Cas d'utilisation

NetApp Solutions

NetApp
April 12, 2024

This PDF was generated from https://docs.netapp.com/fr-fr/netapp-solutions/ai/ai-protopia-overview.html on April 12, 2024. Always check docs.netapp.com for the latest.

Sommaire

Cas d'utilisation . . . 1
IA responsable et inférence confidentielle : NetApp ai avec Protopia image transformation . . . 1
Analyse des sentiments avec NetApp ai . . . 27
Formation distribuée dans Azure - prévision de taux par clic . . . 45
Tr-4896: Formation distribuée dans Azure: Détection de voie - conception de la solution . . . 69
Tr-4841 : système d'exploitation pour l'IA dans le cloud hybride avec mise en cache des données . . . 98
Inférence d'IA en périphérie - NetApp avec Lenovo ThinkSystem - conception de la solution . . . 121
WP-7328 : la stratégie d'IA de NetApp basée sur NVIDIA Jarvis . . . 145
Tr-4858 : solution d'orchestration NetApp avec Run:ai . . . 166
Tr-4799 : architecture de référence NetApp ONTAP ai pour les workloads de conduite autonome . . . 187
Tr-4811 : Architecture de référence NetApp ONTAP ai pour le secteur de la santé : imagerie diagnostique - conception de la solution . . . 188
Tr-4807 : architecture de référence NetApp ONTAP ai pour les workloads de services financiers - conception de la solution . . . 188
Valeur de l'IA générative et de l'NetApp . . . 189
Tr-4785 : déploiement de l'IA avec NetApp E-Series et BeeGFS . . . 200
NVA-1150-DESIGN : guide de design Quantum StorNext avec les systèmes NetApp E-Series . . . 200
NVA-1150-DEPLOY : guide de déploiement Quantum StorNext avec les systèmes NetApp E-Series . . . 201

Cas d’utilisation

IA responsable et inférence confidentielle : NetApp ai avec Protopia image transformation

Tr-4928 : IA responsable et inférence confidentielle : NetApp ai avec Protopia image et la transformation des données

Sathish Thyagarajan, Michael Oglesby, NetApp Byung Hoon Ahn, Jennifer Cwagenberg, Protopia

Les interprétations visuelles font désormais partie intégrante de la communication avec l’émergence de la capture d’images et du traitement d’images. L’intelligence artificielle (IA) dans le traitement d’images numériques offre de nouvelles opportunités commerciales, notamment dans le domaine médical du cancer et de l’identification de maladies, dans l’analyse visuelle géospatiale pour l’étude des risques environnementaux, la reconnaissance de modèles, dans le traitement vidéo pour la lutte contre le crime, etc. Toutefois, cette opportunité est également liée à des responsabilités extraordinaires.

Plus les entreprises sont aux prises avec l’IA, plus elles acceptent de risques liés à la confidentialité et à la sécurité des données, ainsi qu’aux questions juridiques, éthiques et réglementaires. Une approche IA responsable permet aux entreprises et aux administrations de gagner la confiance et de la gouvernance nécessaires à l’IA à grande échelle. Ce document décrit une solution d’inférence d’IA validée par NetApp dans trois scénarios différents en utilisant les technologies de gestion de données NetApp avec le logiciel Protopia Data obfuscation pour privatiser les données sensibles et réduire les risques et les préoccupations d’ordre éthique.

Des millions d’images sont générées chaque jour avec différents appareils numériques, aussi bien par des particuliers que des entités commerciales. Avec l’explosion massive des données et des charges de travail de calcul qui en résulte, les entreprises se tournent vers les plateformes de cloud computing pour améliorer leur évolutivité et leur efficacité. Dans le même temps, la confidentialité des données sensibles contenues dans les données d’images est sujet à la transmission vers un cloud public. Le manque d’assurances de sécurité et de protection de la vie privée devient le principal obstacle au déploiement de systèmes d’IA de traitement des images.

De plus, il y a le "droit d’effacement" Par le RGPD, le droit d’une personne de demander à une entreprise d’effacer toutes ses données personnelles. Il y a également le "Protection de la vie privée", qui établit un code de pratiques équitables en matière d’information. Les images numériques telles que les photographies peuvent constituer des données personnelles selon le RGPD, qui régit la collecte, le traitement et l’effacement des données. Le non-respect de cette règle est un non-respect du RGPD, qui peut entraîner de lourdes sanctions pour non-respect des conformités qui peuvent être sérieusement préjudiciables à des entreprises. Les principes de confidentialité constituent l’une des piliers de l’implémentation d’une IA responsable qui garantit l’équité dans les prévisions des modèles de machine learning (ML) et de deep learning (DL) et réduit les risques associés à la violation de la confidentialité ou à la conformité réglementaire.

Ce document présente une solution de conception validée dans trois scénarios différents, avec et sans confusion d’images concernant la préservation de la confidentialité et le déploiement d’une solution d’IA responsable :

- **Scénario 1.** inférence à la demande dans le bloc-notes Jupyter.
- **Scénario 2.** inférence par lot sur Kubernetes.

- **Scénario 3.** serveur d'inférence NVIDIA Triton.

Pour cette solution, nous utilisons le jeu de données de détection de visage et le banc d'essai (FDDB), un jeu de zones de face conçues pour étudier le problème de la détection de visage sans contrainte, combiné au cadre d'apprentissage machine PyTorch pour la mise en œuvre de FaceBoxes. Ce jeu de données contient les annotations pour 5171 faces dans un ensemble de 2845 images de différentes résolutions. Ce rapport technique présente également certains domaines de solutions et plusieurs utilisations pertinentes, recueillies par les clients et par les ingénieurs de terrain de NetApp, dans les situations où la solution est applicable.

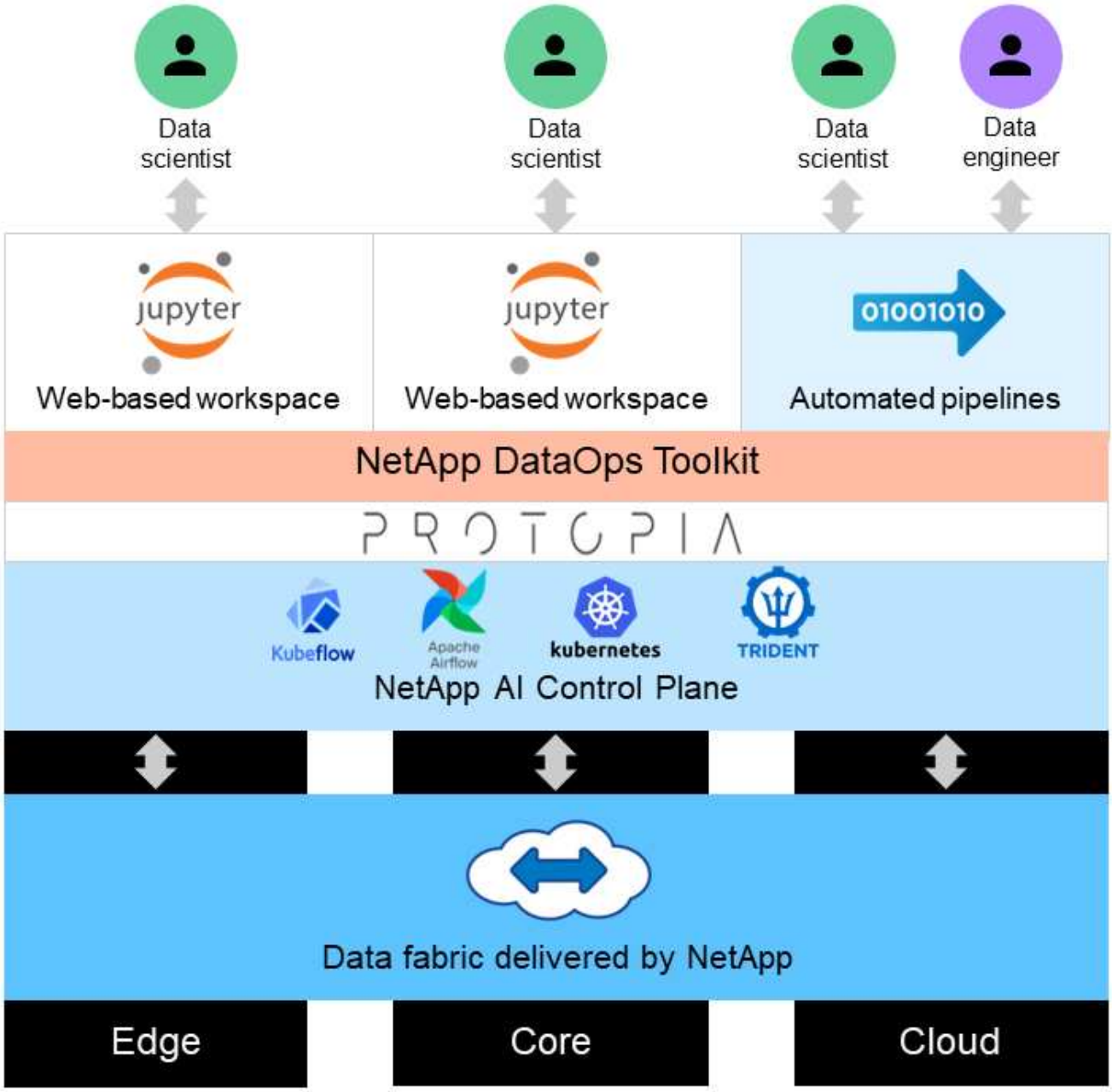
Public visé

Ce rapport technique est destiné aux publics suivants :

- Des dirigeants d'entreprise et des architectes d'entreprise qui souhaitent concevoir et déployer une IA responsable et résoudre des problèmes de protection des données et de confidentialité concernant le traitement d'images faciales dans les espaces publics.
- Les data Scientists, les ingénieurs en données, les chercheurs d'IA et de machine learning (ML), et les développeurs de systèmes d'IA/ML qui cherchent à protéger et à préserver la confidentialité.
- Les architectes d'entreprise qui conçoivent des solutions de fusion des données pour des modèles et des applications d'IA/DE ML conformes aux normes réglementaires telles que le RGPD, la loi CCPA ou la loi sur la confidentialité des données du ministère de la Défense (DoD) et des administrations.
- Les data Scientists et les ingénieurs d'IA recherchent des méthodes efficaces pour déployer le deep learning (DL) et les modèles d'inférence d'IA/AM/AP qui protègent les informations sensibles.
- Des gestionnaires de terminaux et des administrateurs de serveurs en périphérie sont chargés du déploiement et de la gestion des modèles d'inférence à la périphérie.

