ADITYA ENGINEERING COLLEGE
An Autonomous Institution
Approved by AICTE • Permanently Affiliated to JNTUK • Accredited by NAAC with 'A' Grade
Recognised by UGC under sections 2(f) and 12(B) of UGC Act, 1956
Aditya Nagar, ADB Road, Surampalem - 533437, Near Kakinada, E.G.Dt., Ph:99498 76662

Department of Computer Science and Engineering

B.Tech - V Semester (2022-23)

UNIT WISE IMPORTANT QUESTIONS

Course Code : 201CS5T03
Name of the Course: Data Warehousing and Data Mining (AR 20)

UNIT –1

1.Define Data warehouse, Data Mining?

A data warehouse is a database that is designed to store and manage large amounts of data from various sources. It is used to support business intelligence and decision-making activities. A data warehouse typically includes data from multiple sources, such as operational databases, transactional systems, and external sources, and it is designed to be queried and analyzed in support of data-driven decision making.

Data mining is the process of using algorithms and other data analysis tools to discover and extract insights and patterns from large data sets. It is a type of predictive modeling that involves finding patterns and trends in data, and using those patterns to make predictions or decisions. Data mining can be used in a variety of contexts, including fraud detection, marketing, and scientific research.

A data warehouse is a database designed to store and analyze large amounts of data. It typically contains historical data that is extracted from transactional systems and other operational databases, and is used to support business intelligence and decision-making activities.

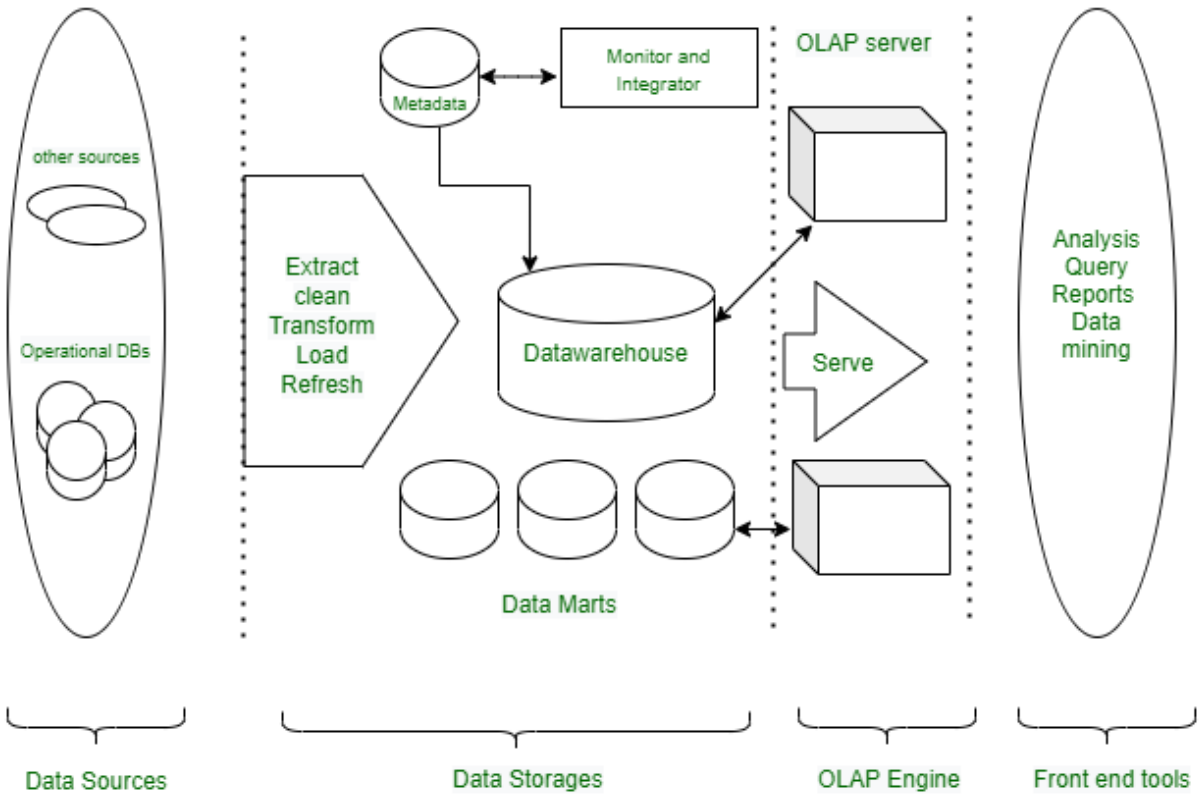
Data mining is the process of using algorithms and other statistical techniques to discover patterns and relationships in large datasets. These patterns and relationships can then be used to make predictions, identify trends, and provide other useful insights. Data mining is often used in conjunction with a data warehouse, as it allows organizations to extract meaningful information from their data and use it to make better decisions.

