Data Mining:
Concepts and Techniques

— Unit 1 —

— Introduction —

Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Classification of data mining systems
- Major issues in data mining
- Overview

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes

  (1000 Terabytes = 1 Petabyte)

  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube

Disk Storage

- 1 Bit = Binary Digit
  - 8 Bits = 1 Byte
  - 1024 Bytes = 1 Kilobyte
  - 1024 Kilobytes = 1 Megabyte
  - 1024 Megabytes = 1 Gigabyte
  - 1024 Gigabytes = 1 Terabyte
  - 1024 Terabytes = 1 Petabyte
  - 1024 Petabytes = 1 Exabyte
  - 1024 Exabytes = 1 Zettabyte
  - 1024 Zettabytes = 1 Yottabyte
  - 1024 Yottabytes = 1 Brontobyte
  - 1024 Brontobytes = 1 Geopbyte

- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”
- Data mining—Automated analysis of massive data sets

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
- 1950s-1990s, **computational science –** simulation , models
- 1990-now, **data science**
  - The flood of data
  - The ability to economically store and manage data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Here is the concept of Data mining.

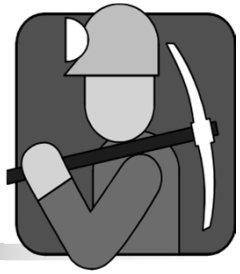
Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

What Is Data Mining?

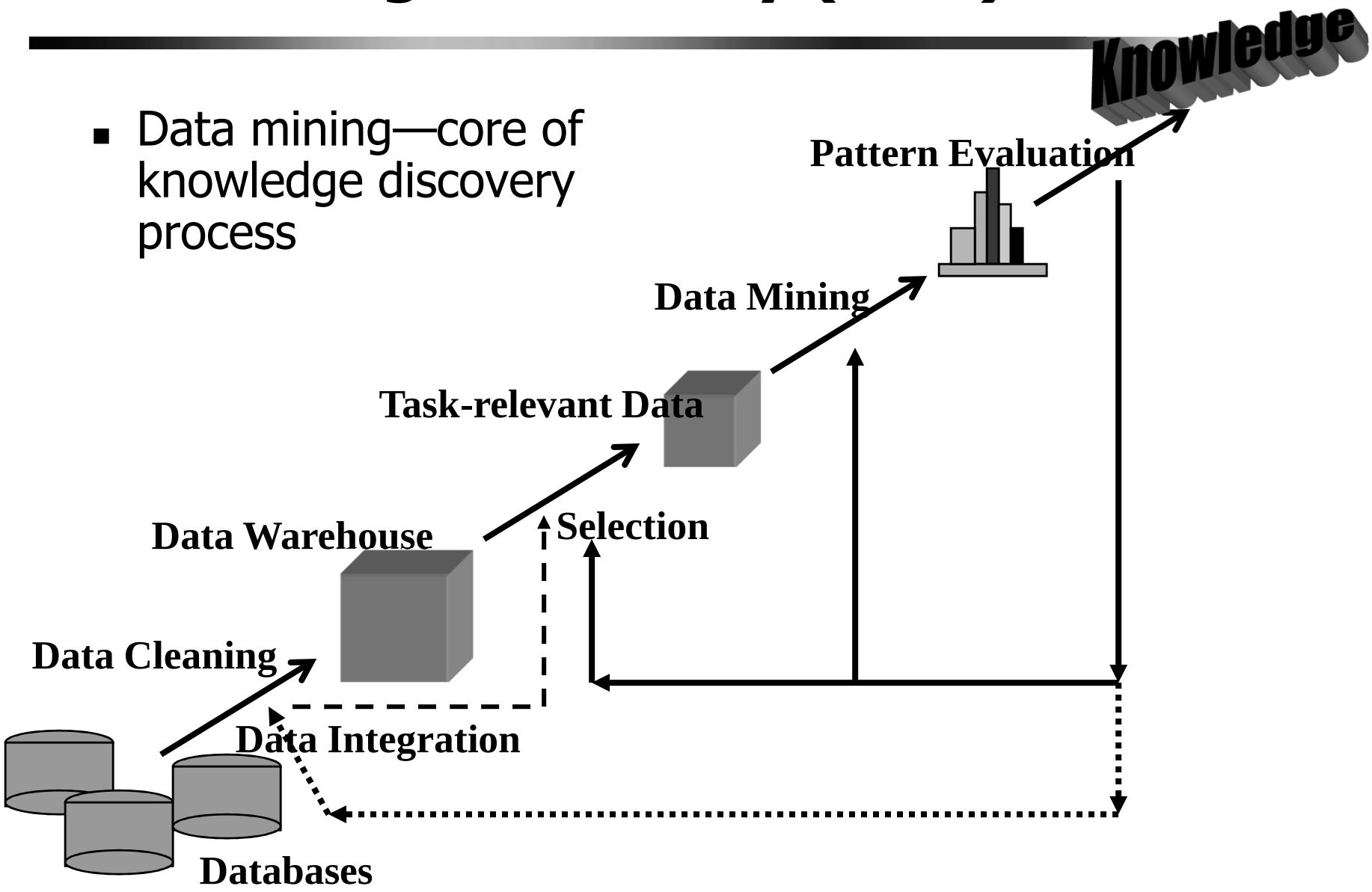


- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

- Alternative names
- Knowledge discovery (mining) in databases (KDD),
- knowledge extraction
- Data/pattern analysis
- Data archeology
- Data dredging
- Information harvesting
- Business intelligence, etc

Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



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 the analysis task are retrieved fromthe database)

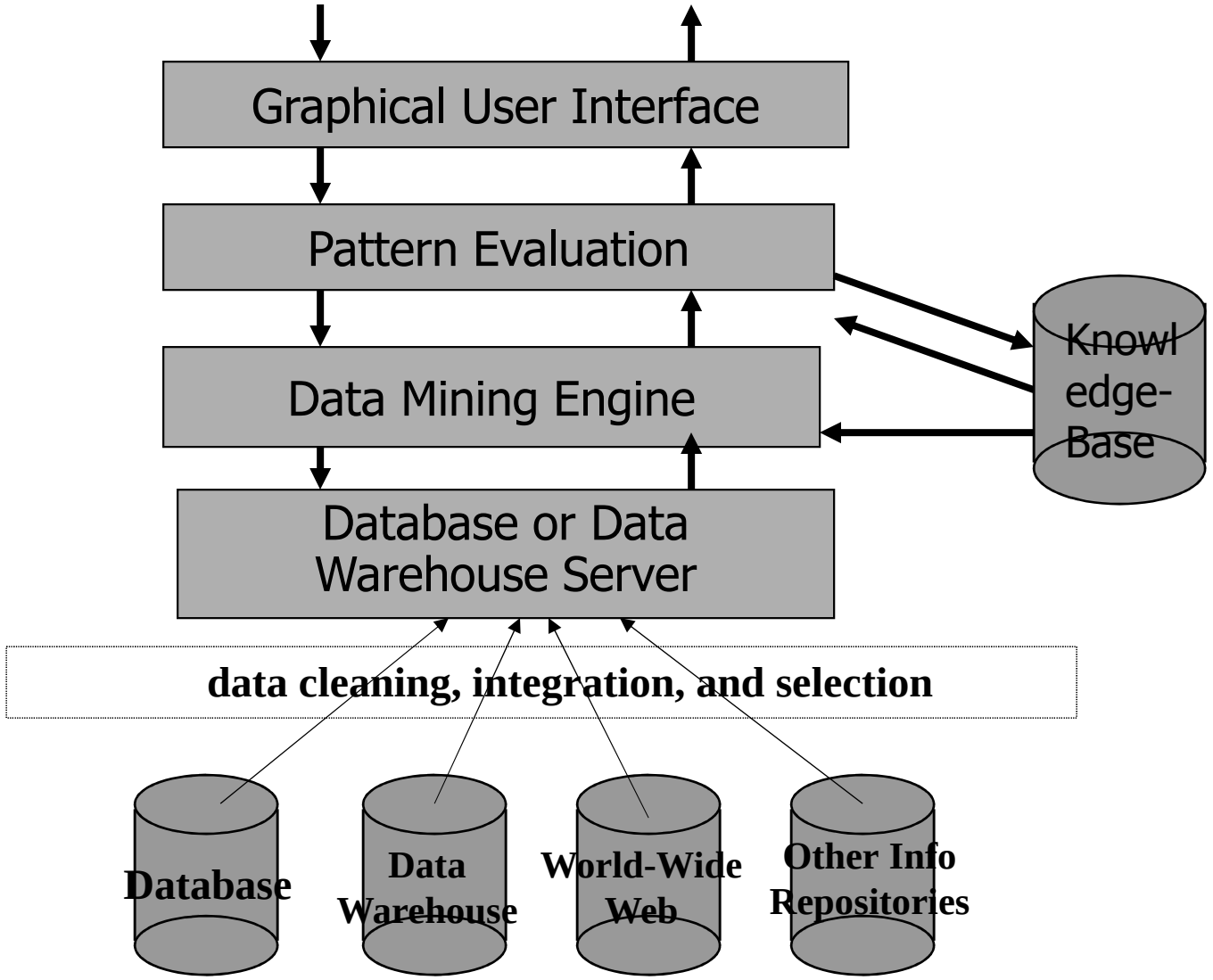
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

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)

Architecture: Typical Data Mining System



- Database, data warehouse, WorldWideWeb, or other information repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

- Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

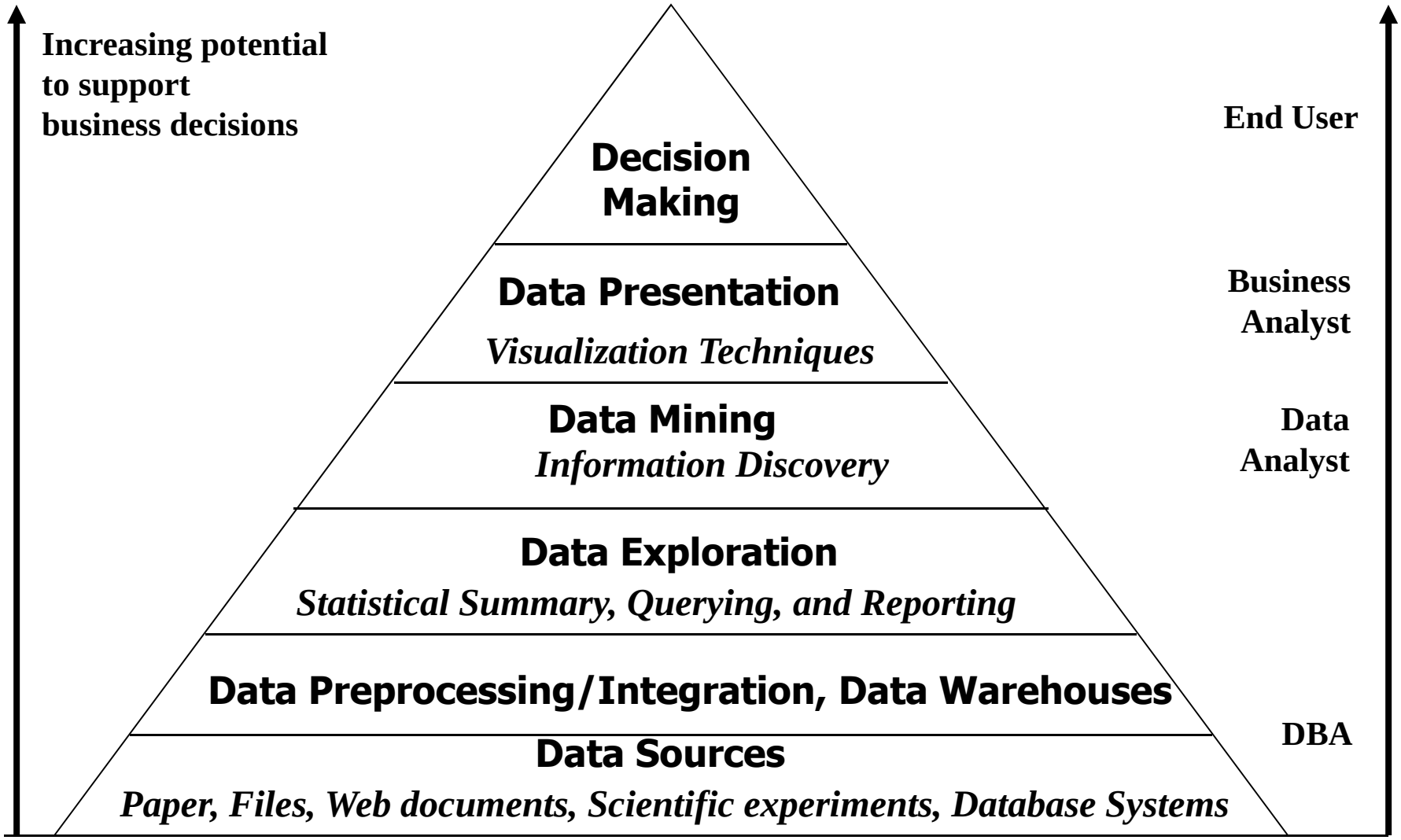
- Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
- Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included.
- Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

- Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

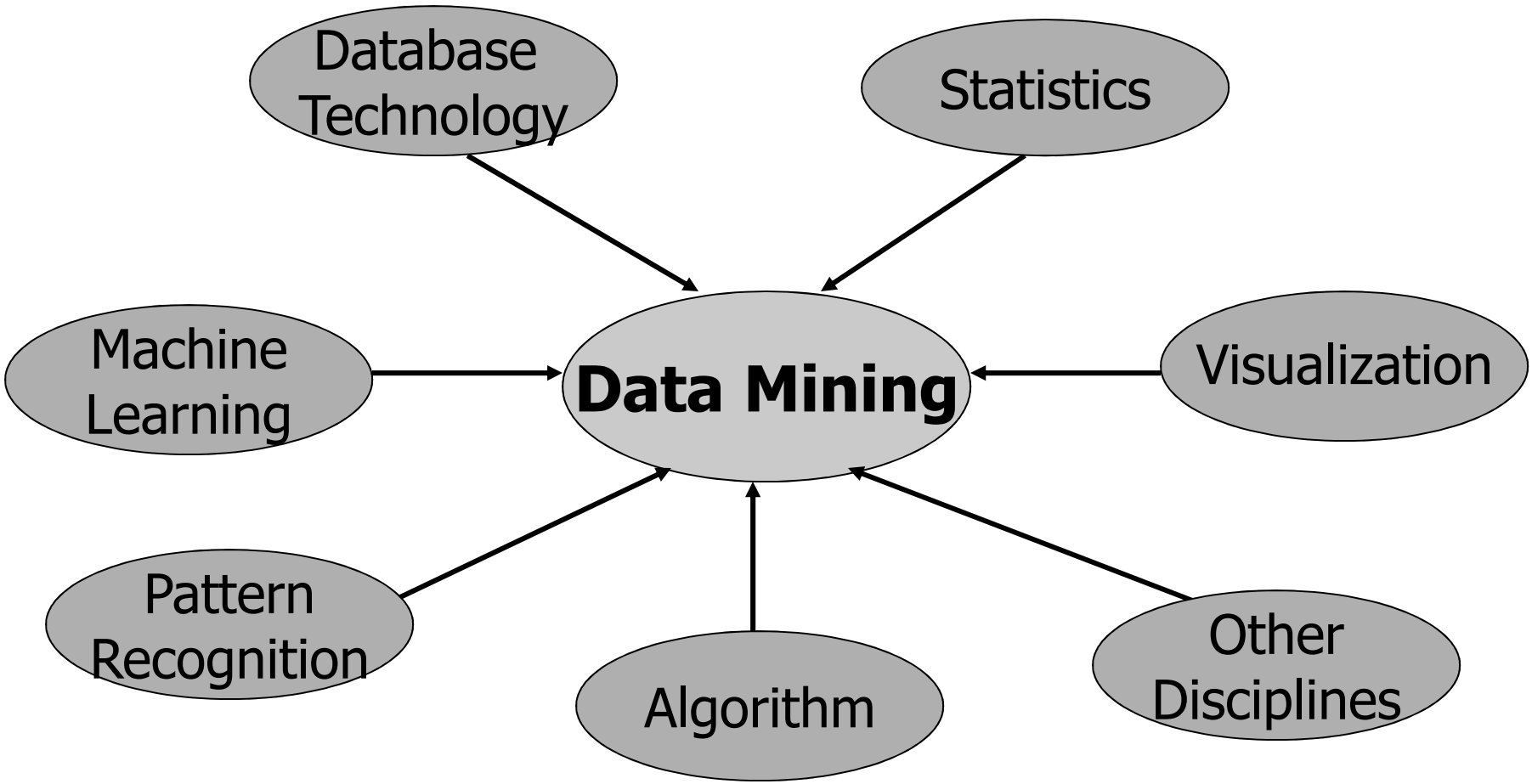
- Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns.
- It may use interestingness thresholds to filter out discovered patterns.
- Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push

- User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.
- In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data Mining and Business Intelligence



Data Mining: Confluence of Multiple Disciplines



Why Not Traditional Data Analysis?

- Tremendous amount of data
- High-dimensionality of data
- High complexity of data
- New and sophisticated applications

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Data view: Kinds of data to be mined
  - Knowledge view: Kinds of knowledge to be discovered
  - Method view: Kinds of techniques utilized
  - Application view: Kinds of applications adapted

Data Mining: On What Kinds of Data?

a. Database-oriented data sets and applications

- Relational database
- Data warehouse
- Transactional database

Relational Databases

- A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- 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 a large set of tuples (records or rows
- A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases.
- An ER data model represents the database as a set of entities and their relationships.

- A relational database for *AllElectronics*. The *AllElectronics* company is described by the following relation tables: *customer, item, employee*, and *branch*.
- Relational data can be accessed by database queries written in a relational query language, such as SQL, or with the assistance of graphical user interfaces.

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

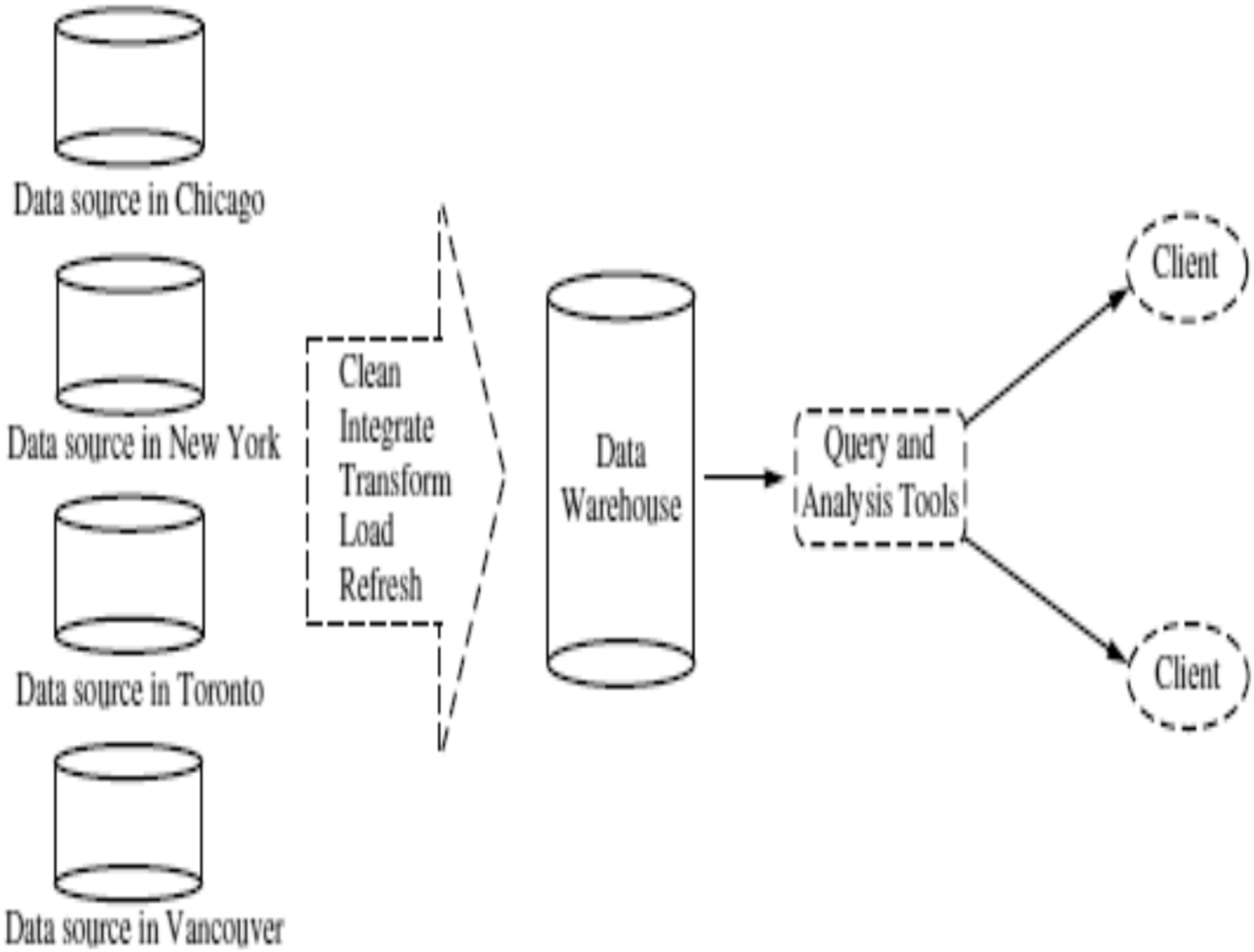
works_at

empl_ID	*branch_ID*
E55	B1
...	...

