GCA (S5) 04A
Exam. Code: DWD

Data Warehousing and Data Mining

SEMESTER V

BACHELOR OF COMPUTER APPLICATION

Block 1

KRISHNA KANTA HANDIQUI STATE OPEN UNIVERSITY

Subject Experts

Prof. Anjana Kakati Mahanta, Gauhati University

Prof. (Retd.) Pranhari Talukdar, Gauhati University

Dr. Jyotiprokash Goswami, Associate Professor, Assam Engineering College

Course Co-ordinators

Dr. Tapashi Kashyap Das, Assistant Professor, KKHSOU

Ms. Sruti Sruba Bharali, Assistant Professor, KKHSOU

SLM Preparation Team

UNITS	CONTRIBUTORS
1	Mr. Atowar Islam, Royal Global University
2	Dr. Swapnanil Gogoi, IDOL, Gauahati University
3 & 4	Mr. Biswajit Das, Cotton University
5	Mr. Dwipen Laskar, Gauhati University
6, 7 & 8	Ms. Daisy Kalita, USTM

Editorial Team

Content : Mr. Gautam Chakrabarty, NERIM

Language : Prof. (Retd.) Robin Goswami, Cotton College

Structure, Format & Graphics: Ms. Sruti Sruba Bharali, KKHSOU

June, 2019 ISBN: 978-93-89123-31-9

Printed and published by Registrar on behalf of the Krishna Kanta Handiqui State Open University.

Headquarters : Patgaon, Rani Gate, Guwahati-781 017

City Office : Housefed Complex, Dispur, Guwahati-781 006; Web: www.kkhsou.in

The university acknowledges with thanks the financial support provided by the **Distance Education Bureau, UGC** for the preparation of this study material.

BACHELOR OF COMPUTER APPLICATION

Data Warehousing and Data Mining

Block-1

CONTENTS

COURSE INTRODUCTION

The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as “knowledge discovery in databases”, the term “data mining” wasn’t coined until the 1990s. Over the last decade, advances in processing power and speed have enabled us to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights. Retailers, banks, manufacturers, telecommunications providers and insurers, among others, are using data mining to discover relationships among everything from pricing, promotions and demographics to how the economy, risk, competition and social media are affecting their business models, revenues, operations and customer relationships.

The course is divided into two blocks:

Block 1 introduces the learners to data mining and data warehousing. Different topics like data preprocessing, multidimensional data are also covered in this block. The data warehouse architecture and representation of data and knowledge along with attribute oriented analysis has also been covered in this block.

Block 2 starts with association rule mining. The block also covers important topics in data mining like classification, prediction, evaluation and clustering in detail. In addition to these, web mining, spatial mining and temporal data mining has also been introduced in this block.

BLOCK INTRODUCTION

This is the first block of the course ***Data Warehousing and Data Mining***. After completing this block, learners will be able to understand the concepts of data mining and data warehousing. Learners will be able to apply the data preprocessing methods and different schema's for multidimensional data and architecture's in different applications.

This block comprises the following eight units:

Unit 1 gives us an introduction to data mining. Classification of data mining and data mining primitives along with major issues in data mining are covered in this unit.

Unit 2 introduces us to data warehousing. The characteristics of data warehousing, its role and structure etc are discussed in this unit.

Unit 3 introduces us to OLAP. The different OLAP operations, differences between OLAP and OLTP are discussed in this unit.

Unit 4 deals with data preprocessing. Different topics like data cleaning, data summarization, data transformation and data reduction are dicussed in this unit.

Unit 5 describes multidimensional data. Multidimensional data models and different schema's like star schema, snowflake schema etc are discussed in this unit.

Unit 6 deals with data warehouse architecture. Topics like the threee tier architecture, indexing and quering in OLAP, OLAM etc and implementation from data warehouse to data mining are discussed in this unit.

Unit 7 deals with data mining knowledge representation. Topics like task relevant data, representation of input and output data along with visualization techniques are described in this unit.

Unit 8 gives us an introduction to attribute oriented analysis. Topics like attribute generalization, attribute relevance, class comparison and different statistical measures are covered in detail in this unit.

Each unit of this block includes some along-side boxes to help you know some of the difficult, unseen terms. Some "EXERCISES" have been included to help you apply your own thoughts. You may find some boxes marked with: "LET US KNOW". These boxes will provide you with some additional interesting and relevant information. Again, you will get "CHECK YOUR PROGRESS" questions. These have been designed for you to self-check your progress of study. It will be helpful for you if you solve the problems put in these boxes immediately after you go through the sections of the units and then match your answers with "ANSWERS TO CHECK YOUR PROGRESS" given at the end of each unit.

UNIT 1: INTRODUCTION TO DATA MINING

UNIT STRUCTURE

1.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define data mining
- describe the different types of data
- describe data mining task primitives
- explain integration of data mining system
- describe the major issues of data mining.

1.2 INTRODUCTION

In this unit, we will learn about data mining. We will also learn about the different types of data like non-dependency oriented data and dependency

oriented data. Besides data mining task primitives and how data mining system can be integrated. In addition to this, the major issues of data mining will be discussed in this unit while in the next unit, we will provide an introduction to data warehousing.

1.3 DATA MINING

Data Mining is a non-trivial process of discovering knowledge from huge amount of data, as mining of gold from rocks or sand is called gold mining, similarly data mining is appropriately named as knowledge mining. To extract the knowledge from large data set knowledge discovery from data (KDD) is used.

Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data. A wide variation exists in terms of the problem domains, applications, formulations, and data representations that are encountered in real applications. Therefore, “data mining” is a broad umbrella term that is used to describe these different aspects of data processing.

In the modern age, virtually all automated systems generate some form of data either for diagnostic or analysis purposes. Some examples of different kinds of data are as follows:

- **World Wide Web:** The number of documents on the indexed Web is now in the order of billions, and the invisible Web is much larger. User accesses to such documents create Web access logs at servers and customer behavior profiles at commercial sites. Furthermore, the linked structure of the Web is referred to as the Web graph, which itself is a kind of data. These different types of data are useful in various applications. For example, the Web documents and link structure can be mined to determine associations between different topics on the Web. On the other hand, user access logs can be mined to determine frequent patterns of accesses or unusual patterns of possibly unwarranted behavior.
- **Financial Interactions:** Most common transactions of everyday life, such as using an automated teller machine (ATM) card or a credit

card, can create data in an automated way. Such transactions can be mined for many useful insights such as fraud or other unusual activity.

- **User Interactions:** Many forms of user interactions create large volumes of data. For example, the use of a telephone typically creates a record at the telecommunication company with details about the duration and destination of the call. Many phone companies routinely analyze such data to determine the relevant patterns of behavior that can be used to make decisions about network capacity, promotions, pricing, or customer targeting.
- **Sensor Technologies and the Internet of Things:** A recent trend is the development of low-cost wearable sensors, smart phones, and other smart devices that can communicate with one another. By one estimate, the number of such devices exceeded the number of people on the planet in 2008. The implications of such massive data collection are significant for mining algorithms.

The deluge of data is a direct result of advances in technology and the computerization of every aspect of modern life. It is, therefore, natural to examine whether one can extract concise and possibly actionable insights from the available data for application-specific goals. This is where the task of data mining comes in. The raw data may be arbitrary, unstructured, or even in a format that is not immediately suitable for automated processing. For example, manually collected data may be drawn from heterogeneous sources in different formats and yet somehow needs to be processed by an automated computer program to gain insights. To address this issue, data mining analysts use a pipeline of processing, where the raw data are collected, cleaned, and transformed into a standardized format. The data may be stored in a commercial database system and finally processed for insights with the use of analytical methods. In fact, while data mining often conjures up the notion of analytical algorithms, the reality is that the vast majority of work is related to the data preparation portion of the process. This pipeline of processing is conceptually similar to that of an actual mining process from a mineral ore to the refined end product. The term “mining” derives its roots from this analogy. From an analytical perspective, data

mining is challenging because of the wide disparity in the problems and data types that are encountered. For example, a commercial product recommendation problem is very different from an intrusion-detection application, even at the level of the input data format or the problem definition. Even within related classes of problems, the differences are quite significant. For example, a product recommendation problem in a multidimensional database is very different from a social recommendation problem due to the differences in the underlying data type. Nevertheless, in spite of these differences, data mining applications are often closely connected to one of four "super problems" in data mining: association pattern mining, clustering, classification, and outlier detection. These problems are so important because they are used as building blocks in a majority of the applications in some indirect form or the other. This is a useful abstraction because it helps us conceptualize and structure the field of data mining more effectively. The data may have different formats or types. The type may be quantitative (e.g., age), categorical (e.g., ethnicity), text, spatial, temporal, or graph-oriented. Although the most common form of data is multidimensional, an increasing proportion belongs to more complex data types. While there is a conceptual portability of algorithms between many data types at a very high level, this is not the case from a practical perspective. The reality is that the precise data type may affect the behavior of a particular algorithm significantly. As a result, one may need to design refined variations of the basic approach for multidimensional data, so that it can be used effectively for a different data type. Therefore, this book will dedicate different chapters to the various data types to provide a better understanding of how the processing methods are affected by the underlying data type.

1.3.1 Various types of Data

One of the interesting aspects of the data mining process is the wide variety of data types that are available for analysis. There are two broad types of data, of varying complexity, for the data mining process: non-dependency-oriented data and dependency-oriented data.

Table 1.1: Example of Multidimensional Data Set

Name	Age	Gender	Race	ZIP Code
Decock S	34	M	Australian	05139
Kohli B	40	M	Indian	10598
Sahni A	35	F	Asian	90201

1.3.1.1 Non-dependency-oriented data

This typically refers to simple data types such as multidimensional data or text data. These data types are the simplest and most commonly encountered. In these cases, the data records do not have any specified dependencies between either the data items or the attributes. An example is a set of demographic records about individuals containing their age, gender, and ZIP code.

Non-dependency-oriented data are the simplest form of data and typically refers to multidimensional data. This data typically contains a set of records. A record is also referred to as a data point, instance, example, transaction, entity, tuple, object, or feature-vector, depending on the application at hand. Each record contains a set of fields, which are also referred to as attributes, dimensions,and features.

The Non-dependency-oriented data are divided into the following categories–

- **Quantitative Multidimensional Data:** The attributes in Table 1.1 are of two different types. The age field has values that are numerical in the sense that they have a natural ordering. Such attributes are referred to as continuous, numeric, or quantitative. Data in which all fields are quantitative is also referred to as quantitative data or numeric data. In the data mining literature, this particular subtype of data is considered the most common. This subtype is particularly convenient for analytical processing because it is much easier to work with quantitative data from a statistical perspective.

- **Categorical and Mixed Attribute Data:** Many data sets in real applications may contain categorical attributes that take on discrete unordered values. For example, in Table 1.1, the attributes such as gender, race, and ZIP code, have discrete values without a natural ordering among them. In the case of mixed attribute data, there is a combination of categorical and numeric attributes. The full data in Table 1.1 are considered mixed-attribute data because they contain both numeric and categorical attributes. The attribute corresponding to gender is special because it is categorical, but with only two possible values. In such cases, it is possible to impose an artificial ordering between these values and use algorithms designed for numeric data for this type. This is referred to as binary data, and it can be considered a special case of either numeric or categorical data.
- **Binary and Set Data:** Binary data can be considered a special case of either multidimensional categorical data or multidimensional quantitative data. It is a special case of multidimensional categorical data, in which each categorical attribute may take on one of at most two discrete values. It is also a special case of multidimensional quantitative data because an ordering exists between the two values.
- **Text Data:** Text data can be viewed either as a string, or as multidimensional data, depending on how they are represented. In its raw form, a text document corresponds to a string. This is a dependency-oriented data type. Each string is a sequence of characters (or words) corresponding to the document. However, text documents are rarely represented as strings.

1.3.1.2 Dependency-oriented data

In this type of data, implicit or explicit relationships may exist between data items. For example, a social network data set contains a set of *vertices* (data items) that are connected together by a set of *edges* (relationships). On the other hand, time series contains implicit dependencies. For example, two successive values collected from a sensor are likely to be related to one another. Therefore, the time attribute implicitly specifies a dependency between successive readings.

The knowledge about preexisting dependencies greatly changes the data mining process because data mining is all about finding relationships between data items. The presence of preexisting dependencies, therefore, changes the expected relationships in the data, and this may be considered interesting from the perspective of these expected relationships. Several types of dependencies may exist that may be either implicit or explicit:

- **Implicit dependencies:** In this case, the dependencies between data items are not explicitly specified but are known to "typically" exist in that domain.
- **Explicit dependencies:** This typically refers to graph or network data in which edges are used to specify explicit relationships. Graphs are a very powerful abstraction that is often used as an intermediate representation to solve data mining problems in the context of other data types.

The different dependency-oriented data types are discussed in detail.

- **Time-Series Data:** Time-series data contain values that are typically generated by continuous measurement over time. For example, an environmental sensor will measure the temperature continuously, whereas an electrocardiogram (ECG) will measure the parameters of a subject's heart rhythm. Such data typically have implicit dependencies

built into the values received over time. For example, the adjacent values recorded by a temperature sensor will usually vary smoothly over time, and this factor needs to be explicitly used in the data mining process.

- **Discrete Sequences and Strings:** Discrete sequences can be considered the categorical analog of time-series data. As in the case of time-series data, the contextual attribute is a time stamp or a position index in the ordering. The behavioral attribute is a categorical value. Therefore, discrete sequence data are defined in a similar way to time-series data.
- **Spatial Data:** In spatial data, many non spatial attributes (e.g., temperature, pressure, image pixel color intensity) are measured at spatial locations. For example, sea-surface temperatures are often collected by meteorologists to forecast the occurrence of hurricanes. In such cases, the spatial coordinates correspond to contextual attributes, whereas attributes such as the temperature correspond to the behavioral attributes. Typically, there are two spatial attributes. As in the case of time-series data, it is also possible to have multiple behavioral attributes. For example, in the sea-surface temperature application, one might also measure other behavioral attributes such as the pressure.
- **Network and Graph Data:** In network and graph data, the data values may correspond to nodes in the network, whereas the relationships among the data values may correspond to the edges in the network. In some cases, attributes may be associated with nodes in the network. Although it is also possible to associate attributes with edges in the network, it is much less common to do so.

1.4 DATA MINING FUNCTIONALITIES

Data mining functionalities and variety of knowledge are presented as follows:

- **Classification:** Classification analysis is the organization of data in given classes or in short classification is to partition the given data into predefined disjoint groups. For example, a bank loan officer want to analyze which loan applicant are appropriate and which can create risk.
- **Clustering:** Data items are grouped according to logical relationship or customer preferences. For example, data can be mined to identify market segment or customer affinities.
- **Outlier Analysis:** Outlier are data elements that cannot grouped in a given class or cluster. Basically outliers are considered as noise and are always removed from applications.
- **Association:** Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*, identifies the frequent item sets. Another threshold, *confidence*, which is the conditional probability that an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.
- **Data Characterization:** Data characterization is a summarization of general features of objects in a target class, and it produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may want to characterize the OurVideoStore customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute-oriented induction method can be used, for example, to carry out data summarization. Note that with a data

cube containing summarization of data, simple OLAP operations fit the purpose of data characterization.

- **Data Discrimination:** Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.
- **Prediction:** Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data.
- **Evolution and Deviation Analysis:** Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

1.5 CLASSIFICATION OF DATA MINING SYSTEM

Data mining systems can be categorized according to various criteria as follows:

- **Classification of data mining systems according to the type of data sources mined:** This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.
- **Classification of data mining systems according to the database involved:** This classification based on the data model involved such as relational database, object oriented database, data warehouse, transactional database, etc.
- **Classification of data mining systems according to the kind of knowledge discovered:** This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.
- **Classification of data mining systems according to mining techniques used:** This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

1.6 DATA MINING TASK PRIMITIVES

Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed. A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to inter- actively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

- **The set of task-relevant data to be mined:** This specifies the portions of the database or the set of data in which the user is interested.

This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).

- **The kind of knowledge to be mined:** This specifies the data mining functions to be per- formed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.
- **The background knowledge to be used in the discovery process:** This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of back- ground knowledge, which allow data to be mined at multiple levels of abstraction. User beliefs regarding relationships in the data are another form of back- ground knowledge.
- **The interestingness measures and thresholds for pattern evaluation:** They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.
- **The expected representation for visualizing the discovered patterns:** This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes. A data mining query language can be designed to incorporate these primitives, allowing users to flexibly interact with data mining systems. Having a data mining query language provides a foundation on which user-friendly graphical interfaces can be built.

1.7 INTEGRATION OF DATA MINING SYSTEM

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme. In this scheme, the main

focus is on data mining design and on developing efficient and effective algorithms for mining the available data sets.

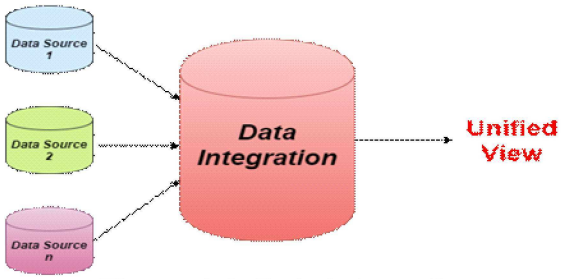


Figure 1.1: Data Integration

The list of integration schemes is as follows:

- **No Coupling:** In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in another file.
- **Loose Coupling:** In this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory managed by these systems and performs data mining on that data. It then stores the mining result either in a file or in a designated place in a database or in a data warehouse.
- **Semi-tight Coupling:** In this scheme, the data mining system is linked with a database or a data warehouse system. In addition to that, efficient implementations of a few data mining primitives can be provided in the database.
- **Tight coupling:** In this coupling scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

1.8 MAJOR ISSUES OF DATA MINING

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. In this section, we will discuss the major issues regarding–

- Mining methodology and user interaction
- Performance issues
- Diverse data types issues

The following diagram describes the major issues.

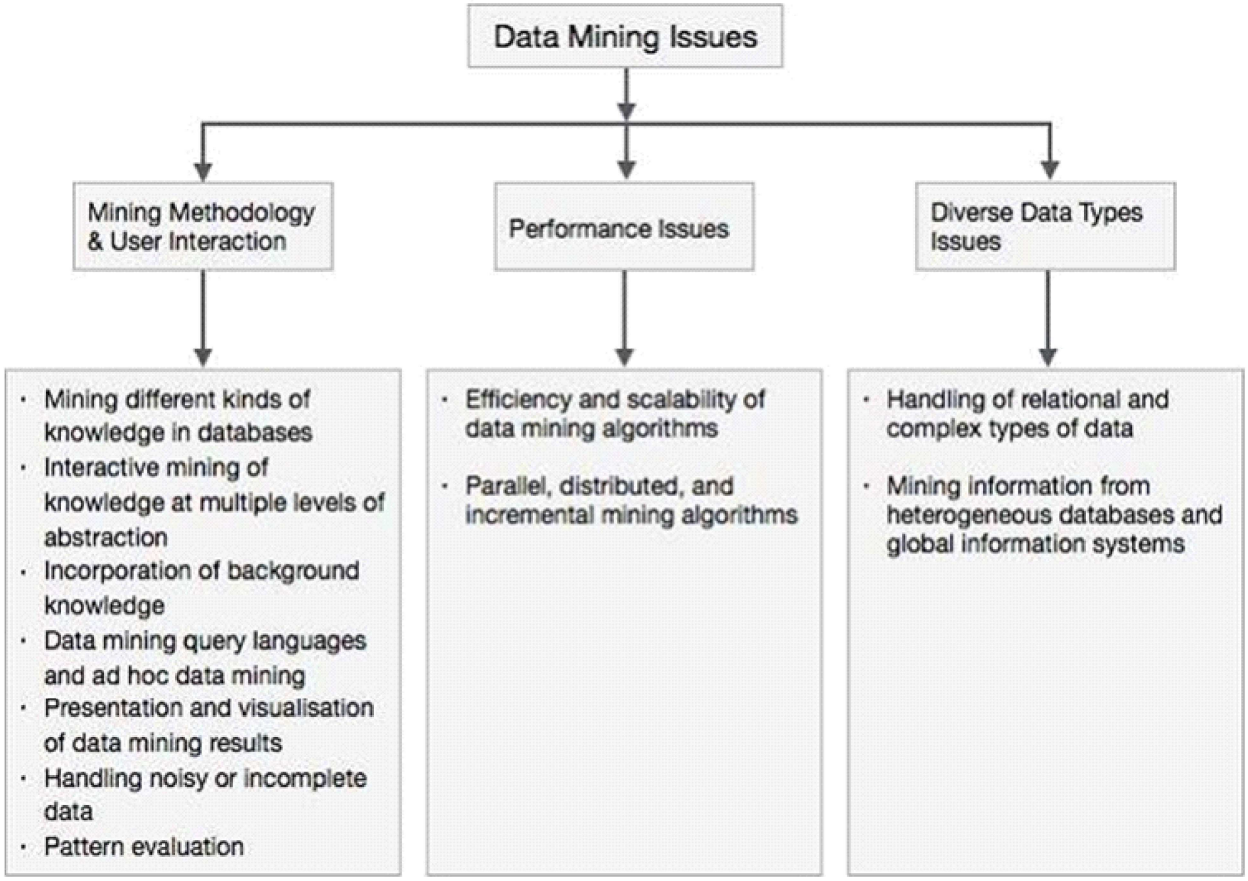


Figure 1.2: Data Mining Issues

1.8.1 Mining Methodology and User Interaction Issues

It refers to the following kinds of issues–

- **Mining different kinds of knowledge in databases:** Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction:** The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge:** To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining:** Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results:** Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data:** The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation:** The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

1.8.2 Performance Issues

There can be performance-related issues such as follows:

- **Efficiency and scalability of data mining algorithms:** In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms:** The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions are merged. The incremental algorithms, update databases without mining the data again from scratch.

1.8.3 Diverse Data Types Issues

- **Handling of relational and complex types of data:** The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems:** The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

CHECK YOUR PROGRESS

Q.1: Answer the following multiple choice questions:

..... is an essential process where intelligent methods are applied to extract data patterns.

i) Data warehousing ii) Data mining
iii) Text mining iv) Data selection

Q.2: Which of the following is not a data mining functionality?

i) Characterization and Discrimination
ii) Classification and regression
iii) Selection and interpretation
iv) Clustering and Analysis

Q.3: ..... is a summarization of the general characteristics or features of a target class of data.

i) Data characterization ii) Data classification
iii) Data discrimination iv) Data selection

Q.4: ..... is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

i) Data characterization ii) Data classification
iii) Data discrimination iv) Data selection

Q.5: ..... is the process of finding a model that describes and distinguishes data classes or concepts.

i) Data characterization ii) Data classification

iii) Data discrimination iv) Data selection

1.9 LET US SUM UP

- Data Mining is a non-trivial process of discovering knowledge from huge amount of data, as mining of gold from rocks or sand is called gold mining, similarly data mining is appropriately named as knowledge mining.
- To extract the knowledge from large data set Knowledge Discovery from Data (KDD) is used.
- Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data.
- Non-dependency-oriented data are the simplest form of data and typically refers to multidimensional data. This data typically contains a set of records.
- A record is also referred to as a data point, instance, example, transaction, entity, tuple, object, or feature-vector, depending on the application at hand.
- In case of dependency-oriented data, implicit or explicit relationships may exist between data items.
- Classification analysis is the organization of data in given classes or in short classification is to partition the given data into predefined disjoint groups.
- Association analysis is the discovery of what are commonly called *association rules*.

1.10 FURTHER READING

1) Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

2) Pujari, A. K. (2001). *Data Mining Techniques.* Universities Press.

