cap5771sp25
February 26, 2025

Exam 1 part 1 version b

Question 1. According to the lecture on data exploration, why might modern data-intensive organizations prefer a schema-on-read approach instead of schema-on-write when managing rapidly evolving datasets?

(a) `Schema-on-read allows only structured data to be stored efficiently.`

(b) `Schema-on-read enforces strict governance during data collection, ensuring higher data quality.`

(c) `Schema-on-read enables data ingestion without transformation, supporting agile decision-making when new data types emerge.`

(d) `Schema-on-write simplifies the integration of social network data with relational systems by delaying schema creation.`

(e) `Schema-on-read accelerates query performance by preloading metadata definitions.`

Question 2. According to lecture materials on exploratory data analysis, a data scientist observes two histograms with identical five-number summaries but differing bin structures during a quality control audit of pharmaceutical batch processing times. Which fundamental analytical limitation of boxplots does this scenario best demonstrate?

(a) `Boxplots confuse stakeholders by using medians instead of means`

(b) `Boxplots obscure distribution modality and density variations across value ranges`

(c) `Boxplots inaccurately represent continuous data as discrete quartiles`

(d) `Boxplots exaggerate outlier impacts through 1.5xIQR whisker thresholds`

(e) `Boxplots require minimum sample sizes exceeding 100 observations`

Question 3. According to the lecture on attribute types in Data Exploration, which of the following best describes an ordinal attribute?

(a) `A special type of numeric attribute that always follows a normal distribution.`

(b) `A categorical attribute with a meaningful order but without a known magnitude between successive values, such as rankings.`

(c) `A binary attribute with only two states, such as positive and negative test results .`

(d) A categorical attribute with distinct labels but no meaningful order, such as hair color.

(e) A numeric attribute with real values that can be measured continuously, such as temperature.

Question 4. Which of the following best describes a data object in the context of data science?

(a) A special type of binary data structure used for machine learning models.

(b) A representation of an entity that consists of multiple attributes.

(c) A single attribute that defines the characteristics of an entity.

(d) A temporary dataset that is used only for intermediate calculations.

(e) A unique identifier used to differentiate records in a database.

Question 5. A dataset has a highly right-skewed (positively skewed) distribution. How will the mean and median compare in this case?

(a) The mean will be less than the median.

(b) The median will be more affected by extreme values than the mean.

(c) The mean and median will be equal.

(d) The mean will be greater than the median.

(e) The mean is always a more reliable measure of central tendency in skewed data.

Question 6. In the context of data management, what is a data dictionary?

(a) A collection of data facts stored in a csv format.

(b) A centralized repository that contains metadata about data, including its structure , definitions, and relationships.

(c) A visualization tool used to represent data in graphical formats.

(d) A programming library used for managing and manipulating databases.

(e) A type of dictionary data structure in Python used to store key-value pairs.

Question 7. According to the lecture on data exploration, how do traditional databases stores load data into a table?

(a) `noSQL.`

(b) `Schema-On-Read.`

(c) `Schema-On-Write.`

(d) `Schemaless.`

(e) `Schemamore.`

Question 8. According to the lecture, if you have a table named Students with columns (sid, name, age, gpa), then in relational database terminology, each row in the Students table is referred to as which of the following?

(a) `Column`

(b) `Schema`

(c) `Attribute`

(d) `Instance`

(e) `Database`

Question 9. According to lecture data exploration demo, when using Python to interact with a database, the cursor.execute(...) method is commonly used to execute SQL queries or commands. Which of the following statements correctly describes the usage of cursor.execute(...)?

(a) `cursor.execute(...) is used to create new database tables.`

(b) `cursor.execute(...) returns a database connection object.`

(c) `cursor.execute(...) takes an SQL query as an argument and executes it.`

(d) `cursor.execute(...)  is used to close the database connection.`

(e) `cursor.execute(...) is used to commit changes to the database.`

Question 10. Which of the following is true about semi-structured data models like XML?

(a) They are self-describing documents that can represent tree structures and free-text .

(b) They represent data using a set of rigid tables with rows and columns.

(c) They rely on predefined schemas that must be applied before data is stored.

(d) They do not allow for hierarchical relationships or nesting of elements.

(e) They are mostly used for numeric data and require precise data types for every field.

Question 11. What is a five-number summary used in boxplot analysis?

(a) Minimum, Q1, Mean, Q3, Maximum

(b) Q1, Median, Q3, IQR, Outliers

(c) Range, Variance, Standard Deviation, IQR, Mean

(d) Minimum, Q1, Median, Q3, Maximum

(e) Mean, Median, Mode, Variance, Standard Deviation

Question 12. According to the lecture on database querying, what is the primary reason for using a cursor in SQL operations?

(a) A cursor stores query results permanently in the database for future use.

(b) A cursor automatically optimizes SQL queries to improve performance.

(c) A cursor allows for row-by-row processing of query results, making it useful for handling large datasets.

(d) A cursor is required to execute any SQL query, including CREATE TABLE statements.

(e) A cursor prevents SQL injection attacks by default, without needing parameterized queries.

Question 13. Scenario: A sports analytics company is analyzing player performance statistics from multiple basketball games. Each game is recorded with numerical values such as points scored, assists, and rebounds for each player. According to the lecture on data exploration from slide 27, which type of numeric attribute best describes the number of points a player scores in a game?

(a) `Discrete numeric attribute`

(b) `Continuous numeric attribute`

(c) `Ordinal attribute`

(d) `Nominal attribute`

(e) `Binary attribute`

Question 14. According to the data-exploration lecture, which of the following statements best describes the differences between schema-on-write and schema-on-read approaches in data management?

