DAM 301: COURSE TITLE: DATA MINING AND DATA WAREHOUSING

NATIONAL OPEN UNIVERSITY OF NIGERIA

FACULTY OF SCIENCE

COURSE CODE: DAM 301

COURSE TITLE: DATA MINING AND DATA WAREHOUSING

CONTENT ***PAGE***

Module 1 Concepts of Data Mining

Module 2 Data Processes and Trends

Module3 Data Warehousing Concepts

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island
Lagos

Abuja Office
No. 5 Dar es Salaam Street
Off Aminu Kano Crescent
Wuse II, Abuja

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

Published by
National Open University of Nigeria

Printed 2013

ISBN: 978-058-674-1

CONTENT PAGE

INTRODUCTION

The course, data mining and data warehousing, DAM 301 is a three- credit unit course available for students studying towards acquiring the Bachelor of Science degree in Computer Science.

The overall aims of this course is to introduce you to the concepts of data mining and data warehousing. Other topics that will be discussed include the data mining problems, application of data mining, commercial tools of data mining, knowledge discovery, architecture of data warehousing, data marts, data warehousing lifecycle, data modelling, building of data warehouse, OLAP, MOLAP, ROLAP, data warehousing and future views.

For easy study and assimilation, the book is written in an easy-to-read and lingo free manner. The study material is divided into three modules namely: concepts of data mining, data mining and trends, and data warehouse concepts.

WHAT YOU WILL LEARN IN THIS COURSE

The overall aims and objectives of this course is to provide guidance on what you should achieve in the course of your studies. Each unit also has its own objectives which state specifically what you should achieve in the corresponding unit. To evaluate your progress continuously, you are expected to refer to the overall course aims and objectives as well as the corresponding unit objectives upon the completion of each unit.

COURSE AIMS

The overall aims and objectives of this course will help you to:

1. *Develop your knowledge and understanding of the underlying principles of data mining and data warehousing*
2. *Acquaint with the classification of data mining and approaches to data mining problems*
3. *Provide the tools of data mining and its application.*
4. *Develop your capacity in building a simple data warehousing.*

COURSE OBJECTIVES

Certain objectives have been set out to ensure that the course achieves its aims. Apart from the course objectives, each unit of this course has a set of objectives. In the course of the study, you will need to confirm, at the end of each unit, if you have met the objectives set at the beginning of each unit. At the end of this course you should be able to:

- *explain the concepts of data mining*
- *explain data processes and its trends*
- *describe the concept of data warehousing.*

WORKING THROUGH THIS COURSE

In order to have a thorough understanding of the course units, you will need to read and understand the contents of this course and explore the usage of some multimedia applications. This course is designed to be covered in approximately sixteen weeks, and it will require your devoted attention. You should do the exercises in the Tutor-Marked Assignment and submit to your tutors.

COURSE MATERIALS

Basically, we made use of textbooks and online materials. You are expected to search for more literature and web references for further understanding. Each unit has references and web references that were used to develop them.

ONLINE MATERIALS

Feel free to refer to the web sites provided for all the online reference materials required in this course. The website is designed to integrate with the print-based course materials. The structure follows the structure of the units and all the reading and activity numbers are the same in both media.

EQUIPMENT

In order to get the most from this course, it is essential that you make use of a computer system which has internet access.

Recommended System Specifications: Processor

- *2.0 GHZ Intel compatible processor 4GB RAM*
- *2000 GB hard drive with 5 GB free disk CD-RW drive*
- *Operating System Windows window 7 and above*
- *Microsoft Office 2007 and above*
- *Antivirus*

STUDY UNITS

There are eleven study units in this course:

Module 1 Concepts of Data Mining

Unit 1	*Overview of Data Mining*
Unit 2	*Data Description for Data Mining*
Unit 3	*Classification of Data Mining*
Unit 4	*Data Mining Technologies*

Module 2 Data Mining Processes and Trends

Unit 1	*Data Preparation and Preprocesses*
Unit 2	*Data Mining Process*
Unit 3	*Applications and trends in Data Mining*

Module 3 Data Warehousing Concepts

Unit 1	*Overview of Data Warehouse*
Unit 2	*Data Warehouse Architecture*
Unit 3	*Data Warehouse Design*
Unit 4.	*Data Warehouse and OLAP Technology*

TEXTBOOKS AND REFERENCES

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Connolly, A., VanderPlas, J., & Gray, A. (2014). Classification. In *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (pp. 365-402). PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctt4cgbdj.13

Jiawei Han., Micheline Kamber (2019). Data Mining: Concepts and Techniques, Second Edition

Yanchang Zhao, (2015). R and Data Mining Examples and Case Studies

Charu C. Aggarwal, 2015. The Textbook 2015th Edition

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

Shah, C. (2020). A Hands-On Introduction to Data Science. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

An Introduction to Data Mining, Retrieved on 28/07/2009. From: http://www.thearling.com/text/dmwhite/dmwhite.htm.

Data Mining Techniques, Retrieved on 28/07/2009. From: http://www.statsoft.com/TEXTBOOK/stdatmin.html.

Data Mining. Retrieved on 29/07/2009. Available Online: http://en.wikipedia.org/wiki/Data_mining.

Introduction to Data Mining and Knowledge Discovery. (3rd ed.).

Zambreno,J., Ozisikyilmaz,B., Choudhary,A. Accelerating Data Mining Workloads: Current Approaches and Future Challenges in System Architecture.

ASSESMENT

There are two aspects to the assessment of the course. First are the tutor- marked assignments; second, is a written examination. In tackling the assignments, you are expected to apply information and knowledge acquired during this course. The assignments must be submitted to your tutor for formal assessment in accordance with the deadlines stated in the assignment file.

The work you submit to your tutor for assessment will count for 30% of your total course mark. At the end of the course, you will need to sit for a final two-hour examination. This will also count for 70% of your total course mark.

ASSIGNMENT FILE

These are of two types: the self-assessment exercises and the tutor- marked assignment. The self-assessment exercises will enable you monitor your performance by yourself, while the tutor-marked assignment is a supervised assignment. The assignments take a certain percentage of your total score in this course. The tutor-marked assignment will be assessed by your tutor within a specified period. The examination at the end of this course will aim at determining the level of mastery of the subject matter. This course includes twelve (11) tutor- marked assignments and each must be done and submitted accordingly. Your best scores however, will be recorded for you. Be sure to send these assignments to your tutor before the deadline to avoid loss of marks.

PRESENTATION SCHEDULE

The presentation schedule included in your course materials gives you the important dates for the completion of tutor-marked assignments and attending tutorials. Remember, you are required to submit all your assignments by the due date. You should guard against lagging behind in your work.

TUTOR- MARKED ASSIGNMENT (TMA)

There are eleven tutor-marked assignments in this course. You need to submit all the assignments. The total marks for the best three (3) assignments will be 30% of your total course mark. Assignment questions for the units in this course are contained in the assignment file. You should be able to complete your assignments from the information and materials contained in your set textbooks, reading and study units.

However, you may wish to use other references to broaden your viewpoint and provide a deeper understanding of the subject. When you have completed each assignment, send it together with form to your tutor. Make sure that each assignment reaches your tutor on or before the deadline given. If, however, you cannot complete your work on time, contact your tutor before the assignment is done to discuss the possibility of an extension.

FINAL EXAMINATION AND GRADING

The final examination for the course will carry 70% of the total marks available for this course. The examination will cover every aspect of the course, so you are advised to revise all your corrected assignments before the examination. This course endows you with the status of a teacher and that of a learner. This means that you teach yourself and that you learn, as your learning capabilities would allow. It also means that you are in a better position to determine and to ascertain the what, the how, and the when of your course learning. No teacher imposes any method of learning on you.

The course units are similarly designed with the introduction following the contents, then a set of objectives and then the concepts and so on. The objectives guide you as you go through the units to ascertain your knowledge of the required terms and expressions.

COURSE MARKING SCHEME

This table shows how the actual course marking is broken down.

ASSIGNMENTS	*MARKS*
Assignment 1-4	*Four assignments, best three marks of the four count at 10% each: 30% of the course marks.*
End of Course Examination	*70% of overall course marks*
Total	*100% of course marks*

COURSE OVERVIEW

Each study unit consists of two hours work. Each study unit includes introduction, specific objectives, directions for study, reading materials, conclusions, and summary, tutor -marked assignments (TMAs), references / further reading. The units direct you to work on exercise related to the required readings. In general, these exercises test you on the materials you have just covered or require you to apply it in some way and thereby assist you to evaluate your progress and to reinforce your comprehension of the material. Together with TMAs, these exercises will help you in achieving the stated learning objectives of the individual units and of the course as a whole.

HOW TO GET THE MOST FROM THIS COURSE

In distance learning the study units replace the university lecturer. This is one of the great advantages of distance learning; you can read and work through specially designed study materials at your own pace, and at a time and place that suit you best

Think of it as reading the lecture instead of listening to a lecturer. In the same way that a lecturer might set you some reading to do, the study units tell you when to read your set books or other material. Just as a lecturer might give you an in-class exercise, your study units provide exercises for you to do at appropriate points.

Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole. Next is a set of learning objectives. These objectives enable you know what you should be able to do by the time you have completed the unit. You should use these objectives to guide your study. When you have finished the units you must go back and check whether you have achieved the objectives. If you make a habit of doing this you will significantly improve your chances of passing the course. Remember that your tutor's job is to assist you. When you need help, do not hesitate to call and ask your tutor to provide it.

1. *Read this Course Guide thoroughly.*
2. *Organize a study schedule. Refer to the 'Course Overview' for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Whatever method you chose to use, you should decide on it and write in your own dates for working on each unit.*
3. *Once you have created your own study schedule, do everything you can to stick to it. The major reason that students fail is that they lag behind in their course work.*
4. *Turn to unit 1 and read the introduction and the objectives for the unit.*
5. *Assemble the study materials. Information about what you need for a unit is given in the 'Overview' at the beginning of each unit. You will almost always need both the study unit you are working on and one of your set of books on your desk at the same time.*
6. *Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading*
7. *Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.*
8. *When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.*
9. *When you have submitted an assignment to your tutor for marking, do not wait for its return before starting on the next unit. Keep to your schedule. When the assignment is returned, pay particular attention to your tutor's comments, both on the tutor-marked assignment form and also written on the assignment. Consult your tutor as soon as possible if you have any questions or problems.*
10. *After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at*

the beginning of each unit) and the course objectives (listed in this Course Guide

FACILITATORS/TUTORS AND TUTORIALS

There are 12 hours of tutorials provided in support of this course. You will be notified of the dates, times and location of these tutorials, together with the name and phone number of your tutor, as soon as you are allocated a tutorial group. Your tutor will mark and comment on your assignments, keep a close watch on your progress and on any difficulties, you might encounter and provide assistance to you during the course. You must mail or submit your tutor-marked assignments to your tutor well before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible. Do not hesitate to contact your tutor by telephone, or e-mail if you need help.

The following might be circumstances in which you would find help necessary. Contact your tutor if:

- *you do not understand any part of the study units or the assigned readings,*
- *you have difficulty with the self-tests or exercises,*
- *you have a question or problem with an assignment, with your tutor's comments on an assignment or with the grading of an assignment.*

You should try your best to attend the tutorials. This is the only chance to have face to face contact with your tutor and to ask questions which are answered instantly. You can raise any problem encountered in the course of your study. To gain the maximum benefit from course tutorials, prepare a question list before attending them. You will learn a lot from participating in discussions actively.

SUMMARY

This course Data Mining and Data Warehousing is intended to develop your understanding of how you can uncover hidden patterns in their data which they can use to predict the behaviour of customers, products and processes.

We hope that you will find the course enlightening and that you will find it both interesting and useful. In the longer term, we hope you will get familiar with the National Open University of Nigeria and we wish outstanding success in all your studies.

MODULE 1 CONCEPTS OF DATA MINING

UNIT 1 OVERVIEW OF DATA MINING

CONTENT

1.0 INTRODUCTION

From the time immemorial humans have been manually extracting hidden predictive patterns from data, but the increasing volume of data in modern time requires an automatic approach. With the advent of data mining, it provides a new powerful technology with great potential to help private and public focus on the most important information in their data bases. Data mining is a result of a long process of research and product development, and the primary reason is to assist not only in uncovering hidden patterns from databases but also consists of collecting, managing, analysis and prediction of data.

The term data mining derived its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable one so the two processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides. This unit examines the meaning of data mining, the difference between it and knowledge discovery in databases (KDD), evolution of data mining, its scope, architecture and how it works.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- *Define the term data mining*
- *Understand the motivation behind data mining? why it is so important?*
- *Tasks of data mining*
- *Data mining application*
- *Differentiate between data mining and knowledge discovery in databases (KDD)*
- *State the evolution of data mining*
- *Understand data mining and OLAP*
- *Identify the architecture of data mining*
- *Explain how data mining works.*

3.0 MAIN CONTENT

3.1 Definition of Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data. Thus, data mining should have been more appropriately named as knowledge mining which emphasis on mining from large amounts of data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The key properties of data mining are

- *Automatic discovery of patterns*
- *Prediction of likely outcomes*
- *Creation of actionable information*
- *Focus on large datasets and databases*

The information or knowledge extracted can be used for any of the following applications

- *Market Analysis*
- *Fraud Detection*
- *Customer Retention*
- *Production Control*
- *Science Exploration*

❖ *Major Sources of Abundant data are:*

- *Business - Web, E-commerce, Transactions, Stocks*
- *Science - Remote Sensing, Bio informatics, Scientific Simulation*
- *Society and Everyone – News, Digital Cameras, You Tube*

- *Data Mining Applications:*

  - *Market Analysis and Management*
  - *Fraud Detection*
  - *Customer Retention*
  - *Production Control*
  - *Scientific Exploration*
  - *Corporate Analysis & Risk Management*

3.2 What Motivated Data Mining? Why Is It Important?

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data mining can be viewed as a result of the natural evolution of information technology.

The database system industry has witnessed an evolutionary path in the development of data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and advanced data analysis (involving data warehousing and data mining). For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. With numerous database systems offering query and transaction processing as common practice, advanced data analysis has naturally become the next target.

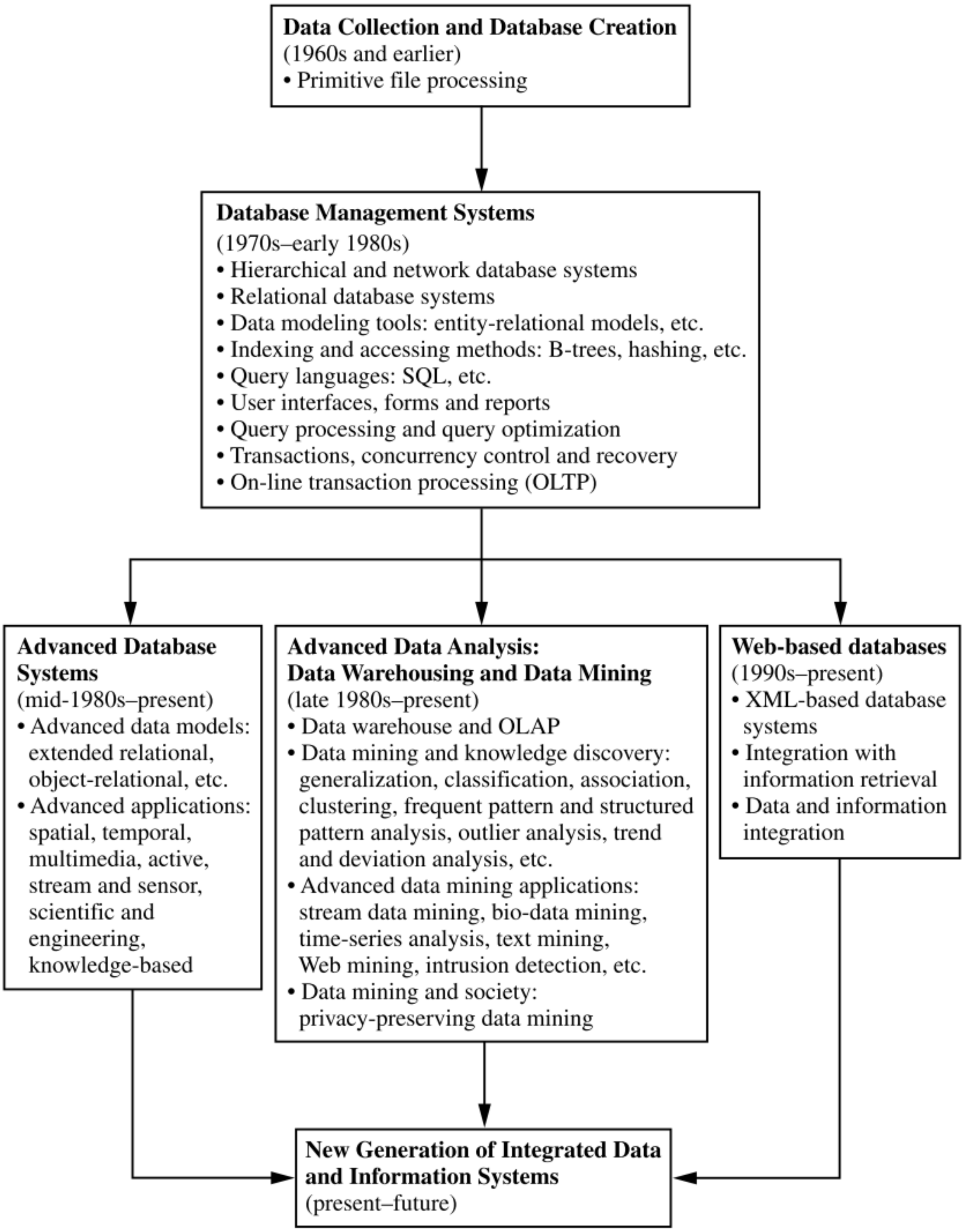

Data Collection and Database Creation
(1960s and earlier)
• Primitive file processing
Database Management Systems
(1970s–early 1980s)
• Hierarchical and network database systems
• Relational database systems
• Data modeling tools: entity-relational models, etc.
• Indexing and accessing methods: B-trees, hashing, etc.
• Query languages: SQL, etc.
• User interfaces, forms and reports
• Query processing and query optimization
• Transactions, concurrency control and recovery
• On-line transaction processing (OLTP)
Advanced Database Systems
(mid-1980s–present)
• Advanced data models: extended relational, object-relational, etc.
• Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based
Advanced Data Analysis: Data Warehousing and Data Mining
(late 1980s–present)
• Data warehouse and OLAP
• Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
• Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
• Data mining and society: privacy-preserving data mining
Web-based databases
(1990s–present)
• XML-based database systems
• Integration with information retrieval
• Data and information integration
New Generation of Integrated Data and Information Systems
(present–future)

Fig. 1.1 *Evolution of database system technology*

Since the 1960s, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems. The research and development in database systems since the 1970s has progressed from early hierarchical and network database systems to the development of relational database systems (where data are stored in relational table structures), data modeling tools, and indexing and accessing methods. In addition, users gained convenient and flexible data access through query languages, user inter- faces, optimized query processing, and transaction management. Efficient methods for on-line transaction processing (OLTP), *where a query is viewed as a read-only transaction, have contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data.*

Database technology since the mid-1980s has been characterized by the popular adoption of relational technology and an upsurge of research and development activities on new and powerful database systems. These promote the development of advanced data models such as extended-relational, object-oriented, object-relational, and deductive models. Application-oriented database systems, including spatial, temporal, multimedia, active, stream, and sensor, and scientific and engineering databases, knowledge bases, and office information bases, have flourished. Issues related to the distribution, diversification, and sharing of data have been studied extensively. Heterogeneous database systems and Internet-based global information systems such as the World Wide Web (WWW) have also emerged and play a vital role in the information industry.

The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful and affordable computers, data collection equipment, and storage media. This technology provides a great boost to the database and information industry, and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis. Data can now be stored in many different kinds of databases and information repositories. One data repository architecture that has emerged is the data warehouse, a repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making. Data warehouse technology includes data cleaning, data integration, and on-line analytical processing (OLAP), that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation as well as the ability to view information from different angles.

3.3 Data Mining Tasks

Data mining involves six common classes of tasks:

***Anomaly detection (Outlier/change/deviation detection) –** The identification of unusual data records, that might be interesting or data errors that require further investigation. Association*

***Association rule learning (Dependency modelling) –** Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.*

***Clustering** – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.*

Classification *– is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".*
Regression – *attempts to find a function which models the data with the least error.*

Summarization *– providing a more compact representation of the data set, including visualization and report generation.*

3.4 Architecture of Data Mining

A typical data mining system may have the following major components.

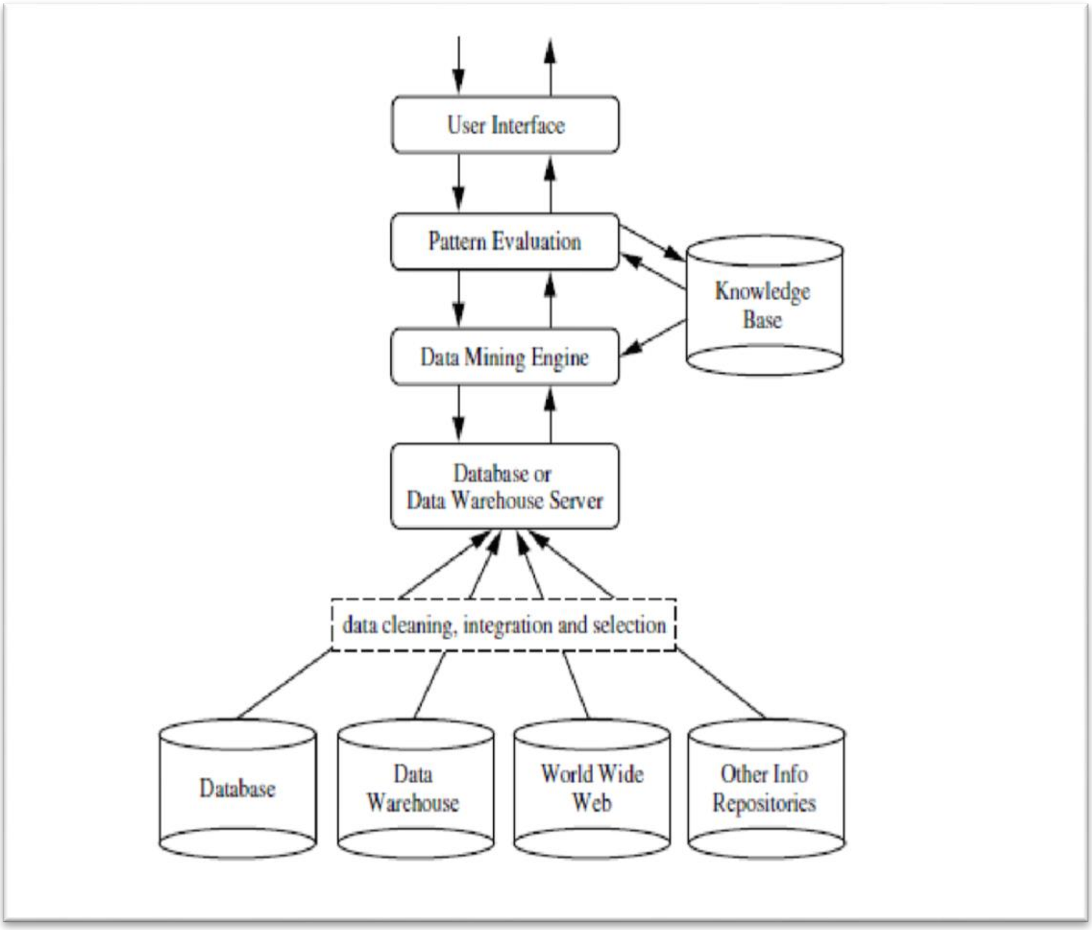


Fig. 1.2 *Architecture of a typical data mining system.*

- ***Knowledge Base:***

This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional *interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).*

- ***Data Mining Engine:***

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

- ***Pattern Evaluation Module***

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

- ***User interface***

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

3.5 Data Mining and Knowledge Discovery in Databases (KDD)

The idea of searching for useful patterns in data has a variety of names such as data mining, knowledge extraction, information discovery, information harvesting, data archeology and data pattern processing. The term data mining as earlier explained in section 3.0 is mostly employed by data analysts MIS specialties, statisticians and database administrators, while Knowledge Discovery in Databases (KDD) refers to the overall process of discovering useful knowledge from data; although, data mining and knowledge discovery in databases (KDD) are frequently treated as synonyms.

The term KDD was first coined by Gregory Piatetsky-Shapiro in 1989 to describe the process of searching for interesting, interpreted, useful and novel data. Reflecting the conceptualization of data mining, it is considered by researchers to be a particular step in a larger process of Knowledge Discovery in Databases (KDD). The knowledge discovery in databases process comprises of a few steps in chronological order that starts from raw data collections to some forms of new knowledge. This include data cleaning, data integration, *data selection, data transformation, data mining, pattern evaluation and knowledge presentation.*

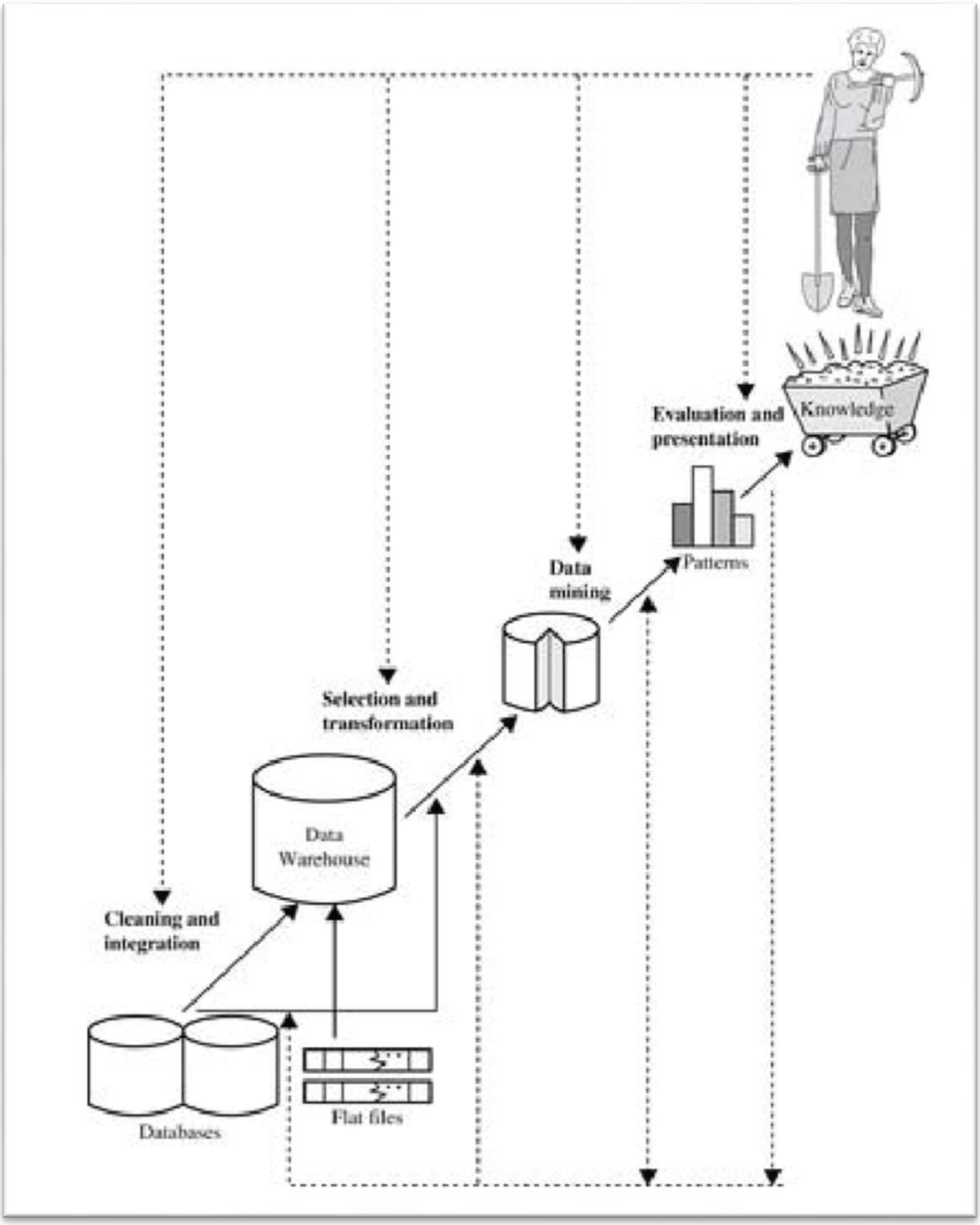


Fig. 1.3. *Data Mining as One of Knowledge Discovery Process*

KDD being an iterative process consists of the following steps:

- ***Data cleaning** (to remove noise and inconsistent data)*
- ***Data integration** (where multiple data sources may be combined) A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step, where the resulting data are stored in a data warehouse.*
- ***Data selection** (where data relevant to the analysis task are retrieved from the database).*

- ***Data transformation*** *(where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations) Sometimes data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing. Data reduction may also be performed to obtain a smaller representation of the original data without sacrificing its integrity*

- ***Data mining*** *(an essential process where intelligent methods are applied to extract data patterns)*

- ***Pattern evaluation*** *(to identify the truly interesting patterns representing knowledge based on interestingness measures*

- ***Knowledge presentation*** *(where visualization and knowledge representation techniques are used to present mined knowledge to users)*

It is common to combine some of these steps together for instance, data cleaning and data integration can be performed together as a pre- processing phase to generate a data warehouse. Also, data selection and data transformation can be combined where the consolidation of the data is the result of the selection, or as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process and can contain loops between any two steps. Once knowledge is discovered it is presented to the user, the evaluation measures are enhanced and the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different and more appropriate results.

3.6 Data Mining and OLAP

The difference between data mining and On-Line Analytical Processing (OLAP) is a very common question among data processing professionals. As we all see, the two are different tools that can complement each other.

OLAP is part of a spectrum of decision support tools. Unlike traditional query and report tools that describe what is in a database, OLAP goes further to answer why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. For example, an analyst may want to determine the factors that lead to loan defaults. He or she might initially hypothesis that people with low incomes are bad credit risks and analyze the database with OLAP to verify or disprove assumption. If that hypothesis were not borne out by the data, the analyst might then look at high debt as the determinant of risk. It the data does not support this guess either he or she might then try debt and income together as the best prediction of bad credit risks (Two Crows Corporation, 2005). In other words, OLAP is used to generate a series of hypothetical patterns and relationships, uses queries against database to verify them or disprove them. OLAP analysis is basically a deductive process. But when the number of variables to be analyzed becomes voluminous it becomes much more difficult and time-consuming to find a good hypothesis, analyze the database with OLAP to verify or disprove it.

Data mining is different from OLAP; unlike OLAP that verifies hypothetical patterns, it uses the data itself to uncover such patterns and is basically an inductive process. For instance, *suppose an analyst wants to identify the risk factors for loan default is to use a data mining tool. The data mining tool may discover people with high dept and low incomes are bad credit risks, it may go further to discover a pattern that the analyst does not consider that age is also a determinant of risk.*

Although data mining and OLAP complement each other in the sense that before acting on the pattern, the analyst needs to know what would be the financial implications using the discovered pattern to govern who gets credit. OLAP tool allows the analyst to answer these kinds of questions. It is also complimentary in the early stages of the knowledge discovery process.

4.0 SELF ASSESSMENT EXERCISE 1

i. *What is data mining?*
ii. *List all the steps involved in KDD process.*
iii. *Describe the steps involved in data mining when viewed as a process of knowledge discovery*

SELF- ASSESSMENT EXERCISE 2

Briefly explain the scope of data mining under the following headings:

i. *Automated prediction of trends and behaviours*
ii. *Present an example where data mining is crucial to the success of a business. What data mining functions does this business need? Can they be performed alternatively by data query processing or simple statistical analysis?*
iii. *How is a data warehouse different from a database? How are they similar?*

5.0 CONCLUSION

With the introduction of data mining technology, individuals and organization can uncover hidden patterns in their data which they can use to predict the behaviour of customers, products and processes.

6.0 SUMMARY

In this unit, you have learnt that:

- *Data mining is the task of discovering interesting patterns from large amounts of data, where the data canbe stored in databases, datawarehouses,or other information repositories. It is a young interdisciplinary field, drawing from areas such as database sys tems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing.*
- *data mining can be distinguished from knowledge discovery in databases (KDD) in a number of ways such as, data mining is a particular step in a large process of knowledge discovery in database (KDD)*

- *some of the scopes of data mining which are automated prediction of trends and behaviours, and automated discovery of previously unknown patterns*
- *data mining is different from OLAP in number of ways; unlike OLAP that verifies hypothetical patterns, it uses the data itself to uncover such patterns and is basically an inductive process.*
- *A knowledge discovery process includes data cleaning, data integration, data selec- tion, data transformation, data mining, pattern evaluation, and knowledge presentation*

TUTOR-MARKED ASSIGNMENT

i. (a) What do you understand by the term data mining?
(b) List and explain in chronological order the steps involved in knowledge discovery in databases.
ii. (a) Differentiate between data mining and OLAP.
(b) Explain how the evolution of database technology led to data mining

7.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

An Introduction to Data Mining, Retrieved on 28/07/2009. From:http://www.thearling.com/text/dmwhite/dmwhite.htm.

Data Mining Techniques, Retrieved on 28/07/2009. From: http://www.statsoft.com/TEXTBOOK/stdatmin.html.

Data Mining. Retrieved on 29/07/2009. Available Online: http://en.wikipedia.org/wiki/Data_mining.

Introduction to Data Mining and Knowledge Discovery, Third Edition.

Introduction to Data Mining, Retrieved on 15/08/2009. From http://www.eas.asu.edu/-mining03/chap2/lesson_2.html

Jeffrey, W. S. (Dec. 2004). *Data Mining: An Overview. From: Congressional Research Service. The Library of Congress.*

Mosud, Y.O. (2009). *Introduction to Data Mining and Data Warehousing. Lagos: Rashmoye Publications.*

Osmar, R. Z.(1999). *Principles of Knowledge Discovery in Databases. Pisharath, J. et al. Accelerating Data Mining Workloads: Current Approaches and Future Challenges in System Architecture.*

Sumathi, S. & Sivanamdam, S.N. (2006*). Introduction to Data Mining Principles, Studies in Computational Intelligence (SCI) 29, 1-20.*

Usama, F., Gregory, P.& Padhraic, S. (1996). *From Data Mining to Knowledge Discovery in Databases, Article of American Association for Artificial Intelligence Press*

UNIT 2 DATA DESCRIPTION FOR DATA MINING

CONTENTS

1.0 INTRODUCTION

In actual sense data mining is not limited to one type of media or data but applicable to any kind of information available in the repository. A repository is a location or set of locations where systems analysts, system designers and system builders keep the documentation associated with one or more systems or projects.

But before we begin to explore the different kinds of data to mine it will be interesting to familiarize ourselves with the variety of information collected in digital form in databases and flat files. Also, to be explored are the types to mine and data mining functionalities.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- *identify the different kinds of information collected in our databases*
- *describe the types of data to mine*
- *explain the different kinds of data mining functionalities and the knowledge they discover.*

3.0 MAIN CONTENT

3.1 Types of Information Collected

We collect on daily basis a myriad of data which ranges from simple numerical measurements and text documents to more complex information such as hypertext documents, scientific data, spatial data and multimedia channels. Here is a different kind of information often collected in digital form in databases and flat files, although not exclusive.

