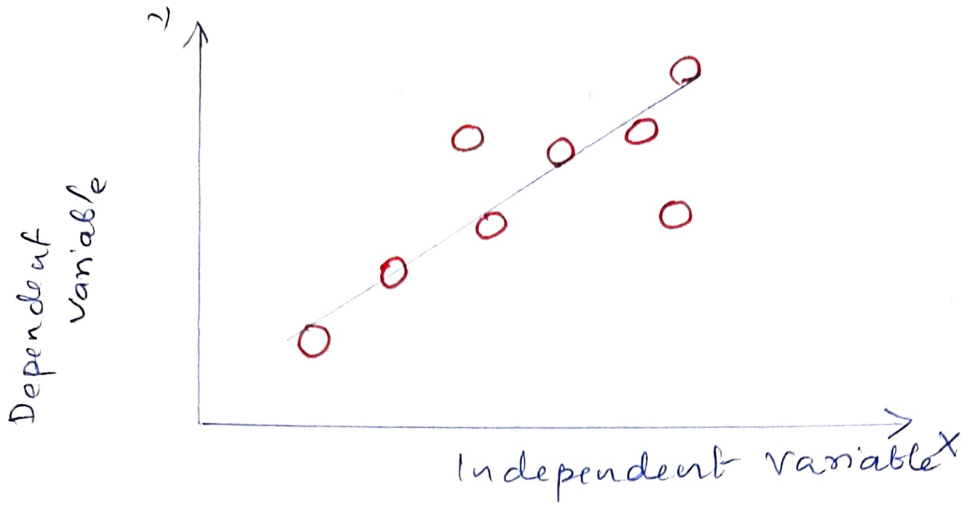
- Regression involves the learning of the function that does this mapping.

- Regression assumes that the target data fit into some known type of function (linear, logistic etc). and then determines the best function of this type that models the given data.

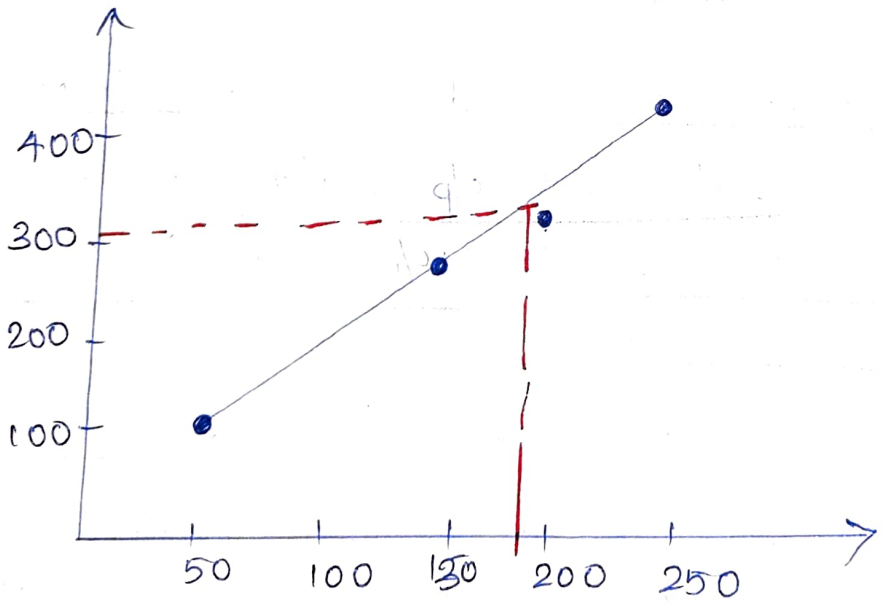
- Linear regression finds the relationship b/w the input and o/p data by plotting a line that fits the o/p data & maps it onto the o/p.

- This line represents the mathematical relationship b/w the independent i/p variables and is called the line of Best Fit.

- It covers as many o/p variables as possible while leaving out the outliers or noise



Advertisement	Sales
50	100
150	275
200	300
250	400
175	??



Simple Linear Regression

- Simplest linear regression that involves only one predictor.
- This model assumes a linear relationship b/w the dependent variable and the predictor model.

$$\text{dependent variable } y = \underbrace{a}_{y \text{ intercept}} + \overbrace{b}^{\text{slope of line}} \underbrace{x}_{\text{independent variable.}}$$

Error in Simple regression.

The regression equation model in ML uses the above slope intercept format.

X and Y values are provided to the machine and it identifies the values of 'a' and 'b' by relating the values of X and Y.

- Finding the exact match of values for a and b is not always possible.
- There will be some error value (ε) associated with it. This is called marginal or residual error

$$Y = (a + bX) + \epsilon$$

If we know the values of 'a' and 'b' then it is easy to predict the value of Y for any given X

But "How to Calculate the values of 'a' and 'b' for a given set of X and Y values?

Ordinary Least Squares (OLS) is a technique used to estimate a line that will minimise the error (ϵ).

This means summing the errors of each prediction or Sum of the squares of Errors SSE i.e

$$\sum_i \epsilon_i^2$$

It is observed that SSE is least when 'b' takes the value

$$b = \frac{\sum_i (X_i - \bar{x})(Y_i - \bar{y})}{\sum_i (X_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)}$$

The corresponding value of 'a' calculated using the above value of 'b' is

$$a = \bar{Y} - b\bar{X}.$$

Q. A college professor believes that if the grade for internal exam is high in a class, the grade for external examination will also be high. A random sample of 15 students in the class was selected and the data is given below.

