Types of databases

Section - 2

We can apply datamining techniques to the following types of databases..

(1) Relational databases.
(2) Data warehouse
(3) Transactional databases.
(4) Spatial databases.
(5) Time-series databases.
(6) Text databases.
(7) multimedia databases.

(4) to (7) ← Advanced database system.

(8) world wide web.

Relational databases

→ A relational database is a collection of tables, each having a unique name.
→ Each table consists of set of attributes (columns or fields) and stores a large set of tuples (records or row).
→ Each tuple represents an object identified by a unique key
→ An E-R model is used to show the relationships.
→ With the help of SQL, we can write database queries to access the relational data.

customer (table eg)

cust ID	Name	address	age	income	credit info

item (table)

item ID	name	brand	category	type	price	place made	supplier	cost

employee (table.

empl_ID	name	category	group	salary	commission

branch

branch_ID	name	address

purchases

trans_ID	cust_ID	empl_ID	date	time	method Paid	Amount

items_sold

trans_ID	item_ID	qty.

works_at

empl_ID	branch_ID

* A given query is transformed into a set of relational operations such as join, selection and projection and is then optimized for processing.
→ A query allows retrieval of specified subsets of the data.
→ For example to show a list of all items that were sold in the last quarter can be done
→ relational languages also include aggregate functions such as sum, average, count, max, and min.
→ This help us to show the total sales of the last month grouped by branch or how many trans. occurred in Dec. etc.
→ When data mining is applied to relational databases, we can search for trends or data patterns.

Ex- Data mining systems can analyze customer data to predict the credit risk of new customers based on their income, age and previous credit information.

→ Data mining also detect deviations
Ex ~~comparison of sales amount~~
~~of this any expected sales~~
actual
amount
such as items whose sales are ~~are~~ from those expected in comparison to the previous year. Such deviations can further be investigated.

Data Warehouse

Let us say an organization is having branches around the world. Each branch has its own set of databases.

The president of the organization asked to provide an analysis of the company's sales / per item / per branch for the 3rd quarter.

This is a difficult job, since the relevant data are spread out over several databases, physically located at different sites.

That is the reason why we are using data warehouse.

A data warehouse → is a repository of information collected from multiple sources, stored under a unified schema, and resides at a single site

→ data are stored the past 5-10 years and are typically summarized.

→ A data warehouse is usually modeled by a multidimensional database

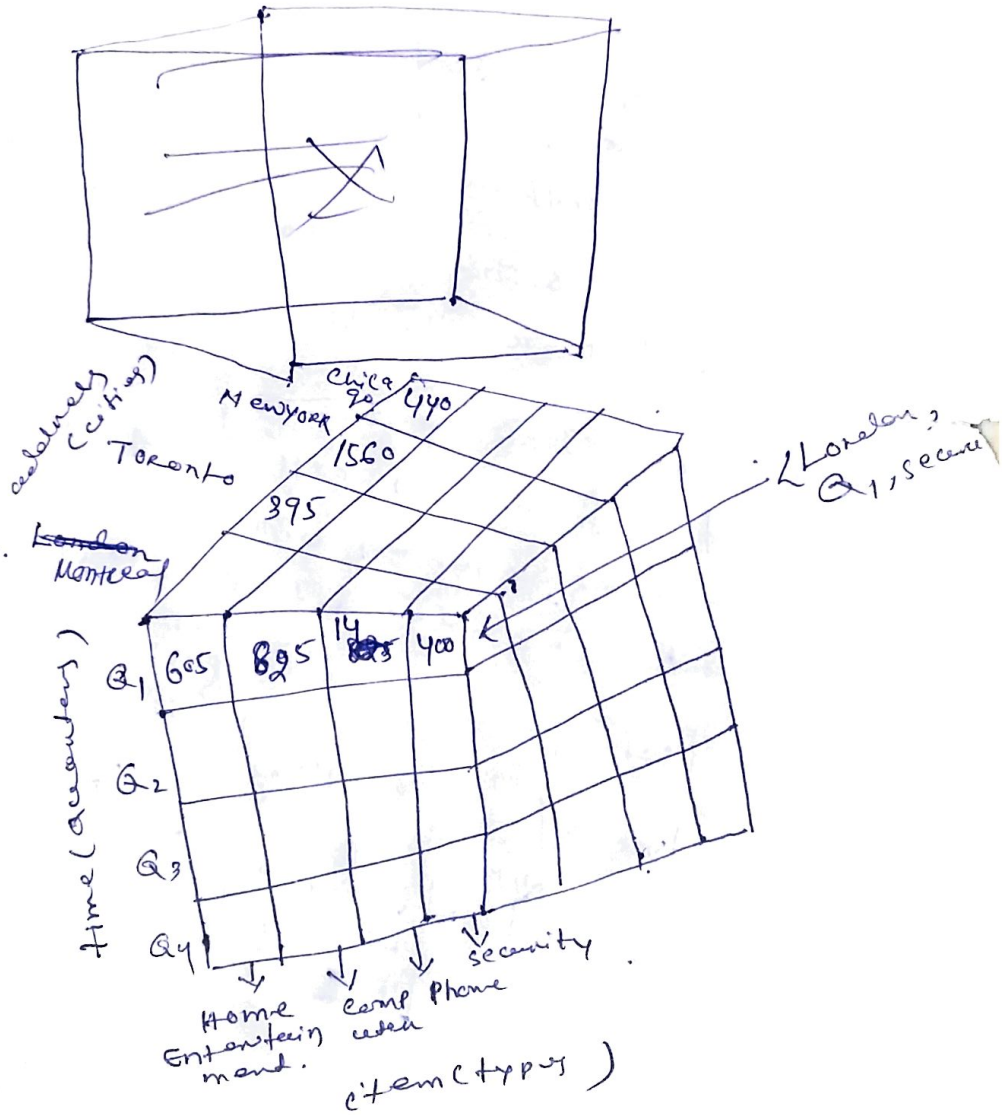
Structure.

