DMW VIVA Q&A :-

1. **Data Warehouse:** A data warehouse is a comprehensive, integrated, and time-variant repository of large volumes of structured and unstructured data, gathered from various sources within an organization. It is designed to support business intelligence and analytical processing by providing a unified view of historical and current data. The primary purpose is to facilitate efficient querying, reporting, and analysis to aid decision-making processes.
2. **ETL Process:** The ETL (Extract, Transform, Load) process is a fundamental component of data integration. It involves extracting raw data from diverse source systems, transforming it to meet the requirements of the target database or data warehouse, and finally loading it into the destination. Extraction involves retrieving data from source systems, transformation encompasses cleaning, aggregating, and structuring the data, while loading involves storing the processed data into the target repository. ETL is essential for ensuring data quality, consistency, and relevance for analytics.
3. **Types of OLAP:** OLAP, or Online Analytical Processing, manifests in two main types: ROLAP (Relational OLAP) and MOLAP (Multidimensional OLAP). ROLAP relies on relational databases, storing multidimensional data in relational tables, while MOLAP uses dedicated multidimensional databases that enable more efficient processing of complex queries. ROLAP is suitable for handling large datasets with complex relationships, whereas MOLAP excels in providing fast query response times for predefined summaries and aggregations.
4. **Data Lake vs. Data Warehouse vs. Data Marts:**
    - A Data Lake is a centralized repository that can store vast amounts of raw, unprocessed data in its native format.
    - A Data Warehouse is a structured, organized database optimized for query and analysis, containing processed and integrated data.
    - Data Marts are smaller, specialized subsets of a data warehouse, tailored to the needs of specific departments or business units.
5. **ROLAP vs. MOLAP:** ROLAP, or Relational OLAP, and MOLAP, or Multidimensional OLAP, represent two distinct approaches to organizing and processing data for analytical purposes. ROLAP relies on relational databases, translating multidimensional data into tables, while MOLAP uses specialized multidimensional databases that directly support the representation of data in a multidimensional cube format. ROLAP offers flexibility in handling complex relationships, while MOLAP provides faster query response times for predefined summaries.
6. **OLTP vs. OLAP:** OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing) serve different purposes in the realm of databases. OLTP focuses on efficiently managing transaction-oriented applications, handling numerous, short-lived transactions involving the insertion, updating, and deletion of records. In contrast, OLAP supports complex, read-intensive

queries that involve aggregations, grouping, and reporting for analytical purposes. OLTP databases are optimized for quick, concurrent transactions, while OLAP databases prioritize efficient querying and reporting.

7. *Example-based question - check q 27:* (Please provide question 27 for a specific answer)
8. **Web Mining and Three Types:** Web mining is a process of extracting useful patterns and information from web data. There are three main types:
   - **Web Content Mining:** Involves analyzing the content of web pages, extracting information such as text, images, and multimedia.
   - **Web Structure Mining:** Focuses on discovering patterns in the link structure of the web, identifying relationships and hierarchies.
   - **Web Usage Mining:** Analyzes user interaction data, such as clicks and navigation patterns, to understand user behavior and preferences.
9. **Web Structure vs. Web Content:**
   - **Web Structure:** This aspect of web mining involves examining the linkages and relationships between web pages. It delves into the organization of information on the web, identifying patterns in the hyperlink structure.
   - **Web Content:** Web content mining, on the other hand, is concerned with the extraction and analysis of information contained within web pages. This includes text, images, multimedia, and other forms of content.
10. **Choosing Seed Points for K-Means:** Selecting seed points for K-Means clustering is a crucial step that can influence the quality of the resulting clusters. Seed points are typically chosen based on either random selection or domain knowledge. Random selection involves picking data points from the dataset as initial centroids, while domain knowledge might guide the choice of points that are likely to represent the characteristics of the clusters well. The goal is to initiate the clustering process with centroids that lead to meaningful and representative cluster formations.
11. **"Append" in ETL Loading:** In the context of ETL (Extract, Transform, Load), the term "append" refers to the process of adding new data to an existing dataset without modifying the current data. This is particularly relevant when dealing with historical data or incremental updates. The append operation ensures that the data warehouse or target database grows over time by incorporating new records while preserving the integrity and historical context of the existing dataset.
12. **DBSCAN vs. Traditional Clustering:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) represents a departure from traditional clustering methods. Unlike traditional methods that often rely on defining clusters based on distance measures, DBSCAN uses density to form clusters. It identifies areas of higher data density and considers points within these areas

as belonging to the same cluster. This allows DBSCAN to discover clusters of arbitrary shapes and handle noise more effectively than traditional methods.