Int Exam	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
Ext Exam	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
15	49	-4.93	-7.8	38.454	24.3049
23	63	3.07	6.2	19.034	9.4249
18	58	-1.93	1.2	-2.316	3.7249
23	60	3.07	3.2	9.824	9.4249
24	58	4.07	1.2	4.884	16.5649
22	61	2.07	4.2	8.694	4.2849
22	60	2.07	3.2	6.624	4.2849
19	63	-0.93	6.2	-5.766	0.8649
19	60	-0.93	3.2	-2.976	0.8649
16	52	-3.93	-4.8	18.864	15.4449
24	62	4.07	5.2	21.164	16.5649
11	30	-8.93	-26.8	239.324	79.7449
24	59	4.07	2.2	8.954	16.5649
16	49	-3.93	-7.8	30.654	15.4449
23	68	3.07	11.2	34.384	9.4249
$\Sigma X = 299$	$\Sigma Y = 852$			$\Sigma(x_i-\bar{x})(y_i-\bar{y})$	$\Sigma(x_i-\bar{x})^2$
$\bar{X} = \frac{299}{15}$	$\bar{Y} = \frac{852}{15}$			$= 429.28$	$= 226.9335$
$= 19.93$	$= 56.8$				

$$\therefore b = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{429.28}{226.93}$$

$$= 1.89$$

$$\therefore a = \bar{y} - b\bar{x}$$

$$= 56.8 - (1.89 \times 19.93)$$

$$= 19.05 \quad (19.18)$$

∴ The Simple Linear Regression Model is

$$M_{Ext} = \overset{19.18}{19.04} + 1.89 \times M_{int}$$

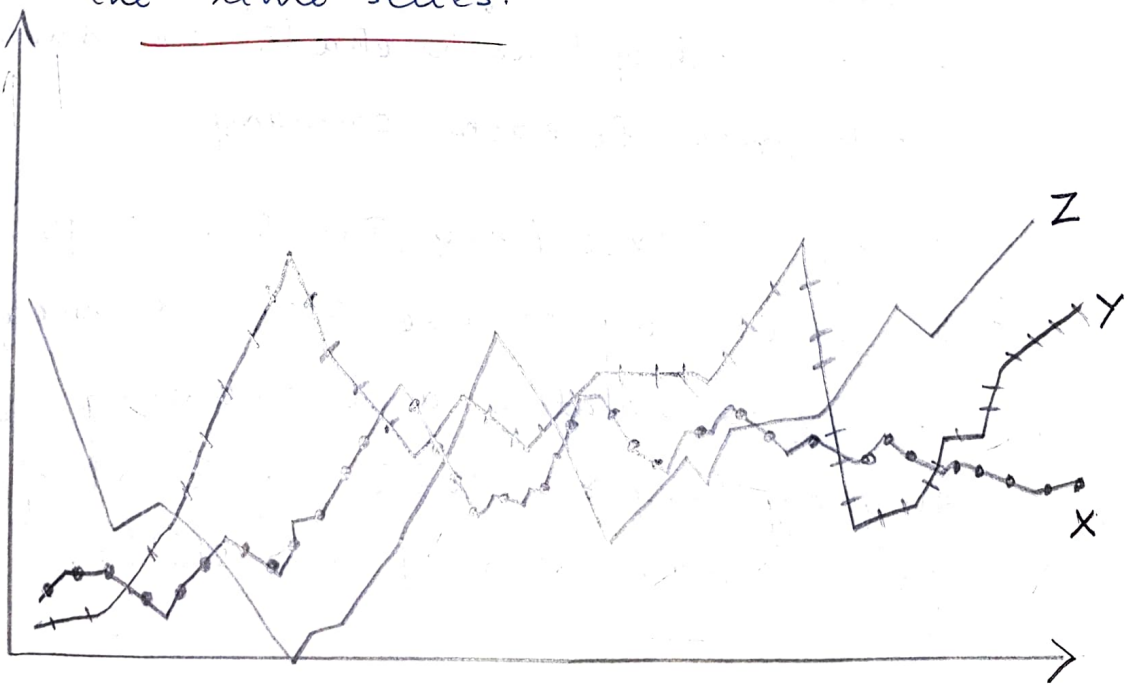
---

③ Time Series Analysis.

- With time series analysis, the value of an attribute is examined as it varies over time.

- The values usually are obtained as evenly spaced time points (daily, weekly, hourly etc).

- A time series plot is used to visualize the time series.



Time Series Plots

From the plots we see that Y and Z have similar behavior while X is not.

Time Series analysis performs 3 basic functions.

1) Distance measures are used to determine the similarity b/w different time series
2) Structure of the line is examined to determine its behavior.
3) use historical time series plot to predict future values.

eg: A person likes to determine whether to purchase stock from companies X, Y or Z. For a period of time he charts the daily stock price for each company.

Using this info (fig: Time Series plots) he decides to purchase stock X because it is less volatile and showing a slightly larger relative amount of growth than either of other stocks.

(4) Prediction

Many real-world data mining applications can be seen as predicting future data. stalis based on past and current data.

Prediction can be viewed as a type of classification.

Prediction applications include flooding, speech recognition, machine learning and pattern recognition.

Although future values may be predicted using time series analysis or regression techniques., other approaches may be used as well.

eg: Predicting flooding is a difficult problem. One approach uses monitors placed at various points of a river. These monitors collect data relevant to flood prediction. water level, rain amount, time, humidity etc. Then the water level at a potential flooding point in the river can be predicted based on data collected by sensors upriver from this point. The prediction ~~can be~~ must be made w.r.t time the data were collected.

Descriptive Models.

① Clustering.

Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone.

- type of unsupervised learning approach
- It can be thought as partioning or segmenting the data into groups that might or might not be disjointed.
- clustering is accomplished by determining the similarity among data on predefined attributes.
- Most similar data are grouped into clusters.

eg: A national department store chain creates special catalogs targeted to various demographic groups based on attributes such as income, location, physical characteristics like age, height weight etc. To determine the target mailings of the various catalogs and to

assist in the creation of new, more specific catalog, a company performs a clustering of potential customers based on the determined attribute values. The results of the clustering exercise are then used by management to create special catalogs and distribute them to correct target population based on the cluster for that catalog.