3) Saxena, A., Saxena, K. Saxena, S. (2015). *Data Mining and Warehousing.* BPB Publications.

4) https://www.tutorialspoint.com/data_mining/dm_quick_guide.htm

1.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: ii) Data mining

Ans. to Q. No. 2: iii) Selection and interpretation

Ans. to Q. No. 3: i) Data characterization

Ans. to Q. No. 4: iii) Data discrimination

Ans. to Q. No. 5: ii) Data classification

1.12 MODEL QUESTIONS

Q.1: What is data mining?

Q.2: What is KDD?

Q.3: Explain the various types of basic data.

Q.4: Explain data mining functionality.

Q.5: Explain the classification of data mining.

Q.6: Explain the major issues of data mining.

*** ***** ***

UNIT 2: INTRODUCTION TO DATA WAREHOUSING

UNIT STRUCTURE

2.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define data warehouse and data base management system
- differentiate between data warehouse and data base management system
- explain the reasons behind the requirement of data warehouse
- describe operational and informational data stores

- list data warehouse characteristics
- describe how a data warehouse is built and what is its structure
- calculate the cost of data warehousing.

2.2 INTRODUCTION

In the later part of 1980, Barry Devlin and Paul Murphy had developed a data warehouse where the flow of information from operational databases to decision support system has been proposed. Requirement of different information to make strategic decisions has become rapidly increased in 1990 due to the increase of businesses and global growth of different corporations. Due to the increase of competition and complexity in businesses, different organizations require more information that will help to make strategic decisions. The operational databases provide information only for executing daily normal operations but it could not be able to provide strategic information.

In 1990, a new concept termed as data warehouse has been evolved which can be able to provide strategic information. Different organizations had started to build data warehouse to make capable enough their decision support mechanism so that it will help to simplify and grow their businesses. In the previous unit we have got an introduction to data mining. In this unit, we will give an introduction to data ware housing. We will also learn about the data base management system and different aspects of data warehouse. In the next unit, we will explore the concept of OLAP in detail.

2.3 DATA BASE MANAGEMENT SYSTEM AND DATA WAREHOUSE

In this section, the definitions of data warehouse and data base management system (DBMS) are presented. The differences between data warehouse and DBMS are also discussed in the later part of this section.

2.3.1 Data Base Management System

A database management system (DBMS) can be defined as a software system that is responsible for data storing, data manipulation

and data validation in a database. It supports data access from a database with an efficient and secured manner. Management of data and data format in a database are also supported by DBMS.

2.3.2 Data Warehouse

A data warehouse is a collection of integrated data from various types of sources that help to prepare analytical reports and in the efficient performance of decision making systems. It also supports structured and ad hoc queries. Data in a data warehouse are structured according to different subjects. The format of data may be different in different sources as they may be received from different applications. But in data warehouse, all data must be stored with a common format and to make it possible, data cleaning and data integration techniques are applied to convert inconsistent data to consistent data. The data warehouse stores data according to a particular time unit. For example, examination results of a college from 2014 to 2018. So historical information are only stored in data warehouse. Physical data storage for data warehouse is always made separate to the operational data bases and so it does not require any transaction processing, recovery and concurrency control mechanism. In most of the times, the data in a data warehouse cannot be modified. Only new data are continuously loaded to it.

2.3.3 Difference between Data Warehouse and DBMS

There are some differences between data warehouse and DBMS. These differences are discussed as below:

- DBMS contains transactional data and it is termed as OnLine Transaction Processing (OLTP) system. On the other hand, data warehouse contains analytical data and it is termed as OnLine Analytical Processing (OLAP) system. So DBMS are constructed to record data and data warehouses are constructed to analyze data.

- DBMS stores application oriented data and at most of the times it is based on single application. But data warehouse stores subject oriented historical data received from multiple heterogeneous sources.
- In DBMS, data are changed regularly as different transactions are performed frequently. But in case of data warehouse, data are not allowed to be changed or modified at most of the times.
- Recovery and concurrency control mechanism are very essential in DBMS as data are changed regularly due to different transaction processing operations. But data warehouse does not require any recovery and concurrency control mechanism.
- Data warehouse deals with long duration historical data but DBMS do not operate on long duration data as most of the times it deals with the current data with short time duration.
- DBMS can provide very less support to the decision making system but decision support mechanism is significantly supported by the data warehouse.

2.4 THE NEED FOR DATA WAREHOUSING

Building and utilization of a data warehouse is termed as data warehousing. Since 1990, data warehousing has become an essential part of every organization due to the increase in requirement of strategic information to make their decision making system efficient. Strategic information always helps the executives and managers to provide better business policies for their companies so that better results in businesses have been achieved by them than their competitors. Because of data warehousing, the time required for data analysis is reduced significantly. Historical and analytical information from data warehouse are utilized by the executive and the managers to achieve the following objectives:

- To learn about different operations related to the organization.
- To learn and identify different primary factors that can affect the business and monitor these factors in a regular period of time so that the business performance of the organizations can be compared to that of its competitors.

- To regularly monitor the requirements of customers and their preferences.
- To continuously monitor the results of sales and marketing.
- To regularly monitor product quality and services.
- To learn about new technologies that can be utilized to make businesses grow and to simplify business related complexities.

Data warehousing provides a common data format for all the stored integrated data that are received from various sources and so for this reason the data analysis and sharing of analyzed data become easier. Because of the common data format, data warehousing can minimize the possibility of error in data interpretation and improve data consistency.

Data warehousing provides easier data accessibility as data are stored separately in one place. It can store multidimensional data and provide support to the users to perform query analysis and analysis of stored data.

Data warehouse also maintains data security by providing secure access to the stored data by authenticated users. In this environment, the users are allowed to access only those data which are specific to them. It can also provide accessibility of corporate data to the authenticated customers and vendors for the development of new business strategies.

In recent times, data warehousing has become an integral part of every large organization. Data warehousing is commonly applied in the following domains:

- Banking and financial services
- Analysis of biological data
- Quality product manufacturing
- Consumer data analysis
- Retail sectors
- E-commerce
- Telecom sectors
- Logistics and Inventory management
- Insurance sectors
- Educational institutions for educational data analysis

2.5 OPERATIONAL DATA STORES

Operational data of different organizations are constructed by various on-line transaction processing operations of operational applications. These data are recent, complete, non redundant and modifiable data. Operational data store (ODS) is a database that stores operational data from various sources and process these data. After processing, ODS transfers these data to operational systems and the data warehouses. So, ODS stores recent integrated operational data about various products, customers etc. These data are not application specific and accessible from all parts of an organization.

ODS has similarities as well as differences with data warehouse. Similarities of ODS and data warehouse are as follows:

- Both ODS and data warehouse contains subject oriented data. In both cases data are not application specific.
- Data in ODS are entirely integrated thus showing a similarity with data warehouse.

Differences of ODS and data warehouse are summarized below:

- First of all, the architecture of ODS and data warehouse is fairly different.
- ODS stores short term current data but on the other hand data warehouse contains long duration historical data.
- In case of ODS, data are regularly changeable or updatable. When new data are transmitted to the ODS, then these recent data will overwrite the older version data of the related fields. So, no historical data are available in ODS. But in case of data warehouse, data are not changeable or updateable. So all historical data are available in data ware house.

2.6 INFORMATIONAL DATA STORES

Informational data store is a collection of summarized and redundant data about different subjects like product, customer, vendor etc. Informational data are utilized to provide appropriate response to the different queries

placed by corporate executives and managers for decision making purpose. Informational data for a corporation can be collected from different applications, databases, computer systems and operational data stores which are available in the corporation. These data are not changeable or updatable.

Differences between informational data store and operational data store are discussed below:

- Data model in case of an operational data store is normalized to maintain ACID properties i.e., atomicity, consistency, isolation, and durability. But it is not required in case of informational data store.
- Informational data store contains current, redundant, summarized and historical data. But operational data store contains only short duration current data.
- Data accessing in case of informational data store is performed primarily on ad hoc basis. On the other hand, only predefined and structured access of data is possible in case of operational data store.
- In case of informational data store, data are not changed or updated frequently. At most of the times, periodic and planned batch wise data updates are observed in informational data store. But in operational data store, data are continuously changed or updated fewer current data.
- The number of concurrent users of informational data store is always fewer than that of the operational data store.

2.7 DATA WAREHOUSE CHARACTERISTICS

The main characteristics of a data warehouse are described as follows:

- Data warehouse is a type of database that stores subject-oriented data achieved from various sources or applications to provide support to the decision making system of a corporation by performing data analysis. Data are stored in data warehouse according to some time period so that data analysis becomes accurate.
- Data warehouse is a collection of integrated and standardized data. All applicable data from various sources and applications are

combined together and stored in data warehouse for efficient decision making purpose. These integrated data may be in different data formats as they are transmitted from different types of operational systems and applications. So, these data must be standardized by removing the inconsistencies from them so that all data are available with a common data format and can be utilized efficiently in the decision making.

- Data warehouse is non-volatile. Data warehouse contains current and long duration historical data. In most of the times, data are not changed or updated in data warehouse. Only new data are included to it. But if it supports change of data then data can be changed or updated periodically.
- Small number of user is supported by data warehouse.
- A data warehouse separates operational data stores and informational data store.
- It has been observed that external data are also very useful for efficient decision making by a corporation. Data warehouse integrates the useful external data from various sources or applications available outside of a company and maps it to the related applications of the company for better decision making.

2.8 DATA WAREHOUSE STRUCTURE

Two types of data warehouse architecture are frequently used by different corporations. These are two-tier architecture and three-tier architecture.

Two-Tier Architecture: The data warehouse two-tier architecture is based on **client–server architecture.** In this architecture, direct communication between the client and the data server is available. The client layer of this architecture is responsible for providing user interface, data access, data aggregation, data analysis, query specification and report formatting. The data warehouse server is the server layer of this architecture. Data logic and data services are executed by this layer. It also stores and maintains metadata. The two-tiered architecture lacks of scalability and flexibility. Large

number of end-users also cannot be supported by this type of architecture. But it is easy to maintain and data modification is also easy in this case.

Three-Tier Architecture: Three-tier architecture is the most popular data warehouse architecture. This architecture consists of three layers that are **bottom tier**, **middle tier** and **top tier**.

- **Bottom-Tier:** The data warehouse database server is placed at the bottom tier in three-tier data warehouse architecture. This database server is a relational database system.
- **Middle-Tier:** The OLAP Server is implemented in the middle-tier.
- **Top-Tier:** The client layer is placed at the top-tier. It consists of different tools for data analysis, query, data mining and reporting.

There are seven basic components of data warehousing architecture as shown in figure 2.1 and discussed in the following parts of this section.

Seven basic components:

i) **Data warehouse Database:** In most of the cases, the data warehouse database is a relational database management system (RDBMS). It is the central database of the data warehousing system. But it has been observed that the traditional RDBMS system cannot be optimized for data warehousing due to some of the factors that can affect the performance of the data warehouse. Some examples of these factors are very large database size, ad-hoc query, aggregates, multi-table joins etc. Some technological approaches that can be used as a solution to this issue are as follows:
   - Parallel relational databases can be used in a data warehouse for better scalability.
   - Multidimensional database can also be used to remove limitations that are available due to the relational data model.
   - Efficient latest index structures can be used to avoid relational table scan which can improve speed.

ii) **Tools for data sourcing, data cleanup, data transformation and data migration:** The tools for data sourcing, data cleanup, transformation, and migration provide the conversion operations, structural modifications, summarizations and key changes. This

component of data warehouse structure is required to convert dissimilar data into information so that it can be utilized by the decision support system. It is required to transmit data from various operational systems to the data warehouse. The metadata is also maintained by this component.

iii) **Metadata:** Metadata is defined as the data about data that describes the data warehouse. It is a very essential component of data warehouse architecture. It is utilized to build and manage the data warehouse. It identifies the data source, data usage, data values and data features of data warehouse. It also describes how data warehouse data can be processed and changed or updated. Metadata repository is responsible for metadata management.

There are two classes of metadata available in a data warehouse *technical metadata* and *business metadata*.

- **Technical Metadata:** Technical metadata is the information about the data which are utilized by data warehouse administrators and designers for data warehouse development and management purpose. Some of the information stored in technical metadata are mentioned as follows:
  - Information about data sources
  - Information about the data transformation methods from operational databases to the data warehouse
  - Information about the procedures that are used for data cleanup and data enhancement
  - Information about data access and access permission
  - Information about data warehouse data structures
- **Business Metadata:** Business metadata stores information which helps the end-users to understand easily the various data contained in a data warehouse. Business metadata stores information about different subject areas of data warehouse. It stores the information that supports all data warehousing components. It also contains information regarding data usage, data history, data ownership etc.

iv) **Data Marts:** A data mart is a structure that stores data that are related to specific subject area and used by the specific group of users and departments of an organization. So data marts are smaller in size and more flexible than data warehouse. It is a subset of the data warehouse. Most of the times, data marts can be used as an alternative to large data warehouses because it is less expensive and less time is required to build it. Data marts can also reduce the response time for end-users by permitting users to be able to access only those specific data that are required by the users. Both data marts and data warehouse can be formed in the same database. Data mart can also be created in a physically separate database.

v) **Access Tools:** Access tools in data warehousing are utilized to provide information to business users for taking efficient strategic decisions. Access tools provide interactions between users and data warehouse system. There are four types of access tools available in data warehousing. These are *query and reporting tools*, *application development tools*, *data mining tools*, *OLAP tools and executive information system tools*.

vi) **Data warehouse administration and management:** Data warehouse administration and management component is responsible for security management, priority management, metadata management, monitoring data quality, data cleaning, data replicating, data distribution, monitoring data updates from various sources, assessment of data warehouse usage, data warehouse usage reporting etc. It is also responsible for managing data warehouse storage and maintaining backup and recovery process.

vii) **Information delivery system:** The information delivery component is responsible for the distribution of data warehouse data and different information items to end-user products and other data warehouses.

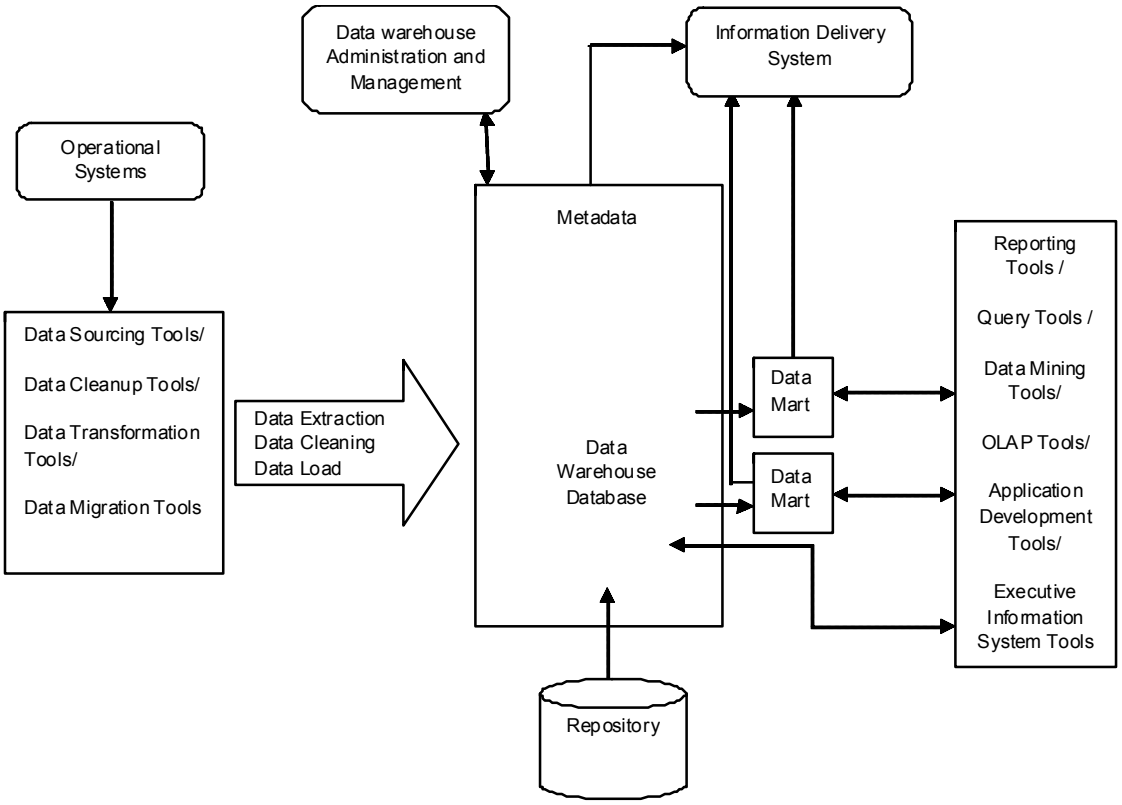


Figure 2.1: Data Warehousing Architecture

2.9 BUILDING A DATA WAREHOUSE

There are two approaches to build a data warehouse. These are top down approach and bottom up approach.

- ***In top down approach***, at first an enterprise data model is developed and then enterprise wise business requirements are collected. After these two processes, the process to build a data warehouse with data marts is started. In this approach, new data mart can be constructed very easily from the data warehouse. But this approach is not flexible enough for the development of a data warehouse where different departmental requirements may be changed. This approach is expensive and it requires more time for initial set up.
- ***In bottom up approach***, at first, individual data marts specific to different subject area and business requirements are constructed. Then these data marts are integrated to build the data warehouse. In this approach, extension of the data warehouse to include new

business area is very easy as it requires only the creation of new data mart and integration of it to the data warehouse. It require less time for initial set up. But in this approach, the integration of data mart to the data warehouse is a difficult process.

For building a data warehouse: design considerations; technical considerations and implementation considerations are discussed in the following subsections.

2.9.1 Design Consideration

All the components of data warehouse, all kind of data sources related to the data warehouse and all kind of information requirements are considered for designing an efficient data warehouse. Data integration from various sources of different types is a very important design consideration of data warehouse. Regular interactions with the end users are also a basic requirement for a successful data warehouse designing. Some other important points for successful data warehouse design are presented below:

- Data model of a data warehouse should be closely related to the data content and structure of the data warehouse. Data warehouse and data marts may have different data models.
- A data warehouse must have some method and technology to maintain metadata repository and to include new information to the metadata regularly.
- Data placement and data distribution strategies should be properly configured in the process of data warehouse designing.
- In recent times, there are various types of tools available that can be used to implement data warehouse. The most appropriate tools related to a particular data warehouse environment should be selected by the data warehouse designers to implement a data warehouse otherwise, it may affect the performance of the data warehouse.
- Data warehouse designer must clearly recognize the end users' requirements to access different data.

2.9.2 Technical Consideration

The technical considerations to build a data warehouse are discussed below:

- The hardware platform for a data warehouse server should be capable enough to store and maintain the required amount of data for decision support systems of an organization. The data warehouse server must be specialized so that it can perform all the related jobs of the data warehouse.
- Metadata repository must be supported by the hardware platform and the associated software.
- A balance between all kinds of computing components should be maintained to design and implement a data warehouse.
- The data warehouse DBMS must be compatible with the data warehouse so that it can easily handle very large size databases and it can also process the complex ad hoc queries efficiently.
- The communications networks should be capable enough to transmit large amount of data and for this purpose the latest hardware and software can be utilized.

2.9.3 Implementation Consideration

Implementation of a data warehouse can be performed by combining all the associated products and objects of a data warehouse. Now to implement a data warehouse, the basic steps to build a data warehouse must be performed. These are stated below:

- At first, all types of information regarding the requirement of a data warehouse by an organization must be collected and analyzed. Then a data model for the data warehouse has to be configured.
- In the next step, various data sources for the data warehouse must be identified. Then an appropriate DBMS and hardware platform has to be selected for the data warehouse server.

- At first, data are collected from the operational databases and then data transformation and data cleaning operation are performed so that these data can be stored in the data warehouse database.
- Different tools and software are selected next for the proper functioning of the data warehouse. For example: database access tools, reporting tools, database connectivity software, data analysis software, presentation software.
- Finally, the data warehouse must be updated when it is required.

Some other implementation considerations to build a data warehouse are as follows:

- Selection of appropriate access tools is one of the important considerations to implement a data warehouse. At the moment, we do not have any tool that can be used for all types of data warehouse access. So the most appropriate set of tools is used for this purpose.
- Data collection, data transformation, data cleanup and data migration operations must be properly performed for successful implementation of a data warehouse.
- There are two data storage approaches available for data warehouse data. In the first approach, a separate storage media is used to store some amount of older warehouse data that are detailed and less important. The other warehouse data are stored in the bulk storage media and it is maintained by the data warehouse server.

In the second approach, data warehouse data are divided depending upon different requirements and data types. Then data are distributed to the related multiple servers. In this approach, metadata must be stored in one source and it must be maintained by a single server.

- Metadata is a very important part of any data warehouse. So metadata must be properly collected and maintained in the process of data warehouse building. The metadata must be

accessible by all the data warehouse users so that they can use the data warehouse efficiently.

- A classification of data warehouse users has to be made for the proper use of a data warehouse. In general, the users are classified into the following categories depending upon their skills and access level of the warehouse.
  - **Casual users:** Casual users can access formatted information from a data warehouse. These users can execute only those queries and reports which are already there in a data warehouse.
  - **Power user:** Power user can create and execute simple ad hoc queries. These users can also examine the results of simple queries and reports. This kind of users requires access tools to perform their tasks.
  - **Experts:** Expert users are capable of developing complex queries. These users can perform complex analysis on the data warehouse information. These users also require different tools to perform their jobs efficiently. They possess a good understanding about the data warehouse data and tools.

2.10 THE COST OF DATA WAREHOUSING

The cost to build a data warehouse is not found to be similar for all types of businesses and corporations. It varies depending upon different business environments and requirements of information. It is not possible to estimate the exact cost to build a data warehouse. The cost of data warehousing is affected by different factors. Most significant factors are discussed below:

- **Hardware Costs:** A compatible hardware platform is required to build an efficient data warehouse. Different types of hardware like storage media, CPUs, data communication infrastructures, workstations etc are required for building a data warehouse. So a significant amount of hardware cost can be estimated for building a

data warehouse. On the other hand, for any data warehouse, amount of data and data usage may be increased day by day and it will also increase the hardware costs. It is also observed that the number of data warehouse users may affect the hardware cost of a data warehouse. If the number of users is increased then the hardware cost is also increased as the hardware requirement becomes more. Sometimes for better performance, additional or the latest technology hardware is required in data warehousing which increases its hardware cost.

- **Software cost:** The hardware platform of data warehousing require different softwares or software tools to perform their jobs efficiently. In recent times, there are various open source softwares available but in case of data warehousing, all possible warehousing tasks cannot be performed by using only open source software. So, a cost related to the required software or software tools is also associated with the process of building a data warehouse. On the other hand, software cost of data warehousing may increase due to the requirement of software maintenance. Software prices also may be increased in future. DBMS is one of the important softwares that is required in data warehousing.
- **Human resource cost:** Data warehousing always requires different types of users and skilled professionals to use the data warehouse for different business purposes and for regular maintenance of it. For example, at least one dedicated system manager, a software engineer and backend developers are required to support the warehouse database. We know that data warehouse must be updated regularly and some skilled professionals are also required to monitor and execute this job. So a cost has to be estimated for building data warehouse related to the employment of these human resources. Different types of trainings to improve efficiency of the human resources are also provided in every corporation. So a cost related to the user trainings is also associated to build a better data warehouse.

CHECK YOUR PROGRESS

Q.1: i) Which of the following is not true for data warehouse?

