DAM301 DATA MINING & DATA WAREHOUSING SUMMARY

Q. What do you understand by the term data mining?

Definition of Data Mining

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.

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. 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.

Q. Differentiate between data mining and knowledge discovery in databases (KDD)

Data Mining and Knowledge Discovery in Databases (KDD)

The term data mining is mostly employed by data analysts MIS specialties, statisticians and database administrators, while Knowledge Discovery in Databases (KDD) refers to the overall process of discovering useful knowledge from data; although, data mining and knowledge discovery in databases (KDD) are frequently treated as synonyms.

The term KDD was first coined by Gregory Piatetsky-Shapiro in 1989 to describe the process of searching for interesting, interpreted, useful and novel data. Reflecting the conceptualisation 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.

Q. List and define seven (7) steps involved in Knowledge Discovery Databases (KDD) process.

KDD being an iterative process consists of the following steps:

1. 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.

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

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

4. 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.

5. 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.

6. Pattern Evaluation: At this stage, patterns that are very interesting and represent knowledge are identified based on given measures.

7. Knowledge Representation: This is the final stage of the KDD process in which the discovered knowledge is visually represented to the user. Visualisation techniques are used to assist the users to have a better understanding and interpret the data mining results.

Q. Differentiate between data mining and OLAP.

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 analyse 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 analysed becomes voluminous it becomes much more difficult and time-consuming to find a good hypothesis, analyse 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.

Q. State the evolution of data mining

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 computerised 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 mature:

I. Massive data collection
II. Powerful multiprocessor computer

Presently commercial databases are growing at an unprecedented rate. In some organisations, 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.

Q. 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.

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:

1. Automated Prediction of Trends and Behaviours: 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 maximise 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.

2. 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 analyse 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 analyse huge quantities of data. Larger databases in turn yield improved predictions.

Q. Identify the architecture of data mining

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 analysing 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 organisation, in areas such as promotional campaign management, fraud detection, and new product rollout and so on.

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 optimised 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 analyse the data as they want to view their business-summarising 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 optimisation. 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 organisation 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).

Q. Briefly explain how does mining methodology works?

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 modelling. Modelling 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.

Q. Identify the different kinds of information collected in our databases

The Types of Information Collected

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

(3) 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 digitise them for future use and analysis.

(7) Satellite Sensing

08024665051

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 analyse them.

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

Most of communications within and between individuals, research organisations 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 interconnected 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.

Q. Briefly explain the types of data that can be mined

Types of Data Mined

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, timesales data, scientific measurements etc.

2. Relational Databases

his 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.

3. Data Warehouses

A data warehouse (a storehouse) is a repository of data gathered from multiple data sources (often heterogeneous) and is designed to be used as a whole under the same unified schema. A data warehouse provides an option of analysing data from different sources under the same roof. The most efficient data warehousing architecture will be

08024665051

able to incorporate or at least reference all management systems using designated technology suitable for corporate database management e.g. Sybase, Ms SQL Server.

4. 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.

5. Spatial Databases

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

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 characterised 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.

7. Time-Series Databases

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

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 organised 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.

Q. What do you understand by data mining functionalities?

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

08024665051

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.

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

Data Mining Functionalities

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.

2. Characterisation

Data characterisation is also called summarisation 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 summarisation 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 visualisation techniques and the discovery of functional relationships between variables. Summarisation techniques are often applied to interactive exploratory data analysis and automated report generation (Usama et al., 1996)

3. Clustering

Clustering is similar to classification and is the organisation 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 maximising the similarity between objects in the same class (that is intra-class similarity) and minimising the similarity between objects of different classes that is inter-class similarity.

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.

5. Discrimination

Data discrimination generates what we call discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For instance, we may want to characterise the rental customers that regularly rent more than 50 movies last year with those whose rental account is lower than 10. The techniques used for data discrimination are similar to that used for data characterisation with the exception that data discrimination results include comparative measures.

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.

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.

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 characterising, 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.

Q. Identify the various classifications of data mining systems

Classification of Data Mining Systems

Data mining systems can be categorised 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 categorises 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 categorises data mining systems according to the kind of knowledge discovered or data mining functionalities such as discrimination, characterisation, association, clustering etc. Some systems tend to 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 categorises data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualisation, database-oriented or data warehouse-oriented.

