**1. What is Data Mining?**

[Data mining](https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/)refers to extracting or mining knowledge from large amounts of data. In other words, Data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns.

**2. What are the different tasks of Data Mining?**

The following activities are carried out during data mining:

* Classification
* Clustering
* Association Rule Discovery
* Sequential Pattern Discovery
* Regression
* Deviation Detection

**3. Discuss the Life cycle of Data Mining projects?**

The life cycle of Data mining projects:

* **Business understanding:**Understanding projects objectives from a business perspective, data mining problem definition.
* **Data understanding:**Initial data collection and understand it.
* **Data preparation:** Constructing the final data set from raw data.
* **Modeling:** Select and apply data modeling techniques.
* **Evaluation:** Evaluate model, decide on further deployment.
* **Deployment**: Create a report, carry out actions based on new insights.

**4. Explain the process of KDD?**

Data mining treat as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. In others view data mining as simply an essential step in the process of knowledge discovery, in which intelligent methods are applied in order to extract data patterns.

*Knowledge discovery from data consists of the following steps:*

* Data cleaning (to remove noise or irrelevant data).
* Data integration (where multiple data sources may be combined).
* Data selection (where data relevant to the analysis task are retrieved from the database).
* Data transformation (where data are transmuted or consolidated into forms appropriate for mining by performing summary or aggregation functions, for sample).
* Data mining (an important process where intelligent methods are applied in order to extract data patterns).
* Pattern evaluation (to identify the fascinating patterns representing knowledge based on some interestingness measures).
* Knowledge presentation (where knowledge representation and visualization techniques are used to present the mined knowledge to the user).

**5. What is Classification?**

Classification is the processing of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Classification can be used for predicting the class label of data items. However, in many applications, one may like to calculate some missing or unavailable data values rather than class labels.

**6. Explain Evolution and deviation analysis?**

Data evolution analysis describes and models regularities or trends for objects whose behavior variations over time. Although this may involve discrimination, association, classification, characterization, or clustering of time-related data, distinct features of such an analysis involve time-series data analysis, periodicity pattern matching, and similarity-based data analysis.

In the analysis of time-related data, it is often required not only to model the general evolutionary trend of the data but also to identify data deviations that occur over time. Deviations are differences between measured values and corresponding references such as previous values or normative values. A data mining system performing deviation analysis, upon the detection of a set of deviations, may do the following: describe the characteristics of the deviations, try to describe the reason behindhand them, and suggest actions to bring the deviated values back to their expected values.

**7. What is Prediction?**

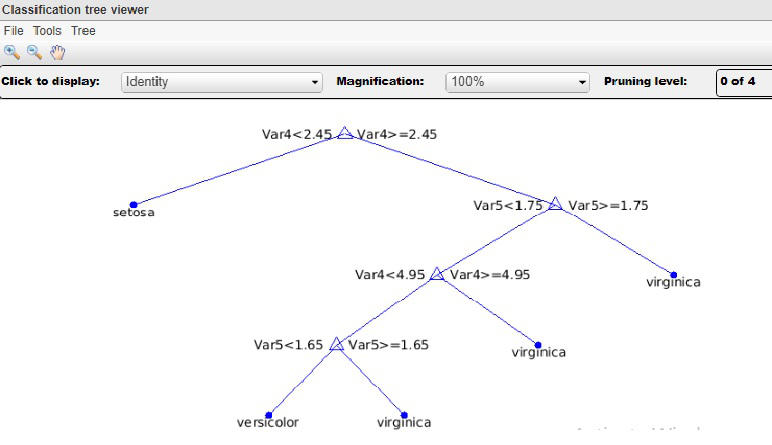
Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled object, or to measure the value or value ranges of an attribute that a given object is likely to have. In this interpretation, classification and regression are the two major types of prediction problems where classification is used to predict discrete or nominal values, while regression is used to predict incessant or ordered values.

**8. Explain the Decision Tree Classifier?**

A Decision tree is a flow chart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test and each leaf node (or terminal node) holds a class label. The topmost node of a tree is the root node.

A Decision tree is a classification scheme that generates a tree and a set of rules, representing the model of different classes, from a given data set. The set of records available for developing classification methods is generally divided into two disjoint subsets namely a training set and a test set. The former is used for originating the classifier while the latter is used to measure the accuracy of the classifier. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified.

In the decision tree classifier, we categorize the attributes of the records into two different types. Attributes whose domain is numerical are called the numerical attributes and the attributes whose domain is not numerical are called categorical attributes. There is one distinguished attribute called a class label. The goal of classification is to build a concise model that can be used to predict the class of the records whose class label is unknown. Decision trees can simply be converted to classification rules.



**9. What are the advantages of a decision tree classifier?**

* Decision trees are able to produce understandable rules.
* They are able to handle both numerical and categorical attributes.
* They are easy to understand.
* Once a decision tree model has been built, classifying a test record is extremely fast.
* Decision tree depiction is rich enough to represent any discrete value classifier.
* Decision trees can handle datasets that may have errors.
* Decision trees can deal with handle datasets that may have missing values.
* They do not require any prior assumptions.  Decision trees are self-explanatory and when compacted they are also easy to follow. That is to say, if the decision tree has a reasonable number of leaves it can be grasped by non-professional users. Furthermore, since decision trees can be converted to a set of rules, this sort of representation is considered comprehensible.

**10. Explain Bayesian classification in Data Mining?**

A Bayesian classifier is a statistical classifier. They can predict class membership probabilities, for instance, the probability that a given sample belongs to a particular class. Bayesian classification is created on the Bayes theorem. A simple Bayesian classifier is known as the naive Bayesian classifier to be comparable in performance with decision trees and neural network classifiers. Bayesian classifiers have also displayed high accuracy and speed when applied to large databases.

**11. Why Fuzzy logic is an** **important area for Data Mining?**

Rule-based systems for classification have the disadvantage that they involve exact values for continuous attributes. Fuzzy logic is useful for data mining systems performing classification. It provides the benefit of working at a high level of abstraction. In general, the usage of fuzzy logic in rule-based systems involves the following:

* Attribute values are changed to fuzzy values.
* For a given new sample, more than one fuzzy rule may apply. Every applicable rule contributes a vote for membership in the categories. Typically, the truth values for each projected category are summed.
* The sums obtained above are combined into a value that is returned by the system. This process may be done by weighting each category by its truth sum and multiplying by the mean truth value of each category. The calculations involved may be more complex, depending on the difficulty of the fuzzy membership graphs.

**12. What are Neural networks?**

A neural network is a set of connected input/output units where each connection has a weight associated with it. During the knowledge phase, the network acquires by adjusting the weights to be able to predict the correct class label of the input samples. Neural network learning is also denoted as connectionist learning due to the connections between units. Neural networks involve long training times and are therefore more appropriate for applications where this is feasible. They require a number of parameters that are typically best determined empirically, such as the network topology or “structure”. Neural networks have been criticized for their poor interpretability since it is difficult for humans to take the symbolic meaning behind the learned weights. These features firstly made neural networks less desirable for data mining.

The advantages of neural networks, however, contain their high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained. In addition, several algorithms have newly been developed for the extraction of rules from trained neural networks. These issues contribute to the usefulness of neural networks for classification in data mining. The most popular neural network algorithm is the backpropagation algorithm, proposed in the 1980s

**13. How Backpropagation Network Works?**

A Backpropagation learns by iteratively processing a set of training samples, comparing the network’s estimate for each sample with the actual known class label. For each training sample, weights are modified to minimize the mean squared error between the network’s prediction and the actual class. These changes are made in the “backward” direction, i.e., from the output layer, through each concealed layer down to the first hidden layer (hence the name backpropagation). Although it is not guaranteed, in general, the weights will finally converge, and the knowledge process stops.

**14. What is a Genetic Algorithm?**

Genetic algorithm is a part of evolutionary computing which is a rapidly growing area of artificial intelligence. The genetic algorithm is inspired by Darwin’s theory about evolution. Here the solution to a problem solved by the genetic algorithm is evolved. In a genetic algorithm, a population of strings (called chromosomes or the genotype of the gen me), which encode candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem, is evolved toward better solutions. Traditionally, solutions are represented in the form of binary strings, composed of 0s and 1s, the same way other encoding schemes can also be applied.

**15. What is Classification Accuracy?**

Classification accuracy or accuracy of the classifier is determined by the percentage of the test data set examples that are correctly classified. The classification accuracyof a classification tree = (1 – Generalization error).

**16. Define Clustering** **in Data Mining?**

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

**17. Write a difference between classification and clustering?[IMP]**

| Parameters | CLASSIFICATION | CLUSTERING |
| --- | --- | --- |
| Type | Used for supervised need learning | Used for unsupervised learning |
| Basic | Process of classifying the input instances based on their corresponding class labels | Grouping the instances based on their similarity without the help of class labels |
| Need | It has labels so there is a need for training and testing data set for verifying the model created | There is no need for training and testing dataset |
| Complexity | More complex as compared to clustering | Less complex as compared to classification |
| Example Algorithms | Logistic regression, Naive Bayes classifier, Support vector machines, etc. | k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering algorithm etc. |

**18. What is Supervised and Unsupervised Learning?[TCS interview question]**

**Supervised learning,** as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

**Unsupervised learning** is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, the machine is restricted to find the hidden structure in unlabeled data by itself.

**19. Name areas of applications of data mining?**

* Data Mining Applications for Finance
* Healthcare
* Intelligence
* Telecommunication
* Energy
* Retail
* E-commerce
* Supermarkets
* Crime Agencies
* Businesses Benefit from data mining

**20. What are the issues in data mining?**

A number of issues that need to be addressed by any serious data mining package

* Uncertainty Handling
* Dealing with Missing Values
* Dealing with Noisy data
* Efficiency of algorithms
* Constraining Knowledge Discovered to only Useful
* Incorporating Domain Knowledge
* Size and Complexity of Data
* Data Selection
* Understandably of Discovered Knowledge: Consistency between Data and Discovered Knowledge.

**21. Give an introduction to data mining query language?**

DBQL or Data Mining Query Language proposed by Han, Fu, Wang, et.al. This language works on the DBMiner data mining system. DBQL  queries were based on SQL(Structured Query language). We can this language for databases and data warehouses as well. This query language support ad hoc and interactive data mining.

**22. Differentiate Between Data Mining And Data Warehousing?**

**Data Mining:** It is the process of finding patterns and correlations within large data sets to identify relationships between data. Data mining tools allow a business organization to predict customer behavior. Data mining tools are used to build risk models and detect fraud. Data mining is used in market analysis and management, fraud detection, corporate analysis, and risk management.  
It is a technology that aggregates structured data from one or more sources so that it can be compared and analyzed rather than transaction processing.

**Data Warehouse**: A data warehouse is designed to support the management decision-making process by providing a platform for data cleaning, data integration, and data consolidation. A data warehouse contains subject-oriented, integrated, time-variant, and non-volatile data.

Data warehouse consolidates data from many sources while ensuring data quality, consistency, and accuracy. Data warehouse improves system performance by separating analytics processing from transnational databases. Data flows into a data warehouse from the various databases. A data warehouse works by organizing data into a schema that describes the layout and type of data. Query tools analyze the data tables using schema.

**23.What is Data Purging?**

The term purging can be defined as Erase or Remove. In the context of data mining, data purging is the process of remove, unnecessary data from the database permanently and clean data to maintain its integrity.

**24. What Are Cubes?**

A data cube stores data in a summarized version which helps in a faster analysis of data. The data is stored in such a way that it allows reporting easily. E.g. using a data cube A user may want to analyze the weekly, monthly performance of an employee. Here, month and week could be considered as the dimensions of the cube.

**25.What are the differences between OLAP And OLTP?[IMP]**

| OLAP (Online Analytical Processing) | OLTP (Online Transaction Processing) |
| --- | --- |
| Consists of historical data from various Databases. | Consists only of application-oriented day-to-day operational current data. |
| Application-oriented day-to-dayIt is subject-oriented. Used for Data Mining, Analytics, Decision making, etc. | It is application-oriented. Used for business tasks. |
| The data is used in planning, problem-solving, and decision-making. | The data is used to perform day-to-day fundamental operations. |
| It reveals a snapshot of present business tasks. | It provides a multi-dimensional view of different business tasks. |
| A large forex amount of data is stored typically in TB, PB | The size of the data is relatively small as the historical data is archived. For example, MB, GB |
| Relatively slow as the amount of data involved is large. Queries may take hours. | Very Fast as the queries operate on 5% of the data. |
| It only needs backup from time to time as compared to OLTP. | The backup and recovery process is maintained religiously |
| This data is generally managed by the CEO, MD, GM. | This data is managed by clerks, managers. |
| Only read and rarely write operation. | Both read and write operations. |

**26. Explain Association Algorithm In Data Mining?**

Association analysis is the finding of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for a market basket or transaction data analysis. Association rule mining is a significant and exceptionally dynamic area of data mining research. One method of association-based classification, called associative classification, consists of two steps. In the main step, association instructions are generated using a modified version of the standard association rule mining algorithm known as Apriori. The second step constructs a classifier based on the association rules discovered.

**27. Explain how to work with data mining algorithms included in SQL server data mining?**

SQL Server data mining offers Data Mining Add-ins for Office 2007 that permits finding the patterns and relationships of the information. This helps in an improved analysis. The Add-in called a Data Mining Client for Excel is utilized to initially prepare information, create models, manage, analyze, results.

**28. Explain Over-fitting?**

The concept of over-fitting is very important in data mining. It refers to the situation in which the induction algorithm generates a classifier that perfectly fits the training data but has lost the capability of generalizing to instances not presented during training. In other words, instead of learning, the classifier just memorizes the training instances. In the decision trees over fitting usually occurs when the tree has too many nodes relative to the amount of training data available. By increasing the number of nodes, the training error usually decreases while at some point the generalization error becomes worse. The  Over-fitting can lead to difficulties when there is noise in the training data or when the number of the training datasets, the error of the fully built tree is zero, while the true error is likely to be bigger.

***There are many disadvantages of an over-fitted decision tree:***

* Over-fitted models are incorrect.
* Over-fitted decision trees require more space and more computational resources.
* They require the collection of unnecessary features.

**29. Define Tree Pruning?**

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over-fitting the data.  So the tree pruning is a technique that removes the overfitting problem. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data. The pruning phase eliminates some of the lower branches and nodes to improve their performance. Processing the pruned tree to improve understandability.

**30. What is a Sting?**

Statistical Information Grid is called STING; it is a grid-based multi-resolution clustering strategy. In the STING strategy, every one of the items is contained into rectangular cells, these cells are kept into different degrees of resolutions and these levels are organized in a hierarchical structure.

**31**. **Define Chameleon Method?**

Chameleon is another hierarchical clustering technique that utilization dynamic modeling. Chameleon is acquainted with recover the disadvantages of the CURE clustering technique. In this technique, two groups are combined, if the interconnectivity between two clusters is greater than the inter-connectivity between the object inside a cluster/ group.