A) Data warehouse contains subject oriented data

B) Data warehouse contains only short duration recent data.

C) Data warehouse provides analytical information for decision making process.

D) Data warehouse is different from traditional database systems.

ii) Data model of ..... is not required to normalize for maintaining ACID properties.

A) Informational data store B) Operational data store

C) DBMS D) Data warehouse

iii) OLAP server of a data warehouse is implemented in the ..... tier of the three-tier structure.

A) Bottom tier B) Second tier

C) Middle tier D) Top tier

iv) Which of the following is not a basic component of data warehouse architecture?

A) Data warehouse database

B) Metadata

C) Data marts

D) Data warehouse users

v) Which of the following is not true for casual users of data warehouse?

A) Casual user cannot create queries.

B) Casual user can access formatted information from a data warehouse.

C) Casual users can perform complex analysis on the data warehouse information.

D) Both (A) and (B)

vi) Which of the following is a basic factor to estimate the cost of a data warehouse?

A) Hardware B) Software

C) Human resources D) All of the above

Q.2: State whether the following statements are true or false:

i) All data in a data warehouse must be stored with a common format.

ii) Data warehouse contains transactional data and it is termed as OnLine Transaction Processing (OLTP) system.

iii) Both operational data store and data warehouse contains subject oriented data.

iv) The number of concurrent users of informational data store is always very less than of the operational data store.

v) A data warehouse combines operational data stores and informational data store.

vi) Data marts are larger in size and less flexible than data warehouse.

2.11 LET US SUM UP

- A data warehouse is a collection of integrated data from different types of sources that help to prepare analytical reports and provide required analytical information to the decision making systems of a corporation.
- All data in a data warehouse must be stored with a common format. Due to the common data format, data warehousing can minimize the possibility of error in data interpretation and improve data consistency.
- The data warehouse stores data according to a particular time unit.
- DBMS contains transactional data and data warehouse contains analytical data.
- DBMS stores application oriented data and data warehouse stores subject oriented historical data received from multiple heterogeneous sources.

- In case of data warehouse, data are not allowed to be changed or modified in most of the times.
- Data warehouse does not require any recovery and concurrency control mechanism.
- Data warehousing provides easier data accessibility as data are stored separately in one place. It can store multidimensional data and provide support users to perform query analysis and analysis of stored data.
- Data warehouse maintains data security by providing secure access to stored data by authenticated users.
- In recent times, data warehousing is applied in different domains like banking and financial services, analysis of biological data, quality product manufacturing, retail sectors, E-commerce, telecom sectors etc.
- Operational data of different organizations are constructed by various on-line transaction processing operations of operational applications. These data are recent, complete, non redundant and modifiable data.
- ODS contains subject oriented data like data warehouse. Data in ODS is entirely integrated.
- The architectures of ODS and data warehouse are different.
- Informational data store is a collection of recent, summarized and redundant data about different subjects. Informational data are used to provide appropriate response to the different queries placed by corporate executives and managers for decision making purpose.
- Data model in case of an informational data store is not required to be normalized for maintaining ACID properties.
- Data accessing In case of informational data store is performed primarily on ad hoc basis.
- In case of informational data store, data are not updated regularly. In most of the times periodic and planned batch wise data updates are observed in informational data store.
- Two-tier architecture and three-tier architecture are the two popular data warehouse architecture.
- The data warehouse two-tier architecture is based on **client–server architecture.**

- Three-tier architecture consists of three layers that are bottom tier, middle tier and top tier.
- A data warehouse architecture contains seven basic components which are data warehouse database, tools for data sourcing, data cleanup, data transformation and data migration, metadata ,data marts, access Tools, data warehouse administration and management and information delivery system
- Metadata is the data about data that describes the data warehouse. Two classes of metadata are available in data warehousing that are technical metadata and business metadata.
- A data mart is a structure that stores data related to specific subject area and used by the specific group of users and departments of a corporation. Data marts are smaller in size and more flexible than data warehouse.
- Four types of access tools available in data warehousing which are query and reporting tools, application development tools, data mining tools, OLAP tools and executive information system tools.
- Two approaches are available to build data warehouse that are top down approach and bottom up approach.
- Data warehouse users are classified into three classes that are casual users, power user and experts.

2.12 FURTHER READING

1) Berson, A., & Smith, S. J. (1997). *Data Warehousing, Data Mining and OLAP*. McGraw-Hill, Inc.
2) Inmon, W. H. (2005). Building the Data Warehouse. John Wiley & Sons.
3) Ponniah, P. (2004). *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. John Wiley & Sons.

2.13 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) B; ii) A; iii) C; iv) D; v) C; vi) D

Ans. to Q. No. 2: i) True; ii) False; iii) True; iv) True; v) False; vi) False

2.14 MODEL QUESTIONS

Q.1: Write down the differences between data warehouse and DBMS.

Q.2: Why is data warehouse required by a corporation?

Q.3: Write down the similarities and differences between operational data store and data warehouse.

Q.4: Write down the characteristics of data warehouse.

Q.5: What is metadata? Why is it required?

Q.6: What is data mart?

Q.7: Write down the different design considerations for building a data warehouse.

Q.8: Write down different technical considerations for building a data warehouse.

Q.9: Write down different implement considerations for building a data warehouse.

*** ***** ***

UNIT 3: INTRODUCTION TO OLAP

UNIT STRUCTURE

3.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the different uses of OLTP and OLAP
- define multi-dimensional data cube
- describe the different operations of OLAP like Roll-Up, Slice, etc.
- differentiate between OLTP and OLAP.

3.2 INTRODUCTION

In the previous unit, we have learned about the data base management system and different aspects of data warehouse. In this unit we will learn about OLAP and OLTP. In this unit, we will also learn how different OLAP operations like Roll-Up, Roll-down, Slice, Dice and Pivot can be performed on a multidimensional data base i.e data cube to analysis the data. In addition to this, we will also learn to differentiate OLTP and OLAP. In the next unit, we will explore the concept of data preprocessing in detail.

3.3 OLTP AND OLAP

OLTP (Online Transactional Processing) is a category of data processing that is focused on transaction-oriented tasks. OLTP is used for business task. It provides a multi-dimensional view of different business tasks. In OLTP, backup and recovery process is maintained religiously. Following are some examples of OLTP.

- Online Banking Transaction
- Online Shopping
- Booking Online ticket
- Order entry etc

OLTP applications typically possess the following characteristics:

- Transactions that involve small amounts of data
- Indexed access to data
- A large number of users
- Frequent queries and updates
- Fast response times

Online Analytical Processing (OLAP) allows the user to view and analyse data in multiple views. This analytical processing enables the user to select and view data from different points of view. OLAP is used to access live data online and to analyze it. It also allows to process data in multi dimensions.

3.4 OLAP OPERATION

There are different types of OLAP operations. They are listed below.

- Roll-Up
- Roll-down
- Slice
- Dice
- Pivot

The above mentioned OLAP operations can be explained with the help of a data cube **C** which is used to store the sales data of an electronic enterprise. Thus the three dimensions are **Year, Items and City.** These

dimensions allow to keep the record of monthly sales of different products or sales of products from different cities in different years etc. as given in figure 3.1.

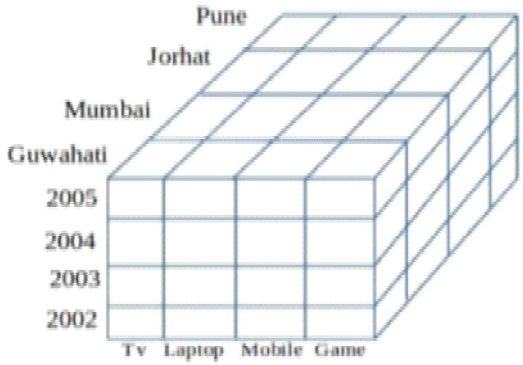


Figure 3.1: Data Cube C[Year, Items, City]

3.4.1 Roll-Up or Drill Up

The roll-up operation performs aggregation on a data cube either by climbing up the hierarchy or by dimension reduction. The figure 3.2 shows the result of **Roll up** operation by climbing up the hierarchy of locations that is **city** to state. In other words, we can say that resulting cuboid group the data by **State** rather than city.

Roll-up C[Year, Items, City] = C[Year, Items, State]

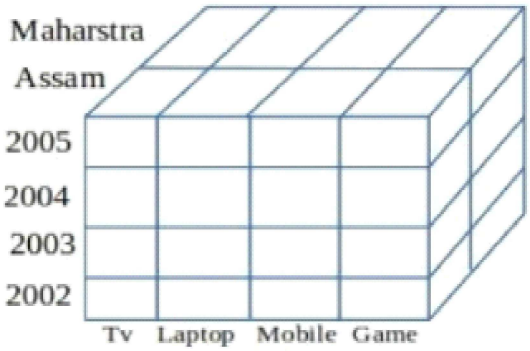


Figure 3.2: Roll-up Operation

Here, each data cell of the cuboid is the aggregation of those data cell that are merged due to roll-up operation. In other word, the result stored in the data cell C[2005, Tv, Assam] is the sum of the data stored in the data cell (figure 3.1) C[2005, Tv, Guwahati] and c[2005, Tv, Jorhat].

When roll-up is performed by dimension reduction, then one or more dimensions from the data cube are removed.

3.4.2 Roll-down or Drill-down

This operation is opposite to roll-up operation. This operation can be performed by stepping down the hierarchy or by adding new dimension. The drill-down operation is concerned with switching from aggregation to more details. The following figure 3.3 shows the result of **Roll-down** operation by stepping down the hierarchy of time that is **Year** to **Month**. In other word we can say that resulting cuboid group the data by **Month** rather than **Year**.

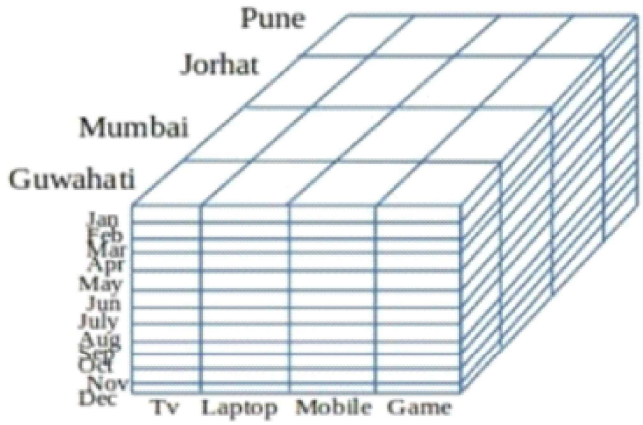


Figure 3.3: Drill-down operation

3.4.3 Slicing

The slice operation selects one particular dimension from the given cube and produces a new sub cube.

The following figure 3.4 shows the slicing operation. Here, the slicing operation is performed for the dimension **Year** using the criterion **Year="2004"**

$$slice_{Year=2004}\ C[Year, Items, City] = C[Items, City]$$

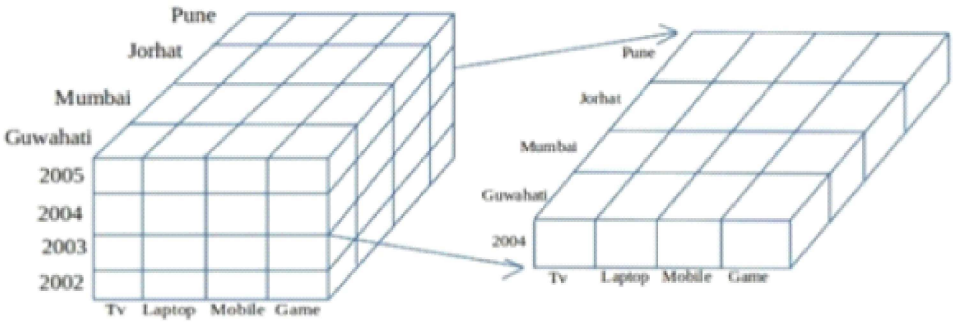


Figure 3.4: Slicing Operation

Each data cell of the resultant sub cube will contain city wise details of all items for the year 2004.

3.4.4 Dicing

The dicing operation is for selecting a smaller cube and it analyzes the cube from different perspectives. The dicing selects two or more dimensions from the original cube and produces a sub cube. The following figure shows the dicing operation. Here, the dicing operation is performed in the following criterion, which involves two dimensions **City** and **Year**.

Dice $_{\text{year=“2003" or “2004 and city= “Guwahati” or “Jorhat”}}$ **C[Year, Item, City] = C[Year', Item, City']**

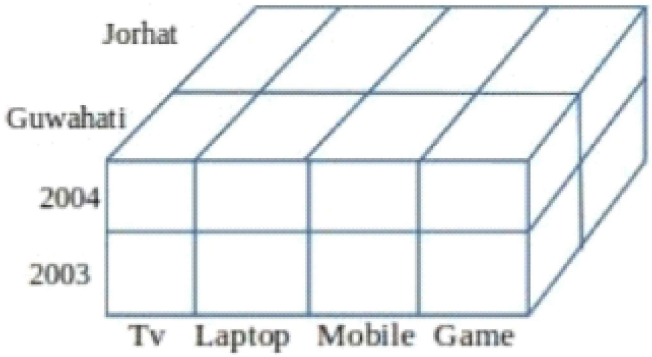


Figure 3.5: Dicing Operation

3.4.5 Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data.

CHECK YOUR PROGRESS

Q.1: State whether the following statements are true (T) or false (F)

i) Roll up operation performs aggregation on data cube.

ii) Roll-down operation can be performed by stepping up the hierarchy or by deleting an existing dimension.

iii) Slice operation selects a particular cell from a data cube and produce a new data cube.

iv) Dicing selects two or more dimensions from the original cube and produces a sub cube.

Q.2: Fill in the blanks:

i) The drill-down operation is concerned with switching from ..... to more details.

ii) Pivot operation ..... the data axes in view in order to provide an alternative presentation of data.

3.5 OLAP VERSUS OLTP

The differences between OLAP and OLTP are listed below.

Sl. No.	Data Warehouse (OLAP)	Operational Database (OLTP)
1	Involves historical processing of information.	Involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	Useful in analyzing the business.	Useful in running the business.
4	It focuses on Information out.	It focuses on Data in.
5	Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.
6	Contains historical data.	Contains current data.
7	Provides summarized and consolidated data.	Provides primitive and highly detailed data.
8	Provides summarized and multidimensional view of data.	Provides detailed and flat relational view of data.
9	Number or users is in hundreds.	Number of users is in thousands.
10	Number of records accessed is in millions.	Number of records accessed is in tens.

11	Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
12	Highly flexible.	Provides high performance.

3.6 LET US SUM UP

- OLTP is an application oriented that provides a multi-dimensional view of different business tasks.
- OLAP is an analytical processing that enables user to select and view data from different point of view.
- Different types of OLAP operations are Roll-Up, Roll-down, Slice, Dice and Pivot.
- Aggregation on a data cube is performed by the roll-up operation.
- The drill-down operation is concerned with switching from aggregation to more details.
- To produce a new sub cube by selecting one particular dimension we can perform Slice operation.
- The dicing selects two or more dimensions from the original cube and produces a sub cube.
- Pivot operation rotates the data axes in view in order to provide an alternative presentation of data.

3.7 FURTHER READING

1) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.
2) Tan, P. N., Steinbach, M. & Kumar, V. (2013). *Data Mining Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining*.

3.8 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) True; ii) False; iii) False; iv) True

Ans. to Q. No. 2: i) aggregation; ii) rotate

3.9 MODEL QUESTIONS

Q.1: What is OLTP?

Q.2: What are the different characteristics of OLTP?

Q.3: What is OLAP?

Q.4: What are the different OLAP operations?

Q.5: Explain the different OLAP operations with a suitable example.

Q.6: How is the **Roll-Up** operation different from **Roll-Down** operation?

Q.7: Differentiate between **Slicing** and **Dicing** operation.

Q.8: Differentiate between OLAP and OLTP.

*** ***** ***

UNIT 4: DATA PREPROCESSING

UNIT STRUCTURE

4.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the different stages of data processing
- define data summarization and data cleaning
- explain data transformation
- describe data reduction.

4.2 INTRODUCTION

In the previous unit, we have learned about OLAP and OLTP. We have also learned how different OLAP operations like Roll-Up, Roll-down, Slice, Dice and Pivot can be performed on a multidimensional data base. In this unit, we will learn about data preprocessing as well as the different stages involved in data preprocessing. Here, we will discuss different data processing techniques involved in data mining such as data cleaning, data transformation, data reduction. We will also discuss about concept hierarchies in this unit. In the next unit, we will explore the concept of multidimensional data in the form of data cube in detail along with different data warehouse schema's.

4.3 DATA PREPROCESSING

In data mining, data preprocessing is a technique which involves transformation of raw data into an understandable format. Data preprocessing is a proven method of resolving real-world data's issues such as inconsistent, incomplete, and/or lacking in certain behaviors or trends.

There are six stages involved in data processing:

- **Data Collection:** Data collection is the first step that is involved in data processing technique. Data is collected from available trustworthy and well-built available resources that include data lakes and data warehouse.
- **Data Preparation:** Data preparation stage is the second stage involved in data processing. Data preparation is also referred to as "pre-processing". At this stage, the raw data that is collected in first stage is cleaned up and organized for the following stage of data processing. During preparation, raw data is checked for any kind of errors and this stage also eliminates bad data (redundant, incomplete, or incorrect data) and begins to create high-quality data for the best business intelligence.
- **Data Input:** The clean data is then entered into its destination and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.
- **Processing:** During this stage, the data input to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, although the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).
- **Data Output/Interpretation:** The output/interpretation stage is the stage at which data becomes finally usable to non-data scientists. It is translated, made readable, and is often in the form of graphs, videos,

images, plain text, etc. Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

- **Data Storage:** The final stage of data processing is storage. After all the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Besides, properly stored data is a necessity for the sake of compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of an organization when needed.

4.4 DATA SUMMARIZATION

Summarization is a key data mining concept which involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied for exploratory data analysis, data visualization and automated report generation.

CHECK YOUR PROGRESS

Q.1: Fill in the blanks:

i) Data preparation is also known as .....

ii) ..... is the first stage of data processing in which raw data begins to take the form of usable information.

iii) Summarization involves techniques for finding a compact description of a .....

4.5 DATA CLEANING

For decision making we need data warehouse and it is very essential that the data in data warehouse be correct. Since large volume of data are collected from heterogeneous sources, so there is a high chance of having error in data. Therefore, to construct an error free and high-quality data warehouse data cleaning is essential. Data cleaning technique includes:

- using transformation rules, e.g., translating attribute name like 'age' to 'DOB'
- using domain-specific knowledge.
- performing parsing and fuzzy matching, e.g., for multiple data sources, one can designate a preferred source as a matching standard, and
- auditing, i.e., discovering facts that flag unusual patterns.

4.6 DATA TRANSFORMATION

Data transformation is a process of transforming heterogeneous data that are collected from different data sources to an uniform structure so that data can be combined and integrated.

Data transformation operations would contribute toward the success of the mining process.

- **Smoothing:** It helps to remove noise from the data.
- **Aggregation:** Summary or aggregation operations are applied to the data. i.e., the weekly sales data is aggregated to calculate the monthly and yearly total.
- **Generalization:** In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.
- **Normalization:** Normalization is performed when the attribute data are scaled up or scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.
- **Attribute Construction:** these attributes are constructed and included in the given set of attributes which are helpful for data mining.

4.7 DATA REDUCTION

Data reduction is a technique that is applied to a data warehouse to obtain a reduced representation of the data set that is much smaller in volume, yet it closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient and yet it must produce the same (or almost the same) analytical results.

There are several data reduction strategies. Those are shown below:

- **Data Cube Aggregation:** Aggregation operations are applied to the data in the construction of a data cube.
- **Dimensionality Reduction:** In dimensionality reduction redundant attributes are detected and removed. This reduces the data set size.
- **Data Compression:** Encoding mechanisms are used to reduce the data set size.
- **Numerosity Reduction:** In numerosity reduction, the data are replaced or estimated by alternative.
- **Concept Hierarchy Generation:** In concept hierarchy, the raw data values for attributes are replaced by ranges or higher conceptual levels.

4.8 CONCEPT HIERARCHIES

Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides the users with the flexibility to view the data from different perspectives.

Data mining on a reduced data set means fewer input/output operations and it is more efficient than mining on a larger data set.

Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining rather than during mining.

CHECK YOUR PROGRESS

Q.2: Fill in the blanks:

i) ..... is a process of transforming heterogeneous data.

ii) ..... helps to remove noise from the data.

iii) ..... technique is applied to a data warehouse to obtain a reduced representation of the data set.

iv) ..... can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

4.9 LET US SUM UP

- Data preprocessing is a technique which involves transformation of raw data into an understandable format.
- Six stages involved in data processing are Data Collection, Data Preparation, Data input, Processing, Data output, Storage.
- Summarization is a technique for finding a compact description of a dataset.
- Data cleaning is essential to construct an error free and high-quality data warehouse.
- Data transformation is a process of transforming heterogeneous data to an uniform structure.
- Normalization is performed when the attribute data are scaled up or scaled down.
- **Data reduction is** a technique that is applied to a data warehouse to obtain a reduced representation of the data set.
- In dimensionality reduction redundant attributes are detected and removed. This reduces the data set size.
- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

4.10 FURTHER READING

1) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.
2) Tan, P. N., Steinbach, M. & Kumar, V. (2013). *Data Mining Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining*.

4.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) Preprocessing; ii) Data input; iii) Data set

Ans. to Q. No. 2: i) Data transformation; ii) Smoothing; iii) Data reduction; iv) Concept hierarchies.

4.12 MODEL QUESTIONS

Q.1: What is data preprocessing? What are the different stages involve in data preprocessing?

Q.2: What is data summarization?

Q.3: Why is data cleaning important? What are the different data cleaning techniques?

Q.4: How do data transformation operations contribute toward the success of the mining process?

Q.5: What is data reduction?

Q.6: What are the different data reduction strategies?

Q.7: Explain the concept hierarchies.

*** ***** ***

UNIT 5: MULTIDIMENSIONAL DATA

UNIT STRUCTURE

5.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the concept of data model and multidimensional representation of data
- define data cube of multidimensional data representation
- explain the concepts of dimension modelling
- describe the components of multidimensional data model
- describe the different data warehouse schema such as star, snowflake and fact constellation.

5.2 INTRODUCTION

We are already familiar with the concepts of data warehousing and OLAP (Online Analytical Processing). We know that the core of the design of the data warehouse lies in a multidimensional view of the data model. A data model is a description of the organization or the structure of data in an information system. Data warehouse users explore data to find useful

patterns by studying how certain attributes of data elements (i.e., *measures*) are related to other attributes (i.e., *dimensions*). Initially, the user has to specify which attributes of the original data is to be treated as measures and which to treated as dimensions. The data is conceptually organized as multi-dimensional array, where each dimension corresponds to a dimension of the warehouse, and the values stored in each cell of the array corresponds to the measure of the warehouse.

In this unit, we will learn about multidimensional view of data, data cubes and different data warehouse schema such as *star schema*, *snowflake schema* and *fact constellation*. In the next unit we will explore data warehouse architecture, data warehouse design, OLAP three-tier architecture, indexing and querying in OLAP, OLAM etc.