Q. Describe the categories of data mining tasks

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 email as legitimate or spam. Common algorithms to use are nearest neighbour, 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."

Q. State the diverse issues coming up in data mining

08024665051

Data Mining Issues

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 of 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.

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 analysed. 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.

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 visualisation 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.

The major issues related to user interface and visualisations 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.

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 we differentiate between what data is important and what data is insignificant.

5. Performance Issues

08024665051

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.

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.

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.

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 feel obligated to make any information available that could be used to prevent a future attack or track a known terrorist

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.

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

Data Mining Challenges

(1) Larger Databases

Databases with hundreds of fields and tables containing millions of records of a multi-gigabyte size are very prevalent, and terabyte (1012 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 identify 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 identify 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 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.

(5) 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, regularisation, and other sophisticated statistical strategies.

Q. Explain what is meant by Data Mining Technologies

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

08024665051

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 customised to meet individual vendor's goal. For instance, most of the vendors sell versions of the CART (Classification and Regression Trees) or CHAID (Chi-Squared Automatic Interaction Detection) decision trees with enhancements to work on parallel computers, while some vendors have proprietary algorithms that will not allow extension or enhancements of any published approach to work well.

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

Q. Identify the various data mining technologies available. Q. Extensively discuss the following data mining techniques: a) Neural networks b) Multivariate Adaptive Repression Splines (MARS)

Data Mining Technologies

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:

a. Sales forecasting
b. Customer research
c. Data validation
d. Risk management
e. Industrial process control
f. 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.

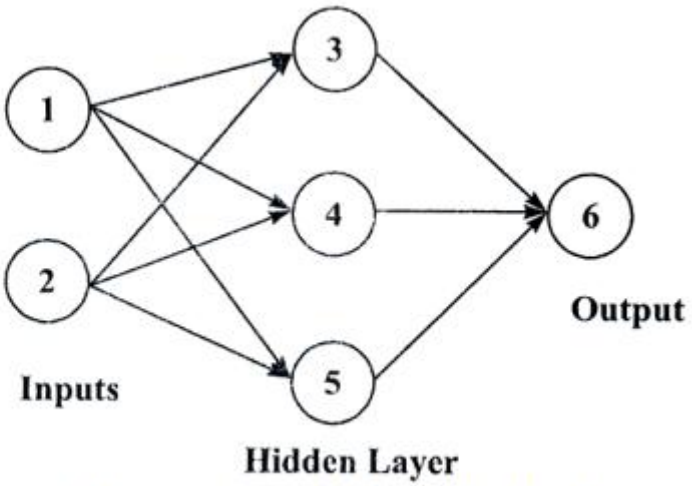


Fig. 4.1: *Neural Network with One Hidden Layer*

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

i. 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).

ii. 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.

Q. What is a Decision tree?

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 nonnumerical data very well.

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.

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.

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

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

5. K-Nearest Neighbour and Memory-Based Reasoning (MBR)

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

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

08024665051

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 optimising models, but add a lot of computational over head.

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.

8. Generalised Additive Models (GAM)

Generalised 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 nonlinear 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.

With the use of computer power in place of theory or knowledge of the functional form, GAM will produce a smooth curve and summarise 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.

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 first published by Freund and Shapire in 1996; it takes multiple random samples from the data and builds a classification model for each. The

08024665051

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.

