DCAI

Implementing Cisco Data Center AI Infrastructure

40 horas | Data Center & Cloud | Cisco

Cisco Continuing Education Credits

38 CE Credits

INTRODUÇÃO

The Implementing Cisco Data Center AI Infrastructure (DCAI) training is designed to equip professionals with the skills to support, secure, and optimize AI workloads within modern data center environments. This comprehensive program delves into the unique characteristics of AI/ML applications, their influence on infrastructure design, and best practices for automated provisioning. Participants will gain in-depth knowledge of security considerations for AI deployments and master day-2 operations, including monitoring and advanced troubleshooting techniques such as log correlation and telemetry analysis. Through hands-on experience, including practical application with tools like Splunk, learners will be prepared to efficiently monitor, diagnose, and resolve issues in AI/ML-enabled data centers, ensuring optimal uptime and performance for critical organizational workloads.

This training prepares you for the 300-640 DCAI v1.0 exam. If passed, you earn the Cisco Certified Specialist - Data Center AI Infrastructure certification and satisfy the concentration exam requirement for the Cisco Certified Network Professional (CCNP) Data Center certification. This training also earns you 38 Continuing Education (CE) credits toward recertification. This training combines content from Operate and Troubleshoot AI Solutions on Cisco Infrastructure (DCAIAOT) and AI Solutions on Cisco Infrastructure Essentials (DCAIE) training.

This training will help you:

• Acquire comprehensive skills to support, secure, and optimize AI workloads within modern data center environments

• Understand the design, implementation, and advanced troubleshooting of AI infrastructure, including network challenges and specialized hardware

• Gain in-depth knowledge of AI/ML concepts, generative AI, and their practical application in network management and automation

• Apply hands-on techniques for monitoring, diagnosing, and resolving issues, leveraging tools like Splunk and utilizing AI for enhanced productivity in network operations

• Prepare for the 300-640 DCAI v1.0 exam

• Earn 38 CE credits toward recertification

Implementing Cisco Data Center AI Infrastructure (300-640 DCAI) v1.0 is a 90-minute exam associated with the Cisco Certified Specialist - Data Center AI Infrastructure certification and satisfies the concentration exam requirement for the CCNP Data Center certification.

This exam tests your knowledge of AI infrastructure, including:

• Design

• Implementation

• Monitoring

• Troubleshooting

OBJETIVO DO CURSO

• Describe key concepts in artificial intelligence, focusing on traditional AI, machine learning, and deep learning techniques and their applications
• Describe generative AI, its challenges, and future trends, while examining the nuances between traditional and modern AI methodologies
• Explain how AI enhances network management and security through intelligent automation, predictive analytics, and anomaly detection
• Describe the key concepts, architecture, and basic management principles of AI-ML clusters, as well as describe the process of acquiring, fine-tuning, optimizing and using pre-trained ML models
• Use the capabilities of Jupyter Lab and Generative AI to automate network operations, write Python code, and leverage AI models for enhanced productivity
• Describe the essential components and considerations for setting up robust AI infrastructure
• Evaluate and implement effective workload placement strategies and ensure interoperability within AI systems
• Explore compliance standards, policies, and governance frameworks relevant to AI systems
• Describe sustainable AI infrastructure practices, focusing on environmental and economic sustainability
• Guide AI infrastructure decisions to optimize efficiency and cost
• Describe key network challenges from the perspective of AI/ML application requirements
• Describe the role of optical and copper technologies in enabling AI/ML data center workloads
• Describe network connectivity models and network designs
• Describe important Layer 2 and Layer 3 protocols for AI and fog computing for Distributed AI processing
• Migrate AI workloads to dedicated AI network
• Explain the mechanisms and operations of RDMA and RoCE protocols
• Understand the architecture and features of high-performance Ethernet fabrics
• Explain the network mechanisms and QoS tools needed for building high-performance, lossless RoCE networks
• Describe ECN and PFC mechanisms, introduce Cisco Nexus Dashboard Insights for congestion monitoring, explore how different stages of AI/ML applications impact data center infrastructure, and vice versa
• Introduce the basic steps, challenges, and techniques regarding the data preparation process
• Use Cisco Nexus Dashboard Insights for monitoring AI/ML traffic flows
• Describe the importance of AI-specific hardware in reducing training times and supporting the advanced processing requirements of AI tasks
• Understand the compute hardware required to run AI/ML solutions
• Understand existing intelligence and AI/ML solutions
• Describe virtual infrastructure options and their considerations when deploying
• Explain data storage strategies, storage protocols, and software-defined storage
• Use NDFC to configure a fabric optimized for AI/ML workloads
• Use locally hosted GPT models with RAG for network engineering tasks