② Summarization:

Summarization maps data into subsets with associated simple descriptions.

barchart pie chart histogram.

- also called generalization or characterization

mean, mode SD.

- extracts or derives representative information about the database.
- This can be accomplished by actually retrieving portions of data
- Summary type information such as mean of some numeric attribute. can be derived from data.
- Summarization succinctly characterizes the contents of the DB.

eg: One of the many criteria used to compare universities by the U.S. News & World Report is the average SAT or ACT Score. This is a summarization used to estimate the type and intellectual level of the student.

③ Association Rules.

- Link analysis, alternatively referred to as affinity analysis or associations refer to data mining task of uncovering relationships among data.
- The best example of this type of application is to determine association rules.
- An association rule is a model that identifies specific type of data associations.
- These associations are often used in the retail sales community to identify items that are frequently purchased together.

eg: Use of association rules in market basket analysis. Data analyzed consists of information about what items a customer purchases.

A grocery store retailer is trying to decide whether to put bread on sale. To help determine the impact of this decision, the retailer generates association rules that show what other products are frequently purchased with bread.

He finds that 60% of the time that bread is sold so are pretzels and that 70% of time jelly is also sold.

Based on these facts, he tries to capitalize on the association b/w bread, pretzels and jelly by placing some pretzels and jelly at the end of the aisle where the bread is placed.

Users of the association rules must be cautioned that these are not causal relationships. They do not represent any relationship inherent in the actual data or in the real world.

There is no relationship b/w bread and pretzels that causes them to be purchased together and there is no guarantee that this association will apply in the future.

Association rules can be used to assist retail store management in effective advertising, marketing and inventory control.

(4) Sequence Discovery.

Sequence analysis or sequence discovery is used to determine sequential patterns in data.

These patterns are based on time sequence of actions.

These patterns are similar to associations in that data (or events) are found to be related, but the relationship is based on time.

Unlike a market basket analysis, which requires the items to be purchased at the same time, in sequence discovery the items are purchased over time in some order.

eg: The WebMaster of at the XYZ Corp. periodically analyzes the web log data to determine how users of XYZ's webpages access them. He is interested in determining what sequence of pages are frequently accessed. He determines that 70% of the users of page A follow one of the following patterns of behavior: <A, B, C> or <A, D, B, C>, or <A, E, B, C>. He then determines to add a link directly from page A to page C.

Data Warehousing (DWH)

- Data warehousing provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions.
- Refers to a DB that is maintained seperately from an operational DB.
- A data warehouse is a subject oriented, integrated, time variant and non volatile collection of data in support of management's decision making process.

a) Subject-Oriented.

- organized around major subjects.
- rather than concentrating on the day-to-day operations & transaction processing of an organization, DW focuses on the modelling and analysis of data for decision makers.
- provide a simple and concise view around particular subject issues.
- exclude data that are not useful in the decision support process.

b) Integrated : A data warehouse is usually constructed by integrating multiple heterogeneous sources such as relational DBs, flat files and online transaction records

- Data cleaning and integration techniques are applied to ensure consistency in naming conventions, encoding structures attribute measures and so on.

c) Time-variant : Data are stored to provide information from a historical perspective (eg: past 5-10 years). Every key structure in the data warehouse contains either implicit or explicit, an element of time.

d) Non-volatile: A data warehouse is always a physically separate store of data found at the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms.

It usually requires only 2 operations in data accessing : initial loading of data and access of data.

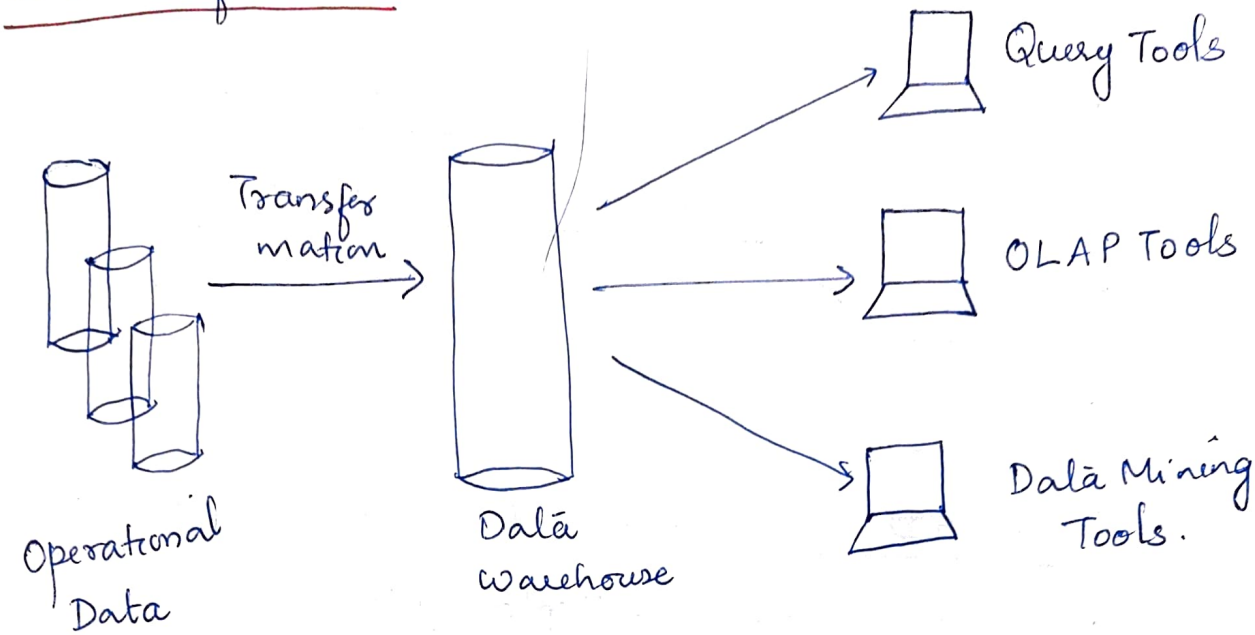


fig : Simple View of Data Warehouse.