2.Demonstrate 3 Tier Data ware house architecture with neat sketch?

A data warehouse is Representable by data integration from multiple heterogeneous sources. It was defined by **Bill Inmon** in 1990. The data warehouse is an integrated, subject-oriented, time-variant, and non-volatile collection of data. A Data Warehouse is structured by data integration from multiple heterogeneous sources. It is a system used for data analysis and reporting. A data warehouse is deliberate a core factor of business intelligence. BI technology provides a historical, current, and predictive view of business operations without data mining many businesses may not be able to perform effective market analysis, the strength and weakness of their competitors, profitable decisions, etc.

Data Warehouse is referred to the data repository that is maintained separately from the organization's operational data. **Multi-Tier Data Warehouse Architecture consists of the following components:**

1. Bottom Tier
2. Middle Tier
3. Top Tier



Bottom Tier(Data sources and data storage) :

1. The bottom Tier usually consists of Data Sources and Data Storage.
2. It is a warehouse database server. For Example RDBMS.
3. In Bottom Tier, using the application program interface(called gateways), data is extracted from operational and external sources.
4. Application Program Interface likes ODBC(Open Database Connection), OLE-DB(Open-Linking and Embedding for Database), JDBC(Java Database Connection) is supported.

Middle Tier :

The middle tier is an OLAP server that is typically implemented using either :
A relational OLAP (ROLAP) model (i.e., an extended relational DBMS that maps operations from standard data to standard data); **or** A multidimensional OLAP (MOLAP) model (ie, a special purpose server that directly implements multidimensional data and operations).

Top Tier :

The top tier is a front-end client layer, which includes query and reporting tools, analysis tools, and/or data mining tools (eg, trend analysis, prediction, etc.).

3. Compare OLAP and OLTP?

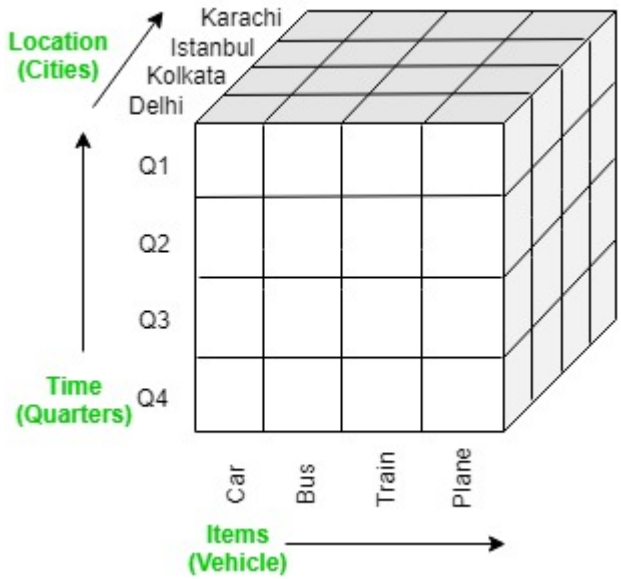
Sr. No.	Category	OLAP (Online analytical processing)	OLTP (Online transaction processing)
1.	Definition	It is well-known as an online database query management system.	It is well-known as an online database modifying system.
2.	Data source	Consists of historical data from various Databases.	Consists of only of operational current data.
3.	Method used	It makes use of a data warehouse.	It makes use of a standard database management system (DBMS).
4.	Application	It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc.	It is application-oriented. Used for business tasks.

5.	Normalized	In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized (3NF).
6.	Usage of data	The data is used in planning, problem-solving, and decision-making.	The data is used to perform day-to-day fundamental operations.
7.	Task	It provides a multi-dimensional view of different business tasks.	It reveals a snapshot of present business tasks.
8.	Purpose	It serves the purpose to extract information for analysis and decision-making.	It serves the purpose to Insert, Update, and Delete information from the database.
9.	Volume of data	A large amount of data is stored typically in TB, PB	The size of the data is relatively small as the historical data is archived. For ex MB, GB
10.	Queries	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.