→ Each dimension corresponds to an attribute or a set of attributes in the schema and each cell stores the value of some aggregate measure

Ex: count or sales_amount.

→ The actual physical structure of a data warehouse is a multidimensional data cube.

→ A data cube gives a multidimensional view of data and allows the pre computation and fast accessing of summarized data.



A multidimensional data cube. showing summarized data for a organization.

Data warehouse → collects information about subjects that cover an entire organization.

Data mart → is a dept. subset of a data warehouse. It focuses on selected subjects and its scope is dept.-wide.

605 825 14 400

those ... year. such deviations can be ... analysis.

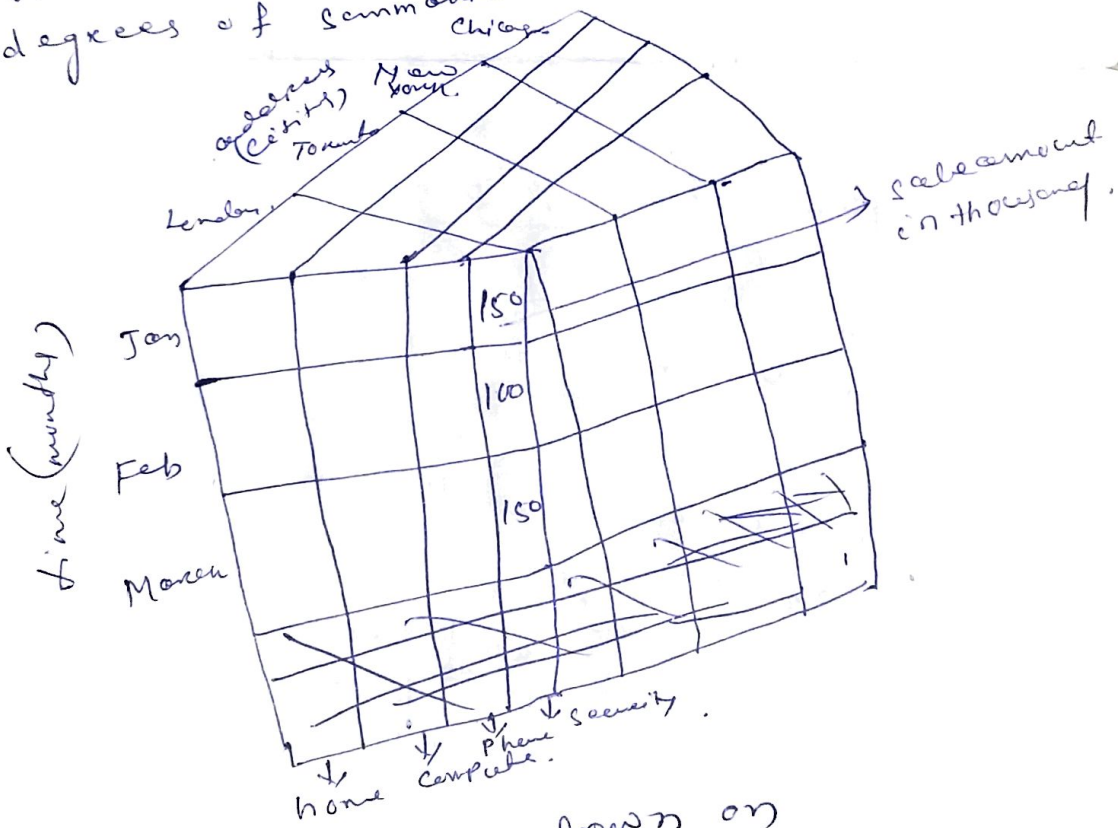
→ The cube has three dimensions; address, time and item type.