Multi Dimensional Data Model.

- Data warehouses and OLAP tools are based on multidimensional data model. This model views data in the form of a data cube.

- A data cube allows data to be modelled and viewed in multiple dimensions.

- It is defined by dimensions and facts.

- Dimensions are entities or perspectives w.r.t which an organization wants to keep records. Each dimension may have a table associated with it called a dimension table.

eg: time: dimension table.

time_key
day
day_of_the_week
month
quarter
year.

- A multidimensional data model is typically organized around a central theme, represented by a fact table.

- Facts are numerical measures — quantities by which we want to analyze relationships b/w dimensions.

eg: 3D Data Cube.

Table 4.3 3-D View of Sales Data for *AllElectronics* According to *time*, *item*, and *location*

	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Note: The measure displayed is *dollars_sold* (in thousands).

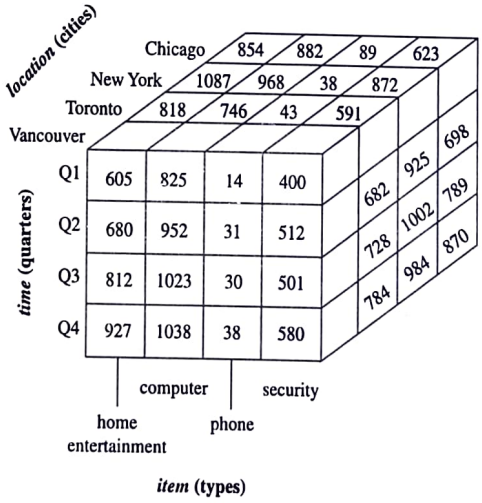


Figure 4.3 A 3-D data cube representation of the data in Table 4.3, according to *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

For the AllElectronic shop, we create a Sales datawarehouse in order to keep records of the store's sales w.r.t dimension time, item, branch and location.

These dimensions allow the store to keep track of things like monthly sales of items and the branches and locations at which the items were sold.

Examples of facts for a sales data warehouse include dollars sold (sales amount in dollars), units sold (number of units sold) and amount budgeted.

Schemas for Multi Dimensional Datawarehouse

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships b/w them. Such a data model is

appropriate for online transaction processing

- A data warehouse requires a concise, subject oriented schema that facilitates online data ~~analysis~~ analysis.

The most popular data model for a data warehouse is a multidimensional model.

Such a model exist in the form of a star schema, snowflake schema, or a fact constellation schema.

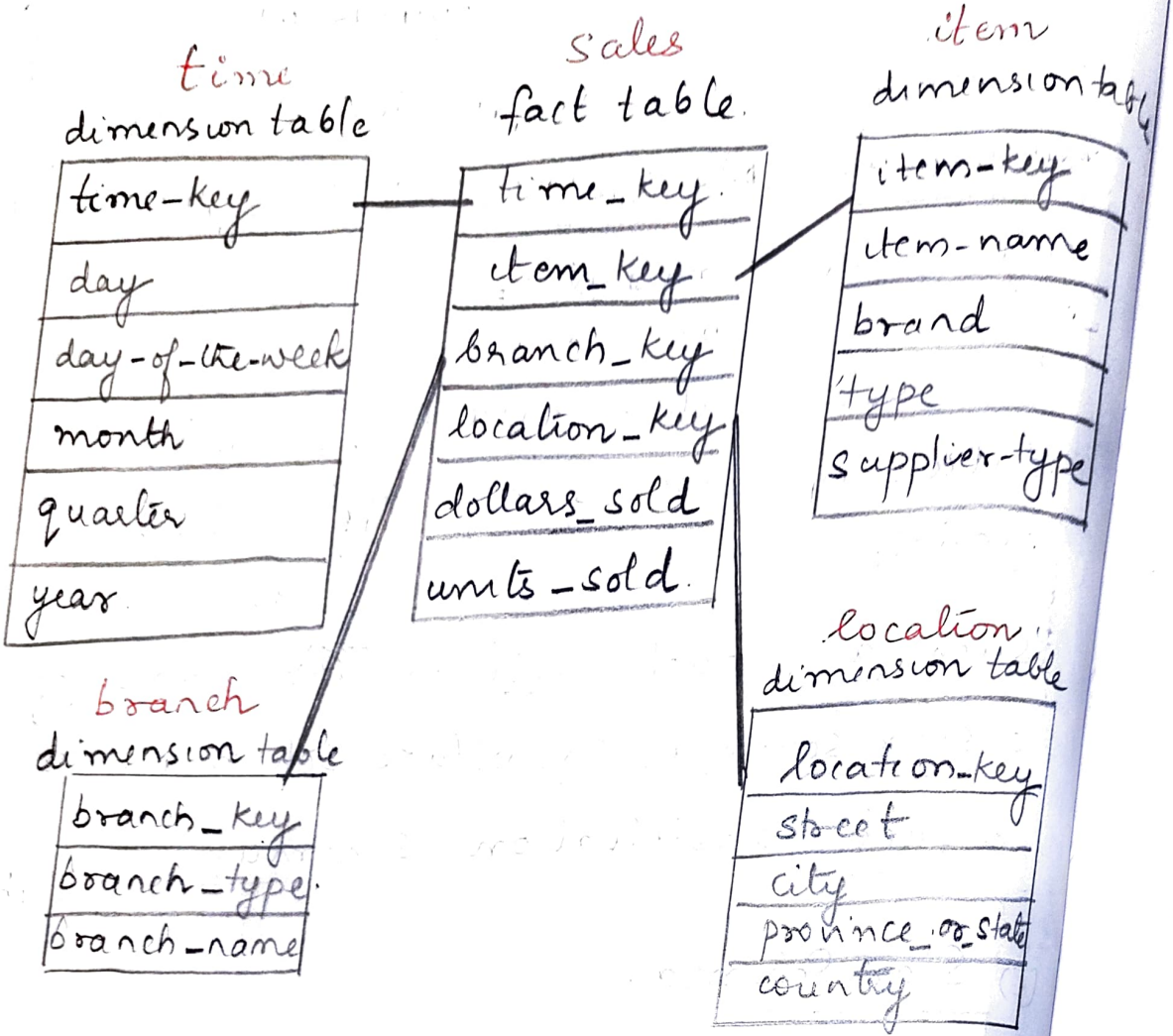
(i) Star Schema.

