DATA WAREHOUSING AND MINING

1. What is Data Warehousing?

A Datawarehouse is the repository of a data and it is used for Management decision support system. Datawarehouse consists of wide variety of data that has high level of business conditions at a single point in time.

In single sentence, it is repository of integrated information which can be available for queries and analysis.

2. What is Business Intelligence?

Business Intelligence is also known as DSS – Decision support system which refers to the technologies, application and practices for the collection, integration and analysis of the business related information or data. Even, it helps to see the data on the information itself.

3. What is Dimension Table?

Dimension table is a table which contain attributes of measurements stored in fact tables. This table consists of hierarchies, categories and logic that can be used to traverse in nodes.

4. What is Fact Table?

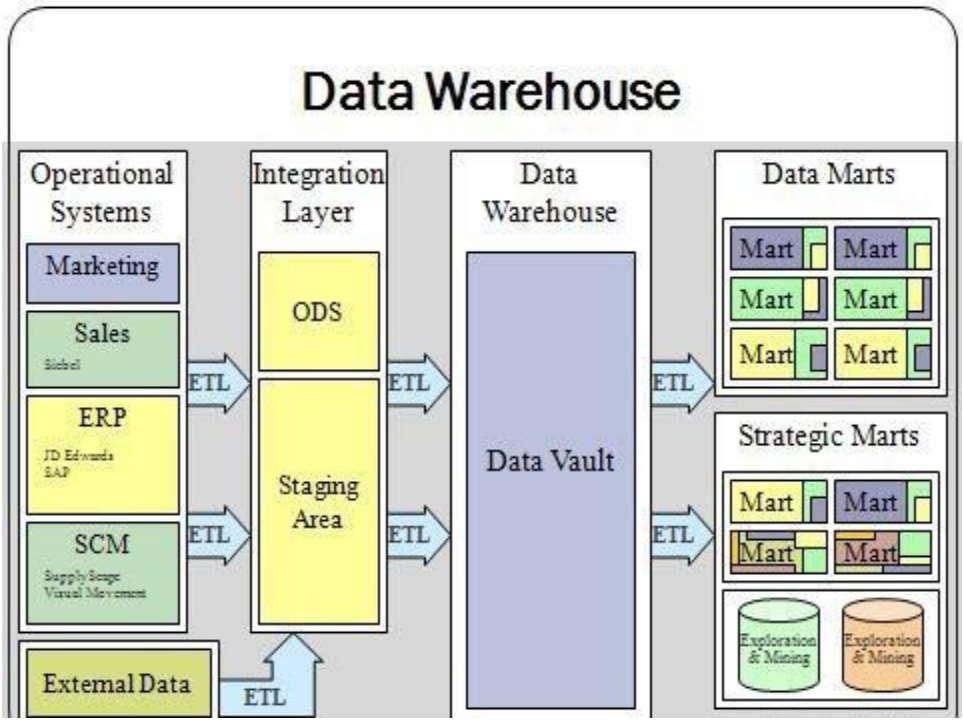
Fact table contains the measurement of business processes, and it contains foreign keys for the dimension tables.

Example – If the business process is manufacturing of bricks

Average number of bricks produced by one person/machine – measure of the business process

5. What are the stages of Data Warehousing?

There are four stages of Data Warehousing:



Data Warehouse

- Offline Operational Database
- Offline Data Warehouse
- Real Time Data Warehouse
- Integrated Data Warehouse

6. What is Data Mining?

Data Mining is set to be a process of analyzing the data in different dimensions or perspectives and summarizing into a useful information. Can be queried and retrieved the data from database in their own format.

7. What is OLTP?

OLTP is abbreviated as On-Line Transaction Processing, and it is an application that modifies the data whenever it received and has large number of simultaneous users.

8. What is OLAP?

OLAP is abbreviated as Online Analytical Processing, and it is set to be a system which collects, manages, processes multi-dimensional data for analysis and management purposes.