3.1.1 Scientific Data

Our society is seriously gathering great amount of scientific data that needs to be analyzed. Examples are in the Swiss nuclear accelerator laboratory counting particles, South Pole iceberg gathering data about oceanic activity, American university investigating human psychology and Canadian forest studying readings from a grizzly bear radio collar. The unfortunate part of it is we can easily capture and store more new data faster than we can analyze the old data that have been accumulated.

3.1.2 Personal and Medical Data

From personal data to medical and government, very large amounts of information are continuously collected. Governments, individuals and organisations such as hospitals and schools are on daily basic stockpiling large quantity of very important personal data to help them manage human resources, better understanding of market, or simply assist client. No matter the private issues this type of data reveals, the information is collected used and even shared. And when compared with other data this information can shed more light on customer behaviour and likes.

3.1.3 Games

The rate at which our society gathers data and statistics about games, players and athletes is tremendous. These ranges from car-racing, swimming, hockey scores, footballs, basketball passes, chess positions and boxers' pushes, all these data are stored. Trainers and athletes make use of this data to improve their performances and have a better understanding of their opponents, but the journalists and commentators use this information to report.

3.1.4 CAD and Software Engineering Data

There are different types of Computer Assisted Design (CAD) systems used by architects and engineers to design buildings and picture system components or circuits. These systems generate a great amount of data. Also, software engineering is a source of data generation with code, function libraries and objects, these needs powerful tools for management and maintenance

3.15 Business Transaction

Every transaction in business is often noted for the sake of continuity. These transactions are usually related and can be inter-business deals such as banking, purchasing, exchanges and stocks or intra-business operations such as management of in-house wares and assets. Large departmental stores for example stores million of transactions on daily basis with the use barcodes.

The storage space does not pose any problem, as the price of hard disks are dropping, but the effective use of the data within a reasonable time frame for competitive decision-making is certainly the most important problem to be solved for businesses that struggle in competitive world.

3.1.6 Surveillance Video and Pictures

With the incredible fall in price of video camera prices, video cameras are becoming very common. The video tapes from surveillance cameras are usually recycled, thereby losing its content. But today there is tendency to store the tapes and even digitise them for future use and analysis.

3.1.7 Satellite Sensing

There are countless numbers of satellites around the globe, some are geo-stationary above a region while some are orbiting round the Earth, but all are sending a non-stop of data to the surface of the earth. NASA which is a body controlling large number of satellites receives more data per second than all NASA engineers and researchers can cope with. Many of the pictures and data captured by the satellite are made public as soon they are received hoping that other researchers can analyse them.

3.1.8 World Wide Web (WWW) Repositories

Since the advent of World Wide Web in 1993, documents of different formats, contents and description have been collected and inter- connected with hyperlinks making it the largest repository of data ever built, The World Wide Web is the most important data collection regularly used for reference because of the wide variety of topics covered and the infinite contributions of resources and authors. Many even believe that the World Wide Web is a compilation of human knowledge.

3.1 Types of Data Mined

Data mining can be applied to any kind of information in the repository, though algorithms and approaches may differs when applied to different types of data. And the challenges posed by different types of data vary extensively. Data Mining is used and studied for databases including relational databases, object-relational databases and object-oriented data, bases data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, and advanced databases such as spatial databases, multimedia database, time-series databases and flat files. Some of these are discussed in more details as follows.

3.2.1 Flat Files

These are the commonest data source for data mining algorithms especially at the research level. Flat files are simply data files in text or binary format with a structured known by the data mining algorithms to be applied. The data in these files can be in form of transactions, time- sales data, scientific measurements etc

3.2.2 Relational Databases

This is the most popular type of database system in use today by computers. It stores data in a series of two-dimensional tables called relation (i.e. tabular form). A relational database consists of a set of tables containing either values of entity attributes, or value of attribute from entity relationships. Tables generally have columns and rows, where columns represent attribute and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. In table 2.1 we present some relations student name, registration number, department and grade in computer representing a fictitious student grade in a class. These relations are just a subset of what could be a database for student score.

Table 2.1: Relational Database

Student Name	*Registration*	*Department*	*Grade in Data Mining*
Olumoye	*BUS/05/MLD/101*	*Business*	*A*
Chantel	*MKT/05/MLD/105*	*Marketing*	*B*
Chukwuma	*BFN/05/MLD/203*	*Banking&Finance*	*A*
Olatunji	*ACC/05/MLD/102*	*Accountancy*	*C*
Victor	*BUS/05/MLD/200*	*Business*	*B*

The most commonly used query language for relational database is Structured Query Language (SQL), it allows for retrieval and manipulation of data stored in the tables as well as the calculation of aggregate function such as sum, min, max and count. The data mining algorithm that uses relational databases can be more versatile than data mining algorithm that is specifically designed for flat files because they can always take advantage of the structure inherent in relational databases, while data mining can benefit from Structured Query Language (SQL) for data selection, transformation and consolidation. Also, it goes beyond what SQL can provide like predicting, comparing and detecting deviations.

3.2.3 Data Warehouses

A data warehouse (a storehouse) is a repository of data gathered from multiple data sources (often heterogeneous) and is designed to be used as a whole under the same unified schema. A data warehouse provides an option of analyzing data from different sources under the same roof. The most efficient data warehousing architecture will be able to incorporate or at least reference all management systems using designated technology suitable for corporate database management e.g. Sybase, Ms SQL Server

3.2.1 Transaction Databases

This is a set of records that represent transactions, each with a time stamp, an identifier and set of items. Also, associated with the transaction files is the descriptive data for the items.

Rentals

Transaction (1)	*Data*	*Time*	*Customer ID*	*Item List*
TI	*14/09/04*	*14.40*	*12*	*10,11,30, 110..*
II.	III.	IV.	V.	VI.
VII.	VIII.	IX.	X.	XI.

Fig. 2.2*: Fragment of a Transaction Database for Rentals in a Store*

Figure 2.2 represents a transaction database, each record shows a rental contact with a customer identifier, a date and list of items rented. But relational database do not allow nested tables that is a set as attribute value, transactions are usually stored in flat files or stored in two normalised transaction tables, one for the transactions and the other one for the transaction items. A typical data analysis on such data is the so-called market basket analysis or association rules in which associations occurring together or in sequence are studied.

3.2.5 Spatial Databases

These are databases that in addition to the usual data stores geographical information such as maps, global or regional positioning, and this type of database also present new challenges to data mining algorithms.

3.2.6 Multimedia Databases

Multimedia databases include audio, video, images and text media. These can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia database is characterized by its high dimension; this makes data mining more challenging. Data mining that comes from multimedia repositories may require vision, computer graphics, images interpretation and natural language processing methodologies

3.2.7 Time-Series Databases

This type of database contains time related data such as stock market data or logged activities. Time-series database usually contain a continuous flow of new data coming in that sometimes causes the need for a challenging real time analysis. Data mining in these types of databases often include the study of trends and correlations between evolutions of different variables, prediction of trends and movements of the variables in time.

3.2.8 World Wide Web

World Wide Web is the most heterogeneous and dynamic repository available. Large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are assessing its resources daily. The data in the World Wide Web are organized in inter-connected documents, which can be text, audio, video, raw data and even applications. The World Wide Web comprises of three major components: the content of the web, which encompasses document available, the structure of the web, which covers the hyperlinks and the relationships between documents the usage of the web, this describe how and when the resources are accessed

.A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or web mining, addresses all these issues and is often divided into web content mining and web usage mining.

SELF- ASSESSMENT EXERCISE 1

List and briefly explain any five kinds of information often collected in digital form in databases and flat files.

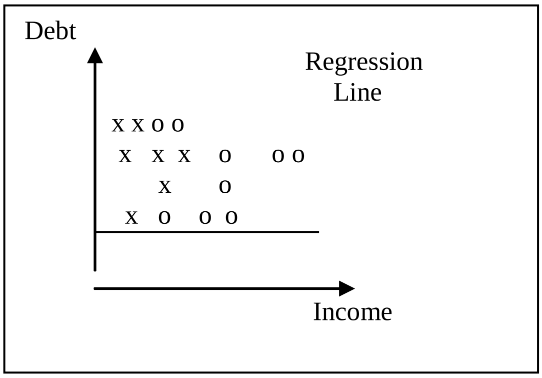
3.3 Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in data mining task. It is a very common phenomenon that many users do not have clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore crucial to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important issue in data mining system.

The data mining functionalities and the variety of knowledge they discover are briefly described in this section. These are as follows:

3.3.1 Classification

This is also referred to as supervised classification and is a learning function that maps (i.e. classifies) item into several given classes. The classification uses given class labels to order the objects in the data collection. Classification approaches normally make use of a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model which is used to classify new objects. Examples of classification method used in data mining application include the classifying of trends in financial markets and the automated identification of objects of interest in large image database. Figure 2.2 shows a simple partitioning of the loan data into two class regions; this may be done imperfectly using a linear decision boundary. The bank may use the classification regions to automatically decide whether future loan applicants will be given loan or not.



The shaped region denotes class with no loan

***Fig. 2.3**: A Simple Linear Classification Boundary for the Loan Data Set*
***Source:** Usama, F. et al. (1996)*

3.3.2 Characterisation

Data characterization is also called summarization and involves methods for finding a compact description (general features) for a subject of data or target class, and produces what is called characteristics rules. The data that is 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. A simple example would be tabulating the mean and standard deviations for all fields. More sophisticated methods involve the deviation of summary rules (Usama et al. 1996; Agrawal et al. 1996), multivariate visualization techniques and the discovery of functional relationships between variables. Summarization techniques are often applied to interactive exploratory data analysis and automated report generation (Usama et al., 1996)

3.3.3 Clustering

Clustering is similar to classification and is the organization of data in classes. But unlike classification, in clustering class tables are not predefined (unknown) and is up to clustering algorithm to discover acceptable classes. Clustering can also be referred to as unsupervised classification because the classification is not dictated by given class tables. We have so many clustering approaches which are all based on the principle of maximizing the similarity between objects in the same class (that is intra-class similarity) and minimizing the similarity between objects of different classes that is inter-class similarity.

3.3.4 Prediction (Regression)

This involves learning a function that maps a data item to a real–valued prediction variable. This method has attracted considerable attention given the potential implication of successful forecasting in a business context. Predictions can be classified into two major types namely: one can either try to predict some unavailable data value or pending trends, or predict a class label for some data (this is tied to classification). The moment a classification model is built based on 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 often refers to forecast of missing numerical value, or increase/decrease trends in time related data. Summarily, the main idea of prediction is to use a large number of past values to consider probable future values.

3.3.5 Discrimination

Data discrimination generates what we call 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 instance, we may want to characterise the rental customers that regularly rent more than 50 movies last year with those whose rental account is lower than 10. The techniques used for data discrimination are similar to that used for data characterisation with the exception that data discrimination results include comparative measures.

3.3.6 Association Analysis

Association analysis is the discovery of what we commonly refer to as 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 that 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 because it searches for relationship between variable. For example, a supermarket might gather data of what each customer buys. With the use of association rule learning, the supermarket can work out what products are frequently bought together, which is useful for marketing purposes. This is sometimes called market basket analysis.

3.3.7 Outlier Analysis

This is also referred to as exceptions or surprises. Outliers are data elements that cannot be grouped in a given class or clusters, and often important to identify, though, outliers can be considered noisy and discarded in some applications. They can reveal important knowledge in other domains; this makes them very significant and their analysis valuable.

3.3.8 Evolution and Deviation Analysis

Evolution and deviation analysis deals with the study of time related data that changes in time. In actual sense evolution analysis models evolutionary trends in data that consent with characterizing, comparing, classifying or clustering of time related data. While deviation analysis is concerned with the differences between measured values and expected values, and attempts to find the cause of the deviations from the expected values.

4.0 SELF- ASSESSMENT EXERCISE

What do you understand by the following data mining terms?

i. *Classification*
ii. *Characterisation*
iii. *Discrimination*
iv. *Association*
v. *Prediction*
vi. *Clustering*

Give examples of each data mining functionality, using a real-life database with which you are familiar

TUTOR- MARKED ASSIGNMENT

i. List and explain any five data mining functionalities and the variety of knowledge they discover.
ii. Research and describe an application of data mining that was not presented in this chapter.
iii. Briefly describe the following advanced database systems and applications: object-relational databases, spatial databases, text databases,multimedia databases, Transaction Databases, stream data, the World Wide Web.

5.0 CONCLUSION

Data mining therefore is not limited to one media or data; it is applicable to any kind of information repository and the kind of patterns that can be discovered depend upon the data mining tasks employed.

6.0 SUMMARY

In this unit, we have learnt that:

- *Data patterns can be mined from many different kinds of databases, such as relational databases, data warehouses, and transactional, and object-relational databases. Interesting data patterns can also be extracted from other kinds of information repositories, including spatial, time-series, sequence, text, multimedia, and legacy databases, data streams, and the World Wide Web.*
- *Data mining functionalities include the discovery of concept/class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. Characterization and discrimination are forms of data summarization.*
- *A Database technology has evolved from primitive file processing to the development of database management systems with query and transaction processing. Further progress has led to the increasing demand for efficient and effective advanced data analysis tools. This need is a result ofthe explosive growth in data collected fromappli- cations, including business and management, government administration, science and engineering, and environmental control.*

- *Data mining functionalities include the discovery of concept/class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. Characterization and discrimination are forms of data summarization.*
- *data mining can be applied to any kind of information in the reporting*
- *data mining system allows the discovery of different kind of knowledge and at different levels of abstraction.*

7.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

An Introduction to Data Mining, Retrieved on 28/07/2009. From: http://www.thearling.com/text/dmwhite/dmwhite.htm.

Data Mining Techniques, Retrieved on 28/07/2009. From: http://www.statsoft.com/TEXTBOOK/stdatmin.html.

Introduction to Data Mining and Knowledge Discovery. (3rd ed.).

Introduction to Data Mining, Retrieved on 15/08/2009. From http://www.eas.asu.edu/-mining03/chap2/lesson_2.html

Mosud, Y. O. (2009). *Introduction to Data Mining and Data Warehousing. Lagos: Rashmoye Publications.*

Sumathi,S. & Sivanamdam,S.N.(2006). *Introduction to Data Mining Principles, Studies in Computational Intelligence (SCI) 29, 1-20.*

Usama, F., Gregory, P. & Padhraic, S.(1996). *From Data Mining to Knowledge Discovery in Databases, Article of American Association for Artificial Intelligence Press.*

UNIT 3 CLASSIFICATION OF DATA MINING

CONTENTS

1.0 INTRODUCTION

There are many data mining systems available or presently being developed. Some are specialized systems dedicated to a given data sources or are confined to limited data mining functionalities, while others are more versatile and comprehensive. This unit examines the various classifications of data mining systems, data mining tasks, the major issues and challenging in data mining.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- *identify the various classifications of data mining systems*
- *describe the categories of data mining tasks*
- *state the diverse issues coming up in data mining*
- *describe the various challenges facing data mining.*
- *describe Issues relating to the diversity of database types*

3.0 MAIN CONTENT

3.1 Classification of Data Mining Systems

The data mining system can be classified according to the following criteria:

- *Statistics*
- *Database Technology*
- *Machine Learning*
- *Information Science*
- *Visualization*
- *Other Disciplines*

Some Other Classification Criteria:

- *Classification according to kind of databases mined*
- *Classification according to kind of knowledge mined*
- *Classification according to kinds of techniques utilized*
- *Classification according to applications adapted*

i. Classification according to kind of databases mined:
We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example, if we classify the database according to data model then we may have a relational, transactional, object- relational, or data warehouse mining system

ii. *Classification according to kind of knowledge mined:*
We can classify the data mining system according to kind of knowledge mined. It is means data mining system are classified on the basis of functionalities such as:
- *Characterization*
- *Discrimination*
- *Association and Correlation Analysis*
- *Classification*
- *Prediction*
- *Clustering*
- *Outlier Analysis*
- *Evolution Analysis*

iii. *Classification according to kinds of techniques utilized*
We can classify the data mining system according to kind of techniques used. We can describe these techniques according to degree of user interaction involved or the methods of analysis employed.

iv. *Classification according to applications adapted:*
We can classify the data mining system according to application adapted. These applications are as follows:
- *Finance*
- *Telecommunications*
- *DNA*
- *Stock Markets*
- *E-mail*

3.2 Data Mining Task

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

- ***Characterization:*** *Data characterization is a summarization of general features of objects in a target class, and 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.*

- ***Association analysis****: 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 than 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.*

- ***Classification**: Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects*

- ***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. The major idea is to use a large number of past values to consider probable future values.*

- ***Clustering**: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).*

- ***Outlier analysis**: Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.*

- ***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.*

3.3 Issues relating to the diversity of database types:

- *Handling of relational and complex types of data: Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data. Therefore, one may expect to have different data mining systems for different kinds of data.*

- *Mining information from heterogeneous databases and global information systems: Local- and wide-area computer networks (such as the Internet) connect many sources ofdata, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semistructured, or unstructured data with diverse data semantics poses great challenges to data mining. Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases. Web mining, which uncovers interesting knowledge about Web contents, Web structures, Web usage, and Web dynamics, becomes a very challenging and fast-evolving field in data mining.*

The above issues are considered major requirements and challenges for the further evolution of data mining technology.

SELF- ASSESSMENT EXERCISE 1

Briefly describe the four data mining tasks list below:
(i) Classification (ii) Clustering
(iii) Regression (iv) Association rule learning
(v) Evolution and deviation analysis

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

A critical question in the design of a data mining system is how to integrate or couple the data mining system with a database system and/or a data warehouse system. If a data mining system works as a stand-alone system or is embedded in an application program, there are no database system or Data warehouse systems with which it has to communicate. This simple scheme is called no coupling, where the main focus of the data mining design rests on developing effective and efficient algorithms for mining the available data sets. However, when a data mining system works in an environment that requires it to communicate with other information system components, such as database and Data warehouse systems, possible integration schemes include no coupling, loose coupling, semi tight coupling, and tight coupling. We examine each of these schemes, as follows:

***No coupling**: No coupling means that a data mining system will not utilize any function of a database or Data warehouse system. It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file. Such a system, though simple, suffers from several drawbacks. First, a database system provides a great deal of flexibility and efficiency at storing, organizing, accessing, and processing data. Without using a database / Data warehouse, a data mining system may spend a substantial amount of time finding, collecting, cleaning, and transforming data. In database and/or Data warehouse systems, data tend to be well organized, indexed, cleaned, integrated, or consolidated, so that finding the task-relevant, high-quality data becomes an easy task. Second, there are many tested, scalable algorithms and data structures implemented in database*

and Data warehouse systems. It is feasible to realize efficient, scalable implementations using such systems. Moreover, most data have been or will be stored in database/Data warehouse systems. Without any coupling of such systems, a data mining system will need to use other tools to extract data, making it difficult to integrate such a system into an information processing environment. Thus, no coupling represents a poor design.

***Loose coupling**: Loose coupling means that a data mining system will use some facilities of a database or Data warehouse system, fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data warehouse. Loose coupling is better than no coupling because it can fetch any portion of data stored in databases or data warehouses by using query processing, indexing, and other system facilities. It incurs some advantages of the flexibility, efficiency, and other features provided by such systems. However, many loosely coupled mining systems are main memory-based. Because mining does not explore data structures and query optimization methods provided by database*

or Data warehouse systems, it is difficult for loose coupling to achieve high scalability and good performance with large data sets.

***Semitight coupling**: Semitight coupling means that besides linking a data mining system to a database/ Data warehouse system, efficient implementations of a few essential data mining primitives (identified by the analysis of frequently encountered data mining functions) can be provided in the database/Data warehouse system. These primitives can include sorting, indexing, aggregation, histogram analysis, multiway join, and precomputation of some essential statistical measures, such as sum, count, max, min, standard deviation, and so on. Moreover, some frequently used intermediate mining results can be precomputed and stored in the database/Data warehouse system. Because these intermediate mining results are either precomputed or can be computed efficiently, this design will enhance the performance of a data mining system.*

***Tight coupling:** Tight coupling means that a data mining system is smoothly integrated into the database/Data warehouse system. The data mining subsystem is treated as one functional component of an information system. Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of a database or Data warehouse system. With further technology advances, data mining, database, and Data warehouse systems will evolve and integrate together as one information system with multiple functionalities. This will provide a uniform* *information processing environment. This approach is highly desirable because it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment.*

With this analysis, it is easy to see that a data mining system should be coupled with a database/ Data warehouse system. Loose coupling, though not efficient, is better than no coupling because it uses both data and system facilities of a database/ Data warehouse system. Tight coupling is highly desirable, but its implementation is nontrivial and more research is needed in this area. Semitight coupling is a compromise between loose and tight coupling. It is important to identify commonly used data mining primitives and provide efficient implementations of such primitives in database or Data warehouse systems.

3.5 Data Mining Issues

3.5.1 Mining different kinds of knowledge in databases:

The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore, it is necessary for data mining to cover broad range of knowledge discovery task.

3.5.2 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 returned results.

3.5.3 ***Incorporation of background knowledge:***

Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

3.5.4 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.

3.5.5 Presentation and visualization of data mining results:

Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans. This is especially crucial if the data mining system is to be interactive. This requires the system to adopt expressive knowledge representation techniques, such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.

3.5.6 Handling noisy or incomplete data:

The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

*3.5.7 **Pattern evaluation**:*

The interestingness problem: A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack nov- elty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations.

3.5.8 Efficiency and scalability of data mining algorithms:

To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases. From a database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems.

*3.5.9 **Parallel, distributed, and incremental mining algorithms**:*

The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged. Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "from scratch." Such algorithms perform knowledge modification incrementally to amend and strengthen what was previously discovered.

4.0 SELF- ASSESSMENT EXERCISE

Briefly discuss the following data mining issues list as follows:

- *i. Mining different kinds of knowledge in databases*
- *ii. Interactive mining of knowledge at multiple levels of abstraction*
- *iii. Incorporation of background knowledge*
- *iv. Data mining query languages and ad hoc data mining*
- *v. Presentation and visualization of data mining results*
- *vi. Handling noisy or incomplete data*
- *vii. Efficiency and scalability of data mining algorithms*
- *viii. Parallel, distributed, and incremental mining algorithms*

TUTOR-MARKED ASSIGNMENT

i. Briefly explain the classification of data mining systems:

(a) Classification according to kind of databases mined
(b) Classification according to kind of knowledge mined
(c) Classification according to kinds of techniques utilized
(d) Classification according to applications adapted

ii. Data mining system can be classified according to different criteria, list and explain them

iii. List and describe the five primitives for specifying a data mining task.

iv. Outliers are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraud- ulent use ofcredit cards. Taking fraudulence detection as an example, propose two meth- ods that can be used to detect outliers and discuss which one is more reliable.

v. Define each of the following data mining functionalities: characterization, discrimina- tion, association and correlation analysis, classification, prediction, clustering, and evo- lution analysis. Give examples ofeach data mining functionality, using a real-life database with which you are familiar.

vi. How is a data warehouse different from a database? How are they similar?

5.0 CONCLUSION

Data mining systems therefore can be categorized into various group using different criteria and there are four major classes of data mining tasks. Also issues and challenges affecting the effective implementation of data mining have to be addressed in order to ensure a successful exercise.

6.0 SUMMARY

In this unit you have learnt that:

- *data mining systems can be categorized according to various criteria such as type of data source mined, data model drawn, kind of knowledge discovered and the mining techniques used*
- *Database technology has evolved from primitive file processing to the development of database management systems with query and transaction processing. Further progress has led to the increasing demand for efficient and effective advanced data analysis tools. This need is a result ofthe explosive growth in data collected from applications, including business and management, government administration, science and engineering, and environmental control.*
- *Data mining systems can be classified according to the kinds ofdatabases mined, the kinds ofknowledge mined, the techniques used, or the applications adapted.*

- *We have studied five primitives for specifying a data mining task in the form of a data mining query. These primitives are the specification of task-relevant data (i.e., the data set to be mined), the kind of knowledge to be mined, background knowledge (typically in the form of concept hierarchies), interestingness measures, and knowl- edge presentation and visualization techniques to be used for displaying the discovered patterns.*
- *Data mining systems can be classified according to the kinds ofdatabases mined, the kinds ofknowledge mined, the techniques used, or the applications adapted.*

7.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Cham: Springer. ISBN: 978-3-319-14141-1

Zaki, M., & Meira, Jr, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511810114

Connolly, A., VanderPlas, J., & Gray, A. (2014). Fast Computation on Massive Data Sets. In *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (pp. 43-66). PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctt4cgbdj.5

Connolly, A., VanderPlas, J., & Gray, A. (2014). Classification. In *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (pp. 365-402). PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctt4cgbdj.13

Data Mining Techniques, Retrieved on 28/07/2009. From: http://www.statsoft.com/TEXTBOOK/stdatmin.html.

Mosud, Y. O. (2009). *Introduction to Data Mining and Data Warehousing. Lagos: Rashmoye Publications.*

Sumathi,S. & Sivanamdam,S.N.(2006). *Introduction to Data Mining Principles, Studies in Computational Intelligence (SCI) 29, 1-20.*

Usama, F., Gregory, P. & Padhraic, *S. From Data Mining to Knowledge Discovery in Databases, Article of American Association for Artificial Intelligence Press, (1996).*

UNIT 4 DATA MINING TECHNOLOGIES

CONTENTS

1.0 INTRODUCTION

In this unit, we shall be exploring some types of models and algorithms used in mining data. Most of the models and algorithms we shall be discussing are generalization of the standard workhorse of the modeling; although we should realize that no one model or algorithm can or should be used exclusively. As a matter of fact, for any given problem, the nature of the data itself will affect the choice of models and algorithm you decide to choose; and there is no best model or algorithm as you will need a variety of tools and technologies for you to find the best possible.

2.0 OBJECTIVE

At the end of this unit, you should be able to:

- *identify the various data mining technologies available.*

3.0 MAIN CONTENT

3.1 Data Mining Technologies

The analytical techniques used in data mining are often well-known mathematical algorithms techniques. But the new thing there is the application of those techniques to general business problems made possible by the increased availability of data and inexpensive storage and processing power. More so, the use of graphical interface has led to tools which are becoming available that business experts can easily use.

Most of the products use variations of algorithms that have been published in statistics or computer science journals with their specific implementations customized to meet individual vendor's goal. For instance, most of the vendors sell versions of the CART (Classification and Regression Trees) or CHAID (Chi-Squared Automatic Interaction Detection) decision trees with enhancements to work on parallel computers, while some vendors have proprietary algorithms that will not allow extension or enhancements of any published approach to work well.

Some of the technologies or tools used in data mining that will be discussed are: Neutral networks, decision trees, rule induction, multivariate adaptive repression splines (MARS), K-nearest neighbour and memory-based reasoning (MBR), logistic regression, discriminant analysis, genetic algorithms, generalized additive models (GAM) and boosting.

3.2 Neural Networks

These are non-linear predictive models that learn through training and resemble biological neutral networks in structure. Neural networks are approach to computing that involves developing mathematical structures with the ability to learn. This method is a result of academic investigations to model nervous system learning and has a remarkable ability to derive its meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either human or computer techniques. A trained neural network can be thought of as an expert in the class of information it wants to analysis. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

Neural networks have very wide applications to real world business problems and have already been implemented in many industries. Because neural networks are very good at identifying patterns or trend in data, they are very suitable for prediction or forecasting needs including the following:

- *Sales forecasting*
- *Customer research*
- *Data validation*
- *Risk management*
- *Industrial process control*
- *Target marketing*

Neural networks use a set of processing elements or nodes similar to neurons in human brain. The nodes are interconnected in a network that can then identify patterns in data once it is exposed to the data, that is to say network learns from experience like human beings. This makes neural networks to be different from traditional computing programs that follow instructions in a fixed sequential order.

The structure of a neural network is shown in figure 4.1. It starts with an input layer, where each mode corresponds to a prediction variable. These input nodes are connected to a number of nodes in a hidden layer. Each of the input nodes is connected to every node in the hidden layer. The nodes in that hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables.

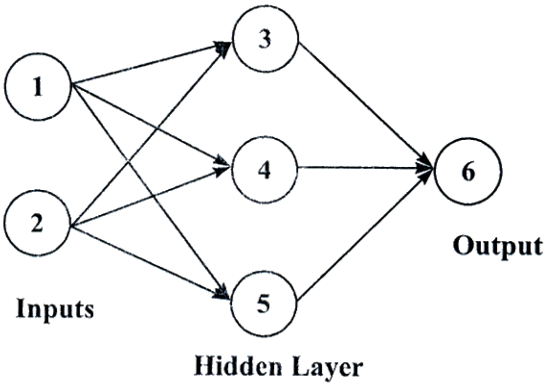


***Fig. 4.**1: Neural Network with One Hidden Layer*

***Source**: Introduction to Data Mining and Knowledge Discovery by Two Crows Corporation*

The commonest type of neural network is the feed-forward back propagation network and it proceeds as:

- *Feed forward: the value of the output made is calculated based on the input node value and a set of initial weights. The value from the input nodes are combined in the hidden layers and the* values of those nodes are combined to calculate the output value (Two Crows Corporation).

  Back-propagation: The error in the output is compiled by finding the difference between the calculated output and desired output that is the actual values found in training set.

This process is repeated for each row in the training set. Each pass through all rows in the training set is called an epoch. The training set is used repeatedly until the error is no longer decreasing. At that point the neutral net is considered to be trained to find the pattern in the test set. One major advantage of neural network models is that they can easily be implemented to run on massive parallel computers with each node simultaneously doing its own calculations.

The problems associated with neural networks as summed up by Arun Swami of Silicon Graphics Computer Systems are the resulting network is viewed as a black box and no explanation of the results is given. This lack of explanation inhibits confidence, acceptance and application of results. Also, neural networks suffered from long learning times which become worse as the volume of data grows.

3.3 Decision Trees

Decisions trees are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. It can also be described as a simple knowledge representation that classifies examples into a finite number of classes; the nodes are labeled with attribute names, the edges labeled with possible values for this attribute and the leaves with different classes. Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object. Decision trees handle non- numerical data very well.

Fig.4.2 illustrates a learned decision tree. We can see that each node represents an attribute or feature and the branch from each node represents the outcome of that node. Finally, its the leaves of the tree where the final decision is made. If features are continuous, internal nodes can test the value of a feature against a threshold

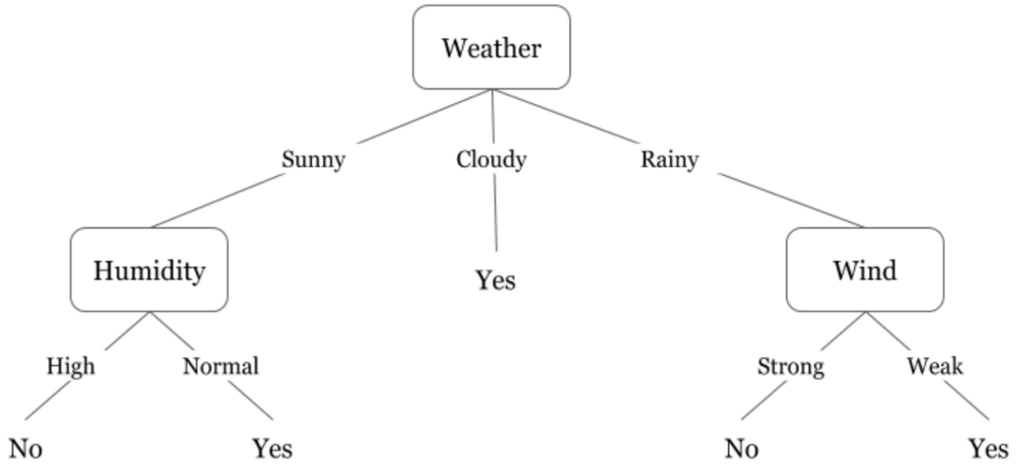


***Fig. 4.2**: Decision Tree Structure*

Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A good number of different algorithms may be used to build decision trees which include Chi-squared Automatic

Interaction Detection (CHAID), Classification and Repression Trees (CART), Quest and C5.0.

Decision trees grow through an iterative splitting of data into discrete groups, where the goal is to maximize the distance between groups at each split. Decision trees that are used to predict categorical variables are called classification trees because they place instances in categories or classes, and the one used to predict continuous variables are called repression trees.

3.4 Rule Induction

This is a method used to derive a set of rules for classifying cases. Although, decision trees can also produce set of rules but induction methods generate set of independent rules which does not force splits at each level but look ahead, it may be able to find different and sometimes better pattern for classification. Unlike trees, the rules generated may not be able to cover all possible situations and there may be conflict in their predictions, in which case it becomes necessary to choose which rule to follow. And one common method used in resolving conflicts is to assign a confidence to rule and used the one in which you are most confidence. An alternative method is if more than two rules conflict you may let them vote, perhaps weighting their votes by the confidence you have in each rule.

3.5 Multivariate Adaptive Repression Splines (MARS)

Jerome H. Friedman one of the inventors of CART (Classification and Regression Trees) developed in the mid-1980s a method designed to address the short coming of CART which are listed as follows:

- *discontinuous predictions (hard splits)*
- *dependence of all splits on previous ones*
- *reduced interpretability due to interactions, especially high-order interactions.*

To this end he developed the MARS algorithm which is able to take care of the CART disadvantages as follows:

- *it replaces the discontinuous branching at a node with continuous transition modelled by a pain of straight lines. At the end of the model-building process, the straight lines at each node are replaced with a very smooth function referred to as a spline*
- *does not require that the new splits be dependent on previous splits.*

The basic idea of MARS is simple, though loses the tree structure of CART and cannot produce rules. On the other hand, it automatically finds and lists the most important predictor variables as well as the interactions among predictor variables. MARS also plots the dependence off the response on each predictor. The result is an automatic non-liner step-wise regression tool.

Just like most neural and decision tree algorithms, MARS has a tendency to overfit the training data which can be addressed in two ways:

(i) Manual cross validation can be performed and the algorithms tuned to provide prediction on the test set.
(ii) There are various tuning parameters in the algorithm itself that can guide internal cross validation.

3.6 K-Nearest Neighbour and Memory-Based Reasoning (MBR)

K-nearest neighbour (k-NN) is a classification technique that uses the same method as when trying to solve new problem, people looks at solutions similar to the problems that they have previously solved. K-NN decides in which class to place a new case by examining some numbers - the "K" in K-nearest neighbour of the most similar cases or neighbours as shown in Figure 4.3. It counts the number of cases for each class and assigns the new case to the same class to which most of its neighbours belong.

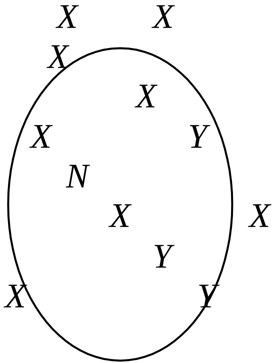


***Fig. 4.3**: K-Nearest Neighbour. N is a new case*
***Source**: Two Crows Corporation, 2005*

It would be assigned to the class X because the seven X's within the ellipse outline outnumber the two Y's

In order to apply K-NN, you must first of all find a measure of the distance between attributes in the data and then calculate it. While this is easy for numerical data, categorical variables need special handling. For example, what is the distance between blue and green? You must then have a way of summing the distance measures for the attributes. Once you can calculate the distance between cases, you then select the set of already classified cases to use as the basis for classifying new cases, decides how large a neighbourhood in which to do the comparisons, and also decide how to count the neighbours themselves. For instance, you might give more weight to nearer neighbours than farther neighbours. (Two crows Corporations)

With K-NN, a large computational load is placed on the computer because the calculation time increases as the factorial of the total number of points. While it's a rapid process to apply a decision tree or neutral net to a new case, K-NN requires that a new calculation be made for each new case. To speed up K-NN frequently all the data is kept in memory. Memory-based reasoning usually refers to a K-NN classifier kept in memory. (Two crow corporation)

The use of K-NN models are very easy to understand when there are few predictor variables, they are also useful for building models that involved non-standard data types, such as text. The only requirement to be able to include a data type is the existence of an appropriate metric.