- Most common modelling paradigm. is the star schema in which the data warehouse contains.

(i) a large central table (fact table) containing the bulk of data, with no redundancy.

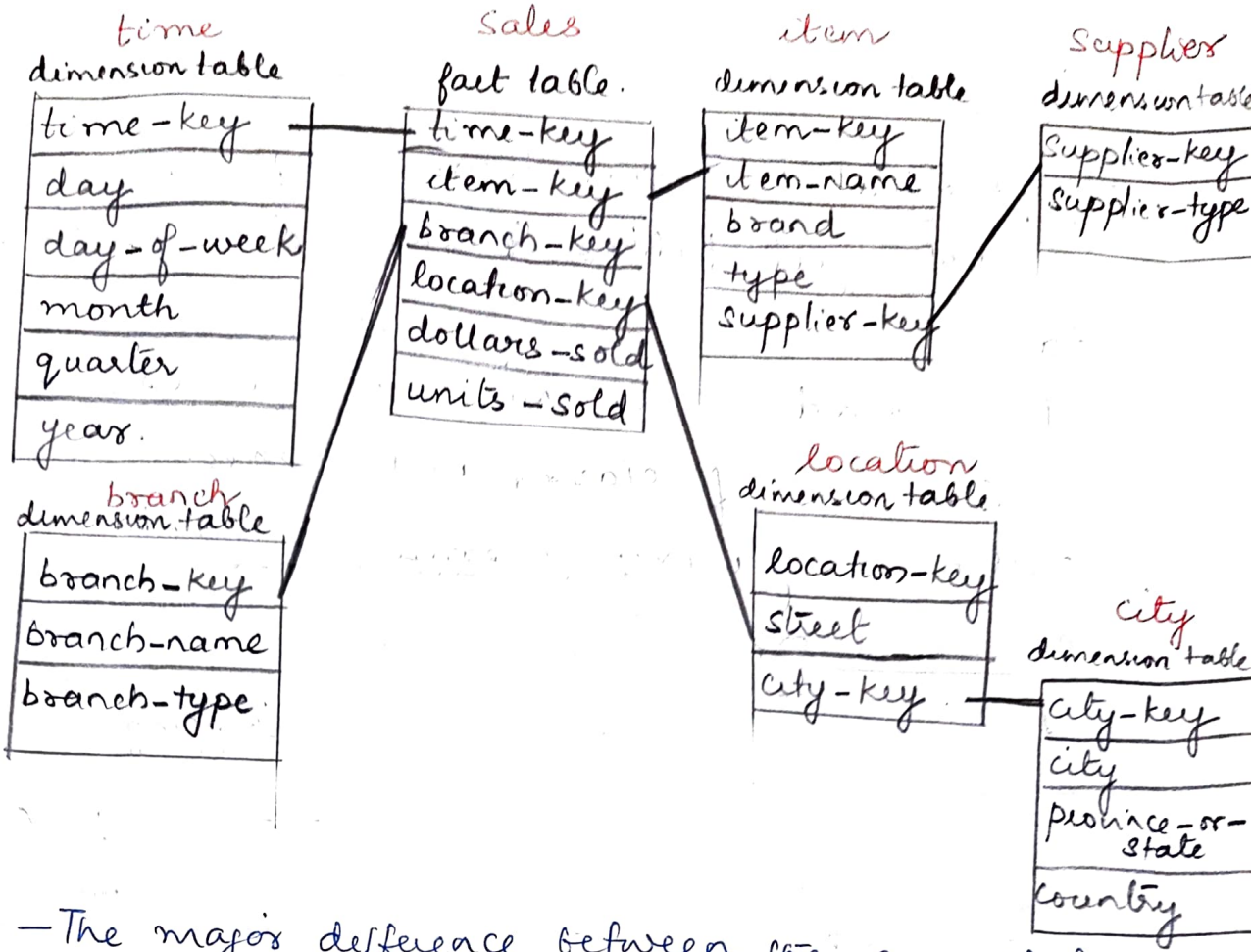
(ii) set of smaller attendant tables (dimension tables) one for each dimension

The schema graph resembles a starburst with the dimension tables displayed in a radial pattern around the central fact table.



(ii) Snow-flakes Schema.

