Expected Viva Questions

1. **What is data warehouse?**
   - A data warehouse is a electronic storage of an Organization's historical data for the purpose of reporting, analysis and data mining or knowledge discovery.

---

2. **What are the benefits of data warehouse?**
   - A data warehouse helps to integrate data and store them historically so that we can analyze different aspects of business including, performance analysis, trend, prediction etc. over a given time frame and use the result of our analysis to improve the efficiency of business processes.

What are some of the tasks of data mining?

A Following activities are carried out during data mining

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]Deviation Detection [Predictive]

---

3. **What do you mean by preprocessing of data in data mining ?**

A Before data is mined it has to be preprocessed. It consists of following three stages

- **Data cleaning** - Real world data is dirty so need to be cleaned
- **Data reduction**- Remove data not useful for mining
- **Data transformation** - Syntactic transformation

---

4. **What is Data cleaning ?**

A **Causes of Dirty Data**

- Missing values
- Noisy data (Human/Machine Errors)
- Inconsistent data

Data cleaning tasks

- Handling missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

---

5. Explain Data reduction?

It consists of following three tasks -

- **Dimensionality reduction** - Attribute subset selection
- **Numerosity reduction** - Tuple subset selection
- **Discretization** - Reduce the cardinality of active domain

---

6. What is Data Transformation? A It consist of following tasks

1. **Generalization** - concept hierarchy climbing
2. **Attribute/feature construction** - New attributes are constructed and added to the tuple
3. **Normalization** - scaled to fall within a small, specified range

---

7. What is the difference between OLTP and OLAP?

- OLTP is the transaction system that collects business data. Whereas OLAP is the reporting and analysis system on that data. OLTP systems are optimized for INSERT, UPDATE operations and therefore highly normalized. On the other hand, OLAP systems are deliberately denormalized for fast data retrieval through SELECT operations.

---

8. What is data mart?

- Data marts are generally designed for a single subject area. An organization may have data pertaining to different departments

like Finance, HR, Marketting etc. stored in data warehouse and each department may have separate data marts. These data marts can be built on top of the data warehouse.

---

9. What is dimension?

- A dimension is something that qualifies a quantity (measure). For an example, consider this: If I just say… "20kg", it does not mean anything. But if I say, "20kg of Rice (Product) is sold to Ramesh (customer) on 5th April (date)", then that gives a meaningful sense. These *product, customer* and *dates* are some dimension that qualified the measure - 20kg.
  Dimensions are mutually independent. Technically speaking, a dimension is a data element that categorizes each item in a data set into non-overlapping regions.

---

10. What is Fact?

- A fact is something that is quantifiable (Or measurable). Facts are typically (but not always) numerical values that can be aggregated.

---

11. Briefly state different between data ware house & data mart?

- Dataware house is made up of many datamarts. DWH contain many subject areas. but data mart focuses on one subject area generally. e.g. If there will be DHW of bank then there can be one data mart for accounts, one for Loans etc. This is high level definitions. Metadata is data about data. e.g. if in data mart we are receving any file. then metadata will contain information like how many columns, file is fix width/elimted, ordering of fileds, dataypes of field etc...

---

12. What are the storage models of OLAP?

- ROLAP, MOLAP and HOLAP

---

13. What are CUBES?

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

---

14. What is MODEL in Data mining world?

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

---

15. Explain how to mine an OLAP cube.

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

---

16. Define Rollup and cube.

- Custom rollup operators provide a simple way of controlling the process of rolling up a member to its parents values.The rollup uses the contents of the column as custom rollup operator for each member and is used to evaluate the value of the member's parents.
  If a cube has multiple custom rollup formulas and custom rollup members, then the formulas are resolved in the order in which the dimensions have been added to the cube.

---

17. Differentiate between Data Mining and Data warehousing.

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

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

---

18. What is Discrete and Continuous data in Data mining world?

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

---

19. What is a Decision Tree Algorithm?

- A decision tree is a tree in which every node is either a leaf node or a decision node. This tree takes an input an object and outputs some decision. All Paths from root node to the leaf node are reached by either using AND or OR or BOTH. The tree is constructed using the regularities of the data. The decision tree is not affected by Automatic Data Preparation.

---

20. What is Naïve Bayes Algorithm?

- Naïve Bayes Algorithm is used to generate mining models. These models help to identify relationships between input columns and the predictable columns. This algorithm can be used in the initial stage of exploration. The algorithm calculates the probability of every state of each input column given predictable columns possible states. After the model is made, the results can be used for exploration and making predictions.

---

21. 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.

---

22. Explain Association algorithm in Data mining?

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

---

23. What are the goals of data mining?

- Prediction, identification, classification and optimization

---

24. Is data mining independent subject?

- No, it is interdisciplinary subject. includes, database technology, visualization, machine learning, pattern recognition, algorithm etc.

---

25. What are different types of database?

- Relational database, data warehouse and transactional database.

---

26. What are data mining functionality?

- Mining frequent pattern, association rules, classification and prediction, clustering, evolution analysis and outlier Analise

---

27. What are issues in data mining?

- Issues in mining methodology, performance issues, user interactive issues, different source of data types issues etc.

28. List some applications of data mining.

- Agriculture, biological data analysis, call record analysis, DSS, Business intelligence system etc

---

29. What do you mean by interesting pattern?

- A pattern is said to be interesting if it is 1. easily understood by human 2. valid 3. potentially useful 4. novel

---

30. Why do we pre-process the data?

- To ensure the data quality. [accuracy, completeness, consistency, timeliness, believability, interpret-ability]

---

31. What are the steps involved in data pre-processing?

- Data cleaning, data integration, data reduction, data transformation.

---

32. List few roles of data warehouse manager.

- Creation of data marts, handling users, concurrency control, updation etc,

---

33. What are the forms of multidimensional model?

- Star schema
- Snow flake schema
- Fact constellation Schema

---

34. What are frequent pattern?

- A set of items that appear frequently together in a transaction data set.
- eg milk, bread, sugar

35. What are the issues regarding classification and prediction?

- Preparing data for classification and prediction
- Comparing classification and prediction

---

36. Define model over fitting.

- A model that fits training data well can have generalization errors. Such situation is called as model over fitting.

---

37. What are the methods to remove model over fitting?

- Pruning [Pre-pruning and post pruning)
- Constraint in the size of decision tree
- Making stopping criteria more flexible

---

38. What is regression?

- Regression can be used to model the relationship between one or more independent and dependent variables.
- Linear regression and non-linear regression

---

39. Compare K-mean and K-mediods algorithm.

- K-mediods is more robust than k-mean in presence of noise and outliers. K-Mediods can be computationally costly.

---

40. What is Baye's Theorem?

- $P(H/X) = P(X/H)* P(H)/P(X)$

---

41. What is concept Hierarchy?

- It defines a sequence of mapping from a set of low level concepts to higher -level, more general concepts.

42. What are the causes of model over fitting?

- Due to presence of noise
- Due to lack of representative samples
- Due to multiple comparison procedure

---

43. What is decision tree classifier?

- A decision tree is an hierarchically based classifier which compares data with a range of properly selected features.

---

44. What are different types of multimedia data?

- image, video, audio

---

45. What is text mining?

- **Text mining** is the procedure of synthesizing information, by analyzing relations, patterns, and rules among textual data. These procedures contains text summarization, text categorization, and text clustering.

---

46. List some application of text mining.

- Customer profile analysis
- patent analysis
- Information dissemination
- Company resource planning

---

47. What do you mean by web content mining?

- **Web content mining** refers to the discovery of useful information from Web contents, including text, images, audio, video, etc.

---

48. Define web structure mining and web usage mining.

- **Web structure mining** studies the model underlying the link structures of the Web. It has been used for search engine result ranking and other Web applications.

Web usage mining focuses on using data mining techniques to analyze search logs to find interesting patterns. One of the main applications of Web usage mining is its use to learn user profiles.

---

49. What are frequent patterns?

- These are the patterns that appear frequently in a data set.
- item-set, sub sequence, etc

---

50. What is data warehouse?

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

---

51. What is data characterization?

- Data Characterization is s summarization of the general features of a target class of data. Example, analyzing software product with sales increased by 10%

---

52. What is data discrimination?

- Data discrimination is the comparison of the general features of the target class objects against one or more contrasting objects.

---

53. What can business analysts gain from having a data warehouse?

- First, having a data warehouse may **provide a competitive** advantage by presenting relevant information from which to measure performance and make critical adjustments in order to help win over competitors.

Second, a data warehouse can **enhance business productivity** because it is able to quickly and efficiently gather information that accurately describes the organization.

Third, a data warehouse **facilitates customer relationship management** because it provides a consistent view of customers and item across all lines of business, all departments and all markets.

Finally, a data warehouse may **bring about cost reduction** by tracking trends, patterns, and exceptions over long periods in a consistent and reliable manner.

---

54. Why is association rule necessary?

- In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.
- It is intended to identify strong rules discovered in database using different measures of interesting.

---

55. What are two types of data mining tasks?

- Descriptive task
- Predictive task

---

56. Define classification.

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.

---

57. What are outliers?

- A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are called **outliers.**

---

58. Define KDD.

- The process of finding useful information and patterns in data.

---

59. What are the components of data mining?

- Database, Data Warehouse, World Wide Web, or other information repository

ØDatabase or Data Warehouse Server

ØKnowledge Based

ØData Mining Engine

ØPattern Evaluation Module

ØUser Interface

---

60. Define metadata.

- A database that describes various aspects of data in the warehouse is called metadata.

---

61. What are the usage of metadata?

- ØMap source system data to data warehouse tables

  ØGenerate data extract, transform, and load procedures for import jobs

  ØHelp users discover what data are in the data warehouse

  ØHelp users structure queries to access data they need

---

62. Define HOLAP.

- The hybrid OLAP approach combines ROLAP and MOLAP technology.