3.7 Genetic Algorithms

Genetic algorithms are methods used for performing a guided search for good models in the solution space. They are not basically used to find patterns per se, but to guide the learning process of data mining algorithms like the neural nets. They are so-called genetic algorithms because they loosely follow the pattern of biological evolution in which the members of one generation of models compete to pass on their characteristic to the next generation, to pass on is contained in 'chromosomes' which contain the parameters for building the model.

For instance, to build a neural net, genetic algorithms can replace back propagation as a way to adjust the weights. The chromosomes would contain the number of hidden layers and the numbers of nodes in each layer. Although, genetic algorithms are interesting approach to optimizing models, but add a lot of computational overhead.

3.8 Discriminant Analysis

This is the oldest classification technique that was first published by R. A. Fisher in 1936 to classify the famous Iris botanical data into three species. Discriminant analysis finds hyper-planes e.g. lines in two dimensions, planes in three etc that separates the classes. The resultant model is very easy to interpret because what the user has to do is to determine on which side of the line (or hyper-plane) a point falls. Training on discriminant analysis is simple and scalable, and the technique is very sensitive to patterns in the data. This technique is applicable in some disciplines such as biology, medicine and social sciences.

3.9 Generalized Additive Models (GAM)

Generalized additive models or GAM is a class of models that extends both linear and logistics repression. They are so-called additive because we assume that the model can be written as the sum of possibly non- linear functions, one for each predictor. GAM can either be used for repression or for classification of a binary response. The response variable can be virtually any function of the predictors as long as there are not discontinuous steps. For example, suppose that payment delinquency is a rather complicated function of income where the probability of delinquency initially declines as income increases. It then turns around and starts to increase again for moderate income, finally peaking before coming down again for higher income card-holders. In such a case, linear model may fail to see any relationship between income and delinquency due to the non-linear behaviour.

With the use of computer power in place of theory or knowledge of the functional form, GAM will produce a smooth curve and summarize the relationship. As with neural nets where large numbers of parameters are estimated, GAM goes a step further and estimates a value of the output for each value of the input-one point, one estimate and generates a curve automatically choosing the amount of complexity based on the data.

3.10 Boosting

The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification. If you are to build a new model using one sample of data, and then build a new model using the same algorithms but on a different sample, you might get a different result. After validating the two models, you could choose the one that best meet your objectives. Better results might be achieved if several models are built and let them vote, making a prediction based on what the majority recommends. Of course any interpretability of the prediction would be lost, but the improved results might be worth it.

Boosting is a technique that was first published by Freund and Shapire in 1996; it takes multiple random samples from the data and builds a classification model for each. The training set is changed based on the result of the previous models. The final classification is the class assigned most often by the models. The exact algorithms for boosting have evolved from the original, but the underlying idea is the same. Boosting has become a very popular addition to data mining packages

3.11 Logistic Regression (Non-Linear Regression Methods)

This is a generalization of linear regression that is used primarily for predicting binary variables (with values such as yes/no or 0/1) and occasionally multi-class variables. Because the response variable is discrete, it cannot be modelled directly by linear regression. Therefore, instead of predicting whether the event itself (i.e. the response variable) will occur, we build the model to predict the logarithm of the odds of its occurrence. The logarithm is called the log odds or the logit transformation.

The odds ratio = probability of an event occurring

probability of the event not occurring

It has the same interpretation as in the more casual use of odds in the games of chance or sporting events. When we say that the odds are 3 to 1 that a particular team will win a soccer game, we mean that the probability of their winning is three times as great as the probability of their loosing. The same terminology can be applied to the chances of a particular type of customer (e.g. a customer with a given gender, income, mental status etc) replying to a mailing. If we say the odds are 3 to 1 that the customer will respond, we mean that the probability of that type of customer responding is three times as great as the probability of him or her not responding. Thus, this method has better chances of providing reliable solutions in such involved applications as financial markets or medical diagnostics.

4.0. SELF- ASSESSMENT EXERCISE

Briefly discuss the following data mining technologies:

i. *Neural networks*
ii. *Decision trees*

SELF- ASSESSMENT EXERCISE 2

Briefly explain any two of the following data mining technologies:

i. Rule induction
ii. Multivariate adaptive regression splines
iii. Genetics algorithm

TUTOR-MARKED ASSIGNMENT

List and explain any five data mining technologies and state an advantage of using such an algorithm.

5.0 CONCLUSION

Therefore, there is no one model or algorithm that should be used exclusively for data mining since there is no best technique. Consequently, one needs a variety of tools and technologies in order to find the best possible model for data mining.

6.0 SUMMARY

In this unit we have learnt that:

- *There are various techniques or algorithm used for mining data, this include neural networks, decisions trees, genetics algorithm, discriminant analysis, rule induction and the nearest neighbour.*

7.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Data Mining Techniques, Retrieved on 28/07/2009. From:

http://www.statsoft.com/TEXTBOOK/stdatmin.html.

Data Mining. Retrieved on 29/07/2009. Available Online: http://en.wikipedia.org/wiki/Data_mining.

Introduction to Data Mining and Knowledge Discovery. (3rd ed.). Mosud, Y.O. (2009). Introduction to Data Mining and Data Warehousing. Lagos: Rashmoye Publications

Sumathi,S. & Sivanamdam,S.N. (2006).*Introduction to Data Mining Principles, Studies in Computational Intelligence (SCI) 29, 1-20).*

MODULE 2 DATA MINING PROCESSES AND TRENDS

UNIT 1 DATA PREPARATION AND PREPROCESSING

CONTENTS

1.0 INTRODUCTION

Data preparation and preprocessing are often neglected but important step in data mining process, the phrase "Garbage in, Garbage out" (G1G1) is particularly applicable to data mining and machine learning projects. Data collection methods are often loosely controlled thereby resulting in out of range values (e.g. income – =N= 400), impossible data combinations (e.g. Gender: Male, Pregnant; yes), missing values and so on. This unit examines meaning and reasons for preparing and preprocessing data, cleaning, data transaction, data reduction and data discretization.

2.0. OBJECTIVES

At the end of this unit, you should be able to:

- *identify the different data formats of an attribute*
- *explain the meaning and importance of data preparation*
- *define term data preprocessing*
- *explain why data is being preprocessed*
- *state the various data pre-processing tasks.*

3.0 MAIN CONTENT

3.1 Data Types and Forms

In data mining, data is usually indicated in the attribute instance format, that is every instance (or data record) will have a certain fixed number of attributes (or fields). In data mining, attributes and instances are the terms used rather than fields or records, which are traditionally databases terminologies. An attribute can be defined as a descriptive property or characteristic of an entity. It may also be referred to as data item or field. An attribute can have different data formats, which can be summarised in the following hierarchy:

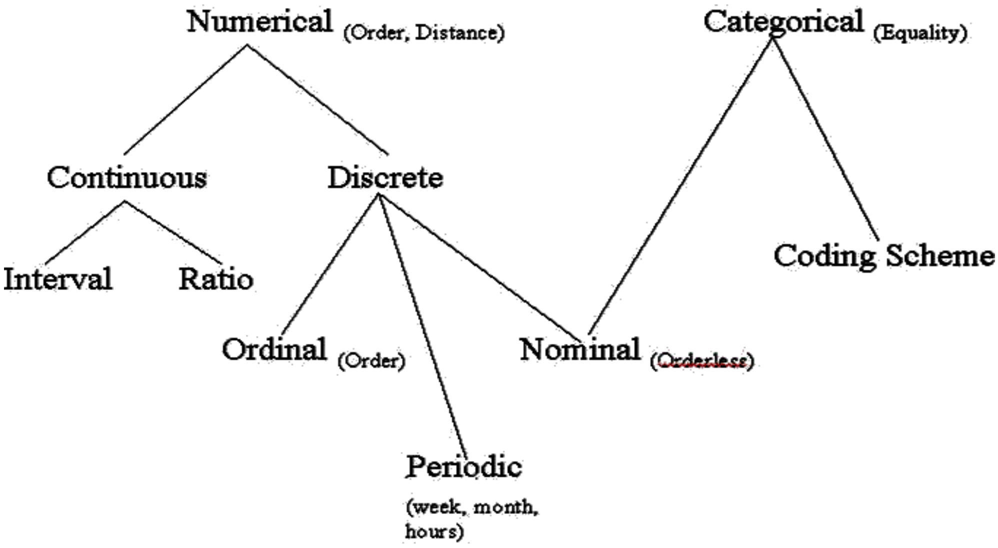


Fig.5.1: *Different Data Formats of an Attribute*
Source: *Introduction to Data Mining:http;//www.eas.asu.edu/ mining03/chap2/lesson_2. html*

Data can also be classified as static or dynamic (temporal). Other types of data that we come across in data mining applications are:

- *Distributed data*
- *Textual data*
- *Web data (e.g. html pages)*
- *Images*
- *Audio /Video*
- *Metadata (information about the data itself)*

3.2 Data Preparation

This is one of the most important tasks in data mining. It is time consuming exercise. The time factor is usually dependent on the size of the data we are concerned with. Datasets could be large in terms of two aspects, dimensionality or high number of instances. High dimensionality affects the time taken more than higher number of instances.

Other problems associated with data preparation are:

- *Missing data*
- *Outliers (data points inconsistent with the majority of the data points)*
- *Erroneous data (inconsistent, misreported or distorted).*

Data preparation is also required when data is to be processed is in the raw format, e.g. pixel format for images. Such data should be converted into appropriate formats which can processed by the data mining algorithms.

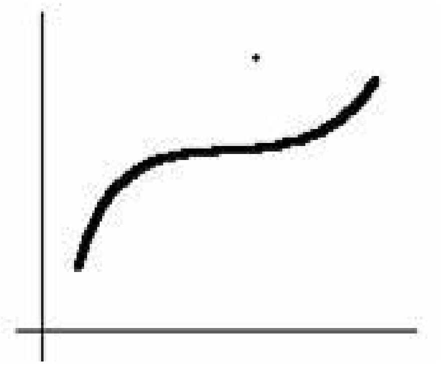
The common types of data preparation methods are:

- *Data normalization (e.g. for image mining)*
- *Dealing with sequential/temporal data*
- *Removing outliers*

3.2.1 Removing Outliers

Outliers are those data points which are inconsistent with the majority of the data points. There can be different kinds of outliers, some valid and some not. A valid example of an outlier is the salary of the CEO in an income attribute; which is normally higher than the other employees. While on the other hand an AGE attribute with value as 200 is obviously noisy and should be removed as an outlier. Some of the general methods used for removing outliers are:

- *Clustering: this can be used to cluster the relevant data points together and then use those cluster centers to find out the data points not close enough to them and then reject them as outliers.*
- *Curve–Fitting: this method initially uses regression analysis to find the curves which fit the data closely. It then removes all points (outliers), which are sufficiently far curve from the fitted curve*



- ❖ *Hypothesis–Testing with Given Model: in this case certain hypothesis are developed which need to be satisfied by the data domain. Then those data points which do not satisfy the hypothesis are rejected as outliers.*

3.3 Data Preprocessing

Data Preprocessing

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. "How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process?"

There are several data preprocessing techniques.

Data cleaning *can be applied to remove noise and correct inconsistencies in data***.**

Data integration *merges data from multiple sources into a coherent data store such as a data warehouse.*

Data reduction *can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.*

Data transformations *(e.g., normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements.*

These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining.

3.4 Data Quality: Why Preprocess the Data?

Data have quality if they satisfy the requirements of the intended use. Incomplete, noisy, and inconsistent data are commonplace properties of large real-world databases and data warehouses. Incomplete data can occur for a number of rea- sons. Attributes of interest may not always be available, such as customer information for sales transaction data. Other data may not be included simply because it was not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the his- tory or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred. There

There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.

The reasons for pre-processing data are stated as follows:

(i) *Real world data are generally dirty which is as a result of the following:*

- ❖ *Incomplete data: missing attributes, lacking attribute values, lacking certain attributes of interest, or containing only aggregated data.*
- ❖ *Inconsistent data: data containing discrepancies in codes or names (such as different coding, different naming, impossible values or out-of-range values)*
- ❖ *Noisy data: data containing errors, outliers, not accurate values*

(ii) *For quality mining results, quality data is needed*

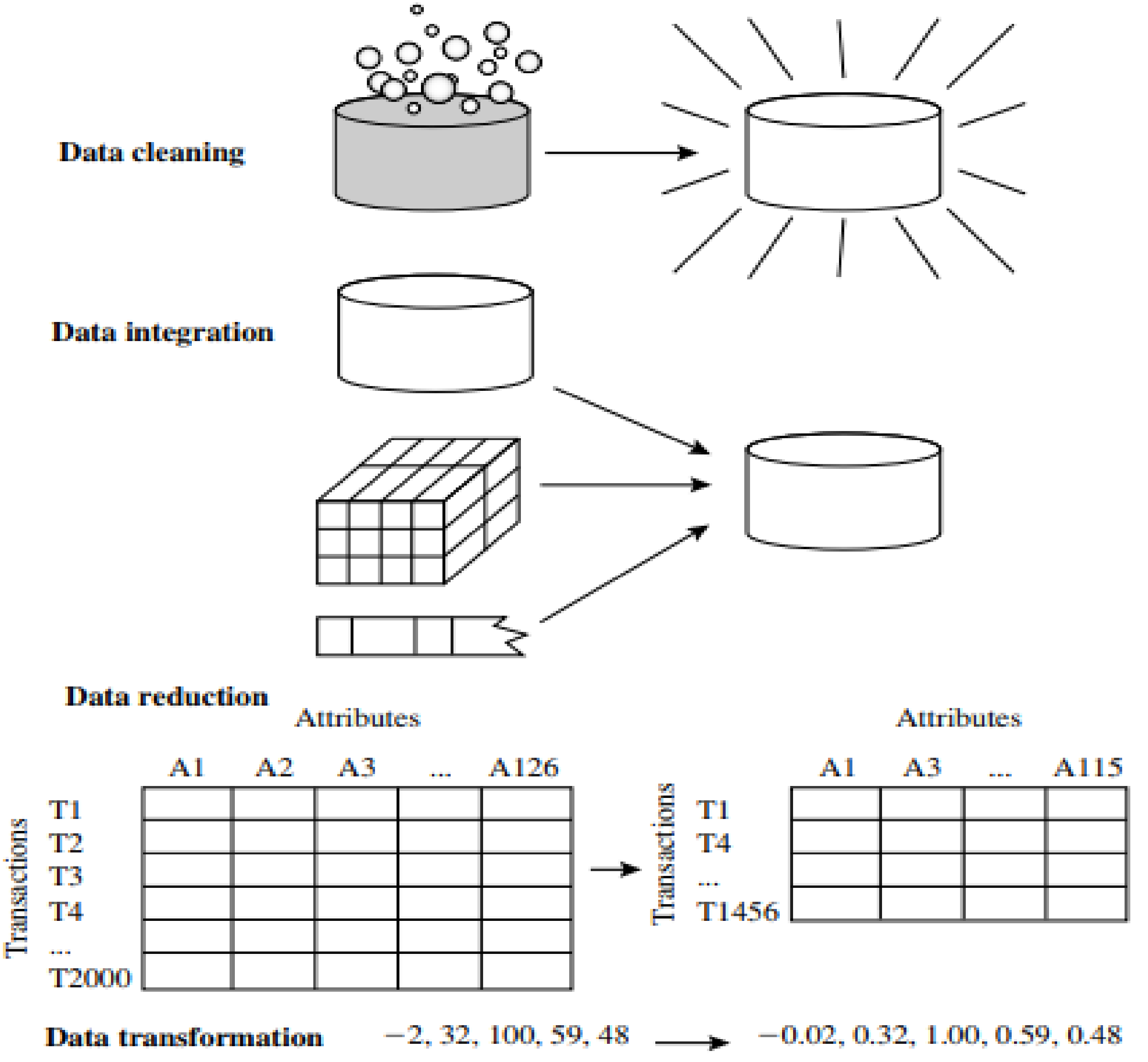
(iii) *Preprocessing is an important step for successful data mining.*

3.5 Major Tasks in Data Preprocessing

We look at the major steps involved in data preprocessing, namely, data cleaning, data integration, data reduction, and data transformation.

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Dirty data can cause confusion in the mining procedure, resulting in unreliable output. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding overfitting the data to the function being modeled. Therefore, a useful preprocessing step is to run your data through some data cleaning routines.

Discretization and concept hierarchy generation can also be useful, where raw data values for attributes are replaced by ranges or higher conceptual levels. For example, raw values for age may be replaced by higher-level concepts, such as youth, adult, or *seniorDiscretization and concept hierarchy generation are powerful tools for data mining in that they allow data mining at multiple abstraction levels. Normalization, data discretization, and concept hierarchy generation are forms of data transformation. You soon realize such data transformation operations are additional data preprocessing procedures that would contribute toward the success of the mining process.*



Forms of data preprocessing.

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. Data reduction strategies include data aggregation (building a data cube for example.) dimension reduction (e.g. Removing irrelevant attributes through correlation analysis), data compression (e.g. using encoding schemes such as minimum length encoding or wavelets) and numerosity reduction (e.g. "replacing" the data by alternative, smaller representations such as clusters or parametric models).

In summary, real-world data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve data quality, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an *important step in the knowledge discovery process, because quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making.*

3.6 Data Quality Measures

Some of the factors used in measuring the quality of a data are:

- ❖ *Accuracy*
- ❖ *Completeness*
- ❖ *Consistency*
- ❖ *Timeliness*
- ❖ *Believability*
- ❖ *Interpretability*
- ❖ *Accessibility*

3.7 Data Preprocessing Tasks

The various tasks involved in data preprocessing are stated as follows:

- ❖ *Data cleaning*
- ❖ *Data transformation*
- ❖ *Attribute/ Feature construction*
- ❖ *Data reduction*
- ❖ *Discretisation and concept hierarchy generation*
- ❖ *Data parsing and standardization*
- ❖ *Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies*
- ❖ *Data integration*
- ❖ *Normalization and aggregation*

SELF- ASSESSMENT EXERCISE 1

i. *What is the importance of data preparation in data mining?*
ii. *List and explain the common types of data preparation method*
iii. *List and explain tasks involved in* data preprocessing
iv. *Why Preprocess the Data in data mining*

3.7.1 Data Cleaning

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Missing Values

The various methods for handling the problem of missing values in data tuples include:

***(a) Ignoring the tuple**:*

This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

(b) Manually filling in the missing value:

In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

(c) Using a global constant to fill in the missing value:

Replace all missing attribute values by the same constant, such as a label like "Unknown," or −∞. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common that of "Unknown." Hence, although this method is simple, it is not recommended.

(a) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple:

For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

***(b) Using the most probable value to fill in the missing value**:*

This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

- ***Noisy data:***

Noise is a random error or variance in a measured variable. Data smoothing tech is used for removing such noisy data.

3.7.2 Several Data smoothing technique

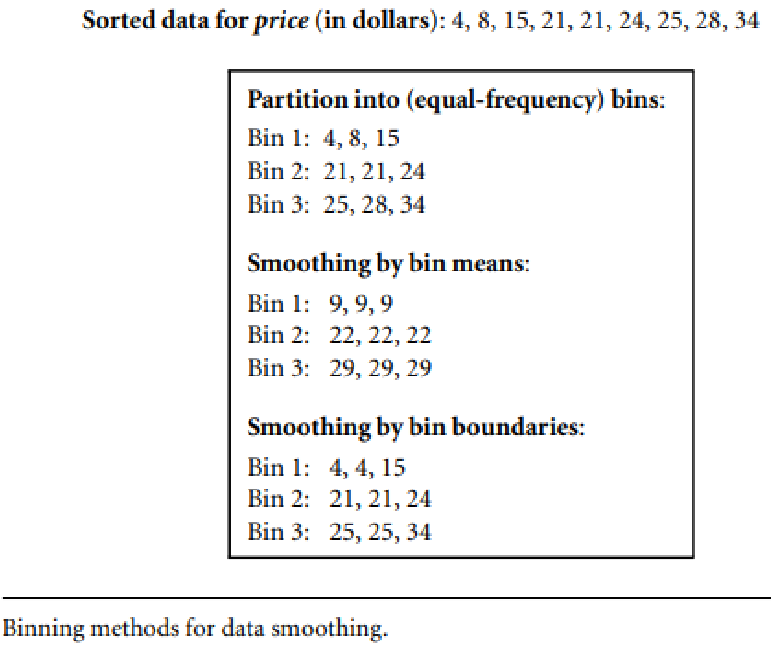
1. Binning methods:

Binning methods smooth a sorted data value by consulting the neighborhood", or values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighborhood of values, they perform local smoothing.
In this technique,

i. The data for first sorted
ii. Then the sorted list partitioned into equi-depth of bins.
iii. Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

a. Smoothing by bin means: Each value in the bin is replaced by the mean value of the bin.
b. Smoothing by bin medians: Each value in the bin is replaced by the bin median.
c. Smoothing by boundaries: The min and max values of a bin are identified as the bin boundaries. Each bin value is replaced by the closest boundary value.

Example:



Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

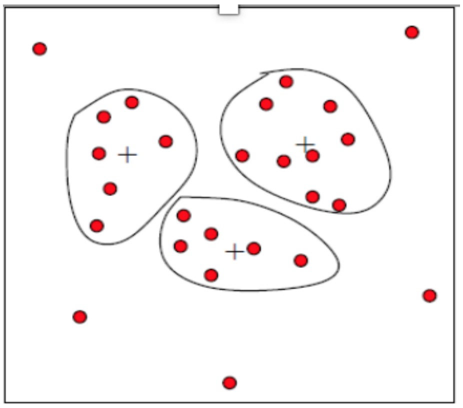
Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Binning methods for data smoothing.

2. Clustering:

Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.



Detecting of outlier by cluster analysis

***Combined computer and human inspection**: Outliers may be identified through a combination of computer and human inspection.*

In one application, for example, an information-theoretic measure was used to help identify outlier patterns in a handwritten character database for classification. The measure's value reflected the "surprise" content of the predicted character label with respect to the known label. Outlier patterns may be informative (e.g., identifying useful data exceptions, such as different versions of the characters "0"or "7"), or "garbage" (e.g., mislabeled characters). Patterns whose surprise content is above a threshold are output to a list. A human can then sort through the patterns in the list to identify the actual garbage ones.

This is much faster than having to manually search through the entire database. The garbage patterns can then be removed from the (training) database. The garbage patterns can be excluded from use in subsequent data mining.

3. Regression:

Data smoothing can also be done by regression, a technique that conforms data values to a function. Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Regression: smooth by fitting the data into regression functions.

Linear regression involves finding the best of line to fit two variables, so that one variable can be used to predict the other.

Multiple Linear regression is an extension of linear regression, where more than two variables are involved and the data are fit to a multidimensional surface.
Using regression to find a mathematical equation to fit the data helps smooth out the noise.

3.7.3 Data Cleaning as a Process

The first step in data cleaning as a process is discrepancy detection. Discrepancies can be caused by several factors, including poorly designed data entry forms that have many optional fields, human error in data entry, deliberate errors (e.g., respondents not wanting to divulge information about themselves), and data decay (e.g., outdated addresses). Discrepancies may also arise from inconsistent data representations and the inconsistent use of codes.
"So, how can we proceed with discrepancy detection?" As a starting point, use any knowledge you may already have regarding properties of the data. Such knowledge or "data about data" is referred to as metadata.
The data should also be examined regarding unique rules, consecutive rules, and null rules.

***Field overloading**: is a kind of source of errors that typically occurs when developers compress new attribute definitions into unused portions of already defined attributes.*

***Unique rule** is a rule says that each value of the given attribute must be different from all other values of that attribute*

***Consecutive rule** is a rule says that there can be no missing values between the lowest and highest values of the attribute and that all values must also be unique.*

***Null rule** specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition and how such values should be handled.*

There are a number of different commercial tools that can aid in the step of discrepancy detection. Data scrubbing tools use simple domain knowledge (e.g., knowledge of postal addresses, and spell-checking) to detect errors and make corrections in the data. These tools rely on parsing and fuzzy matching techniques when cleaning data from multiple sources. Data auditing tools find discrepancies by analyzing the data to discover rules and relationships, and detecting data that violate such conditions. They are variants of data mining tools.

3.7.4 Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Strategies for data transformation include the following:

- *Smoothing, which works to remove noise from the data. Techniques include binning, regression, and clustering.*

- *Attribute construction (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.*
- *Aggregation, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for data analysis at multiple abstraction levels.*
- *Normalization, where the attribute data are scaled so as to fall within a smaller range, such as -1.0 to 1.0, or 0.0 to 1.0.*

- *Discretization, where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The labels, in turn, can be recursively organized into higher level concepts, resulting in a concept hierarchy for the numeric attribute.*
- *Concept hierarchy generation for nominal data, where attributes such as street can be generalized to higher-level concepts, like city or country. Many hierarchies for nominal attributes are implicit within the database schema and can be automatically defined at the schema definition level.*

3.9.9 Data Reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Strategies for data reduction include the following:

1. ***Data cube aggregation**, where aggregation operations are applied to the data in the construction of a data cube.*
2. ***Attribute subset selection**, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.*
3. ***Dimensionality reduction**, where encoding mechanisms are used to reduce the data set size.*
4. ***Numerosity reduction**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.*
5. ***Discretization and concept hierarchy generation**, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction*

3.7.6 Data Integration

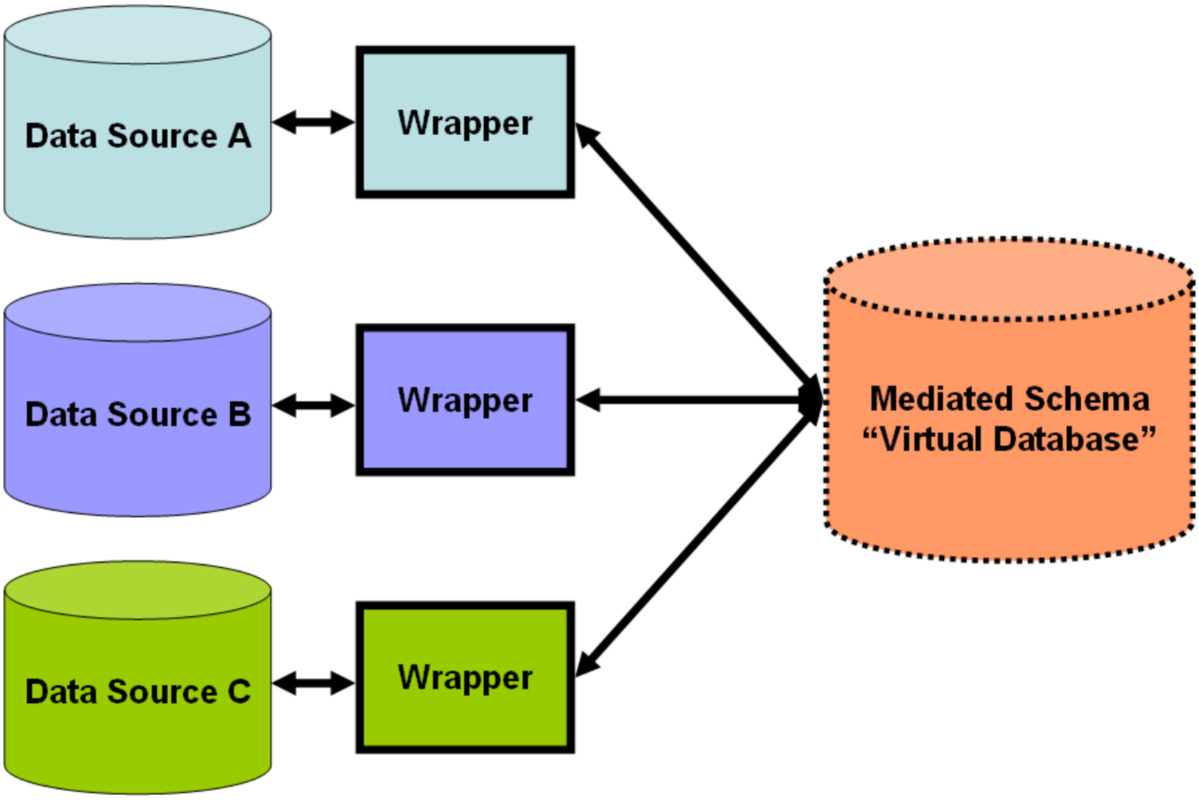
It combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files.

The data integration systems are formally defined as triple<G, S, M>

Where G: The global schema

S: Heterogeneous source of schemas

M: Mapping between the queries of source and global schema



1. ***Schema integration and object matching:***
How can the data analyst or the computer be sure that customer id in one database and customer number in another reference to the same attribute?

2. ***Redundancy:***
An attribute (such as annual revenue, for instance) may be redundant if it can be derived from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

3. ***detection and resolution of data value conflicts:***
For the same real-world entity, attribute values from different sources may differ.

3.8 *Measure the Central Tendency*

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location.

In other words, in many real-life situations, it is helpful to describe data by a single number that is most representative of the entire collection of numbers. Such a number is called a measure of central tendency. The most commonly used measures are as follows. ***Mean, Median, and Mode***

3.8.1 Mean

mean, or average, of numbers is the sum of the numbers divided by n. That is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n} \quad \text{i.e.,} \quad \text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

shortly,

$$\bar{x} = \frac{\sum x}{n}$$

where $\bar{x}$ (read as 'x bar') is the mean of the set of x values,

$\sum x$ is the sum of all the x values, and

n is the number of x values.

Example 1

The marks of seven students in a mathematics test with a maximum possible mark of 20 are given below:

15 13 18 16 14 17 12

Find the mean of this set of data values.

Solution:

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$= \frac{15+13+18+16+14+17+12}{7}$$

$$= \frac{105}{7}$$

$$= 15$$

So, the mean mark is 15

Midrange

The midrange of a data set is the average of the minimum and maximum values.

3.8.2 Median

median of numbers is the middle number when the numbers are written in order. If is even, the median is the average of the two middle numbers.

Example 2 The marks of nine students in a geography test that had a maximum possible mark of 50 are given below:

47 35 37 32 38 39 36 34 35

Find the median of this set of data values.

Solution:

Arrange the data values in order from the lowest value to the highest value:

32 34 35 35 36 37 38 39 47

The fifth data value, 36, is the middle value in this arrangement.

Median = 36

Note:

The number of values, n in the data set is 9

$$\text{Median} = \frac{1}{2}(9+1)\text{th value}$$
$$= 5^{th} \text{ value}$$
$$= \qquad 36$$

$$\text{Median} = \frac{1}{2}(n+1)\text{th value, where } n \text{ is the number of data values in the sample}$$

If the number of values in the data set is even, then the median is the average of the two middle values.

Example 3

Find the median of the following data

set: 12 18 16 21 10 13 17 19

Solution:

Arrange the data values in order from the lowest value to the highest value:

10 12 13 16 17 18 19 21

The number of values in the data set is 8, which is even. So, the median is the average of the two middle values.

$$\text{Median} = \frac{4^{th}\text{ data value} + 5^{th}\text{ data value}}{2}$$

$$= \frac{16+17}{2}$$

$$= \frac{33}{2}$$

$$= 16.5$$

3.8.3 Trimmed mean

A trimming mean eliminates the extreme observations by removing observations from each end of the ordered sample. It is calculated by discarding a certain percentage of the lowest and the highest scores and then computing the mean of the remaining scores.

3.8.4 Mode

mode of numbers is the number that occurs most frequently. If two numbers tie for most frequent occurrence, the collection has two modes and is called bimodal.

The mode has applications in printing. For example, it is important to print more of the most popular books; because printing different books in equal numbers would cause a shortage of some books and an oversupply of others.

Likewise, the mode has applications in manufacturing. For example, it is important to manufacture more of the most popular shoes; because manufacturing different shoes in equal numbers would cause a shortage of some shoes and an oversupply of others.

Example 4

Find the mode of the following data set:

48 44 48 45 42 49 48

Solution:

The mode is 48 since it occurs most often.

- *It is possible for a set of data values to have more than one mode.*
- *If there are two data values that occur most frequently, we say that the set of data values is **bimodal**.*
- *If there are three data values that occur most frequently, we say that the set of data values is **trimodal***

- *If two or more data values that occur most frequently, we say that the set of data values is* ***multimodal***
- *If there is no data value or data values that occur most frequently, we say that the set of data values has no mode.*

The mean, median and mode of a data set are collectively known as measures of ***central tendency*** *as these three measures focus on where the data is centred or clustered. To analyse data using the mean, median and mode, we need to use the most appropriate measure of central tendency. The following points should be remembered:*

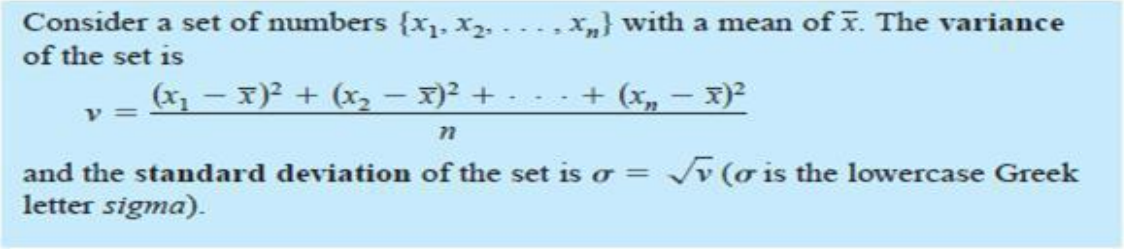
- *The mean is useful for predicting future results when there are no extreme values in the data set. However, the impact of extreme values on the mean may be important and should be considered. E.g. the impact of a stock market crash on average investment returns.*
- *The median may be more useful than the mean when there are extreme values in the data set as it is not affected by the extreme values.*
- *The mode is useful when the most common item, characteristic or value of a data set is required.*

3.9 Measures of Dispersion

Measures of dispersion measure how spread out a set of data is. The two most commonly used measures of dispersion are the variance and the standard deviation. Rather than
showing how data are similar, they show how data differs from its variation, spread, or dispersion.
Other measures of dispersion that may be encountered include the Quartiles, Inter quartile range (IQR), Five number summary, range and box plots

3.9.1 Variance and Standard Deviation

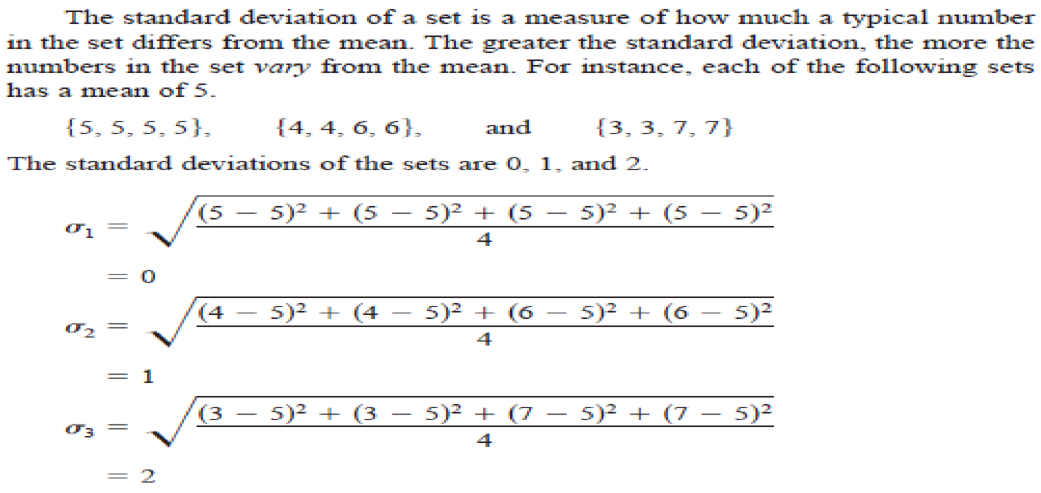
Very different sets of numbers can have the same mean. You will now study two measures of dispersion, which give you an idea of how much the numbers in a set differ from the mean of the set. These two measures are called the variance of the set and the standard deviation of the set



Consider a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with a mean of $\overline{x}$. The **variance** of the set is

$$v = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdot \cdot \cdot + (x_n - \overline{x})^2}{n}$$

and the **standard deviation** of the set is $\sigma = \sqrt{v}$ (σ is the lowercase Greek letter *sigma*).



The standard deviation of a set is a measure of how much a typical number in the set differs from the mean. The greater the standard deviation, the more the numbers in the set *vary* from the mean. For instance, each of the following sets has a mean of 5.

$$\{5, 5, 5, 5\}, \qquad \{4, 4, 6, 6\}, \qquad \text{and} \qquad \{3, 3, 7, 7\}$$

The standard deviations of the sets are 0, 1, and 2.

$$\begin{aligned} \sigma_1 &= \sqrt{\frac{(5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2}{4}} \\ &= 0 \\ \sigma_2 &= \sqrt{\frac{(4-5)^2 + (4-5)^2 + (6-5)^2 + (6-5)^2}{4}} \\ &= 1 \\ \sigma_3 &= \sqrt{\frac{(3-5)^2 + (3-5)^2 + (7-5)^2 + (7-5)^2}{4}} \\ &= 2 \end{aligned}$$

Percentiles are values that divide a sample of data into one hundred groups containing (as far as possible) equal numbers of observations.

The Pth percentile of a distribution is the value such that p percent of the observations fall at or below it.

The most commonly used percentiles other than the median are the 25th percentile and the 75th percentile.

The 25th percentile demarcates the first quartile, the median or 50th percentile demarcates
the second quartile, the 75th percentile demarcates the third quartile, and the 100th percentile
demarcates the fourth quartile.

3.9.2 Quartiles

Quartiles are numbers that divide an ordered data set into four portions, each containing approximately one-fourth of the data. Twenty-five percent of the data values come before the first quartile (Q1). The median is the second quartile (Q2); 50% of the data values come before the median. Seventy-five percent of the data values come before the third quartile (Q3).

*Q1=25th percentile=(n*25/100), where n is total number of data in the given data set*
*Q2=median=50th percentile=(n*50/100)*
*Q3=75th percentile=(n*75/100)*

3.9.3 Range

The range of a set of data is the difference between its largest (maximum) and smallest (minimum) values. In the statistical world, the range is reported as a single number, the difference between maximum and minimum. Sometimes, the range is often reported as "from (the minimum) to (the maximum)," i.e., two numbers.

Example 1:
Given data set: 3, 4, 4, 5, 6, 8

The range of data set is 3–8. The range gives only minimal information about the spread of the data, by defining the two extremes. It says nothing about how the data are distributed between those two endpoints.

Example 2:

In this example we demonstrate how to find the minimum value, maximum value, and range of the following data: 29, 31, 24, 29, 30, 25

Arrange the data from smallest to largest.

24, 25, 29, 29, 30, 31

2. Identify the minimum and maximum values:

Minimum = 24, Maximum = 31

3. Calculate the range:

Range = Maximum-Minimum = 31–24 = 7.
Thus, the range is 7.

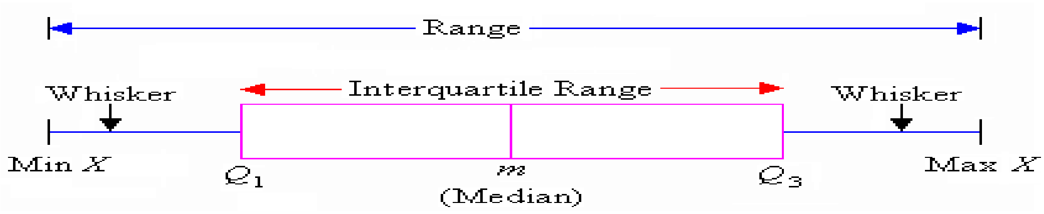
3.9.4 Box plots

Box plots (also called box-and-whisker plots or box-whisker plots) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value.

A box plot is a graph used to represent the range, median, quartiles and inter quartile range of a set of data values.

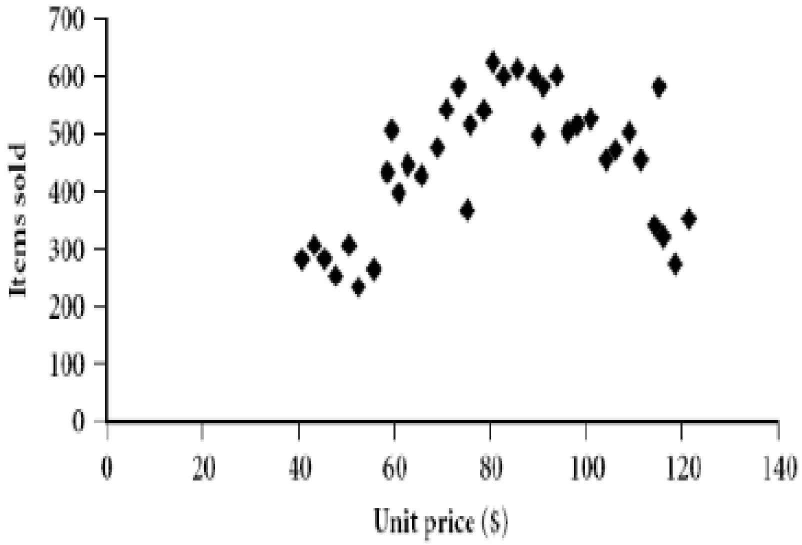
Constructing a Box plot: To construct a box plot:

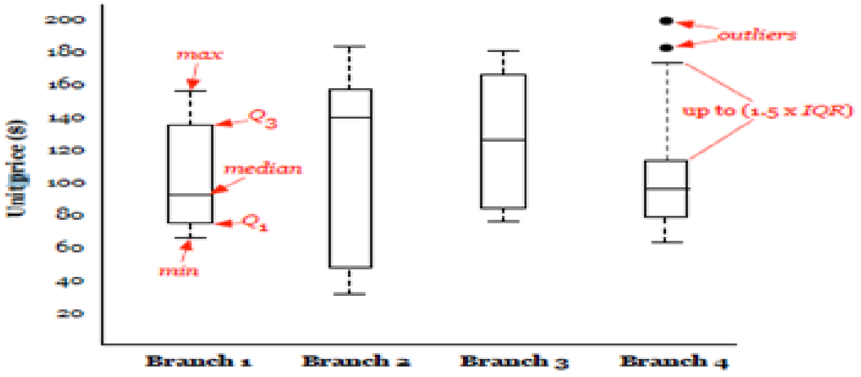
i. *Draw a box to represent the middle 50% of the observations of the data set.*
ii. *Show the median by drawing a vertical line within the box.*
iii. *Draw the lines (called* ***whiskers****) from the lower and upper ends of the box to the minimum and maximum values of the data set respectively, as shown in the following*
diagram.



- X is the set of data values.
- Min X is the minimum value in the data Max X is the maximum value in the data set.

The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median.





Example: Draw a boxplot for the following data set of scores:

76 79 76 74 75 71 85 82 82 79 81

***Step 1**: Arrange the score values in ascending order of magnitude:*

71 74 75 76 76 79 79 81 82 82 85

There are 11 values in the data set.

***Step 2**: Q1=25th percentile value in the given data set*
Q1=11(25/100) th value*

=2.75 =>3rd value

=75

***Step 3:** Q2=median=50th percentile value*

*=11 * (50/100) th value*

=5.5th value => 6th value

=79

Step 4:

Q3=75th percentile value

=11(75/100)th value*

=8.25th value=>9th value

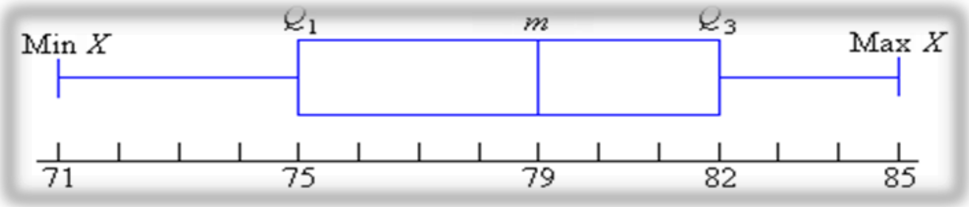
= 82

***Step 5**: Min X= 71*

***Step 6**: Max X=85*

***Step 7**: Range= 85-71 = 14*

***Step 5**: IQR=height of the box=Q3-Q1=9-3=6th value=79*



Since the medians represent the middle points, they split the data into four equal parts. In other words

- *one quarter of the data numbers are less than 75*
- *one quarter of the data numbers are between 75 and 79*
- *one quarter of the data numbers are between 79 and 82*
- *one quarter of the data numbers are greater than 82*

3.9.5 Outliers

Outlier data is a data that falls outside the range. Outliers will be any points below Q1 – 1.5×IQR or above Q3 + 1.5×IQR.

Example:

Find the outliers, if any, for the following data set:

10.2, 14.1, 14.4, 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4

To find out if there are any outliers, I first have to find the IQR. There are fifteen data points, so the median will be at position

(15/2) = 7.5

=8th value

=14.6.

That is, Q2 = 14.6.

Q1 is the fourth value in the list and Q3 is the twelfth: Q1 = 14.4 and Q3 = 14.9.

Then IQR = 14.9 – 14.4 = 0.5.

Outliers will be any points below:

Q1 – 1.5×IQR = 14.4 – 0.75 = 13.65 or above Q3 + 1.5×IQR = 14.9 + 0.75 = 15.65.

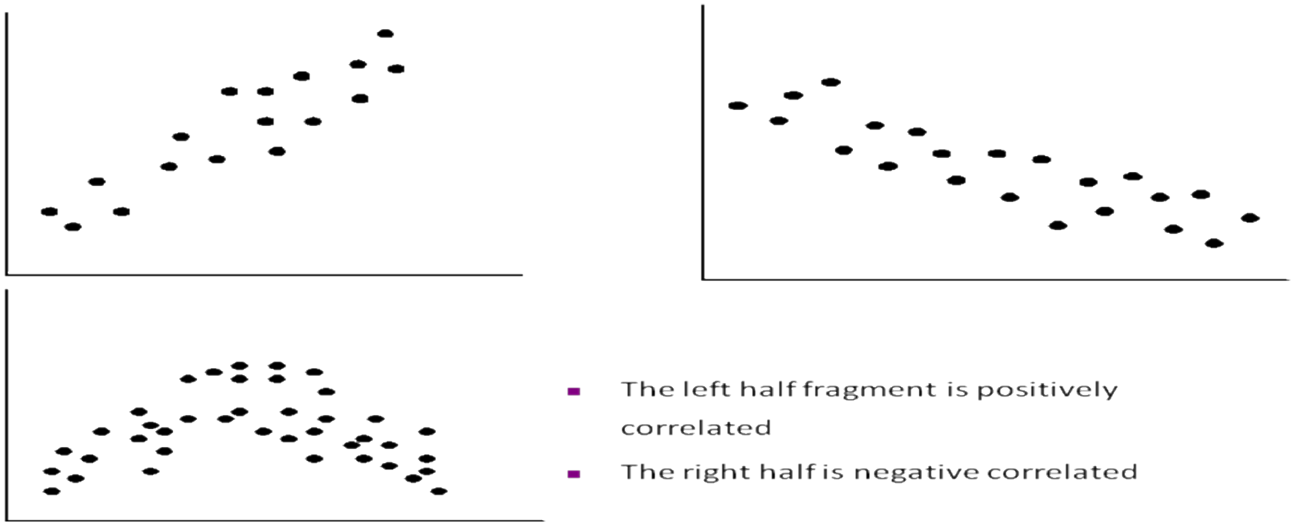
Then the outliers are at 10.2, 15.9, and 16.4.

The values for Q1 – 1.5×IQR and Q3 + 1.5×IQR are the "fences" that mark off the "reasonable" values from the outlier values. Outliers lie outside the fences.

3.9.6 Scatter Plot

A scatter plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.

Each unit contributes one point to the scatter plot, on which points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between the two variables.



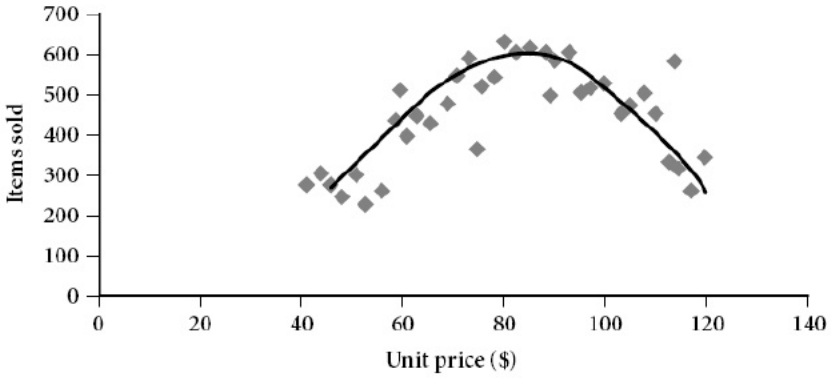
- The left half fragment is positively correlated
- The right half is negative correlated

Positively and Negatively Correlated Data

A scatter plot will also show up a non-linear relationship between the two variables and whether or not there exist any outliers in the data.

3.9.7 Loess curve

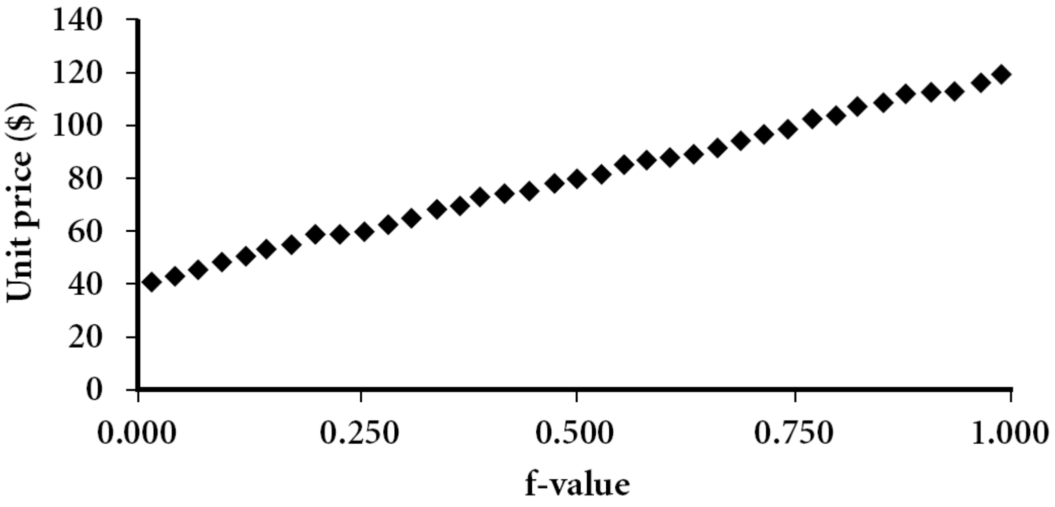
It is another important exploratory graphic aid that adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence. The word loess is short for "local regression."



3.9.8 Quintile plot

- *Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)*
- *Plots quintile information*

- *For a data* xi *data sorted in increasing order,* fi *indicates that approximately 100* fi*% of the data are below or equal to the value* xi



The f *quintile of the data is found. That data value is denoted q(f). Each data point can be*

assigned an f*-value. Let a time series* x *of length* n *be sorted from smallest to largest values,*

such that the sorted values have rank. The f*-value for each observation is computed as . 1,2,..., n . The* f*-value for*

$$f_i = \frac{i - 0.5}{n}$$

each observation is computed as,

4.0 SELF- ASSESSMENT EXERCISE

i. State the reasons for pre-processing a data

ii. Discuss issues to consider during data integration.

iii. List and explain the various tasks involved in data pre-processing

iv. Draw a boxplot for the following data set of scores
7, 9, 8, 8, 8, 6, 6, 5, 4
89, 90, 92, 95, 95, 98, 102, 103
18, 20 ,22 ,23 ,25 ,29 ,30 ,30 ,30, 31
1, 3, 4, 4, 5, 5, 5, 7, 11, 15

v. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
(a) What is the mean of the data? What is the median?
(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
(c) What is the midrange of the data? (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data? (e) Give the five-number summary of the data. (f) Show a boxplot of the data. (g) How is a quantile-quantile plot different from a quantile plot?

TUTOR- MARKED ASSIGNMENT

i. List some of the factors used in measuring the quality of a data.
ii. List and explain some data preparation methods
iii. Discuss issues to consider during data integration.
iv. Briefly discuss the different types of data normalization methods.
v. Outliers are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Taking fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

5.0 CONCLUSION

Therefore, data preparation and preprocessing are very important step in data mining process.

6.0 SUMMARY

In this unit we have learnt that:

- *Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Data cleaning is usually performed as an iterative two-step process consisting of discrepancy detection and data transformation.*
- *Data preprocessing is an important issue for both data warehousing and data mining, as real-world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation, and data reduction.*
- *Descriptive data summarization provides the analytical foundation for data pre- processing. The basic statistical measures for data summarization include mean, weighted mean, median, and mode for measuring the central tendency of data, and range, quartiles, interquartile range, variance, and standard deviation for measur- ing the dispersion of data. Graphical representations, such as histograms, boxplots, quantile plots, quantile-quantile plots, scatter plots, and scatter-plot matrices, facili- tate visual inspection of the data and are thus useful for data preprocessing and mining*
- *data preparation is one of the important tasks in data mining*
- *data has to be prepared because of a lot of reason these include real world data is dirty, incomplete data and noisy data.*
- *Data integration combines data frommultiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution ofsemantic heterogeneity contribute toward smooth data integration*
- *Data transformation routines convert the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between a small range, such as 0.0 to 1.0*

- *Data reduction techniques such as data cube aggregation, attribute subset selection, dimensionality reduction, numerosity reduction, and discretization can be used to obtain a reduced representation ofthe data while minimizing the loss ofinformation content.*
- *Although numerous methods of data preprocessing have been developed, data preprocessing remains an active area of research, due to the huge amount of inconsistent or dirty data and the complexity of the problem.*

7.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Aggarwal, C. C. (2015). Data Mining: The Textbook. Cham: Springer. ISBN: 978-3-319-14141-1

Zaki, M., & Meira, Jr, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511810114

UNIT 2 DATA MINING PROCESS

CONTENTS

1.0 INTRODUCTION

It is very crucial to recognize the fact that a systematic approach is essential for a successful data mining; although, many vendors and consulting organisations have specified a process designed to guide the user, especially someone new to building predictive models through a sequence of steps that will lead to good results. This unit examines the necessary steps in successful data mining using 'the two crows process model'

2.0 OBJECTIVE

At the end of this unit, you should be able to:

- *state the basic steps of data mining for knowledge discovery*

3.0 MAIN CONTENT

3.1 Process Models

The use of a systematic approach is essential for a successful data mining. A lot of vendors and consortium of organisations have specified a process model designed to guide the user in achieving a good result.

Recently, a consortium of vendors and users that consist of NCR Systems Engineering Copenhagen (Denmark), Daimler-Benz AG (Germany), SPSS/ Integral Solutions Ltd. (England) and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands) has been developing a specification called CRISP-DM (Cross Industry Standard Process for Data Mining). SPSS uses the 5'As- Assess, Access, Analyse, Act and Automatic and SAS uses SEMMA- Sample, Explore, Modify, Model, Assess. CRISP-DM is similar to process models from other companies including the one from two crow's corporations. As of September 1999, CRISP-DM was a work in progress. (Two Crows Corporation, 2005)

3.2 The Two Crows Process Model

The Two Crows data mining process model described in this section takes advantage of some insights from CRISP-DM and from its previous version. The Two Crow's process model involves a list of steps but does not connote that data mining process is a linear one, but needs to be looped back to previous steps. For example, what you learn in the 'explore data' step may require you to add new data to the data mining database. The initial models you build may provide insights that lead you to create new variables.

The basic steps of data mining for knowledge discovery are:

- *define business problem*

- *build data mining database*
- *explore data*
- *prepare data for modelling*
- *build model*
- *evaluate model*
- *deploy model and results.*

3.3 Define the Business Problem

The first step which is of course a prerequisite to knowledge discovery is for you to understand your data and your business. Without this understanding, there will be no algorithm or technique regardless of sophistication is going to provide you with a result in which you should have confidence. Without this background you will not be able to identify the problems you are trying to solve, prepare the data for mining or correctly interpret the results.

In order to make best use of data mining you must be wishing to increase the respond to a direct mail campaign. Depending on your specific goal such as "increase the response rate" or 'increasing the value of a response" you will build a very different model. An effective statement of the problem will include a way of measuring the results of the knowledge discovery project, which may also include a cost justification.

3.4 Building a Data Mining Database

These steps together with the next two which are "explore the data" and "prepare the data for modelling" constitute the core of data preparation. This takes more time and effort than all other steps combined. Although, there may be repeated iterations of the data preparation and model building step as one learns something from the model that suggests you modify the data. The data preparation steps may take anything from 50% to 90% of the time and effort of the entire data mining process.

The data to mine should be collected in a database and this does not necessarily mean that a database management system must be used. Depending on the amount of data, complexity of the data and use to which it is to be put, a flat file or even a spreadsheet may be adequate. By and large, it is not a good idea to use your corporate data warehouse for this, it is better to create a separate data mart. Almost certainly you will be modifying the data from the data warehouse. You may also want to bring in data from outside your company to overlay on the data warehouse data or you may want to add new fields computed from existing fields. You may need to gather additional data through surveys.

Other people building different models from the warehouse (some of whom will use the same data as you) may want to make similar alterations to the warehouse. However, data warehouse administrators do not look kindly on having data changed in what is unquestionably a corporate resource (Two Crows, 2005).

Another reason for a separate database is that the structure of the data warehouse may not easily support the kinds of exploration you need to do to understand this data. This include queries summarising the data, multi-dimensional reports (which is sometimes referred to as pivot tables), and many different kinds of graphs or visualisation. Also, you may want to *store this data in different database management system (DBMS) with a different physical design than the one used for your corporate data warehouse.*

The various tasks in building a data mining database are:

- *Data collection*
- *Data description*
- *Selection*
- *Data quality assessment and data cleansing*
- *Consolidation and integration*
- *Metadata construction*
- *Load the data mining database*
- *Maintain the data mining database.*

These aforementioned tasks are not performed in strict sequence, but as the need arises, for example you will start constructing the metadata infrastructure as you collect the data and continue to modify it

SELF -ASSESSMENT EXERCISE 1

i. List all explain the various tasks in building a data mining database
ii. List and explain basic steps of data mining for knowledge discovery

3.4.1 Data Collection

There is need to identify the sources of the data you want to mine, though a data- gathering phase may become very necessary because some of the data you need may never have been collected. Also, you may need to acquire external data from public databases (such as census or whether data) or proprietary databases (such as credit bureau data).

A data collection report (DCR) lists the properties of different source data sets. Some of the elements in this report include the following:

- *Source of data (either internal application or outside vendor)*
- *Owner*
- *Person/organisation responsible for maintaining the data*
- *Database administration (DBA)*
- *Cost (if purchased)*
- *Storage organisation (oracle database, VSAM file etc)*
- *Size in table, rows, records etc.*
- *Size in bytes*
- *Physical storage (CD-ROM, tape, server etc)*
- *Security requirements*
- *Restrictions on use*
- *Privacy requirements*

You should be sure to take note of special security and private issues that your data mining database will inherit from the source data. For example, some countries datasets are constrained in their use by privacy regulations.

3.4.2 Data Description

This describes the contents of each file or database table. Some of the properties that are documented in a typical Data Description Report are:

- *Number of fields / columns*
- *Number / percentage of records with missing values*
- *Field names*

  For each field:
  - *Data type*
  - *Definition*
  - *Description*
  - *Source of field*
  - *Unit of measure*
  - *Number of unique values*
  - *List of values*
  - *Range of values*
  - *Number / percentage of missing values*
  - *Collection information (e.g. how, where, conditions)*
  - *Time frame (e.g. daily, weekly, monthly)*
  - *Specific time data (e.g. every Monday or every Tuesday)*
  - *Primary key / foreign key relationships (two corporation).*

3.4.3 Selection

The next step after describing the data is selecting the subset of data to mine. This is not the same as sampling the database or choosing prediction variables. Instead, it is a gross elimination of irrelevant or unrequired data. Other criteria for excluding data may include resource constraints, cost, restrictions on data use, or quality problems

3.4.4 Data Quality Assessment and Data Cleansing

The term GIGO (Garbage in, Garbage out) is also applicable to data mining, so if you want good models you need to have good data. Data quality assessment identifies the features of the data that will affect the model quality. Essentially, one is trying to ensure the correctness and consistency of values and that all the data you have measures the same thing in the same way.

There are different types of data quality problems. This include, single fields having an incorrect value, incorrect combinations in individual fields (e.g. pregnant males) and missing data such as throwing out every record with a field missing, this may wind up with a very small database or an inaccurate picture of the whole database. Recognizing the fact *that you may not be able to fix all the problems, so you will need to work around them as best as possible; although, it is preferable and more cost-effective to put in place procedures and checks to avoid the data quality problems. However, you must build the models you need with the data you now have, and avoid something you will work toward for the future.*

3.4.5 Integration and Consolidation

The data you need may be residing in a single database or in multiple database and the source database may be transactional database used by the operational systems of your organization. Other data may be in data warehouses or data marts built for specific purposes.

Data integration and consolidation combines data from different sources into a single mining database and requires reconciling differences in data values from the various sources. Improperly reconciled data is a major source of quality problems. There are often large different databases (Two Crows Corporation, 2005). Though, some inconsistencies may not be easy to cover, such as different addresses for the same customer, making it more difficult to resolve. For instance, the same customers may have different names or worse multiple customers' identification numbers. Also, the same name may be used for different entities (homonyms), or different names may be used for the same entity (synonyms)

3.4.6 Metadata Construction

The information in the dataset description and data description is the basic for metadata infrastructure. In essence this is a database about the database itself. It provides information that will be used in the creation of the physical database as well as information that will be used by analysts in understanding the data and building the models. (Two Crows Corporation, 2005)

3.4.7 Load the Data Mining Database

In most cases data are stored in its database. But for large amounts or complex data this will be a DBMS as against to a flat file. After collecting, integrating and cleaning the data, it is now necessary to load the database itself. Depending on the complexity of the database design, this may turn out to be a serious task that requires the expertise of information systems professionals.

3.4.8 Maintain the Data Mining Database

Once a database is created, it needs to be taken care of, to be backed up periodically: its performance should be monitored, and may need occasional reorganization to reclaim disk storage or to improve performance for a large and complex database stored in a DBMS, the maintenance may also require the services of information systems professionals.

3.5 Explore the Data

The goal of this section is to identify the most important fields in predicting an outcome, and determine which derived values may be useful. In a data set with hundreds or even thousands of columns, exploring the data can be time-consuming and labor intensive as it is illuminating. A good interface and fast computer response becomes very important at this stage because of the nature of your exploration may change when you have to wait even 20 minutes for some graphs, let alone a day.

3.6 Prepare Data for Modelling

This is the final data preparation step before building the models. There are four major parts to this step namely:

- ❖ *Select variable*
- ❖ *Select rows*
- ❖ *Construct new variables*
- ❖ *Transform variables.*

(i) Select Variables

Idyllically, you would take all the variables you have, feed them to the data mining tool and let it find those which are the best predictions. (Two Crow Corporations, 2005). Practically, this may not work very well. Some reasons for this are that the time it takes to build a model increases with the number of variables and it blindly includes unrelated columns which can lead to incorrect models. A very common error, for example is to use as a prediction variable data that can only be known if you know the value of the response variable. People have actually used date of birth to predict age without realising it.

(i) Select Rows

Just like in the case of selecting variables, for you to use all the rows you have to build models. However, if you have a lot of data, this may take too long or require buying a bigger computer than you would like. Consequently, it is a good idea to sample the data when the database is large. This yields no loss of information for most business problems, though sample selection must be done carefully to ensure the sample is truly random. Given a choice of either investigating a few models built on all the data or investigating more models on a sample, the latter approach will usually help you develop a more accurate and robust model.

(ii) Construct New Variables

It is necessary to construct new prediction derived from the raw data, for example forecasting credit risk using a debt-to-income ratio rather than just debt and income as prediction variables may yield more accurate results that are also easier to understand. Certain variable that have little effect alone may need to be combined with others using various arithmetic or algebraic operations such as addition and ratio. Some variables that extend over a wide range may be modified to construct a better prediction, such as using the log of income instead of income.

(iii) Transform Variables

The tool chosen may dictate how to represent your data, for example the categorical explosion required by neural nets. Variables may also be scaled to fall within a limited range such as 0 to 1. Many decision trees used for classification require continuous data such as income to be grouped in range (bins) such as high, medium and low. The encoding you select can influence the result of your model. For instance, the cutoff points for the bins may change the outcome of a model.

3.7 Data Mining Model Building

Iterative process is the most important to remember about model building. There is need to explore alternative models of finding the one that is most useful in solving your business problem. What you learn in searching for a good model may lead to go back and make some changes to the data you are using or even modify your problem statement. Once a decision has been made on the type of prediction you want to make (e.g. classification or regression), you then choose a model type for making the prediction. This could be a decision tree, a neural net, a proprietary method, or that old stand, logistic repression. Your choice of model type will influence what data preparation you must do and how you go about it. The tool you want to use may require that the data be in a particular file format, thus requiring you to extract the data into that format. Once the data is ready, you can proceed with training your model.

The process of building a predictive model requires a well-defined training and validation protocol in order to insure the most accurate and robust predictions. This type of protocol is sometimes called Supervised Learning. The reason for supervised learning is to train or estimate your model on a portion of the data, then test and validate it on the remainder of the data. A model is built when the cycle of training and testing is completed. At times a third data set referred to as validation data set is needed because the test data may influence features of the model and the validation set acts an independent measure of the model's accuracy. Training and testing the data mining model requires the data to be split into at least two groups: one for model training (i.e. estimation of the model parameters) and for one model testing. If you fail to use different training and test data, the model is generated using

the training database, it is used to predict the test database, and the resulting accuracy rate is a good estimate of how the model will perform on future database that are similar to the training and test databases.

Simple Validation: simple validation is the most basic testing method. To carry out this, you set aside a percentage of the database, and do not use it in any way in the model building and estimation. The percentage is basically between 5% and 33% for all the future calculations to be correct, the division of the data into two groups must be random, so that the training and test data sets both reflect the data being modelled. In building a single model, this simple validation may need to be performed several times for instance, when using a neural net, sometimes each training pass through the net is nested against a test database.

Cross Validation: if you have only a modest amount of data (a few thousand rows) for building the model, you cannot afford to set aside a percentage of it for simple validation. Cross validation is a method that let you use all your data. The data is randomly divided into two equal sets in order to estimate the predictive accuracy of the model. The first thing is to build a model on the first set and use it to predict the outcomes in the second set and calculate an error rate. Then a model is built on the second set and use to predict the outcomes in the first set and again calculate an error rate. Finally, a model is built using all the data.

Bootstrapping: this is another technique for estimating the error of a model; it is primarily used with very small data sets. As in cross validation the model is built on the entire dataset. Then numerous data sets called bootstrap samples are created by sampling from the original data set. After each case is sampled, it is replaced and a case is selected again until the entire bootstrap sample is created. It should be noted that records may occur more than once in the data sets thus created. A model is built on the data set, and its error rate is calculated. This is called the resubstituting.

3.8 Evaluation and Interpretation

There are two stages involved in evaluation and interpretation namely:

(i) Model Validation

After building a model, the next thing is to evaluate its results and interpret their significance. And it should be remembered that the accuracy rate found during testing is only applicable to the data on which the model is built. In practice the accuracy may very if the data to which the model is applied differs in importance and unknowable ways from the original data. However, accuracy by itself is not necessarily the right metric for selecting the best model.

(ii) External Validation

As earlier described under model validation, that no matter how good the accuracy of a model is estimated to be, there is no guarantee that it reflects the real world. A valid model is not necessarily a correct model; this is because there are always implied assumptions in the model. Moreover the data used to build the model may fail to match the real world in some unknown ways leading to an incorrect model. Therefore it is important to test a model in the real world. If a model is used to select a subset of a mailing list, do a test mailing to verify the model. Also, if a model is used to predict credit risk, try the model on a small set of applicants before full deployment. The higher the risk associated with an incorrect model, the more important it is to construct an experiment to check the model results. (Two Crow corporate, 2005)

3.9 Deploy the Model and Result

Once, a data mining model is built and validated, it can be used in two major ways; the first way is for an analyst to recommend actions based on simply viewing the model and its results. For instance, the analyst may look at the clusters the model has identified and the rules that define the model. The second way is to apply the model to different data sets. The model could be used to flay records based on their classification or assign a score such as the probability of an action. The model can select some records from the database and subject these to further analyses with an OLAP tool.

Data mining model is often applied to one event or transaction at a time, such as scoring a loan application for risk. The amount of time in processing each new transaction and the rate at which new transactions arrive will determine whether a parallelized algorithm is required. Thus, while loan applications can easily be evaluated on modest-sized computers monitoring credit card transaction or cellular telephone calls for fraud would require a parallel system to deal with the high transaction rate.

Model Monitoring: There is need to measure how well your model has worked after using it, even when you think you have finished because your model is working well. You must continually monitor the performance of the model. Thus from time to time the model will have to be rested, restrained and possibly completely rebuilt

SELF- ASSESSMENT EXERCISE 2

i. What is the importance of data collection in building a data mining database?
ii. Explain the concept Data Mining Model Building
iii. Explain the two stages involved in evaluation and interpretation

4.0. CONCLUSION

Therefore, the process of mining data involves seven basic steps which is not linear but needs to be looped back to previous steps for a successful data mining.

5.0 SUMMARY

In this unit we have learnt that:

- *Data mining for knowledge discovery is made up of some basic steps, this include defining the business problem, building the data mining database, explore the data, prepare the data for modelling, build the model, evaluate the model, and deploy model and results.*

6.0 TUTOR-MARKED ASSIGNMENT

i. List the basic steps of data mining for knowledge discovery.

7.0 REFERENCES/FURTHER READING

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Introduction to Data Mining and Knowledge Discovery. Two Crows Corporation, Third Edition.

Jiawei Han., Micheline Kamber (2019). Data Mining: Concepts and Techniques, Second Edition

Yanchang Zhao, (2015). R and Data Mining Examples and Case Studies

UNIT 3 APPLICATIONS AND TRENDS IN DATA MINNG

CONTENTS

1.0 INTRODUCTION

The purpose of this unit is to give the reader some ideas of the types of activities in which data mining is already being used and what companies are using them. The applications areas that would be discussed include data mining in banking and finance, retails, telecommunications, healthcare, credit card company, transportation, surveillance, games, business, science and engineering, and spatial data,

2.0 OBJECTIVE

At the end of this unit you should be able to:

- *identify the various applications of data mining in our societies.*

3.0 MAIN CONTENT

3.1 Applications of Data Mining

Data mining is used for a variety of purposes both in private and public organizations and has been deployed successfully in a wide range of companies. While the early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing; the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships.

Two critical factors for a successful data mining are: a large well integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied such as customer prospecting, retention, campaign management and so on.

3.2 Data Mining for Financial Data Analysis

Most banks and financial institutions offer a wide variety of banking services (such as checking and savings accounts for business or individual customers), credit (such as business, mortgage, and automobile loans), and investment services (such as mutual funds). Some also offer insurance services and stock investment services. Financial data collected in the banking and financial industry are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. Here we present a few typical cases:

- ***Design and construction of data warehouses for multidimensional data analysis and data mining**: Like many other applications, data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, one may like to view the debt and revenue changes by month, by region, by sector, and by other factors, along with maximum, minimum, total, average, trend, and other statistical information. Data warehouses, data cubes, multi feature and discovery-driven data cubes, characterization and class comparisons, and outlier analysis all play important roles in financial data analysis and mining.*

- ***Loan payment prediction and customer credit policy analysis**: Loan payment prediction and customer credit analysis are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance and customer credit rat- ing. Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, fac- tors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus the total monthly income), payment- to-income ratio, customer income level, education level, residence region, and credit history. Analysis of the customer payment history may find that, say, payment-to- income ratio is a dominant factor, while education level and debt ratio are not. The bank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profiles show relatively low risks according to the critical factor analysis.*

- ***Classification and clustering of customers for targeted marketing:** Classification and clustering methods can be used for customer group identification and targeted marketing. For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking. Customers with similar behaviors regarding loan payments may be identified by multi-dimensional clustering techniques. These can help identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.*

- ***Detection of money laundering and other financial crimes**: To detect money laundering and other financial crimes, it is important to integrate information from multiple databases (like bank transaction databases, and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers. Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of customers), linkage analysis tools (to identify links among different customers and activities), classification tools (to filter unrelated attributes and rank the highly related ones), clustering tools (to group different cases), outlier analysis tools (to detect unusual amounts of fund transfers or other activities), and sequential pattern analysis tools (to characterize unusual access sequences). These tools may identify important relationships and patterns of activities and help investigators focus on suspicious cases for further detailed examination.*

3.3 Data Mining for the Retail Industry

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on the Web, or e-commerce. Today, many stores also have websites where customers can make purchases on-line. Some businesses, such as Amazon.com (www.amazon.com), exist solely on-line, without any brick-and-mortar (i.e., physical) store locations. Retail data provide a rich *source for data mining. Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.*

A few examples of data mining in the retail industry are outlined as follows.

- ***Design and construction of data warehouses based on the benefits of data mining**: Because retail data cover a wide spectrum (including sales, customers, employees, goods transportation, consumption, and services), there can be many ways to design a data warehouse for this industry. The levels of detail to include may also vary substantially. The outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures. This involves deciding which dimensions and levels to include and what preprocessing to perform in order to facilitate effective data mining.*
- ***Multidimensional analysis of sales, customers, products, time, and region**: The retail industry requires timely information regarding customer needs, product sales, trends, and fashions, as well as the quality, cost, profit, and service of commodities. It is there- fore important to provide powerful multidimensional analysis and visualization tools, including the construction of sophisticated data cubes according to the needs of data analysis. The multi feature data cube, introduced in Chapter 4, is a useful data structure in retail data analysis because it facilitates analysis on aggregates with complex conditions.*
- ***Analysis of the effectiveness of sales campaigns:** The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers. Careful analysis of the effectiveness Of sales campaigns can help improve company profits. Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of trans- actions containing the sales items during the sales period versus those containing the same items before or after the sales campaign. Moreover, association analysis may dis- close which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.*
- ***Customer retention—analysis of customer loyalty**: With customer loyalty card information, one can register sequences of purchases of particular customers. Customer loyalty and purchase trends can be analyzed systematically. Goods purchased at different periods by the same customers can be grouped into sequences. Sequential pattern mining (Chapter 8) can then be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new ones.*
- ***Product recommendation and cross-referencing of items**: By mining associations from sales records, one may discover that a customer who buys a digital camera is likely to buy another set of items. Such information can be used to form product recommendations. Collaborative recommender systems use data mining techniques to make personalized product*

recommendations during live customer transactions, based on the opinions of other customers. Product recommendations can also be advertised on sales receipts, in weekly flyers, or on the Web to help improve customer service, aid customers in selecting items, and increase sales. Similarly, information such as "hot items this week" or attractive deals can be displayed together

3.4 Data Mining Applications in Telecommunications

The telecommunication industry has quickly evolved from offering local and long distance telephone services to providing many other comprehensive communication services, including fax, pager, cellular phone, Internet messenger, images, e-mail, computer and Web data transmission, and other data traffic. The integration of telecommunication, computer network, Internet, and numerous other means of communication and computing is also underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service. The following are a few scenarios for which data mining may improve telecommunication services:

- ***Multidimensional analysis of telecommunication data**: Telecommunication data are intrinsically multidimensional, with dimensions such as calling-time, duration, location of caller, location of callee, and type of call. The multidimensional analysis of such data can be used to identify and compare the data traffic, system workload, resource usage, user group behavior, and profit. For example, analysts in the industry may wish to regularly view charts and graphs regarding calling source, destination, volume, and time-of-day usage patterns. Therefore, it is often useful to consolidate telecommunication data into large data warehouses and routinely perform multidimensional analysis using OLAP and visualization tools.*
- ***Fraudulent pattern analysis and the identification of unusual patterns**: Fraudulent activity costs the telecommunication industry millions of dollars per year. It is important to (1) identify potentially fraudulent users and their atypical usage patterns; (2) detect attempts to gain fraudulent entry to customer accounts; and (3) discover unusual patterns that may need special attention, such as busy-hour frustrated call attempts, switch and route congestion patterns, and periodic calls from automatic dial-out equipment (like fax machines) that have been improperly programmed. Many of these patterns can be discovered by multidimensional analysis, cluster analysis, and outlier analysis.*
- ***Multidimensional association and sequential pattern analysis:** The discovery of association and sequential patterns in multidimensional analysis can be used to promote telecommunication services. For example, suppose you would like to find usage pat- terns for a set of communication services by customer group, by month, and by time of day. The calling records may be grouped by customer in the following form:*
  (customer ID, residence, office, time, date, service 1, service 2, · · ·)

- ***Mobile telecommunication services:*** *Mobile telecommunication, Web and information services, and mobile computing are becoming increasingly integrated and common in our work and life. One important feature of mobile telecommunication data is its association with spatiotemporal information. Spatiotemporal data mining may become essential for finding certain patterns. For example, unusually busy mobile phone traffic at certain locations may indicate something abnormal happening in these locations. Moreover, ease of use is crucial for enticing customers to adopt new mobile services. Data mining will likely play a major role in the design of adaptive solutions enabling users to obtain useful information with relatively few keystrokes.*
- ***Use of visualization tools in telecommunication data analysis:*** *Tools for OLAP visualization, linkage visualization, association visualization, clustering, and outlier visualization have been shown to be very useful for telecommunication data analysis.*

SELF- ASSESSMENT EXERCISE 1

The applications of data mining in telecommunication industries can be grouped into three areas namely: Fraud detection, marketing/customer profiling and network fault isolation. Briefly discuss these areas.

3.5 Data Mining for Biological Data Analysis

The past decade has seen an explosive growth in genomics, proteomics, functional genomics, and biomedical research. Examples range from the identification and comparative analysis of the genomes of human and other species (by discovering sequencing patterns, gene functions, and evolution paths) to the investigation of genetic networks and protein pathways, and the development of new pharmaceuticals and advances in cancer therapies. Biological data mining has become an essential part of a new research field called bioinformatics. Since the field of biological data mining is broad, rich, and dynamic, it is impossible to cover such an important and flourishing theme in one sub- section. Here we outline only a few interesting topics in this field, with an emphasis on genomic and proteomic data analysis.

DNA sequences form the foundation of the genetic codes of all living organisms. All DNA sequences are comprised off our basic building blocks, called nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). These four nucleotides (or bases) are combined to form long sequences or chains that resemble a twisted ladder. The DNA carry the information and biochemical machinery that can be copied from generation to generation. During the processes of "copying," insertions, deletions, or mutations (also called substitutions) of nucleotides are introduced into the DNA sequence, forming different evolution paths. A gene usually comprises hundreds of individual nucleotides arranged in a particular order. The nucleotides can be ordered and sequenced in an almost unlimited number of ways to form distinct genes. A genome is the complete set of genes of an organism. The human genome is estimated to contain around 20,000 to 25,000 genes. Genomics is the analysis of genome sequences.

The identification of DNA or amino acid sequence patterns that play roles in various biological functions, genetic diseases and evolution is challenging. This requires a great deal of research in computational algorithms, statistics, mathematical programming, data mining, machine learning, information retrieval, and other disciplines to develop effective genomic and proteomic data analysis tools. Data mining may contribute to biological data analysis in the following aspects:

- ***Semantic integration of heterogeneous, distributed genomic and proteomic databases:** Genomic and proteomic data sets are often generated at different labs and by different methods. They are distributed, heterogenous, and of a wide variety. The semantic integration of such data is essential to the cross-site analysis of biological data. Moreover, it is important to find correct linkages between research literature and their associated biological entities. Such integration and linkage analysis would facilitate the systematic and coordinated analysis of genome and biological data. This has promoted the development of integrated data warehouses and distributed federated databases to store and manage the primary and derived biological data. Data cleaning, data integration, reference reconciliation, classification, and clustering methods will facilitate the integration of biological data and the construction of data warehouses for biological data analysis.*

- ***Alignment, indexing, similarity search, and comparative analysis of multiple nucleo- tide/protein sequences**: Various biological sequence alignment methods have been developed in the past two decades. BLAST and FASTA, in particular, are tools for the systematic analysis of genomic and proteomic data. Biological sequence analysis methods differ from many sequential pattern analysis algorithms proposed in data mining research. They should allow for gaps and mismatches between a query sequence and the sequence data to be searched in order to deal with insertions, deletions, and mutations. Moreover, for protein sequences, two amino acids should also be considered a "match" if one can be derived from the other by substitutions that are likely to occur in nature. Sophisticated statistical analysis and dynamic programming methods often play a key role in the development of alignment algorithms. Indices can be constructed on such data sets so that precise and similarity searches can be performed efficiently*

- ***Discovery of structural patterns and analysis of genetic networks and protein path- ways:** In biology, protein sequences are folded into three-dimensional structures, and such structures interact with each other based on their relative positions and the distances between them. Such complex interactions form the basis of sophisticated genetic networks and protein pathways. It is crucial to discover structural patterns and regularities among such huge but complex biological networks. To this extent, it is important to develop powerful and scalable data mining methods to discover approximate and frequent structural patterns and to study the regularities and irregularities among such interconnected biological networks.*
- ***Association and path analysis: identifying co-occurring gene sequences and linking genes to different stages of disease development:** Currently, many studies have focused on the comparison of one gene to another. However, most diseases are not triggered by a single gene but by a combination of genes acting together. Association analysis methods can be used to help determine the kinds of genes that are likely to co-occur in target samples. Such analysis would facilitate the discovery of groups of genes and the study of interactions and relationships*

between them. While a group of genes may contribute to a disease process, different genes may become active at different stages of the disease. If the sequence of genetic activities across the different stages of disease development can be identified, it may be possible to develop pharmaceutical interventions that target the different stages separately, therefore achieving more effective treatment of the disease. Such path analysis is expected

- *Visualization tools in genetic data analysis: Alignments among genomic or proteomic sequences and the interactions among complex biological structures are most effectively presented in graphic forms, transformed into various kinds of easy-to-understand visual displays. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration. Visualization and visual data mining therefore play an important role in biological data analysis.*

3.6 Data Mining in Other Scientific Applications

Previously, most scientific data analysis tasks tended to handle relatively small and homogeneous data sets. Such data were typically analyzed using a "formulate hypothesis, build model, and evaluate results" paradigm. In these cases, statistical techniques were appropriate and typically employed for their analysis (see Section 11.3.2). Data collection and storage technologies have recently improved, so that today, scientific data can be amassed at much higher speeds and lower costs. This has resulted in the accumulation of huge volumes of high-dimensional data, stream data, and heterogenous data, containing rich spatial and temporal information. Consequently, scientific applications are shifting from the "hypothesize-and-test" paradigm toward a "collect and store data, mine for new hypotheses, confirm with data or experimentation" process. This shift brings about new challenges for data mining. Vast amounts of data have been collected from scientific domains (including geo-sciences, astronomy, and meteorology) using sophisticated telescopes, multispectral high-resolution remote satellite sensors, and global positioning systems. Large data sets are being generated due to fast numerical simulations in various fields, such as cli- mate and ecosystem modeling, chemical engineering, fluid dynamics, and structural mechanics. we look at some of the challenges brought about by emerging scientific applications of data mining, such as the following:

- ***Data warehouses and data preprocessing:*** *Data warehouses are critical for information exchange and data mining. In the area of geospatial data, however, no true geospatial data warehouse exists today. Creating such a warehouse requires finding means for resolving geographic and temporal data incompatibilities, such as reconciling semantics, referencing systems, geometry, accuracy, and precision. For scientific applications in general, methods are needed for integrating data from heterogeneous sources (such as data covering different time periods) and for identifying events. For climate and ecosystem data, for instance (which are spatial and temporal), the problem is that there are too many events in the spatial domain and too few in the temporal domain. (For example, El Ni˜no events occur only every four to seven years, and previous data might not have been collected as systematically as today.) Methods are needed for the efficient computation of sophisticated spatial aggregates and the handling of spatial-related data streams.*

- ***Mining complex data types***: *Scientific data sets are heterogeneous in nature, typically involving semi-structured and unstructured data, such as multimedia data and georeferenced stream data. Robust methods are needed for handling spatiotemporal data, related concept hierarchies, and complex geographic relationships (e.g., non- Euclidian distances).*

- ***Graph-based mining***: *It is often difficult or impossible to model several physical phenomena and processes due to limitations of existing modeling approaches. Alter- natively, labeled graphs may be used to capture many of the spatial, topological, geometric, and other relational characteristics present in scientific data sets. In graph- modeling, each object to be mined is represented by a vertex in a graph, and edges between vertices represent relationships between objects. For example, graphs can be used to model chemical structures and data generated by numerical simulations, such as fluid-flow simulations. The success of graph-modeling, however, depends on improvements in the scalability and efficiency of many classical data mining tasks, such as classification, frequent pattern mining, and clustering.*

- ***Visualization tools and domain-specific knowledge***: *High-level graphical user inter- faces and visualization tools are required for scientific data mining systems. These should be integrated with existing domain-specific information systems and database systems to guide researchers and general users in searching for patterns, interpreting and visualizing discovered patterns, and using discovered knowledge in their decision making.*

3.7 Data Mining for Intrusion Detection

The security of our computer systems and data is at continual risk. The extensive growth of the Internet and increasing availability of tools and tricks for intruding and attacking networks have prompted intrusion detection to become a critical component of net- work administration. An intrusion can be defined as any set of actions that threaten the integrity, confidentiality, or availability ofa network resource (such as user accounts, file systems, system kernels, and so on).

Most commercial intrusion detection systems are limiting and do not provide a complete solution. Such systems typically employ a misuse detection strategy. Misuse detection searches for patterns of program or user behavior that match known intrusion scenarios, which are stored as signatures. These hand-coded signatures are laboriously provided by human experts based on their extensive knowledge of intrusion techniques. If a pattern match is found, this signals an event for which an alarm is raised. Human security analysts evaluate the alarms to decide what action to take, whether it be shutting down part of the system, alerting the relevant Internet service provider of suspicious traffic, or simply noting unusual traffic for future reference. An intrusion detection system for a large complex network can typically generate thousands or millions of alarms per day, representing an overwhelming task for the security analysts. Because systems are not static, the signatures need to be updated whenever new software versions arrive or changes in network configuration occur. An additional, major drawback is that misuse detection can only identify cases *that match the signatures. That is, it is unable to detect new or previously unknown intrusion techniques.*

Novel intrusions may be found by anomaly detection strategies. Anomaly detection builds models of normal network behavior (called profiles), which it uses to detect new patterns that significantly deviate from the profiles. Such deviations may represent actual intrusions or simply be new behaviors that need to be added to the profiles. The main advantage of anomaly detection is that it may detect novel intrusions that have not yet been observed. Typically, a human analyst must sort through the deviations to ascertain which represent real intrusions. A limiting factor of anomaly detection is the high per- centage of false positives. New patterns of intrusion can be added to the set of signatures for misuse detection.

As we can see from this discussion, current traditional intrusion detection systems face many limitations. This has led to an increased interest in data mining for intrusion detection. The following are areas in which data mining technology may be applied or further developed for intrusion detection:

- ***Development of data mining algorithms for intrusion detection:*** *Data mining algorithms can be used for misuse detection and anomaly detection. In misuse detection, training data are labeled as either "normal" or "intrusion." A classifier can then be derived to detect known intrusions. Research in this area has included the application of classification algorithms, association rule mining, and cost-sensitive modeling. Anomaly detection builds models of normal behavior and automatically detects significant deviations from it. Supervised or unsupervised learning can be used. In a supervised approach, the model is developed based on training data that are known to be "normal." In an unsupervised approach, no information is given about the training data. Anomaly detection research has included the application of classification algorithms, statistical approaches, clustering, and outlier analysis. The techniques used must be efficient and scalable, and capable of handling network data of high volume, dimensionality, and heterogeneity.*

- ***Association and correlation analysis, and aggregation to help select and build dis- criminating attributes:*** *Association and correlation mining can be applied to find relationships between system attributes describing the network data. Such information can provide insight regarding the selection of useful attributes for intrusion detection. New attributes derived from aggregated data may also be helpful, such as summary counts of traffic matching a particular pattern.*

- ***Analysis of stream data****: Due to the transient and dynamic nature of intrusions and malicious attacks, it is crucial to perform intrusion detection in the data stream environment. Moreover, an event may be normal on its own, but considered malicious if viewed as part of a sequence of events. Thus it is necessary to study what sequences of events are frequently encountered together, find sequential patterns, and identify outliers. Other data mining methods for finding evolving clusters and building dynamic classification models in data streams are also necessary for real-time intrusion detection.*

***Distributed data mining**: Intrusions can be launched from several different locations and targeted to many different destinations. Distributed data mining methods may be used to analyze network data from several network locations in order to detect these distributed attacks.*

- ***Visualization and querying tools:** Visualization tools should be available for viewing any anomalous patterns detected. Such tools may include features for viewing associations, clusters, and outliers. Intrusion detection systems should also have a graphical user interface that allows security analysts to pose queries regarding the network data or intrusion detection results.*
  In comparison to traditional intrusion detection systems, intrusion detection systems based on data mining are generally more precise and require far less manual processing and input from human experts.

3.8 Data Mining System Products and Research Prototypes

Although data mining is a relatively young field with many issues that still need to be researched in depth, many off-the-shelf data mining system products and domain-specific data mining application software's are available. As a discipline, data mining has a relatively short history and is constantly evolving—new data mining systems appear on the market every year; new functions, features, and visualization tools are added to existing systems on a constant basis; and efforts toward the standardization of data min-ing language are still underway. Therefore, it is not our intention in this book to provide a detailed description of commercial data mining systems. Instead, we describe the fea- tures to consider when selecting a data mining product and offer a quick introduction to a few typical data mining systems. Reference articles, websites, and recent surveys of data mining systems are listed in the bibliographic notes.

3.9 How to Choose a Data Mining System

With many data mining system products available on the market, you may ask, "What kind of system should I choose?" Some people may be under the impression that data mining systems, like many commercial relational database systems, share the same well- defined operations and a standard query language, and behave similarly on common functionalities. If such were the case, the choice would depend more on the systems' hardware platform, compatibility, robustness, scalability, price, and service. Unfortunately, this is far from reality. Many commercial data mining systems have little in com- mon with respect to data mining functionality or methodology and may even work with completely different kinds of data sets. To choose a data mining system that is appropriate for your task, it is important to
have a multidimensional view of data mining systems. In general, data mining systems should be assessed based on the following multiple features:

- ***Data types:** Most data mining systems that are available on the market handle formatted, record-based, relational-like data with numerical, categorical, and symbolic attributes. The data could be in the form of ASCII text, relational database data, or data warehouse data. It is important to check what exact format(s) each* *system you are considering can handle. Some kinds of data or applications may require specialized algorithms to search for patterns, and so their requirements may not be handled by off-the-shelf, generic data mining systems. Instead, specialized data mining systems may be used, which mine either text documents, geospatial data, multimedia data, stream data, time-series data, biological data, or Web data, or are dedicated to specific applications (such as finance, the retail industry, or telecommunications). Moreover, many data mining companies offer customized data mining solutions that incorporate essential data mining functions or methodologies.*

- ***System issues:*** *A given data mining system may run on only one operating system or on several. The most popular operating systems that host data mining software are UNIX/Linux and Microsoft Windows. There are also data mining systems that run on Macintosh, OS/2, and others. Large industry-oriented data mining systems often adopt a client/server architecture, where the client could be a personal computer, and the server could be a set of powerful parallel computers. A recent trend has data mining systems providing Web-based interfaces and allowing XML*

- ***data as input and/or output. Data sources:*** *This refers to the specific data formats on which the data mining system will operate. Some systems work only on ASCII text files, whereas many others work on relational data, or data warehouse data, accessing multiple relational data sources. It is important that a data mining system supports ODBC connections or OLE DB for ODBC connections. These ensure open database connections, that is, the ability to access any relational data (including those in IBM/DB2, Microsoft SQL Server, Microsoft Access, Oracle, Sybase, etc.), as well as formatted ASCII text data*

- ***Data mining functions and methodologies:*** *Data mining functions form the core of a data mining system. Some data mining systems provide only one data mining function, such as classification. Others may support multiple data mining functions, such as concept description, discovery-driven OLAP analysis, association mining, linkage analysis, statistical analysis, classification, prediction, clustering, outlier anal- ysis, similarity search, sequential pattern analysis, and visual data mining. For a given data mining function (such as classification), some systems may support only one method, whereas others may support a wide variety of methods (such as decision tree analysis, Bayesian networks, neural networks, support vector machines, rule- based classification, k-nearest-neighbor methods, genetic algorithms, and case-based reasoning). Data mining systems that support multiple data mining functions and multiple methods per function provide the user with greater flexibility and analysis power. Many problems may require users to try a few different mining functions or incorporate several together, and different methods can be more effective than others for different kinds of data. In order to take advantage of the added flexibility, how- ever, users may require further training and experience. Thus such systems should also provide novice users with convenient access to the most popular function and method, or to default settings.*

- ***Coupling data mining with database and/or data warehouse systems****: A data mining system should be coupled with a database and/or data warehouse system, where the coupled components are seamlessly integrated into a uniform information processing environment. In general, there are four forms of such* *coupling: no coupling, loose coupling, semi tight coupling, and tight coupling. Some data mining systems work only with ASCII data files and are not coupled with database or data warehouse systems at all. Such systems have difficulties using the data stored in database systems and handling large data sets efficiently. In data mining systems that are loosely coupled with database and data warehouse systems, the data are retrieved into a buffer or main memory by database or warehouse operations, and then mining functions are applied to analyze the retrieved data. These systems may not be equipped with scalable algorithms to handle large data sets when processing data mining queries. The coupling ofa data mining system with a database or data warehouse system may be semi tight, providing the efficient implementation of a few essential data mining primitives (such as sorting, indexing, aggregation, histogram analysis, multiway join, and the precomputation of some statistical measures). Ideally, a data mining system should be tightly coupled with a database system in the sense that the data mining and data retrieval processes are integrated by optimizing data mining queries deep into the iterative mining and retrieval process. Tight coupling of data mining with OLAP-based data warehouse systems is also desirable so that data mining and OLAP operations can be integrated to provide OLAP-mining features*

- ***Scalability:*** *Data mining has two kinds of scalability issues: row (or database size) scalability and column (or dimension) scalability. A data mining system is considered row scalable if, when the number of rows is enlarged 10 times, it takes no more than 10 times to execute the same data mining queries. A data mining system is considered column scalable if the mining query execution time increases linearly with the number of columns (or attributes or dimensions). Due to the curse of dimensionality, it is much more challenging to make a system column scalable than row scalable.*

- ***Visualization tools:*** *"A picture is worth a thousand words"—this is very true in data mining. Visualization in data mining can be categorized into data visualization, mining result visualization, mining process visualization, and visual data mining. The variety, quality, and flexibility of visualization tools may strongly influence the usability, interpretability, and attractiveness of a data mining system. Data*

- ***Data mining query language and graphical user interface:*** *Data mining is an exploratory process. An easy-to-use and high-quality graphical user interface is essential in order to promote user-guided, highly interactive data mining. Most data mining systems provide user-friendly interfaces for mining. However, unlike relational database systems, where most graphical user interfaces are constructed on top of SQL (which serves as a standard, well-designed database query language), most data mining systems do not share any underlying data mining query language. Lack of a standard data mining language makes it difficult to standardize data mining products and to ensure the interoperability of data mining systems.*

3.10 Theoretical Foundations of Data Mining Research

Research on the theoretical foundations of data mining has yet to mature. A solid and systematic theoretical foundation is important because it can help provide a coherent framework for the development, evaluation, and practice of data mining technology .Several theories for the basis of data mining include the following:

- ***Data reduction**: In this theory, the basis of data mining is to reduce the data representation. Data reduction trades accuracy for speed in response to the need to obtain quick approximate answers to queries on very large databases. Data reduction techniques include singular value decomposition (the driving element behind principal components analysis), wavelets, regression, log-linear models, histograms, clustering, sampling, and the construction of index trees.*

- ***Data compression:** According to this theory, the basis of data mining is to compress the given data by encoding in terms of bits, association rules, decision trees, clusters, and so on. Encoding based on the minimum description length principle states that the "best" theory to infer from a set of data is the one that minimizes the length of the theory and the length of the data when encoded, using the theory as a predictor for the data. This encoding is typically in bits.*

- ***Pattern discovery**: In this theory, the basis of data mining is to discover patterns occurring in the database, such as associations, classification models, sequential pat- terns, and so on. Areas such as machine learning, neural network, association mining, sequential pattern mining, clustering, and several other subfields contribute to this theory. Probability*

- ***Probability theory**: This is based on statistical theory. In this theory, the basis of data mining is to discover joint probability distributions of random variables, for example, Bayesian belief networks or hierarchical Bayesian models.*

- ***Microeconomic view**: The microeconomic view considers data mining as the task of finding patterns that are interesting only to the extent that they can be used in the decision-making process of some enterprise (e.g., regarding marketing strategies and production plans). This view is one of utility, in which patterns are considered interesting if they can be acted on. Enterprises are regarded as facing optimization problems, where the object is to maximize the utility or value of a decision. In this theory, data mining becomes a nonlinear optimization problem*

- ***Inductive databases**: According to this theory, a database schema consists of data and patterns that are stored in the database. Data mining is therefore the problem of performing induction on databases, where the task is to query the data and the theory (i.e., patterns) of the database. This view is popular among many researchers in database systems.*

These theories are not mutually exclusive. For example, pattern discovery can also be seen as a form of data reduction or data compression. Ideally, a theoretical framework should be able to model typical data mining tasks (such as association, classification, and clustering), have a probabilistic nature, be able to handle different forms of data, and consider the iterative and interactive essence of data mining. Further efforts are required toward the *establishment of a well-defined framework for data mining, which satisfies these requirements.*

4.0 SELF- ASSESSMENT EXERCISE

1. Briefly discuss the roles of data mining in the following application areas:

a. Intrusion detection
b. Science and engineering
c. Business
d. Telecommunication.
e. Financial data
f. Retail business

2. Describe theoretical foundation of data mining system

TUTOR- MARKED ASSIGNMENT

i. List and briefly explain any five applications of data mining in our societies.
ii. Describe the criteria for choosing a data mining system

5.0 CONCLUSION

Data mining has become increasingly common in both the private and public sectors. Industries such as banking and finance, retail, healthcare, telecommunication commonly use data mining to reduce costs, enhance research and increase sales. In the public sector, data mining applications initially were used as a means of detecting fraud waste, but have grown to also be used for purposes such as measuring and improving programme performance.

6.0 SUMMARY

In this unit we have learnt that:

- *Many customized data mining tools have been developed for domain-specific applications, including finance, the retail industry, telecommunications, bioinformatics, intrusion detection, and other science, engineering, and government data analysis. Such practice integrates domain-specific knowledge with data analysis techniques and provides mission-specific data mining solutions.*
- *There are many data mining systems and research prototypes to choose from. When selecting a data mining product that is appropriate for one's task, it is important to consider various features of data mining systems from a multidimensional point of view. These include data types, system issues, data sources, data mining functions and methodologies, tight coupling of the data mining system with a database or data warehouse system, scalability, visualization tools, and data mining query language and graphical user interfaces.*
- *Researchers have been striving to build theoretical foundations for data mining. Several interesting proposals have appeared, based on data reduction, data com-*

pression, pattern discovery, probability theory, microeconomic theory, and inductive database.

7.0 REFERENCES/FURTHER READING

Connolly, A., VanderPlas, J., & Gray, A. (2014). Classification. In *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (pp. 365-402). PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctt4cgbdj.13

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

Better Health Care With Data Mining, Philip Baylis (Co-Author), Shared Medical Systems Limited, UK.

Data Management and Data Warehouse Domain Technical Architecture, June 6, 2002.

Data Mining Workloads: Current Approaches and Future Challenges in System Architecture.

Hans-P, K. et al. (2007). Future Trends in Data Mining. Ludwig- Maximilians-Universitat.

Mosud, Y. O. (2009). Introduction to Data Mining and Data Warehousing. Lagos: Rashmoye Publications.

MODULE 3 DATA WAREHOUSE CONCEPTS

UNIT 1 OVERVIEW OF DATA WAREHOUSE

CONTENTS

1.0. INTRODUCTION

Data warehouses usually contain historical data derived from transaction data, but it can include data from other sources. Also, it separates analysis work load from transaction workload and enables an organization to consolidate data from several sources.

This unit examines the meaning of data warehouse, its goals and characteristics, evolution, advantages and disadvantages, its components, applications and users.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

3.1 *define the term data warehouse*
3.2 *state the goals and characteristics of data warehouse*
3.3 *list the major components of data warehouse*
3.4 *state the structure and approaches to storing data in data warehouse*
3.5 *describe the users and application areas of data warehouse.*

3.0 MAIN CONTENT

3.1 Definition of Data Warehouse

A ***Data Warehousing*** *(DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.*

it is a blend of technologies and components which aids the strategic use of data. It is also a electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

Other definitions of data warehouse include:

1. *A data warehouse is a data structure that is optimized for distribution. It collects and stores integrated sets of historical data from multiple operational systems and feeds them to one or more data marts.*

2. *A data warehouse is that portion of an overall is architected data environment that serves as the single integrated source of data for processing information.*

3. *Data warehouse is a repository of an organization's electronically stored data designed to facilitate reporting and analysis.*

Data warehouse system is also known by the following name:

- *Decision Support System (DSS)*
- *Executive Information System*
- *Management Information System*
- *Business Intelligence Solution*
- *Analytic Application*
- *Data Warehouse*

Decision Support System
Executive Information System
Data Warehouse
Other Names for Datawarehouse
Management Information System
Analytic Application
Business Intelligence Solution

3.2 How Datawarehouse works?

A Data Warehouse works as a central repository where information arrives from one or more data sources. Data flows into a data warehouse from the transactional system and other relational databases.

Data may be:

1. *Structured*
2. *Semi-structured*
3. *Unstructured data*

The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. A data warehouse merges information coming from different sources into one comprehensive database.

By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available. Data warehousing makes data mining possible. Data mining is looking for patterns in the data that may lead to higher sales and profits.

3.3 Types of Data Warehouse

- ***Enterprise Data Warehouse:***

Enterprise Data Warehouse is a centralized warehouse. It provides decision support service across the enterprise. It offers a unified approach for organizing and representing data. It also provide the ability to classify data according to the subject and give access according to those divisions

- ***Operational Data Store:***

Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.

- ***Data Mart:***

A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

3.4 General stages of Data Warehouse

Earlier, organizations started relatively simple use of data warehousing. However, over time, more sophisticated use of data warehousing begun.

The following are general stages of use of the data warehouse:

***Offline Operational Database*:**

In this stage, data is just copied from an operational system to another server. In this way, loading, processing, and reporting of the copied data do not impact the operational system's performance.

Offline Data Warehouse:

Data in the Datawarehouse is regularly updated from the Operational Database. The data in Datawarehouse is mapped and transformed to meet the Datawarehouse objectives

***Real time Data Warehouse*:**

In this stage, Data warehouses are updated whenever any transaction takes place in operational database. For example, Airline or railway booking system.

Integrated Data Warehouse:

In this stage, Data Warehouses are updated continuously when the operational system performs a transaction. The Datawarehouse then generates transactions which are passed back to the operational system.

3.5 Components of Data warehouse

Load manager: *Load manager is also called the front component. It performs with all the operations associated with the extraction and load of data into the warehouse. These operations include transformations to prepare the data for entering into the Data warehouse.*

Warehouse Manager: *Warehouse manager performs operations associated with the management of the data in the warehouse. It performs operations like analysis of data to ensure consistency, creation of indexes and views, generation of denormalization and aggregations, transformation and merging of source data and archiving and baking-up data.*

Query Manager: *Query manager is also known as backend component. It performs all the operation operations related to the management of user queries. The operations of this Data warehouse components are direct queries to the appropriate tables for scheduling the execution of queries.*

3.5.1 Steps to Implement Data Warehouse

The best way to address the business risk associated with a Datawarehouse implementation is to employ a three-prong strategy as below

1. ***Enterprise strategy****: Here we identify technical including current architecture and tools. We also identify facts, dimensions, and attributes. Data mapping and transformation is also passed.*
2. ***Phased delivery****: Datawarehouse implementation should be phased based on subject areas. Related business entities like booking and billing should be first implemented and then integrated with each other.*
3. ***Iterative Prototyping****: Rather than a big bang approach to implementation, the Datawarehouse should be developed and tested iteratively.*

Here, are key steps in Datawarehouse implementation along with its deliverables

Step	***Tasks***	***Deliverables***
1	*Need to define project scope*	*Scope Definition*
2	*Need to determine business needs*	*Logical Data Model*
3	*Define Operational Datastore requirements*	*Operational Data Store Model*
4	*Acquire or develop Extraction tools*	*Extract tools and Software*
5	*Define Data Warehouse Data requirements*	*Transition Data Model*
6	*Document missing data*	*To Do Project List*
7	*Maps Operational Data Store to Data Warehouse*	*D/W Data Integration Map*
8	*Develop Data Warehouse Database design*	*D/W Database Design*
9	*Extract Data from Operational Data Store*	*Integrated D/W Data Extracts*
10	*Load Data Warehouse*	*Initial Data Load*
11	*Maintain Data Warehouse*	*On-going Data Access and Subsequent Loads*

3.5.2 Best practices to implement a Data Warehouse

- *Decide a plan to test the consistency, accuracy, and integrity of the data.*
- *The data warehouse must be well integrated, well defined and time stamped.*
- *While designing Datawarehouse make sure you use right tool, stick to life cycle, take care about data conflicts and ready to learn you're your mistakes.*
- *Never replace operational systems and reports*
- *Don't spend too much time on extracting, cleaning and loading data.*
- *Ensure to involve all stakeholders including business personnel in Datawarehouse implementation process. Establish that Data warehousing is a joint/ team project. You don't want to create Data warehouse that is not useful to the end users.*
- *Prepare a training plan for the end users.*

3.5.3 Advantages of Data Warehouse:

- *Data warehouse allows business users to quickly access critical data from some sources all in one place.*
- *Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.*
- *Data Warehouse helps to integrate many sources of data to reduce stress on the production system.*
- *Data warehouse helps to reduce total turnaround time for analysis and reporting.*

- *Restructuring and Integration make it easier for the user to use for reporting and analysis.*
- *Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.*
- *Data warehouse stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.*

3.5.4 Disadvantages of Data Warehouse:

- *Not an ideal option for unstructured data.*
- *Creation and Implementation of Data Warehouse is surely time confusing affair.*
- *Data Warehouse can be outdated relatively quickly*
- *Difficult to make changes in data types and ranges, data source schema, indexes, and queries.*
- *The data warehouse may seem easy, but actually, it is too complex for the average users.*
- *Despite best efforts at project management, data warehousing project scope will always increase.*
- *Sometime warehouse users will develop different business rules.*
- *Organizations need to spend lots of their resources for training and Implementation purpose.*

3.6 Goals of Data Warehouse

The major goals of data warehousing are stated as follows:

i. *To facilitate reporting as well as analysis*
ii. *Maintain an organisations historical information*
iii. *Be an adaptive and resilient source of information*
iv. *Be the foundation for decision making*.