11.	Update	The OLAP database is not often updated. As a result, data integrity is unaffected.	The data integrity constraint must be maintained in an OLTP database.
12.	Backup and Recovery	It only need backup from time to time as compared to OLTP.	Backup and recovery process is maintained rigorously
13.	Processing time	The processing of complex queries can take a lengthy time.	It is comparatively fast in processing because of simple and straightforward queries.
14.	Types of users	This data is generally managed by CEO, MD, GM.	This data is managed by clerks, managers.
15.	Operations	Only read and rarely write operation.	Both read and write operations.
16.	Updates	With lengthy, scheduled batch operations, data is refreshed on a regular basis.	The user initiates data updates, which are brief and quick.
17.	Nature of audience	Process that is focused on the customer.	Process that is focused on the market.
18.	Database Design	Design with a focus on the subject.	Design that is focused on the application.
19.	Productivity	Improves the efficiency of business analysts	Enhances the user's productivity.

4. Explain OLAP Operations with neat sketch?

OLAP stands for ***Online Analytical Processing*** Server. It is a software technology that allows users to analyze information from multiple database systems at the same time. It is based on multidimensional data model and allows the user to query on multi-dimensional data (eg. Delhi -> 2018 -> Sales data). OLAP databases are divided into one or more cubes and these cubes are known as *Hyper-cubes*.

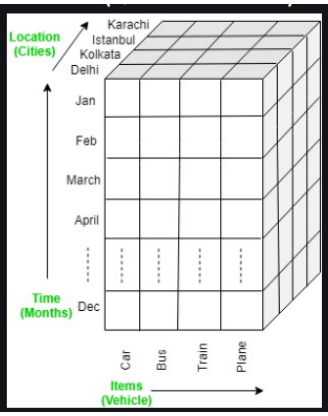


OLAP operations:

There are five basic analytical operations that can be performed on an OLAP cube:

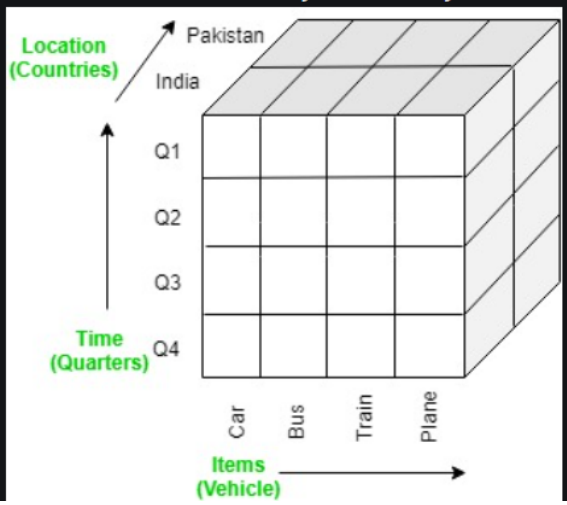
1. **Drill down:** In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:
   - Moving down in the concept hierarchy
   - Adding a new dimension

   In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).

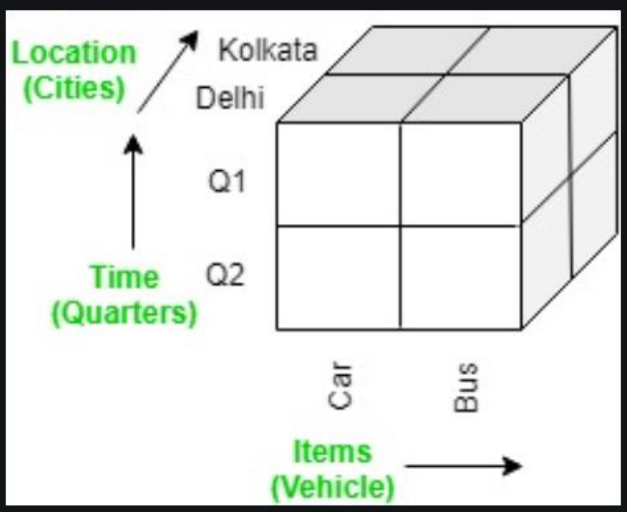


2. **Roll up:** It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:
   - Climbing up in the concept hierarchy
   - Reducing the dimensions

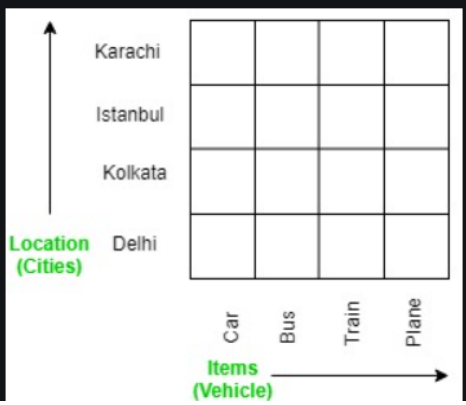
   In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).