5.3 DATA CUBE

Multidimensional data model stores data in the form of data cube. To understand the concept of multidimensional view of data, let us take the data set represented in 2-D in the following table 5.1 (*example is taken from S Choudhury, 2009*) for an employment data warehouse *"employment in California"* in order to keep records of the employment details with respect to the dimensions *sex, year* and *profession.*

A data cube allows data to be modelled and viewed in multiple dimensions. An OLAP data cube is also known as *hypercube*. It is defined by *dimensions* and *facts*. *Dimensions* are the perspectives or entities with respect to which an organization wants to keep records. For example, in *Employment in California*, there are three dimensions namely *sex, year* and *profession*. Each dimension may have a table associated with it, called a *dimension table*. For example, a dimension table for *sex* may contain the attributes item *male* and *female*

A multidimensional data model is typically organized around a central theme, like *employment*, for instance. This theme is represented by a fact table. Facts are numerical measures by which it analyzes relationships between dimensions. Examples of facts for above example data warehouse may include *total male employees*, *total civil engineers and total employees*

in a year etc. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

Table 5.1: Statistical Table: Two-dimensional Representation

			Professional Class					
			Engineer		Secretary		Teaching	
			Profession		Profession		Profession	
			Chemical Engineer	Civil Engineer	Junior Secretary	Executive Secretary	Elementary Teacher	High School Teacher
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	2112	1548
		95	2078	3282	5664	1711	2053	1380
	MALE	91	258	1120	6673	1623	2160	2751
		92	289	1276	6925	1744	2175	2993
		93	312	1398	7152	1889	2189	3125
		94	518	1216	6543	1534	2857	2387
		95	329	1321	6129	1567	2453	3287

The 3-D data cube representation of the above data set information in table 5.1.; can be represented as shown below in the figure: 5.1. (Example is *taken from [A K Pujari, 2009]).*

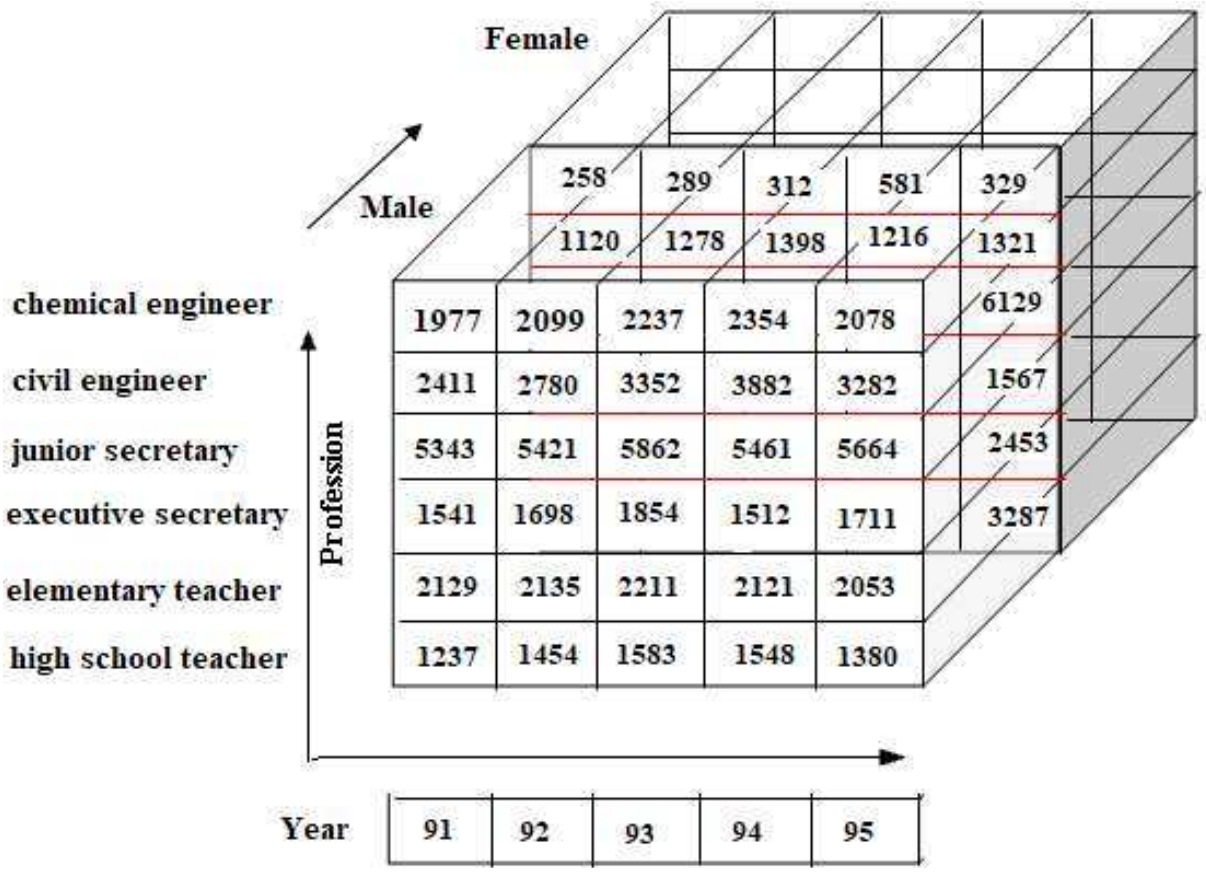


Figure: 5.1. Multidimensional representation of data (data cube)

5.4 LATTICE OF CUBOIDS

The data cube is a metaphor for multidimensional data storage. It helps to represent the dimension hierarchy in the multidimensional view of data i.e. multidimensional data can be represented as a *lattice of cuboids*. The important thing to remember is that data cubes are n-dimensional and do not confine data to 3-D. The cuboid that holds the lowest level of summarization is called the *base cuboid (n-D cuboid)* and it consists of all the data cells. Any n-D data can be displayed as a series of (n”1)-D cubes which are obtained by grouping the cells and computing the numeric measures (facts) of all n-dimensions.

The cuboid consisting of one cell with numeric measures of all *n* dimensions holds the highest level of summarization. It is called the *apex cuboid* (*0-D cuboid)*. All the other cuboids lie between the base cuboid and apex cuboid in the lattice of cuboids.

Let us take another example of a data warehouse for *All Sports* to keep records of the store's sales with respect to the dimensions *time*, *product*, *branch*, and *location*. These dimensions allow the store to keep track of things like monthly sales of products and the branches and locations at which the products were sold.

Table 5.2: A 3-D view of sales data for *All Sports*, according to the dimensions *time*, *product* and *location*. The measure displayed is rupees (in thousands)

	Location="Guwahati"				Location="Nagaon"				Location="Dibrugarh"				Location="Nalbari"			
time	products Cricket iitems	Foot ball items	jerseys	Indoor items	products Cricket ing items	Foot ball items	jerseys	Indoor items	products Cricket ing items	Foot ball items	jerseys	Indoor items	Cricket ing items	Foot ball items	jerseys	Indoor items
Q1	210	234	1245	678	156	987	677	789	891	787	552	771	245	452	345	535
Q2	122	567	4560	342	789	876	689	787	1190	990	256	884	892	245	321	663
Q3	190	788	7789	227	564	556	560	565	567	905	772	340	672	563	1334	245
Q4	567	345	3476	909	667	450	2342	409	4511	1092	644	496	677	566	245	678

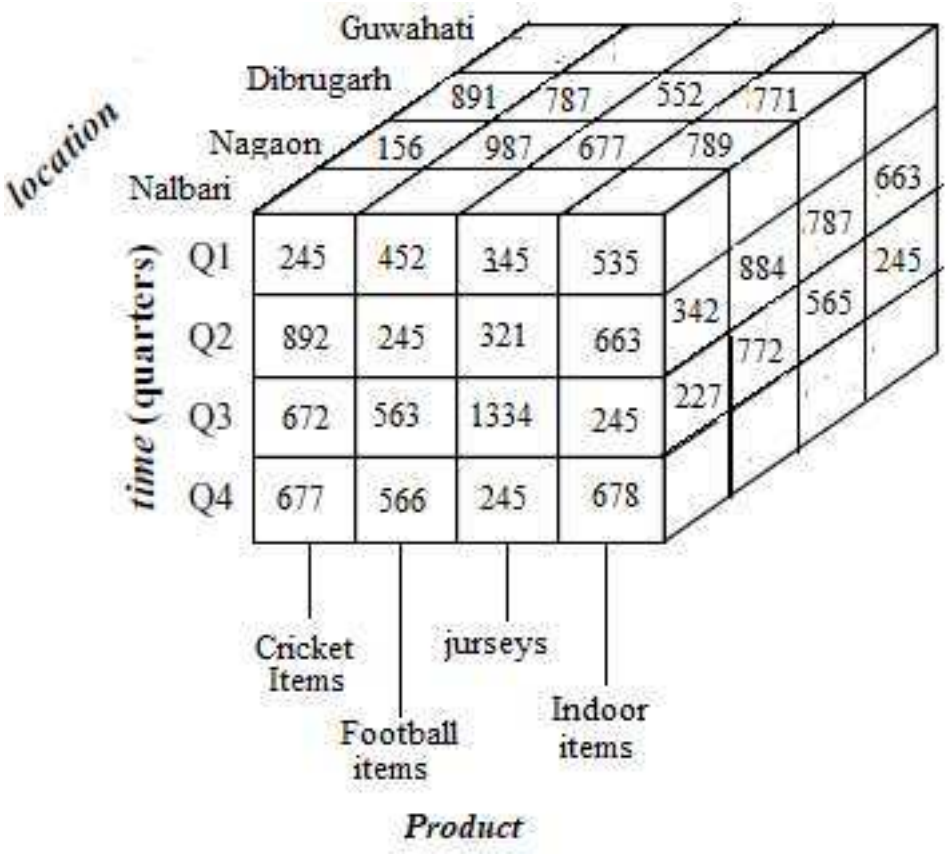


Figure 5.2: A 3-D data cube representation of the data in table 5.2, according to the dimensions *time*, *product* and *location*. The measure displayed is rupees (in thousands)

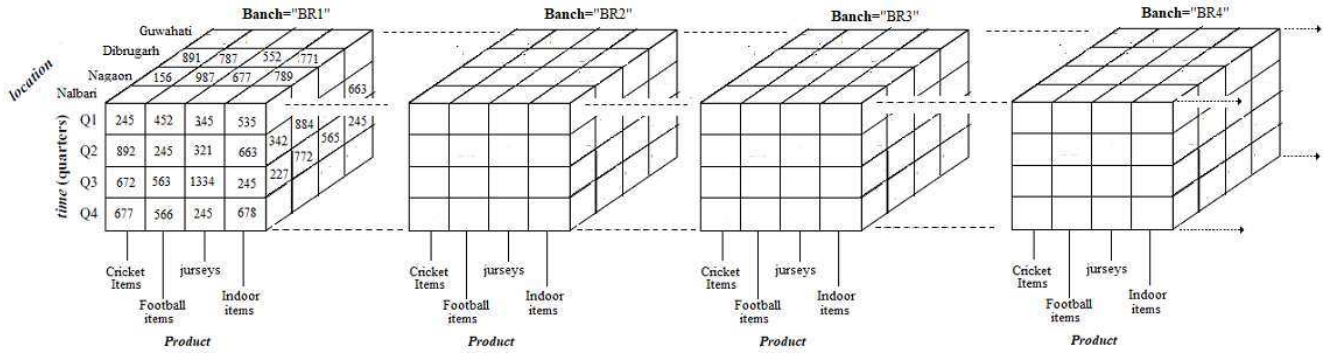


Figure 5.3: A 4-D data cube representation of sales data, according to the dimensions *time, product, location, and Branch.* The measure displayed is Rupees sold (in thousands). For improved readability, only some of the cube values are shown

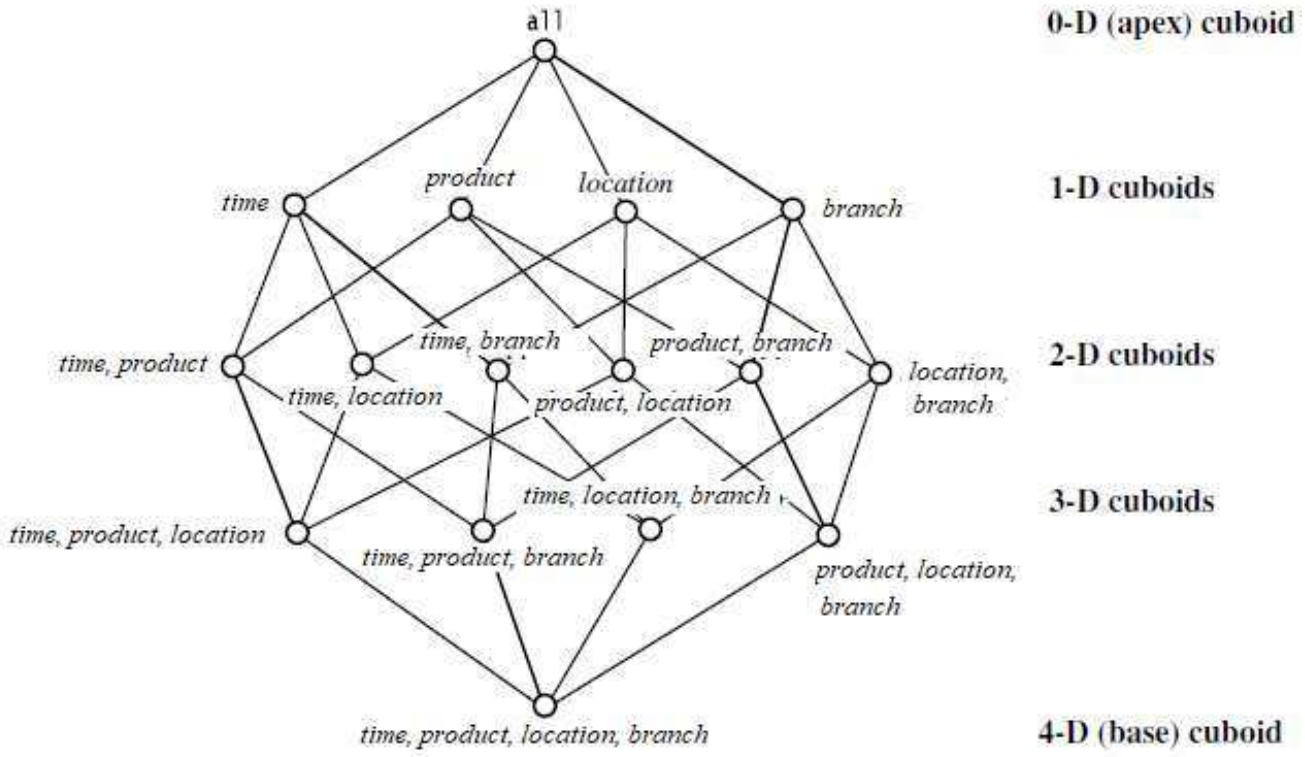


Figure 5.4: Lattice of cuboids, making up a 4-D data cube for the dimensions time, product, location, and branch. Each cuboid represents a different degree of summarization.

CHECK YOUR PROGRESS

Q.1: The core of the multidimensional model is the ....., which consists of a large set of facts and a number of dimensions.

a) Multidimensional cube b) Dimensions cube
c) Data cube d) Data model

Q.2: Which of the following is not a kind of data warehouse application?

a) Information processing b) Analytical processing
c) Data mining d) Transaction processing

Q.3: Data cube can grow n number of dimensions, thus becoming:

a) dimensional cube b) solid cubes
c) star cubes d) hyper cubes

Q.4: Which of the following describes the data contained the data warehouse?

a) Relational data b) Meta data
c) Informational data d) Operational data

Q.5: Data that can be modelled as dimension attributes and measured attributes are called as .....

a) Multidimensional b) Single dimensional
c) Measured data d) dimensional data

Q.6: OLAP stands for–

a) Online analysis processing
b) Online transaction processing
c) Online analytical processing
d) Online aggregate processing

5.5 DATA WAREHOUSE SCHEMA

Multidimensional schema is especially designed to model data warehouse systems. Schema is a logical description of the entire database. The entity-relationship (ER) data model is generally used in the design of

relational databases. A database schema consists of a set of entities and the relationships between them. It includes the name and description of records of all record types including all associated data-items and aggregates. Similar to databases, data warehouses also need to maintain a schema. There are various types of data warehouse schema; they are: *Star schema*, *Snowflake schema*, and *Fact Constellation schema*.

5.5.1 Star Schema

The star schema architecture is the simplest data warehouse schema and is widely used to develop or build a data warehouse and dimensional data marts. This schema is widely used to develop or build a data warehouse and dimensional data marts. It is called a star schema because the diagram resembles a star. It consists of a single large central *fact table* and a set of smaller *dimension table,* one for each dimension. The fact table contains the detailed summary data with no redundancy. It's primary key has one key per dimension. Each dimension table is joined with the fact table using a primary or foreign key. The fact table contains the *fact or subject* of interest for each corresponding dimension in the dimension table. It also stores numerical measures for those co-ordinates which are non-dimensional attributes. The relationship between fact table and dimensional table is 1:N. An example of star schema is shown in figure 5.5.

In the star schema shown in figure 5.5, **SALES** is a fact table having attributes *product_id, time_id, customer_id, employee_id* and *location_id* which references to the dimension tables *product, time, customer, employee* and *location* respectively. It also contains two numerical measures: *price_sold* and *qty_sold*. **Employee** dimension table contains the attributes: *employee_id, first_name, mid_name, last_name* and *dob*. **Time** dimension table contains attributes: *time_id, day, month, quarter* and *week***. Location** dimension table contains attributes: *location_id, district, street_number* and *city_name.* **Product** dimension table contains

the attributes: *product_id, product_name, product_type, category* and *unit_price.* **Customer dimension** table contains the attributes: *customer_id, customer_fname, customer_lname, city, state, pin_no* and *contact_no.*

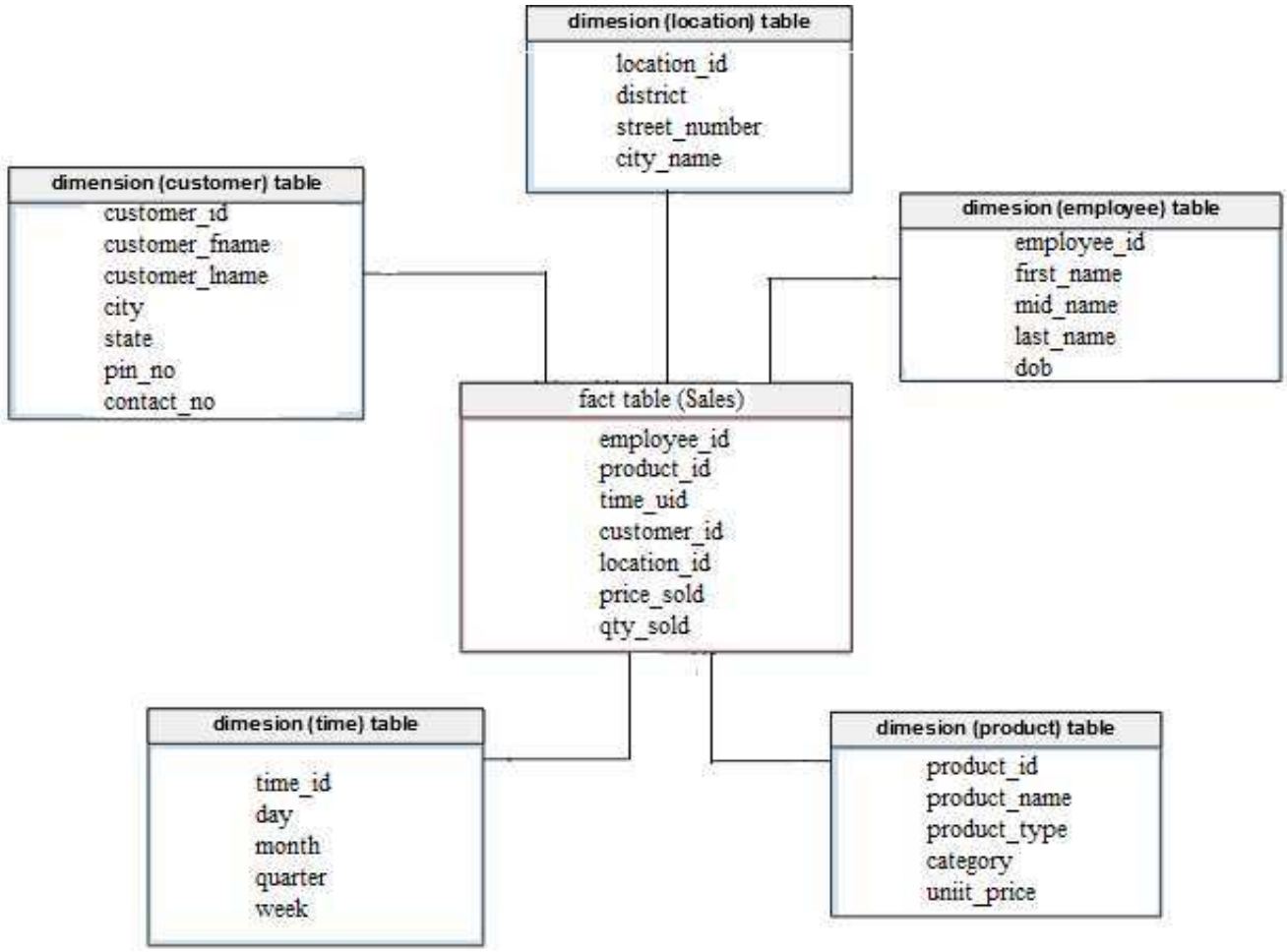


Figure 5.5: Star Schema

Advantages of Star Schema Data Warehouses:

- Easy to understand
- Easy to define dimension hierarchies
- Reduces number of physical joins
- Easy to maintain
- Used very simple metadata

Disadvantages of Star Schema Data Warehouses:

- Data integrity is not enforced well since it is in a highly de-normalized state

- Is not flexible in terms of analytical needs
- Normally do not reinforce many-to-many relationships within business entities.

EXERCISE 5.1

Draw a Star schema for a Library management Data warehouse.

5.5.2 Snowflake Schema

The *snowflake* schema is more complex compared to star schema because the dimension tables of the snowflake are normalized to support attributed hierarchies. A sample snowflake schema is shown in figure 5.6.