**32. Explain the Issues regarding Classification And Prediction?**

Preparing the data for classification and prediction:

* Data cleaning
* Relevance analysis
* Data transformation
* Comparing classification methods
* Predictive accuracy
* Speed
* Robustness
* Scalability
* Interpretability

**33.Explain the use of data mining queries or why data mining queries are more helpful?**

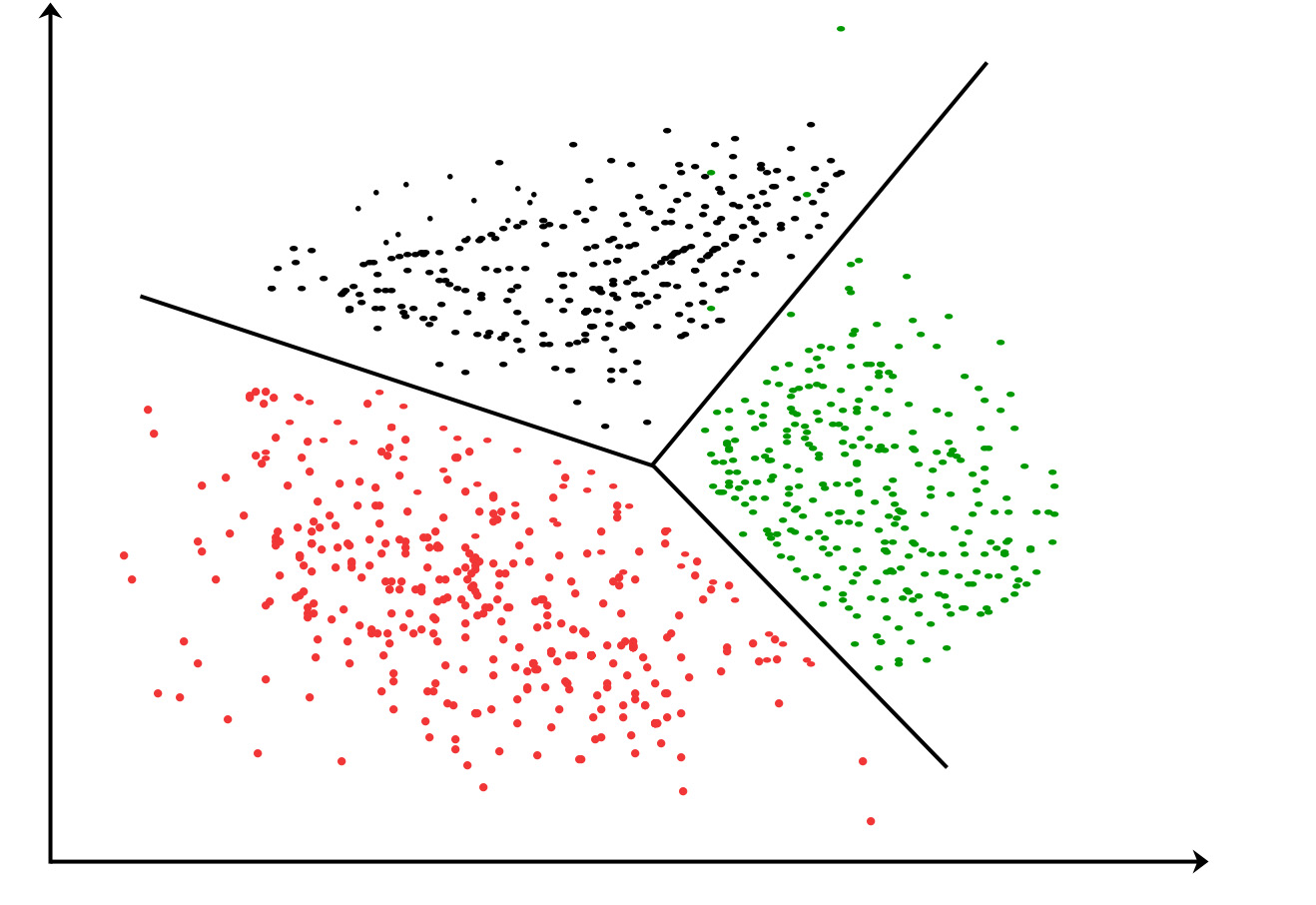
The data mining queries are primarily applied to the model of new data to make single or multiple different outcomes. It also permits us to give input values. The query can retrieve information effectively if a particular pattern is defined correctly. It gets the training data statistical memory and gets the specific design and rule of the common case addressing a pattern in the model. It helps in extracting the regression formulas and other computations. It additionally recovers the insights concerning the individual cases utilized in the model. It incorporates the information which isn’t utilized in the analysis, it holds the model with the assistance of adding new data and perform the task and cross-verified.

**34. What is a machine learning-based approach to data mining?**

This question is the high-level Data Mining Interview Questions asked in an Interview. Machine learning is basically utilized in data mining since it covers automatic programmed processing systems, and it depended on logical or binary tasks. . Machine learning for the most part follows the rule that would permit us to manage more general information types, incorporating cases and in these sorts and number of attributes may differ. Machine learning is one of the famous procedures utilized for data mining and in Artificial intelligence too.

**35.What is the K-means algorithm?**

[K-means clustering algorithm](https://www.geeksforgeeks.org/k-means-clustering-introduction/) – It is the simplest unsupervised learning algorithm that solves clustering problems. K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



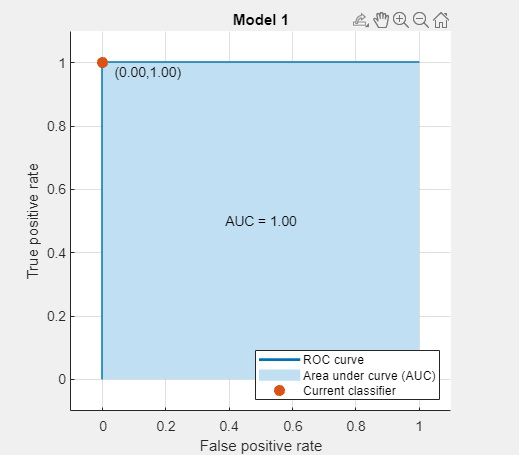
*Figure: K-Means Clustering division of attribute*

**36. What are precision and recall?[IMP]**

Precision is the most commonly used error metric in the n classification mechanism. Its range is from 0 to 1, where 1 represents 100%.

Recall can be defined as the number of the Actual Positives in our model which has a class label as Positive (True Positive)”. Recall and the true positive rate is totally identical. Here’s the formula for it:

**Recall** = (True positive)/(True positive + False negative)



**37. What are the ideal situations in which t-test or z-test can be used?**

It is a standard practice that a t-test is utilized when there is an example size under 30 attributes and the z-test is viewed as when the example size exceeds 30 by and large.

**38. What is the simple difference between standardized and unstandardized coefficients?**

In the case of normalized coefficients, they are interpreted dependent on their standard deviation values. While the unstandardized coefficient is estimated depending on the real value present in the dataset.

**39. How are outliers detected?**

Numerous approaches can be utilized for distinguishing outliers anomalies, but the two most generally utilized techniques are as per the following:

* Standard deviation strategy: Here, the value is considered as an outlier if the value is lower or higher than three standard deviations from the mean value.
* Box plot technique: Here, a value is viewed as an outlier if it is lesser or higher than 1.5 times the interquartile range (IQR)

**40. Why is KNN preferred when determining missing numbers in data?**

K-Nearest Neighbour (KNN) is preferred here because of the fact that KNN can easily approximate the value to be determined based on the values closest to it.

The k-nearest neighbor (K-NN) classifier is taken into account as an example-based classifier, which means that the training documents are used for comparison instead of an exact class illustration, like the class profiles utilized by other classifiers. As such, there’s no real training section. once a new document has to be classified, the k most similar documents (neighbors) are found and if a large enough proportion of them are allotted to a precise class, the new document is also appointed to the present class, otherwise not. Additionally, finding the closest neighbors is quickened using traditional classification strategies.

**41. Explain Prepruning and Post pruning approach in Classification?**

**Prepruning:** In the prepruning approach, a tree is “pruned” by halting its construction early (e.g., by deciding not to further split or partition the subset of training samples at a given node). Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples, or the probability distribution of those samples. When constructing a tree, measures such as statistical significance, information gain, etc., can be used to assess the goodness of a split. If partitioning the samples at a node would result in a split that falls below a pre-specified threshold, then further partitioning of the given subset is halted. There are problems, however, in choosing a proper threshold. High thresholds could result in oversimplified trees, while low thresholds could result in very little simplification.

**Postpruning:**The postpruning approach removes branches from a “fully grown” tree. A tree node is pruned by removing its branches. The cost complexity pruning algorithm is an example of the post pruning approach. The pruned node becomes a leaf and is labeled by the most frequent class among its former branches. For every non-leaf node in the tree, the algorithm calculates the expected error rate that would occur if the subtree at that node were pruned. Next, the predictable error rate occurring if the node were not pruned is calculated using the error rates for each branch, collective by weighting according to the proportion of observations along each branch. If pruning the node leads to a greater probable error rate, then the subtree is reserved. Otherwise, it is pruned. After generating a set of progressively pruned trees, an independent test set is used to estimate the accuracy of each tree. The decision tree that minimizes the expected error rate is preferred.

**42. How can one handle suspicious or missing data in a dataset while performing the** **analysis?**

If there are any inconsistencies or uncertainty in the data set, a user can proceed to utilize any of the accompanying techniques: Creation of a validation report with insights regarding the data in conversation Escalating something very similar to an experienced Data Analyst to take a look at it and accept a call Replacing the invalid information with a comparing substantial and latest data information Using numerous methodologies together to discover missing values and utilizing approximation estimate if necessary.

**43.What is the simple difference between Principal Component Analysis (PCA) and Factor Analysis (FA)?**

Among numerous differences, the significant difference between PCA and FA is that factor analysis is utilized to determine and work with the variance between variables, but the point of PCA is to explain the covariance between the current segments or variables.

**44**. **What is the difference between Data Mining and Data Analysis?**

| **Data Mining** | **Data Analysis** |
| --- | --- |
| Used to perceive designs in data stored. | Used to arrange and put together raw information in a significant manner. |
| Mining is performed on clean and well-documented. | The analysis of information includes Data Cleaning.  So, information is not available in a well-documented format. |
| Results extracted from data mining are difficult to interpret. | Results extracted from information analysis are not difficult to interpret. |

**45.** **What is the difference between Data Mining and Data Profiling?**

* **Data Mining:** Data Mining refers to the analysis of information regarding the discovery of relations that have not been found before. It mainly focuses on the recognition of strange records, conditions, and cluster examination.
* **Data Profiling**: Data Profiling can be described as a process of analyzing single attributes of data. It mostly focuses on giving significant data on information attributes, for example, information type, recurrence, and so on.

**46.** **What are the important steps in the data validation process?**

As the name proposes Data Validation is the process of approving information. This progression principally has two methods associated with it. These are Data Screening and Data Verification.

* **Data Screening**: Different kinds of calculations are utilized in this progression to screen the whole information to discover any inaccurate qualities.
* **Data Verification:** Each and every presumed value is assessed on different use-cases, and afterward a final conclusion is taken on whether the value must be remembered for the information or not.

**47. What is the difference between univariate, bivariate**,**and multivariate analysis?**

The main difference between univariate, bivariate, and multivariate investigation are as per the following:

* **Univariate**: A statistical procedure that can be separated depending on the check of factors required at a given instance of time.
* **Bivariate**: This analysis is utilized to discover the distinction between two variables at a time.
* **Multivariate**: The analysis of multiple variables is known as multivariate. This analysis is utilized to comprehend the impact of factors on the responses.

**48. What is the difference between variance and covariance?**

Variance and Covariance are two mathematical terms that are frequently in the Statistics field. Variance fundamentally processes how separated numbers are according to the mean. Covariance refers to how two random/irregular factors will change together. This is essentially used to compute the correlation between variables.

**49. What are different types of Hypothesis Testing?**

The various kinds of hypothesis testing are as per the following:

* **T-test**: A T-test is utilized when the standard deviation is unknown and the sample size is nearly small.
* **Chi-Square Test for Independence:** These tests are utilized to discover the significance of the association between all categorical variables in the population sample.
* **Analysis of Variance (ANOVA):** This type of hypothesis testing is utilized to examine contrasts between the methods in different clusters. This test is utilized comparatively to a T-test but, is utilized for multiple groups.

Welch’s T-test: This test is utilized to discover the test for equality of means between two testing sample tests.

**50. Why should we use data warehousing and how can you extract data for analysis?**

Data warehousing is a key technology on the way to establishing business intelligence. A data warehouse is a collection of data extracted from the operational or transactional systems in a business, transformed to clean up any inconsistencies in identification coding and definition, and then arranged to support rapid reporting and analysis.

*Here are some of the benefits of a data warehouse:*

* It is separate from the operational database.
* Integrates data from heterogeneous systems.
* Storage a huge amount of data, more historical than current data.
* Does not require data to be highly accurate.

### Bonus Interview Questions & Answers

**1. What is Visualization?**

Visualization is for the depiction of data and to gain intuition about the data being observed. It assists the analysts in selecting display formats, viewer perspectives, and data representation schema.

**2. Give some data mining tools?**

* DBMiner
* GeoMiner
* Multimedia miner
* WeblogMiner

**3. What are the most significant advantages of Data Mining?**

There are many advantages of Data Mining. Some of them are listed below:

* Data Mining is used to polish the raw data and make us able to explore, identify, and understand the patterns hidden within the data.
* It automates finding predictive information in large databases, thereby helping to identify the previously hidden patterns promptly.
* It assists faster and better decision-making, which later helps businesses take necessary actions to increase revenue and lower operational costs.
* It is also used to help data screening and validating to understand where it is coming from.
* Using the Data Mining techniques, the experts can manage applications in various areas such as Market Analysis, Production Control, Sports, Fraud Detection, Astrology, etc.
* The shopping websites use Data Mining to define a shopping pattern and design or select the products for better revenue generation.
* Data Mining also helps in data optimization.
* Data Mining can also be used to determine hidden profitability.

**4. What are ‘Training set’ and ‘Test set’?**

In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. The training set is an example given to the learner, while the Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of examples held back from the learner. The training set is distinct from the Test set.

**5. Explain what is the function of ‘Unsupervised Learning?**

* Find clusters of the data
* Find low-dimensional representations of the data
* Find interesting directions in data
* Interesting coordinates and correlations
* Find novel observations/ database cleaning

**6. In what areas Pattern Recognition is used?**

Pattern Recognition can be used in

* Computer Vision
* Speech Recognition
* Data Mining
* Statistics
* Informal Retrieval
* Bio-Informatics

**7. What is ensemble learning?**

To solve a particular computational program, multiple models such as classifiers or experts are strategically generated and combined to solve a particular computational program Multiple. This process is known as ensemble learning. Ensemble learning is used when we build component classifiers that are more accurate and independent of each other. This learning is used to improve classification, prediction of data, and function approximation.

**8. What is the general principle of an ensemble method and what is bagging and boosting in the ensemble method?**

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm to improve robustness over a single model. Bagging is a method in an ensemble for improving unstable estimation or classification schemes. While boosting methods are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

**9. What are the components of relational evaluation techniques?**

The important components of relational evaluation techniques are

* Data Acquisition
* Ground Truth Acquisition
* Cross-Validation Technique
* Query Type
* Scoring Metric
* Significance Test

**10. What are the different methods for Sequential Supervised Learning?**

 The different methods to solve Sequential Supervised Learning problems are

* Sliding-window methods
* Recurrent sliding windows
* Hidden Markow models
* Maximum entropy Markow models
* Conditional random fields
* Graph transformer networks

**11. What is a Random Forest?**

Random forest is a machine learning method that helps you to perform all types of regression and classification tasks. It is also used for treating missing values and outlier values.

**12. What is reinforcement learning?**

Reinforcement Learning is a learning mechanism about how to map situations to actions. The end result should help you to increase the binary reward signal. In this method, a learner is not told which action to take but instead must discover which action offers a maximum reward. This method is based on the reward/penalty mechanism.

**13. Is it possible to capture the correlation between continuous and categorical variables?**

Yes, we can use the analysis of the covariance technique to capture the association between continuous and categorical variables.

**14. What is Visualization?**

Visualization is for the depiction of information and to acquire knowledge about the information being observed. It helps the experts in choosing format designs, viewer perspectives, and information representation patterns.

**15. Name some best tools which can be used for data analysis.**

The most common useful tools for data analysis are:

* Google Search Operators
* KNIME
* Tableau
* Solver
* RapidMiner
* Io
* NodeXL

**16. Describe the structure of Artificial Neural Networks?**

An artificial neural network (ANN) also referred to as simply a “Neural Network” (NN), could be a process model supported by biological neural networks. Its structure consists of an interconnected collection of artificial neurons. An artificial neural network is an adjective system that changes its structure-supported information that flows through the artificial network during a learning section. The ANN relies on the principle of learning by example. There are, however, 2 classical types of neural networks, perceptron and also multilayer perceptron. Here we are going to target the perceptron algorithmic rule.

**17. Do you think 50 small decision trees are better than a large one? Why?**

Yes,50 small decision trees are better than a large one because 50 trees make a more robust model (less subject to over-fitting) and simpler to interpret.

### 1) What is Data Mining? / What do you understand by Data Mining?

Data Mining is a process of extracting usable data from a more extensive set of raw data by using some methods along with machine learning, statistics, and database systems. It implies analyzing data patterns in large batches of data using one or more software. Data mining is a specific subfield of Computer Science and Statistics. The main goal of Data Mining is to extract information (using intelligent methods) from a data set and transform the information into an understandable structure for further use.

Using Data Mining, businesses can learn more about their customers and develop more effective strategies to expand their various business functions and utilize their resources more optimally and insightfully. Data mining consists of useful data collection and warehousing as well as computer processing. It makes businesses to attain their objective and makes better decisions.

### 2) What are the key features of Data Mining?

Data mining has many applications in multiple fields, like science and research. Following is the list of key features of Data Mining:

* By trend and behavior analysis of the data, we can create automatic pattern predictions.
* We can create decision-oriented information.
* We can focus on large data sets and databases for analysis.
* We can predict the behavior based on the outcomes.
* Clustering based on finding and visually documented groups of facts not previously known.