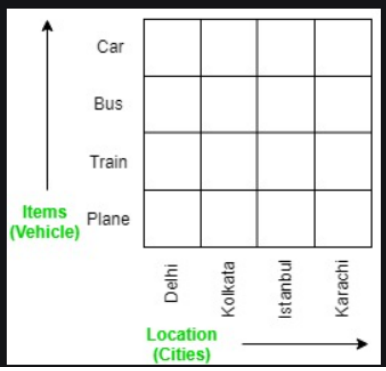
3. **Dice:** It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:
   - Location = "Delhi" or "Kolkata"
   - Time = "Q1" or "Q2"
   - Item = "Car" or "Bus"



4. **Slice:** It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = "Q1".



5. **Pivot:** It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.

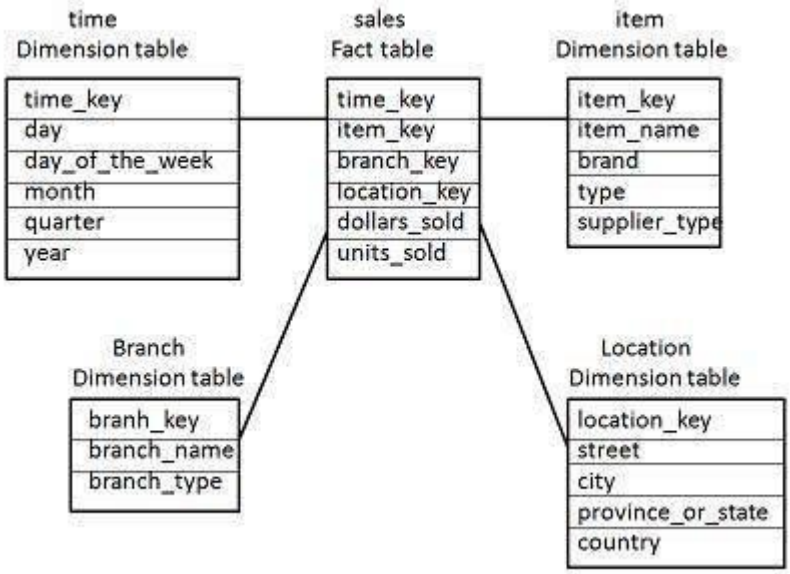


5. Illustrate schemas available in Multidimensional Data model with neat sketch?

Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

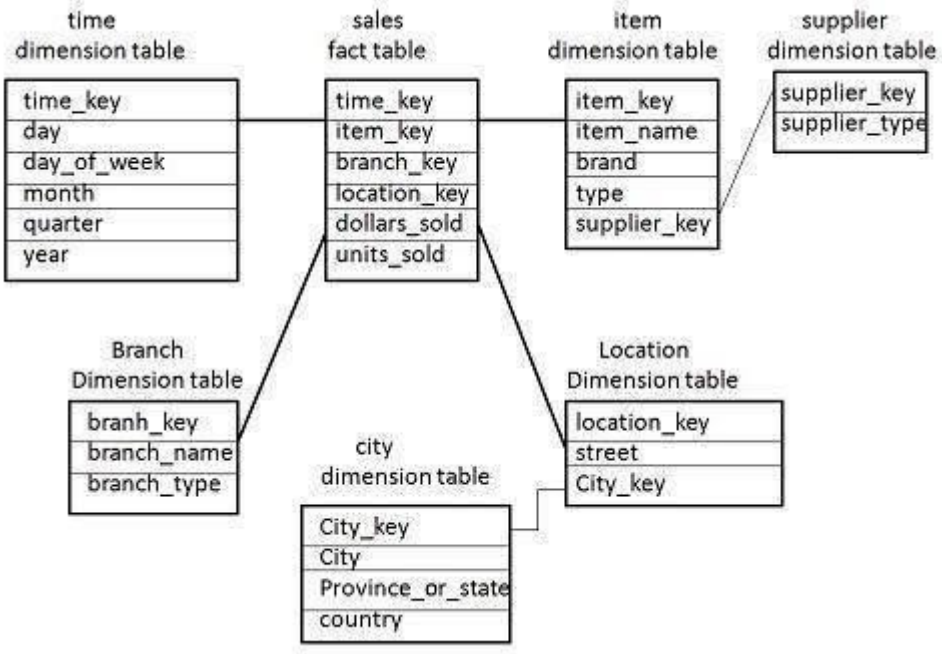


- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Note – Each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

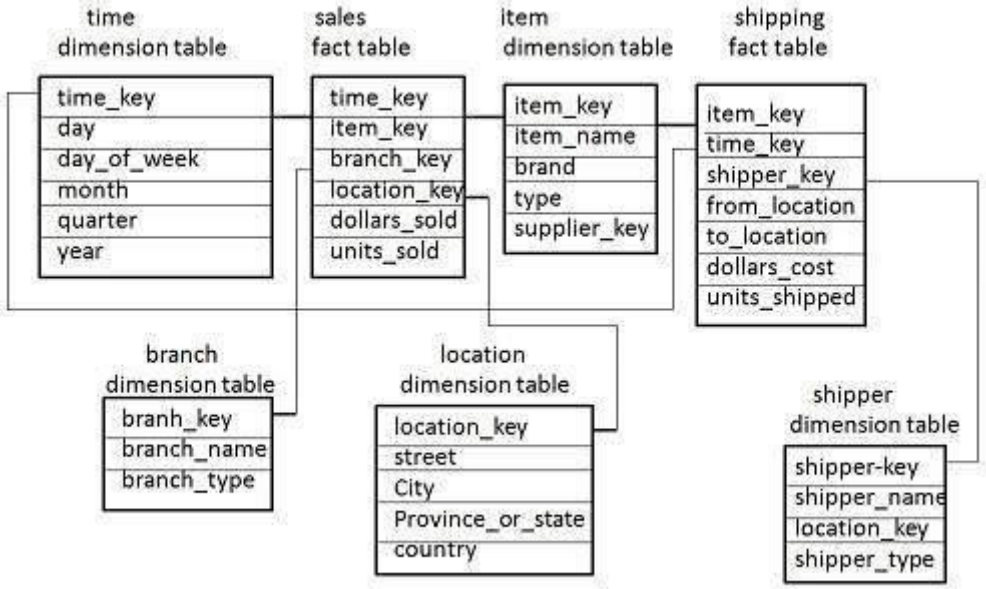


- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

Note – Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

6. Discuss in detail the kind of patterns that can be mined?

There are many different types of patterns that can be mined from data, and the specific patterns that can be found will depend on the data itself and the goals of the analysis. Some common types of patterns that can be mined from data include:

- Clustering patterns: Clustering is a technique used to group data points into clusters based on their similarity. For example, if we have a dataset of customer purchases, we might use clustering to group customers into different segments based on their purchasing behavior.
- Association patterns: Association rules are used to find relationships between variables in a dataset. For example, if we have a dataset of customer transactions, we might use association rules to identify which products are frequently purchased together.
- Sequential patterns: Sequential pattern mining is used to find patterns in sequential data, such as a sequence of events or a sequence of transactions. For example, if we have a dataset of customer transactions, we might use sequential pattern mining to identify common sequences of purchases made by customers.
- Outlier patterns: Outliers are data points that are significantly different from the rest of the data. Outlier detection algorithms are used to identify and analyze these unusual data points. For example, if we have a dataset of customer transactions, we might use outlier detection to identify unusual or suspicious transactions.
- Trend and seasonality patterns: Time series data often contains patterns such as trends and seasonality, which can be important for forecasting and other types of analysis. For example, if we have a dataset of sales data over time, we might use trend and seasonality analysis to identify long-term trends and periodic patterns in the data.

Overall, the specific patterns that can be mined from a dataset will depend on the data itself and the goals of the analysis. The key is to identify the relevant patterns and use them to gain insights and make better decisions.