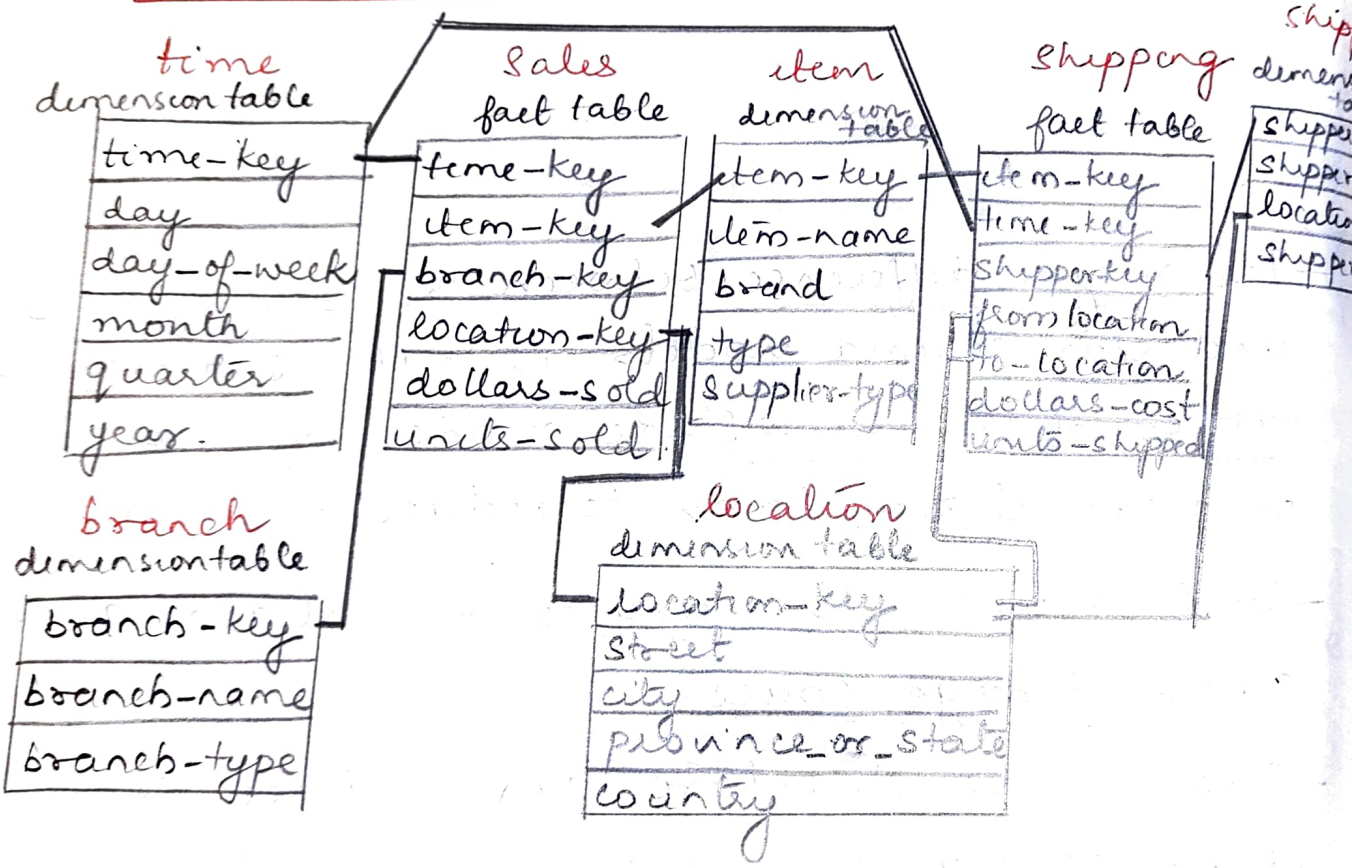
- variant of the star schema model where some dimension tables are normalized thereby further splitting the data into additional tables.
- The resulting schema graph forms a shape similar to a snowflake.



- The major difference between the snowflakes and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies.
- Such a table is easy to maintain and saves storage space.
- Snowflake structure can reduce the effectiveness of browsing since more joins will be needed to execute a query.
- The system performance may be adversely impacted.
- Although the snowflakes schema reduces redundancy, it is not popular as the star schema in data warehouse design.

iii) Fact Constellation:

- Sophisticated applications may require multiple fact tables to share dimension tables
- This kind of schema can be viewed as a collection of stars, and hence it is called a galaxy schema or fact constellation



Typical OLAP operations

1. Roll-Up (Drill up)

- Summarize data by climbing up hierarchy or by dimension reduction

2. Drill Down (Roll-Down)

- Reverse of Roll-Up.
- from higher level summary to lower level summary or detailed data, or introducing new dimensions

3. Slice and Dice.

Slice operation performs a selection on one dimension of the given cube, resulting in a sub cube.

Dice operation performs a selection on two or more dimensions.

4. Pivot:

Rotate (Reorient) the cube, visualization 3D to series of 2D planes.

Other Operations:

Drill across: involving (across) more than one fact table

Drill through: through the bottom level of the cube of its backend relational tables (using SQL)