9. What is the difference between OLTP and OLAP?

Following are the differences between OLTP and OLAP:

OLTP	OLAP
Data is from original data source	Data is from various data sources
Simple queries by users	Complex queries by system
Normalized small database	De-normalized Large Database
Fundamental business tasks	Multi-dimensional business tasks

10. What is ODS?

ODS is abbreviated as Operational Data Store and it is a repository of real time operational data rather than long term trend data.

11. What is the difference between View and Materialized View?

A view is nothing but a virtual table which takes the output of the query and it can be used in place of tables.

A materialized view is nothing but an indirect access to the table data by storing the results of a query in a separate schema.

12. What is ETL?

ETL is abbreviated as Extract, Transform and Load. ETL is a software which is used to reads the data from the specified data source and extracts a desired subset of data. Next, it transform the data using rules and lookup tables and convert it to a desired state.

Then, load function is used to load the resulting data to the target database.

13. What is VLDB?

VLDB is abbreviated as Very Large Database and its size is set to be more than one terabyte database. These are decision support systems which is used to server large number of users.

14. What is real-time data warehousing?

Real-time data warehousing captures the business data whenever it occurs. When there is business activity gets completed, that data will be available in the flow and become available for use instantly.

15. What are Aggregate tables?

Aggregate tables are the tables which contain the existing warehouse data which has been grouped to certain level of dimensions. It is easy to retrieve data from the aggregated tables than the original table which has more number of records.

This table reduces the load in the database server and increases the performance of the query.

16. What is factless fact tables?

A factless fact tables are the fact table which doesn't contain numeric fact column in the fact table.

17. How can we load the time dimension?

Time dimensions are usually loaded through all possible dates in a year and it can be done through a program. Here, 100 years can be represented with one row per day.

18. What are Non-additive facts?

Non-Addictive facts are said to be facts that cannot be summed up for any of the dimensions present in the fact table. If there are changes in the dimensions, same facts can be useful.

19. What is conformed fact?

Conformed fact is a table which can be used across multiple data marts in combined with the multiple fact tables.

20. What is Datamart?

A Datamart is a specialized version of Data Warehousing and it contains a snapshot of operational data that helps the business people to decide with the analysis of past trends and experiences. A data mart helps to emphasizes on easy access to relevant information.

21. What is Active Data Warehousing?

An active data warehouse is a data warehouse that enables decision makers within a company or organization to manage customer relationships effectively and efficiently.

22. What is the difference between Data Warehouse and OLAP?

Data Warehouse is a place where the whole data is stored for analyzing, but OLAP is used for analyzing the data, managing aggregations, information partitioning into minor level information.

23. What is ER Diagram?

ER diagram is abbreviated as Entity-Relationship diagram which illustrates the interrelationships between the entities in the database. This diagram shows the structure of each tables and the links between the tables.

24. What are the key columns in Fact and dimension tables?

Foreign keys of dimension tables are primary keys of entity tables. Foreign keys of fact tables are the primary keys of the dimension tables.

25. What is SCD?

SCD is defined as slowly changing dimensions, and it applies to the cases where record changes over time.

26. What are the types of SCD?

There are three types of SCD and they are as follows:

SCD 1 – The new record replaces the original record

SCD 2 – A new record is added to the existing customer dimension table

SCD 3 – A original data is modified to include new data

27. What is BUS Schema?

BUS schema consists of suite of confirmed dimension and standardized definition if there is a fact tables.

28. What is Star Schema?

Star schema is nothing but a type of organizing the tables in such a way that result can be retrieved from the database quickly in the data warehouse environment.

29. What is Snowflake Schema?

Snowflake schema which has primary dimension table to which one or more dimensions can be joined. The primary dimension table is the only table that can be joined with the fact table.

30. What is a core dimension?

Core dimension is nothing but a Dimension table which is used as dedicated for single fact table or datamart.