### 3) 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.

****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

### 4) What is the difference between Data Mining and Data Warehousing?

Data Warehousing mainly focuses on extracting data from different sources, cleaning the data, and storing it in the warehouses. On the other hand, Data Mining is used to study and explore the data using queries. In this process, the meaning pattern or data is extracted. We can also fire these queries on the data warehouses. After Data Mining, the explored information is used to report, plan strategies, find meaningful patterns, etc.

****Example:**** A company's data warehouse stores all the relevant information of projects and employees. We can apply Data Mining queries to this data warehouse to get useful records.

### 5) What are the different types of Data Mining?

We can classify Data Mining into the following types:

* Selection
* Integration
* Data cleaning
* Pattern evaluation
* Data transformation
* Knowledge representation etc.

### 6) 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.

### 7) What do you understand by Data Purging?

Data Purging is a process that is used in database management systems to maintain relevant data in a database. It is used to clean the junk data by eliminating or deleting the row and columns' unnecessary NULL values. It is essential because whenever we need to load new data in the database, we have to purge the irrelevant data from the database.

Using Data Purging of the database frequently, we can remove the junk data that takes up a fair amount of database memory and slow down the database's performance. So, we can say that data purging is mandatory when the database's size gets too large.

### 8) What are cubes in Data Mining?

In Data Mining, cubes or data cubes are used to store data in a summarized version to analyze this faster when required. The data is stored in such a way that reporting becomes very easy.

For example, Organizations use data cubes to analyze the weekly or monthly performance of their employees. Here, month and week are considered as the dimensions of the cube.

### 9) What is the difference between OLAP and OLTP?

The terms OLAP and OLTP look similar but refer to different kinds of systems. We can divide an IT system into two categories: Analytical Process and Transactional Process.

|  |  |
| --- | --- |
| **OLAP** | **OLTP** |
| OLAP stands for Online Analytical Process. | OLTP stands for Online Transactional Process. |
| OLAP process consists of complex queries that are applied to large amounts of historical data aggregated from OLTP databases and other sources. | The OLTP process captures and maintains transaction data in a database. |
| This process is mainly used in data mining, analytics, and business intelligence projects. | In this process, each transaction involves individual database records made up of multiple fields or columns. For example, banking and credit card activity or retail checkout scanning. |
| In OLAP, the main focus is on response time to these complex queries. Each query involves one or more columns of data aggregated from many rows. | In OLTP, the main focus is on fast processing because OLTP databases are read, written, and updated frequently. If a transaction fails, built-in system logic ensures data integrity. |
| Low volumes of transactions categorize OLAP. | Short online transactions categorize OLTP. |
| An example of OLAP is the year-over-year financial performance or marketing lead generation trends of an organization. | An example of OLTP is banking and credit card activity or retail checkout scanning. |
| The query failure in OLAP does not interrupt or delay transaction processing for customers, but it can delay or impact business intelligence insights' accuracy. | The OLTP databases are read, written, and updated frequently, so if a transaction fails, built-in system logic ensures data integrity. |

### 10) 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

There are some advantages and disadvantages of using the above storage models.

### 11) 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.

### 12) 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.

### 13) What are the advantages and disadvantages of using the HOLAP storage model?

The term HOLAP stands for "Hybrid Online Analytical Processing." It is a combination of MOLAP and ROLAP. This is a hybrid storage model and was built to overcome the MOLAP and ROLAP storage model's limitations.

****Advantages of using the HOLAP storage model:****

* It provides better accessibility in comparison to both ROLAP & MOLAP storage models.
* Because of its cache facility, the querying is faster in this storage model.
* The query performance is moderate. It is faster than ROLAP but slower than MOLAP.
* Its cubes are smaller than MOLAP, so only precise data is fetched for processing.
* It is best when data volume is expected to increase over time.
* Its processing ability is higher as compared to ROLAP and MOLAP systems.

****Disadvantages of using HOLAP storage model:****

* In this storage model, both ROLAP and MOLAP are combined to form HOLAP, so the data volume is large.
* It occupies a lot of storage space, as it contains the data from relational databases and multidimensional databases.
* The processing speed is slow while querying.
* It requires system processing whenever data is updated, inserted, or deleted in the database.
* We need to update the cache whenever an update happens in the database associated with the stored queries and relational data.
* Maintenance is complex in this storage model because it quite often updates.

### 14) What are the different problems that "Data Mining" can solve?

Data Mining can solve the following types of problems:

* Data Mining is mainly used to analyze data and make faster business decisions, increasing revenue with lower costs.
* Data Mining also helps to understand, explore and identify patterns of data.
* Data Mining is used to automate the process of finding predictive information in large databases.
* It is used to identify previously hidden patterns.

### 15) What is Discrete and Continuous data in Data Mining?

In Data Mining, discreet 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 discreet 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.

### 16) What do you understand by a model in Data Mining?

In Data Mining, models help the different algorithms in decision making or pattern matching. In the second stage of Data Mining, we consider various models and choose the best one according to their predictive performance.

### 17) How do Data Mining and Data Warehousing work together?

Generally, Data Mining and Data Warehousing work together. Data Warehousing is used to analyze the business needs by storing data in a meaningful form, and Data Mining is used to forecast the business needs. So, here Data Warehouse can act as a source of this forecasting.

### 18) What are the different stages used in "Data Mining"?

Following are the three different stages used in Data Mining:

* ****Exploration:**** Exploration is the first stage of Data Mining. This stage involves the preparation and collection of different data sets like cleaning, transformation, etc. Based on different types of available data sets, various tools are used to analyze the data.
* ****Model building and validation:**** This is the validation stage where the data sets are validated by applying different models by comparing the data sets for best performance. This particular step is called pattern identification. This is a critical process because the user has to identify which pattern is best suitable for easy predictions.
* ****Deployment:**** This is the last stage where the best-chosen pattern is applied for the data sets. It is used to generate predictions, and it helps in estimating expected outcomes.

### 19) What is a Model in the field of Data Mining?

Model is an essential factor in Data Mining activities. It is used to define algorithms that help in decisions making and pattern matching.

### 20) What is the Naive Bayes Algorithm in Data Mining?

The Naive Bayes Algorithm is widely used in Data Mining to generate mining models. After that, these generated models are generally used to identify the relationship between the input columns and the predicated available columns. This algorithm is mainly used during the initial stages of the explorations.

### 21) What is Clustering Algorithm in Data Mining?

In Data Mining, the clustering algorithm is used to group sets of data with similar characteristics (also known as clusters). By the use of these clusters, we can make faster decisions and explore data. First, this algorithm identifies the relationships in a dataset, and then it generates a series of clusters based on the relationships. The process of creating clusters is also repetitive.

### 22) Which are the most popular areas of applications of Data Mining?

Following is the list of the most popular area of application of Data Mining Applications for Finance.

* Healthcare
* Intelligence
* Telecommunication
* Energy
* Retail
* E-commerce
* Supermarkets
* Crime Agencies
* Businesses Benefit from Data Mining

### 23) Explain the time series algorithm in Data Mining?

In Data Mining, the time series algorithm is mainly used for that type of data where the values are changed continuously based on time. For example, age.

This algorithm is used to predict the data set and then keep track of the continuous data and successfully choose the correct data. It also generates a specific model to predict the data's future trends based on the entire original data sets.

### 24) What do you understand by DMX in the context of Data Mining?

DMX is an acronym that stands for Data Mining Extensions. It is a query language for Data Mining models supported by Microsoft's SQL Server Analysis Services product. Same as SQL also supports a data definition language, data manipulation language, and a data query language, all three with SQL-like syntax.

* ****Data Definition:**** This is used to define and create new models and structures.
* ****Data Manipulation:**** This is used to manipulate data based on the requirement.

### 25) What are the different functions of Data Mining?

Following is the list of different functions of Data Mining:

* Characterization
* Association and correlation analysis
* Classification
* Prediction
* Cluster analysis
* Evolution analysis
* Sequence analysis

### 26) What do you understand by data aggregation and data generalization?

****Data Aggregation:**** Data aggregation is a process where data is aggregated altogether, and we can construct a cube for data analysis purposes.

****Data generalization:**** Data generalization is a process where high-level data replace low-level data to make it more meaningful and generalized.

### 27) What do you understand by Data Mining Interface?

The Data Mining Interface is used to improve the quality of the queries we use in Data Mining. It is nothing but a GUI form for Data Mining activities.

### 28) What do you understand by the term Cluster Analysis?

In the context of Data Mining, the term cluster analysis is an important type of analysis that is used in market research, pattern recognition, data analysis, and image processing, etc.

### 29) 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.

### 30) What are the most significant advantages of Data Mining?

There are many advantages of Data Mining. Some of them are listed below:

* Data Mining is used to polish the raw data and make us able to explore, identify, and understand the patterns hidden within the data.
* It automates finding predictive information in large databases, thereby helping to identify the previously hidden patterns promptly.
* It assists faster and better decision making, which later helps businesses take necessary actions to increase revenue and lower operational costs.
* It is also used to help data screening and validating to understand where it is coming from.
* Using the Data Mining techniques, the experts can manage applications in various areas such as Market Analysis, Production Control, Sports, Fraud Detection, Astrology, etc.
* The shopping websites use Data Mining to define a shopping pattern and design or select the products for better revenue generation.
* Data Mining also helps in data optimization.
* Data Mining can also be used to determine hidden profitability.

Because of the above reasons, Data Mining has become very popular nowadays and used by numerous industries, including marketing, advertising, IT/ITES, business intelligence, and even government intelligence organizations.

### 31) What are the most significant disadvantages of Data Mining?

Besides a lot of advantages, Data Mining has some disadvantages too. Following is the list of some of them:

****Security Issues****

Security is the biggest issue of Data Mining. Companies have information about their employees and customers, including social security numbers, birthdays, payroll, etc. However, this is always in the question that how they take care of this information. Hackers can access and steal customers' information, including personal and financial information, and may misuse it.

****Privacy Issues****

Due to Data Mining, concerns about personal privacy have been increasing enormously recently, especially in the age of the internet with social networks, e-commerce, online banking, etc. People can lose their personal and confidential information, which can cost them big troubles.

****Misuse of information/inaccurate information****

Data Mining doesn't ensure you give the correct information always. Information collected through Data Mining can be intended for ethical purposes and be misused. Hackers or unethical businesses can exploit people by using this information.

### 32) Which are the main prominent fields and areas where Data Mining is used?

Data Mining is mainly used in the following fields:

****Finance & Banking Sectors****

Data Mining is very important in the finance & banking field because data extraction provides financial institutions information on loans and credit reports. It facilitates us to create a model for historic customers by determining their good or bad credits. It is also used to detect fraudulent transactions by credit cards that protect a credit card owner.

****Marketing & Retails****

Marketing companies use Data Mining to create models based on the shopping history of their customers. By using this technique, they can sell profitable products to their targeted customers.

****Increasing Brand Loyalty****

Companies use Data Mining techniques in marketing campaigns after understanding their customers' needs and habits. After getting the right information, the companies can quickly increase their brand loyalty.

****Helps in Decision Making****

Companies use Data Mining techniques to help them in making some decisions in marketing or business. By using this technology, it is effortless to determine all information. Also, the company can decide what is unknown and unexpected.

****To Predict Future Trends****

Data Mining can be used to predict future trends by studying the data patterns for a long time. It can also help people to adopt behavioral changes.

****Increase Company Revenue****

Data mining technology involves collecting information on goods sold online. This can eventually reduce the cost of products and increase the company revenue.

****Determining Customer Groups****

Data Mining provides market analysis so we can get a response directly from customers. It also includes information during the identification of customer groups.

****Increases Website Optimization****

Data Mining can find all kinds of unseen element information, which can help you optimize your website.

### 33) What are the required technological drivers in Data Mining?

In Data Mining, we have to deal with mainly two things, database size, and query complexity.

* ****Database size:**** In Data Mining, we have to maintain and process a vast amount of data, so we must have a robust system with enough storage space.
* ****Query Complexity:**** To analyze the complex and large number of queries, we must require a powerful system with enough RAM.

#### **1. Explain the techniques of Data Mining?**

****Answer:****  
The techniques are sequential patterns, prediction, regression analysis, clustering analysis, classification analysis, association rule learning, anomaly or outlier detection, and decision trees.

#### **2. Explain the advantages of data mining?**

****Answer:****  
Data mining’s main [advantage](https://www.educba.com/advantages-of-data-mining/)is using this in Banks and other financial companies or institutions to check out the defaulters based on last transactions of users and behaviour patterns. It is also used for sending or pushing the correct advertisements over the internet. Based on [machine learning algorithms](https://www.educba.com/machine-learning-algorithms/), the web pages are displayed based on a user’s previous history and interests or search over the internet.

Let us move to the next Data Mining Interview Questions

#### **3. Explain the scope of data mining?**

****Answer:****  
The scope of data mining is an automated prediction of trends and behaviours, automatic discovery of previously unknown patterns. It is used to automate the process of finding predictive information in large databases. [Data mining tools](https://www.educba.com/data-mining-tool/) are used to sweep through databases. It is also being used to identify the previously hidden patterns.

#### **4. List out the types of data mining?**

****Answer:****  
This is the basic Data Mining Interview Questions asked in an interview. Integration, selection, data cleaning, data transformation, pattern evaluation, and knowledge representation are data mining types.

#### **5. Explain the difference between Data Mining and Data Warehousing?**

****Answer:****  
Data mining processes, where it explores the data using queries, or it means to analyse the data and analyze the results or output. This helps in reporting, strategy planning and visualizing the meaningful data sets. Data warehousing is a process where the data is extracted from the various resources, and after that, it is being verified and stored.

### **Part 2 – Data Mining Interview Questions (Advanced)**

Let us now have a look at the advanced Data Mining Interview Questions And Answers.

#### **6. Can you please tell, which problems, in general, the data mining can solve?**

****Answer:****  
Data mining is a critical process because it is being used to validate and shortlist the data from the large volume of data of the system or organizations. How the data is flowing and what is the process, it can be defined based on data mining results. Data mining is widely used in industries like marketing, services, [artificial intelligence (AI)](https://www.educba.com/careers-in-artificial-intelligence/), government intelligence (GI) and advertising. There are other industries like telecom, E-commerce, healthcare, energy, biological data analysis, crime agencies, retail, information retrieval like communication systems, education, and sales.

#### **7. Explain the use of data mining queries or why data mining queries are more helpful?**

****Answer:****  
The data mining queries mainly helped apply the model to the new data to make single or multiple results. It also allows us to provide input values such as parameters in batch. The query can retrieve the cases more effectively, which fits a particular pattern. It gets the training data’s statistical memory and helps get the exact design and rule of the typical case representing a pattern in the model. It helps in extracting the regression formulas and other calculation that explain ways. It also retrieves the details about the individual cases used in the model. It includes the data which is not used in the analysis, and generally, it retains the model with the help of adding fresh information and perform the task and cross-verified.

Let us move to the next Data Mining Interview Questions.

#### **8. Explain clustering in data mining?**

****Answer:****  
Clustering in Data Mining is referred to as a group of abstract objects into classes of similar items is made. In data mining, a cluster of data objects is treated as one group, and while doing the cluster analysis, partition of data is done into groups. The groups are labelled based on similar data. Data clustering is used in image processing, data analysis, pattern recognition and other like market research. It helps identify areas and classifies the document based on the collected data over search information through a web or any other medium. It is mainly used for detecting applications to check the fraud of online transactions. Cluster analysis is required in data mining because of its scalability, ability to deal with different kinds of attributes, interpretability, ability to deal with messy data, and highly dimensional data.

#### **9. What is a machine learning-based approach to data mining?**

****Answer:****  
This is the advanced Data Mining Interview Questions asked in an interview. Machine learning is mainly used in data mining because it covers automatic computing procedures, and it was based on logical or binary operations. We have to focus on decision-tree approaches, and the results are mainly evolved from the logical sequence of steps. Machine learning generally follows the principle that would allow us to deal with more general data types, including cases and in these types and number of attributes may vary. Machine learning is one of the popular technique used for data mining and in Artificial intelligence as well.

#### **10. Explain the major elements of Data Mining?**

****Answer:****  
Data mining mainly helps extract the information, transform and load data into the data warehouse system. It mainly stores and manages the data in a multi-dimensional based database management system. It analyses the data by application software and shows that in a useful format, and this data mainly accessed by the professionals or business analysts.

****Q.1. What are foundations of data mining?****

Generally, we use it for a long process of research and product development. Also, we can say this evolution was started when business data was first stored on computers. We can also navigate through their data in real time. Data Mining is also popular in the business community. As this is supported by three technologies that are now mature: Massive data collection, Powerful multiprocessor computers, and Data mining algorithms.

