

PRACTICE QUESTIONS

NVIDIA

NCP-ADS

NVIDIA-Certified Professional: Accelerated Data Science

SHARED BY [u/warmcloud1984](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 18, 2026
EXAM	NVIDIA NCP-ADS, NVIDIA-Certified Professional: Accelerated Data Science

Shared by [u/warmcloud1984](#) on aiexamcerts

10 practice questions for the NVIDIA NCP-ADS exam, NVIDIA-Certified Professional: Accelerated Data Science, each followed by its answer and a short explanation. More free sets and the full practice exam: aiexamcerts.com/nvidia/ncp-ads. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCP-ADS and the question number.

1. Which file format is best suited for repeated analytic queries that touch only a few columns of a very wide table?
 - A Plain text with fixed-width records
 - B XML documents with deeply nested tags
 - C One JSON object per line that holds every field
 - D Columnar Parquet files with compression

Answer: D

Columnar storage lets an engine read only the required columns and skip the rest, and compression reduces scan time. Row-oriented text and JSON read whole records, which is the tempting but slower pattern. XML is verbose and slow to parse.

2. For a column of integers between 0 and 100, an 8-bit _____ integer type stores each value in one byte.

Answer: unsigned (also accepted: uint8, unsigned 8-bit)

An unsigned 8-bit type holds values from 0 to 255 in one byte, so a column of values up to 100 fits without loss. A signed 8-bit type would also fit these values, but it wastes the range below zero.

3. Which two techniques reduce GPU memory use during training? (Choose two.)
 - A Gradient checkpointing, which recomputes activations instead of storing all of them
 - B Doubling the batch size to average out gradient noise
 - C Mixed precision training with 16-bit tensors for most operations
 - D Keeping a full copy of every epoch's predictions on the GPU for logging
 - E Deleting the optimizer state after every step

Answer: A, C

Gradient checkpointing trades extra compute for memory by recomputing activations, and mixed precision halves the storage of many tensors. Doubling the batch size increases memory use, and keeping prediction copies on the GPU adds to it. Deleting optimizer state would break the updates.

4. Which cuGraph result identifies groups of nodes that can all reach one another in an undirected graph?
- A Connected components of the graph
 - B Learning rate schedule that decays the step size after each epoch during training
 - C Confusion matrix
 - D Batch normalization layer

Answer: A

Connected components partition an undirected graph into groups where every node can reach every other node in its group. The other options are a training schedule, an evaluation table and a neural network layer, so none describes graph structure.

5. Which two statements about hyperparameter search are correct? (Choose two.)
- A Grid search always finds a better model than random search for the same budget
 - B Random search tries more distinct values of each important hyperparameter than a grid with the same number of trials
 - C Bayesian optimization ignores the results of earlier trials
 - D Early stopping can save compute by halting runs whose validation loss stops improving
 - E Hyperparameters are learned directly from the training data

Answer: B, D

Random search samples more distinct values per important dimension, and early stopping halts unpromising runs to save compute. Grid search does not always win at equal budgets, and Bayesian optimization uses past trial results. Hyperparameters are set by the practitioner rather than learned from the weights.

6. Which two statements describe the value of a rolling window in exploratory time-series work? (Choose two.)
- A It removes seasonality automatically without any other step
 - B It smooths short-term noise so the underlying trend is easier to see
 - C It guarantees that the series is stationary
 - D It provides a local baseline against which unusual points can be flagged
 - E It replaces the need for timestamps in the index

Answer: B, D

A rolling window smooths noise so trends stand out, and a rolling mean or spread gives a local baseline for flagging anomalies. A rolling window does not remove seasonality or guarantee stationarity, which are tempting but incorrect claims. It also still depends on an ordered index.

7. Standardizing features to zero mean and unit variance is required before any tree-based model can be trained.
- A True
 - B False

Answer: B (False)

Tree-based models split on thresholds, so a monotonic rescaling does not change their splits, and standardization is optional for them. It is useful for models that depend on distances or gradients, which is why the blanket requirement is the tempting error.

8. In cuDF and pandas, the method that computes a moving aggregate over a fixed number of preceding rows is called _____.

Answer: rolling

The rolling method applies an aggregate such as a mean, sum or standard deviation over a moving window of rows. The resample method changes the time frequency instead, which is a different operation.

9. Which setting most directly determines which GPU a Dask-CUDA worker process uses?
- A The Git branch name used for the project, including its tags and commit history
 - B The CUDA_VISIBLE_DEVICES value that each worker process sees
 - C The Conda channel priority setting
 - D The Jupyter notebook theme

Answer: B

Each Dask-CUDA worker is pinned to the devices visible to it through CUDA_VISIBLE_DEVICES, which is how one GPU is assigned per worker. Branch names, channel priority and notebook themes have no effect on device assignment.

10. A query-latency histogram has a long right tail. Which summary is most informative for reporting typical latency?
- A The total of all latencies added together
 - B The row count of the log table, including every retry and duplicate entry
 - C The median with 95th and 99th percentiles
 - D The most common timestamp in the log

Answer: C

The median describes the typical request, and high percentiles show the tail that users notice under load. The total of latencies is dominated by the slow requests, which is the tempting mistake. Row counts and timestamps say nothing about latency.