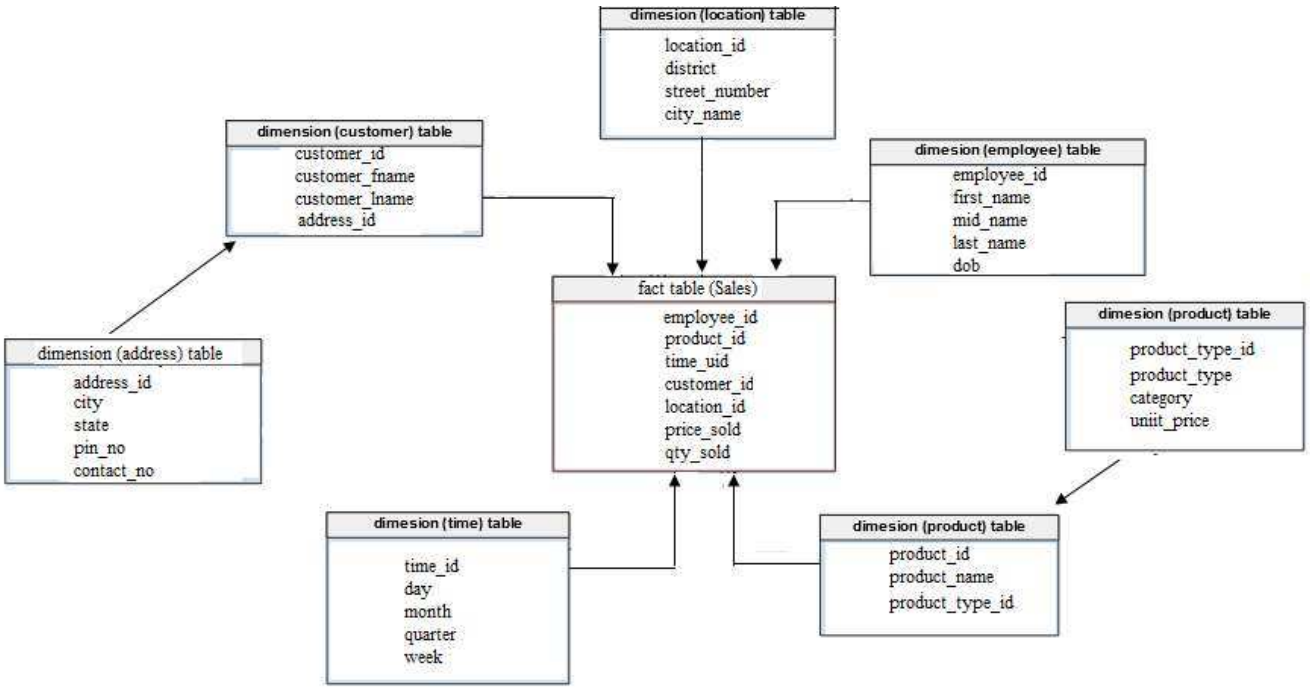


Figure 5.6: Snowflake Schema

The snowflake schema consists of a centralized fact table which is connected to multiple dimension tables and the dimension tables can be normalized into additional dimension tables. Similar to star schema, the fact table in snowflake schema also contains the *fact or subject* of interest for each corresponding dimension in

the dimension table and stores numerical measures for those co-ordinates as well. In contrast to star schema, the dimension table in snowflake schema is normalized.

Advantages of Snowflake Schema Data Warehouses:

- Easy to maintain because dimension tables are normalized and normalizing results in saving storage spaces
- Reduces redundant information storage

Disadvantages of Snowflake Schema Data Warehouses:

- Increase in large number of join operations.

EXERCISE 5.2

Draw a Snowflake schema for a Hospital management Data warehouse.

5.5.3 Fact Constellation

Fact constellation schema is more complex than star or snowflake schema. It contains more than one fact table which share

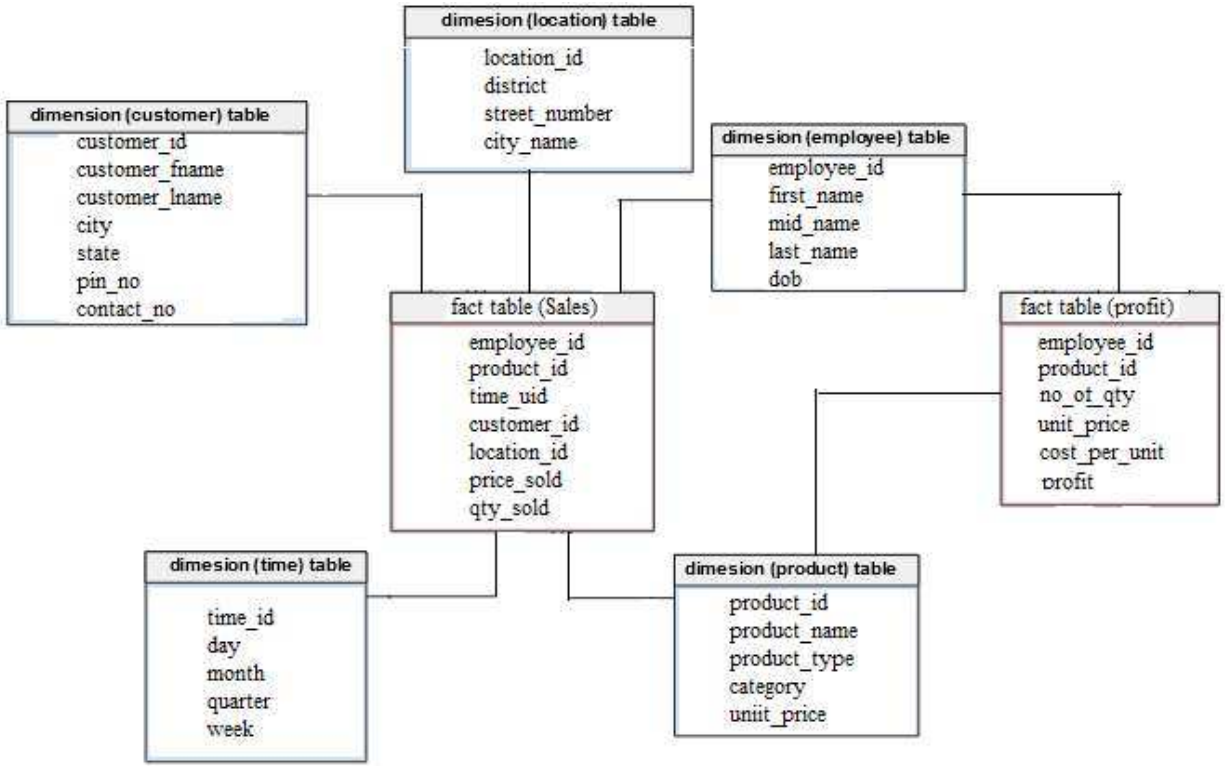


Figure 5.7: Fact Constellation Schema

some dimension tables. It is also referred to a galaxy schema. A sample fact constellation schema is shown in figure 5.7. There are two fact tables *Sales* and *Profit* which share the same dimension tables *employee* and *product*.

Advantages of Fact Constellation Schema Data Warehouses:

- Provides a more flexible schema compared to other schemas
- Different fact tables are explicitly assigned to the dimensions

Disadvantages of Fact Constellation Schema Data Warehouses:

- Hard to maintain
- More Complexity involved due to the involvement of more number of aggregations.

CHECK YOUR PROGRESS

Q.7: Which is a good alternative to the star schema–

a) Star schema b) Snowflake schema

c) Fact constellation d) Star-snowflake schema

Q.8: The type of relationship in star schema is .....

a) many to many b) one to many

c) many to one d) many to many

Q.9: Which statement best describes fact table?

a) fact table describes the transactions stored in a DW

b) fact table is the main store of the descriptions of the transactions

c) fact table describes the granularity of data in a DW

d) fact table is the main store of all of the recorded transactions over time

Q.10: Which of the following is the numeric measurements or values that represents a specific business aspects or activity:

a) dimensions b) schemas

c) facts d) tables

Q.11: Fact tables in a Data Warehouse is:

a) partially normalized b) completely denormalized

c) completely normalized d) partially denormalized

Q.12: In which of the following, a fact table in the centre is directly linked with dimension table?

a) star schema b) snowflake schema

c) fact constellation schema d) relational schema

5.6 LET US SUM UP

- In data warehouse, data is viewed as multidimensional data model which stores data in the form of data cube.
- A data cube allows data to be modelled and viewed in multiple dimensions.
- An OLAP data cube is defined by *dimensions* and *facts*.
- The Dimensions are the perspectives or entities with respect to which an organization wants to keep records and Facts are numerical measures by which it analyzes relationships between dimensions.
- There are various types of data warehouse schema; they are: *Star schema*, *Snowflake schema*, and *Fact Constellation schema*.
- A star schema consists of a single large central *fact table* and a set of smaller *dimension table,* one for each dimension.
- The snowflake schema consists of a centralized fact table which is connected to multiple dimension tables.
- The Fact constellation schema; know as galaxy contains more than one fact table which share some dimension tables.

5.7 FURTHER READING

1) Pudi, V., Krishna R. P. (2008) *Data Mining*. Oxford University Press.

2) Pujari, A. K. (2001). *Data Mining Techniques*. Universities Press.

5.8 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (a) **Ans. to Q. No. 2:** (d)

Ans. to Q. No. 3: (d) **Ans. to Q. No. 4:** (b)

Ans. to Q. No. 5: (a) **Ans. to Q. No. 6:** (c)

Ans. to Q. No. 7: (c) **Ans. to Q. No. 8:** (b)

Ans. to Q. No. 9: (d) **Ans. to Q. No. 10:** (c)

Ans. to Q. No. 11: (c) **Ans. to Q. No. 12:** (a)

5.9 MODEL QUESTIONS

Q.1: What is a data model?

Q.2: What is multidimensional data model?

Q.3: What is data cube? What do you mean by lattice of cuboids?

Q.4: What is fact table? How it is related to dimension table?

Q.5: What is data warehouse schema? What are the different types of data warehouse schema?

Q.6: Explain star schema. State the advantages and disadvantages of star schema?

Q.7: Briefly describe Snowflake schema in data warehouse. State the advantages and disadvantages of star schema? How is it different from star schema?

Q.8: What is fact constellation schema? State any two advantages of fact constellation schema.

*** ***** ***

UNIT 6: DATA WAREHOUSE ARCHITECTURE

UNIT STRUCTURE

6.1 LEARNING OBJECTIVES

After going through this unit you will be able to:

- describe the architecture of data warehouse specially three-tier architecture
- explain the different warehouse models
- describe how to design a warehouse
- describe OLAM and the conversion from OLAP to OLAM.

6.2 INTRODUCTION

In this previous unit we have learned about multidimensional view of data, data cubes and different data warehouse schema such as *star schema*, *snowflake schema* and *fact constellation*. In this unit we will learn about data warehouse architecture, data warehouse design. We will also learn about OLAP three-tier architecture in detail along with indexing &

querying in OLAP, OLAM etc. In the next unit we will explore the concept of data, knowledge and different data visualization techniques.

To design a data warehouse, first we collect the operational data from different source database. Process it for consistency and load in data warehouse. Data warehouse gives the advantages like track the trends and patterns over a long period, can gather information quickly and efficiently, helps us to manage customer relationship. We can design three types of data warehouse server namely data mart, virtual data warehouse and enterprise warehouse. And get another data warehouse as metadata. Data warehouse design is the process of building a solution to integrate multiple sources data that support analytical reporting and data analysis. It consists of requirements gathering, physical environment setup, data modeling, ETL, OLAP cube design, front end development, report development. Indexing the data warehouse can reduce the amount of time it takes to see query results. Indexing can be implemented on dimensions as well as fact table. Integration of OLAP and mining is called as OLAP mining (OLAM) and can easily transform from data warehouse to OLAM.

6.3 DATA WAREHOUSE ARCHITECTURE

Conceptual architectures have been proposed for a data warehouse. Operational data is the data collected and available in the transaction processing system. It resides in the different source databases. Before loading in the warehouse, first process the data for consistency from different sources. The detailed transaction level data that have been cleaned and confirmed for consistency is called reconciled data. It is used as base data for all warehouses.

From the data warehouses, the business analyst takes information to measure the performance and make critical adjustments in order to win over other business holders in the market. A data warehouse offers the following advantages–

- it can enhance business productivity because a data warehouse can gather information quickly and efficiently,
- A data warehouse provides a consistent view of customers and items, since it helps us manage customer relationship.

- A data warehouse also helps in bringing down the costs by tracking trends, patterns over a long period.

We need to understand and analyze the business needs to design an effective and efficient data warehouse and construct a **business analysis framework**. Each person has different views about the design of a data warehouse. These views are as follows–

- **The top-down view:** This view allows the selection of relevant information required for a data warehouse.
- **The data source view:** This view presents the operational system information being captured, stored, and managed.
- **The data warehouse view:** It represents the information stored inside the data warehouse with the help of fact tables and dimension tables.
- **The business query view:** It is the view of the data from the end-user viewpoint.

Data Warehouse architecture is presented in figure 6.1.

Single-Tier Architecture: The objective of a single layer is to minimize the amount of stored data. Remove data redundancy is the main goal of single tier architecture. This architecture is not frequently used in practice.

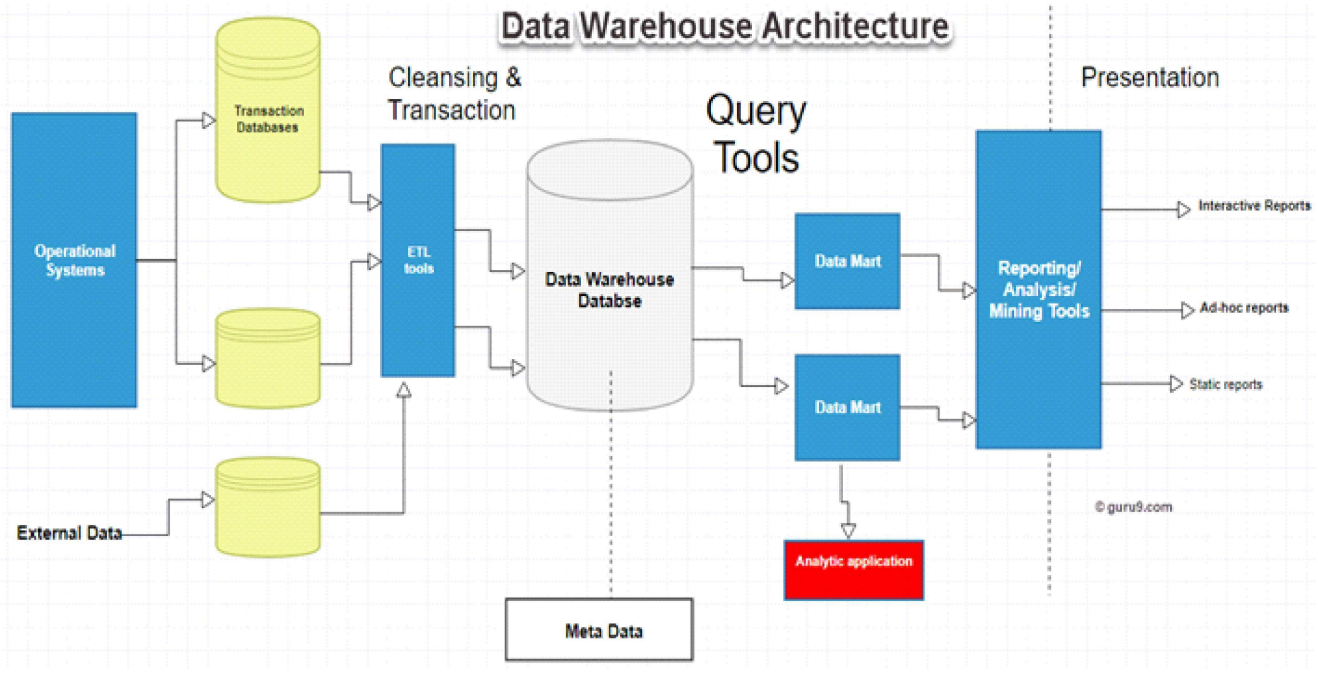


Figure 6.1: Data Warehouse Architecture

Two-Tier Architecture: Two-layer architecture separates physical sources and data warehouse. It is not expandable and also end-user does not support this architecture. Due to network limitations, it has connectivity problems.

Three-Tier Architecture: Generally a data warehouses adopts a three-tier architecture. It has three different tiers namely bottom tier as database server, middle tier as OLAP server and top tier as client or front end.

6.3.1 Data Warehouse Models

From the perspective of data warehouse architecture, we have the following data warehouse models–

- Virtual Warehouse
- Data mart
- Enterprise Warehouse

- **Virtual Warehouse:** The operational data warehouse view is known as a virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse needs excess capacity on operational database servers.
- **Data Mart:** Data mart contains a subset of organizational data. This subset of data is important to specific groups of an organization. In other words, we can claim that group specific data contain in a data mart. For example, data related to items, customers, and sales can be contained in the marketing data mart. Data marts are confined to subjects.

  Points to remember about data marts–

  - Window-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
  - The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.
  - The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.
  - Data marts are small in size.

- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data marts are flexible.

➢ **Enterprise Warehouse:**

- All the information and the subjects spanning an entire organization by an enterprise warehouse.
- It provides us integration of enterprise data.
- The data is integrated from operational systems and external information providers.
- This information can be small or large. It can be vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

6.3.2 Metadata

Metadata is simply defined as data about data. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to detailed data. In terms of data warehouse, we can define metadata as follows.

➢ Metadata is the road-map to a data warehouse.

➢ Metadata in a data warehouse defines the warehouse objects.

➢ Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Note: In a data warehouse, we create metadata for the data names and definitions of a given data warehouse.

Categories of Metadata: Metadata can be broadly categorized into three categories–

➢ **Business Metadata:** It refers to the data ownership information, business definition, and changing policies.

➢ **Technical Metadata:** It consists of database system names, table and column names and sizes, data types and allowed values. Technical metadata also contains structural information such as primary and foreign key attributes and indices.

- **Operational Metadata:** It contains currency of data and data lineage. Currency of data means whether the data is active, archived, or purged.

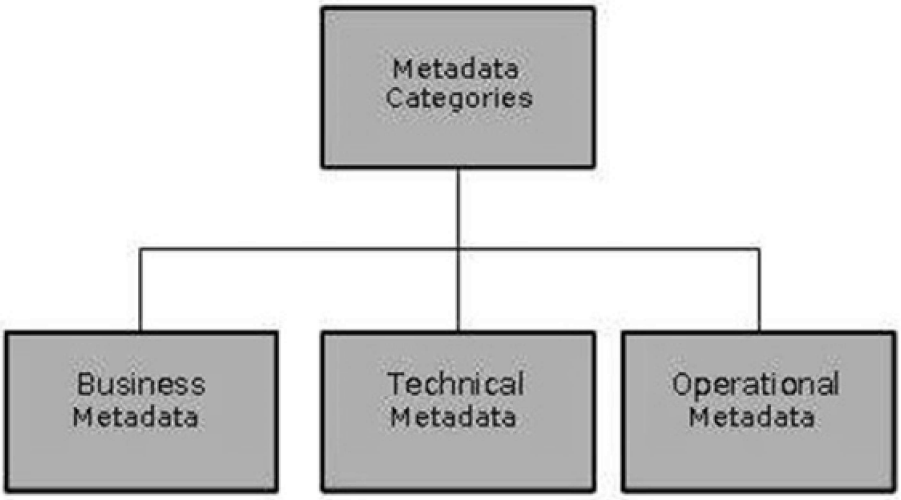


Figure 6.2: Categories of Metadata

Role of Metadata: The role of metadata in a warehouse is different from the warehouse data, and it plays an important role. The various roles of metadata are explained below.

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.
- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata is used for query tools.
- Metadata is used in extraction and cleansing tools.
- Metadata is used in reporting tools.
- Metadata is used in transformation tools.
- Metadata plays an important role in loading functions.

The following diagram shows the roles of metadata.

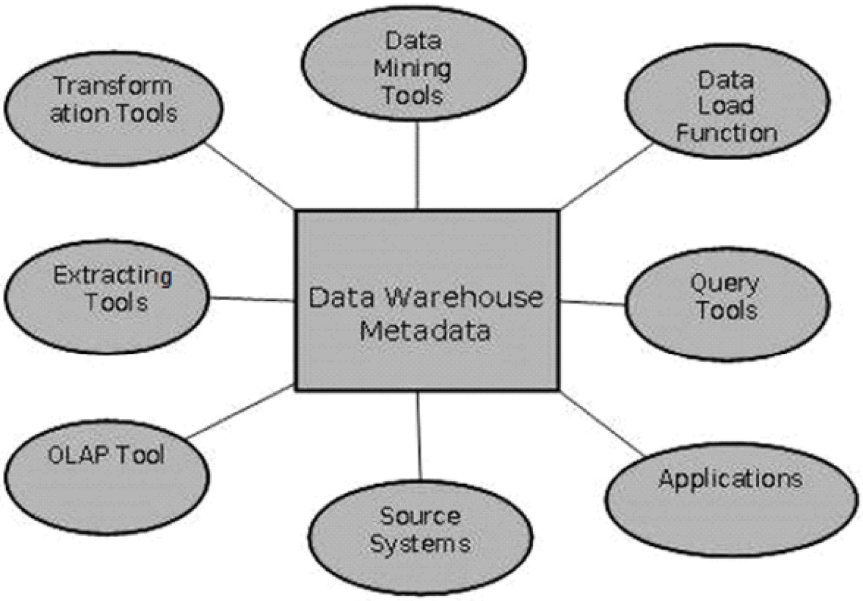


Figure 6.3: Data Warehouse Metadata

Metadata Repository: Metadata repository is an integral part of a data warehouse system. It has the following metadata–

- **Definition of data warehouse:** It includes the structure of data warehouse. Schema, view, hierarchies, derived data definitions, and data mart locations and contents are defined in the structure.
- **Business metadata:** It includes the data ownership information, business definition, and changing policies.
- **Operational metadata:** It contains currency of data and data lineage.
- **Data for mapping from operational environment to data warehouse:** It contains the source databases and their contents, data extraction, data partition cleaning, transformation rules, data refresh and purging rules.
- **Algorithms for summarization:**It consists of dimension algorithms, data on granularity, aggregation, summarizing, etc.

6.4 DATA WAREHOUSE DESIGN

Good Business Intelligence (BI), allows the organization to query data obtained from trusted sources and use the answers to gain a competitive edge in the industry. The first step to effective BI is a well-designed

warehouse. Data warehouse design is the process of building a solution to integrate multiple sources data that support analytical reporting and data analysis. A poorly designed data warehouse can result in acquiring and using inaccurate source data that negatively affect the productivity and growth of the organization.

Requirements Gathering: The first step of the data warehouse design process is gathering requirements. The goal of this phase is to determine the criteria for a successful implementation of the data warehouse. An organization's long-term business strategy is important as the current business and technical requirements. User analysis and reporting requirements is identified as well as hardware, development, testing, implementation, and user training. Once the business and technical strategy has been decided the next step is to address how the organization will backup the data warehouse and how it will recover if the system fails.

Physical Environment Setup: Once the business requirements are set, next step is to determine the physical environment for the data warehouse. There should be separate physical application and database server separate ETL/ELT, OLAP, cube, and reporting processes set up for development, testing, and production. The IT staff can investigate the issue without negatively impacting the production environment if integrity is suspected.

Data Modeling: Once requirements gathering and physical environments have been defined, the next step is to define how data structures will be accessed, connected, processed, and stored in the data warehouse through the process of data modeling. In this phase of data warehouse design, data sources are identified. Once the data sources have been identified, the data warehouse team can begin building the logical and physical structures to fulfill the established requirements.

ETL: The ETL process takes the most time to develop major implementation. Identifying data sources during the data modeling phase is reduce ETL development time. The goal of ETL is to provide optimized load speeds.

OLAP Cube Design: On-Line Analytical Processing (OLAP) provides the infrastructure for ad-hoc user query and multi-dimensional analysis.

Against the query of data OLAP design specification should come. OLAP cube dimensions are specified in the documentation and measures should be obtained during the beginning of data warehouse design process. The three critical elements of OLAP design include:

- Grouping measures– numerical value that want to analyze such as revenue, number of customers, how many products customers purchase, or average purchase amount.
- Dimension– where measures are stored for analysis such as geographic region, month, or quarter.
- Granularity– the lowest level of detail that you want to include in the OLAP dataset.

  The OLAP cube process is optimized during development.

Front End Development: This step is execute to work on how users will access the data warehouse. Front end development is define how users will access the data for analysis and run reports.

Report Development: For most end users, the only contact they have with the data warehouse is through the generated reports. An essential feature for data warehouse report generation is users' ability to select their report criteria quickly and efficiently is. Delivery options are other criteria. A well-designed data warehouse able to handle the new reporting requests with little to no data warehouse system modification.

CHECK YOUR PROGRESS

Q.1: The ..... exposes the information being captured, stored, and managed by operational systems.

a) top-down view b) data warehouse view

c) data source view d) business query view

Q.2: ..... describes the data contained in the data warehouse.

a) Relational data b) Operational data

c) Metadata d) Informational data.

