# DCAI (IMPLEMENTING CISCO DATA CENTER AI INFRASTRUCTURE) 1.0

## Objetivo

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

## PÃºblico Alvo

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

## PrÃ©-Requisitos

There are no prerequisites for this training. However, the knowledge and skills you are recommended to have before attending this training are: â¢ Cisco UCS compute architecture and operations â¢ Cisco Nexus switch

portfolio and features â□¢ Data Center core technologies

## Carga HorÃiria

40 horas (5 dias).

## ConteÃºdo ProgramÃitico

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