- A query allows retrieval of specified subsets of the data.
- Suppose that your job is to analyze the *AllElectronics* data.
- Through the use of relational queries, you can ask things like "Show me a list of all items that were sold in the last quarter."
- Relational languages also include aggregate functions such as sum, avg (average), count, max (maximum), and min (minimum).

DataWarehouses

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

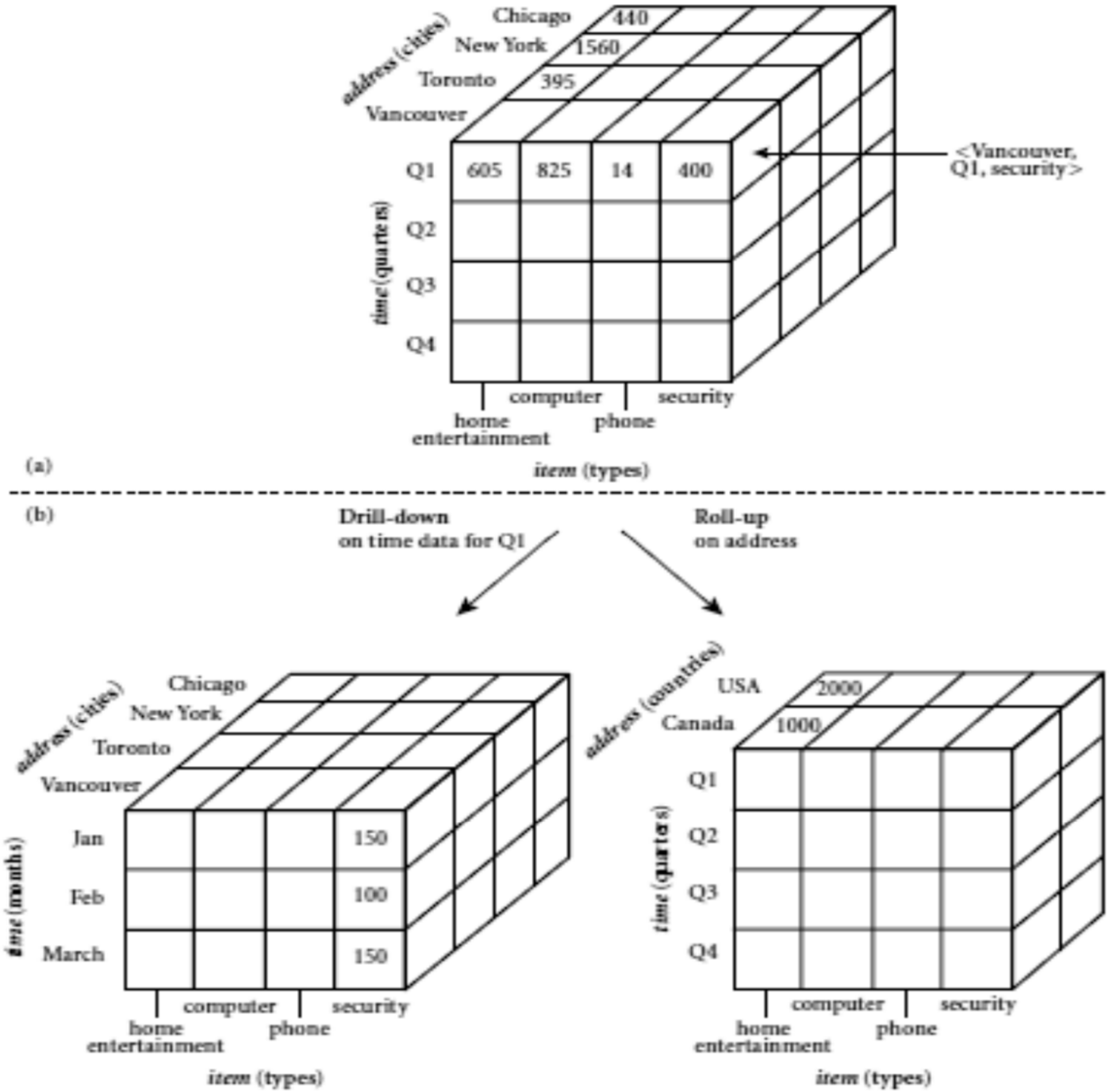

Data source in Chicago
Data source in New York
Data source in Toronto
Data source in Vancouver
Clean
Integrate
Transform
Load
Refresh
Data
Warehouse
Query and
Analysis Tools
Client
Client

- To facilitate decision making, the data in a data warehouse are *organized around major subjects*, such as customer, item, supplier, and activity.
- The data are stored to provide information from a *historical perspective* (such as from 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.
- Each cell stores the value of some aggregate measure, such as *count* or *sales amount*.
- The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube.
- A data cube provides a multidimensional view of data and allows the pre computation and fast accessing of summarized data.

A data cube for *AllElectronics*

- A data cube for summarized sales data of *AllElectronics*
- The cube has three dimensions: *address* (with city values*Chicago, New York, Toronto, Vancouver*), *time* (with quarter values *Q1, Q2, Q3, Q4*), and
- *item*(with itemtype values *home entertainment, computer, phone, security*).
- The aggregate value stored in each cell of the cube is *sales amount* (in thousands).
- For example, the total sales forthefirstquarter,*Q1*, for items relating to security systems in Vancouver is $400,000, as stored in cell (*Vancouver, Q1, security)*


Chicago
New York
Toronto
Vancouver
address (cities)
440
1560
395
Q1
Q2
Q3
Q4
time (quarters)
605
825
14
400
<Vancouver, Q1, security>
computer
security
home entertainment
phone
item (types)
(a)
(b)
Drill-down on time data for Q1
Roll-up on address
Chicago
New York
Toronto
Vancouver
address (cities)
Jan
Feb
March
time (months)
150
100
150
computer
security
home entertainment
phone
item (types)
USA
Canada
address (countries)
2000
1000
Q1
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)

- Data warehouse systems are well suited for on-line analytical processing, or OLAP.
- Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization,
- we can drill down on sales data summarized by quarter to see the data summarized by month. Similarly, we can roll up on sales data summarized by city to view the data summarized by country.

What is the difference between a datawarehouse and a data mart?"

- A data warehouse collects information about subjects that span an *entire organization*, and thus its scope is *enterprise-wide*.
- A data mart, on the other hand, is a department subset of a data warehouse. It focuses on selected subjects, and thus its scope is *department-wide*.

Transactional Databases

- A transactional database consists of a file where each record represents a transaction.
- A transaction typically includes a unique transaction identity number (*trans ID*) and a list of the items making up the transaction (such as items purchased in a store).

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Fragment of a transactional database for sales at *AllElectronics*.

- "Show me all the items purchased by Sandy Smith" or "How many transactions include item number I3?"
- Answering such queries may require a scan of the entire transactional database.
- "Which items sold well together?"
- market basket data analysis would enable you to bundle groups of items together as a strategy for maximizing sales.
- data mining systems for transactional data can do so by identifying frequent itemsets.

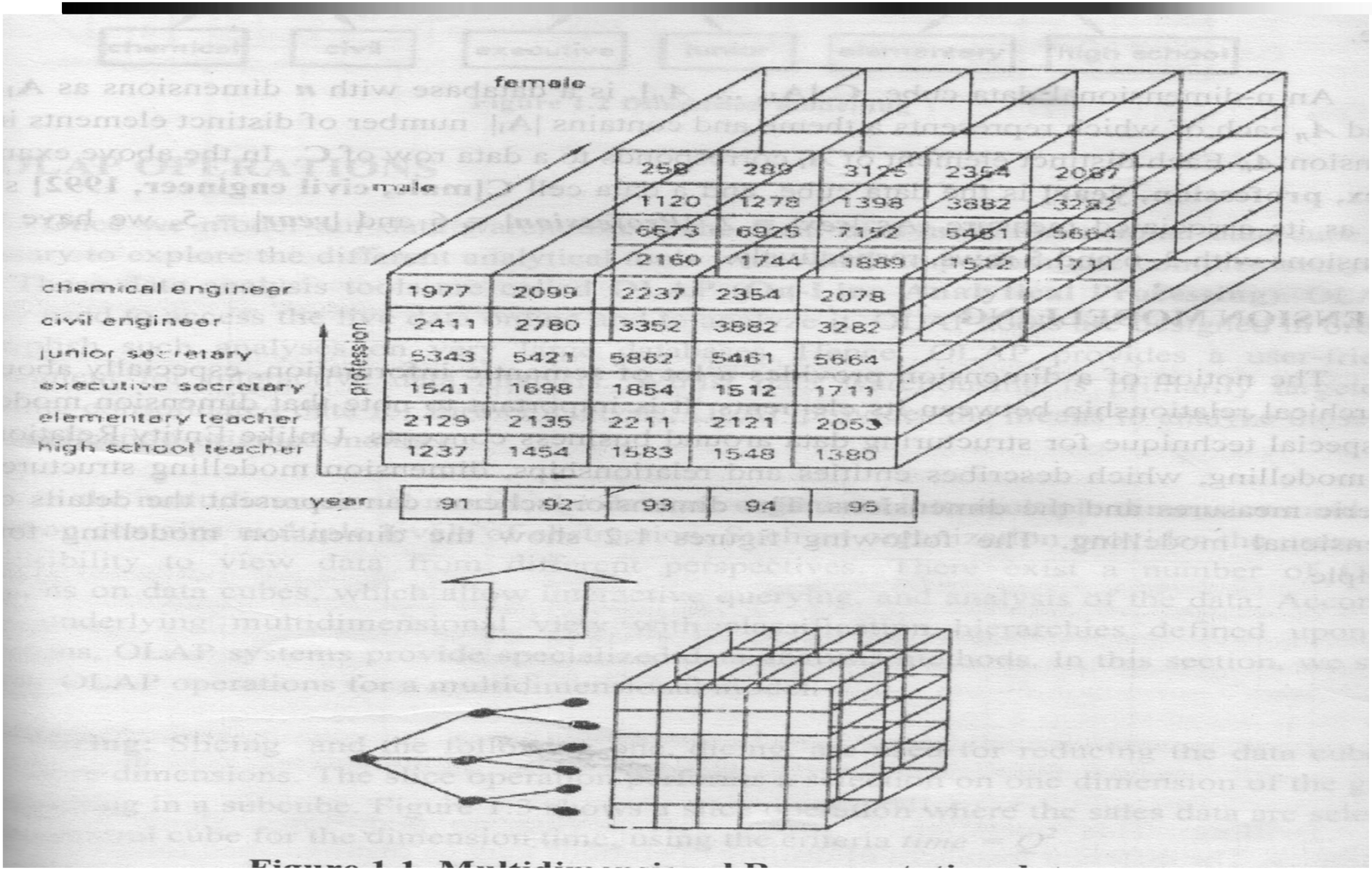
b. Advanced data sets and advanced applications

- Data streams and sensor data
- Time-series data, temporal data, sequence data
- Structure data, graphs, social networks and multi-linked data
- Object-relational databases
- Heterogeneous databases and legacy databases
- Spatial data and spatiotemporal data
- Multimedia database
- Text databases
- The World-Wide Web

Two dimensional representation-table

			Professional Class					
			Engineer		Secretary		Teaching	
			PROFESSON		PROFESSON		PROFESSON	
			Chemical Engineers	Civil Engineers	Junior Secretary	Executive Secretary	Elementary Teachers	High School Teachers
SEX	MALE	91	1977	2411	5343	1541	2129	1237
		92	2099	2780	5421	1698	2135	1457
		93	2237	3352	5862	1854	2211	1583
		94	2354	3882	5461	1512	2121	1548
		95	2078	3282	5664	1711	2053	1380
	FEMALE	91	258	1120	6673	1623	2160	2751
		92	289	1276	6925	1744	2175	1993
		93	312	1398	7152	1889	2189	3125
		94	581	1216	6543	1534	2857	2387
		95	329	1321	6129	1567	2453	3287

Data cube


female
male
profession
chemical engineer
civil engineer
junior secretary
executive secretary
elementary teacher
high school teacher
year
91
92
93
94
95
1977 2099 2237 2354 2078
2411 2780 3352 3882 3282
5343 5421 5862 5461 5664
1541 1698 1854 1512 1711
2129 2135 2211 2121 2053
1237 1454 1583 1548 1380

Dimension Modelling.