(a) `Schema-on-read requires data transformation before loading, while schema-on-write allows data to be copied directly to the file store without transformation.`

(b) `Schema-on-write provides more flexibility than schema-on-read by allowing data to be loaded without a predefined structure.`

(c) `Schema-on-read allows for more flexibility by deferring schema application until data is read, while schema-on-write requires a predefined structure before data can be loaded, resulting in faster read times but less agility.`

(d) `Schema-on-read is only applicable to structured data formats like relational databases, while schema-on-write is used for semi-structured data like XML.`

(e) `Schema-on-write and schema-on-read are interchangeable terms referring to the same data loading approach in traditional databases.`

Question 15. According to the lecture on data exploration from slide 10, which of the following statements about schema-on-read is correct?

(a) `Schema is applied before data is stored.`

(b) `Schema is applied while reading the data.`

(c) `Schema-on-read is commonly used in traditional relational databases.`

(d) `Schema-on-read enforces strict data structure constraints.`

(e) `Schema-on-read makes data loading slower than schema-on-write.`

Question 16. According to Lecture 1 on Data Exploration, which of the following is NOT an attribute type in data classification?

(a) `Ratio`

(b) `Relational`

(c) `Ordinal`

(d) `Nominal`

Question 17. According to lecture on data exploration, What type of distribution does the below description represent?

'When plotting a dataset, the mean is greater than the median and the distribution has a longer tail extending to the right. '

(a) `Positively Skewed Distribution`

(b) `Negatively Skewed Distribution`

(c) `Bimodal Distribution`

(d) `Normal Distribution`

(e) `Symmetric Distribution`

Question 18. In data classification, how are attributes like ID numbers and ZIP codes categorized?

(a) `Quantitative attributes, because they consist of numerical values that can be used in mathematical operations.`

(b) `Both qualitative and quantitative attributes, depending on the context in which they are used.`

(c) `Qualitative attributes, because they represent categorical information without inherent numerical meaning.`

(d) `Neither qualitative nor quantitative, as they do not fit into standard data classification categories.`

(e) `Only ZIP codes are quantitative, while ID numbers are qualitative.`

Question 19. According to the data exploration demo, assume we are given a table called movies with columns (title, director, genre, box office record), which SQL query tells us the genres that have more than 100 movies that have achieved a box office record of $500,000 or more?

(a) SELECT genre, count(*) FROM movies WHERE boxofficerecord >= 500000 GROUP BY genre HAVING count(*) > 100

(b) SELECT title, count (*) FROM movies WHERE boxofficerecord >= 500000 GROUP BY title HAVING count(*) > 100

(c) SELECT genre, count(*) FROM movies WHERE boxofficerecord >= 500000

(d) SELECT genre, count(*) FROM movies GROUP BY genre HAVING count(*) > 100 WHERE boxofficerecord >= 500000

(e) SELECT genre, count(*) FROM movies GROUP BY genre HAVING count(*) > 100

Question 20. According to data exploration demo, what is the meaning of the cursor.fetchall() in this code block?

```
cursor.execute("SELECT tracks.Name, albums.Title, artists.Name
            FROM tracks JOIN albums ON tracks.AlbumId = albums.AlbumId
              JOIN artists ON albums.ArtistId = artists.ArtistId")
tracks_albums_artists = cursor.fetchall()
for item in tracks_albums_artists:
    print(item)
```

(a) It will crash your whole system

(b) It retrieves the output as the a of tuples

(c) It forms a new SQL query

(d) It creates a new SQL table and stores the data

(e) It retrieves only one row from the SQL output

Question 21. According to the lecture on Data Exploration, Which statement correctly describes the cardinality and arity of this data.

Student ID Name Age Major
101 Peter 24 Mathematics
102 Ram 22 Computer Science
103 James 19 Physics

(a) The cardinality of the Students table is 3, and the arity is 4

(b) The Arity of the Student Table is 4, but cardinality refers to the number of relationships between different tables, as we don't have a second table cardinality cannot be defined.

(c) `The cardinality of the Student table is 12, and the arity is 3`

(d) `The cardinality of the Students table is 4,  and the arity is 3`

(e) `The cardinality and arity refer to the same thing, and the value for this table is 4`

Question 22. According to the lecture on data exploration, which one of the following is not apart of ”The 5 Vs of Big Data”?

(a) `Volume`

(b) `Visualization`

(c) `Veracity`

(d) `Velocity`

(e) `Variety`

Question 23. According to the lecture on data exploration, which of the following types of data is considered semi-structured?

(a) `Customer service phone call recordings`

(b) `SQL application database`

(c) `Excel spreadsheets`

(d) `Server log files`

(e) `Video files`

Question 24. According to the data exploration lecture, what is the primary reason for detecting and handling missing data during the data preparation phase?

(a) `To make the dataset look more professional in presentations.`

(b) `To ensure the accuracy and reliability of statistical analyses and machine learning models.`

(c) `To improve data visualization aesthetics by eliminating blank spaces.`

(d) `To reduce the file size of the dataset for faster processing.`

(e) `To automatically generate new data to replace the missing values without validation .`

Question 25. According to the lecture on Data Exploration, which type of attribute is represented by the following scenario? In a race, five athletes finish at different times. You rank them from 1st place to 5th place based on who finishes first to last. What type of attribute is used when ranking athletes from 1st to 5th?

(a) `Binary attribute.`

(b) `Ratio attribute.`

(c) `Ordinal attribute.`

(d) `Interval attribute.`

(e) `Nominal attribute.`