Boost up Your Certification Score

NVIDIA
NCA-AIIO
NVIDIA-Certified Associate AI Infrastructure and Operations

Exams Boost
Boost Up Your Career

For More Information – Visit link below:

https://www.examsboost.com/

Product Version

✓ Up to Date products, reliable and verified.
✓ Questions and Answers in PDF Format.

Visit us at: https://www.examsboost.com/test/nca-aiio

Latest Version: 7.1

Question: 1

A company is implementing a new network architecture and needs to consider the requirements and considerations for training and inference. Which of the following statements is true about training and inference architecture?

A. Training architecture and inference architecture have the same requirements and considerations.
B. Training architecture is only concerned with hardware requirements, while inference architecture is only concerned with software requirements.
C. Training architecture is focused on optimizing performance while inference architecture is focused on reducing latency.
D. Training architecture and inference architecture cannot be the same.

Answer: C

Explanation:
Training architectures are designed to maximize computational throughput and accelerate model convergence, often by leveraging distributed systems with multiple GPUs or specialized accelerators to process large datasets efficiently. This focus on performance ensures that models can be trained quickly and effectively. In contrast, inference architectures prioritize minimizing response latency to deliver real-time or near-real-time predictions, frequently employing techniques such as model optimization (e.g., pruning, quantization), batching strategies, and deployment on edge devices or optimized servers. These differing priorities mean that while there may be some overlap, the architectures are tailored to their specific goals—performance for training and low latency for inference.
(NVIDIA AI Infrastructure and Operations Study Guide, Section on Infrastructure Considerations for AI Workloads; NVIDIA Documentation on Training and Inference Optimization)

Question: 2

For which workloads is NVIDIA Merlin typically used?

A. Recommender systems
B. Natural language processing
C. Data analytics

Answer: A

Explanation:
NVIDIA Merlin is a specialized, end-to-end framework engineered for building and deploying large-scale recommender systems. It streamlines the entire pipeline, including data preprocessing (e.g., feature engineering, data transformation), model training (using GPU-accelerated frameworks), and inference optimizations tailored for recommendation tasks. Unlike general-purpose tools for natural language processing or data analytics, Merlin is optimized to handle the unique challenges of recommendation workloads, such as processing massive user-item interaction datasets and delivering personalized results efficiently.
(NVIDIA Merlin Documentation, Overview Section)

Question: 3

Which NVIDIA parallel computing platform and programming model allows developers to program in popular languages and express parallelism through extensions?

A. CUDA
B. CUML
C. CUGRAPH

Answer: A

Explanation:
CUDA (Compute Unified Device Architecture) is NVIDIA's foundational parallel computing platform and programming model. It enables developers to harness GPU parallelism by extending popular languages such as C, C++, and Fortran with parallelism-specific constructs (e.g., kernel launches, thread management). CUDA also provides bindings for languages like Python (via libraries like PyCUDA), making it versatile for a wide range of developers. In contrast, CUML and CUGRAPH are higher-level libraries built on CUDA for specific machine learning and graph analytics tasks, not general-purpose programming models.
(NVIDIA CUDA Programming Guide, Introduction)

Question: 4

Which of the following aspects have led to an increase in the adoption of AI? (Choose two.)

A. Moore's Law
B. Rule-based machine learning
C. High-powered GPUs
D. Large amounts of data

Answer: C, D

Explanation:
The surge in AI adoption is driven by two key enablers: high-powered GPUs and large amounts of data. High-powered GPUs provide the massive parallel compute capabilities necessary to train complex AI models, particularly deep neural networks, by processing numerous operations simultaneously, significantly reducing training times. Simultaneously, the availability of large datasets—spanning text, images, and other modalities—provides the raw material that modern AI algorithms, especially data-hungry deep learning models, require to learn patterns and make accurate predictions. While Moore's Law (the doubling of transistor counts) has historically aided computing, its impact has slowed, and rule-based machine learning has largely been supplanted by data-driven approaches.
(NVIDIA AI Infrastructure and Operations Study Guide, Section on AI Adoption Drivers)

Question: 5

In training and inference architecture requirements, what is the main difference between training and inference?

A. Training requires real-time processing, while inference requires large amounts of data.
B. Training requires large amounts of data, while inference requires real-time processing.
C. Training and inference both require large amounts of data.
D. Training and inference both require real-time processing.

Answer: B

Explanation:
The primary distinction between training and inference lies in their operational demands. Training necessitates large amounts of data to iteratively optimize model parameters, often involving extensive datasets processed in batches across multiple GPUs to achieve convergence. Inference, however, is designed for real-time or low-latency processing, where trained models are deployed to make predictions on new inputs with minimal delay, typically requiring less data volume but high responsiveness. This fundamental difference shapes their respective architectural designs and resource allocations.
(NVIDIA AI Infrastructure and Operations Study Guide, Section on Training vs. Inference Requirements)

Visit us at: https://www.examsboost.com/test/nca-aiio

Thank You for Trying Our Product

For More Information – Visit link below:

https://www.examsboost.com/

15 USD Discount Coupon Code:

G74JA8UF

FEATURES

- ✓ 90 Days Free Updates
- ✓ Money Back Pass Guarantee
- ✓ Instant Download or Email Attachment
- ✓ 24/7 Live Chat Support
- ✓ PDF file could be used at any Platform
- ✓ 50,000 Happy Customer