10. Logistic Regression (Non Linear Regression Methods)

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

Q. Explain the meaning and importance of data preparation

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

Q. Identify the different data formats of an attribute

Data Types and Forms

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

08024665051

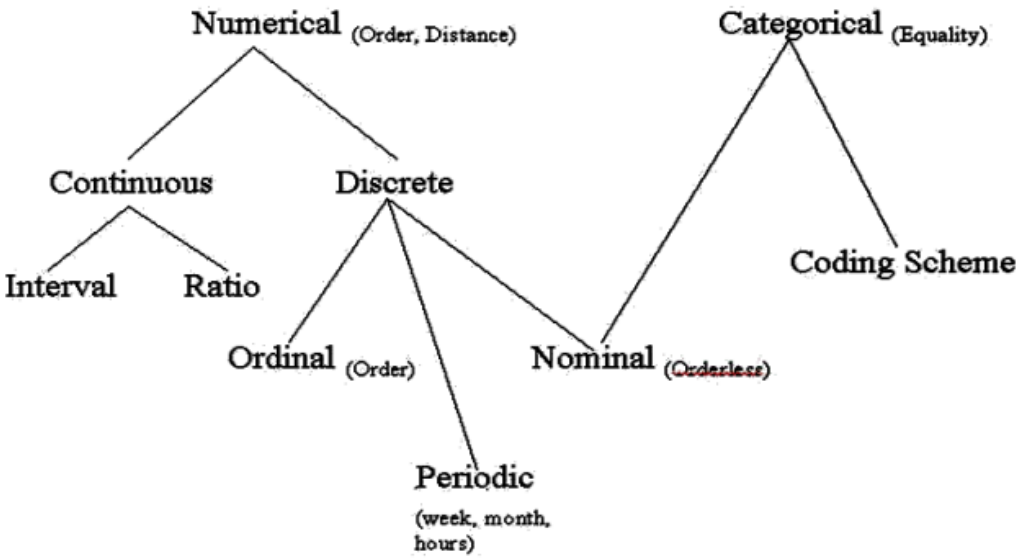


Fig.5.1: *Different Data Formats of an Attribute*

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

I. Distributed data
II. Textual data
III. Web data (e.g. html pages)
IV. Images
V. Audio /Video
VI. Metadata (information about the data itself)

Q. List and explain some data preparation methods

Data Preparation

The common types of data preparation methods are:

I. Data normalisation (e.g. for image mining)
II. Dealing with sequential/temporal data
III. Removing outliers

1. Data Normalisation

The different types of data normalisation methods are:

1. Decimal Scaling: This type of scaling transforms the data into a range between (-1, 1). The transformation formula is v'(i)/10k. 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 Normalisation: This type of normalisation 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

ρανγε [0,1]. ⇒For example: - If v = 73600 in [12000, 98000]

ϖɜ= 0.716 ιν τηε νεω

3. Zero – Mean Normalisation: When you use this type of normalisation, 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 1.225.ϖɜ= ⇒std_devIncome = 16000, then v = 76000

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. Removing Outliers

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

Clustering: This can be used to cluster the relevant data points together and then use those cluster centers to find out the data points not close enough to them and then reject them as outliers.

Curve–Fitting: This method initially uses regression analysis to find the curves which fit the data closely. It then removes all points (outliers), which are sufficiently far curve from the fitted curve

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

Q. Define term data preprocessing

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 world data tends to be incomplete, noisy inconsistent, high dimensional and multi-sensory etc. hence are not directly suitable for

08024665051

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.

Q. State the various data pre-processing tasks

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 analysed 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.

Q. Explain why data is being preprocessed

The Reasons for Data Preprocessing

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

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

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

08024665051

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

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

Q. Explain "why Data Preprocessing"?

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.

Q. List some of the factors used in measuring the quality of a data.

Data Quality Measures

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

I. Accuracy
II. Completeness
III. Consistency
IV. Timeliness
V. Believability
VI. Interpretability
VII. Accessibility

Q. Briefly discuss the different types of data pre-processing tasks

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 be due to:

- Attribute not considered important
- Misunderstanding at data entry
- Inconsistent with other data and thus deleted

08024665051

- Equipment malfunction e.g. EFTPOS down – use cash
- (EFTPOS- Electronic Fund Transfer Point-Of-Sale).

(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)
- 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.

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 (summarisation, data cube construction)
- Generalisation: Moving up in the concept hierarchy on nominal attributes i.e. replacing data with higher level concepts (e.g. city)◊address details
- Attributes Construction: Replacing or adding new attributes inferred by existing attributes.

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.

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).

5. Discretisation and Concept Hierarchy Generation

1. Reduce the number of values for a continuous attribute by dividing the range into intervals.

08024665051

2. Concept hierarchies for numerical attributes can be constructed automatically.

3. Binning (smoothing, distributing values into bins, then replace each value with mean, median or boundaries of the bin)

4. 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

5. Segmentation by natural partitioning (partition into 3,4, or 5 relatively uniform intervals)

6. Entropy (information) – based discretisation

6. Data Parsing and Standardisation

1. Parse free format data into specific, well–defined attributes

2. Standardise using rules and look-up tables (correction and replacement tables), or probabilistically (hidden Markov models)

3. Important for data linkage (based on names, addresses etc).

Q. State the basic steps of data mining for knowledge discovery

Building a Data Mining Database

The various tasks in building a data mining database are:

1. Data Collection

There is need to identify the sources of the data you want to mine, though a data-gathering phase may become very necessary because some of the data you need may never have been collected. Also, you may need to acquire external data from public databases (such as census or whether data) or proprietary databases (such as credit bureau data). A data collection report (DCR) lists the properties of different source data sets. Some of the elements in this report include the following:

- Source of data (either internal application or outside vendor)
- Owner
- Person/organisation responsible for maintaining the data
- Database administration (DBA)
- Cost (if purchased)
- Storage organisation (oracle database, VSAM file etc)

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:

08024665051

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

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.