→ The aggregate value stored in each cell of the cube is sales_amount

Ex- The total sales for the 1st quarter Q1, for items relating to security in London is $400,00.

→ by providing multidimensional data views & the precomputation of summarized data, data warehouse is well suited for OLAP (Online analytical processing)

→ OLAP operations include drill-down and roll-up, which allow the user to view the data at different degrees of summarization.



(Drill-down on time data for Q1)

Drill-down on sales data summarized by quarter to see the data summarized by month.

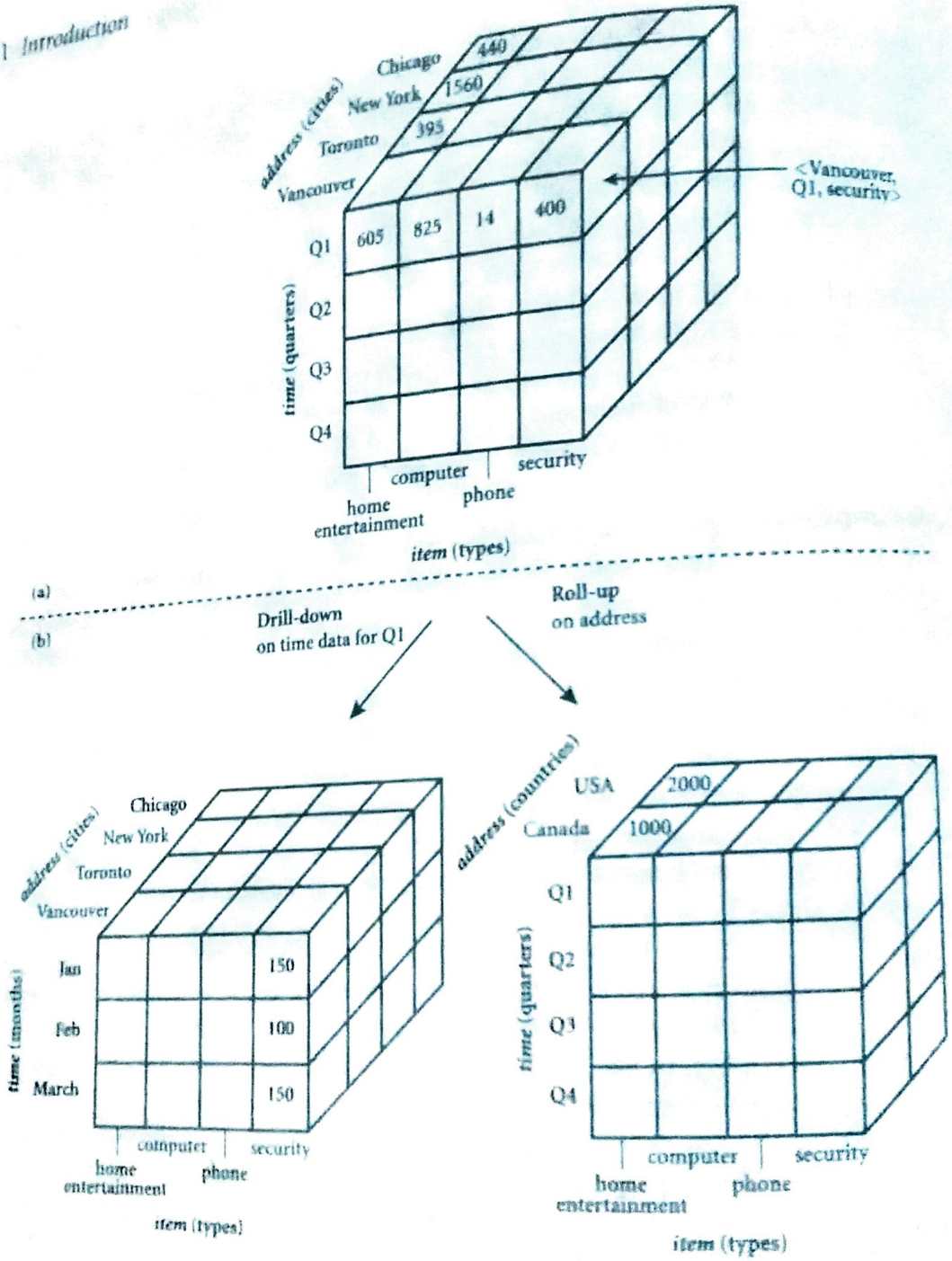
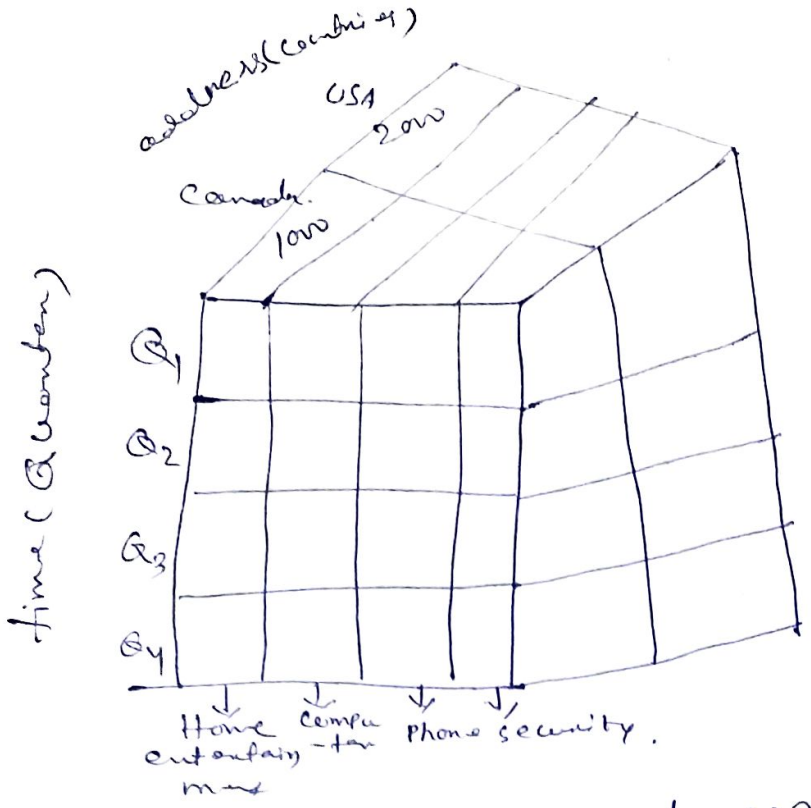