31. What is called data cleaning?

Name itself implies that it is a self explanatory term. Cleaning of Orphan records, Data breaching business rules, Inconsistent data and missing information in a database.

32. What is Metadata?

Metadata is defined as data about the data. The metadata contains information like number of columns used, fix width and limited width, ordering of fields and data types of the fields.

33. What are loops in Data Warehousing?

In data warehousing, loops are existing between the tables. If there is a loop between the tables, then the query generation will take more time and it creates ambiguity. It is advised to avoid loop between the tables.

34. Whether Dimension table can have numeric value?

Yes, dimension table can have numeric value as they are the descriptive elements of our business.

35. What is the definition of Cube in Data Warehousing?

Cubes are logical representation of multidimensional data. The edge of the cube has the dimension members,and the body of the cube contains the data values.

36. What is called Dimensional Modelling?

Dimensional Modeling is a concept which can be used by data warehouse designers to build their own data warehouse. This model can be stored in two types of tables – Facts and Dimension table.

Fact table has facts and measurements of the business and dimension table contains the context of measurements.

37. What are the types of Dimensional Modeling?

There are three types of Dimensional Modeling and they are as follows:

- Conceptual Modeling
- Logical Modeling

- Physical Modeling

38. What is surrogate key?

Surrogate key is nothing but a substitute for the natural primary key. It is set to be a unique identifier for each row that can be used for the primary key to a table.

39. What is the difference between ER Modeling and Dimensional Modeling?

ER modeling will have logical and physical model but Dimensional modeling will have only Physical model.

ER Modeling is used for normalizing the OLTP database design whereas Dimensional Modeling is used for de-normalizing the ROLAP and MOLAP design.

40. What are the steps to build the data warehouse?

Following are the steps to be followed to build the data warehouse:

- Gathering business requirements
- Identifying the necessary sources
- Identifying the facts
- Defining the dimensions
- Defining the attributes
- Redefine the dimensions and attributes if required
- Organize the Attribute hierarchy
- Define Relationships
- Assign unique Identifiers

41. What are the different types of data warehousing?

Following are the different types of Data Warehousing:

- Enterprise Data Warehousing
- Operational Data Store

- Data Mart

42. What needs to be done while starting the database?

Following need to be done to start the database:

1. Start an Instance
2. Mount the database
3. Open the database

43. What needs to be done when the database is shutdown?

Following needs to be done when the database is shutdown:

1. Close the database
2. Dismount the database
3. Shutdown the Instance

44. Can we take backup when the database is opened?

No, We cannot take full backup when the database is opened.

45. What is defined as Partial Backup?

A Partial backup in an operating system is a backup short of full backup and it can be done while the database is opened or shutdown.

46. What is the goal of Optimizer?

The goal to Optimizer is to find the most efficient way to execute the SQL statements.

47. What is Execution Plan?

Execution Plan is a plan which is used to the optimizer to select the combination of the steps.

48. What are the approaches used by Optimizer during execution plan?

There are two approaches:

1. Rule Based
2. Cost Based

49. What are the tools available for ETL?

Following are the ETL tools available:

Informatica

Data Stage

Oracle

Warehouse Builder

Ab Initio

Data Junction

50.What is the difference between metadata and data dictionary?

Metadata is defined as data about the data. But, Data dictionary contain the information about the project information, graphs, abinito commands and server information.

51. What is Data mining ?

Data mining is knowledge discovery in databases. It is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.

52. What is difference between OLAP and data mining ?

OLAP - (On-line Analytical Processing)provides you with a very good view of what is happening, but can not predict what will happen in the future or why it is happening where as data mining is group of techniques that find relationships that have not previously been discovered.

53. What are the types of tasks that are carried out during data mining ?

Data mining involves 2 types of tasks

•Prediction Tasks- Use some variables to predict unknown or future values of other variables

•Description Tasks- Find human-interpretable patterns that describe the data.