Q.3: ..... databases are owned by particular departments or business groups.

a) Informational b) Operational

c) Both informational and operational d) Flat

Q.4: What is the full form of ETL?

.....

Q.5: Write name of the process in data warehouse design.

.....

.....

6.5 OLAP THREE TIER ARCHITECTURE

Generally a data warehouses adopts three-tier architecture. Following are the three different tiers of the data warehouse architecture.

- **Bottom Tier:** The data warehouse database server is the bottom tier of the architecture. It is the relational database system which is uses the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.
- **Middle Tier:** In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.
  - By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
  - By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.
- **Top-Tier:** This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

The following figure 6.4 depicts the three-tier architecture of data warehouse.

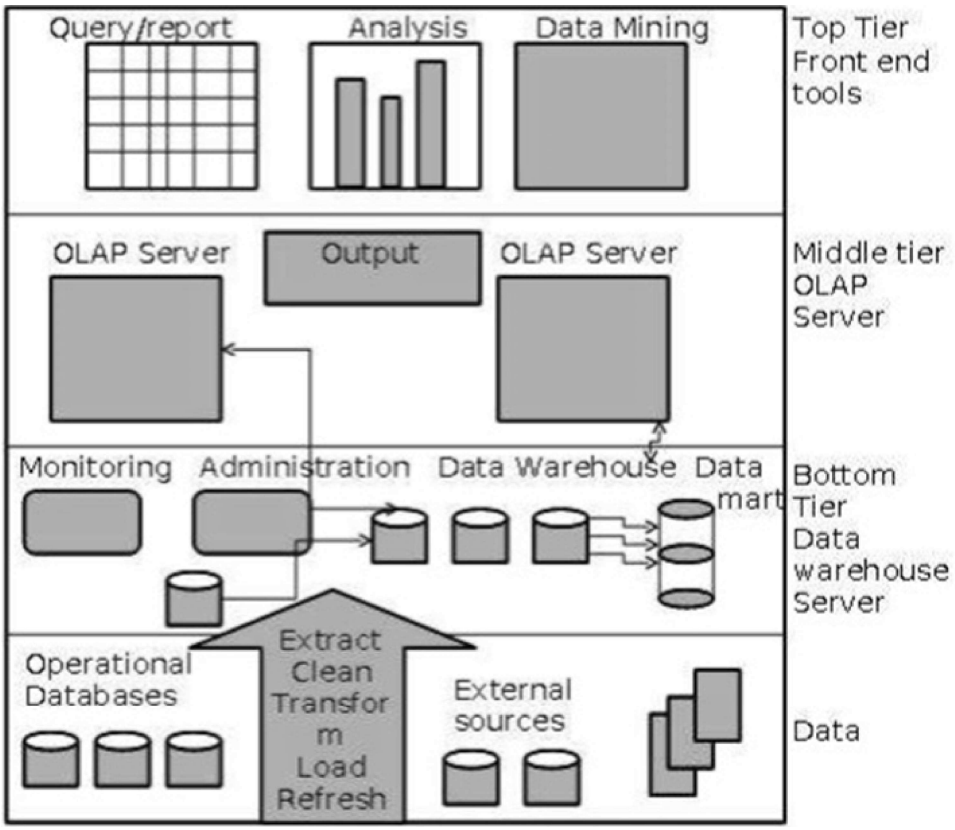


Figure 6.4: Three-tier Architecture of Data Warehouse

6.6 INDEXING AND QUERYING IN OLAP

Indexing the data warehouse can reduce the amount of time to see query results. We can apply indexing on dimensions and on fact table. If too few indexes are applied, the data loads quickly but the query response is slow. If applied indexes are too many, the data loads slowly and your storage requirements go through the roof but the query response is good. Indexing in any database, transactional or warehouse, most often reduces the length of time to see query results. This is especially true with large tables and complex queries that involve in joins operation.

Some of the variables that you'll want to take into account when indexing the data warehouse are the type of data warehouse you have, how large the dimensions and fact tables are, who will be accessing the data and how they'll do so, and whether access will be ad hoc or via structured application interfaces. These variables will determine how indexing scheme should be structured.

Indexing Dimensions: If you want to index the dimension key (primary key), which is not a "natural" or transactional key such as customer name or customer ID where we can not apply clustering.

Indexing the Fact Table: Indexing the fact table is similar to indexing a dimension, although you must account for partitioning.

Modifying Your Indexing Scheme: Over time, you'll have to modify your indexing scheme to show the changes to accommodate what's happening in your organization. And most data warehouse/BI systems will access directly relational tables, so you can use tried-and-true transactional methods for tuning indexes, such as evaluating the query and data mix and adjusting it accordingly.

Querying in OLAP: Online Analytical Processing (OLAP) databases facilitate business-intelligence queries. OLAP is a database technology that has been optimized for querying and reporting, instead of processing transactions. The source data for OLAP is Online Transactional Processing (OLTP) databases stored in data warehouses. OLAP data is derived from this historical data, and aggregated into structures that permit sophisticated analysis for multidimensional structure. OLAP data is also organized hierarchically and stored in cubes. The organization displays high-level summaries using a PivotTable report or PivotChart report, such as sales totals across an entire country or region, and also display the details for sites where sales are particularly strong or weak.

6.7 OLAM

OLAP Mining (OLAM) is an Integration of Data Mining and Data Warehousing–

- On-line analytical mining of data warehouse data is represent as integration of mining and OLAP technologies.
- Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing etc.
- OLAM is Interactive characterization, comparison, association, classification, clustering, prediction.

- Integration of data mining functions, e.g., first clustering and then association.

Importance of OLAM: OLAM is important for the following reasons–

- **High quality of data in data warehouses:** The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in the preprocessing of data. The data warehouses constructed by such preprocessing are valuable sources of high quality data for OLAP and data mining as well.
- **Available information processing infrastructure surrounding data warehouses:** Information processing infrastructure refers to accessing, integration, consolidation, and transformation of multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.
- **OLAP–based exploratory data analysis:** Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various subset of data and at different levels of abstraction.
- **Online selection of data mining functions:** Integrating OLAP with multiple data mining functions and online analytical mining provide users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

6.8 IMPLEMENTATION FROM DATA WAREHOUSING (OLAP) TO DATA MINING (OLAM)

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. The figure 6.5 shows the integration of both OLAP and OLAM–

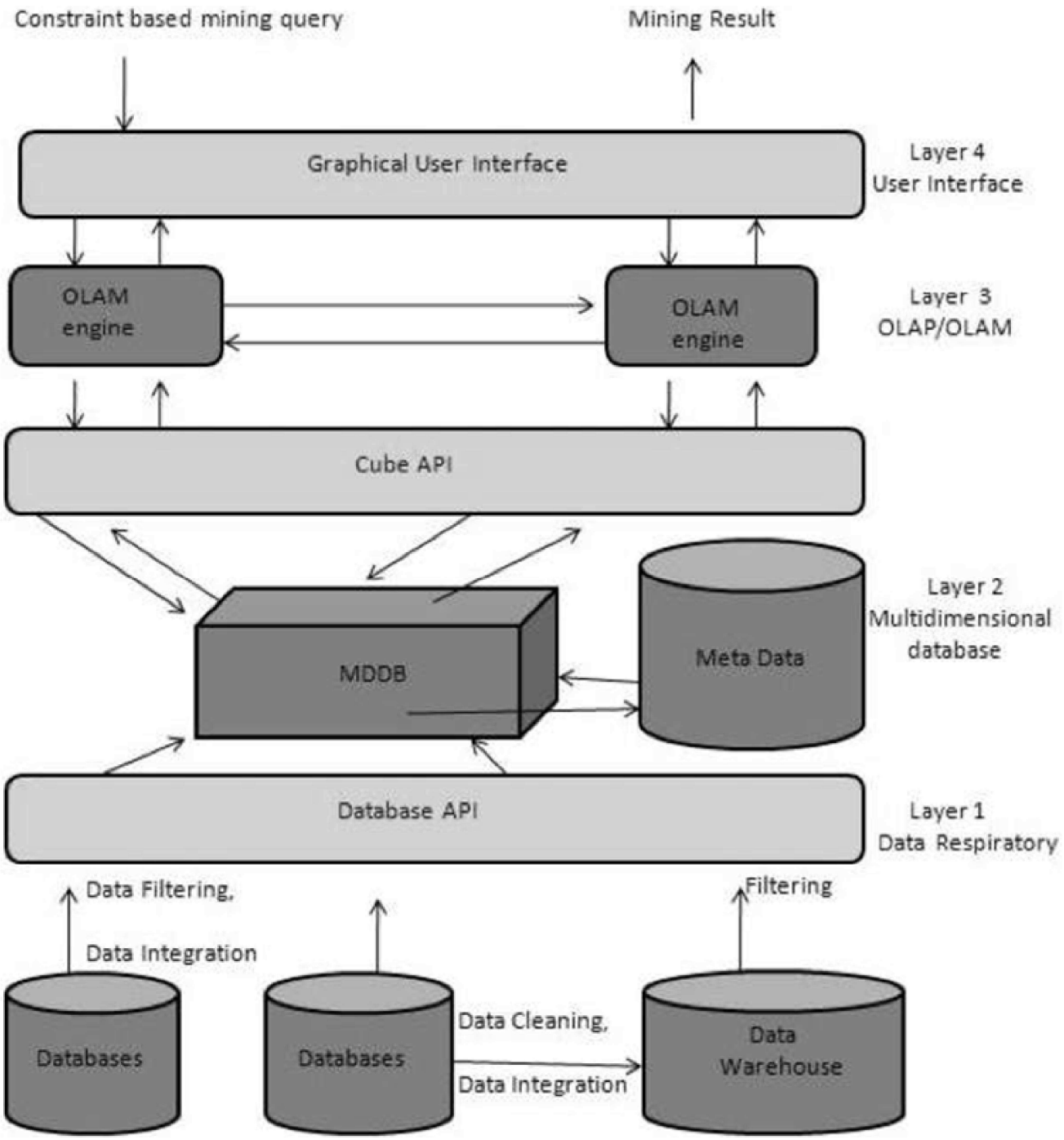


Figure 6.5: Integration of OLAP and OLAM

CHECK YOUR PROGRESS

Q.6: The load and index is–

a) A process to reject data from the data warehouse and to create the necessary indexes.

b) A process to load the data in the data warehouse and to create the necessary indexes.

c) A process to upgrade the quality of data after it is moved into a data warehouse.

d) A process to upgrade the quality of data before it is moved into a data warehouse.

Q.7: The active data warehouse architecture includes–

a) at least one data mart

b) data that can extracted from numerous internal and external sources

c) near real-time updates

d) all of the above.

Q.8: Reconciled data is–

a) data stored in the various operational systems throughout the organization.

b) current data intended to be the single source for all decision support systems.

c) data stored in one operational system in the organization.

d) data that has been selected and formatted for end-user support applications.

Q.9: What are advantages of OLAM?

.....

.....

.....

.....

6.9 LET US SUM UP

- Metadata is simply defined as data about data.
- The operational data warehouse view is known as a virtual warehouse.
- Data mart contains a subset of organizational data.
- Data warehouse design is the process of building a solution to integrate multiple sources data that support analytical reporting and data analysis.
- OLAP is a database technology that has been optimized for querying and reporting, instead of processing transactions.
- The source data for OLAP is Online Transactional Processing (OLTP) databases stored in data warehouses.

6.10 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
2) Ponniah, P. (2011). *Data Warehousing Fundamentals for IT Professionals*. John Wiley & Sons.
3) Pujari, A. K. (2001). Data Mining Techniques. Universities Press.

6.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (c)

Ans. to Q. No. 2: (c)

Ans. to Q. No. 3: (b)

Ans. to Q. No. 4: Extract, Transform and Load

Ans. to Q. No. 5: Requirement gathering, Physical environment setup, Data modeling, ETL, OLAP cube design, Front end development, Report development.

Ans. to Q. No. 6: (b)

Ans. to Q. No. 7: (d)

Ans. to Q. No. 8: (b)

Ans. to Q. No. 9: High quality of data in data warehouses, Available information processing infrastructure surrounding data warehouses, OLAP–based exploratory data analysis, Online selection of data mining functions.

6.12 MODEL QUESTIONS

Q.1: Define Indexing.

Q.2: Write about the different models of data warehouse.

Q.3: Define data warehouse.

Q.4: Write about the component of data warehouse.

Q.5: Define OLAM.

Q.6: Discuss the importance of OLAM.

Q.7: List out the steps of data warehouse design.

Q.8: Describe the three-tier architecture of data warehouse.

Q.9: Define Metadata.

*** ***** ***

UNIT 7: DATA MINING KNOWLEDGE REPRESENTATION

UNIT STRUCTURE

7.1 LEARNING OBJECTS

After going through this unit, you will be able to:

- describe different primitives of data mining task
- represent data and knowledge
- describe basic interestingness measures
- describe different visualization techniques.

7.2 INTRODUCTION

In this previous unit we have learned data warehouse architecture and data warehouse design. We have also learned about OLAP three -tier architecture in detail along with indexing & querying in OLAP, OLAM etc. In this unit, we will learn about data mining tasks and different visualization techniques.

Generally, we use data mining for a long process of research and product development. Also, we can say this evolution was started when business data was first stored on computers. We can also navigate through

their data in real time. Data Mining is also popular in the business community. As this is supported by three technologies that are now mature: Massive data collection, Powerful multiprocessor computers, and Data mining algorithms.

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process. Data mining is based on different task relative primitive like which portion of the transactional database to be mined, what kind of data to be mined, what background knowledge are important to represent the output knowledge, which would be visualize with different techniques. In the next unit, we will explore the concept of attribute generalization, attribute revelance and discuss many statistical measures.

7.3 TASK RELEVANT DATA

Each user will have a data mining task, that is, some form of data analysis that he or she would like to have performed. Data mining query decides the data mining task, which is input to the data mining system. A data mining query is basically defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

Here is the list of data mining task primitives–

- **Set of task relevant data to be mined:** This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).
- **Kind of knowledge to be mined:** This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.
- **Background knowledge to be used in discovery process:** This knowledge about the domain to be mined is useful for guiding the

knowledge discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.

- **Interestingness measures and thresholds for pattern evaluation:** They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.
- **Representation for visualizing the discovered patterns:** This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

7.4 BACKGROUND KNOWLEDGE

Background knowledge consists both of domain-specific knowledge as well as general knowledge about the behavior of the world. Use of background knowledge in the process of identifying general patterns within a database leads to patterns that are more useful and significant. Many data mining problems can be solved better if more background knowledge is added: predictive models can become more accurate, and descriptive models can reveal more interesting endings. Collecting and integrating background knowledge is a manual work.

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in data mining–

1) Descriptive
2) Classification and Prediction

1) Descriptive Function: The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions–

- Class/Concept Description
- Mining of Frequent Patterns

- Mining of Associations
- Mining of Correlations
- Mining of Clusters

a) **Class/Concept Description:** Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways–
   - **Data Characterization:** This refers to summarizing data of class under study. This class under study is called as Target Class.
   - **Data Discrimination:** It refers to the mapping or classification of a class with some predefined group or class.

b) **Mining of Frequent Patterns:** Frequent patterns are occur frequently in transactional data. Here is the list of kind of frequent patterns–
   - **Frequent Item Set:** It refers to a set of items that frequently appear together, for example, milk and bread.
   - **Frequent Subsequence:** A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
   - **Frequent Sub Structure:** Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

c) **Mining of Association:** Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules. For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

d) **Mining of Correlations:** It is a kind of additional analysis performed to uncover interesting statistical correlations between

associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

e) **Mining of Clusters:** Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

2) **Classification and Prediction:** Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms–

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

The list of functions involved in these processes are as follows–

a) **Classification:** It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The derived model is based on the analysis set of training data i.e. the data object whose class label is well known.

b) **Prediction:** It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.

c) **Outlier Analysis:** Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.

d) **Evolution Analysis:** Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

CHECK YOUR PROGRESS

Q.1: Background knowledge referred to–

a) Additional acquaintance used by a learning algorithm to facilitate the learning process.

b) A neural network that makes of a hidden layer.

c) It is form of automatic learning

d) None of these

Q.2: ..... is not a data mining functionality.

a) Clustering and Analysis

b) Selection and Interpretation

c) Classification and Regression

d) Characterization and Discrimination

Q.3: Classification is–

a) A subdivision of a set of examples into a number of classes.

b) A measure of accuracy, of the classification of a concept that is given by certain theory

c) The task of assigning a classification to a set of examples

d) None of these

Q.4: Prediction is–

a) The result of the application of a theory or a rule in a specific case

b) One of several possible enters within a database table that is chosen by the designer as the primary means of accessing the data in the table.

c) Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.

d) None of these

7.5 INTERESTINGNESS MEASURE

Interestingness measures have an important role in data mining, regardless of the kind of patterns being mined. These measures are using

for selecting and ranking patterns according to their potential interest to the user. Good measures also allow the time and space costs of the mining process to be reduced.

Measuring the interestingness of discovered patterns is an active and important area of data mining. Based on the diversity of definitions presented to-date, interestingness is perhaps best treated as a broad concept that emphasizes conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility, and actionability. These nine specific criteria are used to determine whether or not a pattern is interesting. They are described as follows:

- **Conciseness:** If pattern contains relatively few attribute-value pairs then it is concise, while a set of patterns is concise if it contains relatively few patterns. A concise pattern or set of patterns is relatively easy to understand and remember and thus is added more easily to the user's knowledge
- **Generality/Coverage:** If a pattern covers a relatively large subset of a dataset then it is general. Generality (or coverage) measures the comprehensiveness of a pattern, that is, the fraction of all records in the dataset that matches the pattern. If a pattern characterizes more information, it tends to be more interesting. Frequent item sets are the most studied general patterns in the data mining literature. Generality frequently coincides with conciseness because concise patterns tend to have greater coverage.
- **Reliability:** A pattern is reliable if the relationship described by the pattern occurs in a high percentage of applicable cases. For example, a classification rule is reliable if its predictions are highly accurate, and an association rule is reliable if it has high confidence.
- **Peculiarity:** A pattern is peculiar if it is far away from other discovered patterns according to some distance measure. Peculiar patterns are generated from peculiar data (or outliers), which are relatively few in number and significantly different from the rest of the data
- **Diversity:** A pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if the patterns in the set

differ significantly from each other. Diversity is a common factor for measuring the interestingness of summaries.

- **Novelty:** A pattern is novel to a person if he or she did not know it before and is not able to infer it from other known patterns. No known data mining system represents everything that a user knows, and thus, novelty cannot be measured explicitly with reference to the user's knowledge.
- **Surprisingness:** A pattern contradicts a person's existing knowledge or expectation is termed as surprising (or unexpected). A pattern that is an exception to a more general pattern which has already been discovered can also be considered surprising. Surprising patterns are interesting because they identify failings in previous knowledge and may suggest an aspect of the data that needs further study.
- **Utility:** A pattern uses by a person contributes to reaching a goal is called utility. Different people have different goals concerning the knowledge that can be extracted from a dataset. This kind of interestingness is based on user-defined utility functions in addition to the raw data.
- **Actionability/Applicability:** A pattern is actionable in some domain if it enables decision making about future actions in this domain. Action ability is sometimes associated with a pattern selection strategy.

7.6 REPRESENTING INPUT DATA AND OUTPUT KNOWLEDGE

i) **Concept:** This concept is introduced as what things are to be mined using following categories of mining:

- **Classification mining/learning:** predicting a discrete class, a kind of supervised learning, success is measured on new data for which class labels are known (test data).
- **Association mining/learning:** detecting associations between attributes, can be used to predict any attribute value and more than one attribute values, hence more rules can be generated, therefore we need constraints.

- **Clustering:** grouping similar instances into clusters, a kind of unsupervised learning, success is measured subjectively or by objective functions.
- **Numeric prediction:** predicting a numeric quantity, a kind of supervised learning, success is measured on test data.
- **Concept description:** output of the learning scheme.

ii) **Instance:** Instances are defined as what things to be classified, associated, or clustered. Individual and independent examples of the concept to be learned (target concept). Instance is described by predetermined set of attributes. Input to the learning scheme is defined as set of instances (dataset), represented as a single relation (table), independence assumption, positive and negative examples are taking for a concept.

iii) **Attributes:** Attributes (features) of input data are predefined set of features to describe an instance. They are nominal (distinct and no relation between them), structured and numeric.

iv) **Output knowledge:** Output knowledge is represented as the output of Association rules, Classification rules, Rules with relation, used the different prediction schemes like nearest neighbor, Bayesian classification, Neural networks, Regression. And output are represented as decision trees where knowledge is portioned as cluster on the basis of structure, concept or statistics.

7.7 VISUALIZATION TECHNIQUE

Visual Data Mining is the process of discovering implicit but useful knowledge from large data sets using visualization techniques.

Data visualization aims to communicate data clearly and effectively through graphical representation. Data visualization has been used extensively in many applications. For e.g., at work for reporting managing business operations and tracking progress of tasks. More popularly, we can take advantage of visualization techniques to discover data relationships that are otherwise not easily observable by looking at the raw data.

Different Data Visualization techniques are as follows:

1) **Pixel oriented visualization techniques:**

- A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value.
- For a data set of m dimensions pixel oriented techniques create m windows on the screen, one for each dimension.
- The m dimension values of a record are mapped to m pixels at the corresponding position in the windows.
- The color of the pixel reflects other corresponding values.
- Inside a window, the data values are arranged in some global order shared by all windows
- Example: All Electronics maintains a customer information table, which consists of 4 dimensions: income, credit_limit, transaction_volume and age. We analyze the correlation between income and other attributes by visualization.
- We sort all customers in income in ascending order and use this order to layout the customer data in the 4 visualization windows as shown in figure 7.1.
  - The pixel colors are chosen so that the smaller the value, the lighter the shading.

Using pixel based visualization we can easily observe that credit_limit increases as income increases customer whose income is in the middle range are more likely to purchase more from All Electronics, there is no clear correlation between income and age.

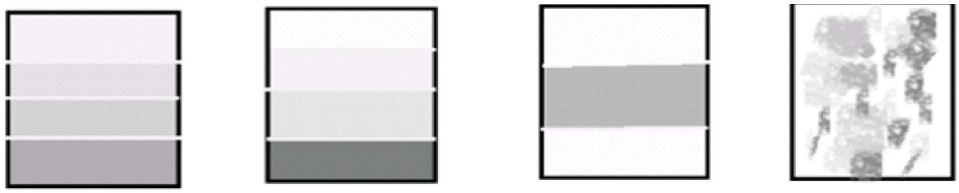


Figure 7.1: Pixel oriented visualization of 4 attributes by sorting all customers in income Ascending order

2) Geometric Projection visualization techniques:

- A drawback of pixel-oriented visualization techniques is that they cannot help us much in understanding the distribution of data in a multidimensional space.
- Geometric projection techniques help users find interesting projections of multidimensional data sets.
- A scatter plot displays 2-D data point using Cartesian co-ordinates. A third dimension can be added using different colors of shapes to represent different data points.
- Eg: Where x and y are two spatial attributes and the third dimension is represented by different shapes.
  Through this visualization, we can see that points of types "+" & "X" tend to be collocated.