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.

5. Integration and Consolidation

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 (hom

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.

7. Load the Data Mining Database

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

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 reorganisation to reclaim disk storage or to improve performance for a large and complex database stored in a

08024665051

DBMS, the maintenance may also require the services of information systems professionals.

Q. Enumerate the shortcomings of CART

1. Discontinuous predictions (hard splits)
2. Dependence of all splits on previous ones
3. Reduced interpretability due to interactions, especially high-order interactions.

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

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 analyse 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.

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.

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.

3. Data Mining Applications in Telecommunications

The telecommunications industry has undergone one of the most dramatic makeovers of any industry. These industries generate and store 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.

4. Data Mining Applications in Healthcare

08024665051

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.

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.

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 analyse its own customer experience, this company can build a unique segmentation to identify attributes of highvalue prospects. Applying this segmentation to a general business database can yield a prioritised list of prospects by regions.

Q. Briefly explain the following applications of data mining in surveillance:

(a) Terrorism Information Awareness (TIA)

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:

I. Language translation
II. Data search with pattern recognition and privacy protection
III. Advanced collaborative and decision support tools

(b) Computer-Assisted Passenger Prescreening System (CAPPS)

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.

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

(a) 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 visualisation 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 organisation possessing huge databases with thematic and geographically referenced data begin to realise the huge potential of the information hidden there.

(b) 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-Organising 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.

(c) 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 optimise 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.

(d) Telecommunication

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.

(2) Marketing/ Customer Profiling: Telecommunication industries maintain a great 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 analysed 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.

Q. Briefly explain the following data mining techniques used for hypertext and hypermedia data mining:

i. 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 modelling, and maximum entropy.

ii. 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 organise 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.

iii. 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.

iv. 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.

Q. Briefly explain the following categories of constraint-based data mining:

(i) 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.

(ii) 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.

(iii) Rule 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.

(iv) Dimension/Level Constraints

Dimension constraints specifies the specific rules which should be applied and used for a particular data mining query or application.

Q. Briefly discuss the following data mining trends in terms of technologies and methods:

1. Ubiquitous data mining

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 analyse 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 minimising the cost of ubiquitous presence. Another challenging aspect of UDM is the human-computer interaction.

2. 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

08024665051

multimedia data mining; although, this field is rather new, but holds much promise for the future.

3. Hypertext and hypermedia

Hypertext and hypermedia data mining can be characterised 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.

4. 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 analysed 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.

Q. Define the term data warehouse

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. A data warehouse is a data structure that is optimised 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 organisation's electronically stored data designed to facilitate reporting and analysis.

Q. State the goals and characteristics of data warehouse

Goals of Data Warehouse

The major goals of data warehousing are stated as follows:

1. To facilitate reporting as well as analysis

2. Maintain an organisations historical information

08024665051

3. Be an adaptive and resilient source of information

4. Be the foundation for decision making.

Q. What are the characteristics of a data warehouse as set forth by William Inmon?

1. Subject–oriented
2. Integrated
3. Nonvolatile
4. Time variant

Characteristics of Data Warehouse

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

(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.

(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'.

Evolution in Organisational Use of Data Warehouses

Data warehousing which is a process of centralised data management and retrieval, just like data mining it is a relatively new term, although the concept itself has been around for years. Organisations 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:

08024665051

i. 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.

ii. 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.

iii. 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.

iv. 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.

Q. What are the advantages and disadvantages of implementing a data warehouse?

Advantages of Data Warehouse

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

(i) Facilitate Decision–Making: 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 analyse 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 safely for extended periods of time.

(v) 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.

Disadvantages of Data Warehouses

(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 organisation.

(v) Because there is often a fine line between data warehouse and operational system, duplicate and expensive functionality may be developed.