PÚBLICO-ALVO

• Network Designers
• Network Administrators
• Storage Administrators
• Network Engineers
• Systems Engineers
• Data Center Engineers
• Consulting Systems Engineers
• Technical Solutions Architects
• Cisco Integrators/Partners
• Field Engineers
• Server Administrators
• Network Managers
• Program Managers
• Project Managers

PRÉ-REQUISITOS

There are no prerequisites for this training. However, the knowledge and skills you are recommended to have before attending this training are:

• Cisco UCS compute architecture and operations
• Cisco Nexus switch portfolio and features
• Data Center core technologies

CONTEÚDO PROGRAMÁTICO

Fundamentals of AI

- Introduction to Artificial Intelligence
- Traditional AI
- Traditional AI Process Flow
- Traditional AI Challenges
- Modern Applications of Traditional AI
- Machine Learning vs. Deep Learning
- ML vs. DL Techniques and Methodologies
- ML vs. DL Applications and Use Cases

Generative AI

- Generative AI
- Generative Adversarial Frameworks
- GenAI Use Cases
- Generative AI Inference Challenges
- GenAI Challenges and Limitations
- GenAI Bias and Fairness
- GenAI Resource Optimization
- Generative AI vs. Traditional AI
- GenAI vs. Traditional AI Data Requirements
- Future Trends in AI
- AI Language Models
- LLMs vs. SLMs

AI Use Cases

- Analytics
- Network Optimization
- Network Automation and Self-Healing Networks
- Capacity Planning and Forecasting
- Cybersecurity
- Predictive Risk Management
- Threat Detection
- Incident Response
- Collaboration and Communication
- Internet of Things (IoT)

AI-ML Clusters and Models

- AI-ML Compute Clusters
- AI-ML Cluster Use Cases
- Custom AI Models–Process
- Custom AI Models–Tools
- Prebuilt AI Model Optimization
- Pre-Trained AI Models
- AI Model Parameters
- Service Placements – On-Premises vs. Cloud vs. Distributed

AI Toolset—Jupyter Notebook

AI Infrastructure

- Traditional AI Infrastructure
- Modern AI Infrastructure
- Cisco Nexus HyperFabric AI Clusters

AI Workloads Placement and Interoperability
- Workload Mobility
- Multi-Cloud Implementation
- Vendor Lock-In Risks
- Vendor Lock-In Mitigation

AI Policies
- Data Sovereignty
- Compliance, Governance, and Regulations

AI Sustainability
- Green AI vs. Red AI
- Cost Optimization
- AI Accelerators
- Power and Cooling

AI Infrastructure Design
- Project Description
- Your Role
- Activity 1: AI Workload Type
- Activity 2: Cloud vs. On-Prem
- Activity 3: The Choice of Network
- Activity 4: Choice of Platform and Sustainability
- Activity 5: Power Considerations

Key Network Challenges and Requirements for AI Workloads
- Bandwidth and Latency Considerations
- Scalability Considerations
- Redundancy and Resiliency Considerations
- Visibility
- Nonblocking Lossless Fabric
- Congestion Management Considerations

AI Transport
- Optical and Copper Cabling
- Organizing Data Center Cabling
- Ethernet Cables
- InfiniBand Cables
- Ethernet Connectivity
- InfiniBand Connectivity
- Hybrid Connectivity

