🎯HERE YOU GET ALL (MUMBAI UNIVERSITY) ENGINEERING NOTES 🎯

FIRST YEAR ENGINEERING NOTES (MU)📚

😊 JOIN TELEGRAM 😊

🎗️ KEEP SHARE THIS LINKS 🎗️

👇👇👇👇👇👇👇👇👇👇👇👇
JOIN US TELEGRAM ALL LINKS FOR NOTES

⭐1.CLICK TO JOIN CHANNEL 👉 @engineeringnotes_mu

⭐2.CLICK TO JOIN GROUP 👉 @engineering_notes_mu

⭐3.CLICK TO JOIN 1ST YEAR NOTES BOT👉@engineeringnotes_mubot

4.CLICK TO JOIN 2ND YEAR COMPUTER ENGINEERING NOTES BOT 👉 @computerengineeringmu_notes_bot

5.CLICK TO JOIN CODING CHANNEL 👉 @codingnewbeginners


Engineering notes
MUMBAI UNIVERSITY

Data Warehouse & Mining Viva Questions

Data Warehousing & Mining (University of Mumbai)

1. **What is Data warehousing?**

A **Data Warehousing** (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from various sources.

2. **What is data warehouse?**

A data warehouse is an electronic storage of an organization's historical data for the purpose of reporting, analysis and data mining or knowledge discovery.

3. **What Is Data Purging?**

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.

4. **What Are the Different Problems That "data Mining" Can Solve?**

• 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.

5. **What is Dimension Table?**

A dimension table is a table in star schema and snowflake schema of a data warehouse. A dimension table stores attributes, or dimensions, that describe the objects in a fact table.

6. **What is Fact Table?**

A fact table is the central table in a star schema and snowflake schema of a data warehouse.

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

7. What is data mining?

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis.

8. Difference between OLAP and OLTP

OLAP	OLTP
OLAP is an acronym Online analytical processing	OLTP is an acronym for Online transaction processing
Consists of historical data from various Databases.	Consists only operational current data.
OLAP has long transactions.	OLTP has short transactions.
Based on SELECT commands to aggregate data for reporting	Based on INSERT, UPDATE, DELETE commands
Complex queries.	Simpler queries.

9. 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 transforms 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.

10. What is Datamart

A data mart is a subset of data stored within the overall data warehouse, for the needs of a specific team, section, or department within the business enterprise.

Data marts make it much easier for individual departments to access key data insights more quickly and helps prevent departments within the business organization from interfering with each other's data.

11. 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

12. What is Star Schema?

A star schema is a data warehousing architecture model where one fact table references multiple dimension tables, which, when viewed as a diagram, looks like a star with the fact table in the center and the dimension tables radiating from it. It is the simplest among the data warehousing schemas and is currently in wide use.

13. What is Snowflake Schema

The snowflake schema is an extension of a star schema. The main difference is that in this architecture, each dimension table can be linked to one or more-dimension tables as well. The aim is to normalize the data.

14. 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.

15. What is a Decision Tree Algorithm?

Decision tree is a supervised learning algorithm used for classification. It uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

16. What is Naïve Bayes Algorithm?

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. It is one of the simple and most effective Classification algorithms.

17. Explain clustering algorithm.

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

18. Explain Association algorithm in Data mining?

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

19. Differentiate Star Schema and Snowflake Schema

Star Schema	Snowflake Schema
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
Simple DB Design.	Very Complex DB Design.
High level of Data redundancy	Very low-level of data redundancy
Denormalized Data structure and query also run faster.	Normalized Data Structure.
Single Dimension table contains aggregated data	Data Split into different Dimension Tables.

20. What are the characteristics of data warehouse?

- **Subject Oriented**
  - A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations.
  - These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.
- **Integrated**
  - A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc.
  - This integration enhances the effective analysis of data.
- **Time Variant**
  - The data collected in a data warehouse is identified with a particular time period.
  - The data in a data warehouse provides information from the historical point of view.
- **Non-volatile**
  - Non-volatile means the previous data is not erased when new data is added to it.

21. What are typical data mining techniques?

1. **Classification:**

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

2. **Clustering:**

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. **Regression:**

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. **Association Rules:**

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

Q1. Data warehouse architecture

Answer: A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise.

- Each data warehouse is different, but all are characterized by standard vital components.

Q2. E-R Modeling versus Dimensional Modeling

Answer:

ER Data Modeling	Dimensional Data Modeling
Data Redundancy is not desired	Data Redundancy is desired
ER Modeling is used for OLTP application design.	Dimensional Modeling is used for OLAP Applications design.
The ER modelling is for databases that are OLTP databases that use normalized data using 1st or 2nd or 3rd normal forms	Dimensional Modeling is used in data warehouses that use the 3rd normal form. It contains denormalized data.

