

Mis à jour le 26/10/2026

S'inscrire

Formation Kueue

3 jours (21 heures)

Présentation

Kueue est un système natif de Kubernetes permettant de gérer les files de tâches et les quotas de ressources pour les traitements batch. Avec les files d'attente et le partage équitable des quotas, Kueue facilite l'entraînement simultané de plusieurs modèles de Machine Learning sur un cluster de GPU partagé tout en limitant la contention.

Notre formation Kueue vous permettra de maîtriser les mécanismes essentiels pour organiser vos entraînements et améliorer l'utilisation des ressources disponibles entre plusieurs équipes.

Vous apprendrez à configurer les ClusterQueues et les LocalQueues pour structurer vos files, à définir les profils matériels avec les ResourceFlavors et à adapter les quotas aux besoins de vos modèles. Vous mettrez en œuvre les priorités, les cohortes et les politiques de préemption.

À l'issue de la formation, vous serez en mesure de piloter plusieurs entraînements concurrents, de vérifier leurs conditions d'admission et de réduire les goulots d'étranglement liés aux ressources. Vous distinguerez la réservation des quotas du placement des pods par Kubernetes.

Cette formation aborde également les entraînements distribués avec PyTorch, la supervision et le diagnostic afin de fiabiliser vos traitements sur une infrastructure mutualisée.

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

Objectifs

- Installer Kueue et configurer les files nécessaires à la gestion des entraînements
- Définir les profils matériels et les quotas de CPU, de mémoire et de GPU

- Organiser l'entraînement simultané de plusieurs modèles sur un cluster de GPU partagé
- Appliquer des politiques de priorité, de partage équitable et de préemption entre équipes
- Intégrer des entraînements PyTorch simples et distribués aux files de traitement
- Diagnostiquer les blocages et ajuster les configurations à partir des métriques

Public visé

- Ingénieur MLOps
- Ingénieur Machine Learning
- Ingénieur plateforme
- Administrateur Kubernetes
- Data Scientist

Pré-requis

- Pratique de Kubernetes pour manipuler les pods, les Jobs et les namespaces
- Maîtrise des commandes courantes de kubectl et de la lecture de manifestes YAML pour déployer des ressources
- Connaissance des conteneurs et des demandes et limites de ressources dans Kubernetes pour dimensionner les traitements
- Notions de programmation en Python et d'entraînement de modèles avec PyTorch pour comprendre les exemples fournis

Pré-requis techniques

- Ordinateur avec 8 Go de RAM minimum, navigateur récent et terminal sous Linux ou macOS ou sous Windows avec WSL2 pour accéder aux exercices
- Installation de kubectl dans une version compatible avec le cluster, de Helm et d'un éditeur tel que Visual Studio Code pour manipuler les manifestes
- Accès à un cluster pédagogique Kubernetes compatible avec la version de Kueue retenue, disposant d'au moins deux GPU et du NVIDIA Device Plugin avec les pilotes adaptés
- Compte disposant des droits nécessaires pour installer les contrôleurs et gérer les ressources du cluster, connexion Internet stable autorisant les accès au cluster et aux registres de conteneurs
- Environnement pédagogique comprenant Kubeflow Training Operator pour les exercices distribués, des images PyTorch compatibles avec les GPU et une supervision Prometheus et Grafana accessible

Programme de notre formation Kueue

[Jour 1 - Matin]

Comprendre et installer Kueue

- Enjeux des GPU mutualisés : contention, ressources inutilisées et entraînements concurrents

- Rôle de Kueue : mise en file, réservation des quotas et admission des traitements
- Complémentarité avec le scheduler Kubernetes et les contrôleurs responsables du cycle de vie des Jobs
- Architecture et objets fondamentaux : Workload, LocalQueue, ClusterQueue et ResourceFlavor
- Installation avec Helm, vérification du contrôleur et découverte des ressources disponibles
- Atelier pratique : Installer Kueue sur le cluster pédagogique, créer une première file et observer l'admission d'un Job simple.

[Jour 1 - Après-midi]

Configurer les quotas et les profils GPU

- Organisation des équipes avec les namespaces et rattachement des LocalQueues aux ClusterQueues
- Définition des ResourceFlavors et correspondance avec les labels et les caractéristiques des nœuds GPU
- Dimensionnement des quotas nominaux de CPU, de mémoire et de GPU par profil matériel
- Demandes de ressources des Jobs et distinction entre quota réservé et capacité physique disponible
- Comparaison de StrictFIFO et BestEffortFIFO face à une tâche volumineuse en attente
- Atelier pratique : Configurer deux files et leurs quotas GPU, soumettre plusieurs Jobs et vérifier leur attente puis leur démarrage à la libération des ressources.

[Jour 2 - Matin]

Partager les ressources entre équipes

- Mutualisation des quotas inutilisés entre ClusterQueues au sein d'une cohorte
- Encadrement des échanges avec les limites de prêt et d'emprunt de quota
- Principes de Fair Sharing et pondération du partage des ressources empruntables
- Hiérarchisation des traitements avec WorkloadPriorityClass et distinction avec la priorité des pods
- Politiques de préemption, récupération des quotas et conséquences sur les entraînements interrompus
- Atelier pratique : Mettre deux équipes en concurrence, observer l'emprunt des GPU disponibles et tester une politique de priorité et de préemption.

[Jour 2 - Après-midi]

Entraîner plusieurs modèles simultanément

- Préparation de Jobs d'entraînement PyTorch à partir de scripts et d'images fournis
- Association des entraînements aux LocalQueues et déclaration des besoins en CPU, mémoire et GPU
- Exécution simultanée de plusieurs modèles lorsque les quotas et la capacité physique le permettent

- Distinction entre concurrence sur plusieurs GPU et partage d'un même GPU, qui nécessite des mécanismes complémentaires
- Suivi des résultats et préparation de points de contrôle applicatifs pour limiter les pertes lors d'une interruption
- Atelier pratique : Lancer simultanément deux entraînements de modèles distincts sur les GPU partagés, mettre un troisième en attente et vérifier son admission après libération des ressources.

[Jour 3 - Matin]

Fiabiliser les entraînements distribués

- Différences entre plusieurs entraînements indépendants et un entraînement distribué sur plusieurs workers
- Intégration de PyTorchJob avec Kubeflow Training Operator et activation de l'intégration correspondante dans Kueue
- Calcul des besoins globaux du Workload à partir des ressources demandées par les workers
- Configuration de waitForPodsReady, délais de démarrage et remise en file des traitements bloqués
- Introduction au Topology-Aware Scheduling pour tenir compte de la proximité réseau des workers
- Atelier pratique : Exécuter un entraînement distribué, provoquer un défaut de disponibilité d'un worker et analyser la remise en file avec waitForPodsReady.

[Jour 3 - Après-midi]

Superviser et optimiser le cluster partagé

- Lecture des conditions des Workloads, des événements et des journaux pour expliquer les décisions d'admission
- Suivi des files, des temps d'attente et des quotas réservés avec Prometheus et Grafana
- Distinction entre réservation de quota et utilisation réelle des GPU mesurée par une instrumentation dédiée
- Diagnostic des goulots d'étranglement : quotas, capacité GPU, CPU, mémoire, stockage et placement des pods
- Bonnes pratiques d'exploitation : droits RBAC, périmètre des Jobs gérés et validation des changements de configuration
- Atelier pratique : Piloter une campagne concurrente de plusieurs modèles, identifier un blocage et ajuster les quotas ou les politiques de file en comparant les métriques avant et après modification.

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.