**[Read to know more about Data Mining](https://data-flair.training/blogs/data-mining/)**

****Q.2. What is the scope of data mining?****

* Automated prediction of trends and behaviours- We use to automate the process of finding predictive information in large databases. Also, questions that required extensive hands-on analysis can now be answered from the data. Moreover, targeted marketing is a typical example of predictive marketing. As we also use data mining on past promotional mailings.
* Automated discovery of previously unknown patterns – As we use data mining tools to sweep through databases. Also, to identify previously hidden patterns in one step. Basically, there is a very good example of pattern discovery. As it is the analysis of retail sales data. Moreover, that is to identify unrelated products that are often purchased together.

****Q.3 What are advantages of data mining?****

Basically, to find probable defaulters, we use data mining in banks and financial institutions. Also, this is done based on past transactions, user behaviour and data patterns.

Generally, it helps advertisers to push the right advertisements to the internet. Also, it surfer on web pages based on machine learning algorithms. Moreover, this way data mining benefit both possible buyers as well as sellers of the various products.

Basically, the retail malls and grocery stores peoples used it. Also, it is to arrange and keep most sellable items in the most attentive positions.

**[Read more about data Mining Advantages](https://data-flair.training/blogs/advantages-of-data-mining/)**

****Q.4. What are the cons of data mining?****

Security: The time at which users are online for various uses, must be important. They do not have security systems in place to protect us. As some of the data mining analytics use software. That is difficult to operate. Thus they require a user to have knowledge based training. The techniques of data mining are not 100% accurate. Hence, it may cause serious consequences in certain conditions.  
**[Read more about data mining Disadvantages](https://data-flair.training/blogs/disadvantages-of-data-mining/)**

****Q.5 Name Data mining techniques?****

a. Classification Analysis

b. Association Rule Learning

c. Anomaly or Outlier Detection

d. Clustering Analysis

e. Regression Analysis

f. Prediction

g. Sequential Patterns

h. Decision trees

**[Read more about Data Mining Techniques](https://data-flair.training/blogs/data-mining-techniques/)**

****Q.6. Give a brief introduction to data mining process?****

Basically, data mining is the latest technology. Also, it is a process of discovering hidden valuable knowledge by analyzing a large amount of data. Moreover. we have to store that data in different databases. As data mining is a very important process. It becomes an advantage for various industries.  
**[Read more about Data Mining Process](https://data-flair.training/blogs/data-mining-process/)**

****Q.7. Name types of data mining?****

a. Data cleaning

b. Integration

c. Selection

d. Data transformation

e. Data mining

f. Pattern evaluation

g. Knowledge representation

****Q.8. Name the steps used in data mining?****

a. Business understanding

b. Data understanding

c. Data preparation

d. Modeling

e. Evaluation

f. Deployment

****Q.9. Name areas of applications of data mining?****

a. Data Mining Applications for Finance

b. Healthcare

c. Intelligence

d. Telecommunication

e. Energy

f. Retail

g. E-commerce

h. Supermarkets

i. Crime Agencies

j. Businesses Benefit from data mining

**[Read more applications of data mining](https://data-flair.training/blogs/data-mining-applications/)**

****Q.10. What is required technological drivers in data mining?****

Database size: Basically, as for maintaining and processing the huge amount of data, we need powerful systems.

Query Complexity: Generally, to analyze the complex and large number of queries, we need a more powerful system.

****Data Mining Interview Questions Answers for Freshers – Q. 1,2,3,4,5,7,8,9****

****Data Mining Interview Questions Answers for Experience – Q. 6,10****

****Q.11. Give an introduction to data mining query language?****

It was proposed by Han, Fu, Wang, et al. for the DBMiner data mining system. Although, it was based on the Structured Query Language. These query languages are designed to support ad hoc and interactive data mining. Also, it provides commands for specifying primitives. We can use DMQL to work with databases and data warehouses as well. We can also use it to define data mining tasks. Particularly we examine how to define data warehouses and data marts in DMQL.

**[Read more about data query language](https://data-flair.training/blogs/data-mining-query-language/)**

****Q.12. What is Syntax for Task-Relevant Data Specification?****

The Syntax of DMQL for specifying task-relevant data −  
use database database\_name  
or  
use data warehouse data\_warehouse\_name  
in relevance to att\_or\_dim\_list  
from relation(s)/cube(s) [where condition]  
order by order\_list  
group by grouping\_list

****Q.13. What is Syntax for Specifying the Kind of Knowledge?****

Syntax for Characterization, Discrimination, Association, Classification, and Prediction.

****Q.14. Explain Syntax for Interestingness Measures Specification?****

Interestingness measures and thresholds can be specified by the user with the statement − with <interest\_measure\_name> threshold = threshold\_value

****Q.15. Explain Syntax for Pattern Presentation and Visualization Specification?****

Generally, we have a syntax, which allows users to specify the display of discovered patterns in one or more forms. display as <result\_form>

****Q.16. Explain Data Mining Languages Standardization?****

This will serve the following purposes −

* Basically, it helps the systematic development of data mining solutions.
* Also, improves interoperability among multiple data mining systems and functions.
* Generally, it helps in promoting education and rapid learning.
* Also, promotes the use of data mining systems in industry and society.

****Q.17. Explain useful data mining queries?****

* First of all, it helps to apply the model to new data, to make single or multiple predictions.
* Also, you can provide input values as parameters, or in a batch.
* While it gets a statistical summary of the data used for training. Also, extract patterns and rule of the typical case representing a pattern in the model.
* Also, helps in extracting regression formulas and other calculations that explain patterns.
* Get the cases that fit a particular pattern.
* Further, it retrieves details about individual cases used in the model.
* Also, it includes data not used in the analysis. Moreover, it retrains a model by adding new data or perform cross-prediction.

****Q.18. Give a brief introduction to data mining knowledge discovery?****

Generally, most people don’t differentiate data mining from knowledge discovery. While others view data mining as an essential step in the process of knowledge discovery.

**[Read more about Data Mining From Knowledge Discovery](https://data-flair.training/blogs/data-mining-and-knowledge-discovery/)**

****Q.19. Explain steps involved in data mining knowledge process?****

Data Cleaning −

Basically, in this step, the noise and inconsistent data are removed.

Data Integration −

Moreover, in this step, multiple data sources are combined.

Data Selection −

Furthermore, in this step, data relevant to the analysis task are retrieved from the database.

Data Transformation −

Basically, in this step, data is transformed into forms appropriate for mining. Also, by performing summary or aggregation operations.

Data Mining −

In this, intelligent methods are applied in order to extract data patterns.

Pattern Evaluation −

While, in this step, data patterns are evaluated.

Knowledge Presentation −

Generally, in this step, knowledge is represented

****Q.20. What are issues in data mining?****

A number of issues that need to be addressed by any serious data mining package

Uncertainty Handling

Dealing with Missing Values

Dealing with Noisy data

Efficiency of algorithms

Constraining Knowledge Discovered to only Useful

Incorporating Domain Knowledge

Size and Complexity of Data

Data Selection

Understandably of Discovered Knowledge: Consistency between Data and Discovered Knowledge.

****Data Mining Interview Questions Answers for Freshers – Q. 11,16,17,18,19****

****Data Mining Interview Questions Answers for Experience – Q. 12,13,14,15,20****

****Q.21. What are major elements of data mining, explain?****

Generally, helps in an extract, transform and load transaction data onto the data warehouse system.

While it stores and manages the data in a multidimensional database system.

Also, provide data access to business analysts and information technology professionals.

Generally, analyze the data by application software.

While, it shows the data in a useful format, such as a graph or table

****Q.22. Name different level of analysis of data mining?****

a. **[Artificial Neural Networks](https://data-flair.training/blogs/artificial-neural-network/)**

b. Genetic algorithms

c. Nearest neighbor method

d. Rule induction

e Data visualization

****Q.23. Name methods of classification methods?****

a. Statistical Procedure Based Approach

b **[Machine Learning](https://data-flair.training/blogs/machine-learning-tutorial/)** Based Approach

c. **[Neural Network](https://data-flair.training/blogs/learning-rules-in-neural-network/)**

d. Classification Algorithms

e. ID3 Algorithm

f. C4.5 Algorithm

g. K Nearest Neighbors Algorithm

H. Naïve Bayes Algorithm

i. SVM Algorithm

J. ANN Algorithm

K. 48 Decision Trees

l. Support Vector Machines

M. SenseClusters (an adaptation of the K-means clustering algorithm)

**[Read more about Data Mining Classification](https://data-flair.training/blogs/classification-algorithms/)**

****Q.24. Explain Statistical Procedure Based Approach?****

Especially, there are two main phases present to work on classification. Also, it can be easily identified within the statistical community.

While, the second, “modern” phase concentrated on more flexible classes of models. Also, in which many of which attempt has to take. Moreover, it provides an estimate of the joint distribution of the feature within each class. Further, that can, in turn, provide a classification rule.

Generally, statistical procedures have to characterize by having a precise fundamental probability model and that is used to provides a probability of being in each class instead of just a classification.

Also, we can assume that the techniques will use by statisticians. Hence some human involvement has to assume with regard to variable selection.

Also, transformation and overall structuring of the problem.

****Q.25. Explain Machine Learning Based Approach?****

Generally, it covers automatic computing procedures. Also, it was based on logical or binary operations. Further, we use to learn a task from a series of examples.

Here, we have to focus on decision-tree approaches. Also, ss classification results come from a sequence of logical steps.

Also, its principle would allow us to deal with more general types of data including cases. While, the number and type of attributes may vary.

****Q.26. Explain ID3 Algorithm?****

Generally, the id3 calculation starts with the original set as the root hub. Also, on every cycle, it emphasizes through every unused attribute of the set and figures. Moreover, the entropy of attribute. Furthermore, at that point chooses the attribute. Also, it has the smallest entropy value.

****Q.27. Name methods of clustering?****

They are classified into the following categories −

* Partitioning Method
* Hierarchical Method
* Density-based Method
* Grid-Based Method
* Model-Based Method
* Constraint-based Method

**[Read more about Data Mining Clustering](https://data-flair.training/blogs/cluster-analysis-data-mining/)**

****Q.28. What do OLAP and OLTP stand for?****

Basically, OLAP is an acronym for Online Analytical Processing and OLTP is an acronym for Online Transactional Processing.

****Q.29. Define metadata?****

Basically, metadata is simply defined as data about data. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

****Q.30. List the types of OLAP server?****

Basically, there are four types of OLAP servers, namely Relational OLAP, Multidimensional OLAP, Hybrid OLAP, and Specialized SQL Servers.

### **1. Can you name the data mining techniques? Which one is your preferred technique?**

A hiring manager may ask you to name the data mining techniques because they want to know you're familiar with different approaches, such as classification or regression analysis. A certain role may also require you to specialize in a specific type or have skills in several procedures. They may also desire to know your preferred technique to mine data.

To answer this question effectively, you can address the following points in your response:

Describe the primary components of each data mining technique.

Discuss your preferred technique for data mining tasks.

Share how you discovered your preferred technique.

**Example:** *"There are eight common data mining techniques. For example, we can use classification analysis to evaluate different categories of data, such as an email platform that uses computer algorithms to sort messages into an inbox versus a spam folder. In the retail industry, we might use association learning instead to research and predict purchasing patterns. When developing a fraud sensor network, we can apply the anomaly detection technique to find outlying data points. We can use cluster analysis to make customer profiles, as it allows us to research commonalities between datasets.*

*We can use regression analysis to predict the value of a dependent value, which can help forecast sales. Prediction involves evaluating the relationship between dependent and independent variables, like sales versus costs. The sequential patterns technique identifies similar data arrangements, which can help us study the purchasing habits of different demographics. Decision trees involve organizing information into a chart to determine the best answer for a complex question. I prefer using cluster analysis because it can help me evaluate large datasets quickly. I discovered this preference after increasing my efficiency during the holiday season at my previous job."*

**Related:** **[10 Data Types (With Definitions and Examples)](https://www.indeed.com/career-advice/career-development/data-type-examples" \t "https://www.indeed.com/career-advice/interviewing/_blank)**

### **2. If we were to hire you for this position, how would you use data mining in marketing and sales?**

During an interview for a marketing or sales job, a hiring manager may ask you how you can use data mining to accomplish a role's job responsibilities. They often ask you this question to learn whether you understand your primary duties and their general purpose. It can also allow them to discuss the specifications of the role with you after perceiving your response.

To answer this question, you can use the following points to help plan your answer:

Use relevant keywords from the job description involving data mining methods.

Share how you're prepared to accomplish marketing and sales tasks.

Explain why your approach can benefit a company's specific industry or field.

**Example:** *"Your job description online stated that Global Consumer Enterprises is looking for a professional who helps analyze consumer data so the company can analyze its customers' transactional history. To accomplish this, I would use the prediction technique to forecast the future sales of the company during each season. I may also apply association rule learning to better understand a consumer's buying habits of certain products.*

*For example, specific patterns can indicate whether a customer is more likely to buy more products before a busy season or if they're afterward. They can also show which demographic may purchase more products during a certain timeframe. Using these data mining techniques, I can help a marketing team create personalized emails or social media posts for customers. I can also help a sales team learn which customers may purchase certain items or respond favorably to specific selling techniques."*

**Related:** **[What Is Data-Driven Marketing?](https://www.indeed.com/career-advice/career-development/data-driven-marketing" \t "https://www.indeed.com/career-advice/interviewing/_blank)**

### **3. Are there any challenges you have encountered with data mining?**

A hiring manager may ask how you've approached any issues while data mining because they may want to know how you handle unexpected situations in the workplace. They may also want to observe your ability to use logic when resolving a technical issue involving data mining. Answering this question can also show an ability to take accountability, which is a professional quality that hiring managers often value.

To answer this question, it's often important to share your challenges openly while maintaining a positive perspective. Consider including the following information in your response:

Describe a past situation where you encountered a data mining challenge.

Share what you learned during the experience.

Explain how you applied what you learned to a future data mining issue.

**Example:** *"During my first holiday season at my previous position at a retail company, I had to analyze the consumer data from last year to forecast the upcoming season. As a result, I dealt with a lot of noisy data from recent file transfers and software updates. The noisy data came in large quantities and it was rather overwhelming at first. At first, I set myself behind schedule because I kept sifting through all the noisy data to ensure it was meaningless. I didn't want to overlook any useful data it may have obscured.*

*To ensure I didn't miss my deadline, I asked my manager for some extra assistance and advice\*\*. They showed me a data mining method in our software program that allowed me to search for information based on certain criteria, like demographics or sales data. Afterward, I decided to use this method for all future sales predictions I had to complete during this position. This task is sometimes still challenging, but I'm improving in speed and accuracy each time I complete it."*

**Related:** **[How To Overcome Challenges in the Workplace](https://www.indeed.com/career-advice/career-development/overcome-workplace-challenges" \t "https://www.indeed.com/career-advice/interviewing/_blank)**

### **4. What's your preferred method of clustering?**

When you're in an interview, the hiring manager may ask about your preferred way to cluster data, which is a data mining tool that can help you identify similarities between pieces of information in a dataset. As learning how to categorize these materials can often help you make important business decisions, interviewers often ask your preferences to gauge how much knowledge you have of different techniques. They can also observe if you understand which data mining circumstances may require a certain clustering method.

You can answer this question by sharing the following information in your response:

List your top three favorite clustering methods.

Explain how this method works.

Tell the hiring manager why it's your favorite method.

**Example:** *"When I was in college and learning about the clustering methods for the first time, I took a preference to the hierarchal, grid-based and model-based techniques. If I had to choose a single preference, I would choose hierarchal because you can use both the agglomerative and divisive approaches to sort data effectively. What I like about these approaches is that I can work with data according to the needs of a project.*

*For example, agglomerative clustering allows me to organize data clusters until it becomes a large group, which often helps me identify patterns in demographic data. The divisive approach allows me to separate a large data set into smaller groups, which often helps me perform a comprehensive data analysis and make important business decisions. Hierarchal clustering also allows me to better understand and analyze large quantities of data, which is why it's my favorite."*

### **5. How can you use data mining to resolve challenges?**

A hiring manager may ask you this question to understand if you can use data mining to resolve issues for a company. They may also want to understand your ability to problem-solve technical issues and manage an interface sufficiently, as you might use one to make search queries in a job position. To answer this, you can share a few examples of how you can use data mining to solve several types of challenges that may occur in your industry.

Consider sharing the following information in your answer:

Describe general ways you can use data mining to solve issues.

Share an anecdote about a time you solved an issue using data mining techniques.

Explain how this event benefited a company.

