# Most Asked Data Mining Interview Questions

## 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.

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

**Healthcare and Personal Grooming**

Data mining has a significant impact in the field of healthcare. It uses data and analytics to identify the best practices that can improve care and reduce costs. Scientists use several Data Mining approaches like multi-dimensional databases, machine learning, soft computing, data

visualization, statistics, etc., to make things easy for patients. Using Data Mining, we can predict the volume of patients in every category and make sure that the patients get the appropriate care at the right place and at the right time.

**Market Basket Analysis**

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

**Education & Training**

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

**Manufacturing Engineering**

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

**Fraud Detection**

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

**Customer Relationship Management**

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

Some other areas where data mining is used:

- Intrusion Detection
- Lie Detection
- Customer Segmentation
- Financial Banking
- Corporate Surveillance
- Research Analysis
- Criminal Investigation
- Bio Informatics

## 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.