Architecture de la solution

Cette solution est conçue pour traiter des workloads d'IA d'inférence en temps réel et par lots sur des datasets volumineux, grâce à la puissance de traitement des GPU avec des processeurs classiques. Cette validation démontre l'inférence à la confidentialité pour L'AM et une gestion optimale des données requises pour les entreprises qui souhaitent déployer des déploiements d'IA responsables. Cette solution fournit une architecture adaptée à une plateforme Kubernetes à un ou plusieurs nœuds pour la périphérie et le cloud computing interconnectée avec NetApp ONTAP ai dans le cœur, le kit NetApp DataOps Toolkit et le logiciel Protopia obfuscation via des interfaces Jupyter Lab et CLI. La figure suivante présente l'architecture logique de Data Fabric optimisée par NetApp avec DataOps Toolkit et Protopia.



Le logiciel d'obfuscation Protopia s'exécute en toute transparence au-dessus du kit d'outils NetApp DataOps et transforme les données avant de quitter le serveur de stockage.

Domaines de solutions

Le traitement numérique de l'image présente de nombreux avantages, permettant ainsi à de nombreuses entreprises d'exploiter au mieux les données associées aux représentations visuelles. Cette solution NetApp et Protopia propose une conception unique d'inférence d'IA pour protéger et privatiser les données d'IA/ML tout au long du cycle de vie DU ML/DL. Il permet aux clients de conserver la propriété des données sensibles, d'utiliser des modèles de déploiement de cloud public ou hybride pour l'évolutivité et l'efficacité en allégeant les problèmes de confidentialité et en déployant l'inférence d'IA à la périphérie.

Intelligence environnementale

Il existe de nombreuses façons dont les industries peuvent tirer parti de l'analyse géospatiale dans les domaines des dangers environnementaux. Les gouvernements et le ministère des travaux publics peuvent obtenir des renseignements exploitables sur la santé publique et les conditions météorologiques afin de mieux conseiller le public lors d'une pandémie ou d'une catastrophe naturelle comme les incendies de forêt. Par exemple, vous pouvez identifier un patient COVID-positif dans les espaces publics, tels que les aéroports ou les hôpitaux, sans compromettre la confidentialité de la personne concernée et avertir les autorités respectives et le public à proximité de ces mesures de sécurité.

Les dispositifs de bord sont wearables

Dans l'armée et sur les champs de bataille, vous pouvez utiliser l'inférence d'IA sur la périphérie comme dispositifs portables pour suivre la santé des soldats, surveiller le comportement des conducteurs et alerter les autorités sur la sécurité et les risques associés à l'approche des véhicules militaires tout en préservant et en protégeant la vie privée des soldats. L'avenir de l'armée passe en haute technologie avec l'Internet des objets du champ de bataille (IoBT) et l'Internet des objets militaires (IoMT) pour des engins de combat portables qui aident les soldats à identifier leurs ennemis et à réaliser de meilleures performances dans les combats grâce à l'informatique de pointe rapide. La protection et la préservation des données visuelles collectées depuis des appareils en périphérie tels que les drones et les équipements portables sont essentielles pour que les pirates et l'ennemi soit à la hauteur.

Opérations d'évacuation sans combattant

Les opérations d'évacuation des non-combattants (ONE) sont menées par le DoD pour aider à l'évacuation des citoyens et des ressortissants américains, du personnel civil DoD et des personnes désignées (pays hôte (HN) et des ressortissants de pays tiers (AMT) dont la vie est en danger pour un refuge sûr approprié. Les contrôles administratifs en place utilisent en grande partie les processus de dépistage manuel des évacuées. Cependant, l'exactitude, la sécurité et la rapidité de l'identification des données évacuées, du suivi des évacués et du filtrage des menaces peuvent être améliorées via des outils d'IA/DE ML hautement automatisés combinés aux technologies d'obfuscation vidéo d'IA et DE ML.

Santé et recherche biomédicale

Le traitement des images est utilisé pour diagnostiquer des pathologies à des fins de planification chirurgicale à partir d'images 3D issues d'une tomographie informatisée (TI) ou d'une imagerie par résonance magnétique (IRM). Les règles de confidentialité HIPAA régissent la façon dont les données doivent être collectées, traitées et effacées par les entreprises pour toutes les données personnelles et images numériques, telles que les photographies. Pour que les données puissent être partagées conformément aux réglementations HIPAA Safe Harbor, les images photographiques en plein visage et les images comparables doivent être supprimées. Les techniques automatisées telles que la déidentification ou le décapage du crâne utilisées pour masquer les caractéristiques faciales d'un individu provenant d'images TDM/RM structurelles sont devenues un élément essentiel du processus de partage des données pour les établissements de recherche biomédicale.

Migration vers le cloud de l'analytique d'IA/DE ML

Les clients grands comptes ont traditionnellement entraîné et déployé des modèles d'IA/ML sur site. Pour des raisons d'évolutivité et d'efficacité, ces clients étendent leurs fonctions d'IA/ML aux déploiements dans un cloud public, hybride ou multicloud. Mais elles sont liées aux données qui peuvent être exposées à d'autres infrastructures. Les solutions NetApp répondent à l'ensemble des menaces de cybersécurité requises "protection des données" Enfin, lorsqu'elle est associée à la transformation des données Protopia, l'évaluation de la sécurité a réduit les risques associés à la migration de workloads d'IA/DE ML d'images vers le cloud.

Pour en savoir plus sur les champs d'application du calcul en périphérie et de l'inférence d'IA dans d'autres secteurs, consultez la page "Tr-4886 inférence d'IA à la périphérie" Et le blog NetApp ai, "L'intelligence contre

la confidentialité".

Présentation de la technologie

Cette section présente les différents composants techniques nécessaires pour compléter cette solution.

Protopia

Protopia ai offre une solution logicielle unique et discrète pour l'inférence confidentielle sur le marché actuel. La solution Protopia offre une protection sans précédent pour les services d'inférence en minimisant l'exposition des informations sensibles. L'IA n'est alimentée que les informations qui figurent dans les enregistrements de données. Il est primordial de mener à bien la tâche à bien et rien de plus. La plupart des tâches d'inférence n'utilisent pas toutes les informations qui existent dans chaque enregistrement de données. Que votre IA consomme des images, de la voix, des vidéos ou même des données tabulaires structurées, Protopia fournit uniquement ce dont le service d'inférence a besoin. La technologie de base brevetée fait appel à un bruit mathématiquement choisi pour transformer les données avec beaucoup de place et pour brouiller les informations qui ne sont pas nécessaires à un service DE ML donné. Cette solution ne masque pas les données ; elle modifie plutôt la représentation des données en utilisant un bruit aléatoire adapté.

La solution Protopia formule le problème de la modification de la représentation en tant que méthode d'optimisation des perturbations à base de gradient qui conserve toujours les informations pertinentes dans l'espace de la fonction d'entrée par rapport à la fonctionnalité du modèle. Ce processus de détection s'exécute comme un pass d'ajustement à la fin de l'entraînement du modèle DE ML. Une fois que le pass génère automatiquement un ensemble de distributions de probabilité, une transformation des données à faible surcharge applique des échantillons de bruit de ces distributions aux données, l'obfusant avant de les transmettre au modèle pour l'inférence.

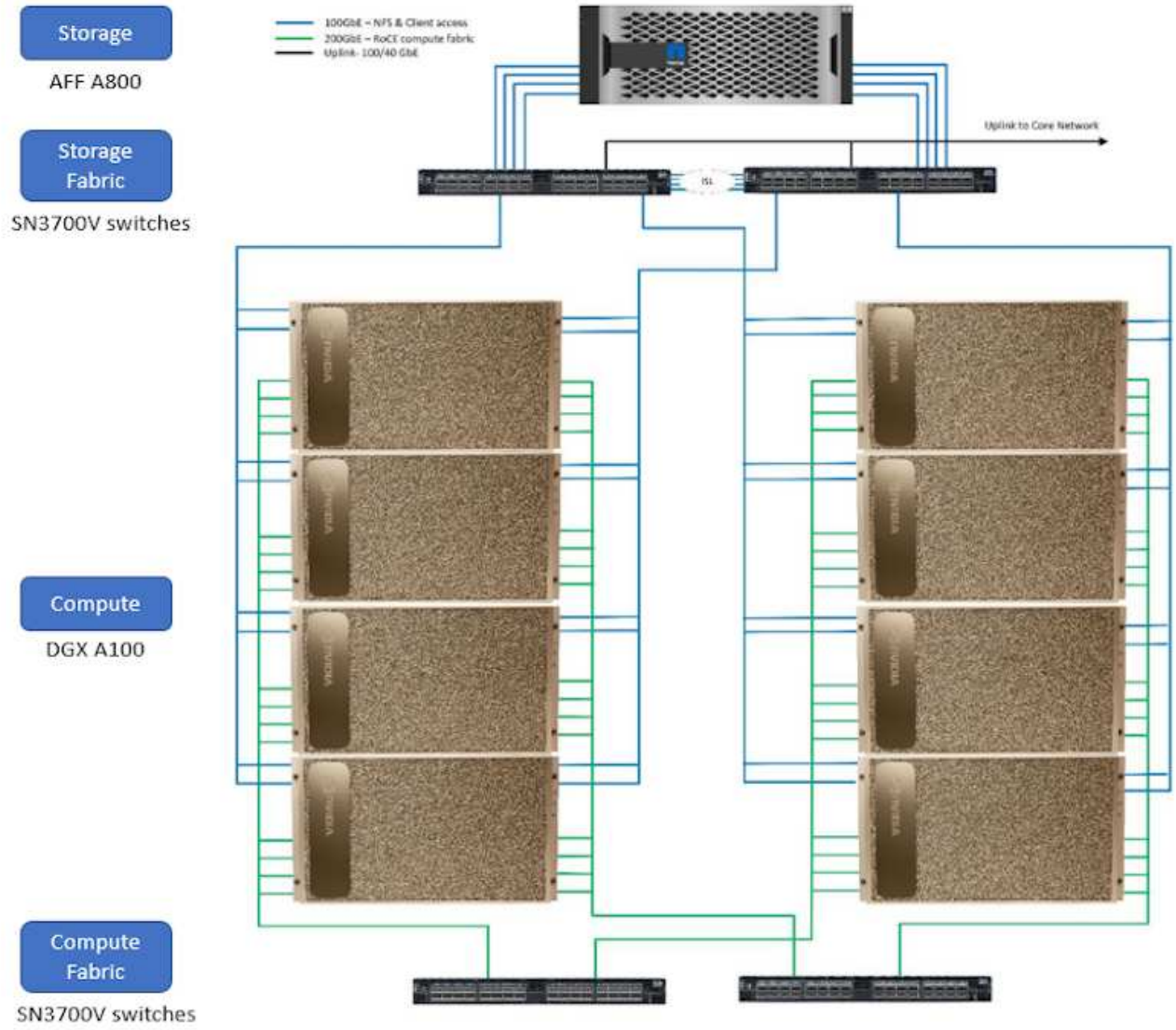
NetApp ONTAP ai

L'architecture de référence NetApp ONTAP ai, optimisée par les systèmes DGX A100 et les systèmes de stockage connectés au cloud de NetApp, a été développée et vérifiée par NetApp et NVIDIA. Et présente plusieurs avantages pour les services IT :