Figure 1.8 A multidimensional data cube, commonly used for data warehousing, (a) showing summarized data for *AllElectronics* and (b) showing summarized data resulting from drill-down and roll-up operations on the cube in (a). For improved readability, only some of the cube cell values are shown.

1.3.3 Transactional Databases

In general, a transactional database consists of a file where each record represents a transaction. A transaction typically



(Roll-up on address)

Roll-up on sales data summarized by city to view the data summarized by country.

Transactional Databases

→ It consists of a file where each record represents a transaction.

→ A transaction includes a unique transaction identity number and a list of items making the transaction.

Sales (Table) —

Fragment of a trans. database for sales.

trans-ID	list of items_IDs
T100	I1, I3, I8, I16
T200	I2, I8

→ It may have additional tables associated with it, which contain other information regarding the sale. For example date of transaction, customer ID no., the ID no. of sales person and so on.

The sales table in the above figure is a nested relation because the attribute list of item_IDs containing a set of items.

→ Relational database systems do not support nested relational structures.

→ Hence, the transactional database is usually either stored in a flat file in a format similar to that of the sales table.

→ or unfolded into a standard relation in a format similar to that of the items_sold table. (Relational table).

→ If we want to know how many transactions include item no. 13? we can answer it by scanning entire database

→ Suppose we ask, which items sold well together?".

→ Market basket data analysis would enable us to bundle groups of items together as a strategy for maximizing sales.

→ Ex The printers are commonly purchased together with computers. we could offer an expensive printers at a discount rate to customers buying computers, in the hopes to increase the selling of the expensive printers.

→ Data mining systems can identify frequent itemsets, i.e sets of items that are frequently sold together.

Object Relational Databases

→ This type of Databases are constructed based on an object-relational data model.

→ The object-relational data model inherits the essential concepts of object-oriented databases, where each entity is considered as an object.

Object oriented concepts
Object
class
encapsulation
inheritance

Example → employees, customers or items are objects.

→ Data and code relating to an object are encapsulated into a single unit.

→ Each object has associated with it the followings :-

(i) A set of variables that describe the objects.

