+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 20/08/2026

S'inscrire

Formation llm-d

2 jours (14 heures)

Présentation

llm-d'est une plateforme cloud native conçue pour industrialiser l inférence distribuée de modèles de langage sur Kubernetes. En combinant routage intelligent, serveurs de modèles et optimisation des ressources matérielles, elle aide les équipes à fournir des services IA performants, fiables et maîtrisés en production.

Notre formation llm-d vous permettra de comprendre les mécanismes nécessaires au déploiement et à l'exploitation de modèles de langage à grande échelle. Vous apprendrez à structurer une architecture d'inférence reposant sur Kubernetes, à intégrer un serveur de modèles tel que vLLM et à exposer des API compatibles avec les usages applicatifs modernes.

Vous apprendrez à configurer les composants fondamentaux de llm-d, notamment le Router, les InferencePools et les serveurs de modèles. Vous mettrez en oeuvre les principes de répartition de charge, de gestion des requêtes et d'exploitation des accélérateurs afin d'obtenir un service d'inférence cohérent avec vos contraintes de capacité et de latence.

La formation aborde également les optimisations essentielles à l'inférence de LLM en production, telles que le routage conscient du cache KV, la séparation prefill et decode, l'autoscaling et le suivi des indicateurs de performance. Vous saurez identifier les compromis entre débit, temps de réponse, consommation GPU et coût d'exploitation.

À l'issue de la formation, vous serez en mesure de déployer, tester et diagnostiquer une plateforme llm-d sur un environnement Kubernetes. Vous disposerez des bonnes pratiques pour préparer des déploiements d'inférence distribuée robustes, observables et adaptés aux charges IA réelles.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Comprendre l'architecture et le positionnement de llm-d dans une plateforme d'inférence distribuée.
- Déployer un serveur de modèles et un InferencePool sur Kubernetes.
- Configurer le llm-d Router pour répartir efficacement les requêtes d'inférence.
- Exploiter le cache KV, le routage intelligent et la séparation prefill et decode.
- Mesurer les performances, analyser la latence et préparer le passage en production.

Public visé

- Ingénieur MLOps
- Data Engineer
- Ingénieur DevOps
- Administrateur Kubernetes
- Architecte cloud

Pré-requis

- Expérience de l'administration ou du déploiement d'applications sur Kubernetes.
- Maîtrise des commandes courantes de Linux et de kubectl.
- Connaissances des conteneurs Docker et des packages Helm.
- Notions sur les API HTTP, les modèles de langage et l'inférence IA.

Pré-requis techniques

- Poste sous Linux, macOS ou Windows 11 avec WSL2.
- Installation de Git, kubectl, Helm 3 et d'un éditeur tel que Visual Studio Code.
- Accès à un cluster Kubernetes 1.29 ou version ultérieure, avec droits de création de ressources dans un namespace dédié.
- Accès à des ressources de calcul GPU ou à un environnement de démonstration préparé, ainsi qu'à un compte Hugging Face si le téléchargement de modèles est nécessaire.
- Connexion Internet stable permettant l'accès aux registres de conteneurs et aux dépôts de code.

Programme de notre formation llm-d

[Jour 1 - Matin]

Fondamentaux de l inférence distribuée avec llm-d

- Panorama de llm-d, cas d usage et positionnement dans une plateforme d IA générative.
- Principes d inférence de LLM, tokens, débit, TTFT, latence inter-token et concurrence.
- Rôle de Kubernetes dans l orchestration des charges d inférence et des accélérateurs.
- Présentation des composants Model Server, InferencePool et llm-d Router.
- Comprendre les interactions entre llm-d, vLLM, SGLang, Gateway API et API compatibles OpenAI.

- Atelier pratique : exploration d un environnement Kubernetes, vérification des outils et lecture des ressources d une architecture llm-d préparée.

[Jour 1 - Après-midi]

Déployer un premier service d inférence

- Préparation du namespace, des secrets et des prérequis nécessaires au chargement d un modèle.
- Installation du llm-d Router en mode standalone avec Helm.
- Déploiement d un serveur vLLM et compréhension des paramètres de modèle, de GPU et de réplication.
- Configuration d un InferencePool pour associer le routage aux endpoints d inférence.
- Validation de l exposition du service et émission de requêtes via une API OpenAI compatible.
- Atelier pratique : déployer un service llm-d complet, appeler un modèle et vérifier l état des pods, services et endpoints.

[Jour 2 - Matin]

Optimiser le routage et les performances

- Identifier les facteurs qui influencent la latence, le débit et l occupation des accélérateurs.
- Comprendre le fonctionnement du cache KV et son impact sur les requêtes partageant un même préfixe.
- Configurer le routage conscient du cache pour favoriser la localité des requêtes.
- Étudier la séparation prefill et decode pour adapter l architecture aux charges conversationnelles.
- Comparer les stratégies de répartition, de réplication et de montée en charge selon les objectifs de service.
- Atelier pratique : modifier une configuration de routage, exécuter des requêtes avec préfixes communs et interpréter les effets observés.

[Jour 2 - Après-midi]

Exploitation, observabilité et préparation à la production

- Suivre les métriques clés de performance d inférence, saturation, erreurs, débit et temps de réponse.
- Diagnostiquer les incidents courants liés aux pods, aux modèles, aux ressources GPU et au réseau.
- Mettre en place des principes d autoscaling et de contrôle de flux adaptés aux charges variables.
- Structurer une démarche de tests de charge et de benchmark pour valider les objectifs de capacité.
- Appliquer les bonnes pratiques de sécurité, de gestion des secrets, de mises à jour et de gouvernance des coûts.
- Atelier pratique : réaliser un diagnostic de performance sur un déploiement llm-d, proposer des optimisations et formaliser un plan de mise en production.

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.