Mise en oeuvre de l'infrastructure IA pour datacenters Cisco (DCAI)

Durée: 5 Jours **Réf de cours: DCAI** **Version: 1.0** **Méthodes d'apprentissage: Blended Learning**

Résumé:

La formation « Mise en œuvre de l'infrastructure IA pour datacenters Cisco (DCAI) » est conçue pour doter les professionnels des compétences nécessaires pour prendre en charge, sécuriser et optimiser les charges de travail liées à l'IA au sein d'environnements de centres de données modernes.

Ce programme complet explore les caractéristiques propres aux applications d'IA et d'apprentissage automatique (ML), leur influence sur la conception de l'infrastructure, ainsi que les meilleures pratiques en matière de provisionnement automatisé.
Les participants acquerront une connaissance approfondie des enjeux de sécurité liés aux déploiements d'IA et maîtriseront les opérations de maintenance courantes (day-2), notamment la surveillance et les techniques avancées de dépannage telles que la corrélation des journaux et l'analyse télémétrique.
Grâce à une expérience pratique, incluant l'utilisation d'outils tels que Splunk, les participants seront en mesure de surveiller, diagnostiquer et résoudre efficacement les problèmes dans les centres de données équipés d'IA et d'apprentissage automatique (ML), garantissant ainsi une disponibilité et des performances optimales pour les workloads critiques de l'entreprise.

Cette formation vous prépare à l'examen 300-640 DCAI v1.0. Si vous le réussissez, vous obtiendrez la certification « Cisco Certified Specialist - Data Center AI Infrastructure » et remplirez les conditions requises pour l'examen de spécialisation de la certification « Cisco Certified Network Professional (CCNP) Data Center ».

Cette formation donne droit à 38 crédits de formation continue (CE).

Mis à jour 13/10/2026
L'offre Blended Live de Global Knowledge combine une formation animée par un instructeur expert et la flexibilité d'un apprentissage à la demande. Elle complète notre offre standard et vise à satisfaire des nouveaux besoins, notamment en termes de disponibilités ou de rythme, pour développer aussi efficacement les compétences.
Blended Live combine : Un apprentissage à la demande Le suivi de sessions synchrones à distance selon un calendrier pré-défini L'échange avec les mentors Des labs à réaliser Des conseils pour compléter l'apprentissage dans un format structuré

Public visé:

Cette formation s'adresse à toute personne souhaitant : Acquérir des compétences approfondies pour prendre en charge, sécuriser et optimiser les charges de travail liées à l'IA au sein d'environnements de datacenters modernes Comprendre la conception, la mise en œuvre et le dépannage avancé des infrastructures d'IA, y compris les défis liés au réseau et au hardware spécialisé Acquérir une connaissance poussée des concepts d'IA/ML, de l'IA générative et de leur application pratique dans la gestion et l'automatisation des réseaux Appliquer des techniques pratiques de surveillance, de diagnostic et de résolution des problèmes, en utilisant des outils tels que Splunk et en exploitant l'IA pour améliorer la productivité des opérations réseau Se préparer à l'examen 300-640 DCAI v1.0

Objectifs pédagogiques:

- À l'issue de ce cours, les participants seront capables de :
- Décrire les concepts clés de l'intelligence artificielle, en mettant l'accent sur l'IA traditionnelle, le machine learning et les techniques de deep learning, ainsi que leurs applications
- Décrire l'IA générative, ses défis et ses tendances futures, tout en examinant les nuances entre les méthodologies traditionnelles et modernes de l'IA
- Expliquer comment l'IA améliore la gestion et la sécurité des réseaux grâce à l'automatisation intelligente, à l'analyse prédictive et à la détection des anomalies
- Décrire les concepts clés, l'architecture et les principes de
- Décrire les principaux protocoles de couche 2 et de couche 3 utilisés pour l'IA et le « fog computing » pour le traitement distribué de l'IA
- Migrer les workloads IA vers un réseau dédié à l'IA
- Expliquer les mécanismes et le fonctionnement des protocoles RDMA et RoCE
- Comprendre l'architecture et les fonctionnalités des structures Ethernet hautes performances
- Expliquer les mécanismes réseau et les outils de qualité de service (QoS) nécessaires à la mise en place de réseaux RoCE hautes performances et sans perte

gestion fondamentaux des clusters IA-ML, ainsi que le processus d'acquisition, de réglage fin, d'optimisation et d'utilisation de modèles ML pré-entraînés

