

PRACTICE QUESTIONS

Snowflake

SnowPro Advanced: Data Scientist

SHARED BY [u/goldenferret1984](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM Snowflake SnowPro Advanced: Data Scientist

Shared by [u/goldenferret1984](#) on aiexamcerts

10 practice questions for the Snowflake SnowPro Advanced: Data Scientist exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/snowflake/snowpro-advanced-data-scientist. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Which statement about data leakage is correct?

- A** Leakage only happens when test rows are shuffled into training
- B** Leakage occurs when information unavailable at prediction time enters the training features
- C** Leakage goes away once more features are added, because extra columns dilute any leaked signal
- D** Leakage affects only deep neural network training runs

Answer: B

Leakage happens when a feature carries information from after the prediction point or from the target itself, which inflates offline scores and collapses in production. Shuffling test rows into training is a different mistake, so the option that limits leakage to that case is too narrow. Adding features makes leakage more likely, not less.

2. When rows are ordered by event time, a random train and test split can let future information reach the training set, so a _____ split that trains on earlier rows and tests on later rows is the safer choice.

Answer: time-based (also accepted: time based, temporal, chronological)

A time-based split mirrors how the model will be used, training on the past and scoring the future. A random split can mix future behavior into training and overstate accuracy. Shuffled cross-validation has the same weakness when rows are time dependent.

3. Which practice keeps raw data ingestion reproducible when a team loads CSV files from an internal stage?

- A** Use INSERT statements typed by hand for every row of every file that arrives in the stage
- B** Load into a raw table with COPY INTO and keep a load history to avoid duplicate loads
- C** Overwrite the raw table with each new file
- D** Convert the files to spreadsheets before loading

Answer: B

COPY INTO tracks which staged files were loaded and avoids reloading them, and a raw table preserves the original values for replay. Hand-typed inserts are not reproducible, and overwriting without records loses the audit trail. Converting to spreadsheets adds a lossy step the pipeline does not need.

4. Snowflake stores data in micro-partitions, and per-column metadata such as minimum and maximum values lets the engine skip partitions that cannot match a filter.

- A True
- B False

Answer: A (True)

The statement is true. Micro-partition pruning uses stored column statistics to avoid scanning partitions that cannot satisfy a predicate, which is why clustering and selective filters improve performance on large tables.

5. Which two statements about categorical encoding are correct? (Choose two.)

- A Frequency encoding can serve as a compact alternative for some high-cardinality features
- B One-hot encoding always preserves the order of categories
- C Target encoding never needs regularization or cross-fitting
- D Ordinal encoding is appropriate for categories with a real rank order
- E Label encoding of nominal categories is always preferred for linear models

Answer: A, D

Ordinal encoding matches categories that have a true order, such as education levels, and frequency encoding can be a compact alternative when counts carry signal. One-hot encoding discards order, target encoding needs regularization to avoid leakage, and label encoding nominal values imposes a fake order on linear models.

6. Which feature is a rolling aggregate that is safe to compute for a customer on day D?

- A Average order value over the 30 days before day D
- B Total orders over the full lifetime including day D and later
- C Whether the customer ultimately churned after day D
- D Refund amount recorded for the order placed on day D+10

Answer: A

A trailing window that closes before the prediction date uses only information that exists at scoring time. Lifetime totals that include later days and churn outcomes leak the future, and a refund recorded after the decision is the same mistake.

7. Which Snowflake resource provides container compute, including GPU options, for training jobs that need more than a standard warehouse offers?
- A A compute pool for container-based workloads
 - B A network policy for IP allowlists and blocklists on the account
 - C A masking policy on the target column
 - D A share of a read-only database

Answer: A

Compute pools provide container compute, including GPU options, for workloads that outgrow a warehouse-based approach. Network policies restrict connections, masking policies protect values, and shares distribute data. None of them adds training compute.

8. Hyperparameter search over a grid of candidate settings is commonly called _____ search.

Answer: grid (also accepted: grid-search)

Grid search exhaustively evaluates each combination in a predefined grid, usually with cross-validation. Random search samples the space, which often finds good settings faster when only a few hyperparameters matter.

9. Which curve plots precision against recall across decision thresholds?
- A Learning curve
 - B Precision-recall curve
 - C Lift chart by decile only
 - D Residual plot of a linear fit

Answer: B

A precision-recall curve traces the tradeoff between precision and recall as the threshold moves, which is especially informative for imbalanced data. A learning curve tracks error against training size, a decile lift chart summarizes ranking gains, and a residual plot inspects regression errors.

10. Which hyperparameter tuning approach samples random combinations instead of testing every combination in a grid?
- A Exhaustive grid enumeration with no sampling
 - B Manual copying of the best SQL query
 - C Random search
 - D Using default settings only with no tuning

Answer: C

Random search draws configurations from the search space, often covering important dimensions more efficiently than a dense grid when few hyperparameters matter. Exhaustive enumeration tests every combination, and copying queries or using defaults only are not sampling strategies.