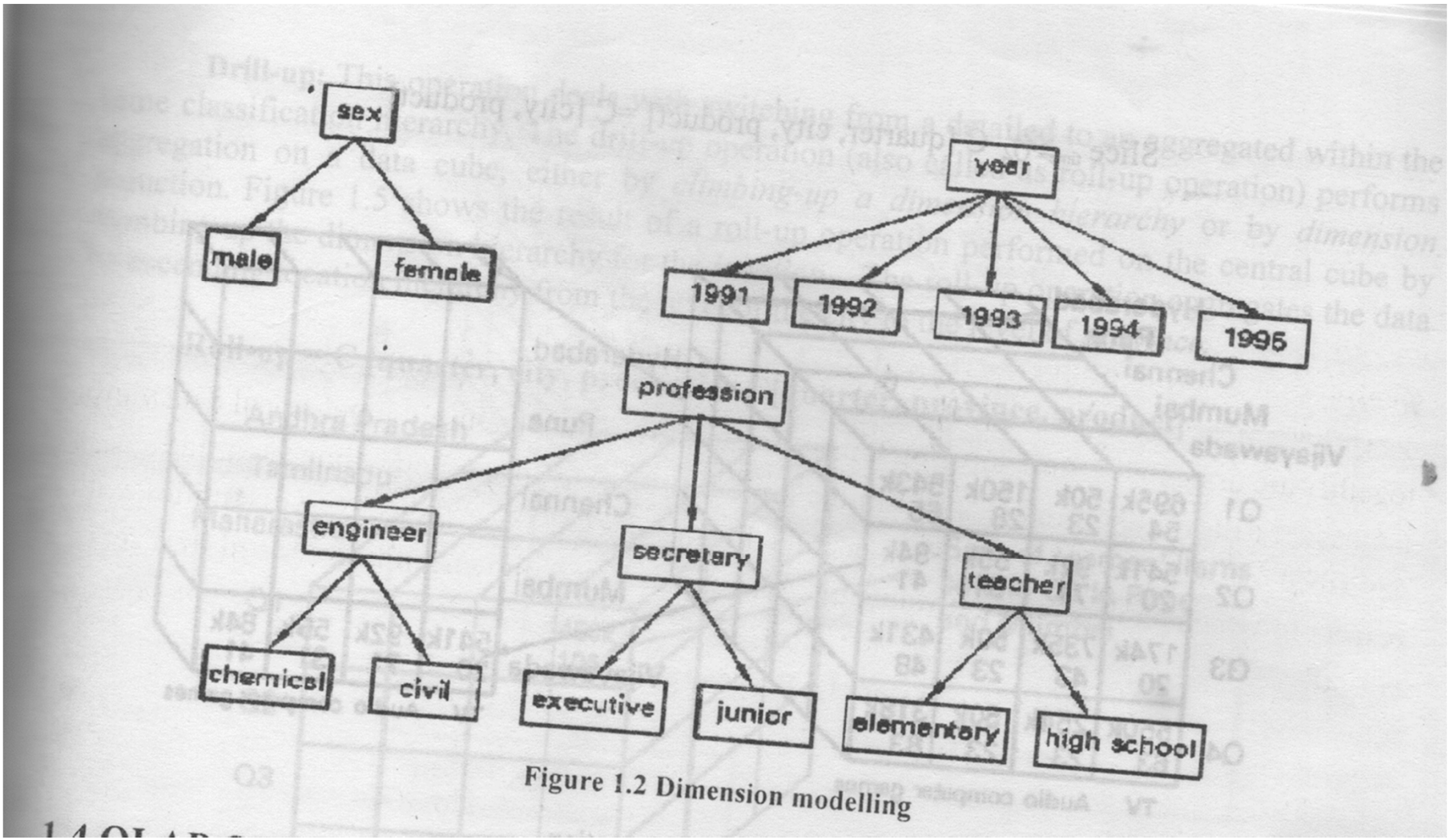


Figure 1.2 Dimension modelling

Data Mining Functionalities

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- In general, data mining tasks can be classified into two categories: descriptive and predictive.
- Descriptive mining tasks characterize the general properties of the data in the database.
- Predictive mining tasks perform inference on the current data in order to make predictions.

Data Mining Functionalities

- **Concept/Class Description: Characterization and Discrimination**
- Data can be associated with classes or concepts.
- For example, in the *AllElectronics* store, classes of items for sale include *computers* and *printers*, and concepts of customers include
- *bigSpenders* and *budgetSpenders*.
- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called **class/concept descriptions**.

- These descriptions can be derived via
- (1) *data characterization*, by summarizing the data of the class under study in general terms, or (2) *data discrimination*, by comparison of the target class with one or a set of comparative classes or (3) both data characterization and discrimination.
- Data characterization is a summarization of the general characteristics or features of a target class of data.

- For example,
- **Data characterization.** summarizing the characteristics of customers who spend more than $1,000 a year at *AllElectronics*.
- **Data discrimination .** compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period.

Data Mining Functionalities

- ii. Frequent patterns, association, correlation
- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- frequent patterns, including item sets, subsequences, and substructures.
- A frequent itemset :milk and bread
- (frequent) sequential pattern: customers tend to purchase first a PC, followed by a digital camera, and then a memory card
- (frequent)structured pattern: substructure occurs frequently
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Association analysis

buys(*X*; "*computer*"))*buys*(*X*; "*software*") [*support* = 1%; *confidence* = 50%]

- A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.
- A 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together.
- Association rules that contain a single predicate are referred to as single-dimensional association rules.

Association analysis

age(*X*, "20:::29")^*income*(*X*, "20K:::29K")) *buys*(*X*, "*CD player*")

[*support* = 2%, *confidence* = 60%]

multidimensional association rule

- association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.

Data Mining Functionalities

iii. Classification and prediction

- Construct models (functions) that describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on (climate), or classify cars based on (mileage)
- Predict some unknown or missing numerical values

(a)

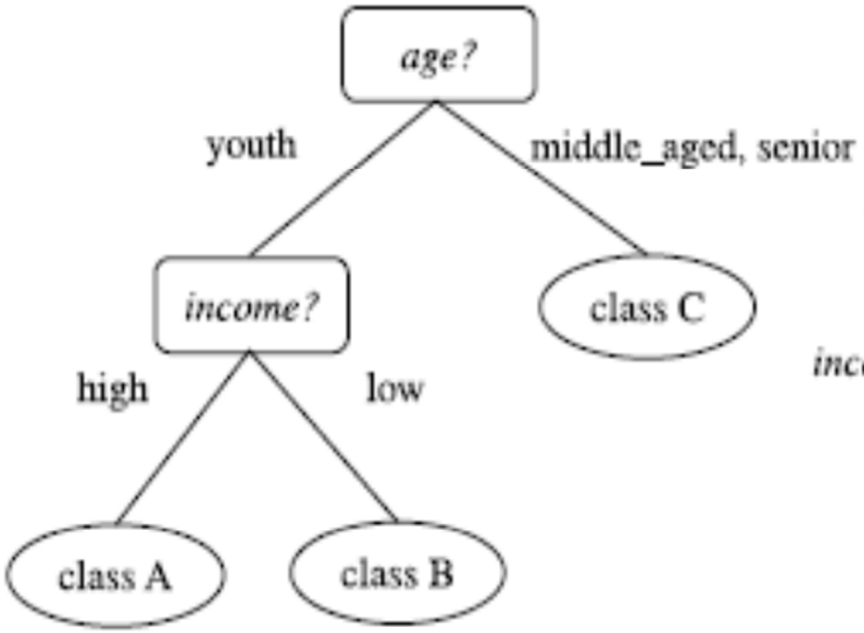
age(X, "youth") AND income(X, "high") ⟶ class(X, "A")

age(X, "youth") AND income(X, "low") ⟶ class(X, "B")

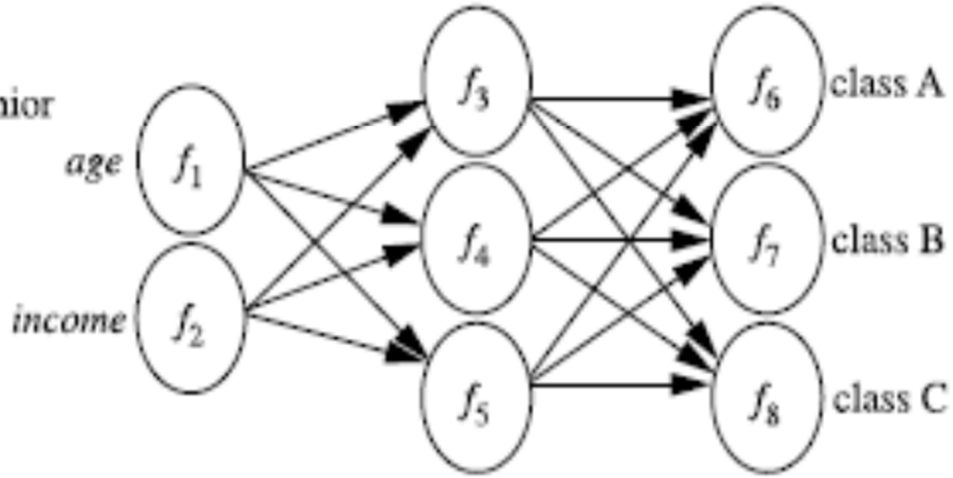
age(X, "middle_aged") ⟶ class(X, "C")

age(X, "senior") ⟶ class(X, "C")

(b)