54. What are some of the tasks of data mining?

Following activities are carried out during data mining

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

55. What do you mean by preprocessing of data in data mining ?

Before data is mined it has to be preprocessed. It consists of following three stages

- Data cleaning - Real world data is dirty so need to be cleaned
- Data reduction- Remove data not useful for mining
- Data transformation - Syntactic transformation

56. What is Data cleaning ?

Causes of Data Cleansing

- Missing values
- Noisy data (Human/Machine Errors)
- Inconsistent data

Data cleaning tasks

- Handling missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

57. Explain Data reduction ?

It consists of following three tasks -

- Dimensionality reduction - Attribute subset selection
- Numerosity reduction - Tuple subset selection
- Discretization - Reduce the cardinality of active domain

58. What is Data Transformation ?

It consist of following tasks

1. Generalization - concept hierarchy climbing
2. Attribute/feature construction - New attributes are constructed and added to the tuple
3. Normalization - scaled to fall within a small, specified range

59. What are the uses of multi feature cubes?

Multi feature cubes, which compute complex queries involving multiple dependent aggregates at multiple granularity. These cubes are very useful in practice. Many complex data mining queries can be answered by multi feature cubes without any significant increase in computational cost, in comparison to cube computation for simple queries with standard data cubes.

60. Explain the difference between star and snowflake schema.

The dimension table of the snowflake schema model may be kept in normalized Form to reduce redundancies. Such a table is easy to maintain and saves storage space.

61. In the context of data warehousing what is data transformation?

`In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

Smoothing, Aggregation, Generalization, Normalization, Attribute construction

62. Define Slice and Dice operation.

The slice operation performs a selection on one dimension of the cube resulting in A sub cube. The dice operation defines a sub cube by performing a selection on two (or) more dimensions.

63. List the characteristics of a data ware house.

There are four key characteristics which separate the data warehouse from

other major operational systems:

1. Subject Orientation: Data organized by subject

2. Integration: Consistency of defining parameters

3. Non-volatility: Stable data storage medium

4. Time-variance: Timeliness of data and access terms

64. What are the various sources for data warehouse?

Handling of relational and complex types of data: Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.

Mining information from heterogeneous databases and global information systems: Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases.

65. Differentiate fact table and dimension table.

Fact table contains the name of facts (or) measures as well as keys to each of the related dimensional tables. A dimension table is used for describing the dimension. (e.g.) A dimension table for item may contain the attributes item_ name, brand and type.

66. Briefly discuss the schemas for multidimensional databases.

Stars schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension.

Snowflakes schema: The snowflake schema is a variant of the star schema model, where Some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

Fact Constellations: Sophisticated applications may require multiple fact tables to share

dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

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

Data warehouse is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making. A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples(records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. Both are used to store and manipulate the data.

68. What is descriptive and predictive data mining?

Descriptive data mining, which describes data in a concise and summarative manner and presents interesting general properties of the data.

Predictive data mining, which analyzes data in order to construct one or a set of models and attempts to predict the behavior of new data sets. Predictive data mining, such as classification, regression analysis, and trend analysis.

69. Differentiate data mining and data warehousing.

Data Mining	Data Warehouse
Data mining is the process of analyzing unknown patterns of data.	A data warehouse is database system which is designed for analytical instead of transactional work.
Data mining is a method of comparing large amounts of data to finding right patterns.	Data warehousing is a method of centralizing data from different sources into one common repository.

