Mt Kenya University

P.O. Box 342-01000 Thika
Email: Info@mku.ac.ke
Web: www.mku.ac.ke

COURSE CODE: BIT 4104

COURSE TITLE: Data Mining and Data Warehousing

Study Units

The study units in this course are as follows:

Module 1: Concepts of Data Mining

Unit 1: Overview of Data Mining

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 pattern s 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
- Differentiate between data mining and knowledge discovery in databases (KDD)
- Understand the difference between data mining and OLAP
- Understand the evolution of data mining
- Know the scope of data mining
- Understand the architecture of data mining
- Understand how data mining works

3.0 Definition of Data Mining

Data mining is an analytical process designed for extracting or exploring hidden and predictive information from large databases which may be business or market related. It can also be described as the process of searching for valuable information in large volumes of data. Data mining is relatively a powerful new technology with great potential to assist companies focus on the most important information in their data warehouses.

Data mining is a cooperative effort of humans and computers; human actually designs the databases, describe the problems and set goals while computers sort through the data and search for patterns that matches the goals.

3.1 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 in1989 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.

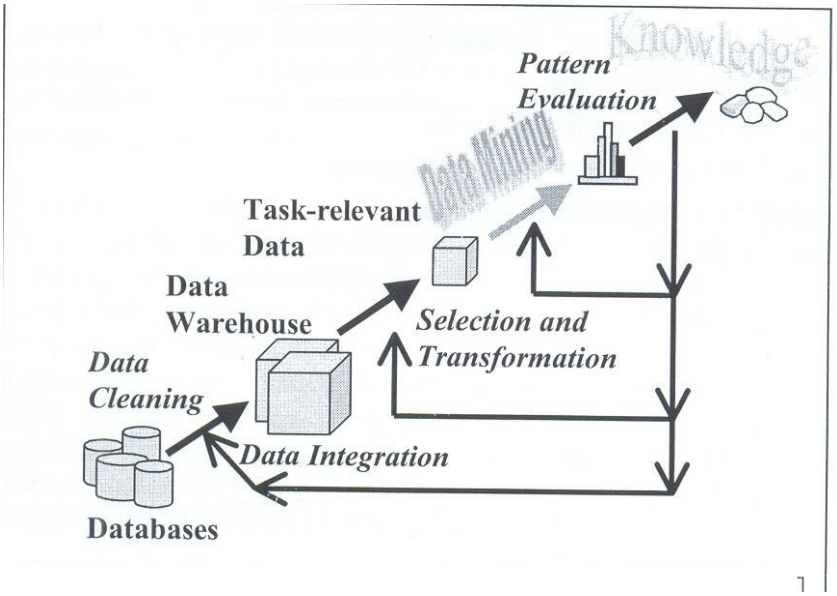


Figure 1.1 Data Mining as One of Knowledge Discovery Process
Source: Principles of Knowledge Discovery in Databases by O. R. Zaiane page 3

KDD being an iterative process consists of the following steps:

- **Data Cleaning:** This is also referred to as data cleansing and is a phase in which noise data and irrelevant data are removed from the collection.

- **Data Integration:** In this phase, multiple data sources which are often heterogeneous may be combined to a common source.

- **Data Selection:** The data that is relevant to the analysis is decided upon and retrieved from the data collection at this stage.

- **Data Transformation:** This is also referred to as data consolidation and is a stage where selected data are transformed into forms that are appropriate for the data mining procedure.

- **Data Mining:** This is an important step in knowledge Discovery in Databases in which clever techniques are applied for the extraction of patterns that are potentially useful.
- **Pattern Evaluation:** At this stage, patterns that are very interesting and represent knowledge are identified based on given measures.
- **Knowledge Representation:** This is the final stage of the KDD process in which the discovered knowledge is visually represented to the user. Visualization techniques are used to assist the users to have a better understanding and interpret the data mining results.

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

3.2 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 variable 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.

3.3 The Evolution of Data Mining

Data mining techniques are the results of a long process of research and product development. The evolution started when business data was first stored on computers with data access improvements and generated technologies that allow users to navigate through their data in real time. This evolutionary process is taken beyond retrospective data access and navigation to prospective and proactive information delivery.

Data mining is a natural development of the increased use of computerized databases to store data and provide answers to business analysis. Traditional query and report tools have been used to describe and extract what is in a database. Data mining is ready for application in the business community because it is supported by these technologies that are now sufficiently matured:

- Massive data collection
- Powerful multiprocessor computer

Presently commercial databases are growing at an unprecedented rate. In some organizations, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost- effective with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have been existing for at least ten years, but have recently been implemented as nature, reliable, understandable tools that consistently outperform older statistical methods.

In the evolution data mining from business data to business information, each new step has built upon the previous one. For example the four steps listed in table 1.1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

Table 1.1: Steps in the Evolution of Data Mining

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	What was my total revenue in the last five years?	Computers, tapes, disk	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	What were unit sales in New England last March?	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data	What were	On line analytic	Pilot,	Retrospective,

Warehousing & Decision Support	unit sales in New England last March? Drill down to Boston.	processing (OLAP), multidimensional databases, data warehouses	Comshare, Arbor, Cognos, Microstrategy	dynamic data delivery at multiple levels
Data Mining (Emerging Today)	What s likely to happen to Boston unit sales next month? Why	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM,SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

(Source: An Introduction to Data Mining, Retrieved From: http:www.thearling.com/text/dmwhite/dmwhite.htm)

The core components of data mining technology have been under development for decades in research areas such as statistic, artificial intelligence and machine learning Today the maturity of these techniques coupled with high performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

Activity A: Student Self Assessment Exercise

1. What is data mining?
2. List all the steps involved in KDD process

3.4 Scope of Data Mining

Data mining derived its name from the similarities between searching for valuable business information in a large database. For example, to search for linked products in gigabytes of stored scanner data and mining a mountain for a vein of valuable ore; the two processes require either sifting through an immense amount of material or intelligently probing it to find exactly where the value resides.

If the database is given a sufficient size and quality, data mining technology can generate new business opportunities by providing the following capabilities:

- **Automated Prediction of trends and Behaviors**: Data mining automates the process of searching for predictive information in large databases. Questions that may traditionally require extensive hands-on analysis can now be answered directly from data very quickly. An example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the most likely target to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- **Automated Discovery of Previously Unknown Patterns** : Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive database in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases in turn yield improved predictions.

3.5 Architecture for Data Mining

In order to best apply this mining technique, it must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Most data mining tools presently operate outside of the warehouse, requiring extra steps for extracting, importing and analyzing data. Moreover, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, and new product rollout and so on. Figure 1.2 shows architecture for advanced analysis in a large data warehouse.

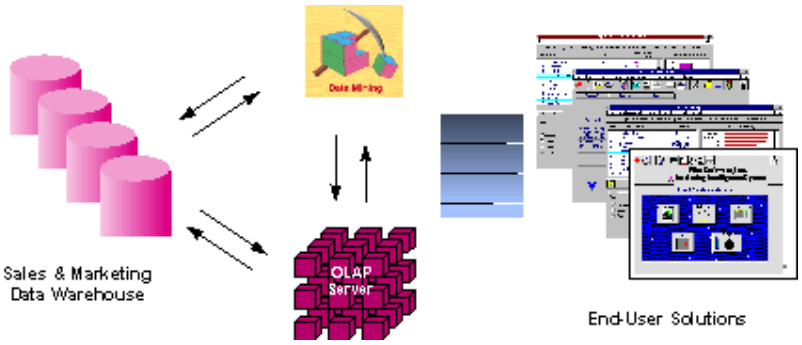


Figure 1.2 Integrated Data Mining Architecture
Source: *An introduction to Data Mining* ***page 5***

The ideal starting point is a data warehouse that contains a combination of internal data tracking all customers contact coupled with external market about competitor s activity. The background information on potential customers also provides an excellent basis for prospecting. The warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick and so on and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business- summarizing by product line, region, and other perspectives of their business. The data mining server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directing into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting and promotion optimization. Integration with data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse continues to grow with

new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

This design represents a fundamental shift from conventional decision support system. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP server by providing a dynamic metadata layer that represents a distilled view of a data. Other analysis tools can then be applied to plan future actions and confirm the impact of those plans (An Introduction to Data Mining)

3.6 How Data Mining Works

How does data mining tells us important things that we do not know or what is going to happen next? The technique used in performing these feats is called **modeling**. *Modeling can simply be defined as an act of building a model based on data from situations where you know the answer and then applying it to another situation where the answer is not known* .

The very act of model building has been around for centuries even before the advent of computers or data mining, technology. What happens in computers does not differ much from the way people build models. Computers are loaded with lots of information about different situations where answer is known and then the data mining software on the computer must run through that data and distill the characteristic of the data that should go into the model. And once the model is built it can be applied to similar situations where you do not know the answer.

For example, as the marketing director of a telecommunication company you have access to a lot of information such as age, sex, credit history, income, zip code, occupation and so on of all your customers; but difficult to discern the common characteristics of his best customers because there are so many variables. From the existing database of customers that contains their information as earlier mentioned; data mining tools such as neural networks can be used to identify the characteristics of those customers that make a lot of long distance calls. This then becomes the director s model for high-value customers, and he would budget his marketing efforts accordingly

Activity B: Student Self Assessment Exercise

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

(i) Automated prediction of trends and behaviours

(ii) Automated discovery of previously unknown patterns

4.0 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.

5.0 Summary

In this unit we have learnt that:

- Data mining is the process of extracting hidden and predictive information from large databases or data warehouse.

- 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 behavious, 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

6.0 Tutor Marked Assignment

1. (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.

(2) (a) Differentiate between data mining and OLAP
   (b) Briefly explain how does mining methodology works?

7.0 Further Reading and Other Resources

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

Osmar R Zaiane (1999)- *Principles of Knowledge Discovery in Databases*

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

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

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

Introduction to Data Mining and Knowledge Discovery, Third Edition.

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

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

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

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House Pvt Ltd.

Module 1: Concepts of Data Mining

Unit 2: Data Description for Data Mining

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:

- Know 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 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.

(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.

(2). Personal and Medical Data

From personal data to medical and government, very large amounts of information are continuously collected. Governments, individuals and organizations 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). 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.

(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.

(5). 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.

(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 digitize them for future use and analysis.

(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 satellite 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 analyze them.

(8). Text Reports and Memos (E-mail Messages)

Most of communications within and between individuals, research organizations and companies based on reports and memos in textual forms are often exchanged by e-mail. These messages are frequently stored in digital form for future use and references which creates digital library.

(9). 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.

3.1.1 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.

3.1.2 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.

3.13 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 a able to incorporate or at least

reference all management systems using designated technology suitable for corporate database management e.g. Sybase, Ms SQL Server.

3.14 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,

Figure 2.1 Fragment of a Transaction Database for Rentals in a Store

Figure 1.1 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 normalized 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.

3.1.5 Spatial Databases

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

3.1.6 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.

3.17 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.

3.1.8 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.

Activity A: Student Self Assessment Exercise

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

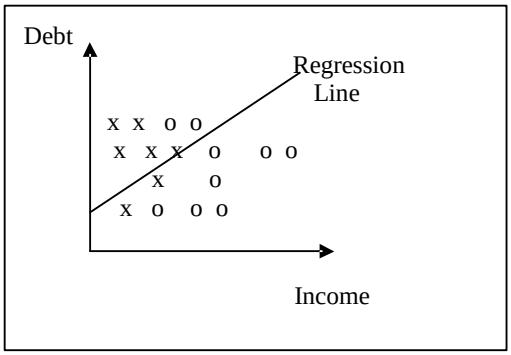
3.2 Data Mining Functionalities

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:

3.2.1 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

Figure 2.2 A Simple Linear Classification Boundary for the Loan Data Set
Source: From Data Mining to Knowledge Discovery in Databases, Ussama, F. et al page 44

3.2.2 Characterization

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)

3.2.3 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.

3.2.4 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.

3.2.5 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 characterize 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 characterization with the exception that data discrimination results include comparative measures.

3.2.6 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.*

3.2.7 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.

3.2.8 Evolution and 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.

Activity B: Student Self Assessment Exercise

What do you understand by the following data mining terms?

(a). Classification
(b). Characterization
(c). Discrimination

4.0 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.

5.0 Summary

In this unit we have learnt that:

- Different kind of information are often collected in digital form in our databases and flat files, these include scientific data, personal and medical data, games,
- 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.

6.0 Tutor Marked Assignment

1. Briefly explain the following types of data that can be mined:
   (i). Flat files (ii). Relational Databases
   (iii). Transaction Databases (iv). Spatial Databases
   (v). Multimedia Databases (vI). World Wide Web

2. List and explain any five data mining functionalities and the variety of knowledge they discover.

7.0 Further Reading and Other Resources

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

Osmar R Zaiane (1999)- *Principles of Knowledge Discovery in Databases*

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

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

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

Introduction to Data Mining and Knowledge Discovery, Third Edition.

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

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

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

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House Pvt Ltd.

Module 1: Concepts of Data Mining

Unit 3: Classification of Data Mining

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:

- Understand the various classifications of data mining systems
- Describe the categories of data mining tasks
- Understand the diverse issues coming up in data mining
- Describe the various challenges facing data mining

3.0 Classification of Data Mining Systems

There are several data mining systems available while some are being developed. Data mining systems can be categorized according to various criteria among other classification are the following:

1. **Classification By The Type Of Data Source Mined** : This classification categories data mining systems according to the type of data handled such as spatial data, multimedia data, time series data, text data, World Wide Web etc

2. **Classification By the Data Model Drawn On:** This class categorizes data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional etc.

3. **Classification By the Kind of Knowledge Discovered:** This classification categorizes data mining systems according to the kind of knowledge discovered or data mining functionalities such as discrimination, characterization, association, clustering etc. Some systems tend be comprehensive systems offering several data mining functionalities together.

4. **Classification By The Mining Techniques Used:** Data mining systems employ and provide different techniques. This class categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database-oriented or data warehouse-oriented. This class also takes into account the degree of user interaction involved in the data mining process such as query-driven systems interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety to data mining technique to fit different situation and options, and offer different degrees of user interaction.

3.1 Data Mining Task

Data mining commonly involves four classes of task:

1. Classification

In this task data will be arranged into predefined groups in terms of attributes, one of which is the class. It will find a model for class attribute as a function of the values of other (predictor) attributes, such that previously unseen records can be assigned a class as accurate as possible. For instance an e-mail program might attempt to classify an e-mail as legitimate or spam. Common algorithms to use are nearest neighbor, Naives Bayes classifier and Neural Network

2. Clustering

Clustering is similar to classification but the groups are not predefined, so the algorithms will try to group similar items together.

3. Regression

This task attempts is similar to find a function which models the data with the least error. A common method is to use Genetic Programming

4. Association Rule Learning

This searches for relationships between variables. For instance, a superstore might gather data of what each customer buys using association rule learning, the superstore can work out what products are frequently bought together that is useful for marketing purposes. This is sometimes called market basket purposes analysis .

Activity A: Student Self Assessment Exercise

Briefly describe the four data mining tasks list below:

(i). Classification (ii). Clustering

(iii). Regression (iv). Association rule learning

3.2 Data Mining Issues

Data mining is still in its infancy, although it is rapidly becoming a trend and ubiquitous, it is lately being applied as reliable and scalable tools that outperform older classical statistical methods. Before the study of data mining develops into a conventional, mature and trusted discipline, several issues have to be addressed; some of these issues are discussed in this section, though not exclusive and are not ordered in any way.

3.2.1 Security and Social Issues

Security is a very social issue in data collection either to be shared or intended to be used for strategic decision making. Also, when data are collected from customers profile, user behaviour understanding, students profile, or correlating personal data with other information, huge amounts of sensitive and private information about individuals or companies are collected and stored. Considering the confidential nature of some the data gathered and the potential illegal access to the information it makes the security issue to be very controversial. In addition data mining may disclose new implicit knowledge about individuals or groups that could violate their private policies, especially if there is potential dissemination of the discovered information. Another important that arises from this concern is the appropriate use data mining. As a result of the value of data, database of all kinds of content are regularly sold and because of the advantage that can be attained from implicit

knowledge discovered, some vital information could be withheld, which other may be widely distributed and used without control.

3.2.2 Data Quality

Data quality refers to the accuracy and completeness of a data. It is a multifaceted issue that represents one of the biggest challenges for data mining. The quality of data can be affected by the structured and consistency of the data being analyzed. The presence of duplicate records, lack of data standards, timeliness of updates and human error can significantly impact the effectiveness of more complex data mining techniques that are sensitive to subtle differences that may exist in the data.

To improve quality of data, it is somehow necessary to clean the data, which may involve the removal of duplicate records, normalizing the values used to represent information in the database (for example, ensuring that no is represented as a 0 throughout the database, and not sometimes as a O, and sometimes N), accounting for missing data points, removing unrequired data fields, identifying anomalous data points (e.g. an individual whose age is shown as 215 years), and standardizing data formats (e.g. changing dates so they include DD/MM/YYYY).

3.2.3 User Interface Issues

The knowledge that is discovered by data mining techniques remains useful as long as it is interesting and understandable by the user. Good data visualization eases the interpretation of data mining results and helps users to have better understanding of their needs. A lot of data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation However, there is still a lot research needed to accomplish a good visualization tools for large datasets that could be used to display and manipulate mined knowledge.

The major issues related to user interface and visualizations are screen real-estate information rendering, and interaction. The interactivity of data and data mining results is very vital since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

3.2.4 Data Source Issues

There are lots of issues related to the data sources; some are practical such as diversity of data types, while others are philosophical like the data excess problem. It is obvious we have an excess of data since we have more data than we can handle and we are still collecting data at an even higher rate. If the spread database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The present practice is to collect as much data as possible now and process it, or try to process it. Our concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether differentiate between what data is important and what data is insignificant.

Regarding the issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We store different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining result on all kinds of data and sources. Different types of data and sources may require distinct algorithms and methodologies. Presently, there is a focus on

relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. Therefore the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.

3.2.5 Performance Issues

A lot of artificial intelligence and statistical methods exist for data analysis and interpretation; though the methods were not actually designed for the very large data set (i.e. terabytes) data mining is dealing with these days. This has raised the issues of scalability and efficiency of the data mining methods when processing large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining; instead linear algorithms are usually the standard. Also, sampling can be used for mining instead of the whole dataset.

Other topic that needs to be considered in performance issues includes completeness and choice of samples, incremental updating and parallel programming. Although parallelism can help solve the size problem if the dataset can be subdivided and the results merged later. And incremental updating is very important for merging results from parallel mining, or updating data mining results when new data because available without necessarily re-analyzing the complete dataset.