- Simplifie la conception
- Offre une évolutivité indépendante des ressources de calcul et de stockage
- Possibilité de faire évoluer de manière fluide une infrastructure initiale de petite taille
- Propose plusieurs options de stockage pour répondre à des exigences variées de coûts et de performance

ONTAP ai intègre étroitement les systèmes DGX A100 et de stockage NetApp AFF A800 avec une connectivité réseau optimale. ONTAP ai simplifie les déploiements d'IA en éliminant la complexité et les approximations. Les clients peuvent commencer avec une petite infrastructure, puis évoluer sans interrompre l'activité, tout en gérant intelligemment leurs données de la périphérie au cœur, et jusqu'au cloud, et inversement.

La figure suivante montre plusieurs variations de la gamme de solutions ONTAP ai avec les systèmes DGX A100. Les performances du système AFF A800 sont vérifiées avec un maximum de huit systèmes DGX A100. En ajoutant des paires de contrôleurs de stockage au cluster ONTAP, l'architecture peut évoluer vers plusieurs racks pour prendre en charge de nombreux systèmes DGX A100 et des pétaoctets de capacité de stockage avec des performances linéaires. Cette approche flexible permet d'ajuster les ratios de calcul/stockage de manière indépendante, en fonction de la taille des modèles d'apprentissage profond utilisés et des metrics de performance requis.



Pour plus d'informations sur ONTAP ai, rendez-vous sur "NVA-1153 : NetApp ONTAP ai avec des systèmes NVIDIA DGX A100 et des switchs Ethernet Mellanox Spectrum."

NetApp ONTAP

ONTAP 9.11, la dernière génération de logiciel de gestion du stockage de NetApp, permet aux entreprises de moderniser l'infrastructure et de passer à un data Center prêt pour le cloud. Avec des capacités de gestion des données à la pointe du secteur, ONTAP permet de gérer et de protéger les données avec un seul ensemble d'outils, quel que soit leur emplacement. Vous pouvez aussi déplacer vos données librement partout où elles sont nécessaires : la périphérie, le cœur ou le cloud. ONTAP 9.11 comprend de nombreuses fonctionnalités qui simplifient la gestion des données, accélèrent et protègent les données stratégiques, et permettent d'utiliser des fonctionnalités d'infrastructure nouvelle génération dans toutes les architectures de cloud hybride.

Kit NetApp DataOps

Le kit NetApp DataOps est une bibliothèque Python qui facilite le clonage quasi instantané d'un volume de données ou d'un espace de travail JupyterLab par exemple, des développeurs, des data Scientists, et des ingénieurs DevOps, Et de réaliser quasi-instantanément des snapshots d'un volume de données ou d'un espace de travail JupyterLab pour la traçabilité ou l'établissement de base. Cette bibliothèque Python peut fonctionner comme un utilitaire de ligne de commande ou une bibliothèque de fonctions que vous pouvez

importer dans n'importe quel programme Python ou ordinateur portable Jupyter.

Serveur d'inférence NVIDIA Triton

NVIDIA Triton Inférence Server est un logiciel d'inférence open source qui permet de standardiser le déploiement et l'exécution des modèles afin d'assurer l'IA rapide et évolutive en production. Triton Inférence Server rationalise l'inférence d'IA en permettant aux équipes de déployer, d'exécuter et de faire évoluer les modèles d'IA entraînés à partir de n'importe quelle infrastructure basée sur le processeur ou le processeur graphique. Triton Inférence Server prend en charge toutes les structures majeures, telles que TensorFlow, NVIDIA TensorRT, PyTorch, MXNet, OpenVINO, etc. Triton s'intègre à Kubernetes pour l'orchestration et l'évolutivité que vous pouvez utiliser dans toutes les principales plateformes d'IA et Kubernetes de cloud public. Il est également intégré à de nombreuses solutions logicielles MLOps.

PyTorch

"PyTorch" Est un framework DE ML open source. Il s'agit d'une bibliothèque Tensor optimisée pour le deep learning, qui utilise les GPU et les processeurs. Le paquet PyTorch contient des structures de données pour les tenseurs multidimensionnels qui fournissent de nombreux utilitaires pour la sérialisation efficace des tenseurs entre autres utilitaires utiles. Il dispose également d'un homologue CUDA qui vous permet d'exécuter vos calculs Tensor sur un GPU NVIDIA offrant des capacités de calcul. Dans cette validation, nous utilisons la bibliothèque OpenCV-Python (cv2) pour valider notre modèle tout en tirant parti des concepts de vision informatique les plus intuitifs de Python.

Gestion simplifiée

La gestion des données est cruciale pour les opérations IT et les data Scientists, de sorte que les ressources appropriées sont utilisées pour les applications d'IA et pour l'entraînement des datasets d'IA/DE ML. Les informations supplémentaires suivantes sur les technologies NetApp ne sont pas incluses dans cette validation, mais elles peuvent être pertinentes en fonction de votre déploiement.

Le logiciel de gestion des données ONTAP comprend les fonctionnalités suivantes pour rationaliser et simplifier les opérations et réduire le coût total d'exploitation :

- Compaction des données à la volée et déduplication étendue La compaction des données réduit le gaspillage d'espace à l'intérieur des blocs de stockage, et la déduplication augmente considérablement la capacité effective. Cela s'applique aux données stockées localement et à leur placement dans le cloud.
- Qualité de service (AQoS) minimale, maximale et adaptative. Les contrôles granulaires de la qualité de service (QoS) permettent de maintenir les niveaux de performance des applications stratégiques dans des environnements hautement partagés.
- NetApp FabricPool Tiering automatique des données inactives vers des options de stockage de cloud public et privé, notamment Amazon Web Services (AWS), Azure et la solution de stockage NetApp StorageGRID. Pour plus d'informations sur FabricPool, voir "Tr-4598 : meilleures pratiques de FabricPool".

Accélération et protection des données

ONTAP offre des niveaux supérieurs de performances et de protection des données et étend ces fonctionnalités aux méthodes suivantes :

- Des performances élevées et une faible latence. ONTAP offre le débit le plus élevé possible à la latence la plus faible possible.
- Protection des données. ONTAP fournit des fonctionnalités de protection des données intégrées avec une gestion commune sur toutes les plateformes.
- NetApp Volume Encryption (NVE). ONTAP offre un chiffrement natif au niveau du volume avec un support

de gestion des clés interne et externe.

- Colocation et authentification multifacteur. ONTAP permet le partage des ressources d'infrastructure avec les plus hauts niveaux de sécurité.

Une infrastructure pérenne

ONTAP permet de répondre aux besoins métier en constante évolution grâce aux fonctionnalités suivantes :

- Évolutivité transparente et opérations non disruptives. ONTAP prend en charge l'ajout non disruptif de capacité aux contrôleurs et l'évolution scale-out des clusters. Les clients peuvent effectuer la mise à niveau vers les technologies les plus récentes, telles que NVMe et FC 32 Gb, sans migration des données ni panne coûteuse.
- Connexion cloud. ONTAP est le logiciel de gestion de stockage le plus connecté au cloud, avec des options de stockage SDS (ONTAP Select) et des instances natives de cloud (NetApp Cloud Volumes Service) dans tous les clouds publics.
- Intégration avec les applications émergentes ONTAP propose des services de données d'entreprise pour les plateformes et applications nouvelle génération, telles que les véhicules autonomes, les Smart cities et Industry 4.0, en utilisant la même infrastructure prenant en charge les applications d'entreprise existantes.

NetApp Astra Control

La gamme NetApp Astra propose des services de stockage et de gestion des données respectueuse des applications pour les applications Kubernetes sur site et dans le cloud public, optimisés par les technologies NetApp de stockage et de gestion des données. Il vous permet de sauvegarder facilement les applications Kubernetes, de migrer des données vers un autre cluster et de créer instantanément des clones d'applications de travail. Si vous devez gérer les applications Kubernetes s'exécutant dans un cloud public, consultez la documentation de "Service Astra Control". Astra Control Service est un service géré par NetApp qui permet la gestion des données intégrant la cohérence applicative des clusters Kubernetes dans Google Kubernetes Engine (GKE) et Azure Kubernetes Service (AKS).

NetApp Astra Trident

Astra "Trident" À partir de NetApp, est un orchestrateur de stockage dynamique open source pour Docker et Kubernetes qui simplifie la création, la gestion et la consommation du stockage persistant. Trident, une application Kubernetes native, s'exécute directement dans un cluster Kubernetes. Trident permet de déployer de manière transparente des images de conteneur d'apprentissage profond sur un système de stockage NetApp et offre une expérience haute performance pour les déploiements de conteneurs d'IA. Les utilisateurs de Kubernetes (développeurs DE ML, data Scientists, etc.) peuvent créer, gérer et automatiser l'orchestration et le clonage pour exploiter des fonctionnalités avancées de gestion des données optimisées par la technologie NetApp.

Copie et synchronisation NetApp BlueXP

"Copie et synchronisation BlueXP" Est un service NetApp qui permet une synchronisation sûre et rapide des données. Que vous ayez besoin de transférer des fichiers entre des partages de fichiers NFS ou SMB sur site, NetApp StorageGRID, NetApp ONTAP S3, NetApp Cloud Volumes Service, Azure NetApp Files, Amazon simple Storage Service (Amazon S3), Amazon Elastic File System (Amazon EFS), Azure Blob, Google Cloud Storage, ou IBM Cloud Object Storage, BlueXP Copy and Sync déplace les fichiers là où vous en avez besoin rapidement et en toute sécurité. Une fois vos données transférées, elles peuvent être utilisées à la source et à la cible. BlueXP Copy et Syncc synchronisent en continu les données en fonction de votre planification prédéfinie, en déplaçant uniquement les données modifiées. Le temps et l'argent consacrés à la réplication des données sont ainsi réduits. BlueXP Copy and Sync est un outil SaaS extrêmement simple à configurer et à utiliser. Les transferts de données déclenchés par la copie et la synchronisation BlueXP sont effectués par les

courtiers de données. Vous pouvez déployer des courtiers de données BlueXP Copy and Sync dans AWS, Azure, Google Cloud Platform ou sur site.