(ii) A set of messages that the object can use to communicate with other objects or rest of the database.

(iii) A set of methods, where each method holds the code to implement a message. After receiving a message, the method returns a value in response.

→ objects that share a common set of properties can be grouped into an object class.

→ Each object is an instance of its class.

→ object classes can be organized into classes/subclasses so that each class represents properties that are common to objects in that class.

Ex An employee class can contain variables like name, address and DOB. Suppose the class, salesperson is a subclass of the class employee.

A sales-person object would inherit all of the variables pertaining to its superclass employee. In addition, it has all of the variables that pertain to salesperson.

Data mining techniques needed for handling complex object structures, complex data types, class and subclass hierarchy, property inheritance etc.

Booking

name	Room	time
DB Group	PC 133	6/1/20 10am

Temporal Databases → It stores relational data that includes time-related attributes. Example: Booking (meeting, room, time)

~~Ex customer shopping sequence~~

A tuple (m, r, t) denotes the fact that meeting m is in room r at time t.

Sequence Databases → stores sequences of ordered events, with or without a concrete notion of time.

Ex customer shopping sequences;
Biological sequences → Formation of DNA

Time series database → stores sequences of values or events obtained over repeated measurements of time (i.e hourly, daily, weekly).

Ex stock exchange data.
Inventory control
Change in Temperature

Data mining techniques can be used to find the trend of changes for objects in the database.

Such information can be useful in decision making and strategy planning.

Spatial databases

It contain spatial (map) related information.

Examples → geographic databases, VLSI, medical and satellite images

The data represents objects defined in geometric shape.

Temporal Database.

Emp ID	Name	Dept	Salary	Valid time Start	Valid time end.
10	John	Research	1100	1985	1990
10	John	Sales	1100	1990	1993
10	John	Sales	1200	1993	till date
11	paul	Research	1000	1988	1995
12	George	Research	1050	1991	till
13	Ringo	Sales	1550	1998	till

time period attached to the data. enquiries when it was valid

Conventional databases do not keep track of past or future DB states.

Raster format
a rectangular grid of pixels (picture elements)

Spatial data are represented in raster format consisting of n-dimensional bit maps or pixel maps.

Ex 2-D satellite image → where each pixel registers the rainfall in a given area.

Vector format
2D points which are connected by lines and curves to form polygons and other shapes

Maps are represented in vector format. where roads, bridges, buildings and lakes are represented as unions of basic geometric shapes.

Geographic databases have numerous applications :-

(1) Forestry and ecology planning.

(2) Providing public service information regarding the location of telephone and electric cables, pipes and sewage system.

(3) Vehicle navigation and dispatching system.

Ex → taxis can store a city map with all information about the city, as well as the current location of each driver.

Data mining may discover patterns which describe the climate of mountains areas

Text Databases

These contain word descriptions for objects. These words are not simple keywords but rather long sentences or paragraphs.

Ex - product specification,
error reports,
warning messages.

Text databases → highly unstructured → Ex - web page
→ semi structured → Ex - e-mail
→ structured → Ex: library catalogue

Multimedia databases

These store image, audio and video data.
Applications → voice-mail, video on demand, WWW, speech based user interface

Heterogeneous Databases

→ It consists of a set of interconnected, autonomous component databases.

→ The components communicate in order to exchange information and answer queries.

→ Objects in one component may differ from objects in other component.

Legacy database

It is a group of heterogeneous databases that combines different kinds of data systems.

Ex - relational or object oriented database.
Network databases.
spread sheets.
multimedia databases.

→ Information exchange across such databases is difficult, as they are in different form.

Data Streams

Here data flow in and out of a window dynamically or continuously.

Ex :- fine series data
power supply.
stock exchange
telecommunication.
weather or environment monitoring

WWW

→ users traverse from one page to another page.

→ Understanding users access patterns will help us to improve marketing decision. we can put some advertisement.

→ Capturing users access patterns in such distributed information environment is called web log mining.

Q.1 What is data mining?

2. Describe the steps involved in data mining when viewed as a process of knowledge discovery (KDD)

3. suppose your job as a S/W Engineer at GAO is to design a data mining system to examine the univ course database, which contains the following information: the name, address and status of each student, the courses taken and the cumulative grade point average (CGPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?

(4) How is data warehouse different from database? How they are similar?

(5) Define each functionality of data mining.

(7) Noisy data → salary = -10
Incomplete data ⇒ salary = " "
inconsistent data ⇒ Age = 29, DOB = 24/7/1986

Second Edition

ELSEVIER

Data Mining

Concepts and Techniques

Jiawei Han and Micheline Kamber