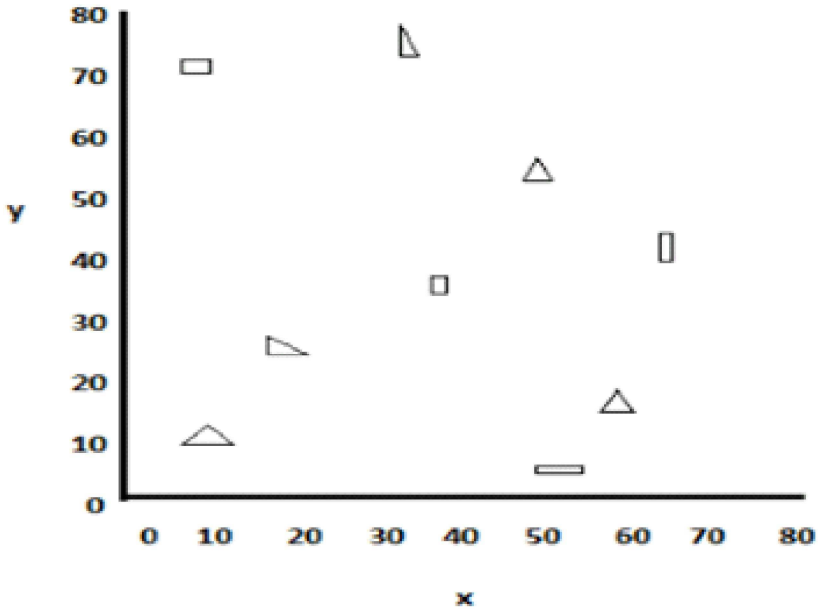


Figure 7.2: Visualization of 2D data set using scatter plot

3) Icon based visualization techniques:

- It uses small icons to represent multidimensional data values.
- Two popular icon based techniques are listed below:
  - **Chernoff faces:** it was introduced in 1973 by Herman Chernoff. They display multidimensional data of up to 18 variables as a cartoon human face. Chernoff faces helps to reveal trends in data. Component of face.

- **Stick figures:** It maps multidimensional data to five-piece stick figure, where each figure has 4 limbs and a body. Two dimensions are mapped to the display axes and the remaining dimensions are mapped to the angle and/or length of the limbs.

Figure 7.3: Chernoff faces each face represents an 'n' dimensional data points (n<18)

4) **Hierarchical visualization techniques: (i.e. subspaces):** These techniques focus on visualizing multiple dimensions simultaneously. A large data set of high dimensionality, it would be difficult to visualize all dimensions at the same time. Hierarchical visualization technique makes subset of the dimensions. "Worlds-within-Worlds" also known as n-vision is representing by Hierarchical visualization technique.
5) **Visualizing Complex data and relation:** There are many new techniques dedicated to non-numerical data. For example, many people on the web tag, blog entries and product reviews. A **tag cloud** is a visualization of statistics of user generated tag where tags are listed alphabetically or user preferred order. Important tags are listed with color. In addition, relation among complex data entries also raises challenges for visualization.

CHECK YOUR PROGRESS

Q.6: An objective measure of pattern interestingness in data mining is/are:

a) Support rule b) Confidence rule

c) Both (a) & (b) d) Neither (a) nor (b)

Q.7: Explain Syntax for Interestingness Measures Specification.

.....

.....

.....

Q.8: Explain Syntax for Pattern Presentation and Visualization Specification.

.....

.....

.....

Q.9: What is tag cloud?

.....

.....

7.8 LET US SUM UP

- Different data mining tasks are the core of data mining process. Different prediction and classification data mining tasks actually extract the required information from the available data sets.
- Background knowledge are either domain specific or based on general knowledge.
- Data mining functions are classified into two main categories: Descriptive, Classification and prediction.
- Interestingness measure is not depends on kind of pattern being mind. Good measure reduces the time and space cost.

- Input data are introduced on the basis of classification, association, clustering and numerical prediction.
- Large data can be easily represent with different visualization techniques.

7.9 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
2) Pujari, A. K. (2001). Data Mining Techniques. Universities Press.

7.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (a)

Ans. to Q. No. 2: (b)

Ans. to Q. No. 3: (a)

Ans. to Q. No. 4: (a)

Ans. to Q. No. 6: (c)

Ans. to Q. No. 7: Interestingness measures and thresholds can be specified by the user with the statement–
with <interest_measure_name> threshold = threshold_value.

Ans. to Q. No. 8: Generally, we have a syntax, which allows users to specify the display of discovered patterns in one or more forms. display as <result_form>

Ans. to Q. No. 9: A tag cloud is a visualization of statistics of user generated tag where tags are listed alphabetically or user preferred order.

7.11 MODEL QUESTIONS

Q.1: Name some data mining techniques?

Q.2: What is the foundation of data mining?

Q.3: Why is background knowledge required for data mining?

Q.4: What are the task related primitives used in data mining?

Q.5: What is input data for data mining?

Q.6: What are attributes used for representing the input data in data mining?

Q.7: Describe the different aspects of the interestingness measures.

Q.8: Write about the different techniques to visualize large data.

*** ***** ***

UNIT 8: ATTRIBUTE-ORIENTED ANALYSIS

UNIT STRUCTURE

8.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- define attribute generalization and attribute relevance
- describe data cube approach
- describe class comparison
- describe the different statistical measures.

8.2 INTRODUCTION

In the previous unit we have discussed topics like data, knowledge and different data visualization techniques. In this unit, we will learn about attribute generalization and attribute relevance along with class comparison in detail. We will also explore the different statistical measures like mean, median, mode etc in detail in this unit.

Data mining usually says about knowledge discovery from data. To know about the data it is necessary to go through the data objects, data attributes and types of data attributes. Mining data also includes relation between data. Data objects are the essential part of a database. A data object represents the entity. Data Objects are like group of attributes of a

entity. For example, a sales data object may represent customer, sales or purchases. When a data object is listed in a database they are called data tuples. A set of attributes used to describe a given object are known as attribute vector. Attributes are mainly categorized as qualitative and quantitative.

In general, *data generalization* summarizes data by replacing relatively low-level values (e.g., numeric values for an attribute *age*) with higher-level concepts (e.g., *young*, *middle-aged*, and *senior*), or by reducing the number of dimensions to summarize data in concept space involving fewer dimensions.

In many applications, users may not be interested in having a single class (or concept) described or characterized, but prefer to mine a description that compares or distinguishes one class (or concept) from other comparable classes (or concepts). Class discrimination or comparison (hereafter referred to as **class comparison**) mines descriptions that distinguish a target class from its contrasting classes. We start with *measures of* ***central tendency***, which measure the location of the middle or center of a data distribution. The most common data dispersion measures are the *range*, *quartiles*, and *inter-quartile range*; the *five-number summary* and *boxplots* and the *variance* and *standard deviation* of the data. Most statistical or graphical data presentation software packages include bar charts, pie charts, and line graphs. Other popular displays of data summaries and distributions include *quantile plots*, *quantile-quantile plots*, *histograms*, and *scatter plots*. In the next unit, in the next block we will explore the concept of association rule mining in detail.

8.3 ATTRIBUTE GENERALIZATION

8.3.1 Attribute

It can be seen as a data field that represents characteristics or features of a data object. For a customer object, attributes can be customer Id, address etc. We can say that a set of attributes used to describe a given object are known as attribute vector or feature vector.

Type of attributes: There are two different types of attributes. The attribute types are:

1) Qualitative [Nominal (N), Ordinal (O), Binary (B)].
2) Quantitative (Discrete, Continuous)

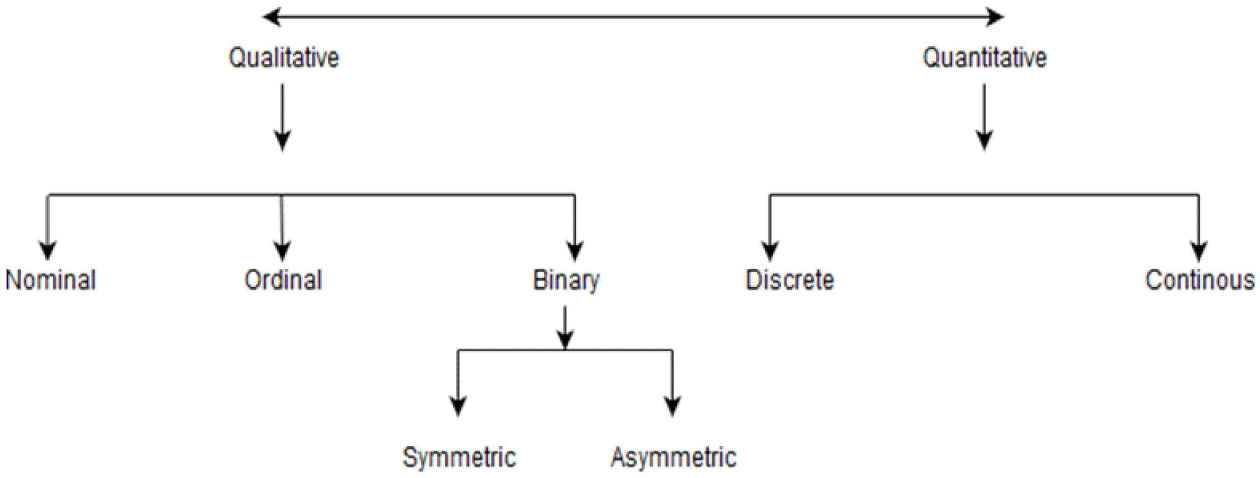


Figure 8.1: Different Types of Attributes

Nominal Attributes-related to names: The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represents some category or state. So these attribute are referred as **categorical attributes** and there is no order (rank, position) among values of nominal attribute. E.g., black, blue are the value of color attribute.

Binary Attributes: Binary data has only two values or states. For example, yes or no, affected or unaffected, true or false.

i) **Symmetric:** Both values are equally important (Gender).

ii) **Asymmetric:** Both values are not equally important (Result).

Ordinal Attributes: The Ordinal Attributes contain values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is. E.g. grade values are A, B, C, D, E.

Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of two types, **interval** and **ratio**.

i) An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point or we can call zero point. Data can be added and subtracted at interval scale but cannot be multiplied or divided. Temperature of two days not comparable.

ii) A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. Mean, median values are the example of this type.

- **Discrete:** Discrete data have finite values it can be numerical and can also be in categorical form. E.g. teacher, business man, peon are the value of profession attribute.
- **Continuous:** Continuous data have infinite no of states. Continuous data is of float type. There can be many values between 2 and 3. E.g., 5.4, 6.3 are the values of height attribute. For data generalization, there are two approaches namely: data cube(OLAP) approach and attribute oriented induction approach. The general idea behind attribute relevance analysis is to compute some measure which is used to quantify the relevance of an attribute with respect to a given class.

8.3.2 Attribute Generalization

Conceptually, the data cube can be viewed as a kind of multidimensional data generalization. In general, *data generalization* summarizes data by replacing relatively low-level values (e.g., numeric values for an attribute *age*) with higher-level concepts (e.g., *young*, *middle-aged*, and *senior*), or by reducing the number of dimensions to summarize data in concept space involving fewer dimensions (e.g., removing *birth date* and *telephone number* when summarizing the behavior of a group of students). Given the large amount of data stored in databases, it is useful to be able to describe concepts in concise and succinct terms at generalized (rather than low) levels

of abstraction. Data generalization is a process that abstracts a large set of task relevant data in a database from relatively conceptual level to high conceptual levels. The generalization of large data sets can be categorized according to two approaches.

1) The data cube(OLAP) approach
2) The attribute oriented induction approach

The Data Cube Approach: For example, *All Electronics* database, sales managers may prefer to view the data generalized to higher levels, such as summarized by customer groups according to geographic regions, frequency of purchases per group, and customer income.

This leads us to the notion of *concept description*, which is a form of data generalization. A concept typically refers to a data collection such as *frequent_buyers, graduate_students*, and so on. **Concept description** generates descriptions for data *characterization* and *comparison*. When concept refers to a class, it is called **class description**. **Characterization** provides a concise and succinct summarization of the given data collection.

We have studied data cube (or OLAP) approaches to concept description using multidimensional, multilevel data generalization in data warehouses. *"Is data cube technology sufficient to accomplish all kinds of concept description tasks for large data sets?"* Consider the following cases.

- **Complex data types and aggregation:** Data warehouses and OLAP tools are based on a multidimensional data model that views data in the form of a data cube, consisting of dimensions (or attributes) and measures (aggregate functions). Furthermore, the aggregation of attributes in a database may include sophisticated data types such as the collection of non-numeric data, the merging of spatial regions, the composition of images, the integration of texts, and the grouping of object pointers. Therefore, OLAP, with its restrictions on the possible dimension and measure types, represents a simplified model for data

analysis. Concept description should handle complex data types of the attributes and their aggregations, as necessary.

- **User control versus automation:** Online analytical processing in data warehouses is a user-controlled process. The selection of dimensions and the application of OLAP operations (e.g., drill-down, roll-up, slicing, and dicing) are primarily directed and controlled by users. The control in most OLAP systems is quite user-friendly. Furthermore, in order to find a satisfactory description of the data, users may need to specify a long sequence of OLAP operations.

  This section presents an alternative method for concept description, called *attribute-oriented induction*, which works for complex data types and relies on a data-driven generalization process.

Attribute-Oriented Induction for Data Characterization: The **attribute-oriented induction (AOI)** approach to concept description was first proposed in 1989, a few years before the introduction of the data cube approach. The data cube approach is essentially based on *materialized views* of the data, which typically have been pre-computed in a data warehouse. In general, it performs offline aggregation before an OLAP or data mining query is submitted for processing. On the other hand, the attribute-oriented induction approach is basically a *query-oriented*, generalization-based, online data analysis technique.

The general idea of attribute-oriented induction is to first collect the task-relevant data using a database query and then perform generalization based on the examination of the number of each attribute's distinct values in the relevant data set. This process is called **data focusing.** Then generalization is performed either by *attribute removal* or *attribute generalization*. Aggregation is performed by merging identical generalized tuples and accumulating their respective counts.

Attribute removal is based on the following rule: *If there is a large set of distinct values for an attribute of the initial working relation, but either (case 1) there is no generalization operator on the attribute (e.g., there is no concept hierarchy defined for the attribute), or (case 2) its higher-level concepts are expressed in terms of other attributes, then the attribute should be removed from the working relation*.

Attribute generalization is based on the following rule: *If there is a large set of distinct values for an attribute in the initial working relation, and there exists a set of generalization operators on the attribute, then a generalization operator should be selected and applied to the attribute*.

Both rules– *attribute removal* and *attribute generalization* – claim that if there is a *large* set of distinct values for an attribute, further generalization should be applied. This raises the question: How large is *"a large set of distinct values for an attribute"* considered to be?

Depending on the attributes or application involved, a user may prefer some attributes to remain at a rather low abstraction level while others are generalized to higher levels. The control of how high an attribute should be generalized is typically quite subjective. The control of this process is called **attribute generalization control**. There are many possible ways to control a generalization process. We will describe here two common approaches. The first technique, called **attribute generalization threshold control**, either sets one generalization threshold for all of the attributes, or sets one threshold for each attribute. If the number of distinct values in an attribute is greater than the attribute threshold, further attribute removal or attribute generalization should be performed. Data mining system have a default attribute threshold value ranging from 2 to 8. The second technique, called **generalized relation threshold control**, sets a threshold for the generalized relation. If the number of (distinct) tuples in the generalized relation is greater than the threshold, further generalization should be performed. Otherwise,

no further generalization should be performed. For data mining system, this threshold value is ranging from 10 to 30.

8.4 ATTRIBUTE RELEVANCE

The general idea behind attribute relevance analysis is to compute some measure which is used to quantify the relevance of an attribute with respect to a given class. Such measures include the information gain, gini index, uncertainty, and correlation coefficients.

Let S be a set of training object (or tuple) where the class label of each tuple is known. Suppose that there are m classes. Let S contain S_i objects of class C_i, for i = 1, ..., m. An arbitrary object belongs to class C_i with probability S_i/s, where s is the total number of objects in set S. The expected information needed to classify given tuple is:

$$I(s_1, s_2, ..., ..., ..., s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s} \qquad \textbf{(8.1)}$$

If an attribute A with values {a1, a2, ..., av} is used to partition S into the subsets {S1, S2, ..., Sv}, where S_j contains those objects in S that have value a_j of A. Let S_j contain S_{ij} objects of class C_i. The expected information based on this partitioning by A is known as the entropy of A. It is the weighted average:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j} + ... + s_{mj}) \qquad \textbf{(8.2)}$$

The information gained by branching on A is defined by:

$$Gain(A) = I(s_1, s_2, ..., ..., ..., s_m) - E(A) \qquad \textbf{(8.3)}$$

The attribute which maximizes gain(A) is selected. Attribute relevance analysis for class description is performed as follows.

1) **Data Collection:** Collect data for both the target class and the contrasting class by query processing. Notice that for class comparison, both the target class and the contrasting class are provided by the user in the data mining query. For class characterization, the target class is the class to be characterized, whereas the contrasting class is the set of comparable data which are not in the target class.

2) **Preliminary Relevance analysis using conservative AOI:** Attribute-oriented induction (AOI) can be used to perform some preliminary relevance analysis on the data by removing or generalizing attributes having a large number of distinct values (such as name and phone#). Such attributes are unlikely to be meaningful for concept description. To be conservative, the AOI should employ attribute generalization thresholds that are set reasonably large. (so as to allow more attributes to be considered in further relevance analysis by selected measure performed in step-3). The relation obtained by such an attribute removal and attribute generalization process is called the candidat relation of the mining task.

3) **Remove irrelevant or weakly relevant attributes using the selected measure:** The selected relevance measure is used to evaluate (or rank) each attribute in the candidate relation. For example, the information gain measure described above may be used. The attributes are then sorted (i.e., ranked) according to their computed relevance measure value. Attribute that are not relevant or weakly relevant are then removed based on the set threshold. The resulting relation is called "Initial Target/Contrast class Working Relation".

CHECK YOUR PROGRESS

Q.1: Identify the example of nominal attribute?

a) Gender b) Temperature

c) Mass d) Salary

Q.2: What is Attribute removal?

.....

.....

.....

.....

.....

Q.3: What is data focusing?

.....

.....

.....

.....

Q.4: Name the process of data relevance analysis.

.....

.....

.....

Q.5: What is nominal attribute?

.....

.....

.....

8.5 CLASS COMPARISON

In many applications, users may not be interested in having a single class (or concept) described or characterized, but prefer to mine a description that compares or distinguishes one class (or concept) from other comparable classes (or concepts). Class discrimination or comparison (hereafter referred to as **class comparison**) mines descriptions that distinguish a target class from its contrasting classes. Notice that the target and contrasting classes must be *comparable* in the sense that they share similar dimensions and attributes. For example, the three classes *person, address*, and *item* are not comparable. However, sales in the last three years are comparable classes, and so are, for example, computer science students versus physics students.

Suppose, for instance, that we are given the *All Electronics* data for sales in 2009 and in 2010 and want to compare these two classes. Consider the dimension *location* with abstractions at the *city*, *province_or_state*, and *country* levels. Data in each class should be generalized to the same *location* level. *"How is class comparison performed?"* In general, the procedure is as follows:

1) **Data collection:** The set of relevant data in the database is collected by query processing and is partitioned respectively into a *target class* and one or a set of *contrasting classes*.
2) **Dimension relevance analysis:** If there are many dimensions, then dimension relevance analysis should be performed on these classes to select only the highly relevant dimensions for further analysis. Correlation or entropy-based measures can be used for this step.
3) **Synchronous generalization:** Generalization is performed on the target class to the level controlled by a user-or expert-specified dimension threshold, which results in a **prime target class relation**. The concepts in the contrasting class(es) are generalized to the same level as those in the prime target class relation, forming the **prime contrasting class(es) relation**.
4) **Presentation of the derived comparison:** The resulting class comparison description can be visualized in the form of tables, graphs, and rules. This presentation usually includes a "contrasting" measure such as count% (percentage count) that reflects the comparison between the target and contrasting classes. The user can adjust the comparison description by applying drill-down, roll-up, and other OLAP operations to the target and contrasting classes, as desired.

8.6 STATISTICAL MEASURES

For a successful data preprocessing, it is essential to have an overall picture of your data. Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

We start with *measures of **central tendency***, which measure the location of the middle or center of a data distribution. The most common data dispersion measures are the *range*, *quartiles*, and *inter-quartile range*; the *five-number summary* and *boxplots* and the *variance* and *standard deviation* of the data. Most statistical or graphical data presentation software packages include bar charts, pie charts, and line graphs. Other

popular displays of data summaries and distributions include *quantile plots*, *quantile–quantile plots*, *histograms*, and *scatter plots*.

Measuring the central tendency: mean, median, and mode

There have various ways to measure the central tendency of data. Suppose that we have some attribute *X*, like *salary*, which has been recorded for a set of objects. Let x_1, x_2, ..., ..., ..., x_N be the set of *observations* for *X*. Measures of central tendency include the mean, median, mode, and midrange.

The most common and effective numeric measure of the "center" of a set of data is the *(arithmetic) mean*. Let be a set of *N* values or *observations*, such as for some numeric attribute *X*, like *salary*. The **mean** of this set of values is:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_i + x_2 + \ldots\ldots\ldots + x_N}{N} \qquad \textbf{(8.4)}$$

This corresponds to the built-in aggregate function, *average* (avg() in SQL), provided in relational database systems.

Example 8.1: Mean

Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Using Eq. (8.4), we have,

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$

$$= \frac{696}{12} = 58$$

Thus, the mean salary is $58,000.

Sometimes, each value x_i in a set may be associated with a weight w_i. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots\ldots\ldots + w_N x_N}{w_1 + w_2 + \ldots\ldots\ldots + w_N} \qquad \textbf{(8.5)}$$

This is called the **weighted arithmetic mean** or the **weighted average**.

For skewed (asymmetric) data, a better measure of the center of data is the **median**, which is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half.

Example 8.2: Median

Let's find the median of the data from Example 8.1. The data are already sorted in increasing order. There is an even number of observations (i.e., 12); therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list). By convention, we assign the average of the two middlemost values as the median; that is the median is \$54,000.

Suppose that we had only the first 11 values in the list. Given an odd number of values, the median is the middlemost value. This is the sixth value in this list, which has a value of \$52000.

We can approximate the median of the entire data set (e.g., the median salary) by interpolation using the formula,

$$\text{median} = L_1 + \left(\frac{N/2 - (\sum \text{freq})_1}{\text{freq}_{\text{median}}} \right) \text{width} \qquad \textbf{(8.6)}$$

where, L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and *width* is the width of the median interval.

The *mode* is another measure of central tendency. The **mode** for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes. Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**. In general, a data set with two or more modes is **multimodal**.

Example 8.3: Mode

The data from Example 8.1 are bimodal. The two modes are \$52,000 and \$70,000.

For unimodal numeric data that are moderately skewed (asymmetrical), we have the following empirical relation:

mean – mode $\approx$ 3 × (mean – median) **(8.7)**

This implies that the mode for unimodal frequency curves that are moderately skewed can easily be approximated if the mean and median values are known.