Classification de NetApp BlueXP

Reposant sur de puissants algorithmes d'IA, "Classification de NetApp BlueXP" permet d'automatiser le contrôle et la gouvernance des données dans l'ensemble de votre environnement de données. Vous pouvez facilement identifier les économies réalisables, identifier les problèmes de conformité et de confidentialité, et trouver des opportunités d'optimisation. Le tableau de bord de classification BlueXP vous donne les informations nécessaires pour identifier les données en double afin d'éliminer la redondance, mapper les données personnelles, non personnelles et sensibles, et activer les alertes pour les données sensibles et les anomalies.

Plan de test et de validation

Pour cette conception de la solution, les trois scénarios suivants ont été validés :

- Une tâche d'inférence, avec et sans projet d'obfuscation Protopia, dans un espace de travail JupyterLab orchestré à l'aide du kit d'outils NetApp DataOps pour Kubernetes.
- Une tâche d'inférence par lot, avec et sans objetProtopia obfuscation, sur Kubernetes avec un volume de données orchestré à l'aide du kit d'outils NetApp DataOps pour Kubernetes.
- Une tâche d'inférence avec une instance de serveur NVIDIA Triton Inférence orchestrée avec le kit d'outils NetApp DataOps pour Kubernetes. Nous avons appliqué l'obfuscation Protopia à l'image avant d'appeler l'API d'inférence Triton pour simuler l'exigence commune selon laquelle les données transmises sur le réseau doivent être brouillées. Ce workflow s'applique aux cas où les données sont collectées dans une zone de confiance mais doivent être transférées en dehors de cette zone de confiance pour l'inférence. Sans l'obfuscation Protopia, il n'est pas possible d'implémenter ce type de flux de travail sans données sensibles quittant la zone de confiance.

Configuration du test

Le tableau suivant présente l'environnement de validation de la conception de la solution.

Composant	Version
Kubernetes	1.21.6
Pilote NetApp Astra Trident CSI	22.01.0
Kit NetApp DataOps pour Kubernetes	2.3.0
Serveur d'inférence NVIDIA Triton	21.11-py3

Procédure de test

Cette section décrit les tâches nécessaires pour terminer la validation.

Prérequis

Pour exécuter les tâches décrites dans cette section, vous devez avoir accès à un hôte Linux ou MacOS avec les outils suivants installés et configurés :

- Kubectl (configuré pour l'accès à un cluster Kubernetes existant)

    - Des instructions d'installation et de configuration sont disponibles "ici".
- Kit NetApp DataOps pour Kubernetes
    - Des instructions d'installation sont disponibles "ici".

Scénario 1 : inférence à la demande dans JupyterLab

1. Créez un namespace Kubernetes pour les workloads d'inférence d'IA/DE ML.

```
$ kubectl create namespace inference
namespace/inference created
```

2. NetApp DataOps Toolkit permet de provisionner un volume persistant pour le stockage des données sur lesquelles vous souhaitez procéder à l'inférence.

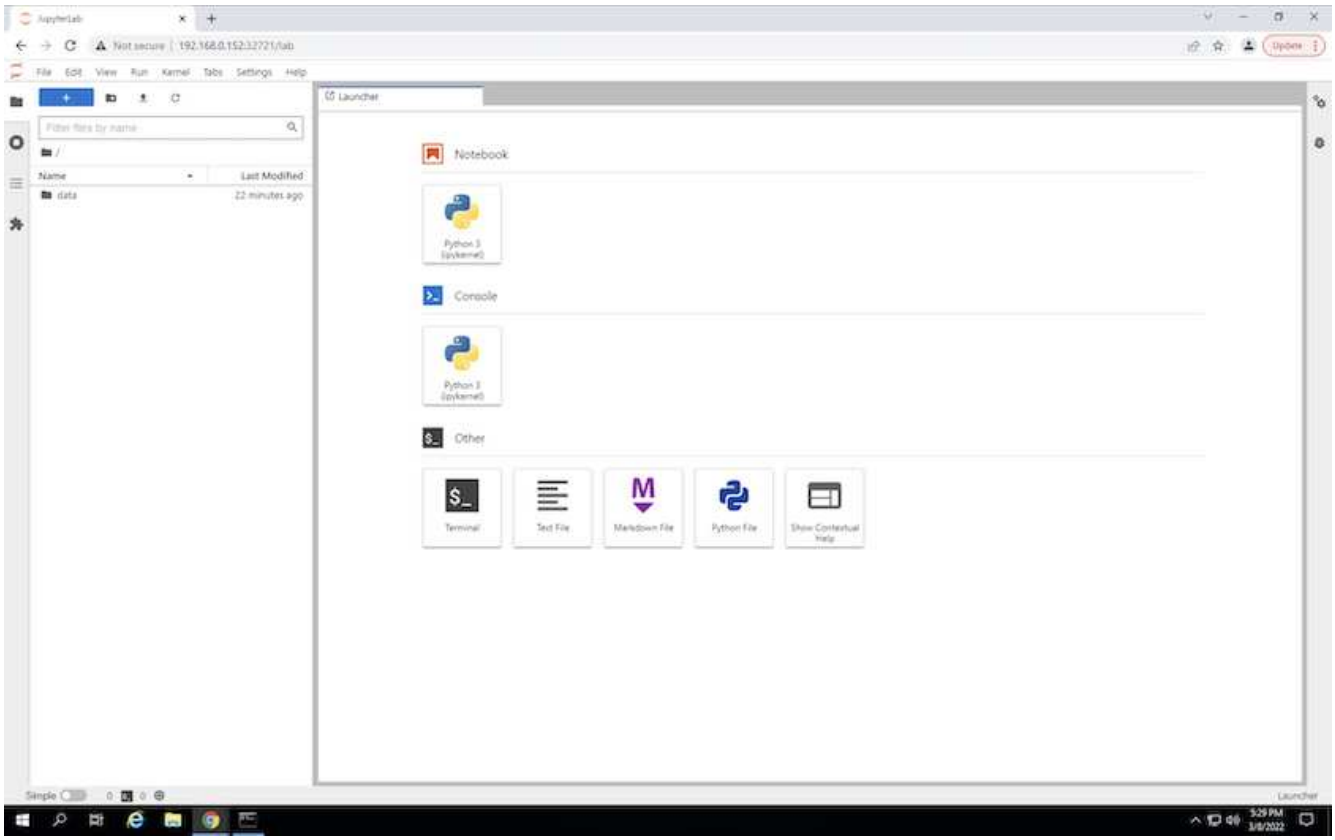
```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=inference-data --size=50Gi
Creating PersistentVolumeClaim (PVC) 'inference-data' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'inference-data' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'inference-data' in namespace 'inference'.
```

3. Utilisez le kit NetApp DataOps pour créer un nouvel espace de travail JupyterLab. Montez le volume persistant créé à l'étape précédente en utilisant le `--mount- pvc` option. Allouez des GPU NVIDIA à l'espace de travail si nécessaire à l'aide de `-- nvidia-gpu` option.

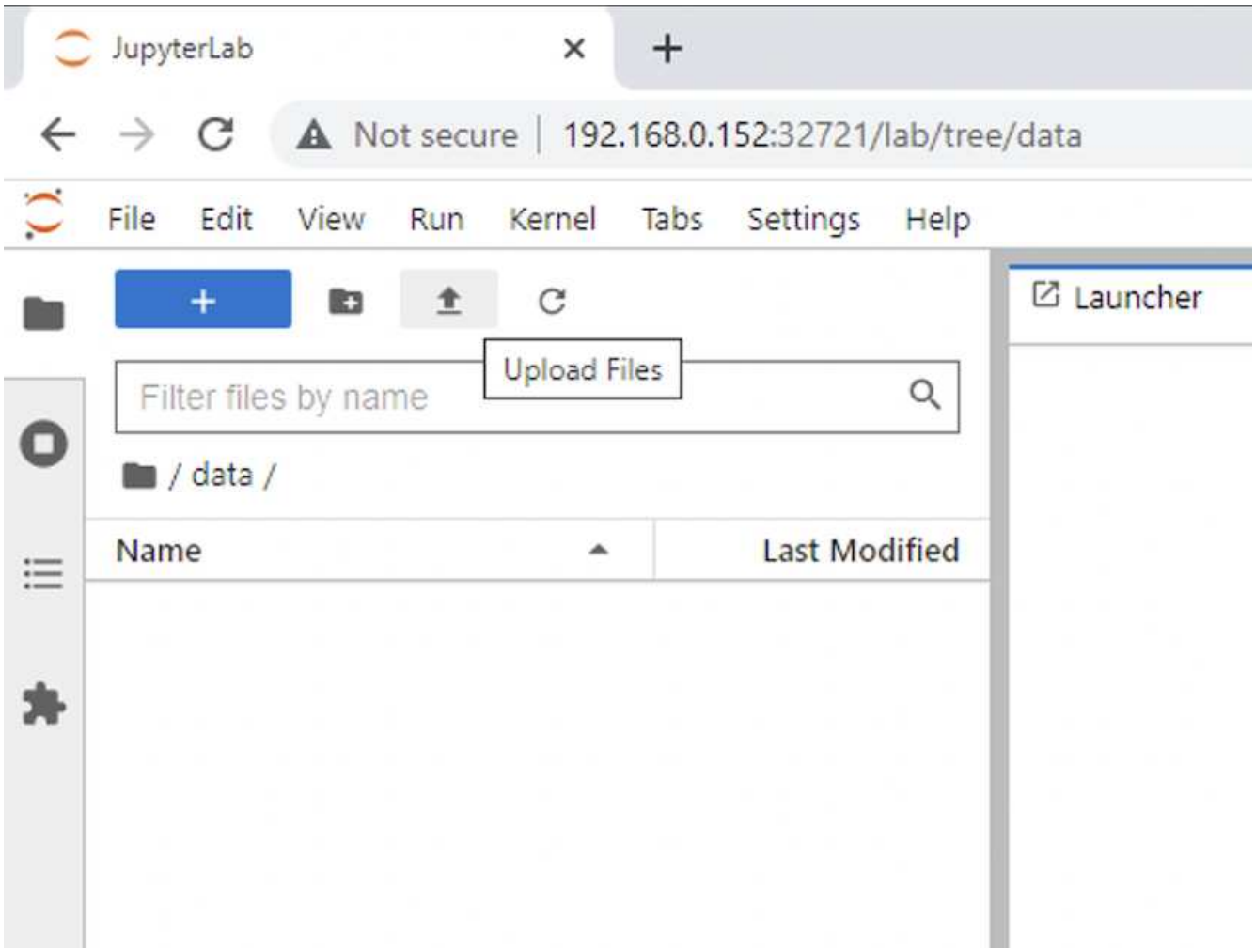
   Dans l'exemple suivant, le volume persistant `inference-data` Est monté sur le conteneur de l'espace de travail JupyterLab à `/home/jovyan/data`. Lors de l'utilisation d'images de conteneur Jupyter officiel de projet, `/home/jovyan` Est présenté comme le répertoire de niveau supérieur dans l'interface Web de JupyterLab.

```
$ netapp_dataops_k8s_cli.py create jupyterlab --namespace=inference
--workspace-name=live-inference --size=50Gi --nvidia-gpu=2 --mount
-pvc=inference-data:/home/jovyan/data
Set workspace password (this password will be required in order to
access the workspace):
Re-enter password:
Creating persistent volume for workspace...
Creating PersistentVolumeClaim (PVC) 'ntap-dsutil-jupyterlab-live-
inference' in namespace 'inference'.
PersistentVolumeClaim (PVC) 'ntap-dsutil-jupyterlab-live-inference'
created. Waiting for Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'ntap-dsutil-jupyterlab-live-inference' in namespace 'inference'.
Creating Service 'ntap-dsutil-jupyterlab-live-inference' in namespace
'inference'.
Service successfully created.
Attaching Additional PVC: 'inference-data' at mount_path:
'/home/jovyan/data'.
Creating Deployment 'ntap-dsutil-jupyterlab-live-inference' in namespace
'inference'.
Deployment 'ntap-dsutil-jupyterlab-live-inference' created.
Waiting for Deployment 'ntap-dsutil-jupyterlab-live-inference' to reach
Ready state.
Deployment successfully created.
Workspace successfully created.
To access workspace, navigate to http://192.168.0.152:32721
```

