AI Soultions on Cisco Infrastructure Essentials

Durée: 4 Jours **Réf de cours: DCAIE** **Version: 1.0**

Résumé:

The **AI Solutions on Cisco Infrastructure Essentials (DCAIE)** course covers the essentials of deploying, migrating, and operating AI solutions on Cisco data center infrastructure. You'll be introduced to key AI workloads and elements, as well as foundational architecture, design, and security practices critical to successful delivery and maintenance of AI solutions on Cisco infrastructure.

This course is worth 34 Continuing Education (CE) credits toward recertification.

Public visé:

This course is for anyone invloved in the implementation, maintenance and troubleshooting of a Cisco Data Center

Objectifs pédagogiques:

- **After completing this course you should be able to:**
- 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
- 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 computer hardware required to run AI/ML solutions
- Understand existing 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

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

Pré-requis:

Attendees should have an understanding of:

- Cisco UCS compute architecture and operations
- Cisco Nexus switch portfolio and features
- Data Center core technologies
- DCCOR - Implementing and Operating Cisco Data Center Core Technologies
- DCNX - Implementing Cisco NX-OS Switches and Fabrics in the Data Center

Test et certification

Recommended as preparation for the following exam:

- There is no exam currently aligned to this course

Contenu:

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
- Generative AI 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 Mastery - Jupyter Notebook

- AI Toolset-Jupyter Notebook

AI Infrastructure

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

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
- Starting Small
- Going Beyond One Server
- Traffic Considerations

Application-Level Protocols

- RDMA Fundamentals
- RDMA Architecture
- RDMA Operations
- RDMA over Converged Ethernet

High-Throughput Converged Fabrics

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

Building Lossless Fabrics

- Traditional QoS Toolset

AI/ML Workload Data Performance

- 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
- GPU Sharing
- Compute Resources Sharing
- Total Cost of Ownership
- AI/ML Clustering

Compute Resources 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 Stirage Strategy
- Fibre Channel and FCoE
- NVMe and NVMe over Fabrics
- Software-Defined Storage

Setting Up AI Cluster

- Setting Up AI Cluster

Deploy and Use Open Source GPT Models

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
- AI Workload Type
- Cloud vs. On-Prem
- The Choice of Network
- Choice of Platform and Sustainability
- Power Considerations

- 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

for RAG

- Deploy and Use Open Source GPTModels for RAG

Labs

- Discovery Lab 1: AI Toolset—Jupyter Not
- Discovery Lab 2: AI/ML Workload Data Performance
- Discovery Lab 3: Setting Up AI Cluster
- Discovery Lab 4: Deploy and Use Open Source GPT Models for RAG

Autres moyens pédagogiques et de suivi:

• Compétence du formateur : Les experts qui animent la formation sont des spécialistes des matières abordées et ont au minimum cinq ans d'expérience d'animation. Nos équipes ont validé à la fois leurs connaissances techniques (certifications le cas échéant) ainsi que leur compétence pédagogique.
• Suivi d'exécution : Une feuille d'émargement par demi-journée de présence est signée par tous les participants et le formateur.
• En fin de formation, le participant est invité à s'auto-évaluer sur l'atteinte des objectifs énoncés, et à répondre à un questionnaire de satisfaction qui sera ensuite étudié par nos équipes pédagogiques en vue de maintenir et d'améliorer la qualité de nos prestations.

Délais d'inscription :

• Vous pouvez vous inscrire sur l'une de nos sessions planifiées en inter-entreprises jusqu'à 5 jours ouvrés avant le début de la formation sous réserve de disponibilité de places et de labs le cas échéant.
• Votre place sera confirmée à la réception d'un devis ou """"booking form""""" signé. Vous recevrez ensuite la convocation et les modalités d'accès en présentiel ou distanciel.
• Attention, si cette formation est éligible au Compte Personnel de Formation, vous devrez respecter un délai minimum et non négociable fixé à 11 jours ouvrés avant le début de la session pour vous inscrire via moncompteformation.gouv.fr.

Accueil des bénéficiaires :

• En cas de handicap : plus d'info sur globalknowledge.fr/handicap
• Le Règlement intérieur est disponible sur globalknowledge.fr/reglement