**Example:** *"There are several ways you can use data mining to resolve challenges for a company. In my previous position, I often processed the data in clusters to analyze large amounts of similar data faster during busier seasons. Using these techniques helped my team reduce expenses and increase revenue in our department, which allowed my supervisor more opportunities to make high-quality business decisions. Some other ways data mining can help solve challenges can include identifying unusual patterns within a set of data, conductive predictive analysis to make forecasts and discovering hidden data patterns in the system.*

*For example, in my last marketing position, the company's sales decreases after a recent email campaign ended. I used regression analysis and outlier detection to analyze this change and discovered our content only targeted 55% of our audience effectively instead of a projected 75%. I communicated this assessment to my supervisor and recommended that we conduct a more thorough analysis of customer data in the future. They agreed with my decision and our next email campaign increased sales by 10%."*

****Question 1. What Is Data Mining?****

****Answer :****

Data mining is a process of extracting hidden trends within a datawarehouse. For example an insurance dataware house can be used to mine data for the most high risk people to insure in a certain geographial area.

****Question 2. Differentiate Between Data Mining And Data Warehousing?****

****Answer :****

Data warehousing is merely extracting data from different sources, cleaning the data and storing it in the warehouse. Where as data mining aims to examine or explore the data using queries. These queries can be fired on the data warehouse. Explore the data in data mining helps in reporting, planning strategies, finding meaningful patterns etc.

E.g. a data warehouse of a company stores all the relevant information of projects and employees. Using Data mining, one can use this data to generate different reports like profits generated etc.

[Data Center Management Interview Questions](https://www.wisdomjobs.com/e-university/data-center-management-interview-questions.html" \o "Data Center Management Interview Questions)

****Question 3. What Is Data Purging?****

****Answer :****

The process of cleaning junk data is termed as data purging. Purging data would mean getting rid of unnecessary NULL values of columns. This usually happens when the size of the database gets too large.

****Question 4. What Are Cubes?****

****Answer :****

A data cube stores data in a summarized version which helps in a faster analysis of data. The data is stored in such a way that it allows reporting easily.

E.g. using a data cube A user may want to analyze weekly, monthly performance of an employee. Here, month and week could be considered as the dimensions of the cube.

[Data Mining Tutorial](https://www.wisdomjobs.com/e-university/data-mining-tutorial-199.html" \o "Data Mining Tutorial)

****Question 5. What Are Olap And Oltp?****

****Answer :****

An IT system can be divided into Analytical Process and Transactional Process.

OLTP – categorized by short online transactions. The emphasis is query processing, maintaining data integration in multi-access environment.

OLAP – Low volumes of transactions are categorized by OLAP. Queries involve aggregation and very complex. Response time is an effectiveness measure and used widely in data mining techniques.

[Clinical SAS Interview Questions](https://www.wisdomjobs.com/e-university/clinical-sas-interview-questions.html" \o "Clinical SAS Interview Questions)

****Question 6. What Are The Different Problems That "data Mining" Can Solve?****

****Answer :****

• Data mining helps analysts in making faster business decisions which increases revenue with lower costs.  
• Data mining helps to understand, explore and identify patterns of data.  
• Data mining automates process of finding predictive information in large databases.  
• Helps to identify previously hidden patterns.

****Question 7. What Are Different Stages Of "data Mining"?****

****Answer :****

****Exploration:**** This stage involves preparation and collection of data. it also involves data cleaning, transformation. Based on size of data, different tools to analyze the data may be required. This stage helps to determine different variables of the data to determine their behavior.

****Model building and validation:**** This stage involves choosing the best model based on their predictive performance. The model is then applied on the different data sets and compared for best performance. This stage is also called as pattern identification. This stage is a little complex because it involves choosing the best pattern to allow easy predictions.

****Deployment:**** Based on model selected in previous stage, it is applied to the data sets. This is to generate predictions or estimates of the expected outcome.

****Question 8. What Is Discrete And Continuous Data In Data Mining World?****

****Answer :****

Discreet data can be considered as defined or finite data. E.g. Mobile numbers, gender. Continuous data can be considered as data which changes continuously and in an ordered fashion. E.g. age.

****Question 9. What Is Model In Data Mining World?****

****Answer :****

Models in Data mining help the different algorithms in decision making or pattern matching. The second stage of data mining involves considering various models and choosing the best one based on their predictive performance.

****Question 10. How Does The Data Mining And Data Warehousing Work Together?****

****Answer :****

Data warehousing can be used for analyzing the business needs by storing data in a meaningful form. Using Data mining, one can forecast the business needs. Data warehouse can act as a source of this forecasting.

****Question 11. What Is A Decision Tree Algorithm?****

****Answer :****

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.

****Question 12. What Is Naive Bayes Algorithm?****

****Answer :****

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.

****Question 13. Explain Clustering Algorithm?****

****Answer :****

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.

****Question 14. What Is Time Series Algorithm In Data Mining?****

****Answer :****

Time series algorithm can be used to predict continuous values of data. Once the algorithm is skilled to predict a series of data, it can predict the outcome of other series. The algorithm generates a model that can predict trends based only on the original dataset. New data can also be added that automatically becomes a part of the trend analysis.

E.g. Performance one employee can influence or forecast the profit.

****Question 15. Explain Association Algorithm In Data Mining?****

****Answer :****

Association algorithm is used for recommendation engine that is based on a market based analysis. This engine suggests products to customers based on what they bought earlier. The model is built on a dataset containing identifiers. These identifiers are both for individual cases and for the items that cases contain. These groups of items in a data set are called as an item set. The algorithm traverses a data set to find items that appear in a case. MINIMUM\_SUPPORT parameter is used any associated items that appear into an item set.

****Question 16. What Is Sequence Clustering Algorithm?****

****Answer :****

Sequence clustering algorithm collects similar or related paths, sequences of data containing events. The data represents a series of events or transitions between states in a dataset like a series of web clicks. The algorithm will examine all probabilities of transitions and measure the differences, or distances, between all the possible sequences in the data set. This helps it to determine which sequence can be the best for input for clustering.

E.g. Sequence clustering algorithm may help finding the path to store a product of “similar” nature in a retail ware house.

****Question 17. Explain The Concepts And Capabilities Of Data Mining?****

****Answer :****

Data mining is used to examine or explore the data using queries. These queries can be fired on the data warehouse. Explore the data in data mining helps in reporting, planning strategies, finding meaningful patterns etc. it is more commonly used to transform large amount of data into a meaningful form. Data here can be facts, numbers or any real time information like sales figures, cost, meta data etc. Information would be the patterns and the relationships amongst the data that can provide information.

****Question 18. Explain How To Work With The Data Mining Algorithms Included In Sql Server Data Mining?****

****Answer :****

SQL Server data mining offers Data Mining Add-ins for office 2007 that allows discovering the patterns and relationships of the data. This also helps in an enhanced analysis. The Add-in called as Data Mining client for Excel is used to first prepare data, build, evaluate, manage and predict results.

****Question 19. Explain How To Use Dmx-the Data Mining Query Language.****

****Answer :****

Data mining extension is based on the syntax of SQL. It is based on relational concepts and mainly used to create and manage the data mining models. DMX comprises of two types of statements: Data definition and Data manipulation. Data definition is used to define or create new models, structures.

Example:  
CREATE MINING SRUCTURE  
CREATE MINING MODEL  
Data manipulation is used to manage the existing models and structures.  
Example:  
INSERT INTO  
SELECT FROM .CONTENT (DMX)

****Question 20. Explain How To Mine An Olap Cube?****

****Answer :****

A data mining extension can be used to slice the data the source cube in the order as discovered by data mining. When a cube is mined the case table is a dimension.

****Question 21. What Are The Different Ways Of Moving Data/databases Between Servers And Databases In Sql Server?****

****Answer :****

There are several ways of doing this. One can use any of the following options:  
- BACKUP/RESTORE,  
- Dettaching/attaching databases,  
- Replication,  
- DTS,  
- BCP,  
- logshipping,  
- INSERT...SELECT,  
- SELECT...INTO,  
- creating INSERT scripts to generate data.

****Question 22. What Are The Benefits Of User-defined Functions?****

****Answer :****

a. Can be used in a number of places without restrictions as compared to stored procedures.  
b. Code can be made less complex and easier to write.  
c. Parameters can be passed to the function.  
d. They can be used to create joins and also be sued in a select, where or case statement.  
e. Simpler to invoke.

****Question 23. Define Pre Pruning?****

****Answer :****

A tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples.

****Question 24. What Are Interval Scaled Variables?****

****Answer :****

Interval scaled variables are continuous measurements of linear scale. For example, height and weight, weather temperature or coordinates for any cluster. These measurements can be calculated using Euclidean distance or Minkowski distance.

****Question 25. What Is A Sting?****

****Answer :****

Statistical Information Grid is called as STING; it is a grid based multi resolution clustering method. In STING method, all the objects are contained into rectangular cells, these cells are kept into various levels of resolutions and these levels are arranged in a hierarchical structure.

****Question 26. What Is A Dbscan?****

****Answer :****

Density Based Spatial Clustering of Application Noise is called as DBSCAN. DBSCAN is a density based clustering method that converts the high-density objects regions into clusters with arbitrary shapes and sizes. DBSCAN defines the cluster as a maximal set of density connected points.

****Question 27. Define Density Based Method?****

****Answer :****

Density based method deals with arbitrary shaped clusters. In density-based method, clusters are formed on the basis of the region where the density of the objects is high.

****Question 28. Define Chameleon Method?****

****Answer :****

Chameleon is another hierarchical clustering method that uses dynamic modeling. Chameleon is introduced to recover the drawbacks of CURE method. In this method two clusters are merged, if the interconnectivity between two clusters is greater than the interconnectivity between the objects within a cluster.

****Question 29. What Do U Mean By Partitioning Method?****

****Answer :****

In partitioning method a partitioning algorithm arranges all the objects into various partitions, where the total number of partitions is less than the total number of objects. Here each partition represents a cluster. The two types of partitioning method are k-means and k-medoids.

****Question 30. Define Genetic Algorithm?****

****Answer :****

Enables us to locate optimal binary string by processing an initial random population of binary strings by performing operations such as artificial mutation , crossover and selection.

****Question 31. What Is Ods?****

****Answer :****

1. ODS means Operational Data Store.  
2. A collection of operation or bases data that is extracted from operation databases and standardized, cleansed, consolidated, transformed, and loaded into an enterprise data architecture. An ODS is used to support data mining of operational data, or as the store for base data that is summarized for a data warehouse. The ODS may also be used to audit the data warehouse to assure summarized and derived data is calculated properly. The ODS may further become the enterprise shared operational database, allowing operational systems that are being reengineered to use the ODS as there operation databases.

****Question 32. What Is Spatial Data Mining?****

****Answer :****

Spatial data mining is the application of data mining methods to spatial data. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasises the importance of developing data driven inductive approaches to geographical analysis and modeling.

Data mining, which is the 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 organizations possessing huge databases with thematic and geographically referenced data begin to realise the huge potential of the information hidden there. Among those organizations are:

\* offices requiring analysis or dissemination of geo-referenced statistical data  
\* public health services searching for explanations of disease clusters  
\* environmental agencies assessing the impact of changing land-use patterns on climate change  
\* geo-marketing companies doing customer segmentation based on spatial location.

****Question 33. What Is Smoothing?****

****Answer :****

Smoothing is an approach that is used to remove the nonsystematic behaviors found in time series. It usually takes the form of finding moving averages of attribute values. It is used to filter out noise and outliers.

****Question 34. What Are The Advantages Data Mining Over Traditional Approaches?****

****Answer :****

Data Mining is used for the estimation of future. For example if we take a company/business organization by using the concept of Data Mining we can predict the future of business interms of Revenue (or) Employees (or) Cutomers (or) Orders etc.

Traditional approches use simple algorithms for estimating the future. But it does not give accurate results when compared to Data Mining.

****Question 35. What Is Model Based Method?****

****Answer :****

For optimizing a fit between a given data set and a mathematical model based methods are used. This method uses an assumption that the data are distributed by probability distributions. There are two basic approaches in this method that are  
1. Statistical Approach  
2. Neural Network Approach.

****Question 36. What Is An Index?****

****Answer :****

Indexes of SQL Server are similar to the indexes in books. They help SQL Server retrieve the data quicker. Indexes are of two types. Clustered indexes and non-clustered indexes. Rows in the table are stored in the order of the clustered index key.  
There can be only one clustered index per table.  
Non-clustered indexes have their own storage separate from the table data storage.  
Non-clustered indexes are stored as B-tree structures.  
Leaf level nodes having the index key and it's row locater.

****Question 37. Mention Some Of The Data Mining Techniques?****

****Answer :****

* + Statistics
  + Machine learning
  + Decision Tree
  + Hidden markov models
  + Artificial Intelligence
  + Genetic Algorithm
  + Meta learning

****Question 38. Define Binary Variables? And What Are The Two Types Of Binary Variables?****

****Answer :****

Binary variables are understood by two states 0 and 1, when state is 0, variable is absent and when state is 1, variable is present. There are two types of binary variables, symmetric and asymmetric binary variables. Symmetric variables are those variables that have same state values and weights. Asymmetric variables are those variables that have not same state values and weights.

****Question 39. Explain The Issues Regarding Classification And Prediction?****

****Answer :****

****Preparing the data for classification and prediction:****

* + Data cleaning
  + Relevance analysis
  + Data transformation
  + Comparing classification methods
  + Predictive accuracy
  + Speed
  + Robustness
  + Scalability
  + Interpretability

****Question 40. What Are Non-additive Facts?****

****Answer :****

Non-Additive: Non-additive facts are facts that cannot be summed up for any of the dimensions present in the fact table.

****Question 41. What Is Meteorological Data?****

****Answer :****

Meteorology is the interdisciplinary scientific study of the atmosphere. It observes the changes in temperature, air pressure, moisture and wind direction. Usually, temperature, pressure, wind measurements and humidity are the variables that are measured by a thermometer, barometer, anemometer, and hygrometer, respectively. There are many methods of collecting data and Radar, Lidar, satellites are some of them.

Weather forecasts are made by collecting quantitative data about the current state of the atmosphere. The main issue arise in this prediction is, it involves high-dimensional characters. To overcome this issue, it is necessary to first analyze and simplify the data before proceeding with other analysis. Some data mining techniques are appropriate in this context.

****Question 42. Define Descriptive Model?****

****Answer :****

It is used to determine the patterns and relationships in a sample data.

****Data mining tasks that belongs to descriptive model:****

* + Clustering
  + Summarization
  + Association rules
  + Sequence discovery

****Question 43. What Is A Star Schema?****

****Answer :****

Star schema is a type of organising the tables such that we can retrieve the result from the database easily and fastly in the warehouse environment.Usually a star schema consists of one or more dimension tables around a fact table which looks like a star,so that it got its name.

****Question 44. What Are The Steps Involved In Kdd Process?****

****Answer :****

* + Data cleaning
  + Data Mining
  + Pattern Evaluation
  + Knowledge Presentation
  + Data Integration
  + Data Selection
  + Data Transformation

****Question 45. What Is A Lookup Table?****

****Answer :****

A lookUp table is the one which is used when updating a warehouse. When the lookup is placed on the target table (fact table / warehouse) based upon the primary key of the target, it just updates the table by allowing only new records or updated records based on the lookup condition.

****Question 46. What Is Attribute Selection Measure?****

****Answer :****

The information Gain measure is used to select the test attribute at each node in the decision tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split.

****Question 47. Explain Statistical Perspective In Data Mining?****

****Answer :****

* + Point estimation
  + Data summarization
  + Bayesian techniques
  + Hypothesis testing
  + Regression
  + Correlation

****Question 48. Define Wave Cluster?****

****Answer :****

It is a grid based multi resolution clustering method. In this method all the objects are represented by a multidimensional grid structure and a wavelet transformation is applied for finding the dense region. Each grid cell contains the information of the group of objects that map into a cell. A wavelet transformation is a process of signaling that produces the signal of various frequency sub bands.

****Question 49. What Is Time Series Analysis?****

****Answer :****

A time series is a set of attribute values over a period of time. Time Series Analysis may be viewed as finding patterns in the data and predicting future values.

****Question 50. Explain Mining Single ?dimensional Boolean Associated Rules From Transactional Databases?****

****Answer :****

The apriori algorithm: Finding frequent itemsets using candidate generation Mining frequent item sets without candidate generation.

****Question 51. What Is Meta Learning?****

****Answer :****

Concept of combining the predictions made from multiple models of data mining and analyzing those predictions to formulate a new and previously unknown prediction.

****Question 52. Describe Important Index Characteristics?****

****Answer :****

The characteristics of the indexes are:  
\* They fasten the searching of a row.  
\* They are sorted by the Key values.  
\* They are small and contain only a small number of columns of the table.  
\* They refer for the appropriate block of the table with a key value.