Q. List the major components of data warehouse

Data Warehouse Components

The major components of a data warehouse are:

1. Summarised data

2. Operational systems of record

3. Integration/Transformation programs

4. Current detail

5. Data warehouse architecture or metadata

6. Archives

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

i. Summarised data

Summarised data is classified into two namely: Lightly summarised data and Highly summarised data

a. **Lightly summarised 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 customised lightly summarised data for every enterprise elements. An enterprise element may have access to both detailed and summarised data, but there will be much less data than the total stored in current detail.

b. Highly summarised data are primarily for enterprise executives. It can come from either the lightly summarised 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 summarised data, executives also have the capability of accessing increasing levels of detail through a "drill down" process.

ii. Operational systems of record

08024665051

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.

iii. Integration program

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:

1. Reformatting, recalculating, or modifying key structures.

2. Adding time elements.

3. Identifying default values

4. Supplying logic to choose between multiple data sources

5. Summarising, tallying and merging data from multiple sources.

iv. 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 organised 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.

v. Metadata

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.

vi. 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.

Q. State the structure and approaches to storing data in data warehouse

Structure of a Data Warehouse

1. 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, organising, packaging and processing the detail data.

2. Logic Data Warehouse: It also contains metadata enterprise rules and processing logic for scrubbing, organising, packaging and processing the data, but does not contain actual data. Instead, it contains the information necessary to access the data wherever they

08024665051

reside. This structure is effective only when there is a single source for the data and they are known to be accurate and timely.

3. Data Mart: this is a data structure that is optimised 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).

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

Approaches for Storing Data in a Warehouse

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

I. The dimensional approach
II. The normalised approach

1. 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 organisation adopting the dimensional approach changes the way in which it does business.

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

Benefits of Normalised Approach

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

Disadvantages of Normalised Approach

1. 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.

Q. Describe the users and application areas of data warehouse

Data Warehouse Users

1. Statisticians: There are usually a handful of sophisticated analysts comprising of statisticians and operations research types in any organisation. 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.

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.

Applications of Data Warehouse

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

1. Credit card churn analysis

08024665051

2. Insurance fraud analysis

3. Call record analysis

4. Logistics management

Q. What do you understand by the term data warehouse architecture?

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.

Q. State the three types of data warehouse architecture.

The Types of Data Warehouse Architectures

1. Data Warehouse Architecture (Basic)

Endusers directly access data derived from several source systems through the data warehouse**.** 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 materialised view.

2. Data Warehouse Architecture (with a staging Area)

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.

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

Even though, the architecture is quite common, you may want to customise your warehouse's architecture for different groups within your organisation. 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 analyse historical data for purchases and sales.

Q. List and briefly explain the seven major components of data warehouse architecture.

Components of Data Warehouse Architecture

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.

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:

a. Extraction
b. Transformation
c. Loading and Indexing

3. Data Warehouse

Database A data warehouse database is a relational data structure that is optimised 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.

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 modelling 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 analyse 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.

08024665051

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.

Q. explains the use of extraction, transformation and load tools

Extraction, Transformation and Load

1. 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.

2. Transformation

Data is transformed from transaction level data into information through several techniques: filtering, summarising, 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.

3. Data Cleansing

Cleansing data is based on the principle of populating the data warehouse with quality data, which is consistent data, which is of a known, recognised 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 standardises 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.

Q. Describe what is meant by resource management

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.

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

Data Warehouse Design Methodologies

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

a) Bottom-up Design

b) Top-down Design

c) Hybrid Design

(i) Bottom-up Design

Ralph Kimball, a well-known author on data warehousing is a proponent of an approach frequently considered as bottom-up to data warehouses design. In this approach smaller local data warehouse, known as data marts are firstly created to provide reporting and analytical capabilities for specific business processes. Data marts contain atomic data and, if necessary, summarised data. These data marts can eventually be merged together to create a comprehensive data warehouse. The combination of data marts is managed through the implementation of what Kimball calls "data warehouse bus architecture". Business value can be returned as quickly as the first data marts can be created. Maintaining tight management over the data warehouse bus architecture is fundamental to maintaining the integrity of the data warehouse.

(ii) Top–down Design

The top-down design methodology generates highly consistent dimensional views of data across data marts since all data marts are loaded from centralised repository. Top-down design has also proven to be robust against business changes. Also, the top-down methodology can be inflexible and indifferent to changing departmental needs during the implementation phases.