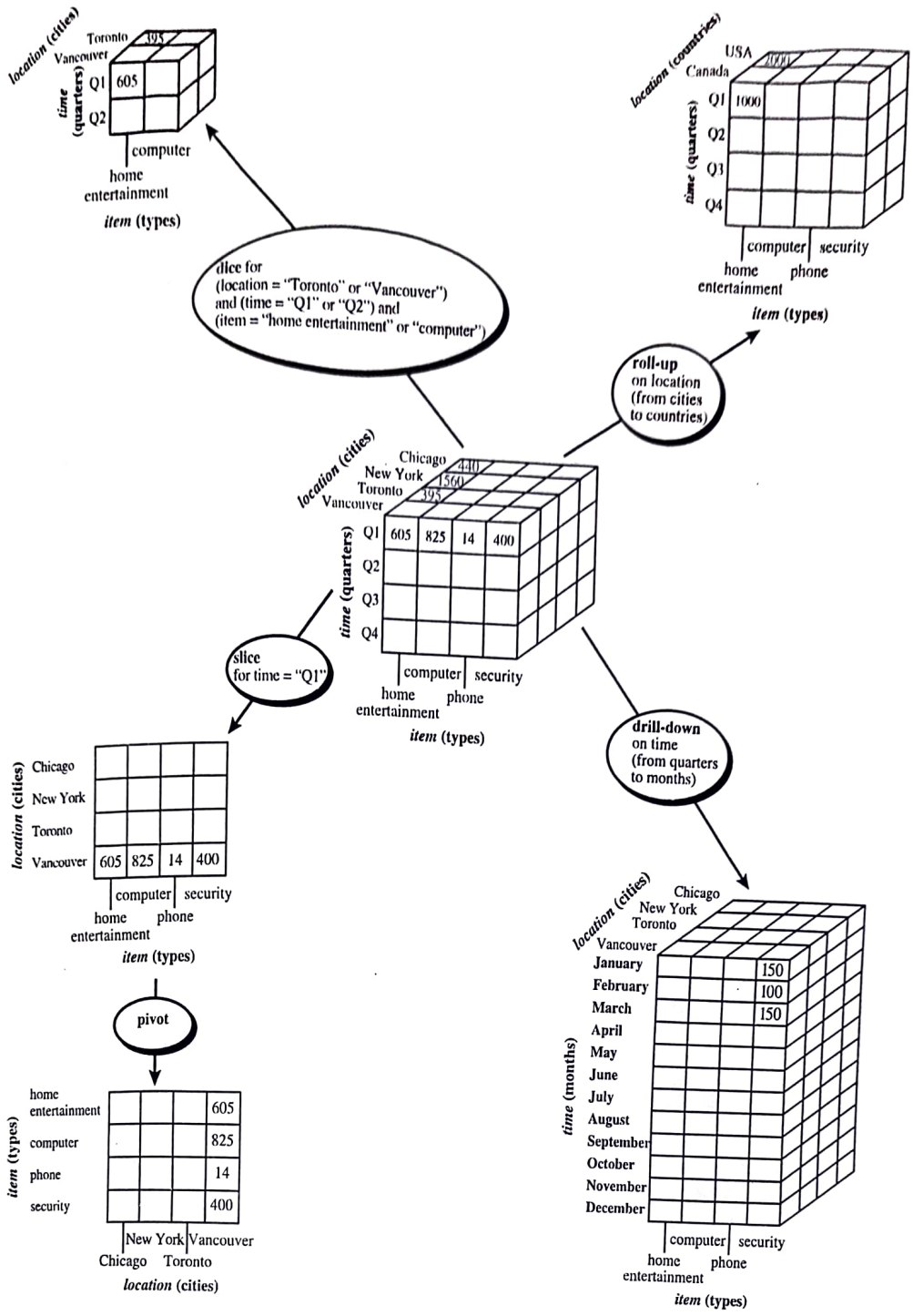
Egs of OLAP operations

1) roll-up.: on location (from cities to countries)

2) drill down: on time (from quarters to months)

3). slice: (for time Q1)

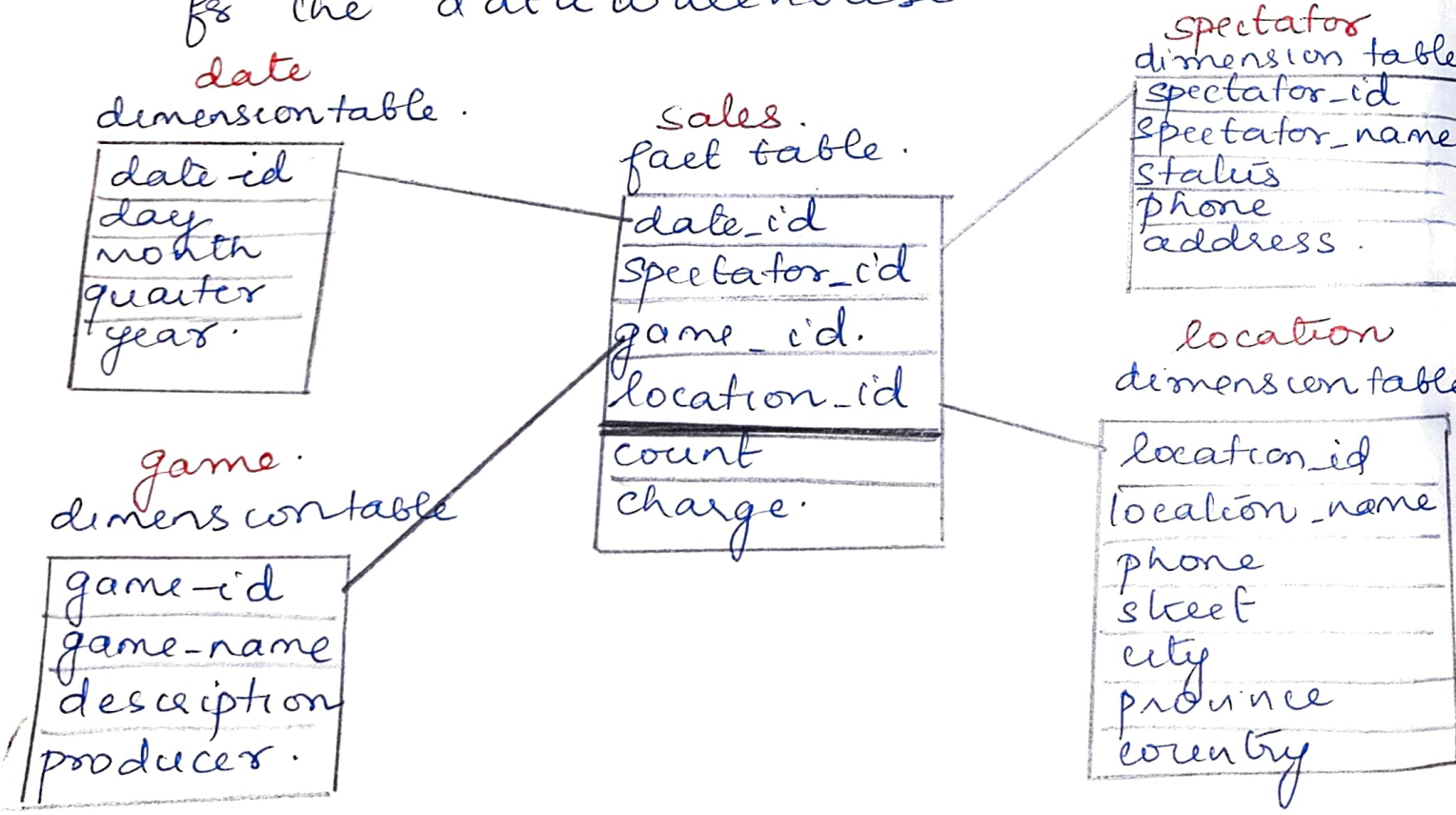
4) dice: (location = Toronto or Vancouver and time = Q1 or Q2 and item = "home entertainment" or "computer")