****Question 53. What Is The Use Of Regression?****

****Answer :****

Regression can be used to solve the classification problems but it can also be used for applications such as forecasting. Regression can be performed using many different types of techniques; in actually regression takes a set of data and fits the data to a formula.

****Question 54. What Is Dimensional Modelling? Why Is It Important ?****

****Answer :****

Dimensional Modelling is a design concept used by many data warehouse desginers to build thier data warehouse. In this design model all the data is stored in two types of tables - Facts table and Dimension table. Fact table contains the facts/measurements of the business and the dimension table contains the context of measuremnets ie, the dimensions on which the facts are calculated.

****Question 55. What Is Unique Index?****

****Answer :****

Unique index is the index that is applied to any column of unique value.  
A unique index can also be applied to a group of columns.

****Question 56. What Are The Foundations Of Data Mining?****

****Answer :****

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:  
\* Massive data collection  
\* Powerful multiprocessor computers  
\* Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

****Question 57. What Snow Flake Schema?****

****Answer :****

Snowflake Schema, each dimension has a primary dimension table, to which one or more additional dimensions can join. The primary dimension table is the only table that can join to the fact table.

****Question 58. Differences Between Star And Snowflake Schemas?****

****Answer :****

Star schema - all dimensions will be linked directly with a fat table.  
Snow schema - dimensions maybe interlinked or may have one-to-many relationship with other tables.

****Question 59. What Is Hierarchical Method?****

****Answer :****

Hierarchical method groups all the objects into a tree of clusters that are arranged in a hierarchical order. This method works on bottom-up or top-down approaches.

****Question 60. What Is Cure?****

****Answer :****

Clustering Using Representatives is called as CURE. The clustering algorithms generally work on spherical and similar size clusters. CURE overcomes the problem of spherical and similar size cluster and is more robust with respect to outliers.

****Question 61. What Is Etl?****

****Answer :****

ETL stands for extraction, transformation and loading.

ETL provide developers with an interface for designing source-to-target mappings, ransformation and job control parameter.  
\*Extraction  
Take data from an external source and move it to the warehouse pre-processor database.  
\*Transformation  
Transform data task allows point-to-point generating, modifying and transforming data.  
\*Loading  
Load data task adds records to a database table in a warehouse.

****Question 62. Define Rollup And Cube?****

****Answer :****

Custom rollup operators provide a simple way of controlling the process of rolling up a member to its parents values.The rollup uses the contents of the column as custom rollup operator for each member and is used to evaluate the value of the member’s parents.

If a cube has multiple custom rollup formulas and custom rollup members, then the formulas are resolved in the order in which the dimensions have been added to the cube.

****Question 63. What Are The Different Problems That "data Mining" Can Solve?****

****Answer :****

\*Data mining helps analysts in making faster business decisions which increases revenue with lower costs.

\*Data mining helps to understand, explore and identify patterns of data.

\*Data mining automates process of finding predictive information in large databases.

\*Helps to identify previously hidden patterns.

****Question 64. What Are Different Stages Of "data Mining"?****

****Answer :****

Exploration: This stage involves preparation and collection of data. it also involves data cleaning, transformation. Based on size of data, different tools to analyze the data may be required. This stage helps to determine different variables of the data to determine their behavior.

Model building and validation: This stage involves choosing the best model based on their predictive performance. The model is then applied on the different data sets and compared for best performance. This stage is also called as pattern identification. This stage is a little complex because it involves choosing the best pattern to allow easy predictions.

Deployment: Based on model selected in previous stage, it is applied to the data sets. This is to generate predictions or estimates of the expected outcome.

****Question 65. Explain How To Use Dmx-the Data Mining Query Language?****

****Answer :****

Data mining extension is based on the syntax of SQL. It is based on relational concepts and mainly used to create and manage the data mining models. DMX comprises of two types of statements: Data definition and Data manipulation. Data definition is used to define or create new models, structures.

Example:  
CREATE MINING SRUCTURE  
CREATE MINING MODEL

Data manipulation is used to manage the existing models and structures.

Example:  
INSERT INTO  
SELECT FROM .CONTENT (DMX)

### **1. Explain what are the different storage models that are available in OLAP?**

The different storage models that are available in OLAP are as follows:

1. ****MOLAP:**** Multidimensional Online Analytical Processing
2. ****ROLAP:****Relational online Analytical processing
3. ****HOLAP:**** Hybrid Online Analytical Processing

They are advantages and disadvantages of each of these storage models that are available in OLAP.

### **2. Explain in detail what is MOLAP? What are the advantages and disadvantages?**

As the name itself depicts “MOLAP”, i.e. Multidimensional.

In this type of data storage, the data is stored in multidimensional cubes and not in the standard relational databases.

****The advantage of using MOLAP is:****

The query performance is excellent, this is because the data is stored in multidimensional cubes. Also, the calculations are pre-generated when a cube is created.

****The disadvantage of using MOLAP is:****

* Only a limited amount of data can be stored. Since the calculations are triggered at the cube generation process it cannot withstand a huge amount of data.
* Needs a lot of skill to utilize this.
* Also, it has a licensing cost associated with it.

### **3. Explain in detail what is ROLAP? What are the advantages and disadvantages?**

As the name suggests that, the data is stored in the form of a relational database.

****The advantages of using ROLAP is:****

1. As the data is stored in relational databases, it can handle a huge amount of data storage.  
2. All the functionalities are available as this is a relational database.

****The disadvantages of using ROLAP is:****

1. It is comparatively slow.   
2. All the limitations that apply to SQL, the same applies to ROLAP too

### **4. Explain in detail what is HOLAP? What are the advantages of using this type of data storage?**

* HOLAP stands for Hybrid online analytical processing.
* Actually, it is a combination of MOLAP and  ROLAP.

****The advantages of using MOLAP is:****

1. In this model, the cube is used to get summarized information.
2. For drill-down capabilities, it uses the ROLAP structure.

### **5. Explain the main difference between Data Mining and Data Warehousing?**

#### **Data Warehousing:**

It is a process where the data is extracted from various sources. Further, the data is cleansed and stored.

|  |
| --- |
| ****Related Article:**[Data Warehousing Introduction](https://mindmajix.com/data-warehousing/what-is-data-warehousing" \o "What is Data Warehousing?" \t "https://mindmajix.com/_blank)** |

#### **Data Mining:**

* It is a process where it explores the data using the queries.
* Basically, the queries are used to explore a particular data set and examine the results. This will help the individual in reporting, strategy planning, visualizing meaningful data sets.

The above can be explained by taking a simple example:

Let’s take a software company where all of their project information is stored. This is nothing but Data Warehousing.

Accessing a particular project and identifying the Profit and Loss statement for that project can be considered as Data Mining.

|  |
| --- |
| ****Related Article:**[Define Data Mining](https://mindmajix.com/what-is-data-mining" \o "What is Data Mining?" \t "https://mindmajix.com/_blank)** |

### **6. Explain in detail what is Data Purging?**

Data purging is an important step in maintaining appropriate data in the database.

Basically deleting unnecessary data or rows which have NULL values from the database is nothing but data purging. So if there is a need to load fresh data into the database table we need to utilized database purging activity. This will clear all unnecessary data in the database and helps in maintaining clean and meaningful data.

Data purging is a process where junk data that exists in the database gets cleared out.

[](https://bit.ly/3if9dmk)

### **7. Explain in detail what does CUBE means?**

Cube is nothing but a data storage place where the data can be stored and makes it easier for the user to deal with his/her reporting tasks. It helps expedite the data analysis process.

****For example:****

Let’s say the data related to an employee is stored in the form of a cube. If you are evaluating the user performance based on a weekly, monthly basis then week and month are considered to be the dimensions of the cube.

### **8. What are the different problems that “Data Mining” can solve in general?**

Data Mining is a very important process where it could be used to validate and screen the data how it is coming through and the process can be defined based on the data mining results. By doing these activities, the existing process can be modified.

They are widely used in the following industries :

1. Marketing
2. Advertising
3. Services
4. Artificial Intelligence
5. Government intelligence

By following the standard principles a lot of illegal activities can be identified and dealt with. As the internet has evolved a lot of loops holes also evolved at the same time.

### **9. Explain the difference between OLAP and OLTP?**

#### ****OLTP:****

1. OLTP stands for Online Transaction and Processing.
2. This is useful in applications that involve a lot of transactions and high volumes of data. This type of application is mainly observed in the Banking sectors, Air ticketing, etc. The architecture used in OLTP is Client-server architecture. It actually supports the transactions cross-network as well.

#### ****OLAP:****

1. OLAP stands for Online Analytical Processing.
2. It is widely used in applications where we need to support business data where complex calculations happen. Most of the time, the data is in low volumes. As this is being multidimensional database, the user will have an insight into how the data is coming through the various sources.

|  |
| --- |
| ****Related Article:**[Difference between OLAP and OLTP](https://mindmajix.com/oltp-vs-olap" \o "OLTP VS OLAP - Difference between OLTP and OLAP" \t "https://mindmajix.com/_blank)** |

### **10. Explain the different stages of “Data Mining”?**

They are three different stages in Data Mining, they are as follows:

1. Exploration
2. Model building and validation
3. Deployment

#### ****Exploration:****

Exploration is a stage where a lot of activities revolve around the preparation and collection of different data sets. So activities like cleaning, transformation are also included. Based on the data sets available, different tools are necessary to analyze the data.

#### ****Model Building and validation:****

In this stage, the data sets are validated by applying different models where the data sets are compared for best performance. This particular step is called pattern identification. This is a tedious process because the user has to identify which pattern is best suitable for easy predictions.

#### ****Deployment:****

Based on the previous step, the best pattern is applied for the data sets and it is used to generate predictions and helps in estimating expected outcomes.

### **11. Explain what is Discrete and continuous data concepts in the Data Mining world?**

Discrete data can be classified as a defined data or finite data. That has meaning to itself. For example Mobile numbers, gender.

Continuous data is nothing but data that continuously changes in an orderly fashion. The example for continuous data is “Age”.

### **12. Explain what is MODEL in terms of Data Mining subject?**

Model is an important factor in Data Mining activities, it defines and helps the algorithms in terms of making decisions and pattern matching. The second step is that they evaluate different models that are available and select the best suitable model for validating the data sets.

### **13. Explain what is Naive Bayes Algorithm?**

****Ans:****The Naive Bayes Algorithm is widely used to generate mining models. These models are generally used to identify the relationship between the input columns and the predicated columns that are available. This algorithm is widely used during the initial stages of the explorations.

### **14. Explain in detail the Clustering Algorithm?**

* The clustering algorithm is actually used on groups of data sets that are available with common characteristics, they are called clusters.
* As the clusters are formed, it helps to make faster decisions, and exporting the data is also fast.
* First of all the algorithm identifies the relationships that are available in the dataset and based on that it generates clusters. The process of creating clusters is also repetitive.

### **15. Explain what is time series algorithm in data mining?**

* This algorithm is a perfect fit for the type of data where the values change continuously based on time. For example Age
* If the algorithm is skilled and tuned to predict the data set, then it will be successfully keep a track of the continuous data and predict the right data.
* This algorithm generates a specific model which is capable of predicting the future trends of the data based on the real original data sets.
* In between the process, new data can also be added in part of trend analysis.

### **16. Explain in detail about association algorithm in Data mining?**

This algorithm is mainly used in recommendation engines for a specific market-based analysis.   
So the input for this algorithm would be the products or items that are bought by a specific customer, based on that purchase a recommendation engine will predict the best suitable products for the customers.

### **17. What is a sequence clustering algorithm?**

As the name itself states that the data is collected at different points which occurs at the sequence of events. The different data sets are analyzed based on the sequence of data sets that occur based on the events. The data sets are analyzed and then the best possible data input will be determined for clustering.

****Example****:

A sequence clustering algorithm will help the organization to specify a particular path to introduce a new product that has similar characteristics in a retail warehouse.

### **18. What are the different concepts and capabilities of Data Mining?**

So Data Mining is primarily responsible to understand and get meaningful data from the data sets that are stored in the database.   
In terms of exploring the data in data mining is definitely helpful because it can be used in the following areas:

****Reporting****

1. Planning
2. Strategies
3. Meaningful Patterns etc.

A large amount of data is cleaned as per the requirement and can be transformed into meaningful data which can be helpful for decision making at the executive level.

****Data mining is really helpful with the following types of data:****

1. Data sets which are in the form of sales figures
2. Forecast values for the business projection
3. Cost
4. Metadata etc

Based on the data analyzed, the information can be analyzed and appropriate relationships are defined.

### **19. What is the best way to work with data mining algorithms that are included in SQL Server data mining?**

With the use of SQL Server, data mining offers an add-on for MS office 2007. This will help to identify and discover the relationships with the data. This data is helpful in the future for enhanced analysis.

The add-on is called “ Data Mining client for excel”. With this the users will be able to first prepare data, build and further manage and evaluate the data where the final output will predicting results.

### **20. How to use DMX- the data mining query language in detail?**

DMX consists of two types of statements in general.

****Data Definition:****

This is used to define and create new models and structures.

****Data Manipulation:****

As the name itself depicts, the data is manipulated based on the requirement.

The usage is explained in detail by picking up an example:

* Create Mining Structure
* Create Mining Model
* Data Manipulation that is used in existing structures and models.

With the syntax, it is

INSERT INTO  
SELECT FROM. CONTENT (DMX)

### **21. What are the different functions of data mining?**

The different functions of data mining are as follows:

1. Characterization
2. Association and correlation analysis
3. Classification
4. Prediction
5. Cluster analysis
6. Evolution analysis
7. Sequence analysis

### **22. Explain in detail what is data aggregation and Generalization?**

#### **Data Aggregation:**

As the name itself is self-explanatory, the data is aggregated altogether where a cube can be constructed for data analysis purposes.

#### **Generalization:**

It is a process where low-level data is replaced by high-level concepts so the data can be generalized and meaningful.

### **23. Explain in detail about In Learning and In classification:**

#### **In Learning:**

This is a model which is primarily used to analyze a particular training data set and it has training data samples that are selected from a selected population.

#### **In Classification:**

This model is primarily used for providing estimation for a particular class by selecting test samples randomly.  The term classification is usually determined by identifying a known class for specific unknown data.

### **24. Explain in detail what is Cluster Analysis?**

The term cluster analysis is an important human activity that is widely used in different applications. To be specific, this type of analysis is used in market research, pattern recognition, data analysis, and image processing.

### **25. Explain about data mining interface?**

* The data mining interface is usually used for improving the quality of the queries that are used.
* The data mining Interface is nothing but the GUI form for data mining activities.

### **26. Why Tuning data warehouse is needed, explain in detail?**

The main aspect of a data warehouse is that the data evolves based on the time frame and it is difficult to predict the behavior because of its ad hoc environment. The database tuning is much difficult in an OLTP environment because of its ad hoc and real-time transaction loads. Due to its nature, the need for data warehouse tuning is necessary and it will change the way how the data is utilized based on the need.

## ****Top 25 Data Mining Interview Questions****

These data mining and data warehousing interview questions and answers could most probably be asked in your interview.

****1.    What is the Scope of Data Mining?****

It helps automate the process of analyzing and identifying predictive information in a huge amount of databases and datasets. Data Mining tools can help scrape and sweep through a diverse range of data in order to identify a pattern that was previously hidden.

****2.    What are the different stages of Data Mining?****

The three main stages are:

a.    Exploration

b.    Model Building and Validation

c.    Deployment

****3.    Define the Exploration Stage in Data Mining?****

The Exploration stage is mainly focused on collecting data from various sources and preparing it for later transformation and cleaning activities.

****4.    Define metadata?****

Metadata can simply be defined as data about data. Metadata is the summarized data that takes us to the detailed data.

****5.    Name a few Data Mining Techniques.****

* Classification Analysis
* Association Rule Learning
* Anomaly or Outlier Detection
* Clustering Analysis
* Regression Analysis
* Prediction
* Sequential Patterns
* Decision Trees

****6.    What are the different types of Data Mining?****

You can classify Data Mining into the following types:

* Integration
* Selection
* Data cleaning
* Pattern evaluation
* Knowledge representation
* Data transformation

****7.    Could you give a small introduction to Data Mining processes?****

It is a process of discovering hidden information that is valuable by analyzing a huge collection of data. It is very advantageous to many industries.

****8.    Why are the Model Building and Validation stage important in Data Mining?****

It is important since, in this stage, data is validated by using different models and is compared to finalize the model with the best performance.

****9.    In Data Mining, what are “Continuous” and “Discrete” data?****

“Continuous data” is the data that changes continuously in a well-structured manner. The perfect example of this is age. “Discrete data” is when data is finite and has a specific meaning present in it. The most suitable example of this is gender.