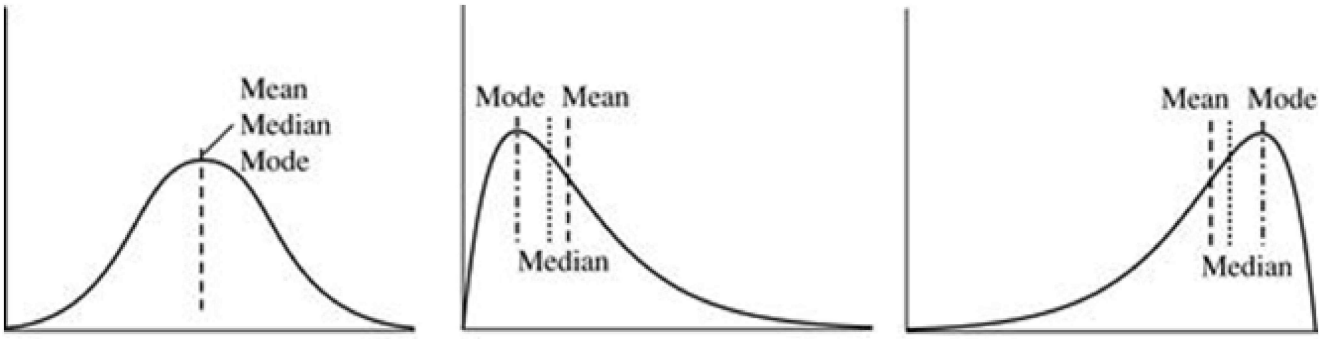
The **midrange** can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set.

Example 8.4: Midrange

The midrange of the data of Example 8.1 is,

$$\frac{30{,}000 + 110{,}000}{2} = \$70{,}000$$

In a unimodal frequency curve with perfect **symmetric** data distribution, the mean, median, and mode are all at the same center value, as shown in Figure 8.1(a).



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed date

Figure 8.2: Mean, median, and mode of symmetric versus positively and negatively skewed data.

Data in most real applications are not symmetric. They may instead be either **positively skewed**, where the mode occurs at a value that is smaller than the median (Figure 8.2b), or **negatively skewed**, where the mode occurs at a value greater than the median (Figure 8.2c).

Measuring the dispersion of data: range, quartiles, variance, standard deviation, and interquartile range: We now look at measures to assess the dispersion or spread of numeric data. The measures include range, quantiles, quartiles, percentiles, and the interquartile range. The five-

number summary, which can be displayed as a boxplot, is useful in identifying outliers. Variance and standard deviation also indicate the spread of a data distribution.

Range, Quartiles, and Interquartile Range: To start off, let's study the *range*, *quantiles*, *quartiles*, *percentiles*, and the *interquartile* range as measures of data dispersion. Let $x_1, x_2, \ldots, ..., ..., x_N$ be a set of observations for some numeric attribute, X. The range of the set is the difference between the largest (max()) and smallest (min()) values.

Suppose that the data for attribute *X* are sorted in increasing numeric order. Imagine that we can pick certain data points so as to split the data distribution into equal-size consecutive sets, as in figure 8.3. These data points are called *quantiles*. **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.

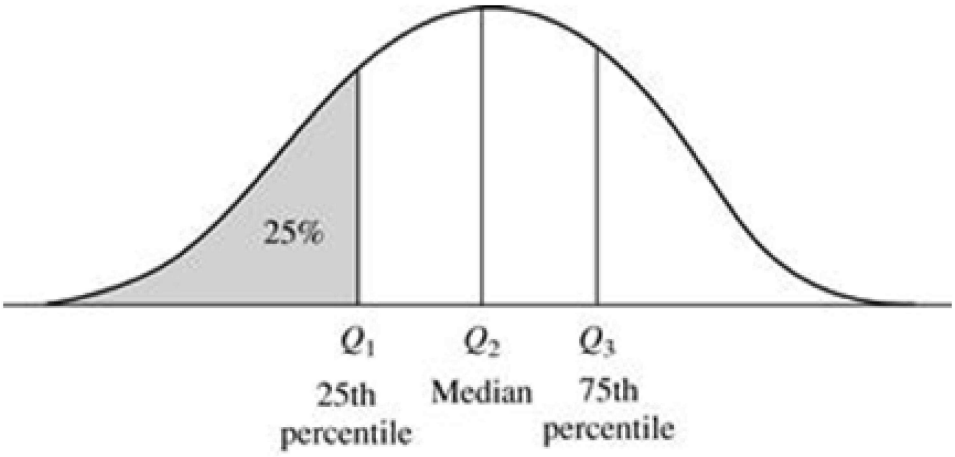


Figure 8.3: A plot of the data distribution for some attribute *X*

The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median. The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**. The 100-quantiles are more commonly referred to as **percentiles.** The **first quartile**, denoted by Q_1, is the 25^{th} percentile. It cuts off the lowest 25% of the data. The **third quartile**, denoted

by Q_3, is the 75th percentile– it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **interquartile range** (**IQR**) and is defined as–

$$IQR = Q_1 - Q_3 \quad \textbf{(8.8)}$$

Example 8.5: Interquartile range

The quartiles are the three values that split the sorted data set into four equal parts. The data of Example 8.1 contain 12 observations, already sorted in increasing order. Thus, the quartiles for this data are the third, sixth, and ninth values, respectively, in the sorted list. Therefore, Q_1 = \$47,000 and Q_3 is \$63,000. Thus, the interquartile range is $IQR = 63 - 47$ = \$16,000.

Five-Number Summary, Boxplots, and Outliers: In the symmetric distribution, the median (and other measures of central tendency) splits the data into equal-size halves. This does not occur for skewed distributions. Therefore, it is more informative to also provide the two quartiles Q_1 and Q_3, along with the median. A common rule of thumb for identifying suspected **outliers** is to single out values falling at least 1.5 × IQR above the third quartile or below the first quartile.

Because Q_1, the median, and Q_3 together contain no information about the endpoints (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the *five-number summary*. The **five-number summary** of a distribution consists of the median (Q_2), the quartiles Q_1 and Q_3, and the smallest and largest individual observations, written in the order of *Minimum*, Q_1, *Median*, Q_3, *Maximum*.

Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:

- Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
- The median is marked by a line within the box.

- Two lines (called *whiskers*) outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations.

For exambple, let us take *All Electronics* data during a given time period. For branch 1, we see that the median price of items sold is \$80, Q_1 is \$60, and Q_3 is \$100. Notice that two outlying observations for this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR here of 40.

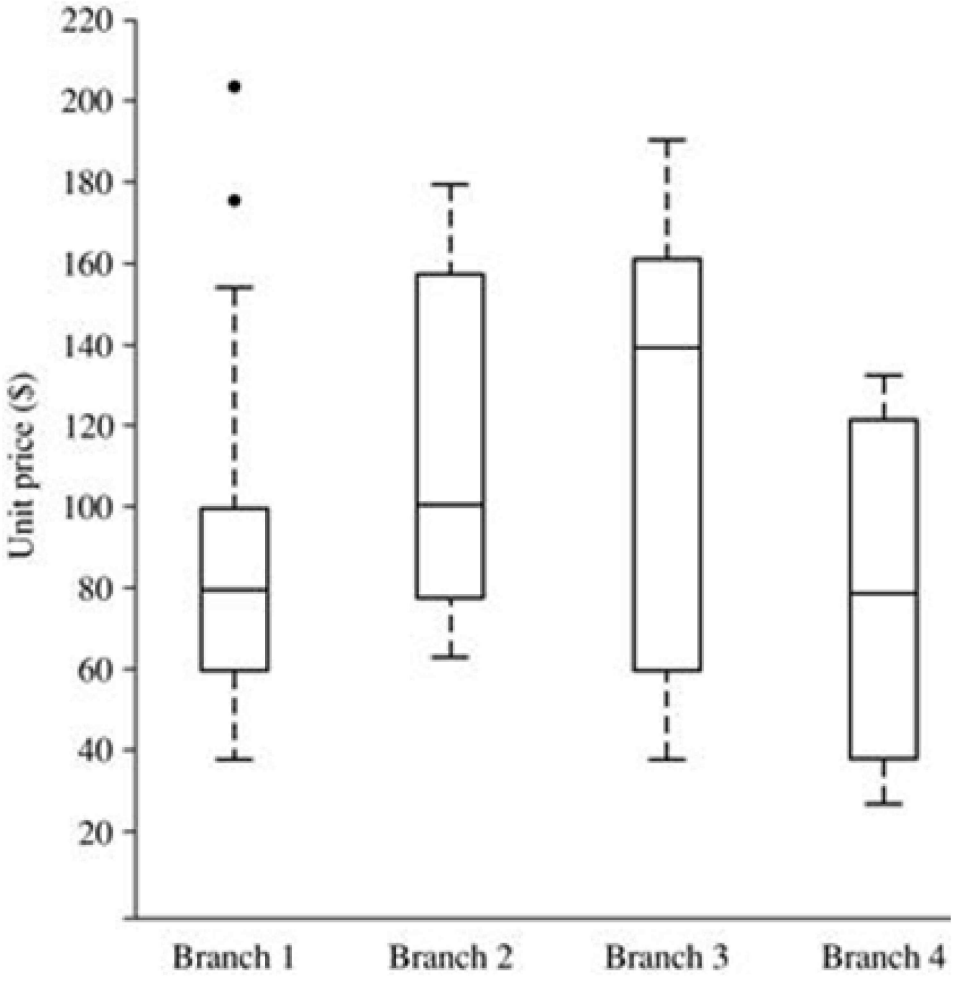


Figure 8.4: Boxplot for the unit price data for items sold at four branches of All Electronics during a given time period

Variance and Standard Deviation: Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

The **variance** of *N* observations, x_1, x_2, ..., ..., ..., x_N, for a numeric attribute *X* is:

$$\sigma^{2=1/N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2=\left(\frac{1}{N}\sum_{i=1}^{N}x_2{}^2\right)-\bar{x}^2 \qquad \textbf{(8.9)}$$

where, $\bar{x}$ is the mean value of the observations. The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

The basic properties of the standard deviation, σ, as a measure of spread are as follows:

- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

The standard deviation is a good indicator of the spread of a data set. The computation of the variance and standard deviation is scalable in large databases.

Graphic displays of basic statistical descriptions of data: These include *quantile plots*, *quantile-quantile plots*, *histograms*, and *scatter plots*. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

Quantile Plot: A **quantile plot** is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, it plots quantile information. Let x_i, for $i = 1$ to N, be the data sorted in increasing order so that x_1 is the smallest observation and x_N is the largest for some ordinal or numeric attribute X.

Let, fi = (i – 0.5) / N **(8.10)**

These numbers increase in equal steps of 1/*N*, ranging from $\frac{1}{2N}$ (which is slightly above 0) to $1-\frac{1}{2N}$ (which is slightly below 1). On a quantile plot, x_i is graphed against f_i. This allows us to compare different distributions based on their quantiles. For example, given the quantile plots of sales data

for two different time periods, we can compare their Q_1, median, Q_3, and other f_i values at a glance.

Figure 8.5 shows a quantile–quantile plot for *unit price* data of items sold at two branches of *All Electronics* during a given time period. Each point corresponds to the same quantile for each data set and shows the unit price of items sold at branch 1 versus branch 2 for that quantile.

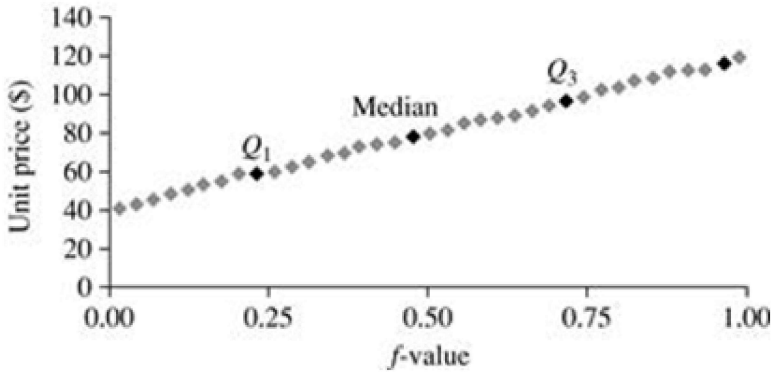


Figure 8.5: A quantile plot for the unit price data of Table 8.1.

Table 8.1: A Set of Unit Price Data for Items Sold at a Branch of All Electronics

Unit Price ($)	Count of Units Sold
40	2765
43	300
47	250
–	–
74	360
75	515
78	540
–	–
115	320
117	270
120	350

Quantile-Quantile Plot: A **quantile-quantile plot**, or **q-q plot**, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

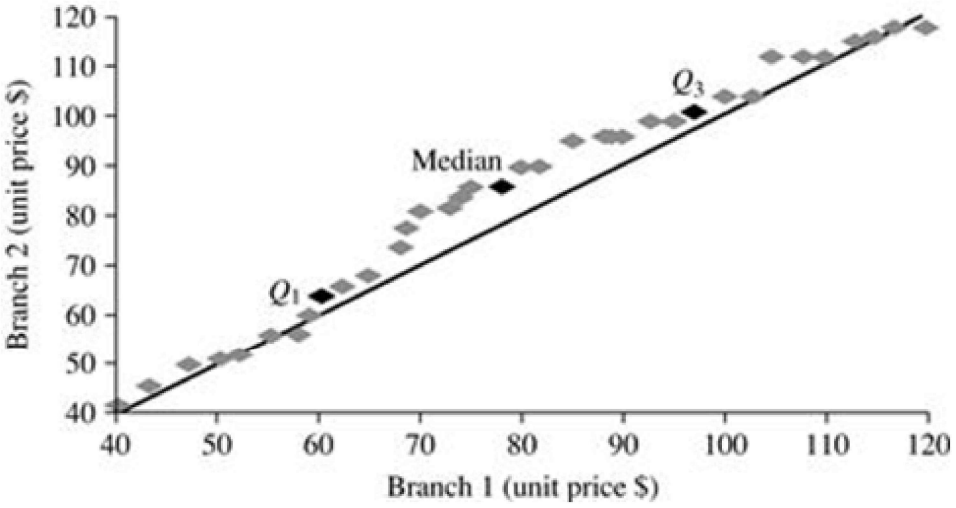


Figure 8.6: A q-q plot for unit price data from two *All Electronics* branches

Histograms: Histograms (or **frequency histograms**) are at least a century old and are widely used. “Histos” means pole or mast, and “gram” means chart, so a histogram is a chart of poles. Plotting histograms is a graphical method for summarizing the distribution of a given attribute, X. If X is nominal, such as *automobile_model* or *item_type*, then a pole or vertical bar is drawn for each known value of X. The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a **bar chart**.

. Figure 8.7 shows a histogram for the data set of Table 8.1, where buckets (or bins) are defined by equal-width ranges representing $20 increments and the frequency is the count of items sold.

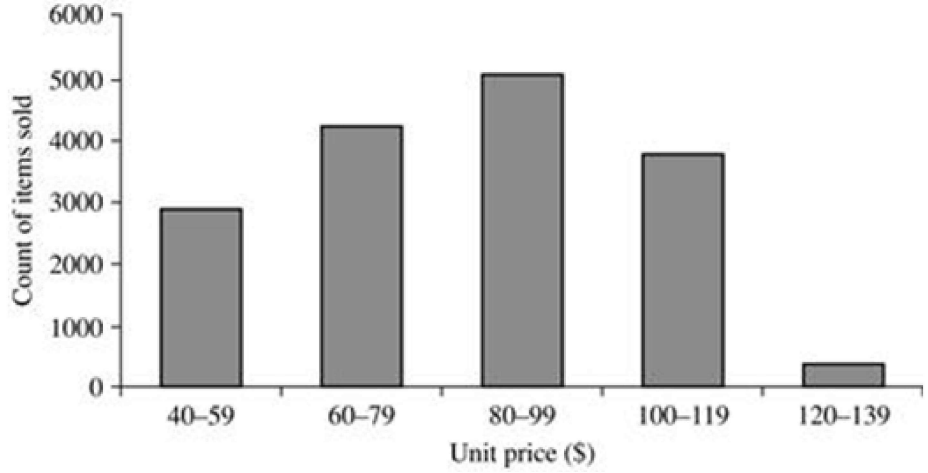


Figure 8.7: A histogram for the Table 8.1 data set

Scatter Plots and Data Correlation: A **scatter plot** is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane. Figure 8.8 shows a scatter plot for the set of data in Table 8.1.

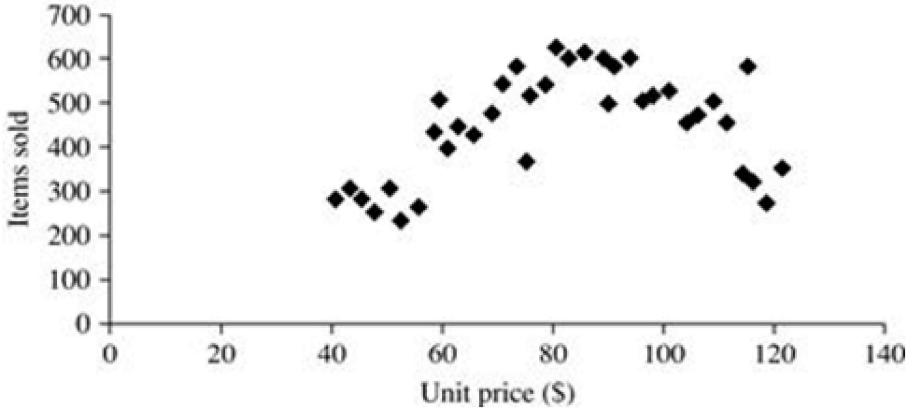


Figure 8.8: A scatter plot for the Table 8.1 data set

Two attributes, *X*, and *Y*, are **correlated** if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated). Figure 8.9 shows examples of positive and negative correlations between two attributes.

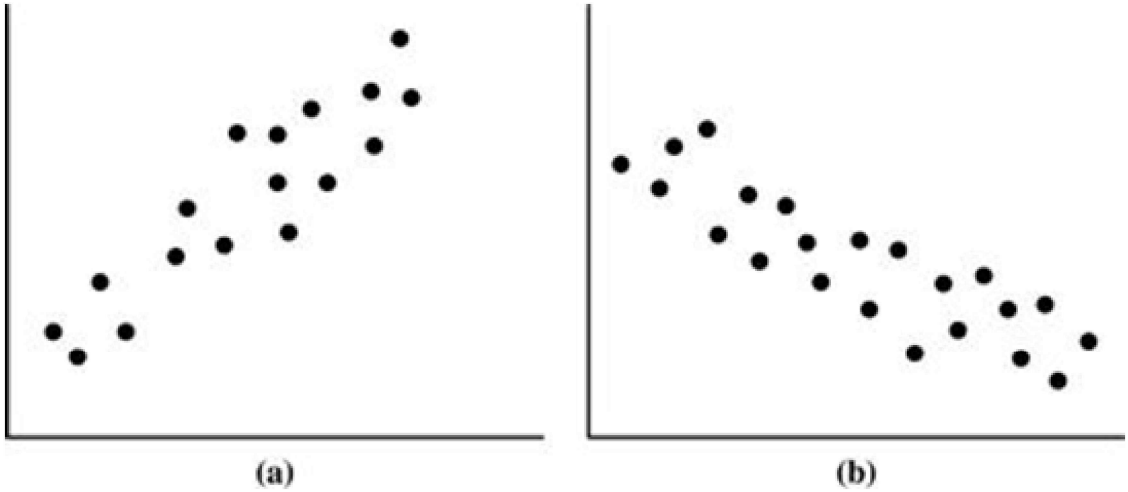


Figure 8.9: Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

Figure 8.10: Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

CHECK YOUR PROGRESS

Q.6: What is Class compassion?

.....

.....

.....

Q.7: Define the properties of standard deviation.

.....

.....

.....

.....

.....

Q.8: What is histogram?

.....

.....

.....

Q.9: What are methods used for graphic display of statistical data?

.....

.....

8.7 LET US SUM UP

- Attributes are used to describe the objects.
- Quantitative and qualitative are the two main categories of attributes.
- Data Generalization have two approaches. Namely data cube approach and attribute oriented induction approach.
- With attribute relevance, we can quantify the attributes.
- With attribute relevance, information gain, gini index, uncertainty, correlation coefficient are calculated.
- Class comparison describes the difference between targeted class from its constructing class.
- Mean, mode, median are the measure of central tendency of statistical data.
- Dispersion of data we get by calculating range, quartiles, variance, standard deviation and interquratile range.
- Quartile plot, quartile-quartile plot, histogram and scattered plot are the methods used for graphics display of the statistical data.

8.8 FURTHER READING

1) Han, J., Pei, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
2) Pujari, A. K. (2001). Data Mining Techniques. Universities Press.

8.9 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (a) Gender

Ans. to Q. No. 2: **Attribute removal** is a process based on the following rule: *If there is a large set of distinct values for an attribute of the initial working relation, but either (case 1) there is no generalization*

operator on the attribute or (case 2) its higher-level concepts are expressed in terms of other attributes, then the attribute should be removed from the working relation.

Ans. to Q. No. 3: The general idea of attribute-oriented induction is to first collect the task-relevant data using a database query and then perform generalization based on the examination of the number of each attribute's distinct values in the relevant data set. This Process is called **data focusing.**

Ans. to Q. No. 4: Data collection, Preliminary Relevance analysis using conservative AOI, Remove irrelevant or weakly relevant attributes using the selected measure.

Ans. to Q. No. 5: The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represents some category or state. So its called as categorical attribute.

Ans. to Q. No. 6: Class discrimination or comparison (hereafter referred to as **class comparison**) mines descriptions that distinguish a target class from its contrasting classes

Ans. to Q. No. 7: Properties of standard deviation are:

- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

Ans. to Q. No. 8: **Histograms** (or **frequency histograms**) are at least a century old and are widely used. "Histos" means pole or mast, and "gram" means chart, so a histogram is a chart of poles.

Ans. to Q. No. 9: Q*uantile plots*, Q*uantile–quantile plots*, *histograms*, and *scatter plots*.

8.10 MODEL QUESTIONS

Q.1: Define attribute.

Q.2: What is data generalization?

Q.3: Describe OLAP and attribute oriented induction approach of data generalization.

Q.4: Explain the reason behind attribute relevance.

Q.5: Explain the process of attribute relevance.

Q.6: Why is class comparison required?

Q.7: Define the different types of central tendency.

Q.8: Describe the different types of dispersion of data.

Q.9: Write about the methods to represent statistical data graphically.

*** ***** ***

Name of the Paper: .....
Course Code: .....

Centre for Internal Quality Assurance (CIQA)
Krishna Kanta Handiqui State Open University
City Office: Housefed Complex, Guwahati-781006

Learner's Feedback on Course

Dear Learner,

Regarding the course as mentioned above, we would like to know your opinions and comments so as to improve the quality of self learning materials in future. Please respond to the following statements by ticking the number you feel most reflect your opinion. After completion of the additional comments, please detach the page and send/mail the same to us at the address given below.

The Director, Centre for Internal Quality Assurance, KKHandiqui State Open University
Housefed Complex, Dispur, Guwahati-781006
(E-mail id: ciqa@kkhsou.in)

1) Approximately how many hours did you spend for studying the units in the course? ☐☐☐

2) Please give your opinions (by ✓ mark) to the following items based on your reading of the block:

Sl. No.	Statements	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I)	The SLMs of the course fulfil the learning objectives set out in the beginning of each unit					
II)	The units of the SLM could satisfy my academic needs and expectations					
III)	The Unit writers have excellent knowledge about the course contents					
IV)	Language and contents of the units were presented at a level which I could readily understand					
V)	Ample opportunity for participation in the activities provided in the units					
VI)	Used enough Illustrations (Diagrams, tables etc.) for conceptual clarity					
VII)	Quality of content is engaging, relevant, and up-to-date					
VIII)	The self check questions are very helpful					
IX)	The Possible/model questions and the answers to check my progress have benefited me a lot					

Additional Comments: *(Please feel free to provide your open comments)*

1) Which aspects of the SLM, according to you, worked well?

.....

.....

2) What sort of changes/improvements do you feel KKHSOU could implement to improve the overall quality of the SLM?

.....

.....

Thank you for taking the time to complete this form.