Data mining is usually done by business users with the assistance of engineers.	Data warehousing is a process which needs to occur before any data mining can take place.
Data mining is the considered as a process of extracting data from large data sets.	On the other hand, Data warehousing is the process of pooling all relevant data together.
One of the most important benefits of data mining techniques is the detection and identification of errors in the system.	One of the pros of Data Warehouse is its ability to update consistently. That's why it is ideal for the business owner who wants the best and latest features.
Data mining helps to create suggestive patterns of important factors. Like the buying habits of customers, products, sales. So that, companies can make the necessary adjustments in operation and production.	Data Warehouse adds an extra value to operational business systems like CRM systems when the warehouse is integrated.
The Data mining techniques are never 100% accurate and may cause serious consequences in certain conditions.	In the data warehouse, there is great chance that the data which was required for analysis by the organization may not be integrated into the warehouse. It can easily lead to loss of information.
The information gathered based on Data Mining by organizations can be misused against a group of people.	Data warehouses are created for a huge IT project. Therefore, it involves high maintenance system which can impact the revenue of medium to small-scale organizations.
After successful initial queries, users may ask more complicated queries which would increase the workload.	Data Warehouse is complicated to implement and maintain.

Organisations can benefit from this analytical tool by equipping pertinent and usable knowledge-based information.	Data warehouse stores a large amount of historical data which helps users to analyze different time periods and trends for making future predictions.
Organisations need to spend lots of their resources for training and Implementation purpose. Moreover, data mining tools work in different manners due to different algorithms employed in their design.	In Data warehouse, data is pooled from multiple sources. The data needs to be cleaned and transformed. This could be a challenge.
The data mining methods are cost-effective and efficient compares to other statistical data applications.	Data warehouse's responsibility is to simplify every type of business data. Most of the work that will be done on user's part is inputting the raw data.
Another critical benefit of data mining techniques is the identification of errors which can lead to losses. Generated data could be used to detect a drop-in sale.	Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
Data mining helps to generate actionable strategies built on data insights.	Once you input any information into Data warehouse system, you will unlikely to lose track of this data again. You need to conduct a quick search, helps you to find the right statistic information.

70. Define CLARANS.

CLARANS(Cluster Large Applications based on Randomized Search) to improve the quality of CLARA we go for CLARANS. It Draws sample with some randomness in each step of search. It overcome the problem of scalability that K-Medoids suffers from.

71. What is meant by web usage mining?

Web usage mining is the process of extracting useful information from server logs i.e. users history. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

72. What is mean by the frequency item set property?

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of the itemset.

73. What is mean by web content mining?

Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query.

74. What is the need for preprocessing the data?

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

75. What is dimensionality reduction?

In dimensionality reduction, data encoding or transformations are applied so as to obtain a reconstructed from the compressed data without any loss of information, the data reduction is called lossless.

76. Mention the various tasks to be accomplished as part of data pre-processing.

1. Data cleaning

2. Data Integration

3. Data Transformation

4. Data reduction

77. What is data cleaning?

Data cleaning means removing the inconsistent data or noise and collecting necessary information of a collection of interrelated data. 6. Define Data mining. (Nov/Dec 2008) amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more

78. Why is it important to have data mining query language?

The design of an effective data mining query language requires a deep understanding of the power, limitation, and underlying mechanisms of the various kinds of data mining tasks. A data mining query language can be used to specify data mining tasks. In particular, we examine how to define data warehouses and data marts in our SQL-based data mining query language, DMQL.

79. List the five primitives for specifying a data mining task.

The set of task-relevant data to be mined the kind of knowledge to be mined: The background knowledge to be used in the discovery process the interestingness measures and thresholds for pattern evaluation The expected representation for visualizing the discovered pattern

80. What is data generalization?

It is process that abstracts a large set of task-relevant data in a database from relatively low conceptual levels to higher conceptual levels 2 approaches for Generalization. 1) Data cube approach 2) Attribute-oriented induction approach

81. How do you clean the data?

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. For Missing Values 1. Ignore the tuple 2. Fill in the missing value manually 3. Use a global constant to fill in the missing value 4. Use the attribute mean to fill in the missing value: 5. Use the attribute mean for all samples belonging to the same class as the given tuple 6. Use the most probable value to fill in the missing value For Noisy Data 1. Binning: Binning methods smooth a sorted data value by consulting its values around it. 2. Regression: Data can be smoothed by fitting the data to a

function, such as with Regression 3. Clustering: Outliers may be detected by clustering, where similar values are organized into groups, or 3.