7. What are the applications of data mining? Explain ?

Data mining is the process of extracting useful information and insights from large datasets. This can be done using a variety of techniques, including machine learning, statistical analysis, and data visualization. Data mining has many applications, including:

- Customer segmentation and targeting: Data mining can be used to group customers into different segments based on their characteristics and behavior. This can help businesses tailor their marketing and sales efforts to specific groups of customers, leading to more effective and efficient marketing campaigns.
- Fraud detection: Data mining can be used to identify unusual patterns in data that may indicate fraudulent activity. For example, a credit card company might use data mining to identify unusual patterns in customer spending habits that could indicate fraud.
- Predictive maintenance: Data mining can be used to identify patterns in data that can help predict when equipment or machinery is likely to fail. This can help businesses schedule maintenance and repairs more efficiently, reducing downtime and improving overall operations.
- Text and sentiment analysis: Data mining can be used to analyze text data, such as customer reviews or social media posts, to identify sentiment and extract insights. For example, a company might use data mining to analyze customer reviews to identify common themes and trends in customer feedback.
- Supply chain optimization: Data mining can be used to identify patterns in data that can help improve supply chain efficiency. For example, a company might use data mining to identify bottlenecks and inefficiencies in its supply chain, and then use this information to make changes that improve overall operations.

Overall, data mining has many applications across a wide range of industries, including retail, finance, healthcare, manufacturing, and more. By extracting valuable insights from large datasets, data mining can help businesses make better decisions and improve their operations.

8. Discuss the issues faced in data mining?

Data mining can be a powerful tool for extracting valuable insights from large datasets. However, like any other technology, data mining also has its challenges and limitations. Some of the issues faced in data mining include:

- Data quality and quantity: The accuracy and usefulness of data mining depends heavily on the quality and quantity of the data. If the data is incomplete, inconsistent, or biased, the results of data mining will be less reliable. Additionally, if the dataset is too small or not representative of the population, the results may not be generalizable.
- Privacy and ethical concerns: Data mining often involves analyzing large amounts of sensitive personal data, which raises concerns about privacy and ethics. It is important to ensure that data mining is done in a way that respects individuals' privacy and complies with relevant laws and regulations.
- Model bias: Data mining algorithms can sometimes produce biased results, either because the algorithm itself is biased or because the dataset used to train the algorithm is biased. For example, if a dataset contains more data from one group than another, the resulting model may be biased towards that group.
- Overfitting: Overfitting is a common problem in data mining, where the model is too complex and fits the training data too closely, but does not generalize well to new data. This can lead to poor performance on unseen data and can reduce the model's usefulness in practice.
- Interpretability: Data mining algorithms can sometimes produce complex and hard-to-interpret models, which can make it difficult to understand why a particular result was produced or how to use the model in practice. This can make it difficult to trust the results of data mining and can limit the usefulness of the insights obtained.

Overall, data mining is a powerful tool for extracting insights from large datasets, but it is not without its challenges and limitations. These issues can be addressed by carefully designing data mining algorithms, using high-quality and diverse datasets, and properly evaluating and interpreting the results of data mining.

UNIT– 2

1. What is an attribute? What are various attributes types? Give examples?
2. Explain various data representation mechanisms in data mining?
3. How could we measure the data similarity and dissimilarity?
4. Explain Data mining as a KDD Process with neat sketch.
5. Illustrate role of Data preprocessing in Data Mining?
6. What is Data cleaning? Explain techniques used in Data Cleaning
7. What is Data Transformation? Explain different Data TransformationTechniques with examples.
8. Explain Data Reduction Techniques?

UNIT – 3

1. Illustrate the process of general classification problem?
2. What are the various attribute selection measures? Explain?
3. Explain the concept of decision tree induction?
4. Discuss the classification algorithm with an example?
5. What is classification in decision tree?
6. Implement classification algorithm on database?
7. What is Decision tree? With an example, briefly describe the algorithm for generating decision tree.
8. Explain various measures used to find the best split node in decision tree

UNIT-4

1. What is frequent Item Set? Give example?
2. What is rule generation? Discuss?
3. Explain the concept of Apriori Algorithm?
4. Implement Apriori on transactional database?
5. What is Support, Minimum support and minimum confidence?
6. With an example explain the concept of confident based pruning?
7. What is an association rule? Give example?

8. What is association rule mining?

9. Explain the process of constructing frequent pattern tree?

10. Explain FP- Growth algorithm with an example?

UNIT-5

1. What is the importance of cluster analysis?
2. Explain different types of clusters with examples?
3. What are various clustering techniques? Explain any one?
4. Discuss basic k-means algorithm in detail with graphical notations?
5. Explain the steps in k-means algorithm?
6. Discuss Bi- secting k-means with an example?
7. Compare and contrast basic K- means and Bi- secting K- means algorithm?