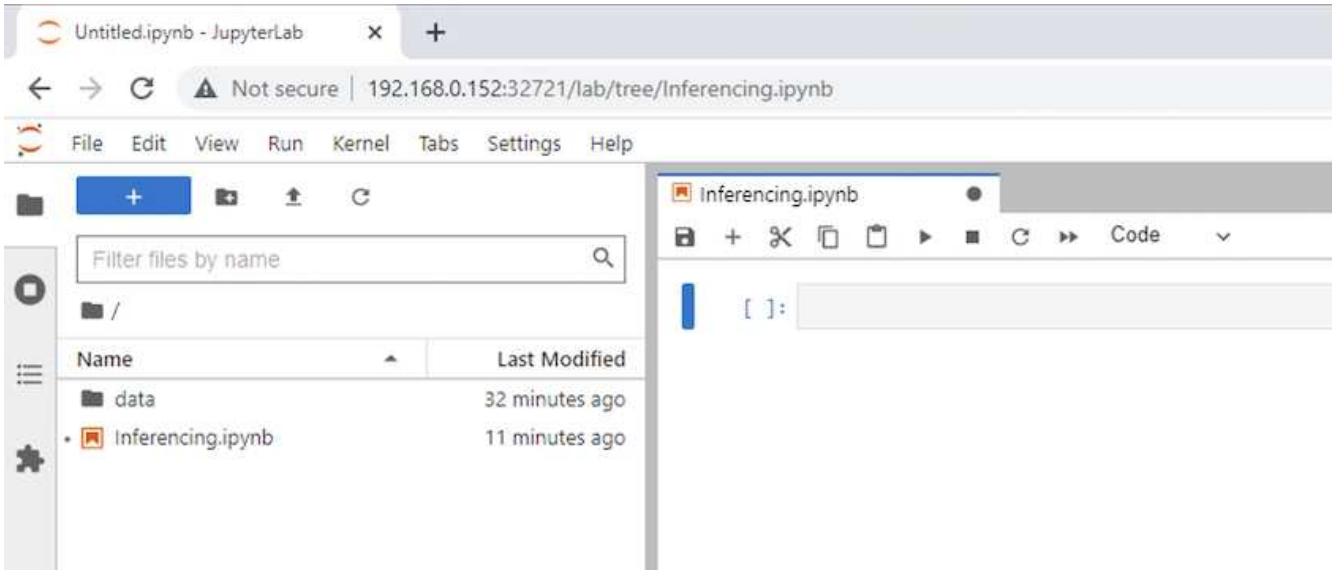
4. Accédez à l'espace de travail JupyterLab à l'aide de l'URL spécifiée dans la sortie du `create jupyterlab` commande. Le répertoire de données représente le volume persistant monté dans l'espace de travail.



5. Ouvrez le `data` directory et télécharge les fichiers sur lesquels l'inférence doit être effectuée. Lorsque les fichiers sont chargés dans le répertoire de données, ils sont automatiquement stockés sur le volume persistant monté sur l'espace de travail. Pour télécharger des fichiers, cliquez sur l'icône Télécharger des fichiers, comme illustré dans l'image suivante.



6. Retournez au répertoire de premier niveau et créez un nouveau bloc-notes.



7. Ajoutez le code d'inférence à l'ordinateur portable. L'exemple suivant montre le code d'inférence pour un cas d'utilisation de détection d'image.

Launcher | image-demo-pytorch.ipynb

Code | Python 3 (ipykernel)

STEP 3-1: Clean (Without obfuscation) detection

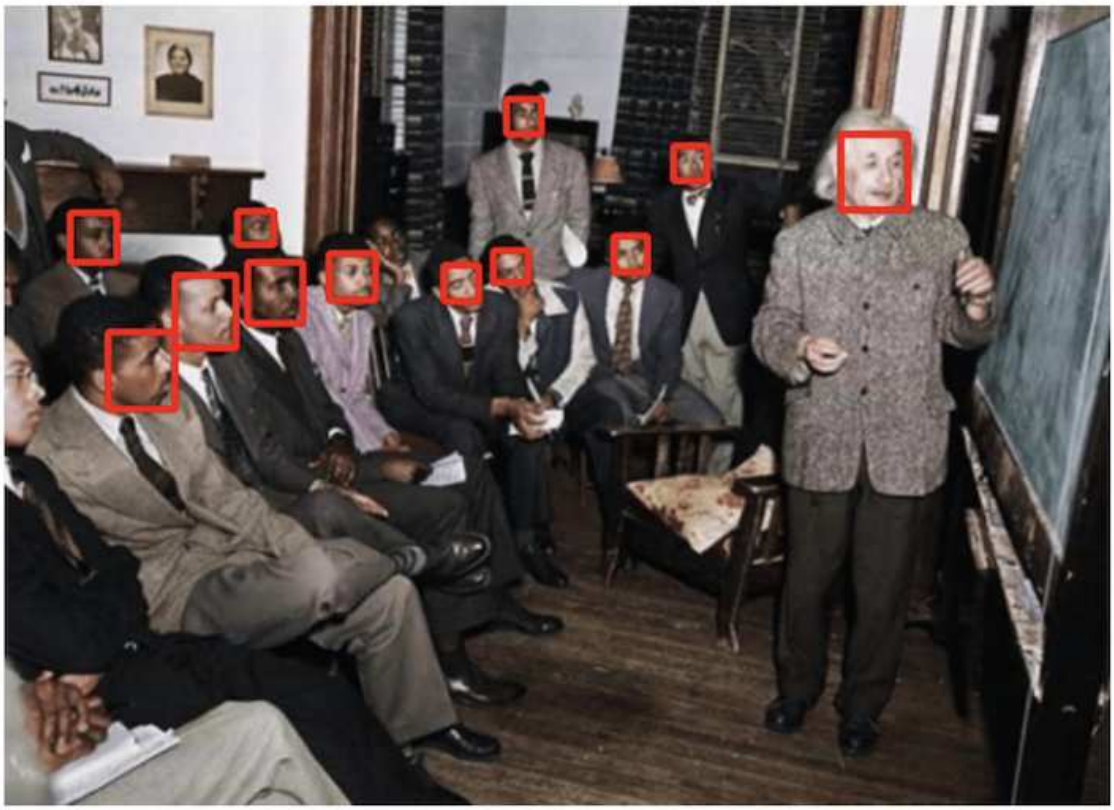
[9]:
```
# get current frame
frame = input_image

# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)

# run forward pass
clean_activation = clean_model.forward_head(preprocessed_input)  # runs the first few layers
loc, pred = clean_model.forward_tail(clean_activation)  # runs rest of the layers

# postprocess output
clean_pred = (loc.detach().cpu().numpy(), pred.detach().cpu().numpy())
clean_outputs = postprocess_outputs(
    clean_pred, [[input_image_width, input_image_height]], priors, THRESHOLD
)

# draw rectangles
clean_frame = copy.deepcopy(frame)  # needs to be deep copy
for (x1, y1, x2, y2, s) in clean_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(clean_frame, (x1, y1), (x2, y2), (0, 0, 255), 4)
```

Launcher | image-demo-pytorch.ipynb

Code | Python 3 (ipykernel)

Visualize Clean (Without obfuscation) detection

[10]:
```
show_cv2_image(clean_frame, scale=2)
```



8. Ajoutez Protopia obfuscation à votre code d'inférence. Protopia travaille directement avec les clients pour fournir une documentation spécifique à l'utilisation et n'est pas dans le cadre du présent rapport technique. L'exemple suivant montre le code d'inférence pour un cas d'utilisation de détection d'image avec Protopia obfuscation ajouté.

