NVIDIA-Certified Associate: AI Infrastructure and Operations Exam Study Guide

NVIDIA-Certified Associate: Infrastructure and Operations Exam Study Guide

Contents

This study guide provides an overview of each topic covered on the NVIDIA AI Infrastructure and Operations (NCA-AIIO) certification exam, recommended training, and suggested reading to prepare for the exam.

Information about NVIDIA certifications can be found **here**.

Job Description

The NCA-AIIO certification is an entry-level credential that validates the foundational concepts of AI computing related to infrastructure and operations. This exam is for IT professionals who are new to AI operations and infrastructure but who are required to understand and describe the different components and aspects of adopting AI in data center environments and on prem environments. This certification is appropriate for a wide variety of job roles from technical pre-sales to data center operations. Candidates must be knowledgeable about AI and ML in an enterprise, be able to describe essential tasks and requirements for AI operations. They will be able to articulate the various software and hardware one would need in an AI deployment. The candidate should know NVIDIA software, hardware and networking that would be used in these environments.

Job Responsibilities

1. Understand AI workloads and use cases.
2. Differentiate between AI and machine learning concepts.
3. Describe the key concepts of data center operations specifically for AI.
4. Have a basic understanding of networking requirements for AI environments.
5. Have a solid understanding of GPUs and DPUs and how they differ from CPU architecture.
6. Contribute to the operations of an AI data center in collaboration with a professional administrator.
7. Understand cluster orchestration and management, job scheduling, and monitoring essentials.
8. Describe how to use virtualized environments in AI workloads.
9. Understand NVIDIA software and hardware used in AI deployments.

Recommended Qualifications and Experience

1. Bachelor’s degree in computer science, software engineering, AI, or a related field
2. Experience in enterprise-level data centers and on-prem compute environments
3. Solid understanding of AI and machine learning concepts

Certification Topics and References

Essential AI knowledge: Exam Weight 38%

1.1	Describe the NVIDIA software stack used in an AI environment.
1.2	Compare and contrast training and inference architecture requirements and considerations.
1.3	Differentiate the concepts of AI, machine learning, and deep learning.
1.4	Explain the factors contributing to recent rapid improvements and adoption of AI.
1.5	Explain the key AI use cases and industries.
1.6	Explain the purpose and use case of various NVIDIA solutions.
1.7	Describe the software components related to the life cycle of AI development and deployment.
1.8	Compare and contrast GPU and CPU architectures.

Recommended Training (Optional)

Course reference: **AI Infrastructure and Operations Fundamentals**

> Unit 1: AI Transformation Across Industries
> Unit 2: Introduction to Artificial Intelligence
> Unit 4: Accelerating AI With GPUs
> Unit 5: AI Software Ecosystems
> Unit 7: Compute Platforms for AI
> Unit 14: Orchestration, MLOps, and Job Scheduling

Suggested Readings

- **NVIDIA® TensorRT™**, NVIDIA Developer
- **Deep Learning Training vs. Inference: Do You Know the Difference?**, by AI TutorMaster, Medium
- **Understanding Machine Learning Inference**, Run:ai
- **Tips on Scaling Storage for AI Training and Inferencing**, NVIDIA Technical Blog
- **What Is Machine Learning (ML)?**, IBM
- **Machine Learning: What It Is and Why It Matters**, SAS
- **What Are Large Language Models Used For?**, NVIDIA Blog
- **NVIDIA GPU Operator: Simplifying GPU Management in Kubernetes,** NVIDIA Technical Blog
- **CPU vs. GPU: What's the Difference?**, Intel

AI Infrastructure: Exam Weight 40%

Inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

2.1 Understand the process of extracting insights from large datasets using data mining, data visualization, and similar techniques.

2.2 Compare models using statistical performance metrics, such as loss functions or proportion of explained variance.

2.3 Conduct data analysis under the supervision of a senior team member.

2.4 Create graphs, charts, or other visualizations to convey the results of data analysis using specialized software.

2.5 Identify relationships and trends or any factors that could affect the results of research.

Recommended Training (Optional)

Course reference: **AI Infrastructure and Operations Fundamentals**

- Unit 4: Accelerating AI With GPUs
- Unit 7.1: Data Center Platform
- Unit 10: Energy-Efficient Computing
- Unit 7.4: Data Center Transformation With NVIDIA DPUs
- Unit 8: Networking for AI
- Unit 11: Energy-Efficient Computing
- Unit 12.4: AI in the Cloud Considerations

Suggested Readings

- **Offloading and Isolating Data Center Workloads With NVIDIA Bluefield® DPU**, NVIDIA Technical Blog
- **NVIDIA DGX SuperPOD™ Reference Architecture**, NVIDIA Docs Hub
- **Power Constraints and AI Workloads: The Hidden Challenges of High-Density Data Centers**, by Jeff Barber, NetZero News, LinkedIn
- **High-Density Servers: Maximizing Efficiency and Performance in Data Centers**, FS
- **Introduction to the NVIDIA DGX™ H100 System**, NVIDIA Docs Hub
- **InfiniBand Key Features**, NVIDIA Academy Vimeo
- **Modernizing GPU Network Data Transfer With NVIDIA NVSwitch™**, AMAX
- **Accelerating IO in the Modern Data Center: Network IO**, NVIDIA Technical Blog

AI Operations: Exam Weight 22%

3.1 Describe AI data center management and monitoring essentials.

3.2 Describe AI cluster orchestration and job scheduling essentials.

3.3 Articulate the key measures and criteria related to monitoring GPUs.

3.4 Identify the key considerations for virtualizing accelerated infrastructure.

Recommended Training (Optional)

Course reference: **AI Infrastructure and Operations Fundamentals**

- Unit 5: AI Software Ecosystem
- Unit 8: Networking for AI
- Unit 13: AI Data Center Management and Monitoring
- Unit 14: Orchestration, MLOps, and Job Scheduling

Suggested Reading List

- **Baseboard Management Controller**, NVIDIA Docs Hub
- **NVIDIA Base Command™**, NVIDIA
- **NVIDIA DCGM**, NVIDIA Developer
- **Out-of-Band Management Networks**, Management Networks for Dell EMC Networking Configuration Guide, Dell Technologies
- **Kubernetes Documentation**, Kubernetes
- **Let's Explore the Importance of Job Scheduling in a Cloud Environment**, by Changju Lee, Samsung SDS
- **What Is a Container?**, Docker
- **Slurm Workload Manager—Overview**, SchedMD
- **6 Reasons for Low GPU Utilization and How to Improve It**, Run:ai
- **NVIDIA Multi-Instance GPU (MIG)**, NVIDIA

Questions?

Contact us **here**.