

PRACTICE QUESTIONS

NVIDIA

NCP-ADS

NVIDIA-Certified Professional: Accelerated Data Science

SHARED BY [u/mistybeacon1997](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 18, 2026
EXAM	NVIDIA NCP-ADS, NVIDIA-Certified Professional: Accelerated Data Science

Shared by [u/mistybeacon1997](#) 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. A model that scores well on a validation set can still degrade in production because of data drift or training-serving skew.

- A True
- B False

Answer: A (True)

Input distributions can change after deployment, and preprocessing that differs between training and serving produces skew. The tempting error is assuming that good validation scores guarantee production performance.

2. Which Dask-CUDA component starts one worker per visible GPU on a single machine?

- A `pandas.read_excel` with multiple threads
- B Matplotlib subplot grids
- C Git submodules
- D LocalCUDACluster

Answer: D

LocalCUDACluster starts one worker per visible GPU on the machine, so Dask-cuDF partitions land on different devices. The pandas reader loads files on the CPU, and the other two options are a plotting utility and a version control feature.

3. Which practice best prevents a preprocessing step from leaking test information into training?

- A Fitting the imputer and scaler on the training split only, then applying them to the test split
- B Fitting the scaler on the full dataset before splitting
- C Using the test labels to choose which rows to drop before any split is made, which tunes the data to the evaluation
- D Computing target statistics from every row, including test rows

Answer: A

Fitting preprocessing on the training split and only applying it to the test split keeps the evaluation independent of the held-out rows. Fitting the scaler on the full dataset is the tempting shortcut, and it leaks test statistics into training.

4. Which technique lets a model train with an effective batch larger than GPU memory allows?
- A Storing the entire dataset in one giant text cell so every batch is read from a single string
 - B Gradient accumulation over several micro-batches before each optimizer step
 - C Disabling backpropagation for every layer
 - D Using a smaller validation split

Answer: B

Gradient accumulation sums gradients across several small micro-batches and then applies one optimizer step, which emulates a larger batch. Disabling backpropagation stops learning entirely, and a smaller validation split does not change the training batch at all.

5. Which tool packages a GPU application together with its system libraries so it runs the same way on another host?
- A A spreadsheet listing package names and their versions for each machine in the cluster
 - B A Jupyter theme file
 - C A container image built from a pinned base image
 - D A screenshot of the desktop

Answer: C

A container image bundles the application with its user-space libraries on a pinned base, which is how one environment is reproduced on another host. A spreadsheet of package names omits system libraries, which is the tempting but incomplete option.

6. A Dask collection that is computed once and kept in worker memory for later stages is said to be _____.

Answer: persisted

Persisting keeps the computed partitions in worker memory so later stages reuse them instead of recomputing the full task graph. This helps when several downstream steps read the same intermediate result.

7. Which step most reduces the amount of data a distributed groupby must shuffle?
- A Adding a random column to each row
 - B Converting every value to a string
 - C Running the aggregation twice in a loop over the full raw table each time it is needed
 - D Filtering out unneeded columns and rows before the groupby

Answer: D

Removing unneeded columns and rows before the groupby shrinks the data that has to move between workers. Adding columns or repeating the aggregation increases the work, and string conversion gives no benefit.

8. Which change most directly reduces the memory needed for model weights and activations?
- A Lowering numeric precision, such as FP16 or INT8 with calibration
 - B Increasing the number of layers in the network to add more trainable parameters to it
 - C Storing the dataset in XML format
 - D Adding more validation folds

Answer: A

Lower-precision formats store each value in fewer bytes, which shrinks both weights and activations. Adding layers increases memory, which is the tempting but opposite effect. XML storage and extra validation folds do not change model memory.

9. Which plot best compares the distributions of latency across five server types?
- A Pie chart of server names
 - B Side-by-side box plots on a shared axis
 - C 3D bar chart with no scale
 - D Network graph of the servers

Answer: B

Side-by-side box plots on a shared axis show the median, spread and outliers for each server type, so the distributions can be compared directly. A pie chart shows shares of a total, and a network graph shows links rather than distributions.

10. Which setup best checks whether a GPU benchmark result will hold for production?
- A Testing once on a developer laptop with a small sample of rows
 - B Using the largest GPU available regardless of price or whether its memory matches the workload
 - C Repeating the benchmark on production hardware and data, with warm-up runs
 - D Comparing only the file names of the outputs

Answer: C

Repeating the benchmark on the production hardware with production-sized data and warm-up runs gives results that transfer to the real workload. A single laptop run is the tempting shortcut, and it does not reflect production memory or throughput.