Launcher | image-demo-pytorch.ipynb

Code | Python 3 (ipykernel)

STEP 3-2: Protopia AI (With obfuscation) detection

[11]:
```
# get current frame
frame = input_image

# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)

# run forward pass
not_noisy_activation = noisy_model.forward_head(preprocessed_input)  # runs the first few layers
##################################################################
#          SINGLE ADITIONAL LINE FOR PRIVATE INFERENCE           #
##################################################################
noisy_activation = noisy_model.forward_noise(not_noisy_activation)
##################################################################
loc, pred = noisy_model.forward_tail(noisy_activation)  # runs rest of the layers

# postprocess output
noisy_pred = (loc.detach().cpu().numpy(), pred.detach().cpu().numpy())
noisy_outputs = postprocess_outputs(
    noisy_pred, [[input_image_width, input_image_height]], priors, THRESHOLD * 0.5
)

# get reconstruction of the noisy activation
noisy_reconstruction = decoder_function(noisy_activation)
noisy_reconstruction = noisy_reconstruction.detach().cpu().numpy()[0]
noisy_reconstruction = unpreprocess_output(
    noisy_reconstruction, (input_image_width, input_image_height), True
).astype(np.uint8)

# draw rectangles
for (x1, y1, x2, y2, s) in noisy_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(noisy_reconstruction, (x1, y1), (x2, y2), (0, 0, 255), 4)
```

Launcher | image-demo-pytorch.ipynb

Code | Python 3 (ipykernel)

Visualize Protopia AI (With obfuscation) detection

[12]:
```
show_cv2_image(noisy_reconstruction, scale=2)
```

Scénario 2 – inférence par lots sur Kubernetes

1. Créez un namespace Kubernetes pour les workloads d'inférence d'IA/DE ML.

```
$ kubectl create namespace inference
namespace/inference created
```

2. NetApp DataOps Toolkit permet de provisionner un volume persistant pour le stockage des données sur lesquelles vous souhaitez procéder à l'inférence.

```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=inference-data --size=50Gi
Creating PersistentVolumeClaim (PVC) 'inference-data' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'inference-data' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'inference-data' in namespace 'inference'.
```

3. Remplissez le nouveau volume persistant avec les données sur lesquelles vous souhaitez procéder à l'inférence.

   Il existe plusieurs méthodes pour charger des données sur un volume persistant. Si vos données sont actuellement stockées dans une plateforme de stockage objet compatible S3, comme NetApp StorageGRID ou Amazon S3, vous pouvez utiliser "Fonctionnalités du kit de données NetApp Data Mover S3". Une autre méthode simple consiste à créer un espace de travail JupyterLab, puis à télécharger des fichiers via l'interface Web de JupyterLab, comme indiqué dans les étapes 3 à 5 de la section «Scénario 1 : inférence à la demande dans JupyterLab. »

4. Créez une tâche Kubernetes pour la tâche d'inférence des lots. L'exemple suivant montre une tâche d'inférence par lot pour un cas d'utilisation de détection d'image. Ce travail effectue l'inférence sur chaque image d'un ensemble d'images et écrit les metrics de précision d'inférence sur le mode stdout.

```
$ vi inference-job-raw.yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: netapp-inference-raw
  namespace: inference
spec:
  backoffLimit: 5
  template:
    spec:
      volumes:
      - name: data
        persistentVolumeClaim:
          claimName: inference-data
      - name: dshm
        emptyDir:
          medium: Memory
      containers:
      - name: inference
        image: netapp-protopia-inference:latest
        imagePullPolicy: IfNotPresent
        command: ["python3", "run-accuracy-measurement.py", "--dataset",
"/data/netapp-face-detection/FDDB"]
        resources:
          limits:
            nvidia.com/gpu: 2
        volumeMounts:
        - mountPath: /data
          name: data
        - mountPath: /dev/shm
          name: dshm
      restartPolicy: Never
$ kubectl create -f inference-job-raw.yaml
job.batch/netapp-inference-raw created
```

5. Vérifiez que la tâche d’inférence a été correctement terminée.

```
$ kubectl -n inference logs netapp-inference-raw-255sp
100%|██████████| 89/89 [00:52<00:00,  1.68it/s]
Reading Predictions : 100%|██████████| 10/10 [00:01<00:00,  6.23it/s]
Predicting ... : 100%|██████████| 10/10 [00:16<00:00,  1.64s/it]
==================== Results ====================
FDDB-fold-1 Val AP: 0.9491256561145955
FDDB-fold-2 Val AP: 0.9205024466101926
FDDB-fold-3 Val AP: 0.9253013871078468
FDDB-fold-4 Val AP: 0.9399781485863011
FDDB-fold-5 Val AP: 0.9504280149478732
FDDB-fold-6 Val AP: 0.9416473519339292
FDDB-fold-7 Val AP: 0.9241631566241117
FDDB-fold-8 Val AP: 0.9072663297546659
FDDB-fold-9 Val AP: 0.9339648715035469
FDDB-fold-10 Val AP: 0.9447707905560152
FDDB Dataset Average AP: 0.9337148153739079
==================================================
mAP: 0.9337148153739079
```

6. Ajoutez Protopia obfuscation à votre travail d'inférence. Vous trouverez des instructions spécifiques à chaque cas pour ajouter des objets de Protopia directement à partir de Protopia, qui ne sont pas dans le cadre de ce rapport technique. L'exemple suivant montre un travail d'inférence par lot pour un cas d'utilisation de détection de face avec l'obfuscation Protopia ajouté à l'aide d'une valeur ALPHA de 0.8. Cette tâche applique l'obfuscation de Protopia avant d'effectuer l'inférence pour chaque image d'un ensemble d'images, puis écrit les metrics de précision de l'inférence dans le système.

   Nous avons répété cette étape pour les valeurs ALPHA 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9 et 0.95. Les résultats sont présentés dans la "« Comparaison de la précision de l'inférence »."

```
$ vi inference-job-protopia-0.8.yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: netapp-inference-protopia-0.8
  namespace: inference
spec:
  backoffLimit: 5
  template:
    spec:
      volumes:
      - name: data
        persistentVolumeClaim:
          claimName: inference-data
      - name: dshm
        emptyDir:
          medium: Memory
      containers:
      - name: inference
        image: netapp-protopia-inference:latest
        imagePullPolicy: IfNotPresent
        env:
        - name: ALPHA
          value: "0.8"
        command: ["python3", "run-accuracy-measurement.py", "--dataset",
"/data/netapp-face-detection/FDDB", "--alpha", "$(ALPHA)", "--noisy"]
        resources:
          limits:
            nvidia.com/gpu: 2
        volumeMounts:
        - mountPath: /data
          name: data
        - mountPath: /dev/shm
          name: dshm
      restartPolicy: Never
$ kubectl create -f inference-job-protopia-0.8.yaml
job.batch/netapp-inference-protopia-0.8 created
```

7. Vérifiez que la tâche d’inférence a été correctement terminée.

```
$ kubectl -n inference logs netapp-inference-protopia-0.8-b4dkz
100%|██████████| 89/89 [01:05<00:00,  1.37it/s]
Reading Predictions : 100%|██████████| 10/10 [00:02<00:00,  3.67it/s]
Predicting ... : 100%|██████████| 10/10 [00:22<00:00,  2.24s/it]
==================== Results ====================
FDDB-fold-1 Val AP: 0.8953066115834589
FDDB-fold-2 Val AP: 0.8819580264029936
FDDB-fold-3 Val AP: 0.8781107458462862
FDDB-fold-4 Val AP: 0.9085731346308461
FDDB-fold-5 Val AP: 0.9166445508275378
FDDB-fold-6 Val AP: 0.9101178994188819
FDDB-fold-7 Val AP: 0.8383443678423771
FDDB-fold-8 Val AP: 0.8476311547659464
FDDB-fold-9 Val AP: 0.8739624502111121
FDDB-fold-10 Val AP: 0.8905468076424851
FDDB Dataset Average AP: 0.8841195749171925
==================================================
mAP: 0.8841195749171925
```

Scénario 3 – NVIDIA Triton Inférence Server

1. Créez un namespace Kubernetes pour les workloads d'inférence d'IA/DE ML.

```
$ kubectl create namespace inference
namespace/inference created
```

2. Utilisez le kit NetApp DataOps Toolkit pour provisionner un volume persistant à utiliser comme référentiel de modèles pour le serveur NVIDIA Triton Inférence.

```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=triton-model-repo --size=100Gi
Creating PersistentVolumeClaim (PVC) 'triton-model-repo' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'triton-model-repo' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'triton-model-repo' in namespace 'inference'.
```

3. Stockez votre modèle sur le nouveau volume persistant dans un "format" C'est reconnu par le serveur NVIDIA Triton Inférence Server.

   Il existe plusieurs méthodes pour charger des données sur un volume persistant. Une méthode simple consiste à créer un espace de travail JupyterLab, puis à télécharger des fichiers via l'interface Web de JupyterLab, comme indiqué dans les étapes 3 à 5 de la section «Scénario 1 : inférence à la demande dans

JupyterLab. ”

4. Utilisez le kit NetApp DataOps pour déployer une nouvelle instance NVIDIA Triton Inférence Server.

```
$ netapp_dataops_k8s_cli.py create triton-server --namespace=inference
--server-name=netapp-inference --model-repo-pvc-name=triton-model-repo
Creating Service 'ntap-dsutil-triton-netapp-inference' in namespace
'inference'.
Service successfully created.
Creating Deployment 'ntap-dsutil-triton-netapp-inference' in namespace
'inference'.
Deployment 'ntap-dsutil-triton-netapp-inference' created.
Waiting for Deployment 'ntap-dsutil-triton-netapp-inference' to reach
Ready state.
Deployment successfully created.
Server successfully created.
Server endpoints:
http: 192.168.0.152: 31208
grpc: 192.168.0.152: 32736
metrics: 192.168.0.152: 30009/metrics
```

5. Utilisez un SDK client Triton pour effectuer une tâche d’inférence. L’extrait de code Python suivant utilise le SDK client Triton Python pour effectuer une tâche d’inférence pour un cas d’utilisation de détection de visage. Cet exemple appelle l’API Triton et transmet une image pour l’inférence. Le serveur Triton Inférence reçoit ensuite la requête, appelle le modèle et renvoie la sortie d’inférence dans le cadre des résultats de l’API.

```
# get current frame
frame = input_image
# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)
# run forward pass
clean_activation = clean_model_head(preprocessed_input)  # runs the
first few layers
#######################################################################
##############
#          pass clean image to Triton Inference Server API for
inferencing           #
#######################################################################
##############
triton_client =
httpclient.InferenceServerClient(url="192.168.0.152:31208",
verbose=False)
model_name = "face_detection_base"
inputs = []
```

```
outputs = []
inputs.append(httpclient.InferInput("INPUT__0", [1, 128, 32, 32],
"FP32"))
inputs[0].set_data_from_numpy(clean_activation.detach().cpu().numpy(),
binary_data=False)
outputs.append(httpclient.InferRequestedOutput("OUTPUT__0",
binary_data=False))
outputs.append(httpclient.InferRequestedOutput("OUTPUT__1",
binary_data=False))
results = triton_client.infer(
    model_name,
    inputs,
    outputs=outputs,
    #query_params=query_params,
    headers=None,
    request_compression_algorithm=None,
    response_compression_algorithm=None)
#print(results.get_response())
statistics =
triton_client.get_inference_statistics(model_name=model_name,
headers=None)
print(statistics)
if len(statistics["model_stats"]) != 1:
    print("FAILED: Inference Statistics")
    sys.exit(1)

loc_numpy = results.as_numpy("OUTPUT__0")
pred_numpy = results.as_numpy("OUTPUT__1")
#######################################################################
##############
# postprocess output
clean_pred = (loc_numpy, pred_numpy)
clean_outputs = postprocess_outputs(
    clean_pred, [[input_image_width, input_image_height]], priors,
THRESHOLD
)
# draw rectangles
clean_frame = copy.deepcopy(frame)  # needs to be deep copy
for (x1, y1, x2, y2, s) in clean_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(clean_frame, (x1, y1), (x2, y2), (0, 0, 255), 4)
```

6. Ajoutez Protopia obfuscation à votre code d'inférence. Vous trouverez des instructions propres à chaque cas pour ajouter des obfuscations Protopia directement à partir de Protopia ; cependant, ce processus n'est pas dans le cadre de ce rapport technique. L'exemple suivant montre le même code Python que celui

indiqué à l'étape 5 précédente, mais avec l'obfuscation Protopia ajouté.