13. **Sequential Pattern Mining:** Sequential pattern mining is a data mining technique focused on identifying patterns that occur sequentially in a dataset. This is particularly useful for analyzing time-ordered data or sequences of events. The goal is to uncover relationships and dependencies between events, enabling the discovery of patterns that occur in a specific order or with certain temporal constraints.
14. **Spatial Data:** Spatial data refers to information that has a spatial or geographical component. It is used to represent and store data associated with physical locations on the Earth's surface. This includes coordinates, latitude, and longitude, allowing for the analysis and visualization of spatial relationships. Spatial data finds applications in various fields, such as geography, urban planning, environmental science, and geographic information systems (GIS).
15. **False Positive vs. False Negative:** In the context of classification models, a false positive occurs when the model incorrectly predicts the positive class, while a false negative occurs when the model incorrectly predicts the negative class. The relative danger of false positives versus false negatives depends on the specific application. In situations where missing a positive instance has more severe consequences, false negatives are considered more dangerous. For instance, in medical diagnoses, a false negative could mean failing to identify a serious condition.
16. **Nature of Outcomes for Classifications and Precision Models:** The outcomes for classification and precision models are discrete in nature. In classification, the model assigns instances to predefined categories or classes. Precision models, often used in scenarios where the focus is on correctly predicting specific outcomes, also produce discrete results. These models aim to provide accurate and precise estimates for given inputs, emphasizing the quality and reliability of predictions.
17. **Correlation Coefficient vs. Chi-Square:**
    - **Correlation Coefficient:** This statistical measure quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation.
    - **Chi-Square:** Chi-Square is a statistical test used to assess the independence of two categorical variables. It compares observed and expected frequencies in a contingency table to determine if there is a significant association between the variables.
18. **Confusion Matrix:** A confusion matrix is a tool used to evaluate the performance of a classification model. It provides a detailed breakdown of predictions by categorizing them into four components:

- **True Positive (TP):** Instances correctly predicted as positive.
- **True Negative (TN):** Instances correctly predicted as negative.
- **False Positive (FP):** Instances incorrectly predicted as positive.
- **False Negative (FN):** Instances incorrectly predicted as negative. The matrix aids in assessing the model's accuracy, precision, recall, and F1 score.

19. **Conditional Independency:** Conditional independence is a concept in probability theory stating that two events are independent given the occurrence of a third event. For example, consider events A and B. If the occurrence of event C makes events A and B independent, then P(A and B | C) = P(A | C) * P(B | C). Conditional independence is crucial in probabilistic graphical models and Bayesian networks, where understanding the relationship between variables given certain conditions is essential for accurate modeling.
20. **Difference between Classification and Prediction:**
    - **Classification:** In classification, the goal is to categorize data into predefined classes or labels. The model learns from labeled training data and assigns new instances to one of the known classes.
    - **Prediction:** Prediction, on the other hand, involves estimating a numerical value for a given input. Instead of assigning to discrete classes, prediction models aim to provide a continuous output. Regression analysis is a common technique used for prediction, where the model predicts a numeric value based on input features. The key distinction lies in the nature of the output: discrete categories for classification and continuous values for prediction.
21. **Calculating Number of Clusters:** The Elbow Method is a common technique to determine the optimal number of clusters in K-Means clustering. It involves plotting the variance explained as a function of the number of clusters and selecting the "elbow" point where the rate of decrease sharply changes.
22. **Selecting Attribute Subset:** Feature selection is a method to reduce data dimensionality. Techniques include filter methods (based on statistical measures), wrapper methods (using a specific classifier to evaluate subsets), and embedded methods (incorporated into the training of the model).
23. **Hypercube in Higher Dimensions:** In higher dimensions, a hypercube extends the concept of a cube. In three dimensions, it is a cube; in four dimensions, it becomes a hypercube. The term extends to describe structures in n-dimensional space, representing a generalization of a cube.
24. **Drawbacks of K-Means:**
    - Sensitivity to Initial Centroids
    - Difficulty with Non-Globular Shapes
    - Need to Specify Number of Clusters
    - Impact of Outliers on Centroid Calculation