82. List the techniques to improve the efficiency of Apriori algorithm.

Hash based technique

Transaction

Reduction

Portioning

Sampling

Dynamic item counting

83. What is FP growth?

FP-growth, which adopts a divide-and-conquer strategy as follows. First, it compresses the database representing frequent items into a frequent-pattern tree, or FP-tree,which retains the itemset association information. It then divides the compressed database into a set of conditional databases (a special kind of projected database), each associated with one frequent item or pattern fragment and mines each such database separately.

84. What Is Naive Bayes Algorithm?

Naive Bayes Algorithm is used to generate mining models. These models help to identify relationships between input columns and the predictable columns. This algorithm can be used in the initial stage of exploration. The algorithm calculates the probability of every state of each input column given predictable columns possible states. After the model is made, the results can be used for exploration and making predictions.

85.Explain Clustering Algorithm?

Clustering algorithm is used to group sets of data with similar characteristics also called as clusters. These clusters help in making faster decisions, and exploring data. The algorithm first identifies relationships in a dataset following which it generates a series of clusters based on the relationships. The process of creating clusters is iterative. The algorithm redefines the groupings to create clusters that better represent the data.

86. Mention Some Of The Data Mining Techniques?

- Statistics
- Machine learning

- Decision Tree
- Hidden markov models
- Artificial Intelligence
- Genetic Algorithm
- Meta learning

87. What Are The Steps Involved In KDD Process?

- Data cleaning
- Data Mining
- Pattern Evaluation
- Knowledge Presentation
- Data Integration
- Data Selection
- Data Transformation

88. What are the different fields where data mining is used?

Data Mining is mainly used by big consumer-based companies that focus on retail, financial, communication, and marketing fields. It is used to get the consumer's transactional data pattern to determine price, customer preferences, and product positioning, which later impact sales, customer satisfaction, and corporate profits.

Following is the list of most important areas where data mining is widely used:

Healthcare and Personal Grooming

Data mining has a significant impact in the field of healthcare. It uses data and analytics to identify the best practices that can improve care and reduce costs. Scientists use several Data Mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization, statistics, etc., to make things easy for patients. Using Data Mining, we can predict the volume of patients in every category and make sure that the patients get the appropriate care at the right place and at the right time.

Market Basket Analysis

This modeling technique follows the theory that if you buy a specific group of items, you are more likely to buy another group of items. Using this technique, the retailer can understand the purchase behavior of a buyer and change the store's layout according to the buyer's needs.

Education & Training

Educational Data Mining is used to identify and predict the students' future learning behavior. If a student is studying a particular course, then the institutes can know which related course they may apply later by using Data Mining. This is also beneficial to make focus on what to teach and how to teach. The institutes can capture the learning pattern of the students and use to develop techniques to teach them.

Manufacturing Engineering

By using Data mining tools, we can discover patterns in complex manufacturing processes. We can use this to predict the product development span time, cost, and dependencies, among other tasks.

Fraud Detection

Data Mining can be used as a perfect fraud detection system to protect the information of all users. By Data Mining, we can classify fraudulent or non-fraudulent data and make an algorithm to identify whether the record is fraudulent or not.

Customer Relationship Management

We can use Data Mining to maintain a proper relationship with a customer.

Some other areas where data mining is used:

- Intrusion Detection
- Lie Detection
- Customer Segmentation
- Financial Banking
- Corporate Surveillance

- Research Analysis
- Criminal Investigation
- Bio Informatics

89. What are the different techniques used for Data Mining?

Following is the list of most important Data Mining techniques:

Prediction: This technique specifies the relationship between independent and dependent instances. For example, while considering sales data, if we want to predict the future profit, the sale acts as a separate instance, whereas the payoff is the dependent instance. Accordingly, based on sales and profit's historical data, the associated profit is the predicted value.