location (cities)
Toronto 395
Vancouver
time (quarters)
Q1 605
Q2
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (countries)
USA 2000
Canada
Q1 1000
Q2
Q3
Q4
time (quarters)
computer security
home entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago 440
New York 1560
Toronto
Vancouver
Q1 605 825 14 400
Q2
Q3
Q4
time (quarters)
computer security
home entertainment
phone
item (types)
slice
for time = "Q1"
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver 605 825 14 400
computer security
home entertainment
phone
item (types)
pivot
item (types)
home entertainment 605
computer 825
phone 14
security 400
New York Vancouver
Chicago Toronto
location (cities)
location (cities)
Chicago
New York
Toronto
Vancouver
January 150
February 100
March 150
April
May
June
July
August
September
October
November
December
time (months)
computer security
home entertainment
phone
item (types)

Q. Suppose that a data warehouse consists of the 4 dimensions : date, spectator, location and game, and the two measures, count and charge; where charge is the fee that a spectator pays when watching a game on a given date.

Spectators may be students, adults or seniors, with each category having its own charge rate.

a). Draw a star schema diagram for the data warehouse.



b) Starting with the base cuboid [date; spectator; location; game] what specific OLAP operations should be performed in order to list the total charge paid by student spectators at GM Place in 2004?.

The Specific OLAP operations to be performed are.

→ Roll-Up on date from date id to year.
→ Roll-Up on spectator from spectator-id to status.
→ Roll-Up on location from location-id to locationname.
→ Roll-Up on game from game-id to all.
on spectator, date, location
→ Dice with status = "students",
location name = "GM Place" and year = "2004".

never pen and pencil are purchased together

rent class

Q2. Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester and instructor and two measures count and avg grade. When at the lowest conceptual level (eg: for a given student, course, semester and instructor combination) the avg. grade measure stores the actual course grade of the student.

At higher conceptual level, avg. grade stores the average grade for the given combination.

a) Draw a Snowflake schema diagram for the data warehouse.

b) Starting with the base cuboid [student; course; semester; instructor] what specific OLAP operations should be performed in order to list the average grade of CS courses for each Big University student?

(a)

course dimension table

course-id
course-nam
departmen

semester dimension tab

semester
semester
year.

(b)

→ Roll

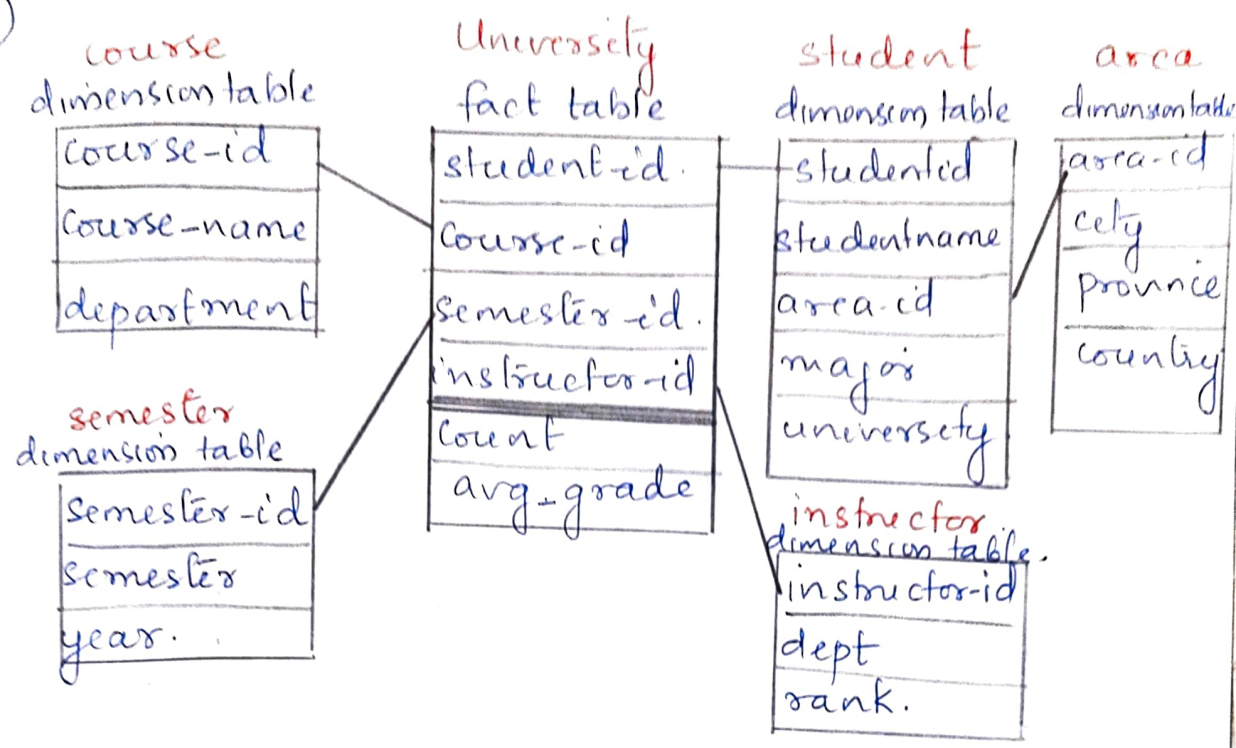
→ Roll-

→ Roll

→ Dic

an

(a)



(b)

→ Roll-Up on course from course-id to department

→ Roll-Up on student from studentid to university

→ Roll-Up on semester from semester-id to all.

→ Dice on course, student with department = "CS" and university = "Big University".

+Module I
printed Notes