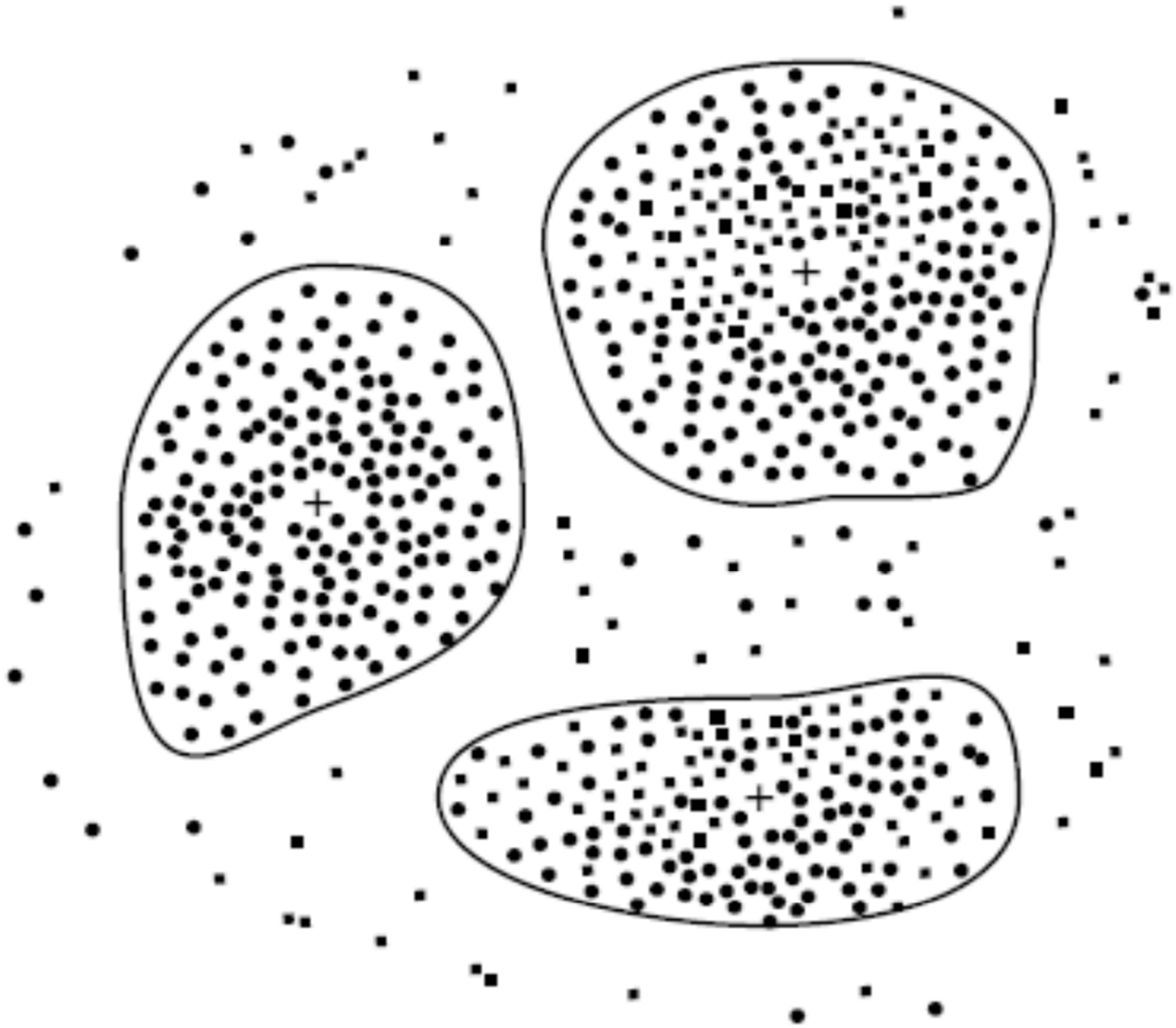
(c)



) A classification model can be represented in various forms, such as (a) IF-THEN rules, (b) a decision tree, or a (c) neural network.

iv. Cluster analysis

- e.g., cluster houses to find distribution patterns
- Maximizing intra-class similarity & minimizing interclass similarity



11 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster "center" is marked with a "+".

v. Outlier analysis

- Outlier: Data object that does not comply with the general behavior of the data

vi. Trend and evolution analysis

- Trend and deviation: e.g., regression analysis (eg: height & wait , price & demand)
- Sequential pattern mining: e.g., digital camera → large SD memory
- Periodicity analysis
- Similarity-based analysis

vii. statistical analyses

Integration of a Data Mining System with a Database or Data Warehouse System

- Integration scheme includes
  - No coupling
  - Loose coupling
  - Semi tight coupling
  - Tight Coupling

No coupling

- No coupling means that a DM system will not utilize any function of a DB or DW system.
- Advantages
  - Simple to use
- Drawbacks
  - Without using a DB/DW system, a DM system may spend a substantial amount of time finding, collecting, cleaning, and transforming data
  - DM system will need to use other tools to extract data, making it difficult to integrate such a system into an information processing environment.
  - Thus, no coupling represents a poor design.

Loose coupling

- Loose coupling means that a DM system will use some facilities of a DB or DW system.
- Fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data warehouse.

Advantages

- Loose coupling is better than no coupling.
- it can fetch any portion of data stored in databases or data warehouses by using query processing.
- It incurs some advantages of the flexibility, efficiency, and other features provided by such systems

Disadvantages

- many loosely coupled mining systems are main memory-based.
- it is difficult for loose coupling to achieve high scalability and good performance with large data sets

Semitight coupling

- *Semitight coupling* means that besides linking a DM system to a DB/DW system, efficient implementations of a few essential data mining primitives can be provided in the DB/DW system.

- These primitives can include sorting, indexing, aggregation, histogram analysis, multiway join, and pre computation of some essential statistical measures, such as sum, count, max, min, standard deviation, and so on.

- this design will enhance the performance of a DM system.

Tight coupling

- *Tight coupling* means that a DM system is smoothly integrated into the DB/DW system.
- The data mining subsystem is treated as one functional component of an information system.
- This approach is highly desirable because it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment.

Major Issues in Data Mining

- Mining different kinds of knowledge in databases.
- Interactive mining of knowledge at multiple levels of abstraction.
- Incorporation of background knowledge.
- Data mining query languages and ad hoc data mining.
- Presentation and visualization of data mining results.
- Handling noisy or incomplete data
- Pattern evaluation-the interestingness problem

Major Issues in Data Mining

- Performance Issues
  - Efficiency and scalability of data mining algorithms
  - Parallel , distributed and incremental mining algorithms
- Issues relating to the diversity of database types
  - Handling of relational and complex types of data.
  - Mining information from heterogeneous database and global information systems

Data warehouse

Definition

Data Warehouse
A collection of corporate information, derived directly from operational systems and some external data sources.
Its specific purpose is to support business decisions, not business operations.

The Purpose of Data Warehousing

- **Realize the value of data**
  - Data / information is an asset
  - Methods to realize the value, (Reporting, Analysis, etc.)
- **Make better decisions**
  - Turn data into information
  - Create competitive advantage
  - Methods to support the decision making process(DSS)

- A data ware house refers to database that is maintained separately from an organization’s operational database.
- It allows integration of variety of application systems. It gives the opportunity for historical data analysis.

- **Subject oriented**: A data ware house is subject oriented rather than Transaction oriented. For eg: customer, product, sales etc.
- **Non volatile**: data ware house is Physically separate ,So it not perishable.

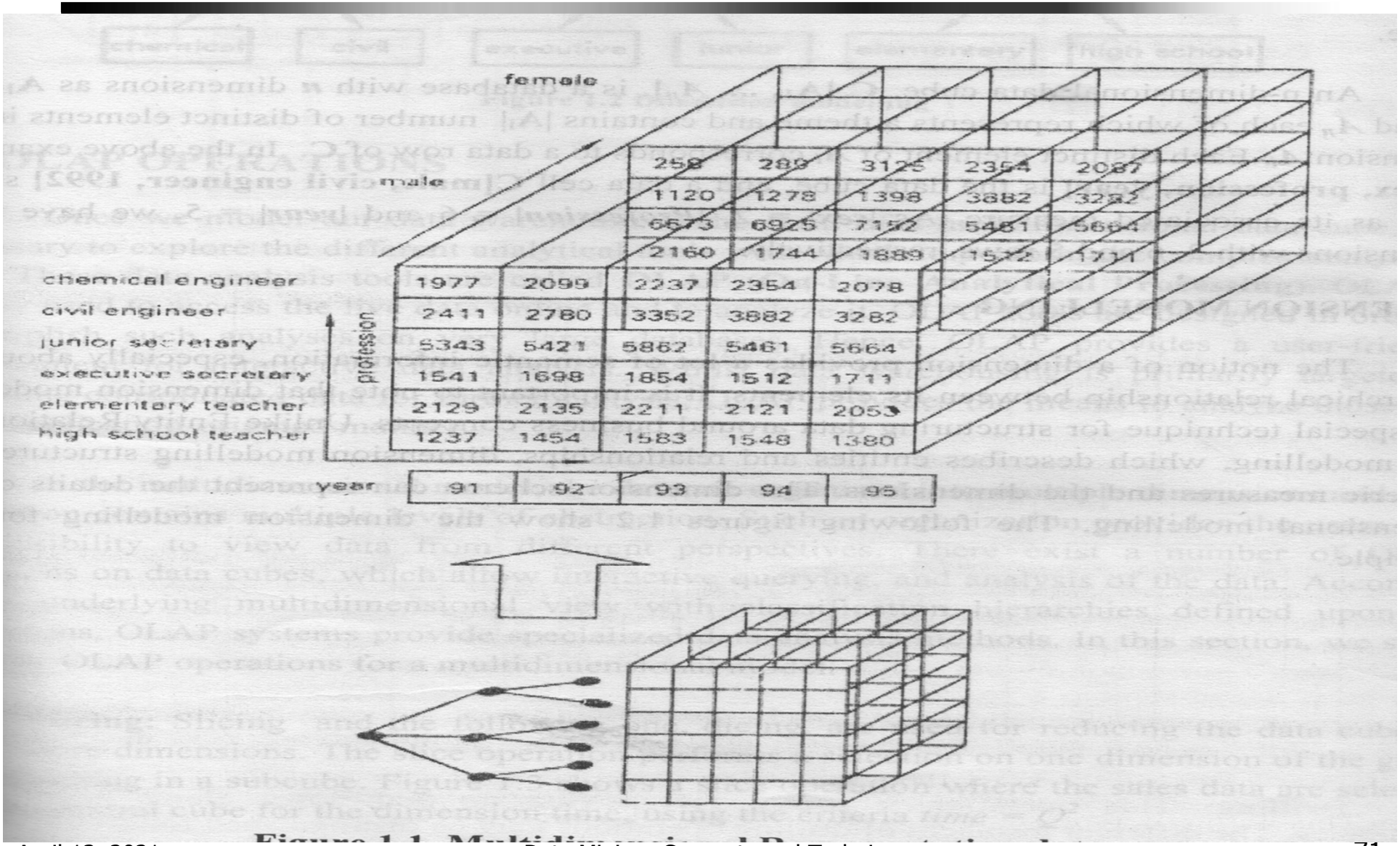
- **Integrated:** A data ware house is constructed by integrating heterogeneous sources such as rdbms, files,OLTPfiles.
- data cleaning and data integration techniques are used for ensuring consistency in naming conventions, encoding structures, attribute measures and so on.
- **Time variant**: Data is stored in historical perspective (5-10years).
- Implicit or explicit time variant will be there in constructing data ware house.