3.2.6 Interoperability

Data quality is related to the issue of interoperability of different databases and data mining software. *Interoperability refers to the ability of a computer system and data to work with other systems or data using common standards or processes* . It is a very critical part of the larger efforts to improve or enhance interagency collaboration and information sharing through government and homeland security initiatives. In data mining, interoperability of databases and software is important to enable the search and analysis of multiple databases simultaneously and to help ensure the compatibility of data mining activities of different agencies.

The data mining projects that want to take the advantage of existing legacy databases or trying to initiate first-time collaborative efforts with other agencies or levels of government such as police may experience interoperability problems. Also, as agencies advances with the creation of new databases and information sharing efforts, they will need to address interoperability issues during their planning stages to better ensure the effectiveness of their data mining projects.

3.2.7 Mining Methodology Issues

These issues relates to the different data mining approaches applied and their limitations. Issues such as versatility of the mining approaches, diversity of data available, dimensionality of the domain, the assessment of the knowledge discovered; the exploitation of background knowledge and metadata, the control and handling of noise in data etc. are all examples that would dictate mining methodology choices. For example it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand.

More so, different approaches may suit and solve user s needs differently. Most algorithms used in data mining assume the data to be noise-free, which of course is a strong assumption. Most of the datasets contain exceptions, invalid or incomplete information which may

probably complicate, if not obscure the analysis process and in many cases compromise the accuracy of the results. Consequentially, data preprocessing (i.e. data cleaning and transformation) becomes very essential. Data cleaning preprocessing is often seen as time consuming and frustrating but is one of the most important phase in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information.

3.2.8 Mission Creep

Mission creep refers to the use of data for purpose other than that for which the data was originally collected. Mission creep is one of the highest risks of data mining as cited by civil libertarians, and represents how control over one s information can be a fragile proposition. This can occur regardless of whether the data was provided voluntarily by the individual or was collected through other means. In fighting terrorism, this take on an acute sense of urgency, because it create pressure on both data holders and official that accesses the data. To abound an available resources unused may appear to be negligent and data holders may holders may feel obligated to make any information available that could be used to prevent a future attack or track a known terrorist.

Similarly, government officials that are responsible for ensuring the safety of others may be pressurized to use or combine existing databases to identify potential threats. Unlike physical searches, or the detention of individuals, accessing information for purposes other than originally intended may appear to be a victimless or harmless exercise. However, such information use can lead to unintended outcome and produce misleading results

One of the primary reasons for misleading results is inaccurate data. All data collection efforts suffer accuracy concerns to some degree. Ensuring the accuracy of information can require costly protocols that may not be cast effective if the data is not of inherently high economic value (Jeffrey W. Seifert, 2004).

3.2.9 Privacy

Privacy focuses on both actual projects proposed as well as concerns about the potential for data mining applications to be expanded beyond their original purposes (mission creep). As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy. For instance, some experts suggest that anti-terrorism data mining applications might be for combating other types of crime as well.

So far there has been little consensus about how data mining should be carried out with several competing points of view being debated. Some observers suggest that existing laws and regulations regarding privacy protections are adequate and that these initiatives do not pose any threats to privacy. Others argue that not enough is known about how data mining projects will be carried out and that greater oversight is needed. As data mining efforts continues to advance, Congress may consider a variety of questions including, the degree to which government data, whether data sources are being used for purpose other than those for which they were originally designed, and the possible application of the privacy Act to these initiatives (Jeffrey W. Seifert, 2004)

3.3 Data Mining Challenges

In this section we shall be describing some of the current research and application challenges for data mining. These challenges are by no means exhaustive but to acquaint the students with the types of problems facing data mining practitioner.

(1). Larger Databases

Databases with hundreds of fields and tables containing millions of records of a multi-gigabyte size are very prevalent, and terabyte (10^{12} bytes) databases are also becoming common. The methods for dealing with large data volume include more efficient algorithms, sampling, and approximation and massively parallel processing.

(2). High Dimensionality

At times there might be no large number of records in the database, but there can be a large number of fields (attributes, variables); so the dimensionality of the problem becomes high. A high-dimensional data set creates problems in terms of increasing the size of the search space for model induction in a combinatorial explosive manner (Usama et at, 1996). Also, it increases the chances of data mining algorithm finding spurious that may not be valid in general. Approaches to this challenge include the methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identity irrelevant variables.

(3). Missing and Noisy Data

This is a very serious challenge especially in business databases. Some important attributes can be missing if the database is not designed with discovery in mind. Possible solutions include the use of more sophisticated statistical strategies to identity hidden variables and dependencies.

(4). Complex Relationship between Fields

Hierarchically structured attributes or values, relations between attributes, and more sophisticated means for representing knowledge about the about of a database will require algorithms that can effectively use such information. Historially, data mining algorithms have been developed for simple attribute-value records, although new technique for deriving relations between variables are being developed.(Usama 1996;Dzeroski 1996;Djoko,Cook,and Holder 1995

(v). Understanding of the Patterns

In many applications, it is very important to make the discovered patterns more understandable to humans. The possible solutions include graphic representation, rule structuring, natural language generation and technique for visualization of data and knowledge. Rule-refinement strategies can be used to address a related problem. This discovered knowledge might be implicitly or explicitly redundant.

(vi). Over Fitting

When the algorithm searches for the best parameters or one particular model using a limited set of data, it can model not only the general patterns in the data but also any noise specific to the data set resulting in poor performance of the model on test data. Possible solutions include cross-validation, regularization, and other sophisticated statistical strategies (Usama et al, 1996)

(vii). User Interaction and Prior Knowledge

Most of the data mining techniques are not truly interactive and cannot easily incorporate prior knowledge about a problem except in simple ways. Some employ deductive database capabilities to discover knowledge that is used to guide the data mining search while others uses Bayesian approaches, which uses prior probabilities over data and distributions as one form of encoding prior knowledge.

(viii). Assessing Statistical Significance

Problems similar to that of over fitting occur when the system is searching over many possible models. For example, if a system tests models at the 0.001 significance level, then on average, with purely random data, N/1000 of these models will be accepted as significant (Usama, 1996). One way of addressing this problem is to use methods that adjust the test statistic as a function of the search.

(ix). Changing Data and Knowledge

The incessant change in data make previously discovered patterns invalid; and the variables measured in a given application database can be modified, deleted or augmented with new measurements over time. The possible solutions to these challenges includes incremental methods for updating the patterns and treating change as an opportunity for discovery by using it to signal the search for patterns of change above (Usama, 1996).

(x). Integration with Other Systems

A stand-alone discovery system might not be very useful, a typical integration issues include integration with a database management systems; for example, through a query interface, integration with spreadsheets and visualization tools, and accommodation of real-time sensor readings.

Activity B: Student Self Assessment Exercise

Briefly discuss the following data mining issues list as follows:

(i). Security and social issues
(ii). Data quality
(iii). User interface issues
(iv). Data source issues
(v). Performance issues

4.0 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.

5.0 Summary

In this unit we 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.
- Data mining tasks can be grouped into four major classes, and these classes include classification, clustering, repression and association rule learning.
- There are lots of data mining issues affecting the implementation of data mining among these are security and social issues, data quality, user interface issues, data source issues, performance issues among others.
- There are lots of challenges facing data mining, among these are larger databases, high dimensionality, missing and noisy data.

6.0 Tutor Marked Assignment

1. Briefly explain the classification of data mining systems:
   (i) Classification by the type of data source mined
   (ii) Classification by the data model drawn on
   (iii) Classification by the kind of knowledge discovered
   (v) Classification by the mining techniques used

2. List and explain any five data mining challenges affecting the implementation of data mining.

7.0 Further Reading and Other Resources

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

Osmar R Zaiane (1999)- *Principles of Knowledge Discovery in Databases*

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

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

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

Introduction to Data Mining and Knowledge Discovery, Third Edition.

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

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

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

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House Pvt Ltd.

Module 1: Concepts of Data Mining

Unit 4: Data Mining Technologies

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 Objectives

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

- Understand the various data mining technologies available

3.0 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 sells 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 neighbor and memory based reasoning (MBR), logistic regression, discriminant analysis, genetic algorithms, generalized additive models (GAM) and boosting.

3.1 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.

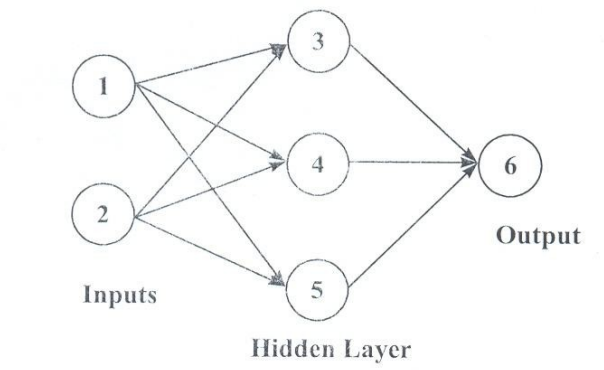


Figure 4.1 Neural Network with One Hidden Layer
Source: Introduction to Data Mining and Knowledge Discovery by Two Crows Corporation

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 complied 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 decreases. 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.

3.2 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.

Figure 4.2 shows a simple decision tree that describes the weather at a given time while illustrating all the basic components of a decision tree: the decision node branches and leaves. The objects contain information on the outlook, humidity, rain etc. Some of the objects are positive examples denoted by **P** while others are negative **N**. Classification in this case is the construction of a tree structure, illustrated in figure 4.2 which can be used to classify all the objects correctly.

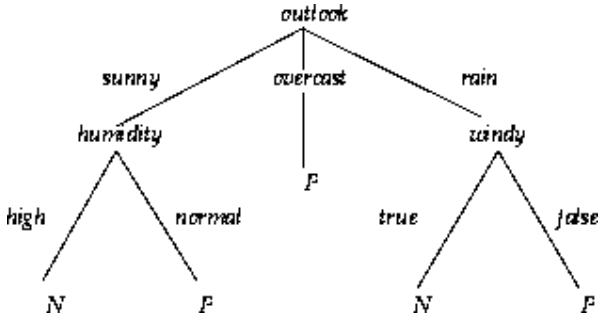


Figure 4.2 Decision Tree Structure
Source: Data Mining Techniques

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*.

3.3 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.

3.4 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 modeled 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.

Activity A: Student Self Assessment Exercise

Briefly discuss the following data mining technologies:

(i) Neural networks

(ii) Decision trees

3.5 K-Nearest Neighbor and Memory-Based Reasoning (MBR)

K-nearest neighbor (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 neighbor of the most similar cases or neighbors 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 neighbors belong.

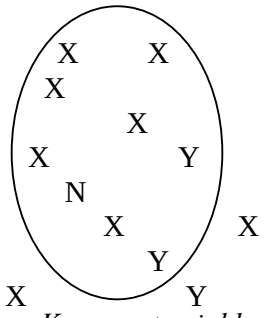


Figure 4.3 K-nearest neighbor. N is a new case.
Source: Two Crows Corporation, 2005

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, decide how large a neighborhood in which to do the comparisons, and also decide how to count the neighbors themselves. For instance you might give more weight to nearer neighbors than farther neighbors. (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 involve 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.

3.6 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 over head.

3.7 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.

3.8 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 decline 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.

3.9 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 fist 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

3.10 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 modeled 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.

Activity B: Student Self Assessment Exercise

Briefly explain any two of the following data mining technologies:

(i). Rule induction
(ii). Multivariate adaptive regression splines
(iii). Genetics algorithms

4.0 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.

5.0 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 neighbor.

6.0 Tutor Marked Assignment

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

7.0 Further Reading and Other Resources

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

Osmar R Zaiane (1999)- *Principles of Knowledge Discovery in Databases*

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

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

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

Introduction to Data Mining and Knowledge Discovery, Third Edition.

Jeffrey W. Seifert, (Dec. 2004), *Data Mining: An Overview* . From: Congressional Research Service, The Library of Congress.
Data Mining. Retrieved on 29/07/2009. Available Online: http://en.wikipedia.org/wiki/Data_mining.

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Limited and Vikas Publishing House Pvt Ltd.

Module 1: Concepts of Data Mining

Unit 5: Data Preparation and Preprocessing

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:

- Know the different data formats of an attribute
- Explain the meaning and importance of data preparation
- Define term data preprocessing
- Why data is being preprocessed
- Understand the various data pre-processing tasks

3.0 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 summarized in the following hierarchy:

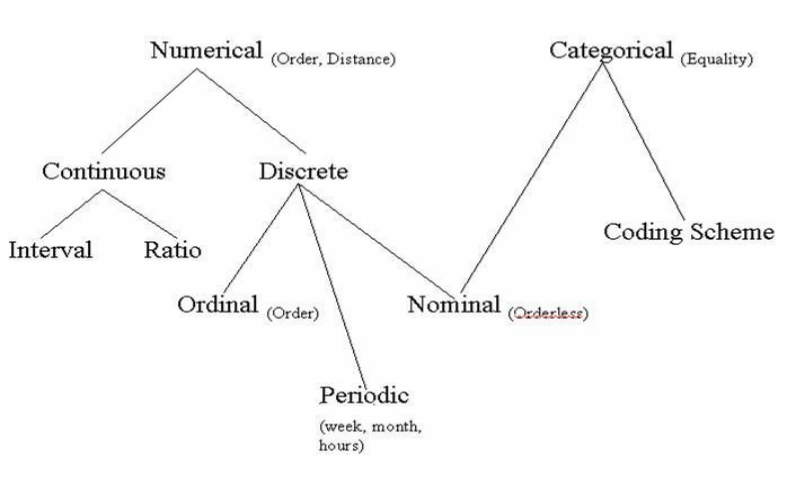


Figure 5.1 Different Data Formats of an Attribute
Source: Introduction to Data Mining:http;//www.eas.asu.edu/mining03/chap2/lesson_2. html

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)

3.1 Data Preparation

This is one of the most important tasks in data mining. It is a 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

3.1.1 Data Normalization

The different types of data normalization methods are:

1. **Decimal Scaling**: This type of scaling transforms the data into a range between (-1, 1). The transformation formula is v (i)/10^k.
   For the smallest k such that max (|V (i)|) 1

e.g. -For the initial range [-991, 99], k =3 and v=-991 becomes v =-0.991

2. **Min-Max Normalization** : This type of normalization transforms the data into a desired range, usually [0.1].
The transformation formula is:
v (i) = (v (i)-min A)/ (maxA- minA)*(new_maxA new_minA) + new_minA
where, [minA, maxA] is the initial range and [new_minA new_maxA] is the new range.
For example: - If v = 73600 in [12000, 98000] ⇒ .]1,0[εγναρ ωεν ετ νι 617.0 =ɔϖ

3. **Zero Mean Normalization:** When you use this type of normalization, the mean of the transformed set of data points is reduced to zero. For this, the mean and standard deviation of the initial set of data value are required. The transformation formula is
v = (v - meanA)/std_devA
where meanA and std_devA are the mean and standard deviation of the initial data values, e.g. - If meanIncome = 54000, and std_devIncome = 16000, then v = 76000 ⇒ =ɔϖ 1.225.

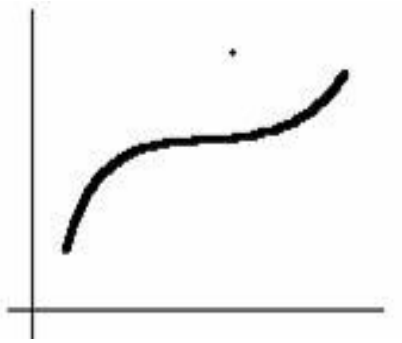
3.1.2 Dealing with Temporal Data

In case of temporal data, the goal is to forecast the $(n+1)^{th}$ value or t(n+1) from the previous n values. Given, x = { + t(1) , + t(2), ., t(n) }, predict the value for t (n+ 1)

3.1.3 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 for 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.
- **CurveFitting:** 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



- **HypothesisTesting 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.

3.2 Data Preprocessing

Data preprocessing is an important step in ensuring the data quality and to improve the efficiency and ease of the mining process real would data tends to be incomplete, noisy, inconsistent, high dimensional and multi-sensory etc. hence are not directly suitable for mining. It is a preliminary processing of data in order to prepare it for the primary processing or further analysis. The term can be applied to any first or preparatory processing stage when there are several steps required to prepare data for the user. For example, extracting data from a larger set.

Web usage data is collected in various ways each mechanism collecting attributes relevant for purpose. There is a need to preprocess the data to make it easier to mine for knowledge. Specifically, the following issues need to be addressed in data preprocessing:

(i) Instrumentation & Data Collection

Clearly improved data quality can improve the quality of any analysis on it. A problem in the Web domain is the inherent conflict between the analysis needs of the analysts (that want more detailed usage of data collected), and the privacy needs of users (who want as little data collected as possible). However, it is not clear how much compliance to this can be expected. Hence, there will be a continuous need to develop better instrumentation and data collection techniques, based on whatever is possible and allowable at any point in time.

(ii) Data Integration

The portion of Web usage data exist in sources as diverse as Web server logs, referral logs, registration files, and index server logs. Intelligent integration and correlation of information from these diverse sources can reveal usage information which may not be evident from any of them. The technique from data integration should be examined for this purpose.

(iii) Transaction Identification

Web usage data collected in various logs is at a very fine granularity. Hence, while it has the advantage of being extremely general and fairly detailed, it cannot be analyzed directly, since the analysis may start to focus on micro trends rather than on the macro trends. On the other hand, the issue of whether a trend is micro or macro depends on the purpose of a specific analysis. Hence it becomes very imperative to group individual data collection events into groups called Web transactions, before feeding it to the mining system.

Activity A: Student Self Assessment Exercise

1. (a) What is the importance of data preparation in data mining
   (b) List and explain the common types of data preparation method

3.3 Why Data Preprocessing?

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.

3.4 Data Quality Measures

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

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

3.5 Data Preprocessing Tasks

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

- Data cleaning
- Data transformation
- Attribute/ Feature construction
- Data reduction
- Discretization and concept hierarchy generation
- Data parsing and standardization

3.5.1 Data Cleaning

Data cleaning task consist of dealing with the following:

- Filling in (input) missing values/data
- Detecting and correcting inconsistent data
- Identifying outliers/smooth noisy data

(i). Missing Data

This may due to:

- Attribute not considered important
- Misunderstanding at data entry
- Inconsistent with other data and thus deleted
- Equipment malfunction e.g. EFTPOS down use cash. (EFTPOS- Electronic Fund Transfer Point-Of-Sale)

How to Handle Missing Data

This may be due to the following:

- Ignoring the record
- Fill in missing value manually (often impracticable)
- Fill in with a global constant (e.g. unknown, or N/A). Not recommended (data mining algorithm will see this as a normal value)
- Fill in with attribute mean or median
- Fill in with class mean or median (classes need to be known)
- Fill in with most likely value (using regression, decision trees, most similar records)
- Use other attributes to predict value (e.g. if a postcode is missing use suburb value and external look-up table).

