

Exam : NCA-AIIO

**Title : AI Infrastructure and
Operations**

Version : DEMO

1.An enterprise is deploying a large-scale AI model for real-time image recognition. They face challenges with scalability and need to ensure high availability while minimizing latency.

Which combination of NVIDIA technologies would best address these needs?

- A. NVIDIA CUDA and NCCL
- B. NVIDIA Triton Inference Server and GPUDirect RDMA
- C. NVIDIA DeepStream and NGC Container Registry
- D. NVIDIA TensorRT and NVLink

Answer: D

2.A company is using a multi-GPU server for training a deep learning model. The training process is extremely slow, and after investigation, it is found that the GPUs are not being utilized efficiently. The system uses NVLink, and the software stack includes CUDA, cuDNN, and NCCL.

Which of the following actions is most likely to improve GPU utilization and overall training performance?

- A. Increase the batch size
- B. Update the CUDA version to the latest release
- C. Disable NVLink and use PCIe for inter-GPU communication
- D. Optimize the model's code to use mixed-precision training

Answer: A

3.In an AI data center, you are responsible for monitoring the performance of a GPU cluster used for large-scale model training.

Which of the following monitoring strategies would best help you identify and address performance bottlenecks?

- A. Monitor only the GPU utilization metrics to ensure that all GPUs are being used at full capacity.
- B. Focus on job completion times to ensure that the most critical jobs are being finished on schedule.
- C. Track CPU, GPU, and network utilization simultaneously to identify any resource imbalances that could lead to bottlenecks.
- D. Use predictive analytics to forecast future GPU utilization, adjusting resources before bottlenecks occur.

Answer: C

4.You are assisting a senior data scientist in analyzing a large dataset of customer transactions to identify potential fraud. The dataset contains several hundred features, but the senior team member advises you to focus on feature selection before applying any machine learning models.

Which approach should you take under their supervision to ensure that only the most relevant features are used?

- A. Select features randomly to reduce the number of features while maintaining diversity.
- B. Ignore the feature selection step and use all features in the initial model.
- C. Use correlation analysis to identify and remove features that are highly correlated with each other.
- D. Use Principal Component Analysis (PCA) to reduce the dataset to a single feature.

Answer: C

5.You are evaluating the performance of two AI models on a classification task. Model A has an accuracy of 85%, while Model B has an accuracy of 88%. However, Model A's F1 score is 0.90, and Model B's F1

score is 0.88.

Which model would you choose based on the F1 score, and why?

- A. Model A - The F1 score is higher, indicating better balance between precision and recall.
- B. Model B - The higher accuracy indicates overall better performance.
- C. Neither - The choice depends entirely on the specific use case.
- D. Model B - The F1 score is lower but accuracy is more reliable.

Answer: A