- Utiliser les fonctionnalités de Jupyter Lab et de l'IA générative pour automatiser les opérations réseau, écrire du code Python et exploiter des modèles d'IA afin d'améliorer la productivité
- Décrire les composants essentiels et les considérations à prendre en compte pour mettre en place une infrastructure d'IA robuste
- Évaluer et mettre en œuvre des stratégies efficaces de répartition des charges de travail et garantir l'interopérabilité au sein des systèmes d'IA
- Explorer les normes de conformité, les stratégies et les référentiels de gouvernance applicables aux systèmes d'IA
- Décrire les pratiques durables en matière d'infrastructure d'IA, en mettant l'accent sur la durabilité environnementale et économique
- Oriente les décisions relatives à l'infrastructure d'IA afin d'optimiser l'efficacité et les coûts
- Décrire les principaux défis réseau du point de vue des exigences des applications d'IA et d'apprentissage automatique
- Décrire le rôle des technologies optiques et cuivre dans la prise en charge des charges de travail des datacenters d'IA et d'apprentissage automatique
- Décrire les modèles de connectivité réseau et les né
- Décrire les modèles de connectivité réseau et les architectures réseau
- Décrire les mécanismes ECN et PFC, présenter Cisco Nexus Dashboard Insights pour la surveillance de la saturation, explorer l'impact des différentes étapes des applications d'IA/ML sur l'infrastructure des centres de données, et inversement
- Présenter les étapes de base, les défis et les techniques liés au processus de préparation des données
- Utiliser Cisco Nexus Dashboard Insights pour surveiller les flux de trafic IA/ML
- Décrire l'importance du hardware spécifique à l'IA pour réduire les temps d'entraînement et répondre aux exigences de traitement avancées des tâches d'IA
- Comprendre la puissance de calcul requise pour exécuter des solutions d'IA/ML
- Comprendre les solutions existantes en matière d'IA et d'IA/ML
- Décrire les options d'infrastructure virtuelle et les considérations à prendre en compte lors de leur déploiement
- Expliquer les stratégies de stockage des données, les protocoles de stockage et le stockage défini par logiciel
- Utiliser NDFC pour configurer une structure optimisée pour les charges de travail d'IA/ML
- Utiliser des modèles GPT hébergés localement avec RAG pour les tâches d'ingénierie réseau

Pré-requis:

Les participants doivent posséder les compétences et connaissances suivantes :

- Architecture et fonctionnement des ressources de calcul Cisco UCS
- Gamme et fonctionnalités des commutateurs Cisco Nexus
- Technologies de base des datacenters
- DCCOR - Implementing and Operating Cisco Data Center Core Technologies
- DCNDE - Cisco Data Center Nexus Dashboard Essentials
- DCNX - Implementing Cisco NX-OS Switches and Fabrics in the Data Center

Test et certification

Recommandé comme préparation à l'examen suivant :

- 300-640 - DCAI - Mise en œuvre de l'infrastructure d'IA pour centres de données Cisco

Après cette formation, nous vous conseillons le(s) module(s) suivant(s):

Pas de recommandation spécifique

Contenu:

Principes fondamentaux de l'IA

- Introduction à l'intelligence artificielle
- L'IA traditionnelle
- Déroulement du flux de travail de l'IA traditionnelle
- Défis liés à l'IA traditionnelle
- Applications modernes de l'IA traditionnelle
- Machine Learning vs Deep Learning
- Techniques et méthodologies de Machine Learning et de Deep Learning
- Applications et cas d'utilisation du ML et du DL

IA générative

- IA générative
- Cadres génératifs antagonistes
- Cas d'utilisation de l'IA générative
- Défis liés à l'inférence en IA générative
- Défis et limites de l'IA générative
- Biais et équité en IA générative
- Optimisation des ressources en IA générative
- IA générative vs IA traditionnelle
- Exigences en matière de données : IA générative vs IA traditionnelle
- Tendances futures en matière d'IA
- Modèles linguistiques d'IA

LLM vs SLM

Cas d'usage IA

- Analyse de données
- Optimisation du réseau
- Automatisation du réseau et réseaux à réparation automatique
- Planification et prévision des capacités
- Cybersécurité
- Gestion prédictive des risques
- Détection des menaces
- Réponse aux incidents
- Collaboration et communication
- Internet des objets (IoT)

IA-ML Clusters et Modèles

- Clusters de traitement IA-ML
- Cas d'utilisation des clusters IA-ML
- Modèles d'IA personnalisés – Processus
- Modèles d'IA personnalisés – Outils
- Optimisation des modèles d'IA prédéfinis
- Modèles d'IA pré-entraînés
- Paramètres des modèles d'IA
- Déploiement des services : sur site, dans le cloud ou en environnement distribué

Maîtrise des outils d'IA - Jupyter Notebook

Infrastructure IA

Transport IA

- Câblage optique et en cuivre
- Organisation du câblage du centre de données
- Câbles Ethernet
- Câbles InfiniBand
- Connectivité Ethernet
- Connectivité InfiniBand
- Connectivité hybride
- Modèles de connectivité

Types de réseaux : réseau isolé ou réseau dédié

- Architectures réseau : modèle hiérarchique à deux niveaux ou à trois niveaux
- Considérations relatives à la mise en réseau : architecture réseau mono-site ou multi-site

Réseau IA

- Protocoles Layer 2
- Protocoles Layer 3
- Considérations sur la scalabilité pour le déploiement des workloads IA
- Le « fog computing » au service du traitement distribué de l'IA