25. **Attributes:** Attributes are characteristics or properties of data entities. In a database, attributes correspond to the columns in a table, defining the properties of the data stored in each record.
26. **Frequent Itemset:** A frequent itemset is a subset of items in a dataset that appears frequently together. In association rule mining, identifying frequent itemsets is crucial for discovering meaningful associations between items.
27. **Example-Based Question - RIP:** (Please provide the specific question for an accurate response)
28. **Overcoming Drawbacks of K-Means:**
    - Use of K-Means++
    - Implementation of Hierarchical K-Means
    - Applying Silhouette Analysis
    - Utilizing Density-Based Clustering (e.g., DBSCAN)
29. **Error Metrics:** Error metrics assess the performance of machine learning models. Common metrics include Mean Squared Error (MSE) for regression, and accuracy, precision, recall, F1 score, and ROC-AUC for classification.
30. **Cross-Validation Dataset:** In machine learning, a cross-validation dataset is a subset of the data used to assess a model's performance. It helps evaluate how well the model generalizes to new, unseen data by simulating the real-world scenario of training on one subset and validating on another.
31. **Factless Tables:** Factless tables in a data warehouse contain no measures or numerical data. They serve to represent events or activities, capturing relationships between dimensions without associated quantitative values. For example, a date and student enrollment table could be factless, capturing when a student enrolls but not the enrollment count.
32. **Classification vs. Clustering:**
    - **Classification:** Involves assigning predefined labels to instances based on patterns learned from labeled training data.
    - **Clustering:** Involves grouping similar instances together without predefined labels, discovering inherent patterns in the data.
33. **Star Schema vs. Snowflake:**
    - **Star Schema:** Central fact table connected to dimension tables in a star-like pattern, simplifying queries.
    - **Snowflake Schema:** Dimension tables are normalized, forming a snowflake pattern, which can save space but complicates queries.
34. **Top-Down vs. Bottom-Up:**
    - **Top-Down:** Approach starts with a global view and breaks it down into smaller components or subsystems.
    - **Bottom-Up:** Approach starts with individual components and builds up to a complete system.
35. **Slice vs. Dice (OLAP):**
    - **Slice:** Viewing a cube by fixing one dimension and varying others.

- **Dice:** Creating a subcube by fixing two or more dimensions and varying others.

36. **Cardinality of Star Scheme (One to Many):** In a star schema, a one-to-many relationship exists between the central fact table and dimension tables. For instance, one product can be associated with multiple sales records in the fact table.
37. **Data Visualization:** Data visualization is the graphical representation of data to uncover patterns, trends, and insights. It includes charts, graphs, and dashboards that enhance understanding and communication of complex information.
38. **Ensemble Models - Bagging and Boosting:** Ensemble models combine multiple weak learners to create a stronger model.
    - **Bagging (Bootstrap Aggregating):** Trains multiple models independently on random subsets of the data and averages their predictions.
    - **Boosting:** Trains models sequentially, giving more weight to misclassified instances to improve overall accuracy.
39. **Sequential and Parallel in Bagging and Boosting:**
    - **Bagging:** Can be implemented in parallel, as models are trained independently.
    - **Boosting:** Typically sequential, as each model's training depends on the performance of the previous one.
40. **Extrinsic and Intrinsic Methods in Clustering Evaluation:**
    - **Extrinsic Methods:** Use external criteria such as known class labels for evaluation.
    - **Intrinsic Methods:** Assess clustering quality based on internal measures, without relying on external information. Common metrics include Silhouette Score and Davies–Bouldin Index.
41. **Overfitting:** Overfitting occurs when a model learns the training data too well, capturing noise or random fluctuations rather than the underlying patterns. This can lead to poor generalization performance on new, unseen data.
42. **Tree Pruning:** Tree pruning is a technique in decision tree algorithms that involves removing branches or nodes to avoid overfitting. It helps simplify the model and improve its ability to generalize to new data.
43. **ROC Curve:** ROC (Receiver Operating Characteristic) curve is a graphical representation of a classification model's performance across different thresholds. It plots the true positive rate against the false positive rate, providing a visual tool to evaluate the trade-off between sensitivity and specificity.
44. **Voting Method:** Voting methods involve combining the predictions of multiple models to make a final decision. Common types include majority

voting (class with the most votes) and weighted voting (assigning different weights to model predictions).