MULTI DIMENSIONAL DATA MODEL

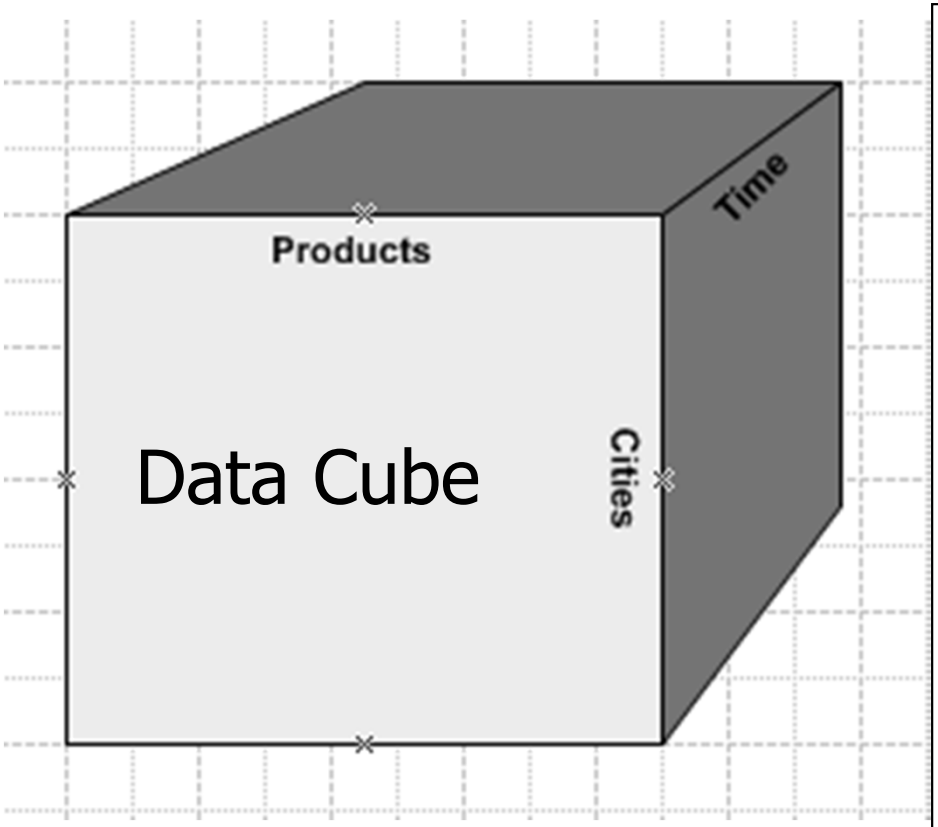
- **Two dimensional representation-table**

			Professional Class					
			Engineer		Secretary		Teaching	
			PROFESSON		PROFESSON		PROFESSON	
			Chemical Engineers	Civil Engineers	Junior Secretary	Executive Secretary	Elementary Teachers	High School Teachers
SEX	MALE	91	1977	2411	5343	1541	2129	1237
		92	2099	2780	5421	1698	2135	1457
		93	2237	3352	5862	1854	2211	1583
		94	2354	3882	5461	1512	2121	1548
		95	2078	3282	5664	1711	2053	1380
	FEMALE	91	258	1120	6673	1623	2160	2751
		92	289	1276	6925	1744	2175	1993
		93	312	1398	7152	1889	2189	3125
		94	581	1216	6543	1534	2857	2387
		95	329	1321	6129	1567	2453	3287

Data cube


female
male
chemical engineer
civil engineer
junior secretary
executive secretary
elementary teacher
high school teacher
profession
year
91
92
93
94
95

Data Ware house implementation



- A multi-dimensional structure called the data cube.
- It is a data abstraction that allows one to view aggregated data from a number of perspectives.

OLAP Operations.

- OLAP Operations are used for retrieving data in a simplified manner from data cube for analysis.
- **Slicing:** Reducing the data cube by one or more dimensions.

  Slice $_{time='Q2'}$ C[quarter,city,product]=C[city, product]

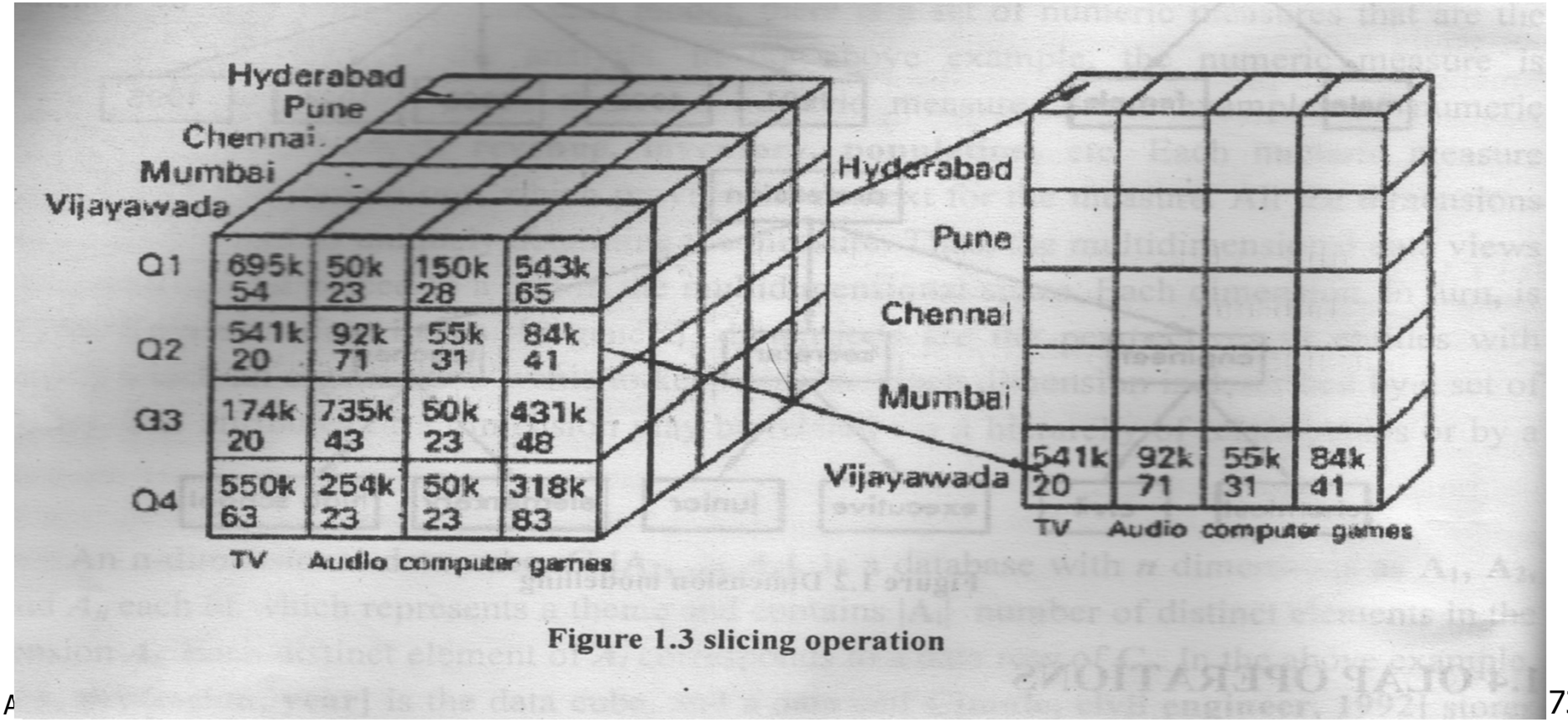


Figure 1.3 slicing operation

- **Dicing:**
- This operation is for selecting a smaller data cube and analyzing it from different perspectives (selection criteria).

- Eg:

dice time=’Q1 or Q2 and location =’Mumbai’ or ‘Pune’
C[Quarter,city,product]=C[quarter,city,product]

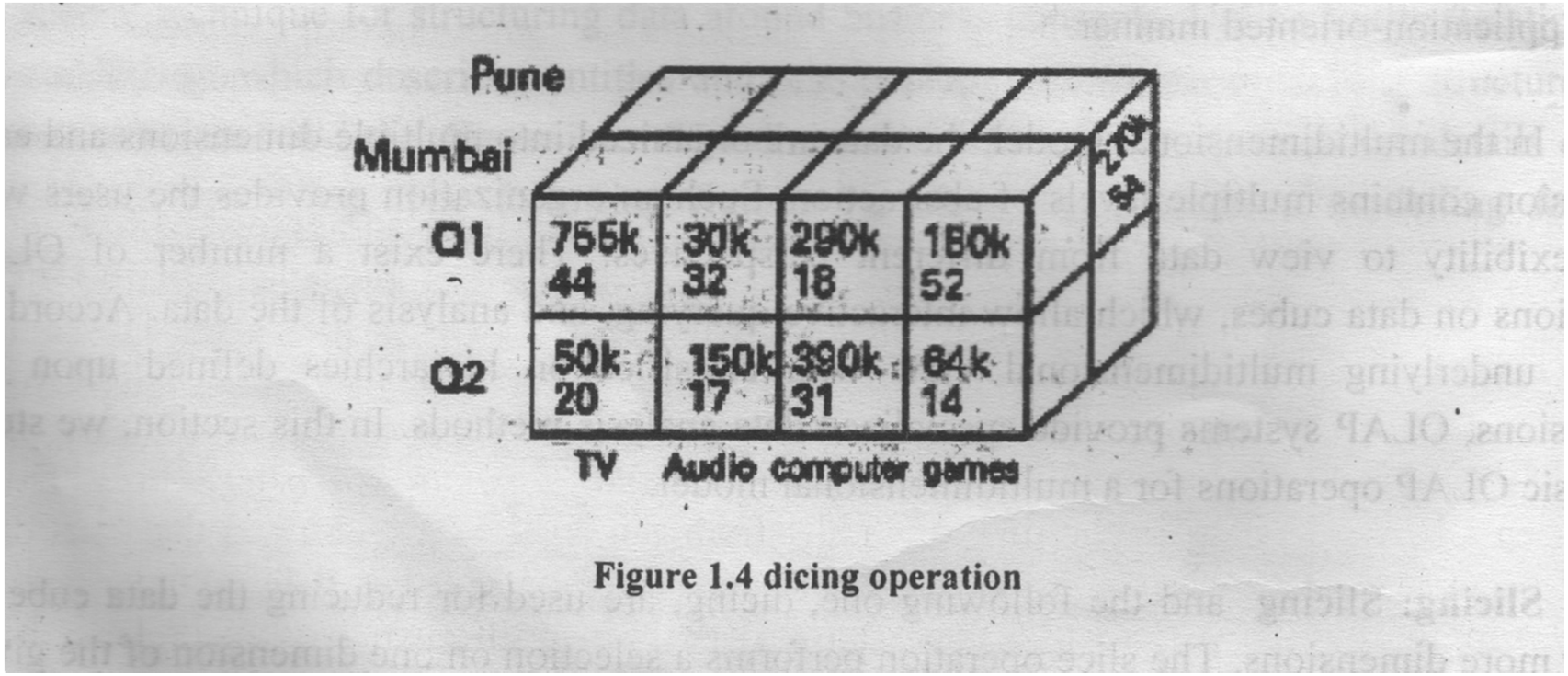


Figure 1.4 dicing operation

- **Drilling:** Moving up and down along classification hierarchies
-
- **Drill up(Roll Up):** This means switching from detailed to an aggregated with in same classification hierarchy.
-
- Eg RollUp=C[quarter,city,product]=C[quarter, province. product]

Roll-up = C [quarter, city, product] = C[quarter, province, product]

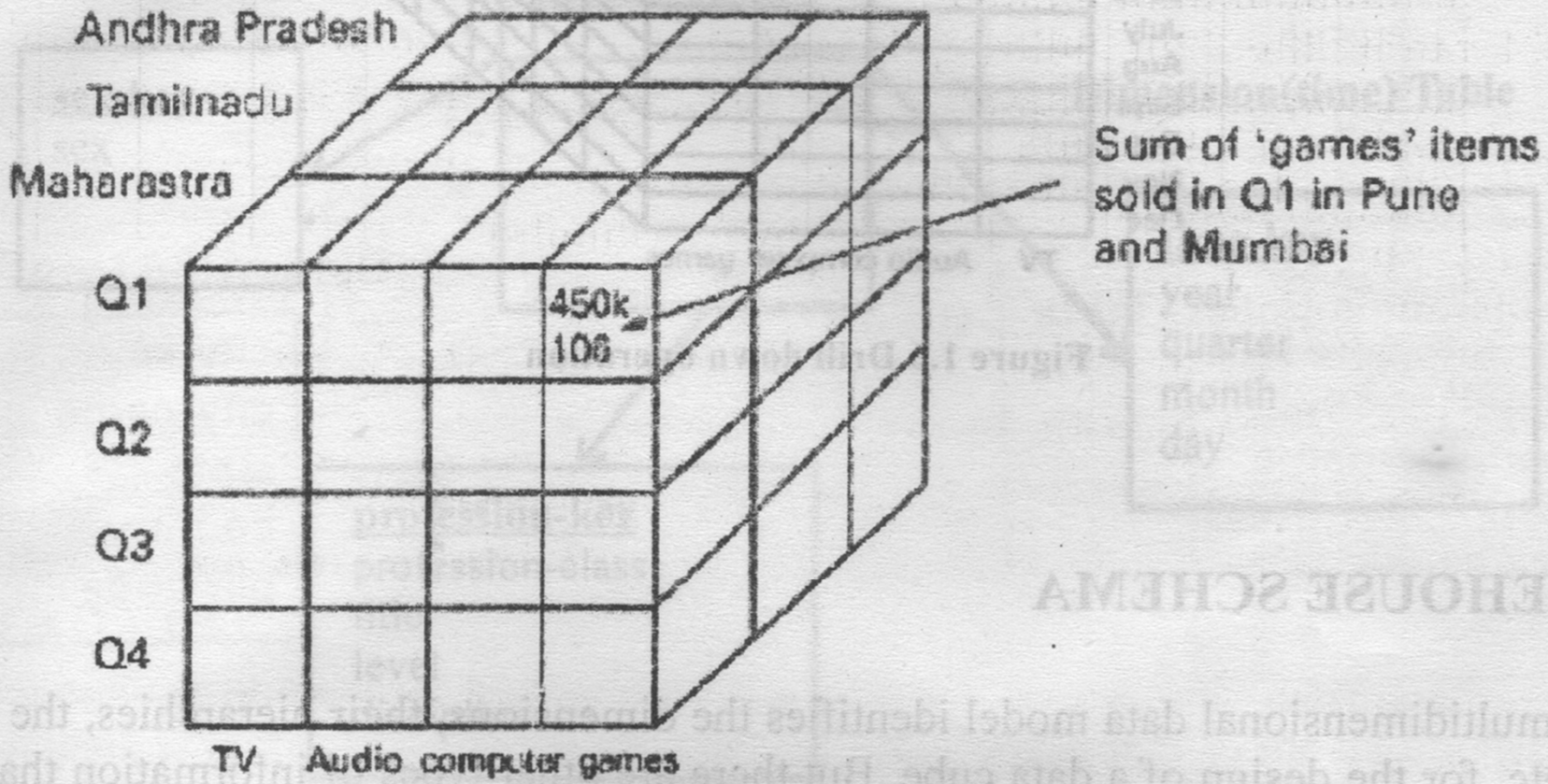


Figure 1.5 roll up operation

<u>Drill Down</u> : This is concerned with switching from aggregated to detailed level.

For eg: day->month->quarter->year,

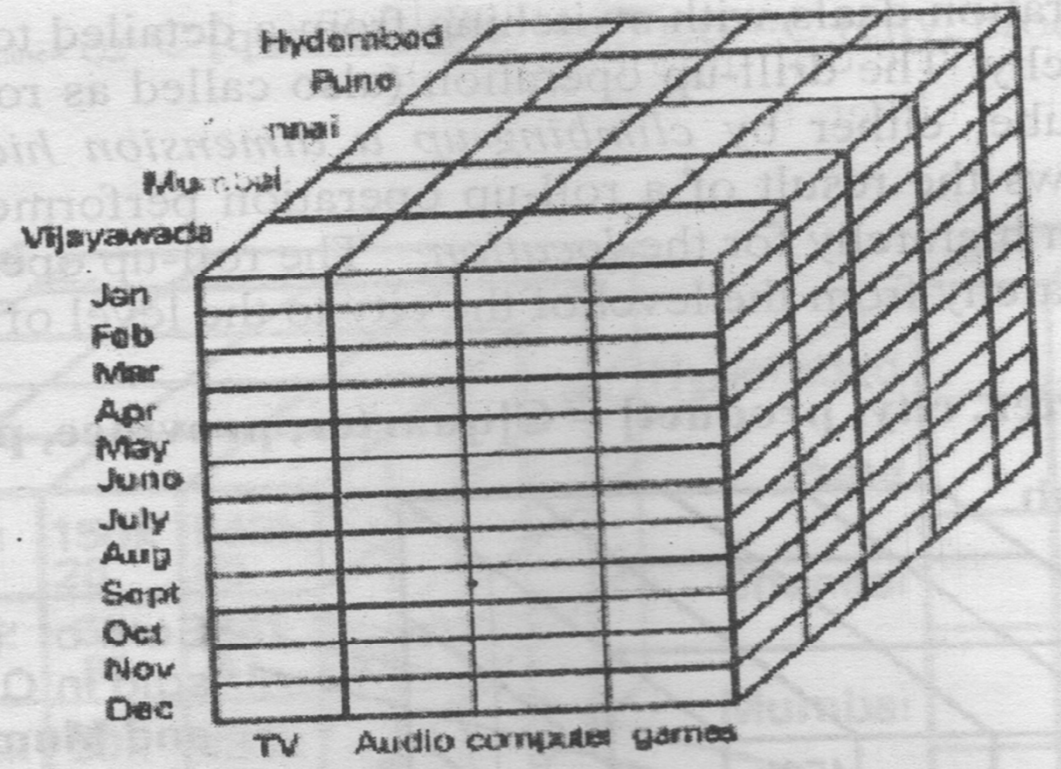


Figure 1.6 Drill down operation

Integration of Data Mining and Data Warehousing:

- **Data warehouse** provides clean, integrated data for fruitful mining.
- **Data mining** provides powerful tools for analysis of data stored in data warehouses.
- **Data mining provides more analysis tools**, e.g.,
  - **association,**
  - **classification,**
  - **clustering,**
  - **pattern-directed, and**
  - **trend analysis.**

- Data mining: the extraction of hidden predictive information from large DB.
- Data might be one of the most valuable assets of your corporation - but only if you know how to reveal valuable knowledge hidden in raw data.
- Data mining allows you to extract diamonds of knowledge from your historical data and predict outcomes of future situations.

- The actual need of data warehouse is

  - To store heterogeneous data for managerial decision purpose.

  - To store data in various dimensions with in a data warehouse.

  -it is easy to analyze the data and to take decisions.

- A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management and decision-making process.

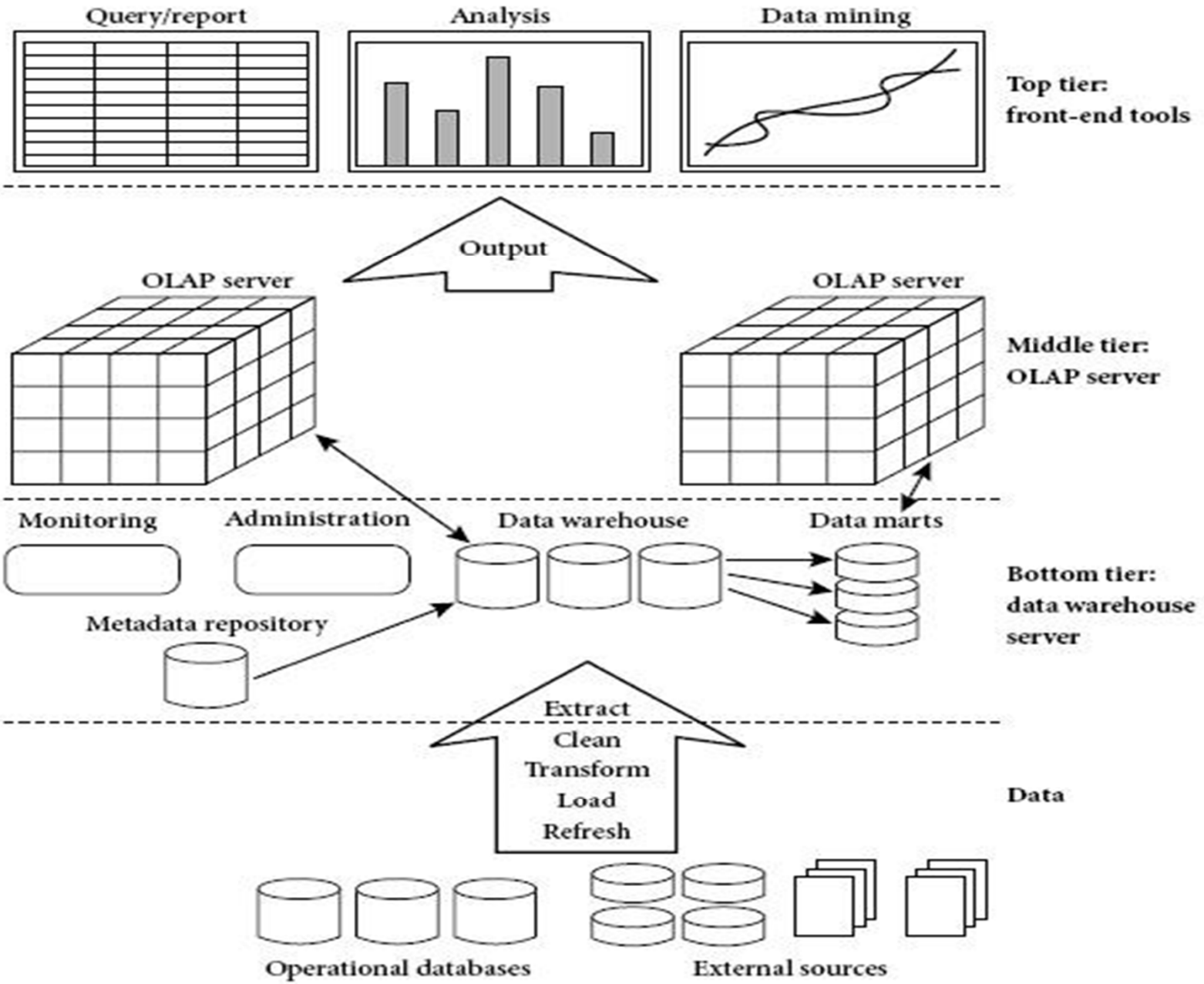
3- Tier Data Warehouse Architecture

3-Tier Data Warehouse Architecture

Data ware house adopt a three tier architecture.

These 3 tiers are:

- Bottom Tier
- Middle Tier
- Top Tier



3.12 A three-tier data warehousing architecture.

Data Sources:

All the data related to any bussiness organization is stored in operational databases, external files and flat files.

- These sources are application oriented

Eg: complete data of organization such as training detail, customer detail, sales, departments, transactions, employee detail etc.

- Data present here in different formats or host format
- Contain data that is not well documented

Bottom Tier: Data warehouse server

Data Warehouse server fetch only relevant information based on data mining (mining a knowledge from large amount of data) request.

Eg: customer profile information provided by external consultants.

- ✓ Data is feed into bottom tier by some backend tools and utilities.

Backend Tools & Utilities:

Functions performed by backend tools and utilities are:

- ✓Data Extraction
- ✓ Data Cleaning
- ✓ Data Transformation
- ✓ Load
- ✓ Refresh

Bottom Tier Contains:

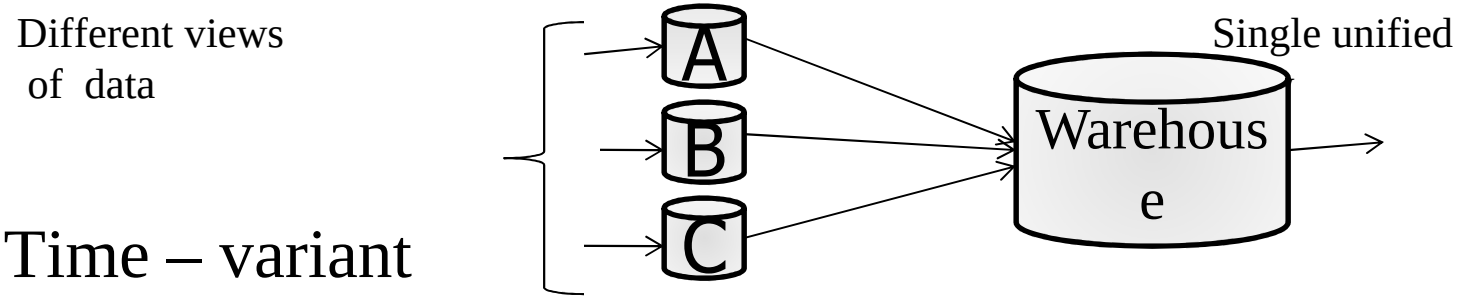
- Data warehouse
- Metadata Repository
- Data Marts
- Monitoring and Administration

Data Warehouse:

It is an optimized form of operational database contain only relevant information and provide fast access to data.

- ✓ Subject oriented

  Eg: Data related to all the departments of an organization
- ✓ Integrated:



- ✓ Time – variant
- ✓ Nonvolatile

Metadata repository:

It figure out that what is available in data warehouse.
It contains:

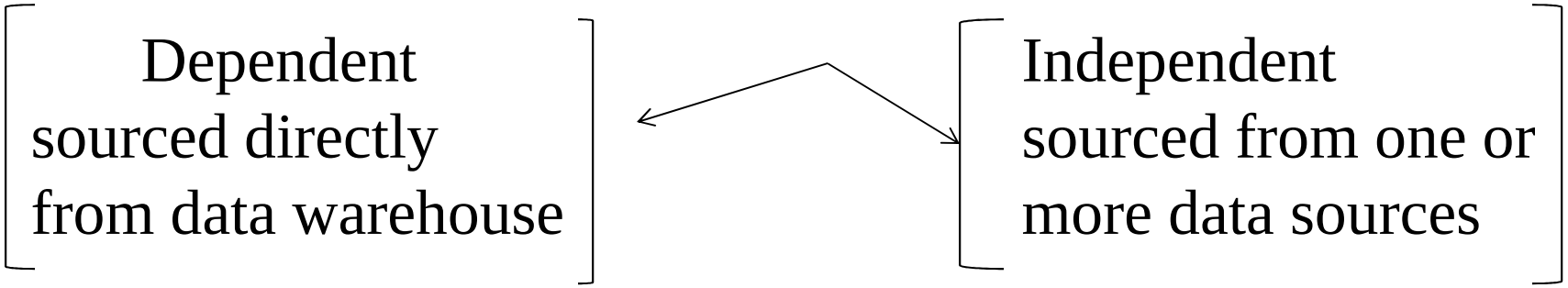
- Structure of data warehouse
- Data names and definitions
- Source of extracted data
- Algorithm used for data cleaning purpose
- Sequence of transformations applied on data
- Data related to system performance

Data Marts:

- Subset of data warehouse contain only small slices of data warehouse

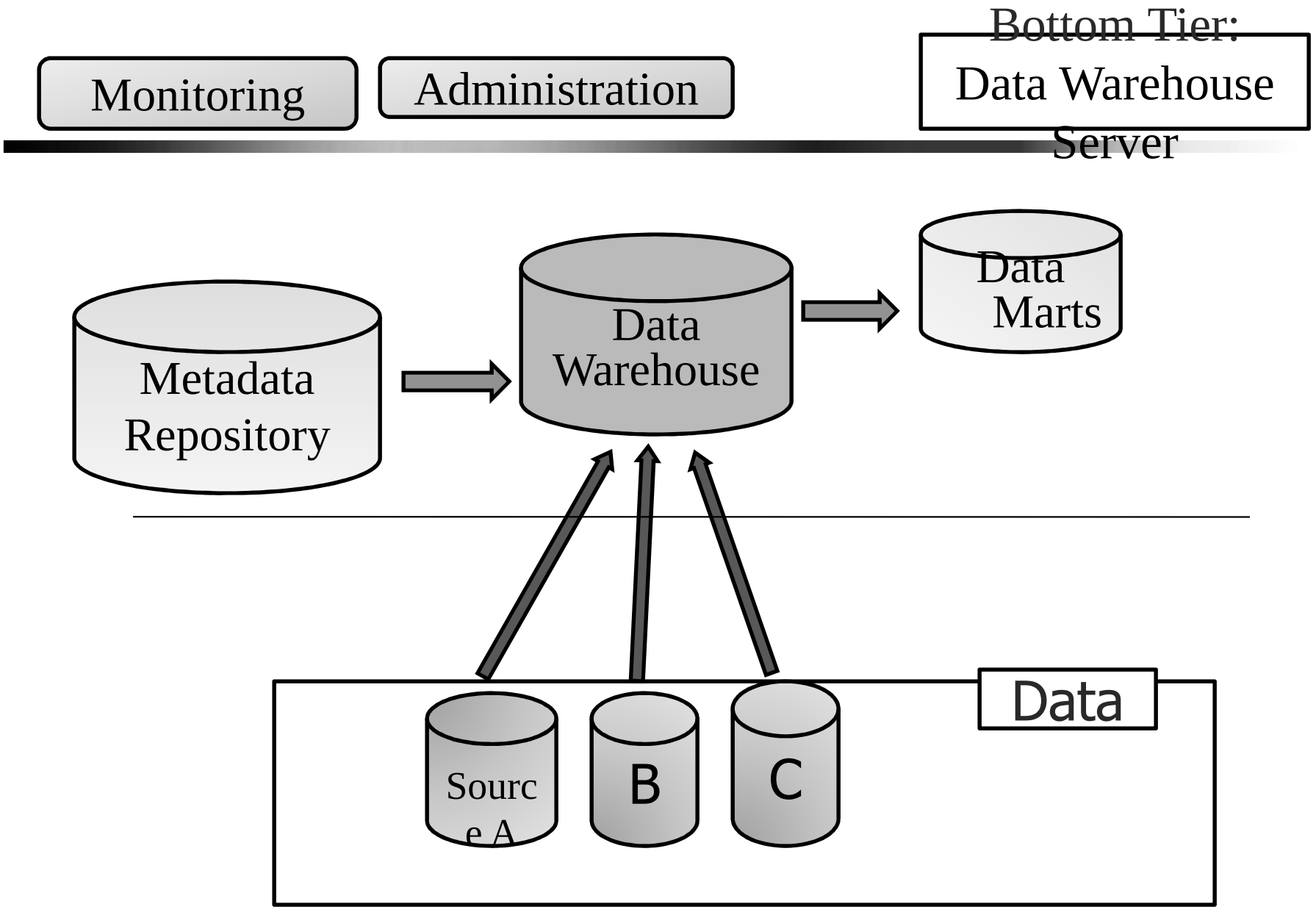
  Eg: Data pertaining to the single department

- Two types of data marts:



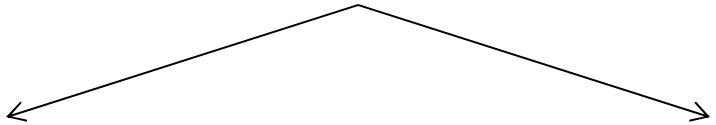
Monitoring & Administration:

- Data Refreshment
- Data source synchronization
- Disaster recovery
- Managing access control and security
- Manage data growth, database performance
- Controlling the number & range of queries
- Limiting the size of data warehouse


Monitoring
Administration
Bottom Tier:
Data Warehouse
Server
Metadata
Repository
Data
Warehouse
Data
Marts
Data
Sourc
e A
B
C

Middle Tier: OLAP Server

- It presents the users a multidimensional data from data warehouse or data marts.
- Typically implemented using two models:



ROLAP Model

Present data in relational tables

MOLAP Model

Present data in array based structures means map directly to data cube array structure.

Top Tier: Front end tools

It is front end client layer.

- Query and reporting tools

Reporting Tools: → Production reporting tools

→ Report writers

Managed query tools: Point and click creation of SQL used in customer mailing list.

- Analysis tools : Prepare charts based on analysis

- Data mining Tools: mining knowledge, discover hidden piece of information, new correlations, useful pattern