Q3. Data warehouse versus Data Marts

Answer: Data Marts:

- Data marts contain repositories of summarized data collected for analysis on a specific section or unit within an organization, for example, the sales department.

Data warehouse:

- A data warehouse is a large centralized repository of data that contains information from many sources within an organization.

Q4. Information Package Diagram

Answer: The Information Package Diagram The first and most generalized level of an information model is its information package diagram.

- This model focuses on the data gathering activities for the users' information packaging requirements.

Q5. Data Warehouse Schemas

Answer: The Data Warehouse Schema is a structure that rationally defines the contents of the Data Warehouse, by facilitating the operations performed on the Data Warehouse and the maintenance activities of the Data Warehouse system, which usually includes the detailed description of the databases, tables, views, indexes.

Q6. Star Schema

Answer: A star schema is a database organizational structure optimized for use in a data warehouse or business intelligence that uses a single large fact table to store transactional or measured data and one or smaller dimensional tables that store attributes about the data.

Q7. Snowflake Schema

Answer: Snowflake Schema in a data warehouse is a logical arrangement of tables in a multidimensional database such that the ER diagram resembles a snowflake shape.

- A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions.
- The dimension tables are normalized which splits data into additional tables.

Q8. Factless Fact Table

Answer: A factless fact table is a fact table that does not have any measures.

- It is essentially an intersection of dimensions (it contains nothing but dimensional keys).
- There are two types of factless tables:
    - For capturing an event,
    - For describing conditions.

Q9. Fact Constellation Schema

Answer: Fact Constellation Schema can implement between aggregate Fact tables or decompose a complex Fact table into independent simplex Fact tables.

- Example:
    - This schema defines two fact tables, sales, and shipping.

Q10. Major steps in an ETL process

Answer: At its most basic, the ETL process encompasses data extraction, transformation, and loading.

- While the abbreviation implies a neat, three-step process – extract, transform, load – this simple definition doesn't capture:
    - The transportation of data.
    - The overlap between each of these stages.

Q11. OLTP versus OLAP

Answer: OLTP:

- Online transaction processing (OLTP) captures, stores, and processes data from transactions in real-time.

OLAP:

- Online analytical processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems.

Q12. OLAP operations are Slice, Dice, Rollup, Drilldown and Pivot.

Answer: There are primarily five types of analytical OLAP operations in data warehouse:

- Three types of widely used OLAP systems are MOLAP, ROLAP, and Hybrid OLAP.
- Desktop OLAP, Web OLAP, and Mobile OLAP are some other types of OLAP

1) Roll-up:

- The roll-up operation (also known as drill-up or aggregation operation) performs aggregation on a data cube, by climbing down concept hierarchies, i.e., dimension reduction.
- Roll-up is like zooming out on the data cubes.
- Roll-up operations are performed on the dimension location.
- It is just the opposite of the drill-down operation

2) Drill-down:

- In drill-down operation, the less detailed data is converted into highly detailed data

3) Slice:

- It selects a single dimension from the OLAP cube which results in a new sub-cube creation.

4) Dice:

- It selects a sub-cube from the OLAP cube by selecting two or more dimensions.

5) Pivot:

- It is also known as rotation operation as it rotates the current view to get a new view of the representation.

Q13. Data Mining Task Primitives.

Answer: Data Mining Task Primitives

- Set of task-relevant data to be mined.
- Kind of knowledge to be mined. Background knowledge to be used in the discovery process.
- Interestingness measures and thresholds for pattern evaluation

Q14. KDD process

Answer: KDD is an iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.

- Preprocessing of databases consists of Data cleaning and Data Integration.

Q15. Issues in Data Mining

Answer: Some of the Data mining challenges are given as under:

- Security and Social Challenges.
- Noisy and Incomplete Data.
- Distributed Data.
- Complex Data.
- Performance.
- Scalability and Efficiency of the Algorithms.
- Improvement of Mining Algorithms.
- Incorporation of Background Knowledge.

Q16. Applications of Data Mining

Answer: Useful applications for data mining

- Future Healthcare. Data mining holds great potential to improve health systems.
- Market Basket Analysis.
- Manufacturing Engineering.
- CRM.
- Fraud Detection.
- Intrusion Detection.
- Customer Segmentation.
- Financial Banking.

Q17. Types of Attributes

Answer: Data Attributes

- Nominal Attribute
- Ordinal Attribute
- Binary Attribute
- Numeric attribute: It is quantitative, such that quantity can be measured and represented in integer or real values, which are of two types.
- Ratio Scaled attribute:

Q18. Statistical Description of Data

Answer: Basic Statistical Descriptions of Data - Mean, Median, Mode & Midrange.

- Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range.

Q19. Data Visualization

Answer: Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals.

- The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

Q20. Data reduction

Answer: Data reduction is a process that reduced the volume of original data and represents it in a much smaller volume.

- Data reduction techniques ensure the integrity of data while reducing the data.
- The time required for data reduction should not overshadow the time saved by the data mining on the reduced data set.

Q21. Decision Tree Induction

Answer: A Decision Tree is a supervised learning method used in data mining for classification and regression methods.

- It is a tree that helps us in decision-making purposes.
- It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed.

Q22. Naïve Bayesian Classification

Answer: The Naive Bayes classification algorithm is a probabilistic classifier.

- It is based on probability models that incorporate strong independence assumptions.
- The independence assumptions often do not have an impact on reality.
- You can derive probability models by using Bayes' theorem (credited to Thomas Bayes).

Q23. Accuracy and Error measures

Answer: The accuracy of measurement or approximation is the degree of closeness to the exact value.

- The error is the difference between the approximation and the exact value.

Q24. Holdout & Random Subsampling

Answer: Let’s understand How Random Sub-sampling works: Random Subsampling performs 'k’s iterations of the entire dataset,i.e. we form 'k’s replica of given data.

- The model is fitted to the training set from each iteration, and an estimate of prediction error is obtained from each test set.

Q25. Cross-Validation

Answer: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

- The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into.

Q26. Bootstrap.

Answer: Bootstrap is a framework to help you design websites faster and easier.

- It includes HTML and CSS based design templates for typography, forms, buttons, tables, navigation, modals, image carousels, etc.
- It also gives you support for JavaScript plugins.

Q27. Types of data in Cluster analysis

Answer: Types Of Data Used In Cluster Analysis - Data Mining

- Types Of Data Structures.
- Nominal or Categorical Variables.
- Ordinal Variables.
- Ratio-Scaled Intervals.
- Variables Of Mixed Type.
- Summary.

Q28. k-Means

Answer: K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups).

- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.
- Data points are clustered based on feature similarity.

Q29. k-Medoids

Answer: k -medoids is a classical partitioning technique of clustering that splits the data set of n objects into k clusters, where the number k of clusters assumed known a priori (which implies that the programmer must specify k before the execution of a k -medoids algorithm).

Q30. Hierarchical Methods of Agglomerative and Divisive

Answer: Agglomerative:

- This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive:

- This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Q31. Market Basket Analysis

Answer: Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.

- It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

Q32. Association Rule

Answer: Association rule mining, at a basic level, involves the use of machine learning models to analyze data for patterns, or co-occurrences, in a database.

- Association rules are created by searching data for frequent if-then patterns and using the criteria support and confidence to identify the most important relationships.

Q33. Frequent Pattern Mining

Answer: Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with a frequency no less than a user-specified threshold.

- For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set, is a frequent itemset.

Q34. Apriori Algorithm

Answer: Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database.

- This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved.
- A minimum support threshold is given in the problem or it is assumed by the user.

Q35. Association Rule Generation

Answer: The goal of association rule generation is to find interesting patterns and trends in transaction databases.

- Association rules are statistical relations between two or more items in the dataset. .
- Forgiving support and confidence levels, there are efficient algorithms to determine all association rules.

Q36. Improving the Efficiency of Apriori

Answer: Inherent defects of the Apriori algorithm, some related improvements are carried out:

1) using a new database mapping way to avoid scanning the database repeatedly;
2) further pruning frequent itemsets and candidate itemsets in order to improve joining efficiency;

Q37. Multilevel Association Rules and Mining

Answer: Association rules created from mining information at different degrees of reflection are called various level or staggered association rules.

- Multilevel association rules can be mined effectively utilizing idea progressions under a help certainty system.

Q38. Multidimensional Association Rules.

Answer: Multidimensional Association Rules :

- Quantitative characteristics are numeric and consolidated order.
- Numeric traits should be discretized.
- The multidimensional affiliation rule comprises more than one measurement.
- Example –buys(X, “IBM Laptop computer”)buys(X, “HP Inkjet Printer”)

Q39. Harvest System

Answer: A rainwater harvesting system comprises components of various stages - transporting rainwater through pipes or drains, filtration, and storage in tanks for reuse or recharge.

- Catchments: The catchment of a water harvesting system is the surface that directly receives the rainfall and provides water to the system.

Q40. Virtual Web View

Answer: The MLDB provides an abstracted and condensed view of a portion of the Web.

- A view of the MLDB, which is called a Virtual Web View (VWV) can be constructed.
- WebML, a web data mining query language is proposed to provide data mining operations on the MLDB.

Q41. Web Usage Mining.

Answer: Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications.

- Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis.