Data Mining & Warehousing

1. What is data mining? In your answer, address the following:
   (a) Is it another hype?
   (b) Is it a simple transformation of technology developed from databases, statistics, and machine learning?
   (c) Explain how the evolution of database technology led to data mining.
   (d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.
2. Present an example where data mining is crucial to the success of a business. What data mining functions does this business need? Can they be performed alternatively by data query processing or simple statistical analysis?
3. How is a data warehouse different from a database? How are they similar to each other?
4. De_ne each of the following data mining functionalities: characterization, discrimination, association, classfication, prediction, clustering, and evolution and deviation analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.
5. Suppose your task as a software engineer at Big-University is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g.,undergraduate or graduate) of each student, and their cumulative grade point average (GPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?
6. Based on your observation, describe another possible kind of knowledge that needs to be discovered by data mining methods but has not been listed in this chapter. Does it require a mining methodology that is quite different from those outlined in this chapter?
7. What is the difference between discrimination and classification? Between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?
8. Describe three challenges to data mining regarding data mining methodology and user-interaction issues.
9. Describe two challenges to data mining regarding performance issues.
10. Write short notes on:
    - o Legacy systems
    - o Data warehouse
    - o Standard Business Applications
11. What is a Data warehouse? How does it differ from a database?
12. Discuss various factors, which lead to Data Warehousing.
13. Briefly discuss the history behind Data warehouse.
14. Write short notes on:
    - o Metadata
    - o Operational systems
    - o OLAP
    - o DSS
    - o Informational Systems
15. What is the need of a Data warehouse in any organization?
16. Discuss various characteristics of a Data warehouse.
17. Explain the difference between non-volatile and Subject-oriented data warehouse.
18. Write short notes on:
    - o Multi-Dimensional Views
    - o Operational Systems
19. What is the significance of an OLTP System?
20. Discuss OLTP related processes used in a Data warehouse.
21. Explain MD views with an example.
22. Identify various benefits of OLTP.
23. "OLAP enables organisations as a whole to respond more quickly to market demands, which often results in increased revenue and profitability". Comment.
24. Who are the primary users of Online Transaction Processing System?
25. "The KPI (key performance indicator) of an OLAP application is to provide just-in-time (JIT) information for effective decision-making". Explain.
26. What are the various kinds of models used in Data warehousing?

27. Discuss the following:
    - o Roll-up display
    - o Drill down operation
    - o Star schema
    - o Snowflake schema
28. Why is the star schema called by that name?
29. State an advantage of the multidimensional database structure over the relational database structure for data warehousing applications.
30. What is one reason you might choose a relational structure over a multidimensional structure for a data warehouse database? .
31. Clearly contrast the difference between a fact table and a dimension table.
32. Your college or university is designing a data warehouse to enable deans, department chairs, and the registrar's office to optimize course offerings, in terms of which courses are offered, in how many sections, and at what times. The data warehouse planners hope they will be able to do this better after examining historical demand for courses and extrapolating any trends that emerge.
    a). Give three dimension data elements and two fact data elements that could be in the database for this data warehouse. Draw a data cube, for this database.
    b). State two ways in which each of the two fact data elements could be of low quality in some respect.
33. You have decided to prepare a budget for the next 12 months based on your actual expenses for the past 12. You need to get your expense information into what is in effect a data warehouse, which you plan to put into a spreadsheet for easy sorting and analysis.
    a). What are your information sources for this data warehouse?
    b). Describe how you would carry out each of the five steps of data preparation for a data warehouse database, from extraction through summarization. If a particular step does not apply, say so and justify your statement.
34. Write short notes on:
    - o Data Quality Management
    - o OLAP
    - o DSS
    - o Data marts
    - o Operational data
35. Discuss Three-Layer Data Architecture with the help of a diagram.
36. What are the various Principles of Data warehouse?
37. What is the importance of a data warehouse in any organization? Where it is required?
38. What is data mining? In your answer, address the following:
    (a) Is it another hype?
    (b) Is it a simple transformation of technology developed from databases, statistics, and machine learning?
    (c) Explain how the evolution of database technology led to data mining.
    (d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.
39. Discuss different approaches used to build a Data Warehousing. Which approach is generally used for building a data warehouse?
40. Discuss various applications of a data warehouse.
41. Write short notes on:
    - o Data cleaning
    - o Back flushing
    - o Heterogeneous Sources
    - o Metadata repository
42. Discuss various steps involved in the acquisition of data for the data warehouse.
43. List out various processes involved in data storage in a data warehouse.
44. What are the important design considerations, which need to be thought of, while designing a data warehouse?
45. Explain the difference between distributed warehouse and the federated warehouse.
46. What are the nine decisions in the design of a data warehouse?
47. Write short notes on:
    - o Tighter Integration
    - o Empowerment
    - o Willingness
48. Briey compare the following concepts. You may use an example to explain your point(s).
    (a) data cleaning, data transformation, refresh.
    (b) discovery-driven cube, multi-feature cube, virtual warehouse.