****10.    Could you list a few areas in which data mining can be applied?****

Data Mining would most likely be used in the fields of:

* Healthcare
* Energy
* Telecommunication
* Retail
* E-commerce

****11.    What is OLAP?****

[OLAP](https://www.jigsawacademy.com/blogs/business-analytics/olap/), Online Analytical Processing, is a technology that involves complex analytical calculations and is used in various Business Intelligence applications. The main purpose of it is to minimize the query response time while enhancing the performance of reporting.

****12.    What is ETL?****

Extract, Transform, and Load (ETL) is a software that reads the data from a certain data source and extracts the required subset of data.

****13.    In data mining, what are the required technological drivers?****

Query Complexity: In order to analyze a large number of complex queries, we require a very powerful system.

Database size: In order to process and maintain a huge collection of data, we require powerful systems.

****14.    What does ODS stand for?****

ODS stands for Operational Data Store.

****15.    What is the Syntax for Interestingness Measures Specification?****

Interestingness thresholds and measures are user-specified with the statement – with <interest\_measure\_name> threshold = threshold\_value

****16.    What is the Dimension Table?****

A table that contains all the attributes of measurements of data that are stored in the fact table is called a Dimension Table.

****17.    What does STING stand for?****

STING stands for Statistical Information Grid. It is a multi-resolution, grid-based clustering method in which objects are stored in rectangular cells.

****18.    What are a few data mining basic issues?****

A few issues of data mining are:

* Uncertainty handling
* Dealing with noisy data
* Dealing with missing values
* Data selection

****19.    What are the few methods of clustering?****

* Portioning method
* Hierarchical method
* Density-based method
* Grid-based method
* Model-based method
* Constraint-based method

****20.    What do OLTP and OLAP stand for?****

OLTP- Online Transactional Processing

OLAP- Online Analytical Processing

****21.    Define Data Warehousing in one sentence?****

It is the repository of integrated information that is readily available for analysis and queries.

****22.    What does ETL stand for?****

ETL stands for Extract, Transform, and Load.

****23.    What is SCD?****

SCD is Slowly Changing Dimensions and is applied to areas where the data record changes over time.

****24.    What are the best ETL tools you can use?****

* Oracle
* Data Stage
* Ab Initio
* Informatica
* Data Junction
* Warehouse Builder

****25.    What are the various levels of Data Mining analysis?****

The various levels are:

* The various levels are:
* Rule induction
* Data visualization
* Genetic algorithms
* Artificial neural network
* Nearest neighbour method

## ****Conclusion****

Going through these data warehousing and data mining important questions while looking at various other data mining objective questions and answers can help you easily get through the interview for a data scientist job.

1. **Name the different Data Mining techniques and explain the scope of Data Mining.**

The different Data Mining techniques are:

* ****Prediction**** – It discovers the relationship between independent and dependent instances. For instance, when considering sales data, if you wish to predict the future profit, the sale acts as an independent instance, whereas the profit is the dependent instance. Accordingly, based on the historical data of sales and profit, the associated profit is predicted value.
* ****Decision trees**** – The root of a decision tree functions as a condition/question having multiple answers. Each answer leads to specific data that helps in determining the final decision based on the data.
* ****Sequential patterns**** – It refers to the pattern analysis used for discovering identical patterns in transaction data or regular events. For example, historical data of customers helps a brand to identify the patterns in the transactions that happened in the past year.
* ****Clustering analysis**** – In this technique, automatically a cluster of objects having similar characteristics is formed. Clustering method defines classes and then places suitable objects in each class.
* ****Classification analysis**** – In this ML-based method, 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 method creates a pattern based on the relationship of the items in a single transaction.

****The scope of Data Mining is to:****

* ****Predict trends and behaviours**** – Data Mining automates the process of identifying predictive information in large datasets/databases.
* ****Discover previously unknown patterns**** – Data Mining tools sweep and scrape through a broad and diverse range of databases to identify the previously hidden trends. This is nothing but a pattern discovery process.

*****upGrad’s Exclusive Data Science Webinar for you –*****

1. **What are the types of Data Mining?**

Data Mining can be classified into the following types:

* Integration
* Selection
* Data cleaning
* Pattern evaluation
* Data transformation
* Knowledge representation

1. **What is Data Purging?**

Data Purging is a crucial procedure in database management systems. It helps to maintain relevant data in a database. It refers to the process of cleaning junk data by eliminating or deleting the unnecessary NULL values of row and columns. Whenever you need to load new data in the database, first, it is essential to purge the irrelevant data.

**Our learners also read**: [Free online python course for beginners](https://www.upgrad.com/blog/python-free-online-course/)!

With frequent Data Purging of the database, you can get rid of the junk data that takes up a substantial amount of database memory, thereby slowing down the performance of the database.

1. **What is the fundamental difference between Data Warehousing and Data Mining?**

Data Warehousing is the technique used for extracting data from disparate sources. It is then cleaned and stored for future use. On the other hand, Data Mining is the process of exploring the extracted data using queries and then analyze the results or outcomes. It is essential in reporting, strategy planning, and visualizing the valuable insights within the data.

1. **Explain the different stages of Data Mining.**

There are three main stages of Data Mining:

Exploration – This stage is primarily focused on collecting data from multiple sources and preparing it for further activities like cleaning and transformation. Once the data is cleaned and transformed, it can be analyzed for insights.

Model Building and validation – This stage involves validating the data by applying different models to it and comparing the results for best performance. This step is also called as pattern identification. It is a time-consuming process since the user has to manually identify which pattern is the best suited for easy predictions.

Deployment – Once the bests-suited pattern for prediction is identified, it is applied to the dataset for obtaining estimated predictions or outcomes.

1. **What is the use of Data Mining queries?**

Data Mining queries help facilitate the application of the model to the new data, either to make single or multiple results. Queries can retrieve cases that fit a particular pattern more effectively. They extract the statistical memory of the training data and help in obtaining the exact pattern along with the rule of the typical case that represents a pattern in the model. Furthermore, queries can extract regression formulas and other calculations to explain patterns. They can also retrieve the details about the individual cases used in a model.

1. **What are “Discrete” and “Continuous” data in Data Mining?**

In Data Mining, discrete data is the data that is finite and has a meaning attached to it. Gender is a classic example of discrete data. Continuous data, on the other hand, is the data that continues to change in a well-structured manner. Age is a perfect example of continuous data.

1. **What is OLAP? How is it different from OLTP?**

OLAP (Online Analytical Processing) is a technology used in many Business Intelligence applications that involve complex analytical calculations. Apart from complex computations, OLAP is used for trends analysis and advanced data modelling. The primary purpose of using OLAP systems is to minimize the query response time while simultaneously boosting the effectiveness of reporting. The OLAP database stores aggregated historical data in a multidimensional schema. Being a multidimensional database, OLAP allows a user to understand how the data is coming through different sources.

OLTP stands for Online Transaction and Processing. It is inherently different from OLAP since it is used in applications that involve bulk transactions and large volumes of data. These applications are primarily found in the BFSI sector. OLTP architecture is a client-server architecture that can support cross-network transactions.

1. **Name the different storage models that are available in OLAP?**

The different storage models available in OLAP are:

* MOLAP (Multidimensional Online Analytical Processing) – This is a type of data storage where the data is stored in multidimensional cubes instead of standard relational databases. It is this feature which makes the query performance excellent.
* ROLAP (Relational Online Analytical Processing) – In this data storage, the data is stored in relational databases, and hence, it is capable of handling a vast volume of data.
* HOLAP (Hybrid Online Analytical Processing) – This is a combination of MOLAP and ROLAP. HOLAP uses the MOLAP model to extract summarized information from the cube, whereas for drill-down capabilities, it uses the ROLAP model.

1. **What is “Cube?”**

In Data Mining, the term “cube” refers to a data storage space where data is stored. Storing data in a cube helps expedite the process of data analysis. Essentially, cubes are the logical representation of multidimensional data. While the edge of the cube has the dimension members, the body of the cube contains the data values.

Let’s assume that a company stores its employee data (records) in a cube. When it wishes to evaluate the employee performance based on a weekly or monthly basis, then the week/month becomes the dimensions of the cube.

1. **What is Data Aggregation and Generalization?**

Data Aggregation is the process wherein the data is combined or aggregated together to create a cube for data analysis. Generalization is the process of replacing the low-level data with high-level concepts so that the data can be generalized and produce meaningful insights.

1. **Explain the Decision Tree and Time Series algorithms.**

In the Decision Tree algorithm, each node is either a leaf node or a decision node. Every time you input an object in the algorithm, it produces a decision. A Decision Tree is created using the regularities of the data. All the paths connecting the root node to the leaf node are reached either by using ‘AND’ or ‘OR’ or ‘BOTH.’ It is important to note that the Decision Tree remains unaffected by Automatic Data Preparation.

The Time-Series algorithm is used for data types whose values keep changing continually based on time (for example, a person’s age). When you trained the algorithm and tune it to predict the dataset, it can successfully keep track of the continuous data and make accurate predictions. The Time-Series algorithm creates a specific model that can predict the future trends of the data based on the original dataset.

1. **What is clustering?**

In Data Mining, clustering is the process used to group abstract objects into classes containing similar objects. Here, a cluster of data objects is treated as one group. Thus, during the analysis process, data partition happens in groups which are then labelled based on identical data. Cluster analysis is pivotal to Data Mining because it is highly scalable and dimensional, and it can also deal with different attributes, interpretability, and messy data.

Data clustering is used in several applications, including image processing, pattern recognition, fraud detection, and market research.

1. **What are the common issues faced during Data Mining?**

During the Data Mining process, you can encounter the following issues:

* Uncertainty handling
* Dealing with missing values
* Dealing with noisy data
* Efficiency of algorithms
* Incorporating domain knowledge
* Size and complexity of data
* Data selection
* Inconsistency between the data and discovered knowledge.

1. **Presentation and Visualization Specification, and Task-Relevant Data Specification.**

The syntax for Interestingness Measures Specification is:

*with <interest\_measure\_name> threshold = threshold\_value*

The syntax for Pattern Presentation and Visualization Specification is:

*display as <result\_form>*

The syntax for Task-Relevant Data Specification is:

*use database database\_name*

*or*

*use data warehouse data\_warehouse\_name*

*in relevance to att\_or\_dim\_list*

*from relation(s)/cube(s) [where condition] order by order\_list*

*group by grouping\_list*

1. **Name the different level of analysis in Data Mining?**

The various levels of analysis in Data Mining are:

* Rule induction
* Data visualization
* Genetic algorithms
* Artificial neural network
* Nearest neighbour method

1. **What is STING?**

STING stands for Statistical Information Grid. It is a grid-based, multi-resolution clustering method in which all the objects are contained into rectangular cells. While the cells are kept in various levels of resolutions, these levels are further arranged in a hierarchical structure.

1. **What is ETL? Name some of the best ETL tools.**

ETL stands for Extract, Transform and Load. It is a software that can read the data from the specified data source and extract a desired subset of data. After this, it transforms the data using rules and lookup tables and converts it to the desired form. Finally, it uses the load function to load the resulting data into the target database.

The best ETL tools are:

* Oracle
* Ab Initio
* Data Stage
* Informatica
* Data Junction
* Warehouse Builder

1. **What is Metadata?**

In simple words, metadata is the summarized data that leads to the larger dataset. Metadata contains important information like the number of columns used, the order of the fields, the data types of the fields, fix width and limited width, and so on.

1. **What are the advantages of Data Mining?**

Data Mining has four core advantages:

* It helps make sense of raw data and explore, identify, and understand the patterns hidden within the data.
* It helps automates the process of finding predictive information in large databases, thereby helping to promptly identify the previously hidden patterns.
* It helps to screen and validate the data and understand where it is coming from.
* It promotes faster and better decision making, thereby helping businesses to take necessary actions to increase revenue and lower operational costs.

These are the reasons why Data Mining has become an integral part of numerous industries, including marketing, advertising, IT/ITES, business intelligence, and even government intelligence.

We hope these Data Mining interview questions and their answers help you to break the ice with Data Mining. Although these are just a few basic level questions you must know, they will help you to get in the flow and dig deeper into the subject matter.

If you are curious to learn about data science, check out IIIT-B & upGrad’s [Executive PG Program in Data Science](https://www.upgrad.com/data-science-pgd-iiitb/?utm_source=BLOG&utm_medium=BODY&utm_campaign=DV_DA_PGD_BLOG_BODY_74734) which is created for working professionals and offers 10+ case studies & projects, practical hands-on workshops, mentorship with industry experts, 1-on-1 with industry mentors, 400+ hours of learning and job assistance with top firms.

## **What are the drawbacks of using a decision tree algorithm?**

Even a minor change in the data can cause a significant change in the structure of the decision tree, resulting in instability. When compared to other algorithms, the calculation of a decision tree might be rather complex at times. Decision tree training is relatively expensive due to the complexity and time required. The Decision Tree technique fails when it comes to applying regression and predicting continuous values.

## **What is the difference between data mining clustering and classification?**

Clustering is a technique of unsupervised learning, whereas classification is a way of supervised learning. Clustering is the process of grouping data points into clusters based on their commonalities. Classification entails labelling the input data with one of the output variable's class labels. Clustering splits the dataset into subgroups, allowing examples with similar functionality to be grouped together. It doesn't rely on labelled data or a training set to work. Classification, on the other hand, classifies new data based on observations from the training set.

## **Are there any disadvantages of data mining?**

Many privacy problems arise when data mining is used. Despite the fact that data mining has opened the path for simple data collection in its own way. When it comes to precision, it still has certain limits. The data obtained might be incorrect, producing issues with decision-making. The data collecting procedure for data mining uses a lot of technology. Every piece of data created requires its own storage and upkeep. The cost of implementation might skyrocket as a result of this.

窗体顶端

**Q.1Explain the purpose of cluster analysis in Data Warehousing.**

Some of defined purpose of cluster analysis are -  
1. Scalability  
2. Ability to deal with different kinds of attributes  
3. Discovery of clusters with attribute shape  
4. High dimensionality  
5. Ability to deal with noisy  
6. Interpretability

窗体底端

窗体顶端

**Q.2What do you understand by Data Mining?**

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

窗体底端

窗体顶端

**Q.3What do you understand by OLAP?**

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

窗体底端

窗体顶端

**Q.4How would you differentiate between View and Materialized View?**

View is defined as nothing but a virtual table which takes the output of the query and it can be used in place of tables On the other hand materialized view is nothing but an indirect access to the table data by storing the results of a query in a separate schema.

窗体底端

窗体顶端

**Q.5How would you define real-time datawarehousing?**

Real-time datawarehousing is primarily used to capture the business data whenever it occurs. Therefore whenever there is business activity which gets completed, then that data will be available in the flow and become available for use instantly.

窗体底端

窗体顶端

**Q.6State some of the advantages of data mining.**

Some of the advantages of Data mining are -  
1. We use data mining in banks and financial institutions to find probable defaulters. This is done based on past transactions, user behaviour and data patterns.  
2. It assists advertisers to push the right advertisements to the internet. Also, it surfer on web pages based on machine learning algorithms. Therefore data mining benefit both possible buyers as well as sellers of the various products.  
3. It is to arrange and keep most sellable items in the most attentive positions.

窗体底端

窗体顶端

**Q.7What according to you are the disadvantages of data mining?**

The disadvantages of data mining are -  
1. They do not have security systems in place to protect us.  
2. Data mining analytics use software therefore it is difficult to operate.  
3. It require a user to have knowledge based training such that the techniques of data mining are not 100% accurate.

窗体底端

窗体顶端

**Q.8State the required technological drivers in data mining.**

The required technological drivers in data mining include -  
1. Database size: Since for maintaining and processing the huge amount of data, therefore powerful systems are required.  
2. Query Complexity: In order to analyze the complex and large number of queries, therefore powerful system is required.

窗体底端

窗体顶端

**Q.9What is data mining query language?**

Data Mining Query Language (DMQL) was proposed for the DBMiner data mining system such that it was based on the Structured Query Language. These query languages are designed to support ad hoc and interactive data mining. Data mining query language provides commands for specifying primitives. We can use DMQL to work with databases and data warehouses as well. We can also use it to define data mining tasks. In particular we examine how to define data warehouses and data marts in Data mining query language.

窗体底端

窗体顶端

**Q.10What do you understand by Syntax for Specifying the Kind of Knowledge?**

Syntax for Characterization, Discrimination, Association, Classification, and Prediction.

窗体底端

窗体顶端

**Q.11What is the purpose of Data Mining Languages Standardization?**

The purpose of Data Mining Languages Standardization are -  
1. It helps the systematic development of data mining solutions.  
2. It improves interoperability among multiple data mining systems and functions.  
3. It helps in promoting education and rapid learning.  
4. It promotes the use of data mining systems in industry and society.