Connectivity Models
- Network Types: Isolated vs. Purpose-Built Network
- Network Architectures: Two-Tier vs. Three-Tier Hierarchical Model
- Networking Considerations: Single-Site vs. Multi-Site Network Architecture

AI Network

- Layer 2 Protocols
- Layer 3 Protocols
- Scalability Considerations for Deploying AI Workloads
- Fog Computing for AI Distributed Processing

Architecture Migration to AI/ML Network

- Project Description
- Your Role
- Activity 1: Starting Small
- Activity 2: Going Beyond One Server
- Activity 3: Traffic Considerations

Application-Level Protocols

- RDMA Fundamentals
- RDMA Architecture
- RDMA Operations
- RDMA over Converged Ethernet > NEW title RoCE/RoCEv2

High-Throughput Converged Fabrics

- InfiniBand-to-Ethernet Transition
- Cisco Nexus 9000 Series Switches Portfolio

Building Lossless Fabrics

- Traditional QoS Toolset
- Enhanced Transmission Selection
- Intelligent Buffer Management on Cisco Nexus 9000 Series Switches
- AFD with ETRAP
- Dynamic Packet Prioritization
- Data Center Bridging Exchange
- Lossless Ethernet Fabric Using RoCEv2
- Advanced Congestion Management with AFD

Congestion Visibility

- Explicit Congestion Notification
- Priority Flow Control
- Congestion Visibility in AI/ML Cluster Networks Using Cisco Nexus Dashboard Insights
- Pipeline Considerations

Data Preparation for AI

- Data Processing Workflow Overview
- Data Processing Workflow Phases

AI/ML Workload Data Performance

AI-Enabling Hardware

- CPUs, GPUs, and DPUs
- GPU Overview
- NVIDIA GPUs for AI/ML
- Intel GPUs for AI/ML
- DPU Overview

• SmartNIC Overview
• Cisco Nexus SmartNIC Family
• NVIDIA BlueField SuperNIC

Compute Resources
• Compute Hardware Overview
• Intel Xeon Scalable Processor Family Overview
• Cisco UCS C-Series Rack Servers
• Cisco UCS X-Series Modular System
• Mapping AI/ML Workloads to Cisco UCS Servers
• GPU Sharing
• Compute Resources Sharing
• Total Cost of Ownership
• AI/ML Clustering

Compute Resource Solutions
• Cisco Hyperconverged Infrastructure Solutions Overview
• Cisco Hyperconverged Solution Components
• FlashStack Data Center
• Nutanix GPT-in-a-Box
• Run:ai on Cisco UCS

Virtual Resources
• Virtual Infrastructure
• Device Virtualization
• Server Virtualization Defined
• Virtual Machine
• Hypervisor
• Container Engine
• Storage Virtualization
• Virtual Networks
• Virtual Infrastructure Deployment Options
• Hyperconverged Infrastructure
• HCI and Virtual Infrastructure Deployment

Storage Resources
• Data Storage Strategy
• Fibre Channel and FCoE
• NVMe and NVMe over Fabrics
• Software-Defined Storage

Setting Up AI Cluster

Deploy and Use Open Source GPT Models for RAG

AI Infrastructure Operations and Monitoring
• The Need for AI Infrastructure Monitoring
• Monitoring Compute
• Monitoring Storage
• Monitoring the Runtime Layer

• Monitoring AI Fabrics
• The Need for AI Infrastructure Lifecycle Management
• Compute Lifecycle Upgrades
• Fabric Lifecycle Upgrades

Troubleshooting AI Infrastructure
• Log Correlation for AI Applications
• Telemetry Analysis for AI Workloads
• Hands-On Telemetry for AI Workloads
• Timing Protocols

Troubleshoot Common Issues in AI/ML Fabric
• Overview of Splunk Enterprise and Splunk Cloud
• Data Ingestion Methods
• Splunk Applications
• Basics of Splunk SPL
• Troubleshoot Common Issues in AI/ML Fabric