Migration de l'architecture vers un réseau IA/ML

- Description du projet
- Votre rôle

Principes fondamentaux du RDMA

- Architecture RDMA
- Opérations RDMA

RDMA sur Ethernet convergent

Structures convergentes à haut débit

- Transition d'InfiniBand vers Ethernet
- Gamme de commutateurs Cisco Nexus série 9000

Mise en place de structures sans perte (Lossless)

- Ensemble d'outils QoS traditionnels
- Sélection de transmission améliorée
- Gestion intelligente de la mémoire tampon sur les commutateurs Cisco Nexus de la série 9000
- AFD avec ETRAP
- Priorisation dynamique des paquets
- Data Center Bridging Exchange
- Structure Ethernet sans perte utilisant

Ressources de calcul

- Présentation du hardware
- Présentation de la gamme de processeurs Intel Xeon Scalable
- Serveurs en rack Cisco UCS série C
- Système modulaire Cisco UCS série X
- Partage des GPU
- Partage des ressources de calcul
- Coût total de possession (CTO)
- Clustering IA/ML

Solutions de ressources de calcul

- Présentation des solutions d'infrastructure hyperconvergée Cisco
- Composants des solutions hyperconvergées Cisco
- Centre de données FlashStack
- Nutanix GPT-in-a-Box

Exécuter :ai sur Cisco UCS

Ressources virtuelles

- Infrastructure virtuelle
- Virtualisation des périphériques
- Définition de la virtualisation des serveurs
- Machine virtuelle
- Hyperviseur
- Container Engine
- Virtualisation du stockage
- Réseaux virtuels
- Options de déploiement de l'infrastructure virtuelle
- Infrastructure hyperconvergée

Déploiement de l'infrastructure hyperconvergée (HCI) et de l'infrastructure virtuelle

Ressources de stockage

- Stratégie de stockage des données
- Fibre Channel et FCoE
- NVMe et NVMe over Fabrics

Stockage défini par logiciel

Configuration d'un cluster d'IA

- Utilisez NDFC pour configurer une structure optimisée pour les charges de travail d'IA/ML.

Déploiement et utilisation de modèles GPT open source pour RAG

- Utilisez des modèles GPT hébergés localement avec RAG pour les tâches

- Infrastructure IA traditionnelle
- Infrastructure IA moderne
- Clusters IA Cisco Nexus HyperFabric

Localisation des workloads IA et interopérabilité

- Mobilité des workloads
- Mise en œuvre multi-cloud
- Risques de dépendance vis-à-vis d'un fournisseur
- Mesures d'atténuation de la dépendance vis-à-vis d'un fournisseur

Politiques IA

- Souveraineté des données
- Compliance, Gouvernance, et Règlementation

Durabilité de l'IA

- IA « verte » vs IA « rouge »
- Optimisation des coûts
- Accélérateurs d'IA
- Alimentation énergétique et refroidissement

Conception d'une infrastructure d'IA

- Description du projet
- Votre rôle

Principaux défis et exigences réseau pour les workloads IA

- Considérations sur la bande passante et la latence
- Considérations sur l'évolutivité
- Considérations sur la redondance et la résilience
- Visibilité
- Matrice sans perte ni blocage

Considérations sur la prévention de la saturation

RoCEv2

- Gestion avancée de la saturation avec AFD

Visibilité de la saturation

- Notification explicite de saturation
- Contrôle de flux prioritaire
- Visibilité de la saturation dans les réseaux de clusters IA/ML grâce à Cisco Nexus Dashboard Insights
- Considérations relatives au pipeline

Préparation des données pour l'IA

- Présentation du workflow de traitement des données

Phases du workflow de traitement des données

Performances des données des charges de travail IA/ML

- Utiliser Cisco Nexus Dashboard Insights pour surveiller les flux de trafic IA/ML

Hardware compatible avec l'IA

- Processeurs (CPU), processeurs graphiques (GPU) et processeurs de traitement (DPU)
- Présentation des GPU
- GPU NVIDIA pour l'IA/ML
- GPU Intel pour l'IA/ML
- Présentation des DPU
- Présentation des SmartNIC
- Gamme Cisco Nexus SmartNIC
- NVIDIA BlueField SuperNIC

d'ingénierie réseau.

Exploitation et surveillance de l'infrastructure IA

- La nécessité de surveiller l'infrastructure IA
- Surveillance des ressources de calcul
- Surveillance du stockage
- Surveillance de la couche d'exécution
- Surveillance des structures IA
- La nécessité de gérer le cycle de vie de l'infrastructure IA
- Mises à niveau du cycle de vie des ressources de calcul
- Mises à niveau du cycle de vie des structures

Dépannage de l'infrastructure d'IA

- Corrélation des logs pour les applications d'IA
- Analyse de la télémétrie pour les Workloads d'IA
- Pratique de la télémétrie pour les Workloads d'IA
- Protocoles de synchronisation
- Dépannage des problèmes courants dans AI/ML Fabric

Présentation de Splunk Enterprise et Splunk Cloud

- Méthodes d'ingestion des données
- Applications Splunk
- Notions de base sur le langage SPL de Splunk

Méthodes pédagogiques :

Support de cours officiel fourni aux participants

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