45. **Quantile-Quantile Plots:** Q-Q plots compare the quantiles of two datasets to assess if they come from the same distribution. Points on the plot should fall along a straight line if the distributions match.
46. **Box Plot:** A box plot, or box-and-whisker plot, is a graphical representation of the distribution of a dataset. It displays the median, quartiles, and potential outliers, providing a concise summary of the data's spread.
47. **Regression:** Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. Types include linear regression, logistic regression, polynomial regression, and more.
48. **CLARANS:** CLARANS (Clustering Large Applications based on Randomized Search) is a clustering algorithm designed for large datasets. It uses a randomized approach to explore the data space efficiently.
49. **Normal Form of Dimension Tables in Snowflake Schema:** Dimension tables in a snowflake schema are typically in 3NF (Third Normal Form), minimizing redundancy and ensuring data integrity.
50. **Q-Q Plot:** Q-Q (Quantile-Quantile) plots compare the quantiles of a sample to those of a theoretical distribution, helping assess the fit between observed and expected distributions.
51. **Q-Plot:** (Please provide more context or details for a specific response)
52. **Entropy:** Entropy is a measure of the level of randomness or uncertainty in a dataset. In information theory, it quantifies the amount of information needed to describe an event.
53. **Boxplot:** A boxplot is a graphical representation of the distribution of a dataset. It displays the median, quartiles, and potential outliers, providing a concise summary of the data's spread.
54. **Different Ways to Calculate Distance:**
    - Hamming Distance
    - Euclidean Distance
    - Manhattan Distance
    - Minkowski Distance In K-Medoid clustering, Euclidean Distance is commonly used.
55. **Improving Efficiency of Apriori:**
    - Hashing
    - Sampling
    - Transaction Reduction
    - Partitioning
56. **Imbalanced Data:** Imbalanced data refers to datasets where one class significantly outnumbers the others. Handling methods include resampling, using different evaluation metrics, and utilizing specialized algorithms for imbalanced data.

57. **NF in Snowflake Schema:** In a snowflake schema, dimension tables are typically in 3NF (Third Normal Form), ensuring data integrity and reducing redundancy.
58. **Types of Crawlers:**
    - Traditional Crawler
    - Focused Crawler Differences include the scope of crawling (broad vs. focused on specific topics) and the frequency of updates.
59. **Page Rank and Web Mining:** Page Rank falls under the category of structure mining in web mining. It assesses the importance of webpages based on link structure.
60. **Pruning:** Pruning in the context of decision trees involves removing nodes or branches to enhance model generalization. Deleting a child node from a branch is one form of pruning.
61. **Drawbacks of Apriori:** Apriori has limitations in handling large volumes of data as the number of scans required for frequent itemset generation increases, impacting efficiency.
62. **Hierarchical Clustering Types and Termination Condition:** Types include agglomerative (bottom-up) and divisive (top-down) hierarchical clustering. Termination conditions can be based on a specific number of clusters or a similarity threshold.
63. **Adaboost:** Adaboost is an ensemble learning method that combines multiple weak learners to create a strong model. It sequentially trains models, giving more weight to misclassified instances in each iteration.
64. **Boosting Methods - Adaboost & Gradient Boosting:** Both Adaboost and Gradient Boosting are boosting methods in ensemble learning. Adaboost combines weak learners with a focus on misclassified instances, while Gradient Boosting builds models sequentially, minimizing errors.
65. **Slowly Changing Dimensions:** Slowly changing dimensions in data warehousing refer to dimensions that store and manage both current and historical data over time. It is crucial for maintaining a historical record of changes.
66. **Data Warehouse:** A data warehouse is a centralized repository for storing, managing, and analyzing large volumes of structured and unstructured data, supporting business intelligence and decision-making.
67. **Dendrogram:** A dendrogram is a tree-like diagram used in hierarchical clustering to represent the relationships between clusters. It illustrates the merging of clusters as the algorithm progresses.
68. **Strategic Information:** Strategic information is crucial for an enterprise to decide business strategies and establish goals. It guides high-level decision-making processes.
69. **Information Package Diagram (IPD):** An Information Package Diagram defines the relationships between subject matter and key performance

measures. It helps visualize how different components contribute to overall performance.

70. **Semi-Additive Attribute:** Semi-additive attributes are measures that have a different way of aggregation over time. Unlike fully additive measures, they cannot be summed across all dimensions
71. **Apriori and Methods to Improve:** Apriori is an algorithm for generating association rules. Methods to improve it include hash-based techniques, transaction reduction, sampling, and partitioning.
72. **Pre-Pruning:** Pre-pruning is a technique in decision tree algorithms where the tree is pruned during its construction, preventing further splitting of a node if it doesn't meet specific criteria. It helps prevent overfitting.
73. **True Positive:** True Positive (TP) refers to instances that are correctly identified as positive by a classification model. For example, in a medical test, true positives are patients correctly diagnosed with a disease.
74. **Precision, Recall, Specificity:**
    - **Precision:** Proportion of true positives among all predicted positives.
    - **Recall (Sensitivity):** Proportion of true positives among all actual positives.
    - **Specificity:** Proportion of true negatives among all actual negatives. Precision is useful when minimizing false positives is a priority, while recall is crucial for detecting all positives.
75. **Slice and Dice in OLAP:**
    - **Slice:** Viewing a cube by fixing one dimension and varying others.
    - **Dice:** Creating a subcube by fixing two or more dimensions and varying others.
76. **Iceberg Query:** Iceberg queries involve operations like GROUP BY and HAVING clauses, focusing on aggregations with a significant subset of data, filtering out non-substantial results.
77. **Data Reduction:** Data reduction involves decreasing data quantity while maintaining quality, improving efficiency and storage. Techniques include data cube aggregation, dimension reduction, data compression, discretization, and concept hierarchies.
78. **Techniques of Data Reduction:**
    - Data Cube Aggregation
    - Dimension Reduction (Stepwise Forward, Stepwise Backward)
    - Data Compression
    - Discretization
    - Concept Hierarchies
79. **Subject-Oriented vs. Application-Oriented Data Warehouses:**
    - **Subject-Oriented:** Organized around key subjects or business areas.