(ii). Inconsistent Data

Data which is inconsistent with our models should be dealt with. Common sense can also be used to detect such kind of inconsistency. Examples are

- The same name occurring differently in an application
- Different names can appear to be the same (Dennis Vs Denis).
- Inappropriate values (Males being pregnant, or having a negative age)
- A particular bank database had about 5% of its customers born on 11/11/11/, which is usually the default value for the birthday attribute.

How to correct inconsistent Data

- It is important to have data entry verification (check both format and values of data entered).
- Correct with the help of external reference data (Look-up tables, e.g. sydney, new, 7000) or rules (e.g. male/0 M, Female/1 F)

(iii). Identity Outliers and Noisy Data

Normally noise is a minority in the data. This is because noise by nature is a random error.

- Random error or variances in a measurement
- Incorrect attribute values (Faulty data collection, data entry problems, data transmission problems, data conversion errors, inconsistent naming, technology limitations e.g. buffer overflow or field size limits).

How to Handle Noisy Data

Noise can be detected by measuring errors at the source of the data. Other way is to find the inconsistent values for the features or the classes by processing the data after collection, but this is more time consuming.

Noise can be removed by using the following techniques:

- **Binning:** This involves sorting the attribute values and partitioning them into bins (Bins means reducing the number of attributes by grouping them into intervals).
  -Then smooth by bin means, bin median or bin boundaries
- **Clustering**: Group values in clusters and then detect and remove outliers(automatic or manual)
- **Regression**: Smooth by fitting the data into regression function.
- Manual Inspections

3.5.2 Data Transformation

This involves consolidating data into forms suitable for data mining.

Ways of data transformation

- **Smoothing**: This is removing noise
- **Aggregation**: Moving up in the concept hierarchy on numeric attributes (summarization, data cube construction)
- **Generalization**: Moving up in the concept hierarchy on nominal attributes i.e. replacing data with higher level concepts (e.g. address details →city)
- **Attributes Construction** : Replacing or adding new attributes inferred by existing attributes.
- **Normalization**
  - Scaling attribute values to fall within a specified range e.g. Min-max(e.g. 0 1 interval)
  - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers: v = (v - Mean)/St-Dev
  - Decimal scaling (move decimal pint for all values)
- Important to save normalization parameters (in meta-data repository).

3.5.3 Attribute/ Feature Construction

- Sometimes it is helpful or necessary to construct new attributes or features
  - Help for understanding and accuracy
  - For example, create attribute volume based on attributes height, depths and width.
- Construction is based on mathematical or logical operations
- Attribute construction can help to discover missing information about the relationship between data attributes.

3.5.4 Data Reduction

It involves reducing the number of attributes which may be a result of:

- Databases or data warehouses often contain terabyte of data, resulting in (very) long run times for data mining techniques.
- High-dimensionality often prohibits the use of algorithms on the original data (causes of dimensionality)

Data Reduction Techniques:

(i) **Data Cube Aggregation (Roll-up):**

- Applying roll-up, dice or dice operations
- Data warehouses often have data stored at different levels of granularity (eg. day, week, mouth, quarter)
- Use the smallest representation that is enough to solve the problem.

(ii) **Dimensionality Reduction**

- Select a (minimum) sub-set of the available attributes (with similar probability distributions of classes compared to the original data)
- Find correlated, redundant or derived attributes (e.g. age and data of birth)
- Step-wise forward selection (find and select best attribute) or backward elimination (find and eliminate worst attribute)
- Use decision tree induction to find minimum attribute sub-set necessary.

(iii) **Data Compression**

- Data encoding or transformation
- Lossless or loss encoding
- Examples: String compression (e.g. ZIP, only allow limited manipulation of data), wavelet transformation, discrete Fourier transformation, principal component analysis.

(iv) **Numerosity Reduction**

- Parametric methods (e.g. repression and log-linear models) can be computationally expensive.
- Non parametric methods (histograms/binning, clustering, sampling).

3.5.5 Discrimination and Concept Hierarchy Operation

- Reduce the number of values for a continuous attribute by dividing the range into intervals.
- Concept hierarchies for numerical attributes can be constructed automatically.
- Binning (smoothing, distributing values into bins, then replace each value with mean, median or boundaries of the bin)

- ❖ Histogram analysis:
- Equal-interval (equiwidth) binning: split the whole range of numbers intervals with equal size.
- Equal frequency (equidepth) binning: use interval containing equal number of values
- ❖ Segmentation by natural partitioning (partition into 3,4, or 5 relatively uniform intervals)
- ❖ Entropy (information) based discretization

3.5.6 Data Parsing and Standardization

- ❖ Parse free format data into specific, well defined attributes
- ❖ Standardise using rules and look-up tables (correction and replacement tables), or probabilistically (hidden Markov models)
- ❖ Important for data linkage (based on names, addresses etc).

Activity B: Student Self Assessment Exercise

(1) State the reasons for pre-processing a data
(2) List and explain the various tasks involved in data pre-processing

4.0 Conclusion

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

5.0 Summary

In this unit we have learnt that:

- ❖ Attributes can be in different data formats
- ❖ Data preparation is one of the important tasks in data mining
- ❖ Data preprocessing is a preliminary processing of data in order to prepare it for further analysis
- ❖ Data has to be prepared because of a lot of reason these include real world data is dirty, incomplete data and noisy data.
- ❖ Data preprocessing involves a lot number of tasks which include data cleaning, data transformation, attribute/feature construction, data reduction, discretization and concept hierarchy generation, data parsing and standardization

6.0 Tutor Marked Assignment

1. List some of the factors used in measuring the quality of a data.
2. List and explain some data preparation methods
3. Briefly discuss the different types of data normalization methods.

7.0 Further Reading and Other Resources

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

Osmar R Zaiane (1999)- *Principles of Knowledge Discovery in Databases*

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

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

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

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

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

Cross Industry Standard Process for Data Mining , Retrieved on 19/09/2009. Available Online: http: // www.crisp-dm.org/

Data Preprocessing, Retrieved on 18/09/2009. Available Online: http://www.cs .ccsu.edu/markov.ccsucourses/Data Mining-3html.

Data Preprocessing, Retrieved on 18/09/2009. Available Online: http://en.wikipedia.orga/wiki/Data preprocessing

Data Pre-processing for Mining, Retrieved on 18/09/2009. Available Online: http://maya.cs.depaul.edu/%EF%BD%9Emobasher/webminer/survery/node25.htmlSECTION00061000000000000000

Data Preprocessing, Retrieved on 18/09/2009. Available Online: http://searchsqlserver.techtarget.com/sDefinition/o,sid87_gci810056,00.html

Introduction to Data Mining, Retrieved on 18/09/2009. Available Online: http://www.eas.asuedu/mining 03/chap2/lesson_2- htm.

Module 1: Concepts of Data Mining

Unit 6: Data Mining Process

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 organizations 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 Objectives

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

- Understand the basic steps of data mining for knowledge discovery

3.0 Process Models

The use of a systematic approach is essential for a successful data mining. A lot of vendors and consortium of organizations 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, Analyze, 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)

3.1 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 modeling
- Build model
- Evaluate model
- Deploy model and results

3.2 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.

3.3 Building a Data Mining Database

This step together with the next two which are *explore the data* and *prepare the data for modeling* 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 summarizing the data, multi-dimensional reports (which is sometimes referred to as pivot tables), and many different kinds of graphs or visualization. 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

Activity A: Student Self Assessment Exercise

1. (a) List all the various tasks in building a data mining database
   (b) Briefly explain any four of the task in (a)

3.3.1 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/organization responsible for maintaining the data
- Database administration (DBA)
- Cost (if purchased)
- Storage organization (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.

3.3.2 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)

3.3.3 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

3.3.4 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.

3.3.5 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)

3.3.6 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)

3.3.7 Load the Data Mining Database

In most cases data are stored in its own 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.

3.3.8 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.

3.4 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 labour 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.

3.5 Prepare Data for Modeling

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 is that the time it takes time to build a model increases with the number of variables and it blindly include 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 realizing it.

(ii). 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.

(iii). 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.

(iv). 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.

3.6 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 models 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 splitted 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 modeled. 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 **resubstitution.**

3.7 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 to 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)

3.8 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

Activity B: Student Self Assessment Exercise

What is the importance of data collection in building a data mining database?

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 modeling, build the model, evaluate the model, and deploy model and results.

6.0 Tutor Marked Assignment

1. (a). List the basic steps of data mining for knowledge discovery.
   (b). Briefly explain the steps listed in (a).

7.0 Further Reading and Other Resources

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

Osmar R Zaiane (1999)- *Principles of Knowledge Discovery in Databases*

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

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

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

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House P

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

Module 2: Applications and Trends in Data Mining

Unit 1: Data Mining Applications

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 Objectives

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

- Understand the various applications of data mining in our societies

3.0 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.

3.1 Data Mining Applications in Banking and Finance

Data Mining has been extensively used in the banking and financial markets. It is heavily used in the banking industry to model and predict credit fraud, to evaluate risk, to perform trend analysis, to analyze profitability and to help with direct marketing campaigns. In the financial markets, neural networks have been used in forecasting the price of stocks, in option trading in bond rating, portfolio management, commodity price prediction, mergers and acquisitions as well as in forecasting financial disasters.

There are so many software applications available in the market that uses data mining techniques for stock prediction. One of such software developed by neural applications corporation is a stock-prediction application that makes use of neural networks.

In the banking industry, the most widespread use of data mining is in the area of fraud detection. Although the use of the data mining in banking has not been noticed in Nigeria but has been in place in the advanced countries for credit fraud detection to monitor payment card accounts, thereby resulting in a health return on investment. However, finding banking industries that uses data mining is not easy, given their proclivity for silence. But one can assume that most large banks are performing some sort of data mining, though many have policies not to discuss it.

In addition many financial institutions have been the subject of many articles about their sophisticated data mining and modeling of their customers behaviour. The only significant problem with data mining in this aspect is the inability to leverage data-mining studies into actionable result. For example, while a bank may know that customer meeting certain criteria are likely going to close their accounts, it is another thing to figure out a strategy to do something about it.

3.2 Data Mining Application in Retails

Retailers are one of earliest adopted of data mining/ data warehouse. Retailers have seen improved decision-support processes lead directly to improved efficiency in inventory management and financial forecasting. The early adoption of data mining by retailers has given them a better opportunity to take advantage of data mining. Large retail chains and grocery stores store vast amounts of point-of-sale data that is information rich; The forefront of the applications of data mining in retail are direct marketing applications.

The direct-mail industry is an area where data mining, or data modeling, is widely used. It is almost every type of retailer, catalogers, consumer retail chains, grocers, publishers, business-to-business marketers, and packaged goods manufacturer makes use of direct customer segmentation, which is clustering problem in data mining. A lot of vendors offer customers segmentation packages e.g. Pilot Discovery Server segment Viewer (a customer segmentation software). This software uses the customer segmentation to help in direct-mailing campaigns. IBM has also used data mining for several retailers to analyze shopping patterns within stores based on point of sale (POS) information.

Other Types of Retail Data Mining Studies

Retailers are interested in many different types of data mining studies. In the area of marketing, retailers are interested in creating data-mining models to answer questions like:

- How much are customers likely to spend over long periods of time?
- What is the frequency of customer purchasing behaviour?
- What are the best types of advertisements to reach certain segments?
- What advertising media are most effective at reaching customers?
- What is the optimal timing at which to send mailers?

Merchandisers are beginning to profile issues such as:

- What types of customers are buying specific products?
- What determines the best product mix to sell on a regional level?
- What is a merchandise department saturated?
- What are the latest product trends?
- What are the times when a customer is most likely to buy?
- What types of products can be sold together?

Also, in identifying customer profitability, customers may wish to build models to answer questions like:

- How does a retailer retain profitable customers?
- What are the significant customer segments that buy products?

Customer identification is critical to successful retail organizations, and is likely to become more so. Data mining helps to model and identify the traits of profitable customers and reveal the hidden relationship that standard query processed cannot find.

3.3 Data Mining Applications in Telecommunications

The telecommunications industry has undergone one of the most dramatic makeovers of any industry. These industries generate and stores a tremendous amount of data.These include call detail data, this describes the calls that pass through the telecommunication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the telecommunication customers. The amount of data generated in telecommunication is so great that manual analysis of the data is difficult, if not impossible. The need to handle such a large volume of data led to the development of knowledge-based expert systems. This automated system performs important functions such as identifying network faults. The problem associated with this approach is that it is time consuming to obtain the knowledge from human experts (the knowledge acquisition bottleneck) and in many cases the experts do not have the requisite knowledge. The advent of data mining technology promised solutions to these problems for this reason the telecommunications industry were an early adopter of data mining technology (Gary M.Weiss)

Network data: Telecommunication networks are extremely complex configurations of equipment, comprising of thousands of interconnected components. Each network element is capable of generating a lot of error and status messages leading to a tremendous amount of network data. This data must be stored and analyzed in order to support network management functions such as fault isolation. As a result of this enormous number of network message generated, technicians cannot possibly handle every message. For this reason expert systems have been developed to automatically analyze these messages and take appropriate action and only involve a technician when the problem cannot be solved automatically. Data mining

technology therefore is now helping to identify faults by automatically extracting knowledge from the network data.

Call Detail Data: Every time a call is placed on a telecommunications network, descriptive information about the call is saved as a call detail record. The numbers of call detail records generated and stored are always huge. For example, the customers of GSM telecommunication in Nigeria cannot generate less than one million call detail records per day. Call detail records include sufficient information to describe the important characteristics of each call. At minimum, each call detail record includes the originating and terminating phone numbers, the data and time of the call and the duration of the call. Call detail records are generated in real time and therefore will be available almost immediately for data mining.

Customer Data

Just like other large businesses, telecommunication companies have millions of customers. By necessity this implies maintaining a database of information of these customers. This information includes name and address and other information such as service plan and contract information, credit score, family income and payment history. This information may even be supplemented with data from external sources, such as from credit reporting agencies. The customer data is often used in conjunction with other data in order to improve results. For instance, customer data is typically used to supplement call detail data when trying to identify phone fraud.

Now the applications of data mining in telecommunication industries can be grouped into three areas: Fraud detection, marketing/customer profiling and network fault isolation.

(1) Fraud Detection

This poses a very serious threat to telecommunication companies, which are: **Subscription fraud and superimposition fraud**

Subscription fraud occurs when a customer opens an account with the intention of not paying for the account charges. **Superimposition fraud** involves a customer opening a legitimate account with some legitimate activity, but also includes some superimposed illegitimate activity by a person other than the account holder. Superimposition fraud poses a bigger challenge for the telecommunications industry and for this reason we focus on applications for identifying this type of fraud.

The applications should basically operate on real-time using the call detail record and immediately fraud is detected or suspected, should trigger some actions. This action may be to immediately block the call or deactivate the account, or may involve opening an investigation, which will result in a call to the customers to verify the legitimacy of the account activity. The commonest method of identifying fraud is to build a profile of customer calling behaviour and compare recent activity against this behavior. Thus this data mining application relies on deviation detection. The calling behaviour is captured by summarizing the call details for a customer, if call detail summaries are updated in real-time, fraud can be identified soon after it occurs. (Gary M. Weiss)

(2) Marketing/ Customer Profiling

Telecommunication industries maintain a greet deal of data about their customers. In addition to the general customer data that most business collect, telecommunication companies also

store call details record which precisely describe the calling behaviour of each customer. This information can be used to profile the customers and these profiles can then be used for marketing and /or forecasting purposes.

(3). Network Fault Isolation

Telecommunication networks are extremely complex configurations of hardware and software. Most of the network elements are capable of at least limited self-diagnosis and these elements may collectively generate millions of status and alarm messages each month. In order to effectively manage the network, alarms must be analyzed automatically in order to identify network performance. A proactive response is essential to maintaining the reliability of the network. Because of the volume of the data, a single fault may cause many different, which may be unrelated, alarms to be generated; the task of network fault isolation is quite different. Data mining has a role to play in generating rules for identifying faults.
Also, telecommunications industry is interested in answering a wide variety of questions with the help of data mining, for instance:

- How does one recognize and predict when cellular fraud occurs?
- How does one retain customers and keep them loyal when competitions offer special offers and reduced rates?
- Which customers are most likely to churn?
- What characteristics make a customer likely to be profitable or unprofitable?
- How does one predict whether customers will buy additional products like cellular service, call waiting or basic service?
- What are the factors that influence customers to call more at certain times?
- What characteristics indicate high-risk investment such as investing in new fiber-optic lines?
- What products and services yield the highest amount of profit?
- What characteristics differentiate our products from those of our competitions?
- What set of characteristics would indicate companies or customer that will increase their line usage?

Activity A: Student Self Assessment Exercise

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.

3.4 Data Mining Applications in Healthcare

Healthcare industries generates mountains of administrative data and issues ranging from medical research, biotechs, pharmaceutical industry, hospitals, bed costs, clinical trials, electronic patient records and computer supported disease management will increasingly produce mountains of clinical data. This data is a strategic resource for health care institutions.

Data mining has been used extensively in the medical industry already. For example, NeuroMedical Systems used neural networks perform a Pap smear diagnostic aid. Vysis used neural networks to perform a protein analysis for drug development. The University of Rochester Cancer Center and the Oxford Transplant Center use Knowledge-SEEKER, which is a decision tree technology to help with their research. Also, the Southern California Spinal Disorders Hospital uses Information Discovery to data mine. Information Discovery quotes a doctor as saying Today alone, I came up with a diagnosis for a patient who did not even have to go through a physical exam .

With the use of data mining technology a pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competition market activity as well as information about the local healthcare systems. The result can be distributed to the sales force via a wide-area network that enables the representative to review the recommendations from the perspective of the key attributes in the decision process.

3.5 Data mining Application in Credit Card Company

A credit card company can control its vast warehouse of customer transaction data to identify customer most likely to have interest in a new credit product. With the use of a small test mailing, the attributes of customers with an affinity for the product can be identified.

3.6 Data Mining Application in Transportation Company

A diversified transportation company with a large direct sales force can apply data mining in identifying the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation to identify attributes of high-value prospects. Applying this segmentation to a general business database can yield a prioritized list of prospects by regions.