3.7 Characteristics of Data Warehouse

The characteristics of a data warehouse as set forth by William Inmon are stated as follows:

❖ *Subject–oriented*
❖ *Integrated*
❖ *Nonvolatile*
❖ *Time variant*

i. Subject-Oriented

The main objective of storing data is to facilitate decision process of a company, and within any company data naturally concentrates around subject areas. This leads to the gathering of information around these subjects rather than around the applications or processes (Muhammad, A.S.)

ii. **Integrated**

The data in the data warehouses are scattered around different tables, databases or even servers. Data warehouses must put data from different sources into a consistent format.

They must resolve such problems as naming conflicts and inconsistencies among units of measure. When this is achieved, they are said to be integrated.

iii. Non-Volatile

Non-volatile means that information in the data warehouse does not change each time an operational process is executed. Information is consistent regardless of when and how the warehouse is accessed.

iv. Time-Variant

The value of operational data changes on the basis of time. The time based archival of data from operational systems to data warehouse makes the value of data in the data warehouses to be a function of time. As data warehouse gives accurate picture of operational data for some given time and the changes in the data in warehouse are based on time- based change in operational data, data in the data warehouse is called 'time-variant'.

Other characteristics outside the definition of William Inmon are:

- *Accessibility: the primary purpose of a data warehouse is to provide readily accessible information to end-user.*
- *Process-Oriented: data warehousing can be viewed as the process of delivering information; and the maintenance of a data warehouse is continuous and iterative in nature.*

SELF -ASSESSMENT EXERCISE 1

What are the advantages and disadvantages of implementing a data warehouse? (Hint: state 3 points)

3.8. Data Warehouse Components

The major components of a data warehouse are:

1. *Summarized data*
2. *Operational systems of record*
3. *Integration/Transformation programs*
4. *Current detail*
5. *Data warehouse architecture or metadata*
6. *Archives.*

3.8.1 Approaches for Storing Data in a Warehouse

There are two leading approaches to storing data in a data warehouse. These are:

- *The dimensional approach*
- *The normalized approach*

Dimensional Approach: in dimensional approach, transaction data are partitioned into either "facts", which are generally numeric transaction data or "dimensions" which are the reference information which gives context to the facts. For example, a sales transaction *can be broken up into facts such as order date, customer's name, product number, order ship-to and bill-to locations and salesperson responsible for receiving the order.*

3.8.2 Benefits of Dimensional Approach

1. *This approach makes the data warehouse easier for the user to understand and to use.*
2. *The retrieval of data from the data warehouse tends to operate very quickly.*

3.8.3 Disadvantages of Dimensional Approach

1. *In order to maintain the integrity of facts and dimensions, loading the data warehouse with data from different operational systems is complicated.*

2. *It is difficult to modify the data warehouse structure if the organization adopting the dimensional approach changes the way in which it does business.*

The Normalized Approach: In this approach, the data in the data warehouse are stored following to a degree and database normalization rules. Tables are grouped together by subject areas that reflect general data categories e.g. data on customers, products, finances.

3.8.4 Benefits of Normalized Approach

The major benefits derived from this approach is that it is straight forward to add information into the database

3.8.5. Disadvantages of Normalized Approach

Because of the number of tables involved, it can be difficult for users to both join data from different sources into meaningful information and then access the information without a precise understanding of the sources of data and of the data structure of the data warehouse.

3.9 Data Warehouse Users

The successful implementation of a data warehouse is measured solely by its acceptance by users. Without users, historical data might as well be achieved by magnetic tape and stored in the basement. Successful data warehouse design starts with understanding the users and their needs.

Data warehouse users can be divided into four categories:

1. *Statisticians*
2. *Knowledge workers*
3. *Information consumers*
4. *Executives*

Each makes up a portion of the user population as illustrated in this diagram

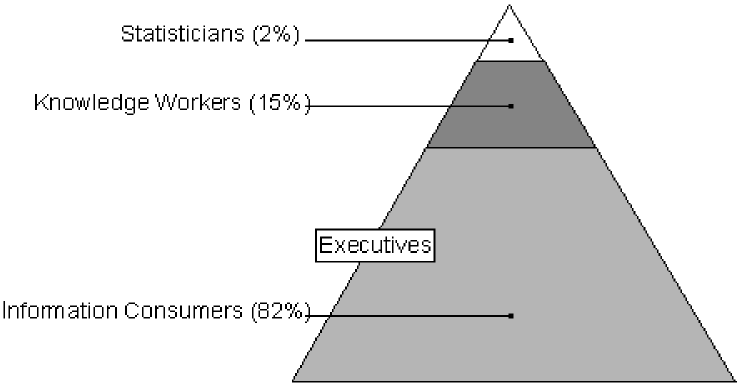


Fig.1.4: The User Pyramid
Source: Dave Browning & Joy Mundy, Dec. 2001

1. *Statisticians*

There are usually a handful of sophisticated analysts comprising of statisticians and operations research types in any organization. Though they are few in number but are best users of the data warehouse, those whose work can contribute to closed loop systems that deeply influence the operations and profitability of the company. It is vital that these users come to love the data warehouse. Generally, that is not difficult: these people are often very self-sufficient and need only to be pointed to the database and given some simple instruction about how to get to the data and what times of the day are best for performing large queries to retrieve data to analyze using their own sophisticated tools.

2. *Knowledge Workers*

A relatively small number of analysts perform the bulk of new queries and analysis against the data warehouse. These are the users who get the "designer" or "analyst" versions of user access tools. They figure out how to quantify a subject area. After a few iterations, their queries and reports typically get published for the benefit of the information consumers. Knowledge workers are often deeply engaged with the data warehouse design and place the greatest demands on the ongoing data warehouse operations team from training and support.

3. Information Consumers

Most users of the data warehouse are information consumers; they will probably never compose a true and ad-hoc query. They use static or simple interactive reports that others have developed. It is easy to forget about these users, because they usually interact with the data warehouse only through the work product of others. Do not neglect these users. This group includes a large number of people, and published reports are highly visible. Set up a great communication infrastructure for distributing information widely, and gather feedback from these users to improve the information sites over time.

4. **Executives**

Executives are a special case of the information customer group. Few executives actually issue their own queries, but an executive's slightest thought can generate an outbreak of activity among the other types of users. An intelligent data warehouse designer/implementer or owner will develop a very cool digital dashboard for executives, assuming it is easy and economical to do so. Usually this should follow other data warehouse work, but it never hurts to impress the bosses

3.10 How Users Query the Data Warehouse

Information for users can be extracted from the data warehouse relational database or from the output of analytical services such as OLAP or data mining. Direct queries to the data warehouse relational database should be limited to those that cannot be accomplished through existing tools, which are often more efficient than direct queries and impose less load on the relational database.

Reporting tools and custom applications often access the database directly. Statisticians extract data for use by special analytical tools. Analysts may write complex queries to extract and compile specific information not readily accessible through existing tools. Information consumers do not interact directly with the relational database but may receive e-mail reports or access web pages that expose data from the relational database. Executives use standard reports or ask others to create specialized reports for them. When using the Analysis Services Tools in SQL servers 2000, Statisticians will often perform data mining, analysts will write MDX queries against OLAP cubes and use data mining, and information consumers will use interactive reports designed by others.

3.11 Applications of Data Warehouse

Some of the areas where data warehousing can be applied are stated as follows:

1. *Credit card churn analysis*
2. *Insurance fraud analysis*
3. *Call record analysis*
4. *Logistics management*

4.0 SELF- ASSESSMENT EXERCISE

1. *Write short notes on the following major components of a data warehouse:*

- i. *Summarized data*
- ii. *Operational systems of record*
- iii. *Integration program*
- iv. *Current detail*
- v. *Metadata*

vi. Archives

2. List and explain different types of Data Warehouse

TUTOR- MARKED ASSIGNMENT

i. (a) What do you understand by the term data warehouse?
(b) Briefly explain the following characteristics of data warehouse:

1) Subject-oriented 2) Integrated
3) Non- volatile 4) Time variant

ii. (a) List the application areas of data warehousing
(b) Differentiate between a data warehouse and data mart.

5.0 CONCLUSION

Therefore, a data warehouse usually contains historical data derived from transaction data and may include data from other sources. Also, it separates analysis workload from transaction workload and enables an organization to consolidate data from several sources.

6.0 SUMMARY

In this unit we have learnt that:

- *a data warehouse is a data structure that is optimized for collecting and storing integrated sets of historical data from multiple operational systems and feeds them to one or more data marts*
- *The major components of a data warehouse, these include summarized data, current detail, system of record, integration/transformation programs, metadata and archives*
- *The structure of a data warehouse consists of the physical data warehouse, logical data warehouse and data mart*
- *The data warehouse users can be divided into four categories namely statisticians, knowledge workers, information consumers and executives. And good numbers of application areas.*
- *The data warehouse works as a central repository where information is coming from one or more data sources.*
- *Three main types of Data warehouses are Enterprise Data Warehouse, Operational Data Store, and Data Mart.*
- *General state of a datawarehouse are Offline Operational Database, Offline Data Warehouse, Real time Data Warehouse and Integrated Data Warehouse.*
- *Four main components of Datawarehouse are Load manager, Warehouse Manager, Query Manager, End-user access tools*
- *Datawarehouse is used in diverse industries like Airline, Banking, Healthcare, Insurance, Retail etc.*
- *Implementing Datawarehosue is a 3 prong strategy viz. Enterprise strategy, Phased delivery and Iterative Prototyping.*

- *Data warehouse allows business users to quickly access critical data from some sources all in one place*

7.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Mosud, Y. O. (2009). Introduction to Data Mining and Data Warehousing, Lagos: Rashmoye Publications.

UNIT 2 DATA WAREHOUSE ARCHITECTURE

CONTENTS

1.0 INTRODUCTION

The term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon, a data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization. An operational database undergoes frequent changes on a daily basis on account of the transactions that take place. Suppose a business executive wants to analyze previous feedback on any data such as a product, a supplier, or any consumer data, then the executive will have no data available to analyze because the previous data has been updated due to transactions. A data warehouses provides us generalized and consolidated data in multidimensional view. Along with generalized and consolidated view of data, a data warehouses also provides us Online Analytical Processing (OLAP) tools. These tools help us in interactive and effective analysis of data in a multidimensional space. This analysis results in data generalization and data mining. Data mining functions such as association, clustering, classification, prediction can be integrated with OLAP operations to enhance the interactive mining of knowledge at multiple level of abstraction. That's why data warehouse has now become an important platform for data analysis and online analytical processing.

2.0. OBJECTIVES

At the end of this unit, you should be able to:

- *explain the term data warehouse architecture*
- *list the three types of data warehouse architecture*
- *describe the components of data warehouse architecture*
- *explain the use of extraction, transformation and load tools*
- *describe what is meant by resource management.*

3.0 MAIN CONTENT

3.1 Definition of Data Warehouse Architecture

A data warehouse as a storehouse, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof. Data warehousing is also the process of constructing and using the data warehouse. A data warehouse is constructed by integrating the data from multiple heterogeneous sources. It supports analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations. To integrate heterogeneous databases, we have the following two approaches:

3.2 Understanding Data Warehouse

- *A data warehouse is a database, which is kept separate from the organization's operational database.*
- *There is no frequent updating done in a data warehouse.*
- *It possesses consolidated historical data, which helps the organization to analyze its business.*
- *A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.*
- *Data warehouse systems help in the integration of diversity of application systems.*
- *A data warehouse system helps in consolidated historical data analysis.*

Data warehouse information are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains:

- ***Tuning Production Strategies** - The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.*

- ***Customer Analysis** - Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.*

- ***Operations Analysis** - Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.*

3.2.1 Why Data Warehouse is Separated from Operational Databases

Because operational databases store huge amounts of data, you may wonder, “why not perform on-line analytical processing directly on such databases instead of spending additional time and resources to construct a separate data warehouse?” A major reason for such a separation is to help promote the high performance of both systems. An operational database is designed and tuned from known tasks and workloads, such as indexing and hashing using primary keys, searching for particular records, and optimizing “canned” queries. On the other hand, data warehouse queries are often complex. They involve the computation of large groups of data at summarized levels, and may require the use of special data organization, access, and implementation methods based on multidimensional views. Processing OLAP queries in operational databases would substantially degrade the performance of operational tasks.

Moreover, an operational database supports the concurrent processing of multiple transactions. Concurrency control and recovery mechanisms, such as locking and logging, are required to ensure the consistency and robustness of transactions. An OLAP query often needs read-only access of data records for summarization and aggregation. Concurrency control and recovery mechanisms, if applied for such OLAP operations, may jeopardize the execution of concurrent transactions and thus substantially reduce the throughput of an OLTP system.

Finally, the separation of operational databases from data warehouses is based on the different structures, contents, and uses of the data in these two systems. Decision sup- port requires historical data, whereas operational databases do not typically maintain historical data. In this context, the data in operational databases, though abundant, is usually far from complete for decision making. Decision support requires consolidation (such as aggregation and summarization) of data from heterogeneous sources, resulting in high-quality, clean, and integrated data. In contrast, operational databases contain only detailed raw data, such as transactions, which need to be consolidated before analysis. Because the two systems provide quite different functionalities and require different kinds of data, it is presently necessary to maintain separate databases. However, many vendors of operational relational database management systems are beginning to optimize such systems to support OLAP queries. As this trend continues, the separation between OLTP and OLAP systems is expected to decrease.

additionally, data warehouses are kept separate from operational databases due to the following reasons:

- *An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contrast, data warehouse queries are often complex and they present a general form of data.*
- *Operational databases support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database*
- *An operational database query allows to read and modify operations, while an OLAP query needs only read only access of stored data.*
- *An operational database maintains current data. On the other hand, a data warehouse maintains historical data.*

3.2.2 Data Warehouse Features

data warehouse exhibits the following characteristics to support the management's decision-making process

- ***Subject Oriented** - A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.*

- ***Integrated*** *– A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.*

- ***Time Variant*** *– The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.*
- ***Non-volatile -*** *Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in operational database is not reflected in the data warehouse.*

3.2.3 Data Warehouse Applications

Data warehouse helps business executives to organize, analyze, and use their data for decision making. A data warehouse serves as a sole part of a plan-execute-assess "closed-loop" feedback system for the enterprise management. Data warehouses are widely used in the following fields:

- *Financial services*
- *Banking services*
- *Consumer goods*
- *Retail sectors*
- *Controlled manufacturing*

3.2.4 Types of Data Warehouse

Information processing, analytical processing, and data mining are the three types of data warehouse applications that are discussed below:

- ***Information Processing*** *– A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.*

- ***Analytical Processing*** *– A data warehouse supports analytical processing of the information stored in it. The data can be analyzed by means of basic OLAP operations, including slice-and-dice, drill down, drill up, and pivoting.*

- ***Data Mini****ng* ***-*** *Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using visualization tools.*

3.2.5 Differences between Operational Database Systems and Data Warehouses

Because most people are familiar with commercial relational database systems, it is easy to understand what a data warehouse is by comparing these two kinds of systems. The *major task of online operational database systems is to perform online transaction and query processing. These systems are called online transaction processing (OLTP) systems. They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting. Data warehouse systems, on the other hand, serve users or knowledge workers in the role ofdata analysis and decision making. Such systems can organize and present data in var- ious formats in order to accommodate the diverse needs of the different users. These systems are known as on-line analytical processing (OLAP) systems. The major difference between OLTP and OLAP are summarized as follows*

Data Warehouse (OLAP)	***Operational Database (OLTP)***
It involves historical processing of information.	*It involves day-to-day processing.*
OLAP systems are used by knowledge workers such as executives, managers, and analysts	*OLTP systems are used by clerks, DBAs, or database professionals.*
It is used to analyze the business.	*It is used to run the business.*
It focuses on Information out.	*It focuses on Data in.*
It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	*It is based on Entity Relationship Model.*
It focuses on Information out.	*It is application oriented.*
It contains historical data.	*It contains current data.*
It provides summarized and consolidated data.	*It provides primitive and highly detailed data.*
It provides summarized and multidimensional view of data.	*It provides detailed and flat relational view of data.*
The number of users is in hundreds.	*The number of users is in thousands*
The number of records accessed is in millions.	*The number of records accessed is in tens.*
The database size is from 100GB to 100 TB.	*The database size is from 100 MB to 100 GB.*

These are highly flexible.	*It provides high performance.*

3.2.6 Comparison between OLTP and OLAP systems.

Feature	***OLTP***	***OLAP***
Characteristic Orientation	*Operational processing transaction*	*Informational processing analysis*
User	*clerk, DBA, database professional*	*knowledge worker (e.g., manager, executive, analyst)*
Function	*day-to-day operations DB*	*long-term informational requirements, decision support*
DB design	*ER based, application-oriented*	*star/snowflake, subject-oriented*
Data	*current; guaranteed up-to-date* *Summarization*	*historical; accuracy maintained over time*
Summarization	*primitive, highly detailed*	*summarized, consolidated*
View	*detailed, flat relational*	*summarized, multidimensional*
Unit of work	*short, simple transaction*	*complex query mostly*
Access	*read/write*	*mostly read* *information*
Focus	*data in*	*Information out*
Operations	*index/hash on primary key tens*	*lots of scans* *millions*
Number of records accessed	*tens*	*millions*
Number of users DB	*thousands*	*hundreds* *100*
DB size	*100 MB to GB*	*100 GB to TB high*
Priority	*high performance, high availability*	*high flexibility, end-user autonomy*
Metric	*transaction throughput*	*query throughput, response time*

3.3 *Data Warehouse Design Process:*

A data warehouse can be built using a top-down approach, a bottom-up approach, or a combination of both.

- *The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood.*

- *The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments.*

- *In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.*

The warehouse design process consists of the following steps:

- *Choose a business process to model, for example, orders, invoices, shipments, inventory, account administration, sales, or the general ledger. If the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.*

- *Choose the grain of the business process. The grain is the fundamental, atomic level of data to be represented in the fact table for this process, for example, individual transactions, individual daily snapshots, and so on.*

- *Choose the dimensions that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transaction type, and status.*
- *Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.*

3.4 A Three Tier Data Warehouse Architecture

Tier-1:

The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the *data warehouse. The data are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.*

Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection). This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

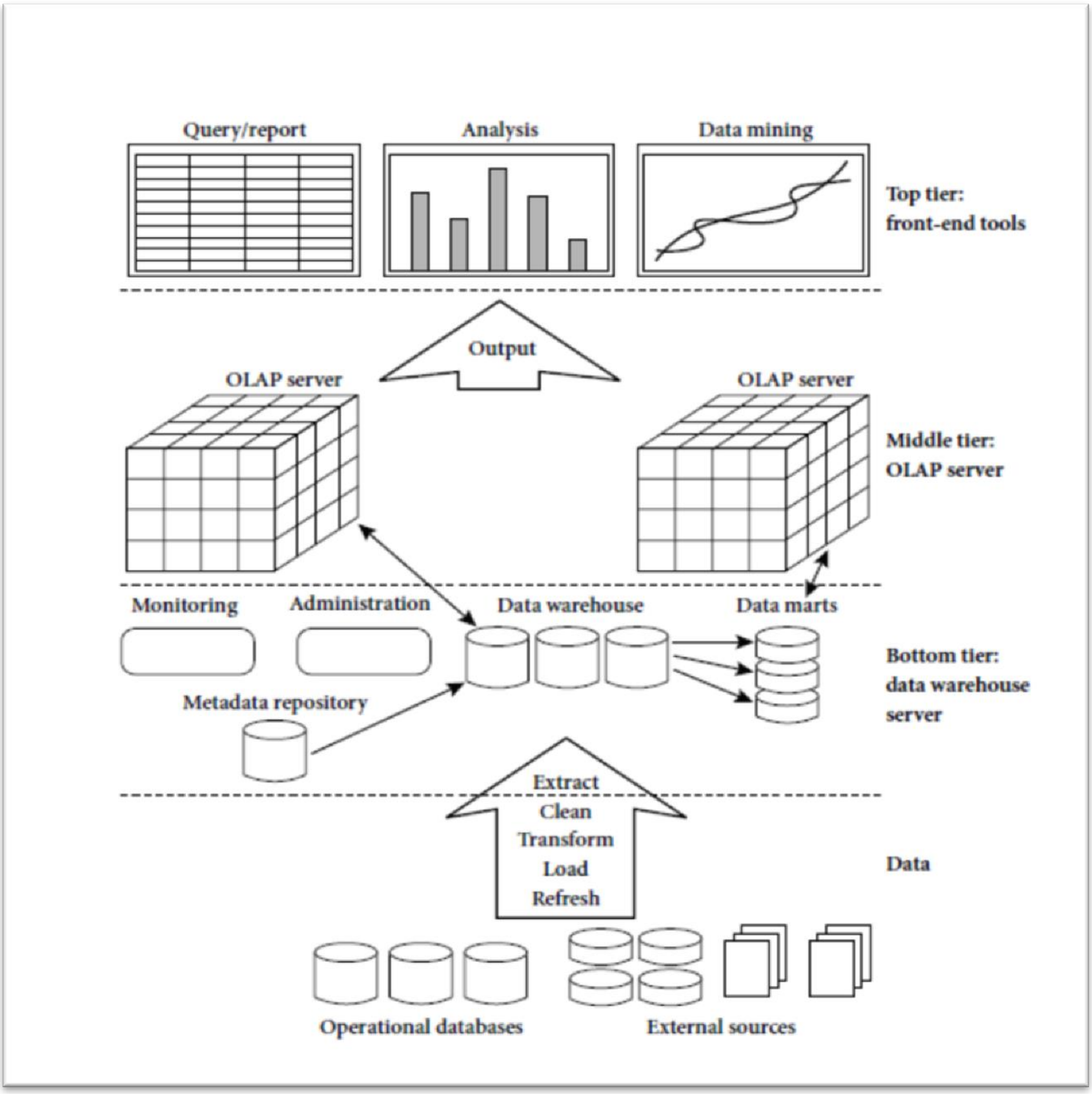


***Fig. 2.1** Three-tier Data warehouse architecture*

Tier-2

The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model or a multidimensional OLAP.

- *OLAP model is an extended relational DBMS that maps operations on multidimensional data to standard relational operations.*

- *A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations*

Tier-3:

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on)

From the architecture point of view, there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse

- ***Enterprise warehouse**: An enterprise warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope. It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond. An enterprise data warehouse may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms. It requires extensive business modeling and may take years to design and build.*

- ***Data mart**: A data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in data marts tend to be summarized. Data marts are usually implemented on low-cost departmental servers that are UNIX/LINUX- or Windows-based. The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years. However, it may involve complex integration in the long run if its design and planning were not enterprise-wide. Depending on the source of data, data marts can be categorized as independent or dependent. Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area. Dependent data marts are sourced directly from enterprise data warehouses.*

- ***Virtual warehouse**: A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capacity on operational database servers.*

"What are the pros and cons of the top-down and bottom-up approaches to data warehouse development?" The top-down development of an enterprise warehouse serves as a systematic solution and minimizes integration problems. However, it is expensive, takes a long time to develop, and lacks flexibility due to the difficulty in achieving consistency and consensus for a common data model for the entire organization. The bottom-up approach to the design, development, and deployment of independent data marts provides flexibility, low cost, and rapid return of investment. It, however, can lead to problems when integrating various disparate data marts into a consistent enterprise data warehouse.

A recommended method for the development of data warehouse systems is to implement the warehouse in an incremental and evolutionary manner. First, a high-level corporate data model is defined within a reasonably short period (such as one or two months) that provides a corporate-wide, consistent, integrated view of data among different subjects and potential usages. This high-level model, although it will need to be refined in the further development of enterprise data warehouses and departmental data marts, will greatly reduce future integration problems. Second, independent data marts can be implemented in parallel with the enterprise warehouse based on the same corporate data model set as above. Third, distributed data marts can be constructed to integrate different data marts via hub servers. Finally, a multitier data warehouse is constructed where the enterprise warehouse is the sole custodian of all warehouse data, which is then distributed to the various dependent data marts.

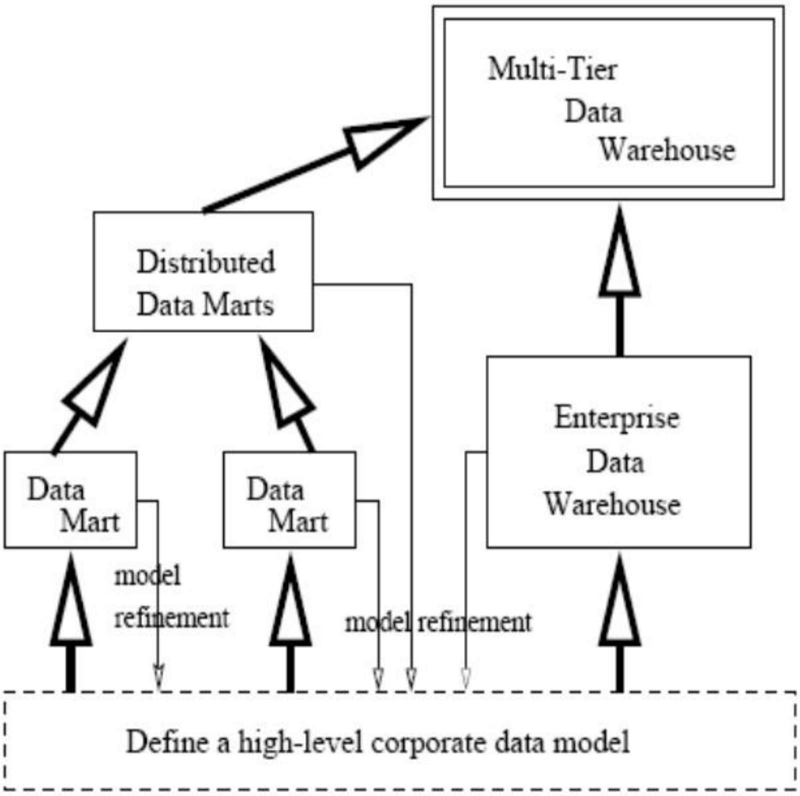


Fig. 2.2 *A recommended approach for data warehouse development.*

3.6 Data warehouse Back-End Tools and Utilities

The ETL (Extract Transformation Load) process

In this section we will discussed about the 4 major process of the data warehouse. They are extract (data from the operational systems and bring it to the data warehouse) transform (the data into internal format and structure of the data warehouse), cleanse (to make sure it is of sufficient quality to be used for decision making) and load (cleanse data is put into the data warehouse).

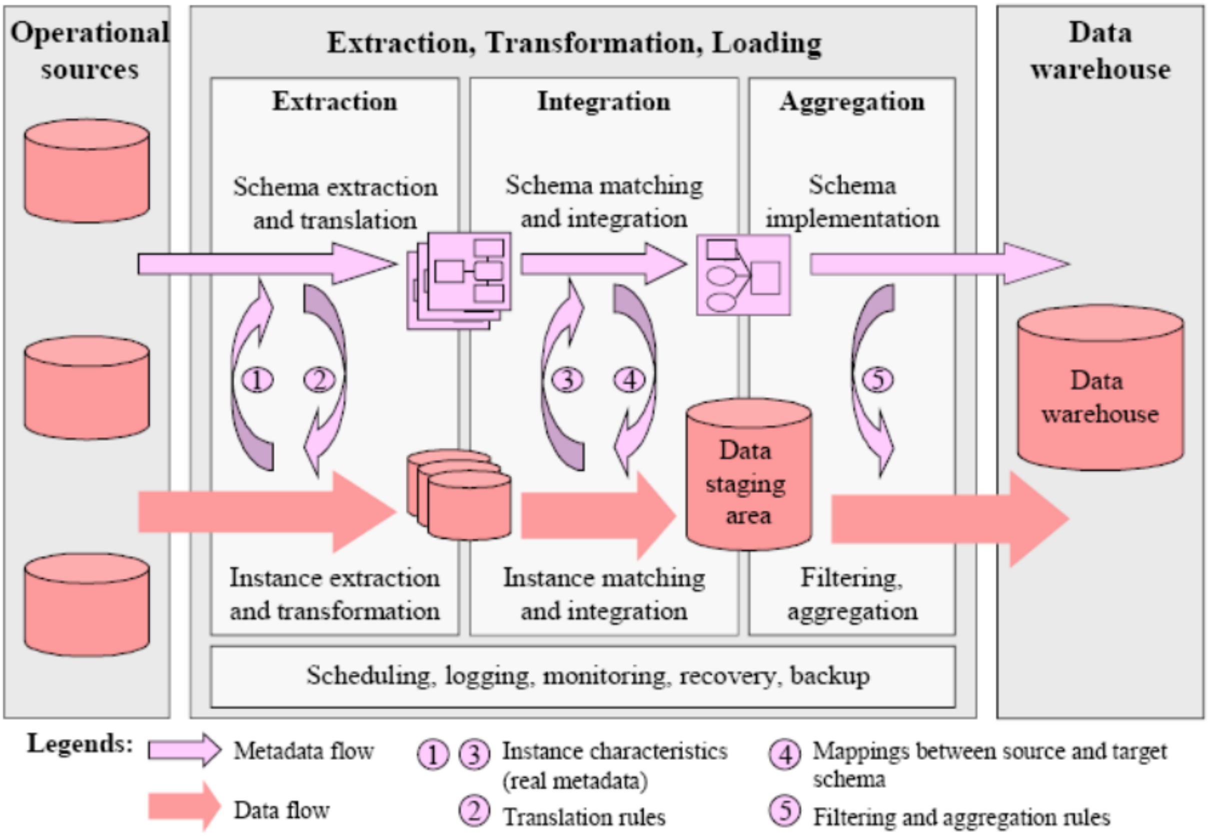

Operational sources
Extraction, Transformation, Loading
Data warehouse
Extraction
Integration
Aggregation
Schema extraction and translation
Schema matching and integration
Schema implementation
① ② ③ ④ ⑤
Data staging area
Data warehouse
Instance extraction and transformation
Instance matching and integration
Filtering, aggregation
Scheduling, logging, monitoring, recovery, backup
Legends:
Metadata flow
Data flow
① ③ Instance characteristics (real metadata)
② Translation rules
④ Mappings between source and target schema
⑤ Filtering and aggregation rules

Fig. 2.3 *Steps of building a data warehouse: the ETL process*

The four processes from extraction through loading often referred collectively as Data Staging.

3.6.1 EXTRACT

Some of the data elements in the operational database can be reasonably be expected to be useful in the decision making, but others are of less value for that purpose. For this reason, it is necessary to extract the relevant data from the operational database before bringing into the data warehouse. Many commercial tools are available to help with the extraction process

Data Junction

is one of the commercial products. The user of one of these tools typically has an easyto-use windowed interface by which to specify the following:

i. Which files and tables are to be accessed in the source database?
ii. Which fields are to be extracted from them? This is often done internally by SQL Select statement.
iii. What are those to be called in the resulting database?
iv. What is the target machine and database format of the output?
v. On what schedule should the extraction process be repeated?

3.6.2. TRANSFORM

The operational databases developed can be based on any set of priorities, which keeps changing with the requirements. Therefore those who develop data warehouse based on these databases are typically faced with inconsistency among their data sources. Transformation process deals with rectifying any inconsistency (if any). One of the most common transformation issues is _Attribute Naming Inconsistency'. It is common for the given data element to be referred to by different data names in different databases. Employee Name may be EMP_NAME in one database, ENAME in the other. Thus one set of Data Names are picked and used consistently in the data warehouse. Once all the data elements have right names, they must be converted to common formats. The conversion may encompass the following:

i. Characters must be converted ASCII to EBCDIC or vice versa.
ii. Mixed Text may be converted to all uppercase for consistency.
iii. Numerical data must be converted in to a common format.
iv. Data Format has to be standardized.
v. Measurement may have to convert. (Rs/ $)
vi. Coded data (Male/ Female, M/F) must be converted into a common format.

All these transformation activities are automated and many commercial products are available to perform the tasks. ***Data MAPPER*** *from Applied Database Technologies is one such comprehensive tool.*

3.6.3 CLEANSING

Information quality is the key consideration in determining the value of the information. The developer of the data warehouse is not usually in a position to change the quality of its underlying historic data, though a data warehousing project can put spotlight on the data quality issues and lead to improvements for the future. It is, therefore, usually necessary to go through the data entered into the data warehouse and make it as error free as possible. This process is known as ***Data Cleansing.***

Data Cleansing must deal with many types of possible errors. These include missing data and incorrect data at one source; inconsistent data and conflicting data when two or more source are involved. There are several algorithms followed to clean the data, which will be discussed in the coming lecture notes.

3.6.4 LOADING

Loading often implies physical movement of the data from the computer(s) storing the source database(s) to that which will store the data warehouse database, assuming it is different. This takes place immediately after the extraction phase. The most common channel for data movement is a high-speed communication link. Ex: Oracle Warehouse Builder is the API from Oracle, which provides the features to perform the ETL task on Oracle Data Warehouse.

3.7 Data cleaning problems

This section classifies the major data quality problems to be solved by data cleaning and data transformation. As we will see, these problems are closely related and should thus be treated in a uniform way. Data transformations are needed to support any changes in the structure, representation or content of data. These transformations become necessary in many situations, e.g., to deal with schema evolution, migrating a legacy system to a new information system, or when multiple data sources are to be integrated. As shown in Fig. 2.4 we roughly distinguish between single-source and multi-source problems and between schema- and instance-related problems. Schema-level problems of course are also reflected in the instances; they can be addressed at the schema level by an improved schema design (schema evolution), schema translation and schema integration. Instance-level problems, on the other hand, refer to errors and inconsistencies in the actual data contents which are not visible at the schema level. They are the primary focus of data cleaning.