Notez que l'obfuscation Protopia est appliquée à l'image avant de la transmettre à l'API Triton. Ainsi, l'image non obfusquée ne quitte jamais la machine locale. Seule l'image masquée est transmise sur le réseau. Ce flux de production s'applique aux cas où les données sont collectées dans une zone de confiance, mais doivent ensuite être transférées en dehors de cette zone de confiance pour l'inférence. Sans l'obfuscation Protopia, il n'est pas possible d'implémenter ce type de flux de travail sans que des données sensibles quittent la zone de confiance.

```
# get current frame
frame = input_image
# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)
# run forward pass
not_noisy_activation = noisy_model_head(preprocessed_input)  # runs the
first few layers
##################################################################
#          obfuscate image locally prior to inferencing          #
#          SINGLE ADITIONAL LINE FOR PRIVATE INFERENCE           #
##################################################################
noisy_activation = noisy_model_noise(not_noisy_activation)
##################################################################
#######################################################################
####################
#          pass obfuscated image to Triton Inference Server API for
inferencing           #
#######################################################################
####################
triton_client =
httpclient.InferenceServerClient(url="192.168.0.152:31208",
verbose=False)
model_name = "face_detection_noisy"
inputs = []
outputs = []
inputs.append(httpclient.InferInput("INPUT__0", [1, 128, 32, 32],
"FP32"))
inputs[0].set_data_from_numpy(noisy_activation.detach().cpu().numpy(),
binary_data=False)
outputs.append(httpclient.InferRequestedOutput("OUTPUT__0",
binary_data=False))
outputs.append(httpclient.InferRequestedOutput("OUTPUT__1",
binary_data=False))
results = triton_client.infer(
    model_name,
    inputs,
    outputs=outputs,
```

```
    #query_params=query_params,
    headers=None,
    request_compression_algorithm=None,
    response_compression_algorithm=None)
#print(results.get_response())
statistics =
triton_client.get_inference_statistics(model_name=model_name,
headers=None)
print(statistics)
if len(statistics["model_stats"]) != 1:
    print("FAILED: Inference Statistics")
    sys.exit(1)

loc_numpy = results.as_numpy("OUTPUT__0")
pred_numpy = results.as_numpy("OUTPUT__1")
#######################################################################
###################

# postprocess output
noisy_pred = (loc_numpy, pred_numpy)
noisy_outputs = postprocess_outputs(
    noisy_pred, [[input_image_width, input_image_height]], priors,
THRESHOLD * 0.5
)
# get reconstruction of the noisy activation
noisy_reconstruction = decoder_function(noisy_activation)
noisy_reconstruction = noisy_reconstruction.detach().cpu().numpy()[0]
noisy_reconstruction = unpreprocess_output(
    noisy_reconstruction, (input_image_width, input_image_height), True
).astype(np.uint8)
# draw rectangles
for (x1, y1, x2, y2, s) in noisy_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(noisy_reconstruction, (x1, y1), (x2, y2), (0, 0, 255),
4)
```

Comparaison de la précision d’inférence

Pour cette validation, nous avons procédé à l’inférence pour une détection d’images à l’aide d’un jeu d’images brutes. Nous avons ensuite effectué la même tâche d’inférence sur le même jeu d’images avec l’obfuscation Protopia ajoutée avant l’inférence. Nous avons répété la tâche en utilisant différentes valeurs ALPHA pour le composant d’obfuscation Protopia. Dans le contexte de l’obfuscation Protopia, la valeur ALPHA représente la quantité d’obfuscation appliquée, avec une valeur ALPHA plus élevée

représentant un niveau d'obfuscation plus élevé. Nous avons ensuite comparé la précision d'inférence sur ces différentes exécutions.

Les deux tableaux suivants détaillent nos cas d'utilisation et présentent les résultats.

Protopia travaille directement avec les clients pour déterminer la valeur ALPHA appropriée pour un cas d'utilisation spécifique.

Composant	**Détails**
Modèle	Coffrets de finition (PyTorch) -
Jeu de données	Jeu de données FDDB

Obfuscation Protopia	**ALPHA**	**Précision**
Non	S/O	0.9337148153739079
Oui.	0.05	0.9028766627325002
Oui.	0.1	0.9024301009661478
Oui.	0.2	0.9081836283186224
Oui.	0.4	0.9073066107482036
Oui.	0.6	0.8847816568680239
Oui.	0.8	0.8841195749171925
Oui.	0.9	0.8455427675252052
Oui.	0.95	0.8455427675252052

Vitesse d'obfuscation

Pour cette validation, nous avons appliqué l'obfuscation Protopia à une image de 1920 x 1080 pixels cinq fois et mesuré le temps nécessaire à l'étape d'obfuscation pour effectuer chaque fois.

Nous avons utilisé PyTorch s'exécutant sur un seul GPU NVIDIA V100 pour appliquer l'obfuscation, et nous avons éliminé le cache GPU entre les exécutions. L'étape d'obfuscation a pris respectivement 5,47 ms, 5,27 ms, 4,54 ms, 5,24 ms et 4,84 ms pour effectuer l'ensemble des cinq passages. La vitesse moyenne était de 5,072 ms.

Conclusion

Les données existent dans trois États : au repos, en transit et dans le calcul. Un service d'inférence d'IA doit contenir des données non menaces lors de l'intégralité du processus. La protection des données pendant l'inférence est essentielle. En effet, le processus peut exposer des informations privées aux clients externes et à l'entreprise fournissant le service d'inférence. Protopia ai est une solution logicielle unique non-intégrée pour l'inférence confidentielle de l'IA sur le marché actuel. Avec Protopia, l'IA est alimenté uniquement les informations transformées dans les dossiers de données, essentielles pour mener à bien les tâches d'IA et DE ML. Cette transformation

stochastique n’est pas une forme de masquage et est basée sur la modification mathématiquement de la représentation des données à l’aide d’un bruit adapté.

Les systèmes de stockage NetApp dotés des fonctionnalités ONTAP offrent des performances identiques ou supérieures au stockage SSD local. Associés au kit NetApp DataOps, ils offrent les avantages suivants aux data Scientists, aux ingénieurs de données, aux développeurs d’IA/AM et aux décideurs IT d’entreprise ou d’entreprise :

- Partage sans effort des données entre des systèmes d’IA, l’analytique et d’autres systèmes métier stratégiques. Ce partage des données permet de réduire la surcharge de l’infrastructure, d’améliorer les performances et de rationaliser la gestion des données à l’échelle de l’entreprise.
- Faites évoluer indépendamment les ressources de calcul et de stockage pour réduire les coûts et améliorer l’utilisation des ressources.
- Rationalisation des workflows de développement et de déploiement à l’aide de copies et de clones Snapshot intégrés pour des espaces de travail utilisateur instantanés et compacts, un contrôle intégré des versions et un déploiement automatisé.
- Protection et gouvernance des données pour la reprise après incident, la continuité de l’activité et les exigences réglementaires.
- Appel simplifié à des opérations de gestion des données ; prise rapide de copies Snapshot d’espaces de travail de data Scientists pour la sauvegarde et la traçabilité à partir du kit NetApp DataOps Toolkit dans des ordinateurs portables Jupyter.

La solution NetApp et Protopia offre une architecture flexible et scale-out, idéale pour les déploiements d’inférence d’IA haute performance. Il protège les données et garantit la confidentialité des informations sensibles pour lesquelles les exigences d’inférence d’IA confidentielles peuvent être respectées grâce à des pratiques d’IA responsables dans les déploiements sur site et dans le cloud hybride.

Où trouver des informations supplémentaires et des remerciements

Pour en savoir plus sur les informations fournies dans ce document, consultez ces documents et/ou sites web :

- Le logiciel de gestion des données NetApp ONTAP : bibliothèque d’informations ONTAP

  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- NetApp Trident pour le stockage persistant pour les conteneurs

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- Kit NetApp DataOps

  "https://github.com/NetApp/netapp-dataops-toolkit"

- NetApp Kubernetes pour les conteneurs : NetApp Astra Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- Protopia ai - Inférence confidentielle

  "https://protopia.ai/blog/protopia-ai-takes-on-the-missing-link-in-ai-privacy-confidential-inference/"

- Copie et synchronisation NetApp BlueXP

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"

- Serveur d'inférence NVIDIA Triton

  "https://developer.nvidia.com/nvidia-triton-inference-server"

- Documentation NVIDIA Triton Inférence Server

  "https://docs.nvidia.com/deeplearning/triton-inference-server/index.html"

- Coffrets en PyTorch

  "https://github.com/zisianw/FaceBoxes.PyTorch"

Remerciements

- Mark Cates, responsable produits senior, NetApp
- Sufian Ahmad, Ingénieur marketing et technique, NetApp
- Hadi Esmaeilzadeh, directeur technique et professeur, Protopia ai

Analyse des sentiments avec NetApp ai

Tr-4910 : analyse du sentiment issue des communications avec le client sur NetApp ai

Rick Huang, Satish Thyagarajan et David Arnette, NetApp Diego Sosa-Coba, SFL Scientific

Ce rapport technique fournit des conseils de conception afin que les clients puissent réaliser une analyse des sentiments dans un centre de support global de niveau entreprise en utilisant les technologies de gestion des données NetApp avec un framework logiciel NVIDIA, par le biais de l'apprentissage par transfert et de la discussion sur l'IA. Cette solution s'applique à tout secteur souhaitant obtenir des informations sur les clients à partir de fichiers vocaux ou texte enregistrés représentant des journaux de conversation, des e-mails et d'autres communications audio ou texte. Nous avons mis en œuvre un pipeline de bout en bout afin de démontrer la reconnaissance vocale automatique, l'analyse des sentiments en temps réel et le réentraînement du modèle de traitement du langage naturel du deep learning. Nous avons également mis en place des fonctionnalités de recyclage sur un cluster de calcul accéléré par processeur graphique grâce au stockage 100 % Flash de NetApp connecté au cloud. Des modèles linguistiques de pointe peuvent être entraînés et optimisés pour inférence rapidement avec le centre de support global. Ils assurent ainsi une expérience client exceptionnelle et permettent d'évaluer de manière objective les performances des employés à long terme.