窗体底端

窗体顶端

**Q.12What do you understand by Active Datawarehousing?**

Active datawarehouse is a datawarehouse which permits decision makers within a company or organization to manage customer relationships effectively and efficiently.

窗体底端

窗体顶端

**Q.13Define 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.

窗体底端

窗体顶端

**Q.14What do you understand by loops in Datawarehousing?**

In datawarehousing, if there is a loop between the tables, then the query generation will take more time and it creates ambiguity. Such that it is advised to avoid loop between the tables.

窗体底端

窗体顶端

**Q.15How do you define Dimensional Modelling?**

Dimensional Modeling is defined as a concept used by dataware house designers to build their own datawarehouse. This model can be stored in two kind of tables – Facts and Dimension table. Such that Fact table has facts and measurements of the business and dimension table contains the context of measurements.

窗体底端

窗体顶端

**Q.16What do you understand by snapshot with reference to data warehouse?**

Snapshot is defined as a complete visualization of data at the time of extraction. This occupies less space and can be used to back up and restore data quickly. In other words, a snapshot is a process of knowing about the activities performed. Also it is stored in a report format from a specific catalog. The report is generated soon after the catalog is disconnected.

窗体底端

窗体顶端

**Q.17Define surrogate key.**

Surrogate key is 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.

窗体底端

窗体顶端

**Q.18What is the goal objective of Optimizer?**

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

窗体底端

窗体顶端

**Q.19What do you understand by Hybrid SCD?**

Hybrid SCDs is a combination of both SCD 1 and SCD 2. It may happen that in a table, some columns are important and we need to track changes for them that is capture the historical data for them whereas in some columns even if the data changes, we do not have to bother. For such tables, we implement Hybrid SCDs, where in some columns are Type 1 and some are Type 2.

窗体底端

窗体顶端

**Q.20Why do we override the execute method as struts Framework ?**

Given Struts Framework, the Action Servlet can be developed, ActionForm servlets (ActionServlet means what class extends the Action class, and ActionForm indicates what class extends the Action Form class) and other servlet classes.  
  
In case of ActionForm class, we can develop validate() method. This method will return the ActionErrors object. In this method we can write the validation code. If this method returns null or ActionErrors with size=0, the web container will call execute() as part of the Action class.  
  
If it returns size > 0, it will not call the execute() method. It will rather execute the jsp, servlet or html file as value for the input attribute as part of the attribute in struts-config.xml file.

窗体底端

窗体顶端

**Q.21What does metadata consists of?**

Metadata consists of : 1. Structure of the data. 2. Algorithm used for summarization. 3. Mapping from the operational environment to the data warehouse.

窗体底端

窗体顶端

**Q.22What is the first step in acquisition of data for the warehouse?**

The first step in acquisition of data for the warehouse is to extract data from multiple, heterogeneous sources.

窗体底端

窗体顶端

**Q.23What does crashing a project imply?**

Project cost and project schedule tradeoffs occur to achieve the maximum schedule compression for the least cost to the project without compromising the intended scope of the project.

窗体底端

窗体顶端

**Q.24What does the unequal partition and join key, results in?**

The unequal partition and join key, results in reduced join performance.

窗体底端

窗体顶端

**Q.25What is learning rate?**

Learning rate is called as the degree of adjustment due to learning trial in neural networks.

窗体底端

窗体顶端

**Q.26What do you understand by Data warehousing?**

A Data warehouse refers to the repository of data used for Management decision support systems. Data warehouse consists of a large variety of data that has a high level of business conditions at a single point in time.

窗体底端

窗体顶端

**Q.27Name the various stages of Data warehousing.**

The stages of Data warehousing are: Offline Operational Database Offline Data Warehouse Real-Time Data warehouse Integrated Data warehouse

窗体底端

窗体顶端

**Q.28Define OLTP and ODS.**

OLTP stands for On-Line Transaction Processing that refers to an application for modifying the data whenever it is received and has a large number of simultaneous users. And, ODS stands for Operational Data Store and it is a repository of real-time operational data rather than long-term trend data.

窗体底端

窗体顶端

**Q.29Define ETL.**

ETL refers to Extract, Transform and Load. This is software that is used for reading the data from the defined data source and extract a desired subset of data. Moreover, it transforms the data using rules and lookup tables and converts it to the desired state.

窗体底端

窗体顶端

**Q.30What do you understand about Junk Dimension?**

This can be considered as a single dimension used for storing the tiny dimensions known as junk attributes. The junk attributes in this phase are a set of text attributes and flags that are transmitted into a different sub-domain known as the junk dimension. It is a dimension table consisting of the properties that do not fit either the truth table or the current dimension tables. Further, these features are basically text or multiple flags like non-generic comments or basic yes/no or true/false markers.

窗体底端

窗体顶端

**Q.31Define Subject-Oriented Data Warehouse.**

This can be defined as the storage of data for a specific field such as product, customer, or sales. The subject-oriented property states that the data in a data warehouse are grouped around major bodies in an organization’s interests. For example, customers, brands, prices, and vendors. Further, this property allows DW users to do in-depth analyses on each topic for operational and strategic decision-making.

窗体底端

窗体顶端

**Q.32What is ETL Testing?**

ETL Testing is carried out for validating data extracted from various sources and destinations. This is used for: Firstly, validating the components of the ETL Data Warehouse. Secondly, putting ETL software through its paces. Thirdly, running the test in the background. Then, identifying and resolving problems. Next, developing the concept and carrying out the test harness and events. Lastly, accepting design criteria and standards.

窗体底端

窗体顶端

**Q.33Define XMLA.**

Using XMLA as XML for testing purposes is an effective method for collecting information from OLAP, data mining, and other online sources. XMLA is a simple object management protocol in which the protocol makes use of two methods: Discover and Execute. The discovery system collects data from the archive and the execution system helps you execute programs against data sources.

窗体底端

窗体顶端

**Q.34Define Snowflake Schema.**

A Data Warehouse snowflake schema is a mathematical structural representation of tables in which the ER diagram resembles a snowflake shape. It is an extension of a Star Schema that adds depth. The dimension tables have been normalized, resulting in the data being divided into additional tables.

窗体底端

窗体顶端

**Q.35Explain the term Cube in Data Warehousing.**

Cubes are used for explaining multidimensional data logically. The dimension members are located on the cube’s edge, and the data values are located on the cube’s body.

窗体底端

窗体顶端

**Q.36What do you know about Business Intelligence?**

Business intelligence can be defined as a decision support system (DSS), and it relates to the technologies, applications, and practices for collecting, integrating, and analyzing business-related knowledge or data. However, data warehousing and business intelligence are concepts used for characterizing the method of storing all of the company’s data in internal or external databases from different sources for analyzing it and creating actionable information through online BI software.

窗体底端

窗体顶端

**Q.37Define Aggregate Tables.**

Aggregate tables contain current warehouse data that have been clustered to a specific degree of dimension. Data is easier to obtain from aggregated tables than from the original table with a larger number of records. Further, aggregate tables roll up data to a degree greater than a base or derived table. This table lowers the burden on the database server and improves query efficiency.

窗体底端

窗体顶端

**Q.38What is an ER Diagram?**

Entity-Relationship Diagram (ER diagram) depicts the interrelationships between the entities in a database. This diagram demonstrates the arrangement of each table as well as the relationships between them. An ER Diagram is a flowchart that displays how entities in a structure interact with one another.

窗体底端

窗体顶端

**Q.39Define the term BUS Schema.**

In a fact table, the BUS schema contains a suite of verified dimensions and uniform definitions. The BUS Schema controls the Dimension Identification for all businesses. And, BUS Schema in ETL has a standardized description of details as well as a conformed dimension.

窗体底端

窗体顶端

**Q.40What is a data lake?**

A data lake refers to a centralized repository for structure and unstructured data storage. It can be used for storing raw data as it is without any structure schema. Moreover, there is no need for performing any ETL or transformation job on it. This can store any type of data like images, text, files, videos, and even it can store machine learning model artifacts, real-time and analytics output, etc.

窗体底端

窗体顶端

**Q.41Define structured and unstructured data.**

Structure data has a known schema and could be fit in a fixed table. It uses the DBMS storage method. Scaling schemas is very difficult. Some of the protocols are ODBS, SQL, and ADO.NET, etc. Whereas, Unstructured data has no schema or structure. It is mostly unmanaged and very easy to scale in runtime and can store any type of data. Some of the protocols are XML,CSV, SMSM, SMTP, JASON etc.

窗体底端

窗体顶端

**Q.42What do you know about data purging?**

Data purging refers to the process involving methods that can erase data permanently from the storage several techniques. And, strategies can be used for data purging the process of data shaping often contrasts with data deletion. However, data purging permanently removes the data to free up more storage and memory space which can be utilized for other purposes. Further, the purging process enables us to archive data even if it is permanently removed from the main source giving us an option to recover that data.

窗体底端

窗体顶端

**Q.43Name some of the data warehouse solutions used in the industry.**

Some of the major solutions are: Snowflakes Oracle Exadata Apache Hadoop SAP BW4HANA Microfocus Vertical Teradata AWS Redshift GCP Big Query

窗体底端

窗体顶端

**Q.44What do you know about slowly changing dimensions?**

A slowly changing dimension (SCD) is one that appropriately manages modifications of dimension members over time. It applies when business entity value changes over time and in an ad-hoc manner.

窗体底端

窗体顶端

**Q.45Name the types of SCD.**

There are six types of Slowly Changing Dimension that are: Type 0: Dimension never changes here, dimension is fixed, and no changes are permissible. Type 1: There’s no record of historical values, only the current state. A kind 1 SCD always reflects the newest values, and when changes in source data are detected, the dimension table is overwritten. Type 2: Row Versioning Track changes as version records which will be identified by current flag & active dates and other metadata. If the source system doesn’t store versions, then it is the info warehouse load process that detects changes and appropriately manages the change during a dimension table. Type 3: Previous Value column Track change to a selected attribute, add a column to point out the previous value, which is updated as further changes occur. Type 4: History Table shows the current value in the dimension table. Hybrid SCD: Hybrid SDC uses techniques from SCD Types 1, 2, and three to trace change.

窗体底端

窗体顶端

**Q.46What is a Factless fact table?**

Factless fact is a fact table with no value. Such a table only contains keys from different dimension tables.

窗体底端

窗体顶端

**Q.47Define the Three-Layer Architecture of ETL.**

1. Staging Layer It stores data collected from various sources. 2. Access Layer In this, end-users focus on data collection in this layer by gaining the information from this layer. 3. Data Integration Layer This imports data from the staging layer into the database. The information may be classified as facts, aggregates, or dimensions. Dimensions and Facts may be combined to form the schema.

窗体底端

窗体顶端

**Q.48Define virtual data warehousing.**

A virtual data warehouse provides a collective view of the finished data. A virtual data warehouse has no historical data. However, it is often considered as a logical data model of the given metadata. This can be considered as the simplest way of translating data and presenting it inside the form which will be employed by decision-makers. Further, it also provides a semantic map that enables users for viewing because the data is virtualized.

窗体底端

窗体顶端

**Q.49Explain the term Star Schema.**

Star schema can be defined as organizing the tables in such a way that results can be retrieved from the database quickly in the data warehouse environment.

窗体底端

窗体顶端

**Q.50Describe the steps to build the data warehouse.**

The steps to be followed for building the data warehouse: Firstly, collecting business requirements. Secondly, identifying the necessary sources. Thirdly, identifying the facts. Then, the specification of the dimensions and the attributes. Next, redefining the dimensions and attributes if required After that, organizing the Attribute hierarchy and defining the relationships. Lastly, allocating unique Identifiers.

窗体底端

窗体顶端

**Q.51Define the term conformed fact.**

A conformed fact is a type of table that will be used over multiple data marts and multiple fact tables.

窗体底端

窗体顶端

**Q.52What is the core dimension?**

The core dimension is a Dimension table which is dedicated for a single fact table or Data Mart.

窗体底端

窗体顶端

**Q.53Define degenerate dimension.**

In a data warehouse, a degenerate dimension refers to a dimension key in the fact table that does not have its own dimension table. Degenerate dimensions usually take place when the fact table’s grain is a single transaction (or transaction line).

窗体底端

窗体顶端

**Q.54What is normalization?**

Normalization can be considered as a multi-step process that puts data into tabular form and eliminates duplicated data from the relation tables. This helps in reducing data redundancy and saves physical database spaces and has minimal write operation cost.

窗体底端

窗体顶端

**Q.55Define Data Pipeline.**

Data Pipeline refers to any set of process elements that move data from one system to a different one. Data Pipeline is often created for an application that uses data to bring value. And, it can also integrate the information across the applications, building the info-driven web products, and completing the data mining activities. Data engineers build the data pipeline.

窗体底端

### **1) What is Data Warehouse?**

Data warehousing (DW) is the repository of a data and it is used for Management decision support system. Data warehouse 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?](https://www.guru99.com/business-intelligence-definition-example.html" \o "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?**

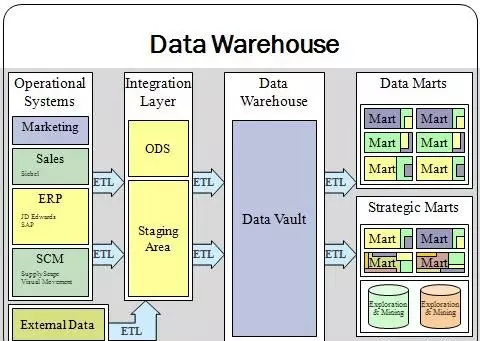
Fact table contains the measurement of business processes, and it contains foreign keys for the dimension tables.

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 Datawarehousing?**

There are four stages of Datawarehousing:

[](https://career.guru99.com/wp-content/uploads/2013/10/Data_warehouse_overview.jpg)Datawarehouse

* Offline Operational Database
* Offline Data Warehouse
* Real Time Datawarehouse
* Integrated Datawarehouse

### **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.

[](https://career.guru99.com/wp-content/uploads/2014/07/data-warehouse-interview-questions.png)Data Warehouse Interview Questions

### **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 datawarehousing?**

Real-time datawarehousing 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 Datawarehousing 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 Datawarehousing?**

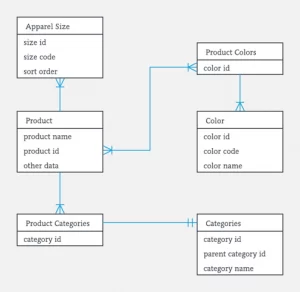
An active datawarehouse is a datawarehouse that enables decision makers within a company or organization to manage customer relationships effectively and efficiently.

### **22) What is the difference between Datawarehouse and OLAP?**

Datawarehouse 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.

[](https://career.guru99.com/wp-content/uploads/2014/07/er-diagram.webp)ER Diagram

### **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

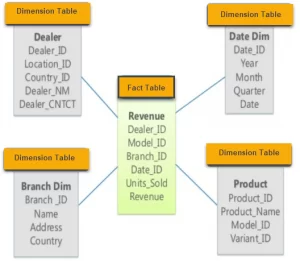
SCD 3 – A original data is modified to include new data

### **27) What is BUS Schema?**

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

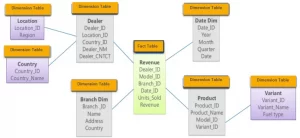
### **28) What is Star Schema?**

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.

[](https://career.guru99.com/wp-content/uploads/2014/07/star-schema.webp)Star Schema

### **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.

[](https://career.guru99.com/wp-content/uploads/2014/07/snowflake-schema.webp)Snowflake Schema

### **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 Datawarehousing?**

In datawarehousing, 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 Datawarehousing?**

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 dataware house designers to build their own datawarehouse. 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?**

Following are the ****Types of Dimensions in Data Warehouse****:

* Conformed Dimension
* Outrigger Dimension
* Shrunken Dimension
* Role-playing Dimension
* Dimension to Dimension Table
* Junk Dimension
* Degenerate Dimension
* Swappable Dimension
* Step Dimension

### **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 datawarehouse?**

Following are the steps to be followed to build the datawaerhouse:

* 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 datawarehosuing?**

Following are the different types of Datawarehousing:

* Enterprise Datawarehousing
* 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?**

Yes, we can take full backup when the database is opened.

### **45) What is defined as Partial Backup?**

A Partial backup in an [operating system](https://career.guru99.com/top-50-operating-system-interview-questions/" \o "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](https://www.guru99.com/sql-server-questions.html" \o "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](https://career.guru99.com/top-50-oracle-interview-questions-and-answers/" \o "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.