Decision trees: It specifies a tree structure where the decision tree's root acts as a condition/question having multiple answers. Each answer sets to specific data that helps in determining the final decision based on the data.

Clustering analysis: This technique specifies that a cluster of objects having similar characteristics is formed automatically. The clustering method defines classes and then places suitable objects in each class.

Sequential Patterns: This technique is used to specify the pattern analysis used for discovering identical patterns in transaction data or regular events. For example, customers' historical data helps a brand identify the patterns in the transactions that happened in the past year.

Classification Analysis: This is a Machine Learning based method in which each item in a particular set is classified into predefined groups. It uses advanced techniques like linear programming, neural networks, decision trees, etc.

Association rule learning: This technique is used to create a pattern based on the items' relationship in a single transaction.

90. What are the different storage models available in OLAP?

There are mainly three storage models available in OLAP. They are:

- MOLAP: Multidimensional Online Analytical Processing

- ROLAP: Relational Online Analytical processing

- HOLAP: Hybrid Online Analytical Processing

91. What are the advantages and disadvantages of using the MOLAP storage model?

The term MOLAP stands for "Multidimensional Online Analytical Processing." As the name shows, it is a multidimensional storage model. This storage model type stores the data in multidimensional cubes and not in the standard relational databases.

Advantages of using the MOLAP storage model:

- It stores the data in multidimensional cubes, so the query performance is excellent.

- The calculations are pre-generated when a cube is created.

Disadvantages of using the MOLAP storage model:

- The most significant disadvantage of using MOLAP is that it can store only a limited amount of data. In this storage model, the calculations are triggered at the cube generation process so, it cannot support a large amount of data.

- It requires a lot of skill to utilize this.

- It is not free. You have to pay the license cost associated with it.

92. What are the advantages and disadvantages of using the ROLAP storage model?

The term ROLAP stands for "Relational Online Analytical Processing." In this storage model, the data is stored in the form of a relational database.

Advantages of using the ROLAP storage model:

- In this storage model, the data is stored in relational databases so, it is easy to handle a large amount of data storage.

- It provides all the functionalities as it is a relational database.

Disadvantages of using the ROLAP storage model:

- The most significant disadvantage of this storage model is that it is comparatively slow.
- All other disadvantages we face in SQL are the same in this storage model also.

93. What is Discrete and Continuous data in Data Mining?

In Data Mining, discrete data is a type of data defined as finite data. This type of information is never changed.

Example: Mobile numbers, gender, etc. are the example of discrete data.

On the other hand, continuous data is a type of data that changes continuously and in an ordered fashion.

Example: Age is an example of continuous data.

94. What are Interval Scaled Variables?

The continuous measurement of linear scale is called Interval Scaled Variable. For example, height and weight, weather temperature, etc. We can calculate these measurements by using Euclidean distance or Minkowski distance.

95. What is factless fact tables?

A factless fact tables are the fact table which doesn't contain numeric fact column in the fact table.

96. What is a Decision Tree Algorithm?

A decision tree is a tree in which every node is either a leaf node or a decision node. This tree takes an input an object and outputs some decision. All Paths from root node to the leaf node are reached by either using AND or OR or BOTH. The tree is constructed using the regularities of the data. The decision tree is not affected by Automatic Data Preparation.

97. Define support and confidence.

The support for a rule R is the ratio of the number of occurrences of R, given all occurrences of all rules.
The confidence of a rule X->Y, is the ratio of the number of occurrences of Y given X, among all other occurrences given X

98. Why is association rule necessary.

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.

It is intended to identify strong rules discovered in database using different measures of interesting.

99. What is data discrimination

Data discrimination is the comparison of the general features of the target class objects against one or more contrasting objects.

100. What is text mining

Text mining is the procedure of synthesizing information, by analyzing relations, patterns, and rules among textual data. These procedures contains text summarization, text categorization, and text clustering.