3.7 Application of Data Mining in surveillance

Recently data mining has been increasingly cited as an important tool for homeland security efforts. Some observers have suggested that data mining should be used as a mean of identifying terrorist activities such as money transfers and communications, and to identify and track individual terrorist themselves such as through travel and immigration records. The initiatives that have attracted significant attention include Terrorism Information Awareness (TIA) program, Secure Flight (formerly known as Computer-Assisted Passenger Prescreening System (CAPPS II), Analysis Dissemination Visualization Insight Semantic Enhancement (ADVISE), and the Multistate Anti-Terrorism Information Exchange (MATRIX)

Terrorism Information Awareness (TIA) Program

Terrorism Information Awareness was conducted by the Defense Advanced Research Projects Agency (DARPA) in U.S. This was a response to the terrorists attack of the September 11, 2001 on the World Trade Center. Information Awareness Office (IAO) was created at DARPA in January 2002 under the leadership of one technical office director, though several existing DARPA programs focused on applying information technology to combat terrorist threats. The mission statement of IAO suggested that emphasis was laid on the use of technology programs to counter asymmetric threats by achieving *total information awareness* useful for preemption, national security warning and national security decision making . To this end the TIA project was to focus on three specific areas of research which were based on:

(1). **Language translation** : The language translation technology would enable the rapid analysis of foreign languages, both spoken and written and allow analysts to quickly search the translated materials for clues about emerging threat.

(2). **Data search with pattern recognition and privacy protection**: The data search, pattern recognition, and privacy protection technologies would permit analysts to search immense quantities of data for patterns that suggest terrorist activity while at the same time controlling

access to data, enforcing laws and police, and ensuring detection of misuse of the information obtained.

(3). **Advanced collaborative and decision support tools:** The advanced collaborative reasoning and decision support technologies would allow analysts from different agencies to share data.

Computer-Assisted Passenger Prescreening System (CAPPS II)

CAPS II is similar to TIA and represented a direct response to the September 11, 2001 terrorist attacks. With the images of airliners that flew into buildings which is still very fresh in people s minds; air travel was now widely viewed not only as a serious valuable terrorist target, but also as weapon for inflicting larger harm. CAPPS II initiative was intended to replace the original CAPPS that are currently being used. Due to the growing number of airplane bombings, the existing CAPPS (originally called CAPS) was developed through a grant provided by the Federal Aviation Administration (FAA) to Northwest Airlines, with a prototype system tested in 1996. In 1997, other major carriers also began work on screening systems and by 1998, most of the U.S based airlines had voluntarily implemented CAPS, with the remaining few working toward implementation.

The current CAPPS system is a rule-based system that uses the information provided by the passenger when purchasing ticket to determine if the passenger belongs into one of the two categories; selectees the one requiring additional security screening, and those that do not. Moreover, CAPPS compares the passenger name to those on a lot of known or suspected terrorist. CAPPS II was described by TSA as an enhanced system for confirming the identities of passengers and to identify foreign terrorist or person with terrorist connections before they can board U.S aircraft. CAPPS II would send the information provided by the passenger in the Passengers Name Record (PNR), including full name, address, phone number and data of birth to commercial data providers for comparison to authenticate the identity of the passenger.

The commercial data provider then transmits a numerical score back to TSA indicating a particular risk level. Passengers with a green score would have undergone normal screening , while passengers with a yellow score would have undergone additional screening, passengers with a red score would not be allowed to board the flight, and would receive the attention of law enforcement . While drawing on information for commercial databases, TSA had stated that it would not see the actual information used to calculate the scores, and that it would not retain the traveler s information.

3.8 Application of Data Mining in Games

Since early 1960 s with the availability of oracles for certain combinational games also referred to as table bases (e.g. for 3X3 chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex; and contain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been opened up. This is the extraction of human usable strategies from these oracles. Current pattern recognition approaches do not seem to fully have the required high level of abstraction in order to be applied successfully. Instead, extensive experimentation with the table-bases, combined with an intensive study of table-base answers to well designed problems and with knowledge of prior art, that is pre-table base knowledge, is used to yield insightful patterns.

3.9 Application of Data Mining in Business

The application of data mining in customer relationship can contribute significantly to the bottom line. Instead of randomly contacting a prospect or customer through a call center or

sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer. More sophisticated methods can be used to optimize resources across campaigns so that one may predict which channel and which offer an individual is most likely to respond to across all potential offers. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

Businesses employing data may see a return on investment but also recognize that the number of predictive models can quickly become very large. Instead of one model predicting which customers will churn, a business could build a separate model for each region and customer type. And instead of sending an offer to all people that are likely to churn, it may only want to send offers to customers that will likely take to offer. And finally, it may also want to determine which customers are going to be profitable over a period of time and only send the offer to those that are likely to be profitable. In order to maintain this quantity of model, they need to mange versions and move to automated data mining.

Data mining is also helpful in human-resources department for identifying the characteristics of their most successful employees. Information obtained, such as the universities attended by highly successful employees, can help human resources focus recruiting efforts accordingly. Another example of data mining, which is often referred to as market basket analysis, relates to its use in retail sales; for example if a clothing store records the purchases of customers, a data mining system could identify those customers that favour silk shirts over cotton ones.

Market basket analysis is also used to identify the purchase patterns of the Alpha consumer. Alpha consumers are people that play a key role in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of the society. Data mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich history of customer transactions on millions of customers dating back several years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

3.10 Applications of Data Mining in Science and Engineering

Data mining is widely used in science and engineering such as in bioinformatics, genetics, medicine, education and electrical power engineering.

In the area of study on human genetics, the important goal is to understand the mapping relationship between the inter-individual variation in human DNA sequences and variability in disease susceptibility. This is very important to help improve the diagnosis, prevention and treatment of the diseases. The data mining technique that is used to perform this task is known as *multifactor dimensionality reduction.*

In electrical power engineering, data mining techniques are widely used for monitoring high voltage equipment. The reason for condition monitoring is to obtain valuable information on the insulation s fitness status of the equipment. Data clustering such as Self-Organizing Map (SOM) has been applied on the vibration monitoring and analysis of transformer On-Load-Tap Changers (OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms.

Other areas of data mining applications are in biomedical data facilitated by domain ontology, mining clinical trial data, traffic analysis using self-organizing map (SOM). And

recently data mining methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.

3.11 Application of Data mining in Spatial Data

Spatial data mining follows along the same functions in data mining with the end objective of finding patterns in geography. Data mining and geographic information systems (GIS) have existed as two separate technologies, each with its own methods, traditions and approaches to visualization and data analysis. Data mining which is a partially automated search for hidden patterns in large databases offers great potential benefits for applied GIS-based decision-making. Recently the task of integrating these two technologies has become critical, especially as various public and private sector organization possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information hidden there.

Activity B: Student Self Assessment Exercise

1. List and briefly explain any five applications of data mining in our societies.

4.0 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 program performance.

5.0 Summary

In this unit we have learnt that:

- The applications of data mining in banking and finance industry, retails, telecommunications, healthcare, credit card company, transportation company, surveillance, games, business, spatial data, science and engineering, include fraud detection, risk evaluation, to forecast the price of stocks and financial disaster. And also for marketing and network fault isolation

6.0 Tutor Marked Assignment

1. Briefly explain the following applications of data mining in surveillance:
   (a). Terrorism Information Awareness (TIA)
   (b). Computer-Assisted Passenger Prescreening System (CAPPS)

2. Briefly discuss the roles of data mining in the following application areas:
   (i). Spatial data
   (ii). Science and engineering
   (iii). Business
   (iv). Telecommunication

7.0 Further Reading and Other Resources

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

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

Hans-P, Kregel, K. M. Borgwardt, P. Kroger, A. Pryakhin, M. Schubert and A. Zimek, (2007), *Future Trends in Data Mining*. Ludwig-Maximilians-Universitat.

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

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

Gary M. Weiss, *Data Mining in Telecommunications*

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

Lean, A. and Lean, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House P

Module 2: Applications and Trends in Data Mining

Unit 2: Future Trends in Data Mining

1.0 Introduction

Over the recent years data mining has been establishing itself as one of the major disciplines in computer science with growing industrial impact. Without any doubt, research in data mining will continue and even increase over coming decades.

In this unit, we shall be examining the present and future trend in the field of data mining with a focus on those which are thought to have the most promising and applicability to future data mining.

2.0 Objectives

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

- Understand the present and the future trend in data mining.
- Know the major trends in technologies and methods.

3.0 The Present and Future of Data Mining

The field of data mining and knowledge discovery in databases (KDD) is growing astronomically and equally showing great potential for the future. As it was earlier discussed in module 2, unit 1, data mining is presently being successfully applied in many areas of human endeavour; this ranges from telecommunication industry, healthcare, banking and finance, transportation company and so on.

What is the future of data mining? Certainly, the field of data mining has made a great stride in the past years, and many industry analysis and experts in the area are optimistic that the future will be bright. There is explicit growth in the area of data mining. A lot of industry analysis and research firms have projected a bright future for the entire data mining/KDD area and its related area Customer Relationship Management (CRM). The spending in the area of business intelligence that encompasses data mining is increasing in U.S. Moreover, data mining projects are expected to grow at geometric ratio because a lot of consumer-based industries with e-commerce orientation will utilize some kinds of data mining model.

As earlier discussed, data mining field is very broad and there are many methods and technologies that have become dominant in the field. Also, not only has there been developments in the conventional areas of data mining there are other areas which have been identified as being especially important as future trends in the field.

3.1 Major Trends in Technologies and Method

There are lots of data mining trends in terms of technologies which are presently being developed and researched. These trends include methods for analyzing more complex forms of data, as well as specific techniques and methods. The trends identified include distributed data mining, hypertext/hypermedia mining, ubiquitous data mining, as well as multimedia, spatial and geographical data mining and time series/sequential data mining.

3.1.1 Distributed/ Collective Data Mining

Distributed/collective data mining is an area that is attracting a good amount of attention data mining. Most of the data mining which is being done today focuses on a databases or data warehouse of information that is physically located in one place. However, situation arises where information may be located in different places in different physical locations. This is basically known as distributed data mining (DDM). Therefore, the goal is to effectively mine distributed data that is located in heterogeneous sites. Examples include biological information located in different databases, data which comes from the databases of two different firms, or analysis of data from different branches of a corporation, combination of which would be an expensive and time-consuming process.

Distributed data mining (DDM) is used to offer a different approach to traditional approach analysis, together with a global data model . In more specific terms, this is specified as:

- Performing local data analysis for generating partial data models, and
- Combining the local data models from different data sites in order to develop the global model. (Jeffrey Hsu).

This global model combines the results of the separate analyses. The global model that is often produced, especially if the data in different locations has different features or characteristics may become incorrect or ambiguous. This problem is especially critical when the data in distributed site is heterogeneous rather than homogeneous. These heterogeneous data sets are known as vertically partition datasets.

An approach proposed by Kargupta et al (2000) speaks of the collective data mining (CDM) approach, which provides a better approach to vertically partition datasets using the notion of orthonormal basis functions, and computes the basis coefficients to generate the global model of the data (Jeffrey Hsu; Kargupta et.al. 2000).

3.1.2 Ubiquitous Data Mining (UDM)

The advent of laptops, palmtops, cell phones and wearable computers is making ubiquitous access to large quantity of data possible. Advanced analysis of data for extracting useful knowledge is the next natural step in the world of ubiquitous computing to access and analyze data from a ubiquitous computing device offer many challenges. For example UDM introduces additional cost due to communication, computation, security and other factors. So, one of the objectives of UDM is to mine data while minimizing the cost of ubiquitous presence. Another challenging aspect of UDM is the human-computer interaction.

To visualize patterns like classifiers, clusters, associations and others in portable devices are usually difficult. The small display areas offer serious challenges to interactive data mining environments. Data management in a mobile environment is also a challenging issue. Moreover, the sociological and psychological aspects of the integration between data mining technology and our lifestyle are yet to be explored. The key issues to consider include theories of UDM, advanced algorithms for mobile and distributed applications, data management issues, mark-up languages and other representation techniques; integration with database applications for mobile environments, architectural issues (architecture, control, security and communication issues), Specialized mobile devices for UDM, agent interaction, cooperation, collaboration, negotiation, organizational behaviour, applications of UDM (Applications in business science, engineering, medicine and other disciplines) location management issues in UDM and technology for web-based applications of UDM (Jeffrey Hsn; Kargupta and Joshi, 2001)

3.1.3 Hypertext and Hyper Media Data Mining

Hypertext and hypermedia data mining can be characterized as mining data that includes text, hyperlinks, text markups and other forms of hypermedia information. As such, it is closely related to both web mining and multi-media mining, which are covered separately in this section, but in reality are quite close in terms of content and applications. While the World Wide Web is substantially composed of hypertext and hypermedia elements, there are other kinds of hypertext/hypermedia data sources which are not found on the web. Examples of these include the information found in online catalogues, digital libraries, online information databases and the likes.

In addition to the traditional forms of hypertext and hypermedia, together with the associated hyperlink structures, there are also inter-document structures which exist on the web, such as the directories employed by such services as Yahoo or the Open Directory project (http://dmoz.org). These taxonomies of topics and subtopics are linked together to form a large network or hierarchical tree of topics and associated links pages.

Some of the important data mining techniques used for hypertext and hypermedia data mining include classification (supervised learning), clustering (unsupervised learning), semi-structured learning and social network analysis.

1. Classification or Supervised Learning

In this type of technique, the process starts off by reviewing training data in which items are marked as being part of a certain class or group. This is the basis from which the algorithm is trained. One application of classification is in the area of web topic directories, which can group similar sounding or spelt terms into appropriate sites. The use of classification can also result in searches which are not only based on keyboards, but also on category and

classification attributes. Methods used for classification include naïve Bayes classification, parameter smoothing, dependence modeling, and maximum entropy (Jeffrey Hsn; Chokrabarth, 2000).

2. Unsupervised Learning

This differs from classification in that classification involve the use of training data, clustering is concerned with the creation of hierarchies of documents based on similarity and organize the documents based on that hierarchy. Intuitively, this would result in more similar documents being placed on the leaf of the hierarchy, with less similar sets of document areas being placed higher up, closer to the root of tree. Techniques that are used for unsupervised learning include k-means clustering, agglomerative clustering, random projections and latent semantic indexing (Jeffrey Hsu; Chakrabarti, 2000).

3. Semi-Supervised Learning

This is an important hypermedia-based data mining. It is the case where there are both labeled and unlabeled documents, and there is a need to learn from both types of documents.

4. Social Network Analysis

Social network analysis is also applicable because the web is considered to be a social network which examines networks formed through collaborative association, whether it is between friends, academicians doing research or service on committees, and between papers through references and citations. Graph distances and various aspects connectivity come into play when working in the area of social works (Jeffrey; Larson, 1996; Mizruchi et.al; 1996).

Activity B: Student Self Assessment Exercise

Briefly explain the following data mining techniques used for hypertext and hypermedia data mining:
(a). Supervised learning
(b). Unsupervised learning
(c). Semi-supervised learning
(d). Social network analysis

3.1.4 Multimedia Data Mining

Multimedia data mining is the mining and analysis of various types of data, including images, video, audio and animation. The idea of mining data that contains different kinds of information is the main objective of multimedia data mining. Multimedia data mining incorporates the areas of text mining as well as hypertext/hypermedia mining, these fields are closely related. Most of the information describing these other areas also apply to multimedia data mining; although, this field is rather new, but holds much promise for the future.

Because multimedia information is a large collection of multimedia objects, it must be represented differently from conventional forms of data. One approach is to create a multimedia-type data into a form which is suited for analysis using one of the main data mining techniques, but considering the unique characteristics of the data this may include the use of measures and dimensions for textures, shape, colour and related attributes. In essence it is possible to create a multidimensional spatial database. The types of analysis that can be conducted on multimedia databases include associations, clustering, classification and similarity search.

Another developing area in multimedia data mining is that of audio data mining (mining music).The idea is basically to use audio, signals to indicate the patterns of data or to represent the features of data mining results. The basic advantage of audio data is that while using a technique such as visual data mining may disclose interesting patterns from observing graphical displays, it does require users to concentrate on watching patterns which can become monotonous. But when representing it as a stream of audio, it is possible to transform patterns into sound and music and listen to pitches, rhythms, tune, and melody in order to identify anything interesting or unusual.

3.1.5 Spatial and Geographic Data Mining

The term spatial data mining can be defined as the extraction of implicit knowledge, spatial relationships, or other patterns that is not explicitly stored in spatial databases. Spatial and geographic data could contain information about astronomical data, natural resources, or even orbiting satellites and spacecraft that transmit images of earth from out in space. Much of this data is image-oriented, and can represent a great deal of information of properly analyzed and mined. Some of the components of spatial data that differentiates it from other kinds includes distance and topological information which can be indexed using multidimensional structures, and requires special spatial data access methods, together with spatial knowledge representation and data access methods, along with the ability to handle geometric calculations.

To analyze spatial and geographic data include such tasks as understanding and browsing spatial data, uncovering relationships between spatial data items and also between non-spatial and spatial items. Applications of these would be useful in such fields as remote sensing, medical imaging, navigation and related uses. Some of the techniques and data structures which are used when analyzing spatial and related types of data include the use of spatial warehouses, spatial data uses and spatial OLAP. Spatial data warehouse can be defined as those which are subject-oriented, integrated, nonvolatile and time-variant. Some of the challenges in constructing a spatial data warehouse include the difficulties of integration of data from heterogeneous sources, and also applying the use of on-line analytical processing which is not only relatively fast, but also offers some forms of flexibility.

By and large, spatial data cubes, which are components of spatial data warehouse, are designed with three types of dimensions and two types of measures. The three types of dimensions include:

- The non-spatial dimension- that is data that is non-spatial in nature
- The spatial to non-spatial dimension - primitive level is spatial but higher level-generalization is non-spatial and
- The spatial-to-spatial dimension both primitive and higher levels are all spatial.

In terms of measures, there are both numerical (numbers only) and spatial (pointers to spatial object) measured used in spatial data cubes.

Beside the implementation of data warehouse for spatial data there is also the issue of analysis which can be done on the data. Some these analysis include association analysis, clustering methods and the mining of raster databases.

3.1.6 Time Series/Sequence Data Mining

This is another important area that centers on the mining of time series and sequence-based data. It involves the mining of a sequence-based data, which can either be referenced by time (time-series, such as stock market and production process data), or is simply a sequence of

data that is ordered in a sequence. Generally, one aspect of mining time series data focuses on the goal of identifying movements or components which exist within the data (trend analysis).these include long-term or trend movements, seasonal variations, cyclical variations, and random movements.

Other techniques that can be used on these kinds of data include similarity search, sequential pattern mining and periodicity analysis.

- **Similarity Search:** This is concerned with the identification of a pattern sequence which is close or similar to a given pattern, and this form of analysis can be broken down into two subtypes: *whole sequences matching* and *subsequence matching*. The whole sequence matching attempts to find all sequences which bear a likeness to each other, while subsequences matching attempts to find those patterns which are similar to a specified given sequences.

- **Sequential Pattern Mining:** It has its focus on the identification of sequences that occurs often in a time series or sequence of data. It is particularly useful in the analysis of customers where certain buying patterns could be identified, such as what might be the likely follow-up purchase to purchasing a certain electronics item or computer.