.

49. Suppose that one needs to record three measures in a data cube: min, average, and median. Design an efficient computation and storage method for each measure given that the cube allows data to be deleted incrementally (i.e., in small portions at a time) from the cube.
50. A popular data warehouse implementation is to construct a multidimensional database, known as a data cube. Unfortunately, this may often generate a huge, yet very sparse multidimensional matrix.
    (a) Present an example, illustrating such a huge and sparse data cube.
    (b) Design an implementation method which can elegantly overcome this sparse matrix problem. Note that you need to explain your data structures in detail and discuss the space needed, as well as how to retrieve data from your structures, and how to handle incremental data updates.
51. In data warehouse technology, amultiple dimensional view can be implemented by a multidimensional database technique (MOLAP), or by a relational database technique (ROLAP), or a hybrid database technique (HOLAP).
    (a) Briey describe each implementation technique.
    (b) For each technique, explain how each of the following functions may be implemented:
        i. The generation of a data warehouse (including aggregation).
        ii. Roll-up.
        iii. Drill-down.
        iv. Incremental updating.
    Which implementation techniques do you prefer, and why?
52. In both data warehousing and data mining, it is important to have some hierarchy information associated with each dimension. If such a hierarchy is not given, propose how to generate such a hierarchy automatically for the following cases:
    (a) a dimension containing only numerical data.
    (b) a dimension containing only categorical data.
53. Consider the following multifeature cube query: Grouping by all subsets of fitem, region, monthg, find the minimum shelf life in 1997 for each group, and the fraction of the total sales due to tuples whose price is less than $100, and whose shelf life is within 25% of the minimum shelf life, and within 50% of the minimum shelf life.
    (a) Draw the multifeature cube graph for the query.
    (b) Express the query in extended SQL.
    (c) Is this a distributive multifeature cube? Why or why not?
54. What are the differences between the three main types of data warehouse usage: information processing, analytical processing, and data mining? Discuss the motivation behind OLAP mining (OLAM).
55. Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.
56. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.
57. Discuss issues to consider during data integration.
58. Using the data for age given in Question 27, answer the following:
    (a) Use min-max normalization to transform the value 35 for age onto the range [0; 1].
    (b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is ?.
    (c) Use normalization by decimal scaling to transform the value 35 for age.
    (d) Comment on which method you would prefer to use for the given data, giving reasons as to why.
59. Use a flow-chart to illustrate the following procedures for attribute subset selection:
    (a) step-wise forward selection.
    (b) step-wise backward elimination
    (c) a combination of forward selection and backward elimination.
60. Using the data for age given in Question 27:
    (a) Plot an equi-width histogram of width 10.
    (b) Sketch examples of each of the following sample techniques: SRSWOR, SRSWR, cluster sampling, stratied sampling.
61. Propose a concept hierarchy for the attribute age using the 3-4-5 partition rule.
62. Propose an algorithm, in pseudo-code or in your favorite programming language, for
    (a) the automatic generation of a concept hierarchy for categorical data based on the number of distinct values of attributes in the given schema,
    (b) the automatic generation of a concept hierarchy for numeric data based on the equi-width partitioning rule, and
    (c) the automatic generation of a concept hierarchy for numeric data based on the equi-depth partitioning rule.
63. Explain the need and importance of a data warehouse.
64. Discuss various design considerations, which are taken into account while building a data warehouse.

65. Describe the organization issues, which are to be considered while building a data warehouse.
66. Explain the need of Performance Considerations
67. “Organizations embarking on data warehousing development can choose one of the two approaches”. Discuss these two approaches in detail.
68. Explain the business considerations, which are taken into account while building a data warehouse.
69. Write short notes on:
    - o Meta data
    - o Data distribution
    - o Data content
    - o CASE tools
    - o Data marts
70. Write short notes on:
    - Access tools
    - Hardware platforms
    - Balanced approach
71. Discuss Data warehouse and DBMS specialization
72. Explain Optimal hardware architecture for parallel query scalability
73. Describe the Technical issues, which are to be considered while building a data warehouse.
74. Explain the Implementation Considerations, which are taken into account while building a data warehouse
75. Write short notes on:
    - o Query Performance
    - o Integrated Dimensional Analysis
    - o Load Performance
    - o Mass User Scalability
    - o Terabyte Scalability
76. Discuss in brief Criteria for a data warehouse
77. Explain Tangible benefits. Provide suitable examples for explanation.
78. Discuss various problems with data warehousing
79. Explain Intangible benefits. Provide suitable examples for explanation.
80. Discuss various benefits of a Data warehouse.