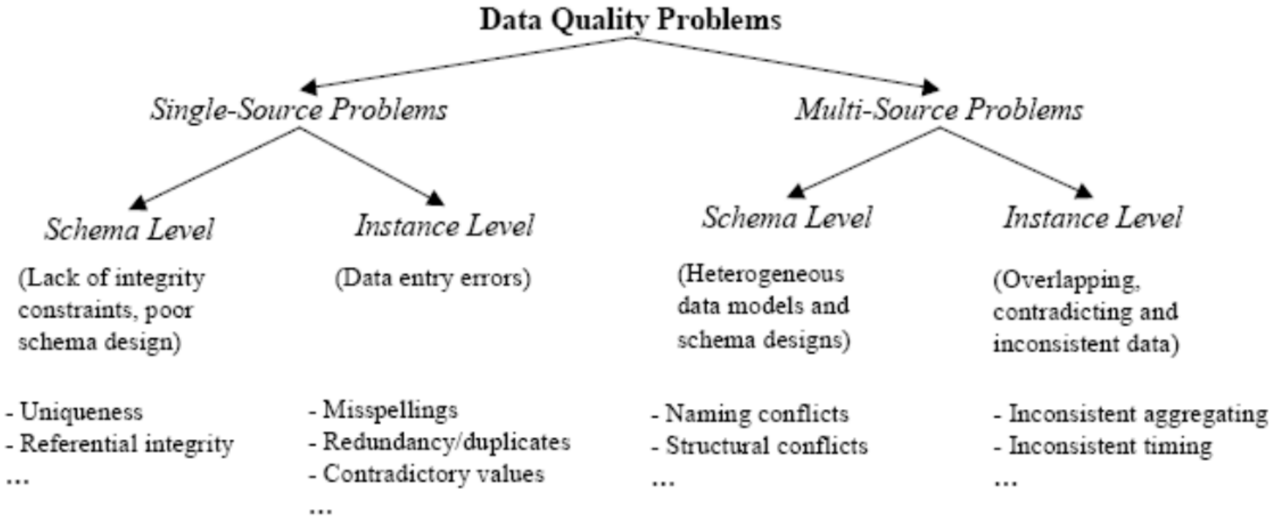
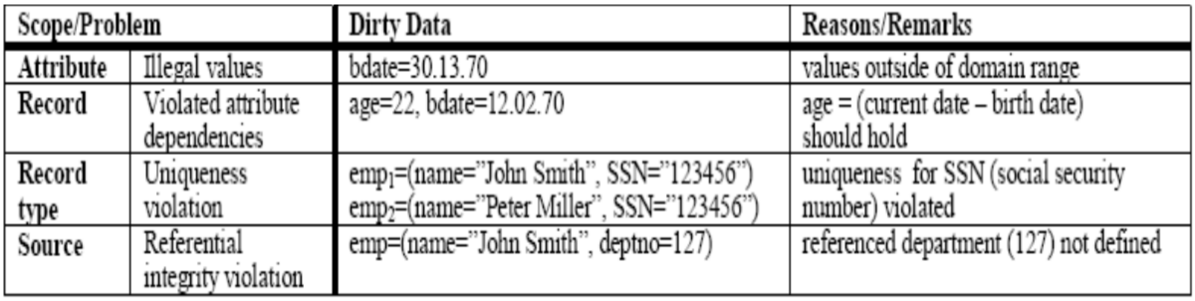


Fig. 2.4 Classification of data quality problem in data source.

3.7.1 Single-Source Problems

The data quality of a source largely depends on the degree to which it is governed by schema and integrity constraints controlling permissible data values. For sources without schema, such as files, there are few restrictions on what data can be entered and stored, giving rise to a high probability of errors and inconsistencies. Database systems, on the other hand, enforce restrictions of a specific data model (e.g., the relational approach requires simple attribute values, referential integrity, etc.) as well as application-specific integrity constraints. Schema-related data quality problems thus occur because of the lack of appropriate model-specific or application-specific integrity constraints, e.g., due to data model limitations or poor schema design, or because only a few integrity constraints were defined to limit the overhead for integrity control. Instance-specific problems relate to errors and inconsistencies that cannot be prevented at the schema level (e.g., misspellings).



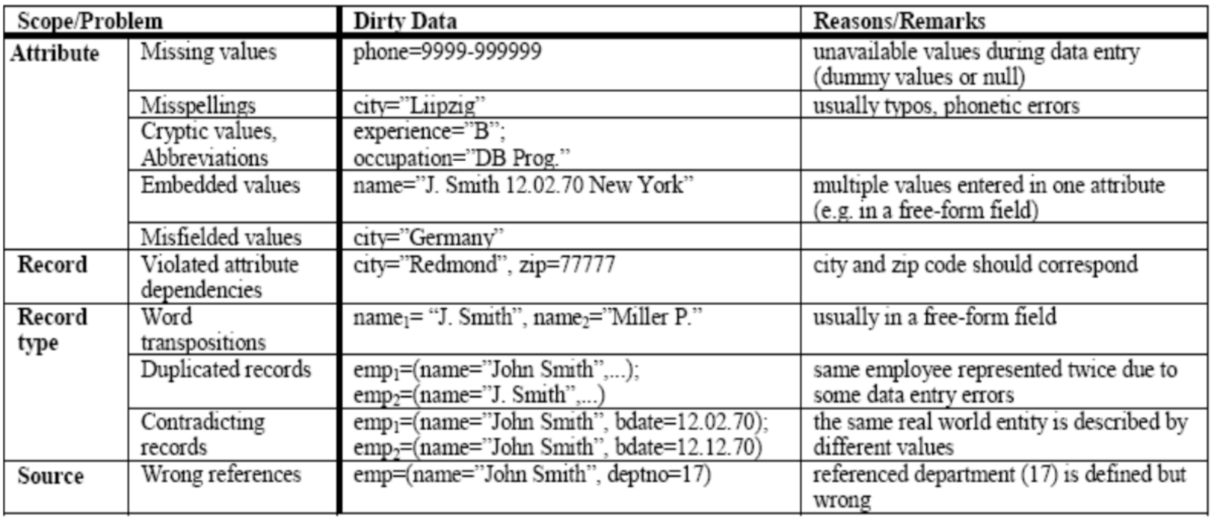
Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = (current date – birth date) should hold

Record type	Uniqueness violation	emp$_1$=(name=”John Smith”, SSN=”123456”)
		emp$_2$=(name=”Peter Miller”, SSN=”123456”)
			uniqueness for SSN (social security number) violated

Source	Referential integrity violation	emp=(name=”John Smith”, deptno=127)	referenced department (127) not defined

Fig. 2.5 Examples for single-source problems at schema level (violated integrity constraints)

For both schema- and instance-level problems we can differentiate different problem scopes: attribute (field), record, record type and source; examples for the various cases are shown in Tables 1 and 2. Note that uniqueness constraints specified at the schema level do not prevent duplicated instances, e.g., if information on the same real world entity is entered twice with different attribute values (see example in Table 2)



Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city=”Liipzig”	usually typos, phonetic errors

	Cryptic values, Abbreviations	experience=”B”;
		occupation=”DB Prog.”	

	Embedded values	name=”J. Smith 12.02.70 New York”	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city=”Germany”	
Record	Violated attribute dependencies	city=”Redmond”, zip=77777	city and zip code should correspond
Record type	Word transpositions	name$_1$= “J. Smith”, name$_2$=”Miller P.”	usually in a free-form field

	Duplicated records	emp$_1$=(name=”John Smith”,...);
		emp$_2$=(name=”J. Smith”,...)
			same employee represented twice due to some data entry errors

	Contradicting records	emp$_1$=(name=”John Smith”, bdate=12.02.70);
		emp$_2$=(name=”John Smith”, bdate=12.12.70)
			the same real world entity is described by different values

Source	Wrong references	emp=(name=”John Smith”, deptno=17)	referenced department (17) is defined but wrong

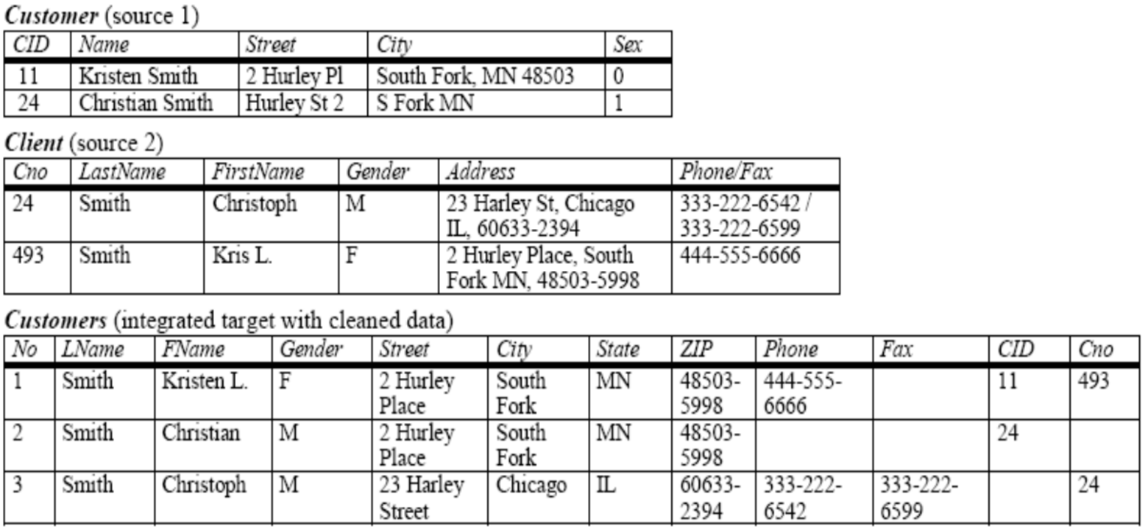
Fig. 2.6 Examples for single-source problems at instance level

3.7.2 Multi Source problems

The problems present in single sources are aggravated when multiple sources need to be integrated. Each source may contain dirty data and the data in the sources may be represented differently, overlap or contradict. This is because the sources are typically developed, deployed and maintained independently to serve specific needs. This results in a large degree of heterogeneity w.r.t. data management systems, data models, schema designs and the actual data.

At the schema level, data model and schema design differences are to be addressed by the steps of schema translation and schema integration, respectively. The main problems w.r.t. schema design are naming and structural conflicts. Naming conflicts arise when the same name is used for different objects (homonyms) or different names are used for the same object (synonyms). Structural conflicts occur in many variations and refer to different representations of the same object in different sources, e.g., attribute vs. table representation, different component structure, different data types, different integrity constraints, etc. In addition to schema-level conflicts, many conflicts appear only at the instance level (data conflicts). All problems from the single-source case can occur with different representations in different sources (e.g., duplicated records, contradicting records,…). Furthermore, even when there are the same attribute names and data types, there may be different value representations (e.g., for marital status) or different interpretation of the values (e.g., measurement units Dollar vs. Euro) across sources. Moreover, information in the sources may be provided at different aggregation levels (e.g., sales per product vs. sales per product group) or refer to different points in time (e.g. current sales as of yesterday for source 1 vs. as of last week for source 2).

A main problem for cleaning data from multiple sources is to identify overlapping data, in particular matching records referring to the same real-world entity (e.g., customer). This problem is also referred to as the object identity problem, duplicate elimination or the merge/purge problem. Frequently, the information is only partially redundant and the *sources may complement each other by providing additional information about an entity. Thus duplicate information should be purged out and complementing information should be consolidated and merged in order to achieve a consistent view of real world entities.*



Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Figure 2.7 Example of multi-source problems at schema and instance level

The two sources in the example of Fig. 2.7 are both in relational format but exhibit schema and data conflicts. At the schema level, there are name conflicts (synonyms Customer/Client, Cid/Cno, Sex/Gender) and structural conflicts (different representations for names and addresses). At the instance level, we note that there are different gender representations ("0"/"1" vs "F"/"M") and presumably a duplicate record (Kristen Smith). The latter observation also reveals that while Cid/Cno are both source-specific identifiers, their contents are not comparable between the sources; different numbers (11/493) may refer to the same person while different persons can have the same number (24). Solving these problems requires both schema integration and data cleaning; the third table shows a possible solution. Note that the schema conflicts should be resolved first to allow data cleaning, in particular detection of duplicates based on a uniform representation of names and addresses, and matching of the Gender/Sex values.

Data cleaning approaches

In general, data cleaning involves several phases

Data analysis*: In order to detect which kinds of errors and inconsistencies are to be removed, a detailed data analysis is required. In addition to a manual inspection of the data or data samples, analysis programs should be used to gain metadata about the data properties and detect data quality problems*

Definition of transformation workflow and mapping rules:
Depending on the number of Data sources ,their degree of heterogeneity and the dirtyness of the data, a large number of data transformation and cleaning steps may have to be executed.

Sometime, a schema translation is used to map sources to a common data model; for data warehouses, typically a relational representation is used. Early data cleaning steps can correct single-source instance problems and prepare the data for integration. Later steps deal with schema/data integration and cleaning multisource instance problems, e.g., duplicates. For data warehousing, the control and data flow for these transformation and cleaning steps should be specified within a workflow that defines the ETL process

The schema-related data transformations as well as the cleaning steps should be specified by a declarative query and mapping language as far as possible, to enable automatic generation of the transformation code. In addition, it should be possible to invoke user written cleaning code and special purpose tools during a data transformation workflow. The transformation steps may request user feedback on data instances for which they have no built-in cleaning logic.

Verification: *The correctness and effectiveness of a transformation workflow and the transformation definitions should be tested and evaluated, e.g., on a sample or copy of the source data, to improve the definitions if necessary. Multiple iterations of the analysis, design and verification steps may be needed, e.g., since some errors only become apparent after applying some transformations.*

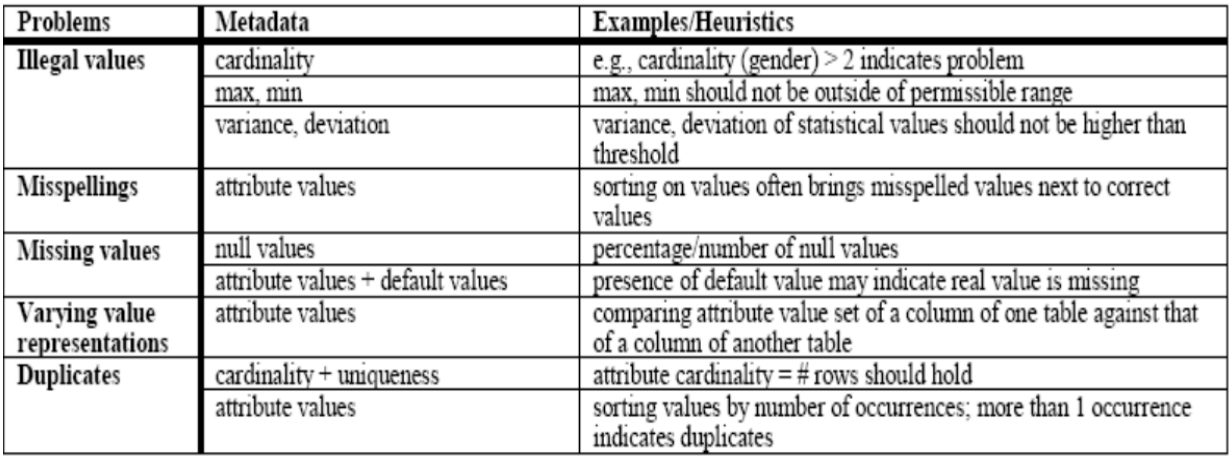
Transformation: *Execution of the transformation steps either by running the ETL workflow for loading and refreshing a data warehouse or during answering queries on multiple sources.*

Backflow of cleaned data: *After (single-source) errors are removed, the cleaned data should also replace the dirty data in the original sources in order to give legacy applications the improved data too and to avoid redoing the cleaning work for future data extractions*

Data analysis

Metadata reflected in schemas is typically insufficient to assess the data quality of a source, especially if only a few integrity constraints are enforced. It is thus important to analyse the actual instances to obtain real (reengineered) metadata on data characteristics or unusual value patterns. This metadata helps finding data quality problems. Moreover, it can effectively contribute to identify attribute correspondences between source schemas (schema matching), based on which automatic data transformations can be derived.

There are two related approaches for data analysis, data profiling and data mining. Data profiling focuses on the instance analysis of individual attributes. It derives information such as the data type, length, value range, discrete values and their frequency, variance, uniqueness, occurrence of null values, typical string pattern (e.g., for phone numbers), etc., providing an exact view of various quality aspects of the attribute.



Problems	Metadata	Examples/Heuristics
Illegal values	cardinality	e.g., cardinality (gender) > 2 indicates problem
	max, min	max, min should not be outside of permissible range
	variance, deviation	variance, deviation of statistical values should not be higher than threshold
Misspellings	attribute values	sorting on values often brings misspelled values next to correct values
Missing values	null values	percentage/number of null values
	attribute values + default values	presence of default value may indicate real value is missing
Varying value representations	attribute values	comparing attribute value set of a column of one table against that of a column of another table
Duplicates	cardinality + uniqueness	attribute cardinality = # rows should hold
	attribute values	sorting values by number of occurrences; more than 1 occurrence indicates duplicates

Figure 2.8 Examples for the use of reengineered metadata to address data quality problems

3.8 Metadata Repository

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Figure 2.8 showed a metadata repository within the bottom tier of the data warehousing architecture. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for timestamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

A metadata repository should contain the following:

- *A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents*
- *Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails)*
- *The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports*
- *The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control)*
- *Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles*
- *Business metadata, which include business terms and definitions, data ownership information, and charging policies*

A data warehouse contains different levels of summarization, of which metadata is one type. Other types include current detailed data (which are almost always on disk), older detailed data (which are usually on tertiary storage), lightly summarized data and highly summarized data (which may or may not be physically housed).

Metadata play a very different role than other data warehouse data and are important for many reasons. For example, metadata are used as a directory to help the decision support system analyst locate the contents of the data warehouse, as a guide to the map- ping of data when the data are transformed from the operational environment to the data warehouse environment, and as a guide to the algorithms used for summarization between the current detailed data and the lightly summarized data, and between the lightly summarized data and the highly summarized data. Metadata should be stored and managed persistently (i.e., on disk).

4.0 SELF ASSESMENT EXERCISE(S)

1. *List and briefly explain the seven major components of data warehouse architecture.*
2. *List and explain data warehouse features*
3. *Explain why data warehouse is separated from operational databases*
4. *Explain other Data Warehouse Applications not stated in this chapter*
5. *Explain the three tier of data warehouse architecture*
6. *Differentiate between data marts and data warehouse*
7. *What is the importance of extract transformation load tool to data warehouse?*
8. *List and explain data cleaning approaches*
9. *Explain what is meant by Data cleaning problems*

5.0 CONCLUSION

Therefore, a data warehouse usually contains historical data derived from transaction data and may include data from other sources. Also, it separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data organized in support of management decision making. Several factors distinguish data warehouses from operational databases. Because the two systems provide quite different functionalities and require different kinds of data, it is necessary to maintain data warehouses separately from operational databases.

6.0 SUMMARY

In this unit we have learnt that:

- *Data warehouses often adopt a three-tier architecture. The bottom tier is a warehouse database server, which is typically a relational database system. The middle tier is an OLAP server, and the top tier is a client, containing query and reporting tools*

- *A data warehouse contains back-end tools and utilities for populating and refresh- ing the warehouse. These cover data extraction, data cleaning, data transformation, loading, refreshing, and warehouse management. Data*
- *On-line analytical processing (OLAP) can be performed in data warehouses/marts using the multidimensional data model. Typical OLAP operations include roll- up, drill-(down, across, through), slice-and-dice, pivot (rotate), as well as statistical operations such as ranking and computing moving averages and growth rates. OLAP operations can be implemented efficiently using the data cube structure.*
- *OLAP query processing can be made more efficient with the use of indexing techniques. In bitmap indexing, each attribute has its own bitmap index table. Bitmap indexing reduces join, aggregation, and comparison operations to bit arithmetic. Join indexing registers the joinable rows of two or more relations from a relational database, reducing the overall cost of OLAP join operations. Bitmapped join indexing, which combines the bitmap and join index methods, can be used to further speed up OLAP query processing. Data*

7.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). *Data Mining: Concepts and Techniques, Second Edition*

Yanchang Zhao, (2015). *R and Data Mining Examples and Case Studies*

Charu C. Aggarwal, 2015. *The Textbook 2015th Edition*

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

Shah, C. (2020). *A Hands-On Introduction to Data Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Anil, R. Data Warehouse and its Applications in Agriculture, Indian Agriculture Statistics Research Institute Library Avenue, New Delhi:110 012.

Data Management and Data Warehouse Domain Technical Architecture, June 6, 2002.

.

UNIT 3 DATA WAREHOUSE DESIGN

CONTENTS

1.0 INTRODUCTION

Data warehouse support business decisions by collecting, consolidating and organizing data for reporting and analysis with tools such as on-line analytical processing (OLAP) and data mining. Though data warehouses are built on relational database technology, but the design of a data warehouse database differs substantially from the design of an online transaction processing system (OLTP) database.

This unit examines the approaches and choices to be considered when designing and implementing a data warehouse. Also to be discussed is the different strategies to test a data warehouse application

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- *differentiate between a logical and physical design*
- *list the basic methodologies used in building a data warehouse*
- *explain the phases involved in developing a data warehouse*
- *state the data warehouse testing life cycle.*

3.0 MAIN CONTENT

3.1 Designing a Data Warehouse

Before embarking on the design of a data warehouse, it is imperative that the architectural goals of the data warehouse be clear and well understood (see also: Module 3, Unit 2: Data Warehouse Architecture Goals). Because the purpose of a data warehouse is to serve users, it is vital to understand the various types of users, their needs, and the characteristics of their interactions with the data warehouse.

Logical Versus Physical Design of Data Warehouses

Once an organization has decided to build a data warehouse and has defined the business requirements, agreed upon the scope of application and has created a conceptual design; the next thing is to translate the requirements into a system deliverable. In order to do this, you create the logical and physical design for the data warehouse.

Logical design involves describing the purpose of a system and what the system will do as against to how it is actually going to be implemented physically. It does not include any specific hardware or software requirements. Also, logical design lays out the system's components and their relationship to one another as they would appear to users. Physical design is the process of translating the abstract logical model into the specific technical design for the new system. It is the actual bolt and nut of the system as it includes the technical specification that transforms the abstract logical design plan into a functioning system.

3.2 Data Warehouse Design Methodologies

The basic techniques used in building a data warehouse are as follows:

a) Bottom-up Design
b) Top-down Design
c) Hybrid Design

(i) Bottom-up Design

Ralph Kimball, a well-known author on data warehousing is a proponent of an approach frequently considered as bottom-up to data warehouses design. In this approach smaller local data warehouse, known as data marts are firstly created to provide reporting and analytical capabilities for specific business processes. Data marts contain atomic data and, if necessary, summarised data. These data marts can eventually be merged together to create a comprehensive data warehouse. The combination of data marts is managed through the implementation of what Kimball calls "data warehouse bus architecture". Business value can be returned as quickly as the first data marts can be created. Maintaining tight management over the data warehouse bus architecture is fundamental to maintaining the integrity of the data warehouse.

(ii) Top–down Design

In this design we first build a data warehouse for the complete organisation and from this select the information needed for our department. Also, Willian Inmon who was one of the leading proponents of the top-down approach to data warehouse design describes it as a data warehouse designed using a normalised enterprise data model. With 'atomic' data, that is data at all the lowest of detail stored in the data warehouse.

The top-down design methodology generates highly consistent dimensional views of data across data marts since all data marts are loaded from centralised repository. Top-down design has also proven to be robust against business changes. Also, the top-down methodology can be inflexible and indifferent to changing departmental needs during the implementation phases.

(iii) Hybrid Design

Over time it has become apparent to proponents of bottom up and top- down data warehouse design that both methodologies have benefits and risks. Hybrid methodologies have evolved to take advantage of the fast turn-around time of bottom-up design and the enterprise-wide data consistency of top-down design.

SELF-ASSESSMENT EXERCISE 1

List and explain the three basic technologies in developing a data warehouse.

3.3 Developing a Data Warehouse

The phases of a data warehouse project listed as follow are similar to those of most database projects, starting with identifying requirements and ending with deploying the system:

- *identify and gather requirements*
- *design the dimensional model*
- *develop the architecture, including the operational data store (ODS)*

3.3.1 Identify and Gather Requirements

You must identify the sponsors. A successful data warehouse project needs a sponsor in the business organisation and usually a second sponsor in the information technology group. Sponsor must understand and support the business value of the project. There is need to understand the business before entering into discussions with users. Then interview and work with the users; it is necessary to learn the need of the users and turn these needs into project requirements. It is also necessary to find out what information they need to be more successful at their jobs, and not what data they think should be in the data warehouse. Moreover, it is the responsibility of data warehouse designers to determine what data is necessary to provide the information.

The issues to discuss are the users' objectives and challenges, and how they go about making business decisions. Business users should be closely tied to the design team during the logical design process; they are the people that understand the meaning of existing data. Many successfully projects include several business users on the design team to act as data experts and sounding boards for design concepts. Whatever the structure of the team, it is important that business users feel ownership for the resulting system.

Interview the data experts after interviewing several users and find out from the experts what data exists and where it resides, but only after understanding the basic business needs of the end users. The information about available data is needed early in the process before completing the analysis of the business needs, but the physical design of existing data should not be allowed to have much influence on discussions about business needs. It is very important to communicate with users often thoroughly so that everyone would participate in the progress of the requirements definition.

3.3.2 Multidimensional Data Model

The most popular data model for data warehouses is a multidimensional model. This model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema. Let's have a look at each of these schema types.

From Tables and Spreadsheets to Data Cubes

"What is a data cube?" A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts. In general terms, dimensions are the perspectives or entities with respect to which
an organization wants to keep records.

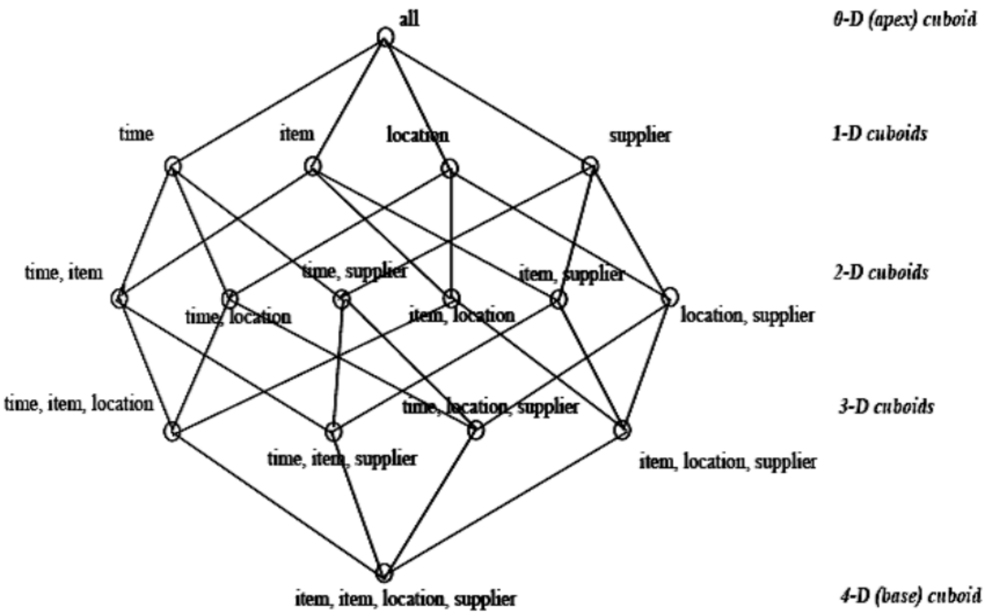


Figure 3.1. *Lattice of cuboids, making up a 4-D data cube for the dimensions time, item, location, and supplier. Each cuboid represents a different degree of summarization*

Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for on-line transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates on-line data analysis. The most popular data model for a data warehouse is a multidimensional model.
Such a model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema. Let's look at each of these schema types.

- ***Star schema**: The star schema is a modeling paradigm in which the data warehouse contains (1) a large central table (fact table), and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.*

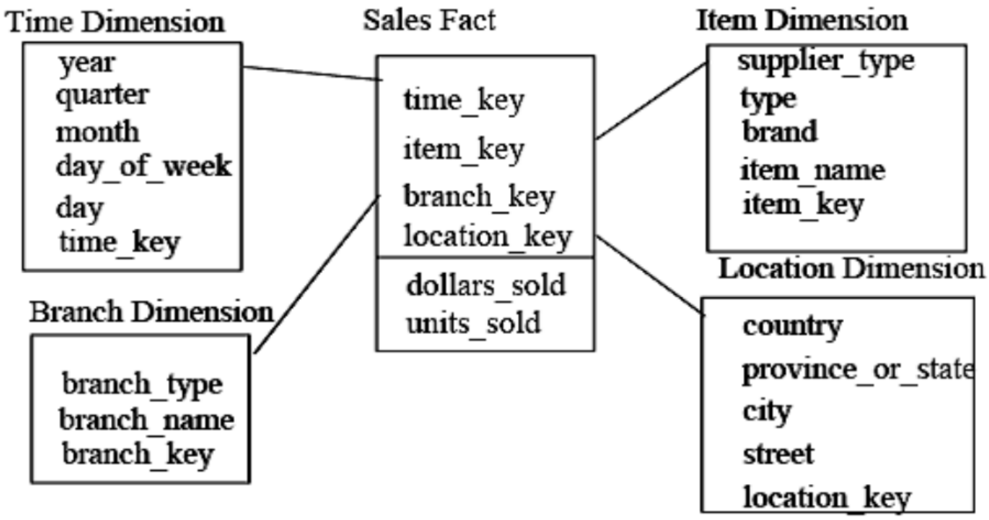


Fig. 3.2 *Star schema of a data warehouse for sales*

- ***Snowflake schema**: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake. The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form. Such a table is easy to maintain and also saves storage space because a large dimension table can be extremely large when the dimensional structure is included as columns*

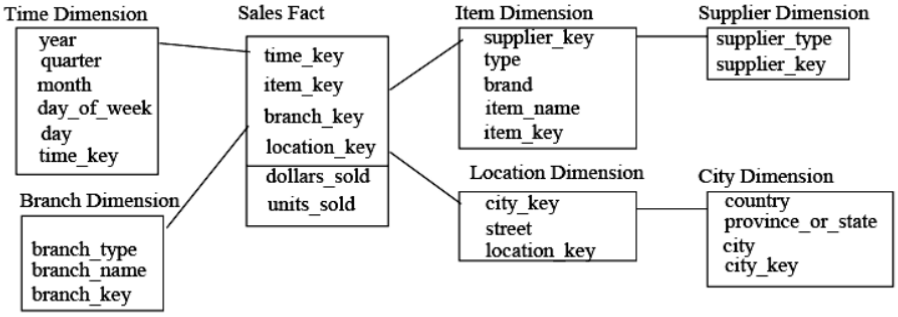


Fig.3.3 *Snowflake schema of a data warehouse for sales.*

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

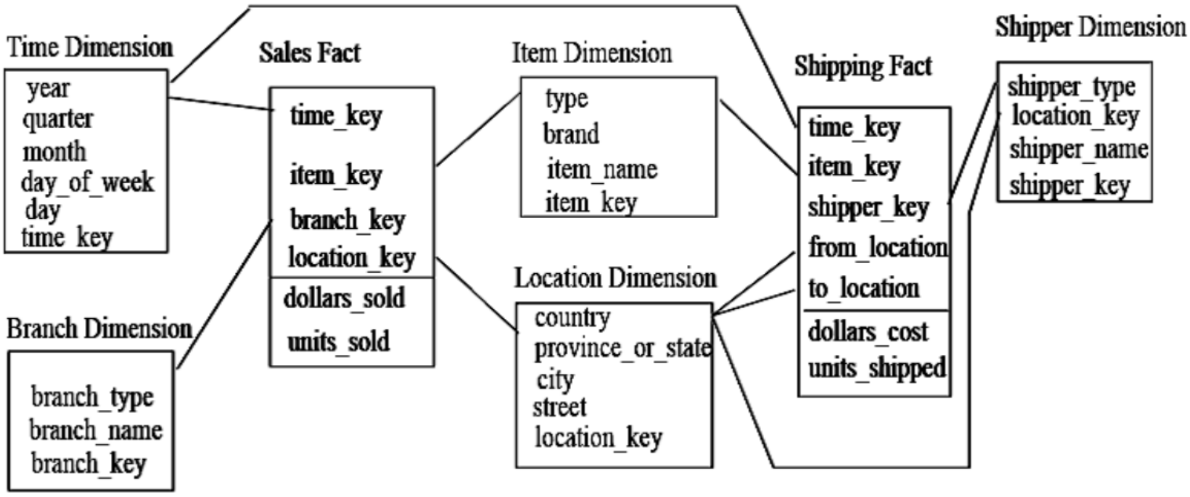


Fig. 3.4. *Fact constellation schema of a data warehouse for sales and shipping.*

Example for Defining Star, Snowflake, and Fact Constellation Schemas

A Data Mining Query Language, DMQL: Language Primitives

- *Cube Definition (Fact Table)*
- *define cube <cube_name> [<dimension_list>]: <measure_list>*
- *Dimension Definition (Dimension Table)*
- *define dimension <dimension_name> as (<attribute_or_subdimension_list>)*
- *Special Case (Shared Dimension Tables)*
  - *First time as "cube definition"*
  - *Define dimension <dimension_name> as <dimension_name_first_name> in cube <cube_name_first_name>*

Defining a Star Schema in DMQL
define cube sales_star [time, item, branch, location]:
dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars), units_sold = count(*)
define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier_type)
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key, street, city, province_or_state, country)

Defining a Snowflake Schema in DMQL
define cube sales_snowflake [time, item, branch, location]:
dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars), units_sold = count(*)
define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier(supplier_key, supplier_type))
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key, street, city(city_key, province_or_state, country))

Defining a Fact Constellation in DMQL

define cube sales [time, item, branch, location]:

dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars), units_sold = count(*)

define dimension time as (time_key, day, day_of_week, month, quarter, year)

define dimension item as (item_key, item_name, brand, type, supplier_type)

define dimension branch as (branch_key, branch_name, branch_type)

define dimension location as (location_key, street, city, province_or_state, country)

define cube shipping [time, item, shipper, from_location, to_location]:

dollar_cost = sum(cost_in_dollars), unit_shipped = count(*)

define dimension time as time in cube sales

define dimension item as item in cube sales

define dimension shipper as (shipper_key, shipper_name, location as location in cube sales, shipper_type)

define dimension from_location as location in cube sales

define dimension to_location as location in cube sales

Measures: Three Categories

Measure: a function evaluated on aggregated data corresponding to given dimension-value pairs.

Measures can be:

- *distributive: if the measure can be calculated in a distributive manner.*
  - *E.g., count(), sum(), min(), max().*
- *algebraic: if it can be computed from arguments obtained by applying distributive aggregate functions.*
  - *E.g., avg()=sum()/count(), min_N(), standard_deviation().*
- *holistic: if it is not algebraic.*
  - *E.g., median(), mode(), rank().*

Concept Hierarchies

A Concept hierarchy defines a sequence of mappings from a set of low level Concepts to higher level, more general Concepts. Concept hierarchies allow data to be handled at varying levels of abstraction. Consider a concept hierarchy for the dimension location.Cityvaluesfor locationincludeVancouver,Toronto,NewYork,andChicago.Each city, however, can be mapped to the province or state to which it belongs. For example, Vancouver canbemapped toBritishColumbia, andChicago to Illinois.Theprovincesand states can in turn be mapped to the country to which they belong, such as Canada or the USA.Thesemappings forma concept hierarchy for the dimension location, mapping a set oflow-level concepts (i.e., cities) to higher-level, more general concepts (i.e., countries). The concept hierarchy described above is illustrated in Figure 3.5.

Many concept hierarchies are implicit within the database schema. For example, suppose that the dimension location is described by the attributes number, street, city, province or state,zipcode,andcountry.These attributes are related by a total order, forming a concept hierarchy such as "street < city < province or state < country". This hierarchy is shown in Figure 3.5. Alternatively, the attributes of a dimension may be organized in a partial order, forming a lattice. An example of a partial order for the time dimension based on the attributes day, week, month, quarter, and year is "day < {month <quarter; week} < year".2 This lattice structure is shown in Figure 3.6. A concept hierarchy that is a total or partial order among attributes in a database schema is called a schema hierarchy.

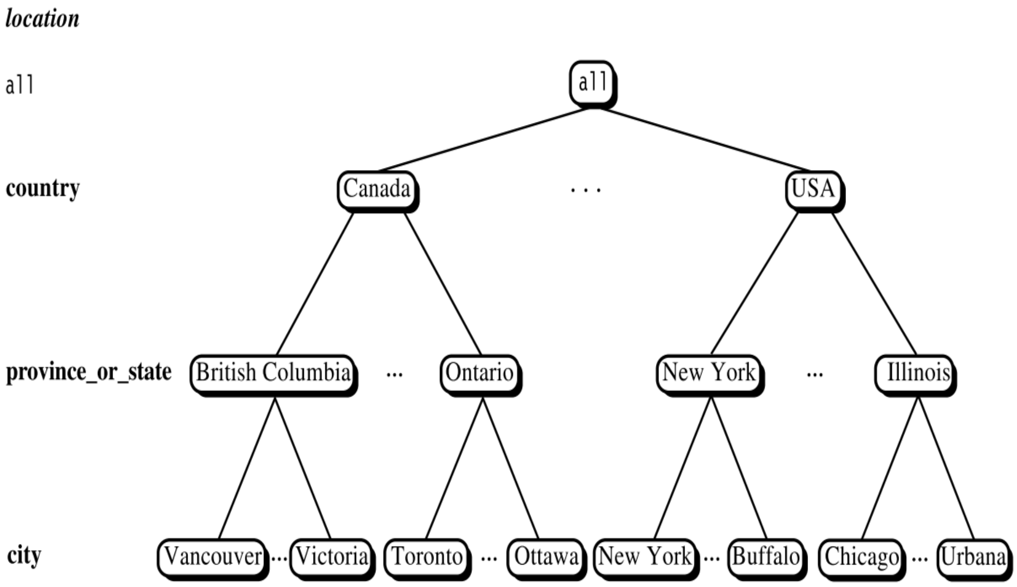


***Fig 3.5** A concept hierarchy for the dimension location. Due to space limitations, not all of the nodes of the hierarchy are shown (as indicated by the use of "ellipsis" between nodes).*

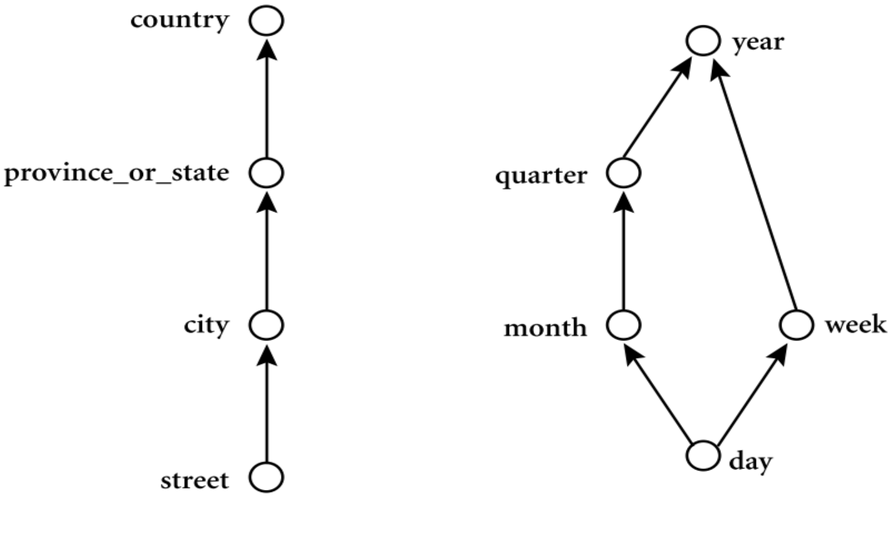


***Fig 3.6** Hierarchical and lattice structures of attributes in warehouse dimensions: (a) a hierarchy for location; (b) a lattice for time.*

Concept hierarchies that are common to many applications may be predefined in the data mining system, such as the concept hierarchy for time. Data mining systems should provide users with the flexibility to tailor predefined hierarchies according to their particular needs. For example, users may like to define a fiscal year starting on April 1 or an academic year starting on September 1.

Hierarchies

These are logical structures that uses ordered levels as a means of organising data. A hierarchy can be used to define data aggregation. For example, in a time dimension, a hierarchy might aggregate data from the month level to the quarter level to the year level: (all time), year quarter, month, day, or (all time), year quarter, week, and day. Also, a dimension may contain multiple hierarchies; a time dimension often contains both calendar and fiscal year hierarchies. Geography is seldom a dimension of its own; it is usually a hierarchy that imposes a structure on sales points, customers, or other geographically distributed dimensions. An example of geography hierarchy for sales points is: (all), country or region, sales-region, state or province, city, store. A hierarchy can also be used to define a navigational drill path and to establish a family structure.

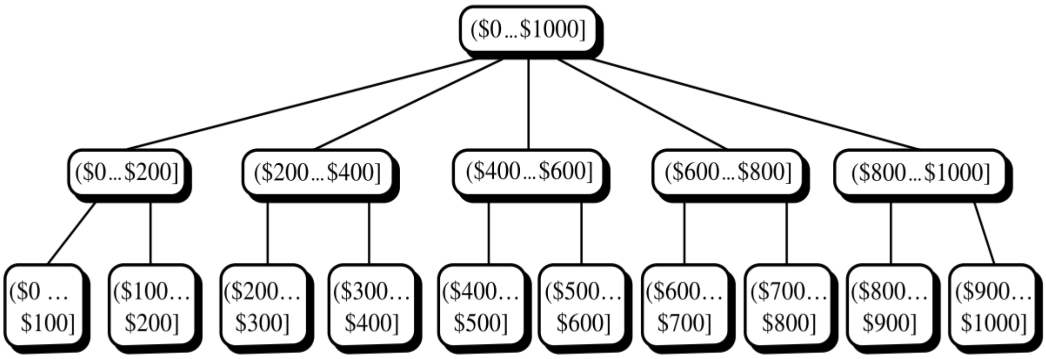


***Fig 3.7** A concept hierarchy for the attribute price.*

Concept hierarchies may also be defined by discretizing or grouping values for a given dimension or attribute, resulting in a set-grouping hierarchy. A total or partial order can be defined among groups ofvalues. An example of a set-grouping hierarchy is shown in Figure 3.7 for the dimension price, where an interval ($X . . .$Y] denotes the range from $X (exclusive) to $Y (inclusive).

There may be more than one concept hierarchy for a given attribute or dimension, based on different user viewpoints. For instance, a user may prefer to organize price by defining ranges for inexpensive, moderately priced, and expensive.

Concept hierarchies may be provided manually by system users, domain experts, or knowledge engineers, or may be automatically generated based on statistical analysis of the data distribution, Concept hierarchies also allow data to be handled at varying levels of abstraction

OLAP operations on multidimensional data

"How are concept hierarchies useful in OLAP?" 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 users with the

flexibility to view data from different perspectives. A number ofOLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. Hence, OLAP provides a user-friendly environment for interactive data analysis.

Example 3.8 *OLAP operations. Let's look at some typical OLAP operations for multidimensional data. Each of the operations described below is illustrated in Figure 3.8. At the center of the figure is a data cube for sales. The cube contains the dimensions location, time, and item, where location is aggregated with respect to city values, time is aggregated with respect to quarters, and item is aggregated with respect to item types. To aid in our explanation, we refer to this cube as the central cube. The measure displayed is dollars sold (in thousands). (For improved readability, only some of the cubes' cell values are shown.) The data examined are for the cities Chicago, New York, Toronto, and Vancouver.*

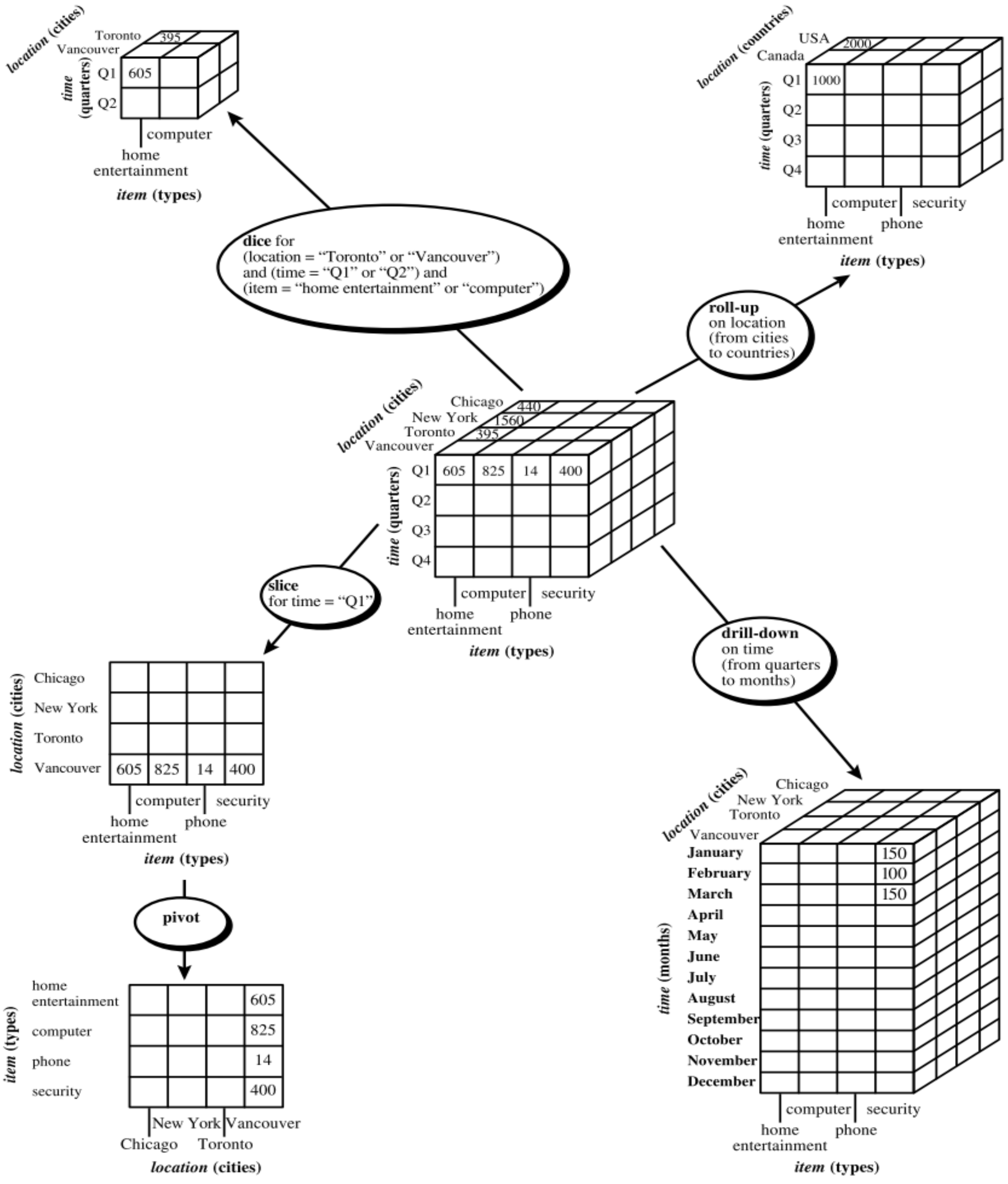


***Fig 3.8** Examples of typical OLAP operations on multidimensional data.*

***Roll-up**: The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. Figure 3.8 shows the result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for location given in Figure 3.7. This hierarchy was defined as the total order "street < city < province or state < country." The roll-up operation shown aggregates the data by ascending the location hierarchy from the level of city to the level of country. In other words, rather than grouping the data by city, the resulting cube groups the data by country.*
When roll-up is performed by dimension reduction, one or more dimensions are removed fromthe given cube. For example, consider a sales data cube containing only the two dimensions location and time. Roll-up may be performed by removing, say, the time dimension, resulting in an aggregation of the total sales by location, rather than by location and by time.

***Drill-down:** Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. Figure 3.10 shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for time defined as "day < month < quarter < year." Drill-down occurs by descending the time hierarchy from the level ofquarter to the more detailed level of month. The resulting data cube details the total sales per month rather than summarizing them by quarter.*
Because a drill-down adds more detail to the given data, it can also be performed by adding newdimensions to a cube. For example, a drill-down on the central cube of Fig 3.8 can occur by introducing an additional dimension, such as customer group.

***Slice and dice**: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. Figure 3.8 shows a slice operation where the sales data are selected from the central cube for the dimension time using the criterion time = "Q1". The dice operation defines a subcube by performing a selection on two or more dimensions. Figure 3.10 shows a dice operation on the central cube based on the following selection criteria that involve three dimensions: (location = "Toronto" or "Vancouver") and (time = "Q1" or "Q2") and (item = "home entertainment" or "computer").*

***Pivot (rotate):** Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data. Figure 3.8 shows a pivot operation where the item and location axes in a 2-D slice are rotated. Other examples include rotating the axes in a 3-D cube, or transforming a 3-D cube into a series of 2-D planes.*

***Other OLAP operations**: Some OLAP systems offer additional drilling operations. For example, drill-across executes queries involving (i.e., across) more than one fact table. The drill-through operation uses relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables.*

Other OLAP operations may include ranking the top N or bottom N items in lists, as well as computing moving averages, growth rates, interests, internal rates of return, depreciation, currency conversions, and statistical functions.

OLAP offers analytical modeling capabilities, including a calculation engine for deriving ratios, variance, and so on, and for computing measures across multiple dimensions. It can generate summarizations, aggregations, and hierarchies at each granularity level and at every dimension intersection. OLAP also supports functional models for forecasting, trend analysis, and statistical analysis. In this context, an OLAP engine is a powerful data analysis tool.

OLAP Systems versus Statistical Databases

Many of the characteristics of OLAP systems, such as the use of a multidimensional data model and concept hierarchies, the association of measures with dimensions, and the notions of roll-up and drill-down, also exist in earlier work on statistical databases (SDBs). A statistical database is a database system that is designed to support statistical applications. Similarities between the two types of systems are rarely discussed, mainly due to differences in terminology and application domains.

OLAP and SDB systems, however, have distinguishing differences. While SDBs tend to focus on socioeconomic applications, OLAP has been targeted for business applications. Privacy issues regarding concept hierarchies are a major concern for SDBs. For example, given summarized socioeconomic data, it is controversial to allow users to view the corresponding low-level data. Finally, unlike SDBs, OLAP systems are designed for handling huge amounts of data efficiently.

4.0 *SELF-ASSESSMENT EXERCISE(S)*

1. *List and briefly explain the phases involved in developing a data warehouse.*
2. *(a). Differentiate between a logical design and physical design (b).*
   Briefly describe the two types of objects commonly used
   in dimensional data warehouse:
   - *(i) Fact tables*
   - *(ii) Dimension tables.*
3. *Explain warehouse design methodologies*
4. *List and explain the process involved in developing a data warehouse*
   (a) Explain multidimensional data models
5. *Explain the different between OLAP systems versus statistical databases*

5.0 CONCLUSION

Therefore data design is the key to data warehousing. The business users know what data they need and how they want to use it. In designing a data warehouse there is need to focus on the users, determine what data is needed, locate sources of data and organize the data in a dimensional model that represents the business needs

6.0 SUMMARY

In this unit we have learnt that:

- *logical design involves describing the purpose of a system and what the system will do as against to how it is actually going to be implemented physically while physical design is the process of translating the abstract logical model into the* *specific technical design for the new system: there are three basic methodologies used in building a data warehouse, this include bottom-up design, top-down design and hybrid design*

- *The process of developing a data warehouse is made up of series of stages which are: Identify and gather requirements, design the dimensional model, develop the architecture, design the relational database and OLAP cubes, develop the maintenance applications, develop analysis applications, test and deploy the systemmthe implementation of data warehouse undergoes the natural cycle of unit testing, system testing, regression testing, integration testing and acceptance testing.*

8.0 REFERENCES/FURTHER READING

Jiawei Han., Micheline Kamber (2019). Data Mining: Concepts and Techniques, Second Edition

Yanchang Zhao, (2015). R and Data Mining Examples and Case Studies

Charu C. Aggarwal, 2015. The Textbook 2015th Edition

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

Shah, C. (2020). A Hands-On Introduction to Data Science. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Jayaprakash, P. et al. Accelerating Data mining Workloads: Current Approaches and Future Challenges in System Architecture Design.

Lean, A. &Lean, M. (1999). Fundamentals of Information Technology.
New Delhi: Leon Press Channel and Vikas Publishing House P.

Mosud, Y. O. (2009). Introduction to Data Mining and Data Warehousing, Lagos: Rashmoye Publications.

Pisharath, J. et al. Accelerating Data Mining Workloads: Current Approaches and Future Challenges in System Architecture.

UNIT 4 DATA WAREHOUSE AND OLAP TECHNOLOGY

CONTENTS

1.0 INTRODUCTION

Data warehousing and on-line analytical processing (OLAP) are essentials elements of decision-support that has become a focus of the database industry. Most of the commercial products and services are now available and all the principal database management system vendors now offer in these areas. Decision support places some rather different requirements on database compared to traditional on-line transaction processing application. This unit examines the differences between OLAP and data warehouse, types of OLAP servers and uses of OLAP.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- *state the meaning of OLAP*
- *differentiate between OLAP and data warehouse*
- *list the different types of OLAP server*
- *describe OLAP as a data warehouse tool and its applications*
- *identify the open issues in data warehouse.*

3.0 MAIN CONTENT

3.1 Meaning of On-Line Analytical Processing (OLAP)

The term On-Line Analytical Processing, OLAP (or Fast Analysis of Shared Multi-dimensional Information –FASMI) refers to the technology that allows users of multidimensional databases to generate on-line descriptive or comparative summaries (i.e. views) of data and other analytical queries.

OLAP was coined in 1993 by Tedd. Codd who is referred to as "the father of the relational database'' as a type of application that allows users to interactively analyse data. An OLAP system is often contrasted to an On-Line Transaction processing (OLTP) system that focuses on processing transaction such as orders, invoice or general ledger transactions. Before OLAP was coined, these systems were often referred to as Decision Support Systems (DSS).

OLAP is now acknowledged as a key technology for successful management in the 90's. It further describes a class of applications that require multidimensional analysis of business data. OLAP systems enable managers and analysts to rapidly and easily examine key performance data and perform powerful comparison and trend analyses, even on very large data volumes.

3.2 OLAP and Data Warehouse

It is important to distinguish the capabilities of a data warehouse from those of an OLAP system. A data warehouse is usually based on relational technology, while OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for further analysis.

OLAP enables analysts, managers and business executives to gain insight into data through fast, consistent and interactive access to a wide variety of possible views of information. Also, OLAP transform raw data so that it reflects the real dimensionality of the enterprise as understood by the user. In addition, OLAP systems have the ability to answer "what if?" and why?" that sets them apart from data warehouses. OLAP enables decision making about future actions. A typical OLAP calculation is more complex than simply summing data.

OLAP and data warehouse are complementary. A data warehouse stores and manages data. OLAP transform data warehouse data into strategic information. OLAP ranges from basic navigation and browsing (this is often referred to as "slice and dice"), to calculations, to more serious *analyses such as time series and complex modelling. As decision- makers exercise more advanced OLAP capabilities, they move from data access to information and to knowledge.*

3.2.1 Benefits of OLAP

Some of the benefits derived from the applications of OLAP systems are as follows:

3.6 *The main benefit of the OLAP is its steadiness in calculations. The reporting is always represented in a coherent presentation irrespective of how fast data is dealt with through the OLAP server or software and this allows the executives and analysts to know exactly to look for where. Other convenience of OLAP is that it allows the manager to tear down data from OLAP database in specific or broad terms. In layman's term, the report can be as simple as comparing two columns or as complex as analysing a huge amount of data. Moreover, it helps to realise relationships that were forgotten earlier.*

3.7 *OLAP helps to reduce the applications backlog still further by making business users self sufficient enough to build their own models. Unlike standalone departmental applications running on PC networks, OLAP applications are dependent on data warehouse and transaction processing systems to refresh their source level data. As a result, ICT gains more self-sufficient users without relinquishing control over the integrity of the data.*

3.8 *Through the use of OLAP, ICT realises more efficient operations by using software designed for OLAP, ICT reduces the query drag and network traffic on transaction systems or the data warehouse.*

3.9 *By providing the ability to model real business problems and a more efficient use of people resources, OLAP enables the organisation as a whole to respond more quickly to market demands. Market responsiveness, in turn often yields improved revenue and profitability.*

3.2.2 Features of OLTP and OLAP

The major distinguishing features between OLTP and OLAP are summarized as follows.

i. **Users and system orientation***: An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.*

ii. **Data contents***: An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for use in informed decision making.*

iii. ***Database design**: An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design*

iv. ***View**: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.*

v. ***Access patterns**: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations although many could be complex queries.*

3.3 OLAP and OLAP Server

On-Line Analytical Processing (OLAP) can further be described as a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent and interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

OLAP functionality is characterised by dynamic multi-dimensional analysis of consolidated enterprise data supporting end-user analytical and navigational activities including the following:

1. *Calculations and modelling applied across dimensions, through hierarchy and/or across member*
2. *Trend analysis over sequential time periods*
3. *Slicing subsets for on-screen viewing*
4. *Drill-down to deeper levels of consolidation*
5. *Reach-through to underlying detail data*
6. *Rotation to new dimensional comparisons in the viewing area.*

OLAP is implemented in a multi-user client/server mode and offers consistently rapid response to queries, regardless of database size and complexity. OLAP helps the user synthesise enterprise information through comparative, personalised viewing, as well as through analysis of historical and projected data in various "what-if" data model scenarios. This is achieved through the use of an OLAP server.

OLAP Server is a high-capacity, multi-user data manipulation engine specifically designed to support and operate on multi-dimensional data structures. A multi-dimensional structure is arranged so that every data item is located and accessed based on the interaction of the dimension members that defines the item. The design of the server and the structure of the data are optimised for rapid ad-hoc information retrieval in any orientation, as well as for fast, flexible calculation and transformation of raw data based on formulaic relationship. The OLAP server may either physically stage the processed multidimensional information to deliver consistent and rapid response times to end users, or it may populate its data structures in real-time from relational or other databases.

3.3.1 Types of OLAP Servers

OLAP systems vary quite a lot, and they have generally been distinguished by a letter tagged onto the front word OLAP, ROLAP, MOLAP and HOLAP. These three are the big players. Other types of OLAP are WOLAP, DOLAP, Mobile-OLAP and SOLAP

i. Relational OLAP (ROLAP) servers

These are the intermediate servers that stand in between a relational back-end server and client front-end tools. ROLAP systems work primarily from the data that resides in a relational database, where the base data and information tables are stored as relational tables. They use a relational or extended relational DBMS to store and manage warehouse data; and OLAP middleware to support missing piece. ROLAP severs include optimisation for each DBMS back end, implementation of aggregation navigation logic, and additional tolls and services. ROLAP technology tends to have greater scalability than MOLAP technology. The DSS server of micro-strategy and meta-cube of informix for example, adopt the ROLAP approach.

One major advantage of ROLAP over the other styles of OLAP analytical tools is that it is deemed to be more scalable in handling huge amounts of data. ROLAP sits on top of relational database therefore enabling it to leverage several functionalities that a relational database is capable of. Another benefit of ROLAP tool is that it is efficient in managing both numeric and textual data. It also permits users to "drill down" to the leaf details or the lowest level of a hierarchy structure. The disadvantage of ROLAP applications is that it displays a slower performance as compared to other style of OLAP tools, since calculations are often times performed inside the server. Another disadvantage of ROLAP tool is that it is dependent on use of SQL for data manipulation, it may not be ideal for performance of some calculations that are not easily translatable into a SQL query.

ii. Multidimensional OLAP (MOLAP)

Multidimensional OLAP with population acronym of MOLAP is widely regarded as the classic form of OLAP. The servers support multidimensional views of data through array-based multidimensional storage engines. They map multidimensional views directly to data cube array structures. This is probably by far the best OLAP tool to use in making analysis *reports since this enable user to easily recognise or rotate the cube structure to view different aspects of data. This is done by way of slicing and dicing.*

One of the major distinctions of MOLAP against a ROLAP tool is that data are pre-summarised and are stored in an optimised format in a multidimensional cube, instead of in a relational database. In this type of model, data are structured into proprietary formats in accordance with a client's reporting requirements with the calculations pre-generated on the cubes. MOLAP analytic tool are capable of performing complex calculations, since calculations are predefined upon cube creation, this results in the faster return of computed data. MOLAP systems also provide users with the ability to quickly write back data into a data set. Moreover when compared with ROLAP, MOLAP is considerably less heavy on hardware due to compression techniques. Summarily, MOLAP is more optimised for fast query performance and retrieval of summarised information.

However, there are certain limitations to the implementation of a MOLAP system; one primary weakness is that MOLAP tool is less scalable than a ROLAP tool as the former is capable of handling only a limited amount of data. Also, MOLAP approach introduces data redundancy. Certain MOLAP products encounters difficulty in updating models with dimensions of very high cardinality.

iii. Hybrid OLAP (HOLAP)

HOLAP is the product of the attempt to incorporate the best features of MOLAP and ROLAP into a single technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP. HOLAP tool bridges the technology gap of both products by enabling access or use to both multidimensional database (MDDB) and Relational Database Management System (RDBMS) data stores. HOLAP also has the capacity to "drill through" from table for delineated data. For example, a HOLAP server may allow large volume of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store or in the pre-calculated cubes. Some of the advantages of HOLAP are better scalability, quick data processing and flexibility in accessing data sources.

Other Types

There are also less popular types of OLAP system upon which could stumble on so often. We have listed some of the less famous existing in the OLAP industry.

iv. Web OLAP (WOLAP)

A Web OLAP is also referred to as Web-enabled OLAP; it pertains to OLAP application that is accessible via the web browser. Unlike traditional client/server OLAP applications, WOLAP is considered to have a three-tiered architecture which consists of three components: a client, a middleware and a database server. Some of the most appealing features of the style of OLAP are the considerably lower investment involved, enhanced accessibility as user only needs an internet connection and a web browser to connect to the data and ease of installation, configuration and deployment process

.But despite all of its unique features, it could still not compare to a conventional client/server machine. Currently, it is inferior in comparison with OLAP applications which involve deployment in client machines in terms of functionality, visual appeal and performance.

v. Desktop OLAP (DOLAP)

Desktop OLAP or "DOLAP" is based on the idea that a user can download a section of the data from the database or source, and work with that dataset locally, or on their desktop. DOLAP is easier to deploy and has a cheaper cost but comes with a very limited functionality in comparison with other OLAP applications.

vi. Mobile OLAP (MOLAP)

Mobile OLAP merely refers to OLAP functionalities on a wireless or mobile device. This enables users to access and work on OLAP data and applications remotely through the use of their mobile devices.

vii. Spatial OLAP (SOLAP)

With the aim of integrating the capabilities of both geographic information systems (GIS) and OLAP into a single user interface, "SOLAP" or Spatial OLAP emerged. SOLAP is created to facilitate management of both spatial and non-spatial data, as data could come not only in an alphanumeric form, but also in images and videos. This technology provides easy and quick exploration of data that resides on a spatial database

Other different blends of an OLAP product like the less popular 'DOLAP' and 'ROLAP' that stands for Database OLAP and Remote OLAP respectively. LOLAP for Local OLAP and RTOLAP for Real Time OLAP are existing but have barely made a noise on the OLAP industry.

3.3.2 From on-line analytical processing to on-line analytical mining

On-Line Analytical Mining (OLAM) (also called OLAP mining), which integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multidimensional databases, is particularly important for the following reasons.

- ***High quality of data in data warehouses.*** *Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data transformation and data integration as preprocessing steps. A data warehouse constructed by such preprocessing serves as a valuable source of high quality data for OLAP as well as for data mining.*

- ***Available information processing infrastructure surrounding data warehouses****.* *Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple, heterogeneous databases ODBC/OLEDB connections, Web-accessing and service facilities, reporting and OLAP analysis tools.*

- ***OLAP-based exploratory data analysis***. *Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms. On-line analytical mining provides facilities for data mining on different subsets of data and at different levels of abstraction, by drilling, pivoting, filtering, dicing and slicing on a data cube and on some intermediate data mining results*

- ***On-line selection of data mining functions.*** *By integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the exibility to select desired data mining functions and swap data mining tasks dynamically.*

3.3.3 Features of OLTP and OLAP

The major distinguishing features between OLTP and OLAP are summarized as follows

1. ***Users and system orientation****: An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.*
2. ***Data contents****: An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for use in informed decision making.*
3. ***Database design****: An OLTP system usually adopts an entity-relationship (ER) data model and an application oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design.*
4. ***View****: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.*
5. ***Access patterns****: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations although many could be complex queries.*

3.3.4 Comparison between OLTP and OLAP systems

Feature	***OLTP***	***OLAP***
Characteristic	*Operational processing*	*Informational processing*
Orientation	*Transaction*	*Analysis*
User	*Clerk, DBA, database professional*	*Knowledge worker (E.g., manager, analyst)*
Function	*Day-to-day operations*	*Long term informational requirements, decision support*
DB design	*E-R based, application-oriented*	*Star/snowflake, subject-oriented*
Data	*Current; guaranteed up-to-date*	*Historical; accuracy maintained overtime*
Summarization	*Primitive, highly detailed*	*Summarized, consolidated*
View	*Detailed, flat rationale*	*Summarized, multidimensional*
Unit of work	*Short, simple transaction*	*Complex query*
Access	*Read/write*	*Mostly read*
Focus	*Data in*	*Information out*
Operations	*Index/hash on primary key*	*Lots of scans*
No of records accessed	*Tens*	*Millions*
No of users	*Thousands*	*Hundreds*
DB size	*100MB to GB*	*100GB to TB*
Priority	*High performance, high availability*	*High flexibility, end user autonomy*
Metric	*Transaction throughput*	*Query throughput, response time*

3.4 Architecture for on-line analytical mining

An OLAM engine performs analytical mining in data cubes in a similar manner as an OLAP engine performs on-line analytical processing. An integrated OLAM and OLAP architecture is shown in Figure 4.1, where the OLAM and OLAP engines both accept users' on-line queries via a User GUI API and work with the data cube in the data analysis via a Cube API.

A metadata directory is used to guide the access of the data cube. The data cube can be constructed by accessing and/or integrating multiple databases and/or by filtering a data warehouse via a Database API which may support OLEDB or ODBC connections. Since an OLAM engine

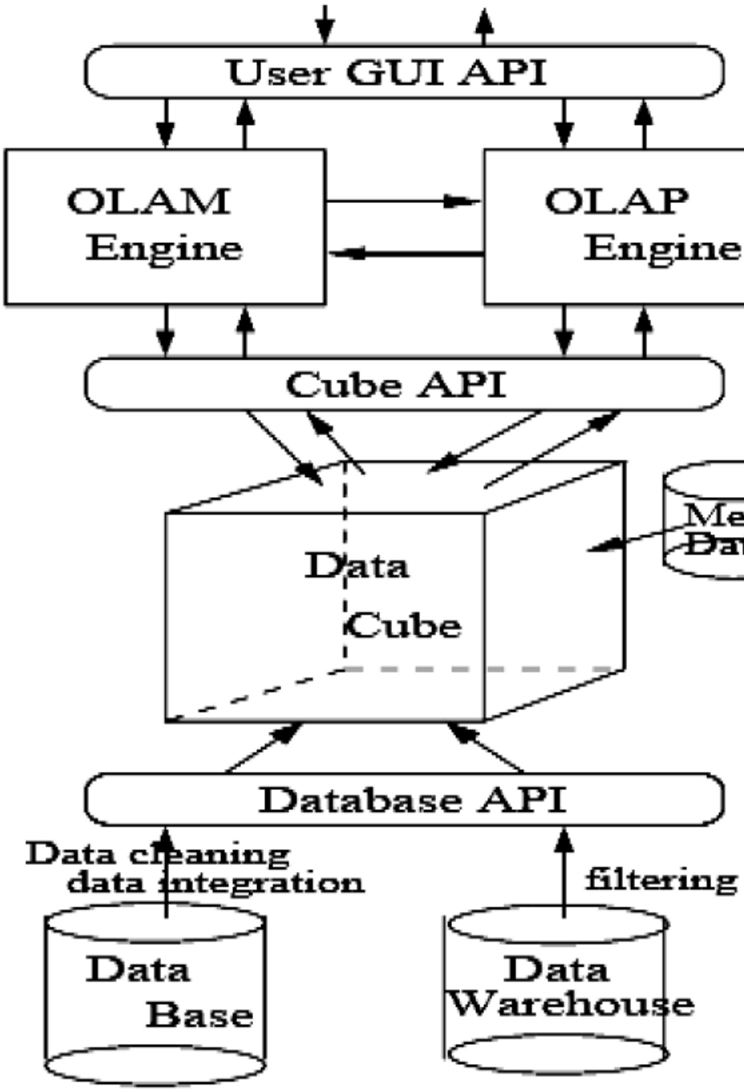


Fig 4.1 An integrated OLAM and OLAP architecture

4.0 SELF -ASSESSMENT EXERCISE

a. Differentiate between the following pairs:

i. OLAP and data warehouse
ii. OLAP and OLAP server
iii. Describe the component of on-line analytical mining Architecture
iv. List the Features of OLTP and OLAP

b. State some of the benefits derived from the applications of OLAP systems.

TUTOR- MARKED ASSIGNMENT

i. (a) What do you understand by the term OLAP?
(b) List and explain the three major types of OLAP.
ii. State some application areas of OLAP.

5.0 CONCLUSION

Therefore, data warehousing and on-line analytical processing (OLAP) are essentials elements of decision-support that has become a focus of the database industry.

6.0 SUMMARY

In this unit we have learnt that:

i. OLAP is a technology that allows users of multidimensional databases to generate on-line descriptive or comparative summaries of data and other analytical queries.
ii. data warehouse is different from OLAP in a number of ways such as data warehouse stores and manages data while OLAP transform data warehouse data into strategic information
iii. there are different types of OLAP which are ROLAP, MOLAP and HOLAP. These three are the big players. Other types of OLAP are WOLAP, DOLAP, Mobile-OLAP and SOLAP
iv. OLAP as a data warehouse tool can be used to provide superior performance for business intelligence queries and to operate efficiently with data organised in accordance with the common dimensional model used in data warehouse.
v. some of the open issues in data warehousing, these include increased research activity in the near future as warehouse and data mart proliferation.

7.0 REFERENCES/FURTHER READING

Aggarwal, C. C. (2015). Data Mining: The Textbook. Cham: Springer. ISBN: 978-3-319-14141-1

Zaki, M., & Meira, Jr, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511810114

Connolly, A., VanderPlas, J., & Gray, A. (2014). Fast Computation on Massive Data Sets. In Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data (pp. 43-66). PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctt4cgbdj.5

Connolly, A., VanderPlas, J., & Gray, A. (2014). Classification. In Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data (pp. 365-402). PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctt4cgbdj.13

Jiawei Han., Micheline Kamber (2019). Data Mining: Concepts and Techniques, Second Edition

Yanchang Zhao, (2015). R and Data Mining Examples and Case Studies

Charu C. Aggarwal, 2015. The Textbook 2015th Edition

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

Shah, C. (2020). A Hands-On Introduction to Data Science. Cambridge: Cambridge University Press. doi:10.1017/9781108560412

Anil, R. *Data Warehouse and its Applications in Agriculture, Indian Agriculture Statistics Research Institute Library Avenue, New Delhi:-110 012.*

Data Management and Data Warehouse Domain Technical Architecture, June 6, 2002

http://www.users.qwest.net/¬lauramh/resume/thorn.htm.

Jayaprakash, *P.et al Accelerating Data mining Workloads: Current Approaches and Future Challenges in System Architecture Design*