- **Periodicity Analysis:** This attempt to analyze the data from the perspective of identifying patterns which repeat or recur in a time series. This form of data mining analysis can be categorized as being full periodic, partial periodic or cyclic periodic. Full periodic is the situation where all of the data points in time contribute to the behavior of the series. This is in contrast to partial periodicity, where only while certain points in time contribute to series behavior, while cyclical periodicity relates to sets of events that occur periodically.

3.1.7 Constraint-Based Data Mining

Most of the data mining techniques which currently exist are very useful but lacks the benefits of user control or guidance. One method of implementing some forms of human involvement into data mining is in the form of constraint-based data mining. This form of data mining incorporates the uses of constraints which has its own characteristics and purpose. These are:

- **Knowledge-Type Constraints**: This type of constraint specifies the type of knowledge which is to be mixed, and is typically specified at the beginning of any data mining query. Some of the types of constraints that can be used include clustering, association and classification.
- **Data Constraints:** This constraint identifies the data which is to be in the specific data mining query. Since constraint-based mining is ideally conducted within the framework of an ad-hoc, query driven system, data constraint can be specified in a form similar to that of a SQL query.
- **Dimension/Level Constraints:** Because most of the information mined is in the form of a database or multidimensional data warehouse, it is possible to specify constraints which specify the levels or dimensions to be included in the current query.
- **Interestingness Constraints:** It would be useful to determine what ranges of a particular variable or measure is considered to be particularly interesting and should be included in the query.
- **Rule Constraints:** This specifies the specific rules which should be applied and used for a particular data mining query or application.

One application of constraint-based approach is in the Online Analytical Mining Architecture (OLAM) which was developed by Han, Lackshamanan, and Ng, 1999, and is designed to support the multidimensional and constraint-based mining of databases and data warehouses. Constraint-based mining of database mining is one of the developing areas that allows for the use of guiding constraints which should make for better data mining.

3.1.8 Phenomenal Data Mining

Phenomenal data mining focuses on the relationships between data and the phenomena which are informed from the data. For example with the use of receipts from cash supermarket purchases, it is possible to identify various aspects of the customers who are making these purchases. Some of these purchases may include age, income, ethnicity and purchasing habits. One aspect of phenomenal data mining, and in particular the goal of inferring phenomena from data, is the need to have access to some facts about the relations between these data and their related phenomena.

3.1.9 Increasing Usability

An ultimate trend that data mining encounters is increased usability to detect understandable patterns and to make data mining methods more user-friendly. Although, future algorithms might handle this complexity, the need for user guidance during preprocessing and data mining will dramatically increase. Even in current data mining algorithms, many established methods employ quite a few different input parameters. In a wider sense, selecting a data mining algorithm and a data transformation method itself can also be considered as a problem of user guidance. Also, to select the best possible methods and find a reasonable parameter settings are often very time consuming.

Other aspects of usability are the intuitiveness when adjusting the parameters and the parameter sensitivity. If the results are not strongly dependent on slight variations of the parametization, adjusting the algorithms becomes less complex. To fulfill these requirements, there is need to distinguish between the two types of parameters and afterwards propose four goals for future data mining methods.

- The first type of parameter, which is called *type I*, is tuning data mining algorithms for deriving useful patterns. For instance, k for a k-NN classifier influences directly the achieved classification.
- The second type of parameter called *type II*, is more or less describing the semantics of the given objects. For instance, the cost matrix used by edit distance has to be based on domain knowledge and this varies from application to application. The important thing of this type of parameter is that the parameters are used to model additional constraints from the real world.

Based on these considerations, the following proposals can be formulated for future data mining solutions:

(i). Avoid type I parameters if possible when designing algorithms.

(ii). If type I parameters are necessary, try to find the optimal parameter settings automatically. For many data mining algorithms, it might be possible to integrate the given parameters into the underlying optimization problem.

(iii). Instead of finding patterns for one possible value of a type II parameter, try to simultaneously derive patterns for each parameter setting and store them for post processing.

(iv). Develop user-friendly methods to integrate domain knowledge where necessary. In most cases the only applicable approach for selecting type II parameters is to include additional domain knowledge into the data mining task.

Activity B: Student Self Assessment Exercise

Briefly explain the following categories of constraint-based data mining:
(i). Knowledge-type Constraints (ii). Data Constraints
(iii). Rule Constraints (iv). Dimension/ Level Constraints

4.0 Conclusion

With the unstoppable and unavoidable growth in data collection in the years ahead, data mining is playing an important role in the way massive data sets are analyzed. Trends clearly indicate that future decision making systems would weigh on even quicker and more reliable models used for data analysis. And to achieve this, current algorithms and computing systems have to be optimized and tuned, to effectively process the large volumes of raw data to be seen in future.

5.0 Summary

In this unit we have learnt that:

- The field of data mining and knowledge discovery has made a giant strides in the past and many experts in the field are optimistic that the future will be bright
- There are lots of data mining trends in terms of technologies and methodologies which include distributed data mining, hypertext/hypermedia mining, ubiquitous data mining, time series/sequence data mining, constraint-based, phenomenal data mining and increasing usability.

6.0 Tutor Marked Assignment

Briefly discuss the following data mining trends in terms of technologies and methods:
(a). Ubiquitous data mining (b). Multimedia data mining
(c). Hypertext and hypermedia (d). Spatial and Geographic Data Mining

7.0 Further Reading and Other Resources

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

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

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

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

Lean, A. and Lean, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House P

Hans-Peter K., Karasten M.B, Peer K., Alexey P. Matthias S. and Arthur Z. (March, 2007) *Future Trends in Data Mining.* Springer Science + business Media, 23 March 2007

Jayaprakash, P, Joseph Z., Berkin O. and Alok C., *Accelerating Data mining Workloads: Current Approaches and Future Challenges in System Architecture Design*

Jeffrey, H., Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century.

Module 3: Data Warehouse Concepts

Unit 1: Overview of Data Warehouse

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:

- Define the term data warehouse
- Understand the goals and characteristics of data warehouse
- Know the major components of data warehouse
- Understand the structure and approaches to storing data in data warehouse
- Describe the users and application areas of data warehouse

3.0 Definition of Data Warehouse

The father of data warehousing William H. Inmon defined data warehouse as follows: A data warehouse is a *subject oriented*, *integrated*, *non-volatile* and *time-variant* collection of data in support of management decisions.

Other definitions of data warehouse include:

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

As mentioned by W.H. Inmon in one of his papers, the data warehouse environment is the foundation of Decision Support Systems (DSS) and is about molding data into information and storing this information based on the subject rather than application.

A data warehouse is designed to house a standardized, constraint, clean and integrated form of data sourced from various operational systems in use in the organization, structured in a way to specifically address the reporting and analytic requirements. One of the primary reasons for developing a data warehouse is to integrate operational data from various sources into a single and consistent architecture that supports analysis and decision making in an organization.

The words operational (legacy) systems as used in the definition are used to create, update and delete production data that feed the data warehouse. A data warehouse is analogous to a physical warehouse. It is the operational systems that create data parts that are loaded into the warehouse. Some of those parts are summarized into information components and are stored in the warehouse.
The users of data warehouse make request and information which are the products that were created from the components and parts stored in the warehouse are delivered. A data warehouse is typically a blending of technologies, including relational and multidimensional databases, client/server architecture, extraction/transformation programs, graphical interfaces and more.

3.1 Goals of Data Warehouse

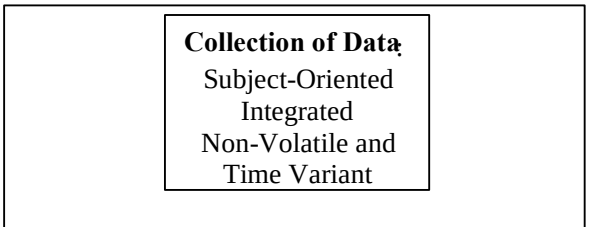
The major goals of data warehousing are stated as follows:

- To facilitate reporting as well as analysis
- Maintain an organizations historical information
- Be an adaptive and resilient source of information
- Be the foundation for decision making.

3.2 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



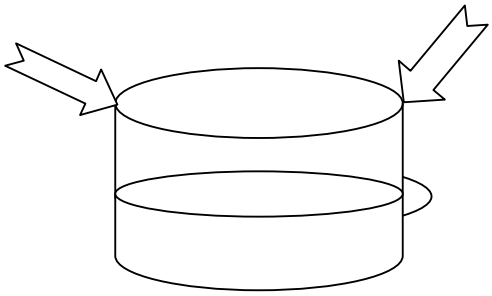


Figure 1.1: Characteristics of Data Warehousing
Source: Data Warehouse with Oracle By M.A Shahzad

(1). 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.)

(2). 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.

(3). 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.

(4). 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.

3.3 Evolution in Organizational Use of Data Warehouses

Data warehousing which is a process of centralized data management and retrieval, just like data mining it is a relatively new term, although the concept itself has been around for years. Organizations basically started with relatively simple use of data warehousing. Over the years, more sophisticated use of data warehousing evolves. The following basic stages of the use of data warehouse can be distinguished:

- **Off Line Operational Database** : The data warehouses at this stage were developed by simply copying the data off an operational system to another server where the processing load of reporting against the copied data does not impact the operational system s performance.
- **Off Line Data Warehouse:** The data warehouses at this stage are updated from data in the operational systems on a regular basis and the data in warehouse data is stored in a data structure designed to facilitate reporting.
- **Real Time Data Warehouse** : The data warehouse at this level is updated every time an operational system performs a transaction, for example, an order or a delivery.
- **Integrated Data Warehouse** : The data warehouses at this level are updated every time an operational system carries out a transaction. The data warehouse then generates transactions that are passed back into the operational systems.

3.4 Advantages and Disadvantages of Data Warehouse

3.4.1 Advantages of Data Warehouse

Some of the significant benefits of implementing a data warehouse are as follows:

(i). **Facilitate decisionmaking:** A data warehouse allows reduction of staff and computer resources required to support queries and reports against operational and production database. The implementation of data warehousing also eliminates the resource use up on production systems when executing long-running, complex queries and reports.

(ii). **Better Enterprise Intelligence**: Increased quality and flexibility of enterprise analysis arises from the multi-tiered data structures of a data warehouse that supports data ranging from detailed transactional level to high-level summary information. Guaranteed data accuracy and reliability result from ensuring that a data warehouse contains only trusted data.

(iii). A data warehouse provides a common data model for all data of interest regardless of the data s source. This makes it easier to report and analyze information than it would be if multiple data models were used to retrieve information such as sales invoices, order receipts, general ledger charges etc.

(iv). Information in the data warehouse is under the control of data warehousing users so that, even if the source system data is purged over time, the information in the warehouse can be stored safety for extended periods of time.

(v). Because data warehouse is separated from operational systems, it provide retrieval of data without slowing down operational systems.

(vi). **Enhanced Customer Service**: Data warehouses can work in conjunction with customer service, hence, enhance the value of operational business applications or better customer relationships notably customer relationship management (CRM) systems by correlating all customer data via a single data warehouse architecture.

(vii). **Business Reengineering:** It allows unlimited analysis of enterprise information and often provide insights into enterprise processes that may yield breakthrough ideas for reengineering the processes. By defining the requirements for data warehouse results in better enterprise goals and measures. Knowing what information is important to an enterprise and will provide direction and priority for reengineering efforts.

(viii). Before loading data into the data warehouse, inconsistencies are identified and resolved. This greatly simplifies reporting and analysis.

(ix). **Cost Effective:** A data warehouse that is based upon enterprise-wide data requirements provides a cost effective means of establishing both data standardization and operational system interoperability. This typically offers significant savings

3.4.2 Disadvantages of Data Warehouses

There are also some disadvantages to the implementation of data warehouse. Some of these are:

(i). Because data must be extracted, transformed and loaded into the warehouse, there is an element of latency in the use of data in the warehouse.

(ii). Data warehouses are not the optimal or most favourable environment for unstructured data.

(iii). Data warehouses have high costs. A data warehouse is usually not static. Maintenance costs are always on the high side.

(iv). Data warehouse can get outdated relatively quickly and there is a cost of delivering suboptimal information to the organization.

(v). Because there is often a fine line between data warehouse and operational system, duplicate and expensive functionality may be developed. Or, functionality may be developed in the data warehouse that in retrospect should have been developed in the operational systems and vice versa.

Activity A: Student Self Assessment Exercise

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

3.5 Data Warehouse Components

The following basic coverage describes each of the components of data warehouse in figure 1.2. This description is based upon the work of William H. Inmon, credited as the father of a data warehousing concept. The major components of a data warehouse are:

- Summarized data
- Operational systems of record
- Integration/Transformation programs
- Current detail
- Data warehouse architecture or metadata
- Archives

1. Summarized Data

Summarized data is classified into two namely:

- Lightly summarized data
- Highly summarized data

Lightly summarized data are the hallmark of data warehouse. All enterprise elements (e.g. department, region, function) do not have the same information requirements, so effective data warehouse design provides for customized lightly summarized data for every enterprise elements. An enterprise element may have access to both detailed and summarized data, but there will be much less data than the total stored in current detail. (Alexis L. et al, 1999)

Highly summarized data are primarily for enterprise executives. It can come from either the lightly summarized data used by enterprise elements or from current detail. Data volume at this level is much less than other levels and represents a diverse collection supporting a wide variety of needs and interests. In addition to access to highly summarized data, executives also have the capability of accessing increasing levels of detail through a drill down process. (Alexis L. et al, 1999).

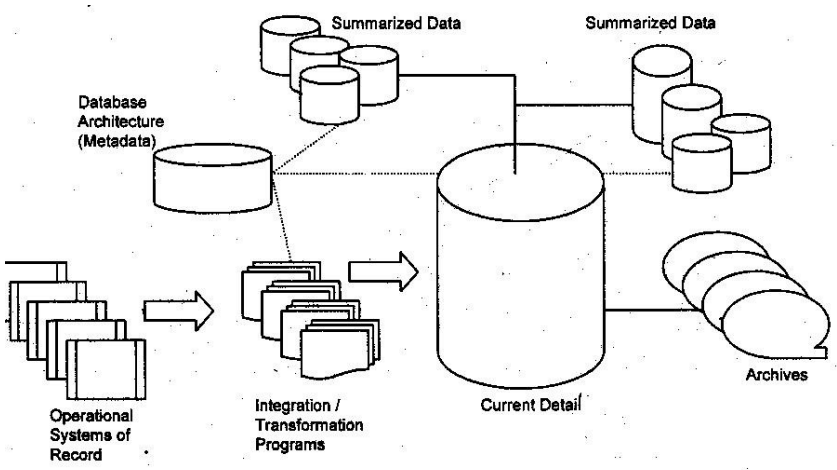


Figure 1.2 Components of a Data Warehouse
Source: Fundamentals of Information Technology By A. Leon et al, page 29.1

2. Current Detail

Current detail is the heart of a data warehouse where bulk of data resides and it comes directly from operational system and may be stored as raw data or as aggregations of raw data. Current detail that is organized by subject area represents the entire enterprise rather than a given application. Current detail is the lowest level of data granularity in the data warehouse. Every data entity in current detail is a snapshot, at a moment in time, representing the instance when the data are accurate. Current detail is typically two to five years old and its refreshment occurs as frequently as necessary to support enterprise requirements (Alexis L. et al, 1999).

3. System of Record

A system of record is the source of the data that feeds the data warehouse. The data in the data warehouse is different from operational systems data in the sense that they can only be read and not modified. Thus, it is very necessary that a data warehouse be populated with the highest quality data available that is most timely, complete, and accurate and has the best structural conformance to the data warehouse. Often these data are closest to the source of

entry into the production. In other cases, a system of record may be containing already summarized data.

4. Integration and Transformation Programs

As the operational data items pass from their systems of record to a data warehouse, integration and transformation programs convert them from application-specific data into enterprise data. These integration and transformation programs perform functions such as:

- Reformatting, recalculating, or modifying key structures.
- Adding time elements.
- Identifying default values
- Supplying logic to choose between multiple data sources
- Summarizing, tallying and merging data from multiple sources.

Whenever either the operational or data warehousing environment changes, integration and transformation programs are modified to reflect that change.

5. Archives

The data warehouse archives contain old data normally over two years old but of significant value and containing interest to the enterprise. There are usually large amount of data stored in the data warehouse archives with a low incidence of access. Archive data are most often used for forecasting and trend analysis. Although archive data may be stored with the same level of granularity as current detail, it is more likely that archive data are aggregated as they are archived. Archives include not only old data in raw or summarized form: they also include the metadata that describes the old data s characteristics (Alexis L. et al, 1999).

6. Meta Data- this is data about data

This is one of the most important parts of a data warehouse; it is metadata or data about data. It is also called data warehouse architecture, metadata is integral to all levels of the data warehouses, but exists and functions in a different dimension from other warehouse data.

Meta data provides data repository. It provides both technical and business view of data stored in the data warehouse. It lays out the physical structure which includes:

- Data elements and their types
- Business definition for the data elements
- How to update data and on which frequency
- Different data elements
- Valid values for each data elements

Meta data plays a very significant role in the definition, building, management and maintenance of data warehouses. In a data warehouse metadata are categorized into two namely:

- Business Metadata
- Technical Metadata

Business metadata describes what is in the warehouse, its meaning in terms of business. The business metadata lies above technical metadata, adding some details to the extracted material. This type of metadata is important as it facilitates business users and increases the accessibility. **Technical metadata** describes the data elements as they exist in the warehouse. This type of metadata is used for data modeling initially, and once the warehouse is erected this metadata is frequently used by warehouse administrator and software tools. (Alexis L. et al, 1999)

3.6 Structure of a Data Warehouse

The structure of a data warehouse is shown in figure 1.3 and consists of the following:

- **Physical Data Warehouse:** This is the physical database in which all the data for the data warehouse is stored, along with metadata and processing logic for scrubbing, organizing, packaging and processing the detail data.
- **Logic Data Warehouse:** It also contains metadata enterprise rules and processing logic for scrubbing, organizing, packaging and processing the data, but does not contain actual data. Instead, it contains the information necessary to access the data wherever they reside. This structure is effective only when there is a single source for the data and they are known to be accurate and timely (Alexis L. et al, 1999).
- **Data Mart** : This is a data structure that is optimized for access. It is designed to facilitate end-user analysis of data. It typically supports a single and analytical application used by a distinct set of workers. Also, a data mart can be described as a subset of an enterprise-wide data warehouse which typically supports an enterprise element (e.g. department, region, function). As part of an iterative data warehouse development process, an enterprise builds a series of physical (or logical) data marts over time and links them via an enterprise-wide logical data warehouse or feeds them from a single physical warehouse

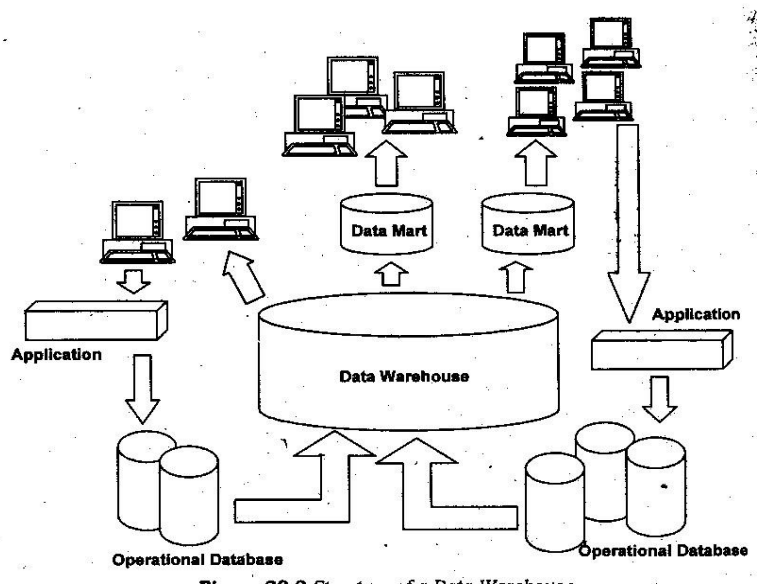


Figure 1.3 Structure of a Data Warehouse
Source: Fundamentals of Information Technology by Alexis Leon et al page 29.2

3.6.1 Differences Between Data Warehouse and Data Mart

Data Warehouse	Data Mart
1. It is a multi-subject information store	It is a single subject data warehouse
2. It is 100 s of gigabytes in size	The size is less than 100 gigabytes
3. It is difficult to build	It is less difficult to build compared with data warehouse

3.7 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.