- **Application-Oriented:** Tailored to specific applications or departments. Similar to the difference between a database and a data warehouse.

80. **Need of a Data Warehouse:** Data warehouses are needed for:
    - Centralized data storage
    - Efficient querying and reporting
    - Historical analysis
    - Decision support
    - Integration of disparate data sources.
81. **Clustering:** Clustering is a data mining task that involves grouping similar instances together based on some similarity metric. It aims to discover inherent structures within the data.
82. **Parameter to Find Frequent Itemsets:** In addition to support and confidence, "Lift" is another parameter used to find frequent itemsets in association rule mining.
83. **Junk Dimension Table:** A junk dimension table combines several low-cardinality attributes into a single table, reducing the number of dimensions in a data warehouse.
84. **Features of Data Warehouse:**
    - Subject-Oriented
    - Integrated
    - Time-Variant
    - Non-Volatile
    - Support for Decision Making
85. **Multiple Concepts in a Single Fact Table:** Yes, more than one concept from a concept hierarchy can exist in a single fact table. For example, a sales fact table can have concepts like product and region.
86. **Data Mining Tasks and Techniques for Apriori Efficiency:**
    - Tasks: Prediction, Classification, Association, Clustering
    - Techniques: Hash-Based, Sampling, Reduction, Partitioning
87. **Confidence and Support:**
    - **Support:** Proportion of transactions containing a particular itemset.
    - **Confidence:** Conditional probability of one itemset given the presence of another.
88. **Spatial Data:** Spatial data includes information with a geographical component, such as coordinates, latitude, and longitude.
89. **Overcoming Disadvantages of K-Means:**
    - Using K-Means++
    - Implementing Hierarchical K-Means
    - Applying Silhouette Analysis
    - Utilizing Density-Based Clustering (e.g., DBSCAN)
    - Euclidean distance is commonly used in K-Means.

90. **Tree Pruning:** Tree pruning involves removing branches or nodes in a decision tree to prevent overfitting and enhance generalization.
91. **Business Intelligence vs. Data Mining:**
    - **Business Intelligence:** Focuses on reporting, querying, and analysis of historical data to support decision-making.
    - **Data Mining:** Involves discovering patterns, trends, and insights in data using various techniques, including machine learning.
92. **Information Delivery Component:** The information delivery component in a data warehouse involves delivering insights and reports to end-users. It includes tools for querying, reporting, and data visualization.
93. **Types of Reports Generated by Information Delivery:**
    - Operational Reports
    - Analytical Reports
    - Exception Reports
    - Key Performance Indicator (KPI) Reports
94. **Granularity:** Granularity refers to the level of detail in the data. It can be fine (detailed) or coarse (aggregated).
95. **Steps in Preprocessing:**
    - Data Cleaning
    - Data Integration
    - Data Transformation
    - Data Reduction
    - Data Discretization
96. **HITs (Hypertext Induced Topic Search):** HITs is an algorithm used by search engines to rank web pages based on their relevance to a given query.
97. **Multilevel Association Mining:** Multilevel association mining involves discovering associations at multiple levels of granularity in the data.
98. **Support in Apriori:** Support is the proportion of transactions containing a particular itemset. In Apriori, it is used to identify frequent itemsets.
99. **Importance of Confidence in Finding Frequent Itemset:** Confidence is important in finding frequent itemsets as it represents the conditional probability of one itemset given the presence of another.
100. **Use of Confidence:** Confidence is used to measure the reliability of association rules. It helps determine how often the antecedent implies the consequent in the dataset. Higher confidence values indicate stronger relationships between items in an association rule.