(iii) Hybrid Design

Over time it has become apparent to proponents of bottom up and topdown 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.

Q. Briefly explain the following data warehouse testing life cycle:

The Data Warehouse Testing Life Cycle

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

08024665051

(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
- Error recovery methods
- Auditing is done properly

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

- No duplications are loaded
- All fields are loaded with full contents i.e. no data field is truncated while transforming
- 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 customised tools of your own to create test data.

iii. Regression testing

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.

v. Acceptance testing

08024665051

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 optimisation.

Q. Differentiate between a logical design and physical design

Logical Versus Physical Design of Data Warehouses

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.

Q. What do you understand by schema?

A schema is a collection of database objects, including tables, views, indices and synonyms.

Q. In the context to a data warehouse, describe the following:

(i) Star schema

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.

(ii) Snowflake schema

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

08024665051

dimension that describes products may be separated into three tables (snake flaked) as illustrated in figure 3.2.

Q. Briefly describe the two types of objects commonly used in dimensional data warehouse: (i) Fact tables (ii) Dimension tables.

Objects Used in Dimensional Data Warehouse Schemas

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

(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 analysed 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.

(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.

Features and Functionality of Index Tuning Wizard

The Index Tuning Wizard provides the following features and functionality:

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

Q. State the meaning of OLAP

Meaning of On-Line Analytical Processing (OLAP)

The term On-Line Analytical Processing, OLAP (or Fast Analysis of Shared Multi-dimensional Information –FASMI) refers to the technology that allows users of

multidimensional databases to generate on-line descriptive or comparative summaries (i.e. views) of data and other analytical queries.

OLAP was coined in 1993 by Tedd. Codd who is referred to as "the father of the relational database'' as a type of application that allows users to interactively analyse data. An OLAP system is often contrasted to an On-Line Transaction processing (OLTP) system that focuses on processing transaction such as orders, invoice or general ledger transactions. Before OLAP was coined, these systems were often referred to as Decision Support Systems (DSS).

OLAP is now acknowledged as a key technology for successful management in the 90's. It further describes a class of applications that require multidimensional analysis of business data. OLAP systems enable managers and analysts to rapidly and easily examine key performance data and perform powerful comparison and trend analyses, even on very large data volumes.

Q. Differentiate between OLAP and data warehouse

OLAP and Data Warehouse

A data warehouse is usually based on relational technology, while OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for further analysis.

OLAP enables analysts, managers and business executives to gain insight into data through fast, consistent and interactive access to a wide variety of possible views of information. Also, OLAP transform raw data so that it reflects the real dimensionality of the enterprise as understood by the user. In addition, OLAP systems have the ability to answer "what if?" and why?" that sets them apart from data warehouses. OLAP enables decision making about future actions. A typical OLAP calculation is more complex than simply summing data.

OLAP and data warehouse are complementary. A data warehouse stores and manages data. OLAP transform data warehouse data into strategic information. OLAP ranges from basic navigation and browsing (this is often referred to as "slice and dice"), to calculations, to more serious analyses such as time series and complex modelling. As decisionmakers exercise more advanced OLAP capabilities, they move from data access to information and to knowledge.

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

The 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.

(ii) OLAP allows the manager to tear down data from OLAP database in specific or broad terms.

08024665051

(iii) OLAP helps to reduce the applications backlog still further by making business users self sufficient enough to build their own models.

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

(v) By providing the ability to model real business problems and a more efficient use of people resources, OLAP enables the organisation as a whole to respond more quickly to market demands.

Q. List the different types of OLAP server

I. Relational OLAP (ROLAP) servers
II. Multidimensional OLAP (MOLAP)
III. Hybrid OLAP (HOLAP)
IV. Web OLAP (WOLAP)
V. Desktop OLAP (DOLAP)
VI. Mobile OLAP (MOLAP)
VII. Spatial OLAP (SOLAP)

Q. Describe OLAP as a data warehouse tool and its applications

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 organised 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 organises data warehouse data into multidimensional cubes based on this dimensional model, and then preprocesses these cubes to provide maximum performance for queries that summarise 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 summarising 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.

Q. Identify the open issues in data warehouse

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.

08024665051

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 optimisation. Incorporation of domain and business rules appropriately into the warehouse creation and maintenance process may take intelligent, relevant and self governing.

08024665051