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.

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.

Benefits of Normalized Approach

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

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.

3.7.1 Interface with Other Data Warehouses

The data warehouse system is likely to be interfaced with other applications that use it as the source of operational system data. A data warehouse may feed data to other data warehouse or smaller data warehouses called data marts. Data warehouse can be a better single and consistent source for data instead of the operational systems. It is important to remember that most of the operational state information is not carried over to the data warehouse. This data warehouse cannot be the source of all operation system interfaces. (Alexis L.et al, 1999)

3.8 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:

- Statisticians
- Knowledge workers
- Information consumers
- Executives

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

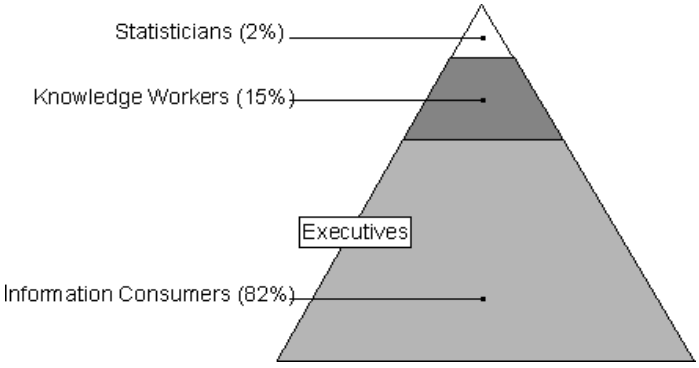


Figure 1.4 The User Pyramid
Source: Data Warehouse Design Considerations by Dave Browning and Joy Mundy, Dec. 2001

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.

3.9 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.

3.10 Applications of Data Warehouse

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

- Credit card churn analysis
- Insurance fraud analysis
- Call record analysis
- Logistics management

Activity B: Student Self Assessment Exercise

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

4.0 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.

5.0 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 characteristics of a data warehouse, these include subject oriented, integrated, non-volatile and time variant

- 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.

6.0 Tutor Marked Assignment

1. (a). What do you understand by the term data warehouse?
   (b). Briefly explain the following characteristics of data warehouse:
   (i). Subject-oriented (ii). Integrated
   (iii). Non- volatile (iii). Time variant

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

7.0 Further Reading and Other Resources

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

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House P

Jayaprakash, P, Joseph Z., Berkin O. and Alok C., *Accelerating Data mining Workloads: Current Approaches and Future Challenges in System Architecture Design*

Dave Browning and Joy Mundy, (Dec., 2001). *Data Warehouse Design Considerations.* Retrieved on 13/10/2009. Available Online: hhtp://msdn.microsoft.com/en-us/library/aa902672(SQL.80)aspx.

Module 3: Data Warehouse Concepts

Unit 2: Data Warehouse Architecture

1.0 Introduction

The term architecture in the content of an organization s data warehousing effort is a conceptualization of how the data warehouse is built. There is no right or wrong architecture; rather multiple architectures exist to support various environments and situations. The worthiness of the architecture can be judged on how the conceptualization aids in the building, maintenance and usage of the data warehouse. This unit examines the meaning of data warehouse architecture, its evolution, components and differentiates between extraction, transformation and load. Also to explored are the relevance of resource and data management in data warehouse architecture.

2.0 Objectives

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

- Understand the term data warehouse architecture
- Know the three types of data warehouse architecture
- Describe the components of data warehouse architecture
- Understand the use of extraction, transformation and load tools
- Describe what is meant by resource management

3.0 Definition of Data Warehouse Architecture

Data warehouse architecture is a description of the elements and services of the warehouse, with details showing how the components will fit together and how the system will grow over times. There is always an architecture, either ad-hoc or planned, but experience shows that planned architectures have a better chance of succeeding.

According to Warren Thornthwaite, a partner with Menlo Park, CA-based InfoDynamics LLC every data warehouse has an architecture . It is either ad-hoc or planned, implied or documented. Unfortunately, many warehouses are developed without an explicit architectural plan, which severely limits flexibility. Without architecture, subject areas do not fit together, connections lead to nowhere, and the whole warehouse is difficult to manage and change. In addition, although it might not seem important, the architecture of a data warehouse becomes the framework for product selection.

3.1 Data Warehouse Architecture Evolution

Architecture provides the mechanism to achieve enterprise integration to support organization and business. It provides an organizing framework that will improve data sharing between

agencies, and in the long run allow for faster development, reuse and consistent data between warehouse projects.

Most importantly, this architecture is an evolutionary process. The architecture as defined in section 3.0 of this unit was initially developed as a place to start. The first enterprise warehouse projects will be based on this architecture. Increments of additional agency projects will cause this architecture to evolve. As technology changes and improves, that too will most likely require us to make adjustments to this architecture. This incremental development of both the architecture and the warehouse offers an opportunity to learn and to minimize the impact of mistakes.

3.2 Types of Data Warehouse Architectures

Data warehouses and their architectures vary depending upon the specifics of an organization s situation. Three common architectures are:

- Data Warehouse Architecture (Basic)
- Data Warehouse Architecture (with a staging Area)
- Data Warehouse Architecture (with a staging Area and Data Marts).

3.2.1 Data Warehouse Architecture (Basic)

Figure 2.1 depicts a very simple architecture for a data warehouse. End-users directly access data derived from several source systems through the data warehouse

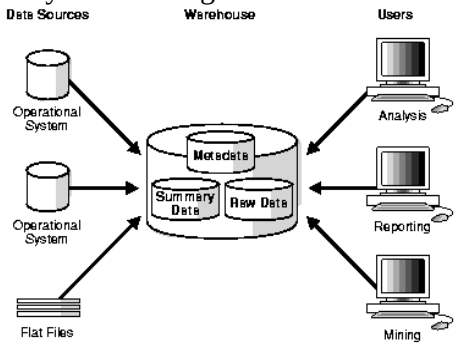


Figure 2.1 Architecture of a data warehouse
Source: Oracle9i Data Warehousing Guide Release 2 (9.2)

In figure 2.1, the metadata and raw data of a traditional OLTP system is present, as additional types of data and summary data. Summaries are very valuable in data warehouses because they pre-compute long operations in advance. For example, a typical data warehouse query is to retrieve something like August sales. A summary in oracle is called a *materialized view*.

3.2.2 Data Warehouse Architecture (With a Staging Area)

In figure 2.1, you need to clean and process your operational data before putting it into the warehouse. This can be done programmatically, though most data warehouse uses a staging area instead. A staging area simplifies building summaries and general warehouse management. Figure 2.2 depicts this typical architecture.

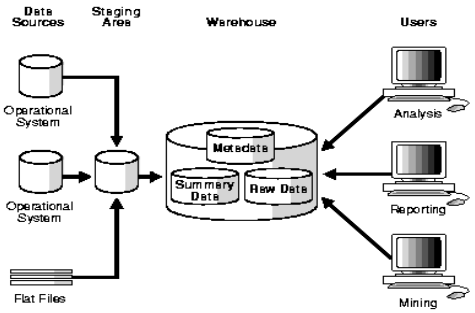


Figure 2.2 Architecture of a Data Warehouse with a Staging Area
Source: Oracle9i Data Warehousing Guide Release 2 (9.2)

3.2.3 Data Warehouse Architecture (With a Staging Area and Data Marts)

Even though, the architecture in figure 2.2 is quite common, you may want to customize your warehouse s architecture for different groups within your organization. This can be done by adding data marts, which are systems designed for a particular line of business. Figure 2.3 shows an example where purchasing, sales and inventories are separated. In this example, a financial analyst might want to analyze historical data for purchases and sales.

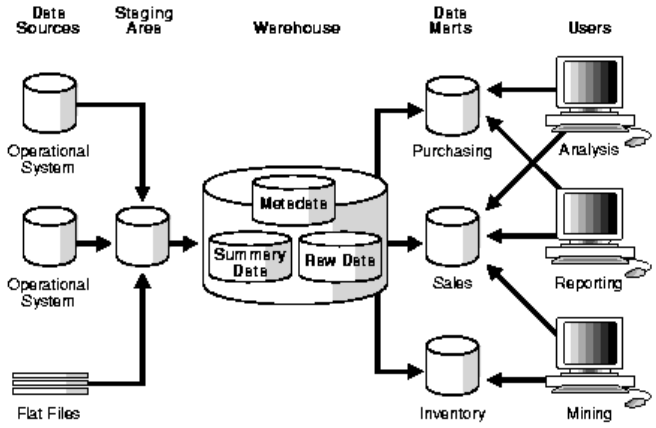


Figure 2.3 Architecture of a Data Warehouse with a Staging Area and Data Marts
Source: Oracle9i Data Warehousing Guide Release 2 (9.2)

Activity A: Student Self Assessment Exercise

(a). What do you understand by the term data warehouse architecture?
(b). State the three types of data warehouse architecture.

3.3 Components of Data Warehouse Architecture

A data warehouse architecture consists of seven major components. These components offer high level of flexibility and scalability for both the enterprise and the agencies wishing to implement a business intelligence solution. These components are: Operational source systems

- Data staging area
- Data warehouse
- One or more conformed data marts
- Extract transform load
- Business intelligence
- Metadata and the metadata repository

1. Operational Source Systems

A data source system is the operational or legacy system of record whose function is to capture and process the original transactions of the business. These systems are designed for data entry, not for reporting, but it is from here the data in data warehouse gets populated. The source systems should be thought of as outside the data warehouse, since we have no control over the content and format of the data. The data in these systems can be in many format from flat files to hierarchical, and relational RDBMS such as MS Access, Oracle, Sybase, UDB and IMS to name a few. Other sources of data already be cleaned and integrated and available from operational data stores. Other sources are PeopleSoft, Web logs and even external information sources.

2. Data Staging Area

The data staging area is that portion of the data warehouse restricted to extracting, cleaning, matching and loading data from multiple legacy systems. The data staging area is the back room and is explicitly off limits to the end users. The data staging area does not support query or presentation services.

Data staging is a major process that includes the following sub procedures:

- **Extraction**: The extract step is the first step of getting data into the data warehouse environment. Extracting means reading and understanding the source data, and copying the pas that are needed to the data staging for further work.

- **Transformation**: Once the data is extracted into the data staging area, there are many transformation steps, including:
  - Cleaning the data by correcting misspellings, resolving domain conflicts, dealing with missing data elements, and passing into standard formats
  - Purging selected fields from the legacy data that are not useful for data warehouse
  - Combining data sources by matching exactly on key values or by performing fuzzy matches on non-key attributes
  - Creating surrogate keys for each dimension record in order to avoid dependency on legacy defined keys, where the surrogate key generation process enforce referential integrity between the dimension tables and fast tables.
  - Building the aggregates for boosting the performance of common queries.

- **Loading and Indexing**: At the end of transformation process, the data is in the form of load record images. Load in the data warehouse environment usually takes the form of replicating the dimensional tables and fast tables and presenting these tables to bulk loading facilities in recipient data mart. Bulk loading is a very important capability that is to be contrasted with record-at-a time loading which is far slower. The target data mart must them index the newly arrived data for query performance.

3. Data Warehouse Database

A data warehouse database is a relational data structure that is optimized for distribution. The warehouse is no special technology in itself. It collects and store integrated sets of historical, non-volatile data from multiple operational systems and feeds them to one or more data marts. Also, it becomes the one source of the truth for all shared data and differs from OLTP databases in the sense that it is designed primarily for reads not writes.

4. Data Marts

Data mart is a logical subset of an enterprise-wide data warehouse. The easiest way to theoretically view a data is that a mart needs to be an extension of the data warehouse. Data is integrated as it enters the data warehouse from multiple legacy sources. Data marts then derive their data from the central data warehouse source. The theory is that no matter how many data marts are created, all the data are drawn from the one and only one version of the truth, which is the data contained in the warehouse.

Distribution of the data from the warehouse to the mart provides the opportunity to build new summaries to fit a particular department s needs. The data marts contain subject specific information supporting the requirements of the end users in individual business units. Data marts can provide rapid response to end-user requests of if most queries are directed to pre-computed and aggregated data stored in the data mart

5. Extract Transform Load

Data Extraction-Transformation-Load (ETL) tools are used to extract data from data sources, cleanse the data, perform data transformations, and load the target data warehouse and then again to load the data marts. The ETL tool is also used to generate and maintain a central metadata repository and support data warehouse administration. The more robust ETL tools is the more it integrates with OLAP tools, data modeling tools and data cleaning tools at the metadata level

6. Business Intelligence (BI)

This is the key area within the business intelligence continuum that provides the tools required by users to specify queries, create arbitrary reports, and to analyze their own data using drill-down and On-line Analytical Processing (OLAP) fractions. One tool however does not fit all. BI tools arena still requires that we match the right tools to the right end user.

7. Metadata and the Metadata Repository

A repository is itself a database containing a complete glossary for all components, database, fields, objects, owners, access, platforms and users within the enterprise. The repository offers a way to understand what information is available, where it comes from, where it is stored, the transformation performed on the data, its currency and other important facts about the data. Also, metadata describes the data structures and the business rules at a level above a data dictionary. Metadata has however taken on a more visible role among day-to-day knowledge workers. Today it serves as the main catalog, or map to a data warehouse. Metadata can be generated and maintained by an ETL tool as part of the specification of the extraction, transformation and load process. The repository can also capture the operational statistics on the operation of the ETL process

3.4 Extraction, Transformation and Load

Transforming data is generally performed as part of the preparation before data is loaded into the data warehouse and data marts. Understanding the business usage of this information and the specific business questions to be analyzed and answered are the keys to determining the transformation necessary to produce the target data mart. ETL tools are used to extract data from operational and external source system, transform the data, and load the transformed data in a data warehouse. This same tool is used to extract and transform the data from the warehouse and distribute it to the data marts. When a schedule is defined for refreshing the data, the extraction and transformation schedule must be carefully implemented so that it both

meets the needs of the data warehouse and does not adversely impact the source systems that store the original data.

Extraction

Extraction is a means of replicating data through a process of selection from one or more source database. Extraction may or not employ some forms of transformation. Data extraction can be accomplished through custom-developed programs. But the preferred method uses vendor- supported data extraction and transformation needs as well as use an enterprise metadata repository that will document the business rules used to determine what data was extracted from the source systems.

Transformation

Data is transformed from transaction level data into information through several techniques: filtering, summarizing, merging, transposing, converting and deriving new values through mathematical and logical formulas. These all operate on one or more discrete data fields to produce a target result having more meaning from a decision support perspective than the source data. This process requires understanding the business focus, the information needs and the currently available sources. Issues of data standards, domains and business terms arise when integrating across operational databases.

Data Cleansing

Cleansing data is based on the principle of populating the data warehouse with quality data, that is consistent data, which is of a known, recognized value and confirms with the business definition as expressed by the user. The cleansing operation is focused on determining those values which violate these rules and either reject, or through a transformation process bring the data into conformance. Data cleansing standardizes data according to specifically defined rules, eliminates redundancy to increase data query accuracy, reduces the cost associated with inaccurate, incomplete and redundant data, and reduces the risk of invalid decisions made against incorrect data.

3.5 Data Management

Components

Data architecture defines all the components, interfaces and processes for implementing and managing an integrated and cohesive data policy. These components are defined below:

1. Data

The term data can be referred to as the collections of raw facts and is stored in multiple application systems on multiple platforms using multiple methods and is catered using online transaction processing (OLTP) systems. While information is the output of a processed and is derived from online analytical processing (OLAP systems used for analysis, planning and management reporting through access to a variety of sources.

Forms of data are:

- **Data Types**: Data types define the domain of values that a data field can have. New technologies are extending the range of data types that can be stored and processed by computers. These offer new ways of interacting and communicating with users and amplify the human/machine interface.

- **Text and Numeric Fields** : Data fields comprise of rows of information containing discrete values related to some business entity. Current operational databases are almost completely text and numeric data fields. Since there are discrete values, these can be individually retrieved, queried and manipulated to support some activities,

reporting need or analysis. These data types will continue to play a significant role in all our databases.

- **Images**: Scanned pictures of documents, photos and other multidimensional forms can be stored in databases. The scanned images is a single data field and is retrieved and updated as a single fact. Software outside of the DBMS is used to manipulate the image.

- **Geographic Data**: Geographic data is information about features on the surface and subsurface of the earth, including their location, share, description and condition. Geographic information includes spatial and descriptive tabular information in tabular and raster (image) formats. A geographic information system (GIS) is a hardware and software environment that captures, stores, analyzes, queries, and displays geographic information. Usually geographic information is the basic for location- based decision making, land-use planning, emerging response, and mapping purposes.

- **Multimedia: Voice, Animation and Video**
  Multimedia is a technology that integrates two or more types of media such as text graphics, sound voice, full-motion video, still video or animation into computer-based application. Multimedia applications are increasing as we employ new modalities of communicating with users. Voice can be stored in a database to capture instructional, informative messages that can then be played back rather displayed as text. This facilitates these situations where keyboards and visual displays are difficult to utilize.

  Graphics, animation and video, likewise, offer an alternative way to inform users where simple text does not communicate easily with the complexity or the relationships between information components. An example might be graphic displays of vessels and equipment allowing drill down too more detailed information related to the part or component. Video may be useful in demonstrating some complex operations as part of a training program.

- **Object**: Objects are composites of other data types and other objects. Objects form a hierarchy of information unlike the relational models. Objects contain facts about themselves and exhibit certain behaviours implemented as procedural code. They also inherit the facts and behavious of their parent objects up through the hierarchy. Relational database stores everything in rows and columns. Although they may support large binary object (LOB) fields that can hold anything an object database can support.

2. Databases

Database is a collection of data organized to service many applications with minimum redundancy. Databases organize data and information into physical structures, which are then accessed and updated through the services of a database management system. Some of the common terms associated with database are:

- **Database Management System (DBMS)** : A database management system is a specialized software that is used to construct, access, expands control and maintain the database.

- **Relational Database Management System (RDBMS:** Relational database management system (RDBMS) is software designed to manage a collection of data. Data is organized into related sets of tables, rows and columns so that relationships between and among data can be established. For example a vehicle database can contain two tables, one for customer information and one for vehicle information. An owns relationship is then established between the two tables.

- **Multi-Dimensional Databases (MDDBMS):** A multi-dimensional database (MDDBMS) is specifically designed for efficient storage and retrieval of large volumes of data. Multi-dimensional databases are organized into fact tables and dimensions that intersect with the facts table to identify to what the fact pertains. Databases of this construction are sued for on-line analytical processing, also referred to as OLAP.

3. Data Warehouse Data Marts

A data warehouse as earlier defined in this course is a database designed to support decision-making in an organization or enterprise. It is refreshed, or batch updated, and can contain massive amount of data. When the database is organized for one department or function it is often referred to as data mart rather than a data warehouse. The data in a data warehouse is typically historical and static in nature. Data marts also contain numerous summary levels. It is structured to support a variety of elaborate analytical queries on large amounts of data that can require extensive searching.

4. Operational Data Stores (ODS)

The operational data store (ODS) is a database that consolidates data from multiple source systems and provides a near real-time, integrated view of volatile and current data. An ODS differs from a warehouse in that ODS s contents are updated in the course of business, whereas a data warehouse contain static data.

5. Data Access

Data access middleware is the layer of communication between a data access level and the Database. The following components are essential for the data access middleware layer for accessing a relational database in an N-tier application environment:

- **Structured Query Language (SQL):** A query language is used to query and retrieve data from relational databases. The industry standard for SQL is ANSI standard SQL. RDBMS vendors implement SQL drivers to enable access to their proprietary databases. Vendors may add extensions to the SQL language for their proprietary databases.

- **Open Database Connectivity (ODBC) Drivers** : Middleware is used to connect database access tools to relational databases using a generic application program interface (API). ODBC drivers are under-provided and allow databases to be connected and used by a generic interface. The ODBC drivers enable access to data and provide insulation between a program and the specific RDBMS language used by each database. Database access tools and programs do not have to be customized for each database, because an ODBC configuration file maintains the database connections. ODBC can be implemented as a client-based solution or a server-based solution.

6. Processing Access

Access to data can be categorized into two major groups:

- **On-line Analytical Processing (OLAP):** This is decision support software that allows the user to quickly analyze information that has been summarized into multidimensional views. OLAP application is a system designed for for few but complex (read only) request. Traditional OLAP products which are also known as **Multidimensional OLAP** or **MOLAP** summarize transactions into multidimensional views ahead of time. User queries on these types of databases are extremely fast because the consolidation has already been done. OLAP places the data into a cube structure that can be rotated by the user, which is particularly suited for financial summaries.

- **On line Transaction Processing (OLTP):** Online transaction processing means that master files are updated as soon as transactions are entered at terminals or received over communication lines. They are considered real time systems. OLTP application is a system designed for many but simple concurrent and updating requests.

7. Replication

Replication is used to keep distributed database up to date with a central source database. Replication uses a database that has been identified as a central source and reproduces the data to distributed target databases. As more and more data is being made available to the public over internet, replication of selected data to locations outside the wire wall is becoming more common. Replication data should be accessed by applications in a read-only mode. If updates were allowed on replicated data, data would quickly become corrupted and out of sync. Updates should be directed to the database access tier in change of updating the authoritative source, rather than to a replicated database. Replication services are available from most relational database vendors for their particular products.

- **Replication Services**: Replication is the process of distributing information access a network of computers. Replication strategies may also employ some forms of transformation such that the information has different content and meaning. When information has a low volatility, replication may be a valid strategy for optimizing performance. Replication will need to be evaluated against our network, our volume of activity and local access requirements.

- **Partial and Full Refresh:** A full refresh simply replaces the existing target with a new copy of the source databases. It is simple to implement, but may not be practical for large databases due to the amount of time involved in the process of dumping and reloading the data.

  A partial refresh replicates only the changes made from the source database to the remote databases. The processing involved in replicating only changes is more complex than a full refresh, but is an optimal solution for a large database. In a partial refresh method, either data or transactions can drive the replication e.g. sending the exact data that was changed on the central database.

- **Mirroring:** Mirroring provides two images of the same database and allows the two databases to be synchronized simultaneously. That is, an update to one causes the mirror to also be updated. This form of replication is the most accurate, but also,

potentially the most difficult to achieve and the most costly to operate (Data Management and Data warehouse, 2002).

3.6 Resource Management

Resource management provides the operational facilities for managing and securing enterprise-wide, distributed data architecture. It provides a common view of the data including definitions, stewardship, distribution and currency and allows those charged with ensuring operational integrity and availability of the tools necessary to do so. Research needs to be done for all components in this category.

Security

Security becomes an increasingly important aspect as access to data and information expands and takes on new forms such as Web pages and dynamic content. Our security policy needs to be examined to ensure that it can be enforced given the move to distributed data and internet access.

Administration

Administration encompasses the creation, maintenance, support, backup and recovery, and archival processes required in managing a database. There is used to be able to centrally manage all of the enterprise databases to ensure consistency and availability. Distributing data to appropriate platform will place more importance on administration and control. This becomes the key to maintaining the overall data architecture. Currently database administration is done using the tools and services native to and provided by most relational database vendors for their particular products. Investing in centrally managed administration products and resources will improve data quality, availability and reliability

Activity B: Student Self Assessment Exercise

List and briefly explain any five components of data management.

4.0 Conclusion

Therefore, data warehouse always has an architecture which can either be ad-hoc or planned, implied or documented without which many warehouses subject areas do not fit together, connections lead to nowhere, and the whole warehouse becomes difficult to manage and change.

5.0 Summary

In this unit we have learnt that:

- The term architecture in the content of an organization s data warehousing effort is a conceptualization of how the data warehouse is built.
- There three common architectures are for data warehouse design, these are: data warehouse architecture (Basic), data warehouse architecture (with a staging Area) and data warehouse architecture (with a staging Area and Data Marts).
- Data warehouse architecture consists of seven major components namely: Operational source systems, data staging area, data warehouse, data marts, extract- transform-load tool, business intelligence and metadata/metadata repository.
- Data extraction transformation load (ETL) tools are used to extract data from data sources, cleanse the data, perform data transformations, and load the target data warehouse.

- Resource management provides a common view for data, this include definitions, stewardship, distribution and currency and allows those charged with ensuring operational integrity and availability of the tools necessary to do so.

6.0 Tutor Marked Assignment

(1). List and briefly explain the seven major components of data warehouse architecture.

(2). (a). Differentiate between data marts and data warehouse

(b). What is the importance of extract transformation load tool to data warehouse?

7.0 Further Reading and Other Resources

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

Surajit C., and Umeshwar D., An Overview of Data Warehousing and OLAP Technology

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

Anil Rai, *Data Warehouse and its Applications in Agriculture* , Indian Agriculture Statistics Research Institute Library Avenue, New Delhi-110 012.

Laura Hadley, *Developing a Data Warehouse Architecture* , Retrieved on 13/10/2009. From: http://www.users.qwest.net/¬lauramh/resume/thorn.htm.

Understanding the Data Warehouse Architecture, Retrieved on: 13/10/2009. From: http://msdn.microsoft.com/en-us/library/ms244687(VS.80).aspx

Data Warehouse, Retrieved on 29/09/2009. From http://en.wikipedia.org/wiki/Data-warehouse.

Data Warehouse. Retrieved on 29/09/2009. From http://www.intranetjournal.com/features/datawarehousing.html.

Muhammad A. Shahzad, Data Warehousing With Oracle. www.oracular.com

What is OLAP?. Retrieved on 13/10/2009. Available online: http://www.tech-faq.com/olap.shtml.

Types of OLAP Systems. Retrieved on 13/10/2009. Available online: http://www.olap.com/w/index.php/Types of OLAP Systems.

OLAP and OLAP Server Definitions. Retrieved on 13/10/2009. Available online: http://www.moulton.com/olap.glossary.html

Dave Browning and Joy Mundy, (Dec., 2001). *Data Warehouse Design Considerations.* Retrieved on 13/10/2009. Available Online: hhtp://msdn.microsoft.com/en-us/library/aa902672(SQL.80)aspx.

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

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House Pvt Ltd.

Module 3: Data Warehousing Concepts

Unit 3: Data Warehouse Design

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
- Understand the basic methodologies used in building a data warehouse
- Explain the phases involved in developing a data warehouse
- Know the data warehouse testing life cycle

3.0 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,

3.1 Data Warehouse Design Methodologies

The basic techniques used in building a data warehouse are as follows:

- Bottom-up Design
- Top-down Design
- 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, summarized 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). Topdown Design

In this design we first build a data warehouse for the complete organization 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 normalized 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 centralized 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.

Activity A: Student Self Assessment Exercise

List and explain the three basic technologies in developing a data warehouse.

3.2 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)
- Design the relational database and OLAP cubes
- Develop the maintenance applications
- Develop analysis applications
- Test and deploy the system

3.2.1 Identify and Gather Requirements

You must identify the sponsors. A successful data warehouse project needs a sponsor in the business organization 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 designer s 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.

3.2.2 Design the Dimensional Model

The user requirements and data realities drive the design of the dimensional model which must address business models, granules of detail, and what dimensions and facts to include. The dimensional model must suit the requirements of the users and support ease of use for direct access. The model must also be designed so that it is easy to maintain and can adapt to future changes. The model design must also result in relational databases that support OLAP cubes to provide instantaneous query results for analysts.

A typical dimensional model uses a star or snowflake design that is easy to understand and relate to business needs, supports simplified queries, and provides superior query performance by minimizing table joins.

Dimensional Model Schemas

The principal characteristic of a dimensional model is a set of detailed business facts surrounded by multiple dimensions that describe those facts. You can arrange schema objects in the schema models designed for data warehouse in a variety of ways, most data warehouse use a dimensional model. When realized in a database, the schema for a dimensional model contains a central fact table and multiple dimension tables. A dimensional model may produce a star schema or a snowflake schema.
A schema is a collection of database objects, including tables, views, indices and synonyms.

Star Schemas

A star schema is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists

of one or more fact tables and the points of the star are the dimension tables as shown in figure 3.1.

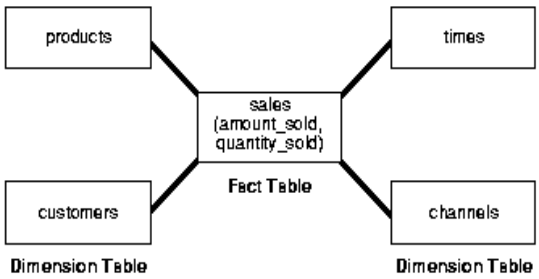


Figure 3.1 Star Schema
Source: Oracle9i: Data Warehousing Guide

The most natural way to model a data warehouse is as a star schema, only one join establishes the relationship between the fact table and any one of the dimension tables. A star schema optimizes performance by keeping queries simple and providing fact response time. All the information about each level is stored in one row.

Snowflake Schemas

A schema is called a snowflake schema if one or more dimension tables do not join directly to the fact table but must join through other dimension tables. For example, a dimension that describes products may be separated into three tables (snake flaked) as illustrated in figure 3.2.

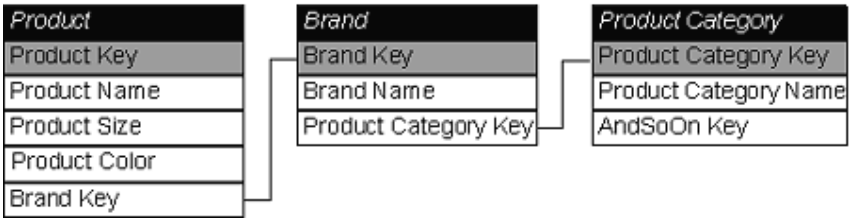


Figure 3.2 Many Dimension Snowflakes
Source: Data Warehouse Design Considerations by Dave Browning and Joy Mundy, Dec. 2001

Star and Snowflake Schemas

Both star and snowflake schemas are dimensional models; the difference is in their physical implementations. Snowflake schemas support ease of dimension maintenance because they are more normalized. Star schemas are easier for direct user access and often support simpler are more efficient queries. The decision to model a dimension as a star or snowflake depends on the nature of the dimension itself, such as how frequently it changes and which of its elements change, and often involves evaluating tradeoffs between ease of use and ease of maintenance.

Objects Used in Dimensional Data Warehouse Schemas

The two types of objects commonly used in dimensional data warehouse schemas are

- Fact tables
- Dimension tables

(i). **Fact Tables:** These are large tables in your warehouse schemas that stores business measurements. Fact tables typically contain facts and foreign keys to the dimension tables.

Fact tables represent data, usually numeric and additive, that can be analyzed and examined. Examples include sales, cost and profit.

A fact table basically has two types of columns: those containing numeric facts (often called measurements), and those that are foreign keys to dimension tables. It also contains aggregated facts that are often called *summary tables*. A fact table usually contains facts with the same level of aggregation. Though most facts are additive, they can also be semi-additive or non-additive. Additive facts can be aggregated by simple arithmetical addition; an example of this is sales. Non-additive facts cannot be added at all, an example is averages. Semi-additive facts can be aggregated along some of the dimensions and not along others. An example of this is inventory levels, where you cannot tell what a level means simply by looking at it

(ii). Dimension Tables

A dimension is a structure often composed of one or more hierarchies which categories data. Dimensional tables encapsulate the attributes associated with facts and separate these attributes into logically distinct groupings, such as time, geography, products, customers and so forth. They are normally descriptive, textual values and may be used in multiple places if the data warehouse contains multiple fact tables or contributes data to data marts. Commonly used dimensions are customers, products and time. This type of dimension that is often used in multiple schemas is called a **conforming dimension** if all copies of the dimension are the same.

Dimension data is typically collected at the lowest level of detail and aggregated into higher level totals that are more useful for analysis. These natural rollups or aggregations within a dimension table are called **hierarchies.**

Hierarchies

These are logical structures that uses ordered levels as a means of organizing 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.

Within a hierarchy, each level is logically connected to the levels above and below it. Data values at lower levels aggregate into the data values at higher level. A dimension can be composed of more than one hierarchy. For example, in the product dimension, there might be two hierarchies, one for product categories and one for product suppliers. Dimension hierarchies also group levels from general to granular. Query tools use hierarchies to enable you to drill down into your area to view different levels of granularity, which is one of the key benefits of a data warehouse. When designing hierarchies, you must consider the relationship in business structures, for example, a dimensional multilevel sales organization.

Levels: A level represents a position in a hierarchy. For example, a time dimension might have a hierarchy that represents data at the month, quarter, and year levels. Levels range from general to specific, with the root level as the highest or most general level.

Level Relationship: Level relationships specify top-to-bottom ordering of levels from most general (the root) to most specific information. They define the parent child relationship between the levels in a hierarchy

Typical Dimension Hierarchy

Figure 3.3 shows a dimension hierarchy based on customers

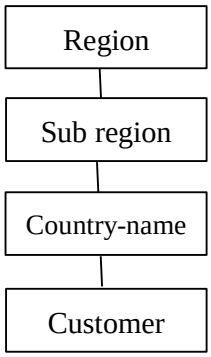


Figure 3.3 Typical Levels in a Dimension Hierarchy

- **Unique Identifiers** : Unique identifiers are specified for one distinct record in a dimension table. Artificial unique identifiers are often used to avoid the potential problem of unique identifiers changing. Unique identifiers are represented with the # character, for example, # customer_id.

- **Relationships**: Relationships guarantee business integrity. An example is that if a business sells something, there is obviously a customer and a product. Designing a relationship between the sales information in the fact table and the dimension tables products and customers enforce the business rules in databases. Figure 3.4 illustrates a common example of a sales fact table and dimension tables customers, products, promotions, times and channels

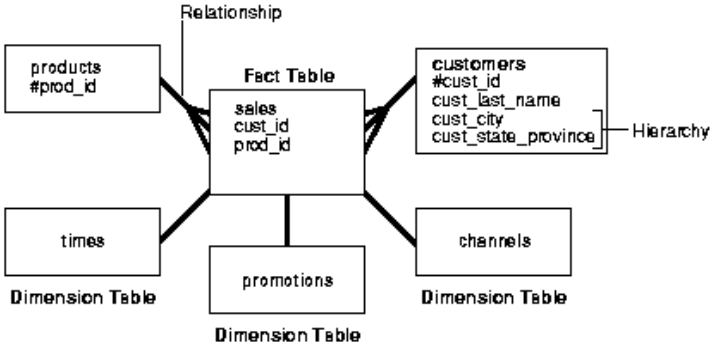


Figure 3.4 typical fact table and Dimension Tables

Source: Data Warehouse Design Considerations by Dave Browning and Joy Mundy, Dec. 2001

3.2.3 Develop the Architecture

The data warehouse architecture reflects the dimensional model developed to meet the business requirements. Dimension design largely determines dimension table design, and fact definitions determine fact table design.

Whether to create a star or snowflake schema depends more on implementation and maintenance considerations than on business needs. Information can be presented to the user in the same way regardless of whether a dimension is snow-flaked. Data warehouse schemas are quite simple and straight forward, in contrast to on-line transaction processing (OLTP) database schemas with their hundreds or thousands of tables and relationships. However, the quantity of data in data warehouses requires attention to performance and efficiency in their design.

Design for Update and Expansion

Data warehouse architectures must be designed to accommodate ongoing data updates, and to allow for future expansion with minimum impact on existing design. Providentially, the dimensional model and its straightforward schemas simplify these activities. Records are added to the fact table in periodic batches often with little effect on most dimensions. For example, a sale of an existing product to an existing customer at an existing store will not affect the product, customer, or store dimensions at all. If the customer is new, a new record is added to the customer dimension table when the fact record is added to the fact table. The historical nature of data warehouse means that records almost never have to be deleted from tables except to correct errors. Errors in source data are often detected in the extraction and transformation processes in the staging area and are corrected before the data is loaded into the data warehouse database.

The data and time dimensions are created and maintained in the data warehouse independent of the other dimension tables or fact tables, updating data and time dimension may involve only a simple annual task to mechanically add the records for the next year. Also the dimensional model lends itself to easy expansion. New dimension attributes and new dimensions can be added, usually without affecting existing schemas other than by extension. Existing historical data should remain unchanged. Data warehouse maintenance applications will need to be extended, but well-designed user applications should still function, though some may need to be updated to make use of the new information.

3.2.4 Design the Relational Database and OLAP Cubes

At this stage, the star or snowflake schema is created in the relational database, surrogate keys are defined (surrogate key is the primary key for a dimension table and is independent of any foreign keys provided by source data systems) and primary an foreign key relationships are established. Views, indexes, and fact table partitions are also defined. OLAP cubes are designed to support the needs of the users.

Key and Relationship

Tables are implemented in the relational database after surrogate keys for dimension tables have been defined, and primary and foreign keys and their relationships have been identified. Primary/foreign key relationships should be established in the database schema. For an illustration of these relationships see the star schema in Dimension Model Schema .

The composite primary key in the fact table is an expensive key to maintain:

- The index alone is almost as large as the fact table.
- The index on the primary key is often created as a clustered index.

In many scenarios a clustered primary key provides excellent query performance. However, all other indexes on the fact table use the large clustered index key. All indexes on the table will be large, the system will require significant additional storage space, and query performance may upgrade.

Due to these, many star schemas are defined with an integer, surrogate primary key or no primary key at all. Therefore, it is recommended that the fact table be defined using the composite primary key. Also, create an IDENTITY column in the fact table that could be used as a unique clustered index, should the database administrator determine this structure would provide better performance.

Indexes

Dimension tables must be indexed on their primary keys, which are the surrogate keys created for the data warehouse tables. The fact table must have a unique index on the primary key. There are scenarios where the primary key index should be clustered and other scenarios where it should not. The large the number of dimensions in the schema, the less beneficial it is to cluster the primary key index. With a large number of dimensions, it is usually more effective to create a unique clustered index on a meaningless IDENTITY column. Elaborating the initial design and development of index plans for end-user queries is not necessary with SQL server 2000, which has sophisticated index techniques and an easy to use index tuning wizard tool to tune indexes to query workload. The SQL server 2000 Index Tuning Wizard allows you to select and create an optimal set of indexes and statistics for a database without requiring an expert understanding of the structure of the database, the workload, or the intervals of SQL server. The wizard analyzes a query workload captured in a SQL Profiler trace or provided by an SQL script, and recommends and index configuration to improve the performance of the database.

The Index Tuning Wizard provides the following features and functionality:

- It can use the query optimizer to analyze the queries in the provided workload and recommend the best combination of index to support the query mix in the mix load.
- It analyzes the effects of the proposed changes, including index usage, distribution of queries among tables, and performance of queries in the work load.
- It can recommend ways to tune the database for a small set of problem queries
- It allows you to customize its recommendation by specifying advanced options, such as disk space constraints.

Views

Views should be created for users that need direct access to data in the warehouse relational database. Users can be granted access to views without having access to the underlying data. Indexed views can be used to improve performance of user queries that access data through views. View definitions should create column and table names that will make sense to business users. If analysis services will be the primary query engine to the data warehouse, it will be easier to create clear and consistent cubes from view with readable column names.

Design OLAP Cubes

An OLAP (Online Analytical Processing) cube is a data structure that allows fast analysis of data. It can also be defined as the capability of manipulating and analyzing data from multiple perspectives. The arrangement of data into cubes overcomes a limitation of relational databases.

OLAP cube design requirements will be a natural outcome of the dimensional model if the data warehouse is designed to support the way users want to query data. In a multidimensional database, a dimensional model is a cube. It holds data more like a 3-D spreadsheet rather than a traditional relational database. A cube allows different views of the data to be quickly displayed. The ability to quickly switch between one slice of data and another allows users to analyze their information in smaller meaningful chunks, at the speed of thought. Use of cubes allows the user to look at data in several dimensions, for example attendance by agency, attendance codes and attendance by date.

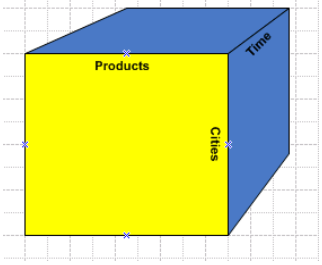


Figure 3.5 OLAP Cubes

Develop the Operational Data Store

Some business problems are best addressed by creating a database deigned to support tactical decision making. The Operational Data Store (ODS) is an operational construct that has elements of both data warehouse and a transaction system. Like a data warehouse, the ODS typically contains data consolidated from multiple systems and grouped by subject area. Like a transaction system, the ODS may be updated by business users, and contains relatively little historical data.

A classic business case for an operational data store is to support the Customer Call Center; call center operators have little need for broad analytical queries that reveal trends in customer behaviour. Rather, their needs are more immediate, the operation should have up to date information about all transactions that involve the complaining customer, this data may come from multiple source systems, but should be presented to the call center operator in a simplified and consolidated way.

The implementations of the ODS vary widely depending on business requirements. There are no strict rules for how the ODS must be implemented. A successful ODS for one business problem may be a replicated mirror of the transaction system; for another business problem a star schema will be most effective operational data stores fall between these two extremes, and include some level of transformation and integration of data. It is possible to architect the ODS so that it serves its primary operational need, and also functions as the close source for the data warehouse staging process.

3.2.5 Develop the Data Maintenance Application

The data maintenance applications, including extraction, transformation, and loading processes must be automated often by specialized applications. Data Transformation Services (DTS) which is SQL Server 2000 is a powerful tool for defining many transformations developed using scripting such as Microsoft Visual Basic Scripting Edition (VBScript) or Microsoft JScript, or languages such as Visual Basic. An extensive discussion on extraction, transformation and loading processes have been provided in Module 3, Unit 2.

3.2.6 Develop Analysis Applications

The applications that support data analysis by the data warehouse users are constructed in this phase of data warehouse development. OLAP cubes and data mining models are constructed using Analysis services tools and client access to analysis data is supported by the Analysis Server. The technique for cube design is covered in Module 3, Unit 2.

Other analysis applications such as Microsoft PivotTables, predefined reports, web sites, and digital dashboards are also developed in this phase as natural language applications using English Query. Specialized third-party analysis tools are also required and implemented or installed. Details of these specialized applications are determined directly by user needs.

3.2.7 Test and Deploy the System

It is important to involve users in the testing phase, after initial testing by development and test groups, users should load the system with queries and use it the way they intend to after the system is brought online. Substantial user involvement in testing will provide a significant number of benefits. Among the benefits are;

- Discrepancies can be found and corrected
- Users become familiar with the system
- Index tuning can be performed

It is important that users exercise the system during the test phase with the kinds of queries they will be using in production. This can enable a considerable amount of empirical index tuning to take place before the system comes online. Additional tuning needs to take place before the after deployment, but starting with a satisfactory performance is a key to success. Users who have participated in the testing and have seen performance continually improve as the system is exercised will be inclined to be supportive during the initial deployment phase as early issues are discovered and addressed.

3.3 The Data Warehouse Testing Life Cycle

Just like any other piece of software, a data warehouse implementation undergoes the natural cycle of unit testing, system testing, regression testing, integration testing and acceptance testing. Although as with other software, there are no off-the-shelf testing products available for a data warehouse.

(i). **Unit Testing**

Traditionally this has been the task of the developer. It is a white-box testing to ensure the module or component is coded as per agreed upon design specifications. The developer should focus on the following:

(a). That all inbound and outbound directory structures are created properly with appropriate permissions and sufficient disk space. All tables used during the ETL are present with necessary privileges.

(b). The ETL routines give expected results:

- All transformation logics work as designed from source till target.
- Boundary conditions are satisfied e.g. check for data field with leap year dates.
- Surrogate keys have been generated properly
- Null values have been populated where expected
- Rejects have occurred where expected and log for, reject is created with sufficient details.
- Error recovery methods

- Auditing is done properly.

(c). That the data loaded into the target is complete:

- All source data that is expected to get loaded into target, actually get loaded, compare counts between source and target and use data profiling tools.
- All fields are loaded with full contents i.e. no data field is truncated while transforming.
- No duplications are loaded
- Aggregations take place in the target properly
- Data integrity constraints are properly taken care of

(ii). System Testing

This is the responsibility of the quality control team (QC). Here we test for the functionality of the application and mostly it is black-box. The major challenge here is preparation of test data. An intelligently designed input dataset can bring out the flows in the application more quickly. Wherever possible use production-like data, you may also use data generation tools or customized tools of your own to create test data. We must test for all possible combinations of input and specifically check out the errors and exception. An unbiased approach is required to ensure maximum efficiency. Knowledge of the business process is an added advantage since we must be able to interpret the results functionally and not just code-wise.

The QA team must test for:

- Data completeness: match source to target counts.
- Data aggregations: match aggregated data against staging tables and/or ODS
- Granularity of data is as per specifications
- Error logs and audit tables are generated and populated properly
- Notifications to IT or business are generated in proper format

(iii). Regression Testing

A data warehouse is not one-time solution. Possibly it is the best example of an incremental design where requirements are enhanced and refined quite often based on business needs and feedbacks. In such a situation it is very critical to test that the existing functionalities of a data warehouse (DW) application are not messed up whenever an improvement is made to it. In general, this is done by running all functional tests for existing code wherever a new piece of code is introduced. However, a better strategy could be to preserve earlier test input data and result sets and running the same again. Now the new results could be compared against the older ones to ensure proper functionality

(iv). Integration Testing

This is done to ensure that the application developed works from an end-to-end perspective. Here we must consider the compatibility of the data warehouse application with upstream and downstream flows. We need to ensure for data integrity across the flow. Our test strategy should include testing for:

- Sequence of jobs to be executed with job dependencies and scheduling
- Re-start ability of jobs in case of failures
- Generation of error logs
- Cleaning scripts for the environment including database.

This activity is a combined responsibility and participation of experts from all related application is a must in order to avoid misinterpretation of results.

(v). Acceptance Testing

This is the most critical part because here the actual users validate your output datasets. They are the best judges to ensure that the application works as expected by them. However, business users may not have proper ETL knowledge. Hence, the development and test team should be ready to provide answers regarding ETL process that relate to data population. The test team must have sufficient business knowledge to translate the results in terms of business. Also, the load windows refresh period for the data warehouse and the views created should be signed off from users.

(vi). Performance Testing

It is very necessary for a data warehouse to go through another phase of testing called performance testing. Any data warehousing application is designed to be scalable and robust. Therefore, when it goes into production environment, it should not cause performance problems. Here, we must test the system with huge volume of data. We must ensure that the load window is met even under such volumes. This phase should involve DBA team, ETL expert and others who can review and validate your code for optimization.

Testing a data warehouse application should be done with a sense of utmost responsibility. A bug in a DW traced at a later stage results in unpredictable losses. It should be remembered tester must go an extra mile to ensure near defect free solutions.

Activity B: Student Self Assessment Exercise

Briefly explain the following data warehouse testing life cycle

(i).	Unit testing	(ii).	System testing
(iii).	Regression testing	(iv).	Integration testing
(v).	Acceptance testing	(vi).	Performance testing

4.0 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.

5.0 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 system.
- The implementation of data warehouse undergoes the natural cycle of unit testing, system testing, regression testing, integration testing and acceptance testing.

6.0 Tutor Marked Assignment

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

7.0 Further Reading and Other Resources

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

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

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

Lean, A. and Lean, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House P

Hans-Peter K., Karasten M.B, Peer K., Alexey P. Matthias S. and Arthur Z. (March, 2007) *Future Trends in Data Mining.* Springer Science + business Media, 23 March 2007

Jayaprakash, P, Joseph Z., Berkin O. and Alok C., *Accelerating Data mining Workloads: Current Approaches and Future Challenges in System Architecture Design*

OLAP Cube. Retrieved on 03/11/2009. Available Online: http://en.wikipedia.org /wiki/OLAP_cube

Oracle9*i* Data Warehousing Guide Release 2 (9.2)

Module 3: Data Warehousing Concepts

Unit 4: Data Warehouse and OLAP Technology

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:

- Understand the meaning of OLAP

- Differentiate between OLAP and data warehouse
- Know the different types of OLAP server
- Describe OLAP as a data warehouse tool and its applications
- Understand the open issues in data warehouse

3.0 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 analyze 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.

3.1 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 modeling. As decision-makers exercise more advanced OLAP capabilities, they move from data access to information and to knowledge.

3.1.1 Benefits of OLAP

Some of the benefits derived from the applications of OLAP systems are as follows:
(i). 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.

(ii). 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 analyzing a huge amount of data. Moreover, it helps to realize relationships that were forgotten earlier.

(iii). 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.

(iv). Through the use of OLAP, ICT realizes more efficient operations by using software designed for OLAP, ICT reduces the query drag and network traffic on transaction systems or the data warehouse.

(v). By providing the ability to model real business problems and a more efficient use of people resources, OLAP enables the organization as a whole to respond more quickly to market demands. Market responsiveness, in turn often yields improved revenue and profitability.

Activity A: Student Self Assessment Exercise

State some of the benefits derived from the applications of OLAP systems

3.2 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 characterized by dynamic multi-dimensional analysis of consolidated enterprise data supporting end-user analytical and navigational activities including the following:

- Calculations and modeling applied across dimensions, through hierarchy and/or across member
- Trend analysis over sequential time periods
- Slicing subsets for on-screen viewing
- Drill-down to deeper levels of consolidation
- Reach-through to underlying detail data
- 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 synthesize enterprise information through comparative, personalized 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 optimized 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.

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,

1. **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 optimization 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 display 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.

2. **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 recognize 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-summarized and are stored in an optimized 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 optimized for fast query performance and retrieval of summarized 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. Some certain MOLAP products encounters difficulty in updating models with dimensions of very high cardinality.

3. 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 stusmble on so often. We have listed some of the less famous existing in the OLAP industry.

(i). 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.

(ii). 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.

(iii). 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.

(iv). 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 OLAP as a Data warehouse Tool

On-line analytical processing (OLAP) is a technology designed to provide superior performance for business intelligence queries. OLAP is designed to operate efficiently with data organized in accordance with the common dimensional model used in data warehouse.

A data warehouse provides a multidimensional view of data in an intuitive model designed to match the types of queries posed by analysts and decision makers. OLAP organizes data warehouse data into multidimensional cubes based on this dimensional model, and then preprocesses these cubes to provide maximum performance for queries that summarize data in various ways. For example, a query that request that total sales income and quantity sold for a range of product in a specific geographic region for a specific time period can typically be answered in a few second or less regardless of how many millions of rows of data are stored in the data warehouse database. OLAP is not designed to store large volumes of text or binary data, nor is it designed to support high volume update transactions. The inherent stability and consistency of historical data in a data warehouse enables OLAP to provide its remarkable performance in rapidly summarizing information for analytical queries. In SQL server 2000, Analysis Services provides tools for developing OLAP applications and a server specifically designed to service OLAP queries.

3.4 Uses of OLAP

The applications of OLAP spans over a variety of organizational functions. Finance departments use OLAP for applications such as budgeting, activity-based costing (allocations), financial performance analysis, and financial modeling. Sales analysis and forecasting are two of the OLAP applications found in sales departments. Also, marketing departments use OLAP for market research analysis, sale forecasting, analysis, customer analysis, and market/customer segmentation. Typical manufacturing OLAP applications include production planning and defect analysis.

3.5 Open Issues in Data Warehousing

Data warehousing which is an active research area is likely to encounter increased research activity in the near future as warehouse and data mart proliferate. Old problems will receive new emphasis; for example, data cleaning, indexing, partitioning and views could receive renewed attention.

Academic research into data warehousing technologies will likely focus on automating aspects of the warehouse, such as the data acquisition, data quality management, selection and construction of appropriate access path and structures, self-maintainability, functionality and performance optimization. Incorporation of domain and business rules appropriately into

the warehouse creation and maintenance process may take intelligent, relevant and self governing.

Activity B: Student Self Assessment Exercise

Differentiate between the following pairs:
(i). OLAP and data warehouse
(ii). OLAP and OLAP server

4.0 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.

5.0 Summary

In this unit we have learnt that:

- OLAP is a technology that allows users of multidimensional databases to generate on-line descriptive or comparative summaries of data and other analytical queries.
- 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
- 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
- OLAP as a data warehouse tool can be used to provide superior performance for business intelligence queries and to operate efficiently with data organized in accordance with the common dimensional model used in data warehouse.
- Some of the open issues in data warehousing, these include increased research activity in the near future as warehouse and data mart proliferation.

6.0 Tutor Marked Assignment

1. (a). What do you understand by the term OLAP?
   (b). List and explain the three major types of OLAP

2. State some application areas of OLAP

7.0 Further Reading and Other Resources

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

Jayaprakash, P, Joseph Z., Berkin O. and Alok C., *Accelerating Data mining Workloads: Current Approaches and Future Challenges in System Architecture Design*

OLAP Cube. Retrieved on 03/11/2009. Available Online: http://en.wikipedia.org /wiki/OLAP_cube

Oracle9*i* Data Warehousing Guide Release 2 (9.2) Activity

Anil Rai, *Data Warehouse and its Applications in Agriculture* , Indian Agriculture Statistics Research Institute Library Avenue, New Delhi-110 012.

Laura Hadley, *Developing a Data Warehouse Architecture* , Retrieved on 13/10/2009. From: http://www.users.qwest.net/¬lauramh/resume/thorn.htm.

Understanding the Data Warehouse Architecture, Retrieved on: 13/10/2009. From: http://msdn.microsoft.com/en-us/library/ms244687(VS.80).aspx

Data Warehouse, Retrieved on 29/09/2009. From http://en.wikipedia.org/wiki/Data-warehouse.

Data Warehouse. Retrieved on 29/09/2009. From http://www.intranetjournal.com/features/datawarehousing.html.

Muhammad A. Shahzad, Data Warehousing With Oracle. www.oracular.com

What is OLAP?. Retrieved on 13/10/2009. Available online: http://www.tech-faq.com/olap.shtml.

Types of OLAP Systems. Retrieved on 13/10/2009. Available online: http://www.olap.com/w/index.php/Types of OLAP Systems.

OLAP and OLAP Server Definitions. Retrieved on 13/10/2009. Available online: http://www.moulton.com/olap.glossary.html

Dave Browning and Joy Mundy, (Dec., 2001). *Data Warehouse Design Considerations.* Retrieved on 13/10/2009. Available Online: hhtp://msdn.microsoft.com/en-us/library/aa902672(SQL.80)aspx.

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

Leon, A. and Leon, M. (1999), *Fundamentals of Information Technology* . New Delhi: Leon Press Channel and Vikas Publishing House Pvt Ltd.