63. What are data mining techniques?

- Association rules
- Classification and prediction
- Clustering
- Deviation detection
- Similarity search
- Sequence Mining

---

64. List different data mining tools.

- Traditional data mining tools
- Dashboards
- Text mining tools

---

65. What is the main goal of data mining?

- Prediction

---

66. List the typical OLAP operations.

- Roll UP
- DRILL DOWN
- ROTATE
- SLICE AND DICE
- DRILL trough and drill across

---

67. Differentiate between star schema and snowflake schema.

- Star Schema is a multi-dimension model where each of its disjoint dimension is represented in single table.

•Snow-flake is normalized multi-dimension schema when each of disjoint dimension is represent in multiple tables.

•Star schema can become a snow-flake

•Both star and snowflake schemas are dimensional models; the difference is in their physical implementations.

•Snowflake schemas support ease of dimension maintenance because they are more normalized.

•Star schemas are easier for direct user access and often support simpler and more efficient queries.

•It may be better to create a star version of the snowflaked dimension for presentation to the users

---

68. List the advantages of star schema.

- •Star Schema is very easy to understand, even for non technical business manager.

  •Star Schema provides better performance and smaller query times

•Star Schema is easily extensible and will handle future changes easily

---

69. What are the characteristics of data warehouse?

- Integrated
- Non-volatile
- Subject oriented
- Time varient

---

70. Define support and confidence.

- The support for a rule R is the ratio of the number of occurrences of R, given all occurrences of all rules.

  The confidence of a rule X->Y, is the ratio of the number of occurrences of Y given X, among all other occurrences given X

---

71. What are the criteria on the basic of which classification and prediction can be compared?

- speed, accuracy, robustness, scalability, goodness of rules, interpret-ability

Tableau:

What is Tableau?

Tableau is a business intelligence software that allows anyone to connect to respective data, and then visualize and create interactive, shareable dashboards.

What are the different types of joins in Tableau?

The joins in Tableau are same as SQL joins. Take a look at the diagram below to understand it.

Left Join

Right Join

Inner Join

Full outer Join

How many maximum tables can you join in Tableau?

You can join a maximum of 32 tables in Tableau

What are sets?

Sets are custom fields that define a subset of data based on some conditions. A **set** can be based on a computed condition, for example, a **set** may contain customers with sales over a certain threshold.

Computed**sets** update as your data changes. Alternatively, a **set** can be based on specific data point in your view.

What are groups?

A group is a combination of dimension members that make higher level categories. For example, if you are working with a view that shows average test scores by major, you may want to group certain majors together to create major categories.

What is a hierarchical field?

A hierarchical field in tableau is used for drilling down data. It means viewing your data in a more granular level.

What is Tableau Data Server?

Tableau server acts a middle man between Tableau users and the data. Tableau Data Server allows you to upload and share data extracts, preserve database connections, as well as reuse calculations and field metadata. This means any changes you make to the data-set, calculated fields, parameters, aliases, or definitions, can be saved and shared with others, allowing for a secure, centrally managed and standardized dataset. Additionally, you can leverage your server's resources to run queries on extracts without having to first transfer them to your local machine.

What are the different filters in Tableau?

- ***Normal Filter*** is used to restrict the data from database based on selected dimension or measure. A Traditional Filter can be created by simply dragging a field onto the 'Filters' shelf.
- ***Quick filter*** is used to view the filtering options and filter each worksheet on a dashboard while changing the values dynamically (within the range defined) during the run time.
- ***Context Filter*** is used to filter the data that is transferred to each individual worksheet. When a worksheet queries the data source, it creates a temporary, flat table that is uses to compute the chart.

How to create a calculated field in Tableau?

- Click the drop down to the right of Dimensions on the Data pane and select "Create > Calculated Field" to open the calculation editor.
- Name the new field and create a formula.

What is the difference between joining and blending in Tableau?

- Joining term is used when you are combining data from the same source, for example, worksheet in an Excel file or tables in Oracle database
- While blending requires two completely defined data sources in your report.

How to add Custom Color to Tableau?

Adding a Custom Color refers to a power tool in Tableau. Restart you Tableau desktop once you save .tps file. From the Measures pane, drag the one you want to add color to **Color**. From the color legend menu arrow, select **Edit Colors**. When a dialog box opens, select the palette drop-down list and customize as per requirement.

Max no of tables we can join in Tableau?

We can join max 32 table, it's not possible to combine more than 32 tables.

What are Dimensions and Facts?

Dimensions is nothing but the descriptive text columns and facts are nothing but measures (numerical values) dimention ex:productname city..facts:sales, profit

Explain Pentaho?

It addresses the blockades that block the organization's ability to get value from all our data. Pentaho is discovered to ensure that each member of our team from developers to business users can easily convert data into value.

Explain the important features of Pentaho.

- Pentaho is capable of creating Advanced Reporting Algorithms regardless of their input and output data format.
- It supports various report formats, whether Excel spreadsheets, XMLs, PDF docs, CSV files.

- It is a Professionally Certified DI Software rendered by the renowned Pentaho Company headquartered in Florida, United States.
- Offers enhanced functionality and in-Hadoop functionality.
- Allows dynamic drill down into larger and greater information.
- Rapid Interactive response optimization.
- Explore and view multidimensional data.