L'analyse des sentiments est un domaine d'étude dans le traitement du langage naturel (NLP) par lequel des sentiments positifs, négatifs ou neutres sont extraits du texte. Les systèmes d'IA conversationnels ont été étendus à un niveau d'intégration quasi global, car de plus en plus de gens viennent à les interagir avec eux. L'analyse des sentiments présente de nombreux cas d'utilisation, allant de la détermination de la performance

des employés du centre de support lors de conversations avec les appelants à la fourniture de réponses de chatbot automatisées appropriées à la prévision du prix des actions d'une entreprise en fonction des interactions entre les représentants des entreprises et le public lors d'appels trimestriels sur les bénéfices. En outre, l'analyse de sentiment peut être utilisée pour déterminer l'opinion du client sur les produits, les services ou l'assistance fournis par la marque.

Cette solution de bout en bout utilise des modèles NLP pour effectuer une analyse de confiance de haut niveau qui permet de créer des cadres analytiques de centre de support. Les enregistrements audio sont traités en texte écrit et le sentiment est extrait de chaque phrase de la conversation. Les résultats, regroupés dans un tableau de bord, peuvent être conçus pour analyser les sentiments de conversation, historiquement et en temps réel. Cette solution peut être généralisée à d'autres solutions avec des modalités de données et des besoins de sortie similaires. Grâce aux données appropriées, d'autres cas d'utilisation peuvent être mis en œuvre. Par exemple, les appels sur les bénéfices de l'entreprise peuvent être analysés pour obtenir un sentiment à l'aide du même pipeline de bout en bout. D'autres formes d'analyses du PNL, telles que la modélisation de sujets et la reconnaissance d'entités nommées (NER), sont également possibles en raison de la nature souple du pipeline.

Ces implémentations d'IA ont été rendues possibles par NVIDIA RIVA, le kit NVIDIA TAO et le kit NetApp DataOps. Les outils NVIDIA sont utilisés pour déployer rapidement des solutions d'IA haute performance à l'aide de modèles et de pipelines prédéfinis. Le kit NetApp DataOps simplifie les tâches de gestion des données pour accélérer le développement.

En valeur pour le client

Les entreprises voient les avantages d'un outil d'évaluation des employés et de réaction des clients pour les conversations texte, audio et vidéo pour l'analyse des sentiments. Les responsables bénéficient des informations présentées dans le tableau de bord, ce qui permet d'évaluer les employés et la satisfaction des clients en fonction des deux côtés de la conversation.

Par ailleurs, le kit NetApp DataOps gère les versions et l'allocation des données au sein de l'infrastructure du client. Cela entraîne des mises à jour fréquentes des analyses présentées au tableau de bord sans générer des coûts de stockage de données qui ne sont pas unwieldy.

Cas d'utilisation

En raison du nombre d'appels traités par ces centres d'assistance, l'évaluation du rendement des appels peut prendre beaucoup de temps si elle est effectuée manuellement. Les méthodes traditionnelles, telles que le comptage des sacs de mots et d'autres méthodes, peuvent atteindre une certaine automatisation, mais ces méthodes ne prennent pas en compte les aspects plus nuancés et le contexte sémantique du langage dynamique. Les techniques de modélisation d'IA peuvent être utilisées pour effectuer certaines de ces analyses plus nuancées de manière automatisée. De plus, avec les outils de modélisation pré-entraînés et publiés par NVIDIA, AWS, Google et autres, un pipeline de bout en bout avec des modèles complexes peut maintenant être mis en place et personnalisé avec une facilité relative.

Un pipeline de bout en bout pour l'analyse des sentiments des centres de support permet d'analyser les fichiers audio en temps réel lorsque les employés convertisse avec les appelants. Ensuite, ces fichiers audio sont traités pour être utilisés dans le composant parole-texte qui les convertit au format texte. Chaque phrase de la conversation reçoit un libellé indiquant le sentiment (positif, négatif ou neutre).

L'analyse des sentiments peut fournir un aspect essentiel des conversations pour évaluer la performance des

appels. Ces sentiments ajoutent une profondeur supplémentaire aux interactions entre les employés et les appelants. Le tableau de bord des sentiments assistés par l'IA fournit aux responsables un suivi en temps réel des sentiments au sein d'une conversation, ainsi qu'une analyse rétrospective des appels passés de l'employé.

Des outils prédéfinis peuvent être combinés de façon puissante afin de créer rapidement un pipeline d'IA de bout en bout pour résoudre ce problème. Dans ce cas, la bibliothèque NVIDIA RIVA peut être utilisée pour effectuer les deux tâches en série : transcription audio et analyse de sentiment. Le premier est un algorithme de traitement du signal d'apprentissage supervisé et le second est un algorithme de classification NLP d'apprentissage supervisé. Ces algorithmes prêts à l'emploi peuvent être ajustés pour toute utilisation pertinente avec les données pertinentes de l'entreprise à l'aide du kit NVIDIA TAO. Cela entraîne la création de solutions plus précises et plus puissantes pour un coût et des ressources réduits. Les clients peuvent intégrer le "NVIDIA Maxine" Infrastructure pour les applications de visioconférence avec accélération graphique dans leur conception de centre de support.

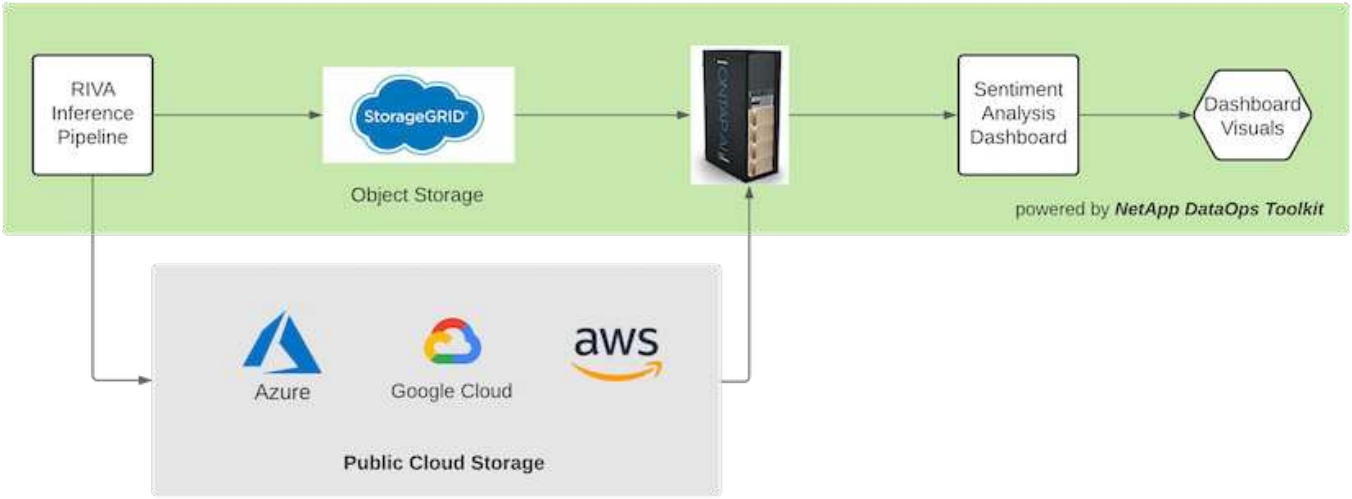
Les utilisations suivantes sont au cœur de cette solution. Dans les deux cas, TAO Toolkit permet de régler avec précision les modèles et RIVA pour le déploiement de modèles.

- Parole-à-texte
- Analyse des sentiments

Pour analyser les interactions entre les employés et les clients du centre de support, chaque conversation client sous forme d'appels audio peut être exécutée dans le pipeline pour extraire des sentiments de phrase. Ces sentiments peuvent ensuite être vérifiés par un humain pour justifier les sentiments ou les ajuster selon les besoins. Les données étiquetées sont ensuite transmises à l'étape de réglage fin pour améliorer les prédictions de sentiment. Si des données de sentiment étiquetées existent déjà, le réglage fin du modèle peut être accéléré. Dans les deux cas, le pipeline est généralisable à d'autres solutions qui exigent l'ingestion de l'audio et la classification des phrases.

Les résultats des sentiments d'IA sont soit téléchargés vers une base de données cloud externe, soit vers un système de stockage géré par l'entreprise. Les sorties de sentiment sont transférées de cette base de

données plus grande vers le stockage local pour une utilisation dans le tableau de bord qui affiche l'analyse de sentiment pour les gestionnaires. La principale fonctionnalité du tableau de bord consiste à interagir avec l'employé du service client en temps réel. Les responsables peuvent évaluer et fournir des commentaires sur les employés lors de leurs appels avec des mises à jour en direct du sentiment de chaque phrase, ainsi qu'un examen historique des performances passées de l'employé ou des réactions des clients.



Le "Kit NetApp DataOps" Peut continuer à gérer les systèmes de stockage des données même après que le pipeline d'inférence RIVA génère des étiquettes de sentiments. Les résultats de cette IA peuvent être téléchargés vers un système de stockage des données géré par le kit NetApp DataOps. Les systèmes de stockage des données doivent être capables de gérer des centaines d'insertions et de sélectionner chaque minute. Le système de stockage de périphériques local interroge le stockage de données plus important en temps réel pour l'extraction. Le tableau de bord peut également être interrogé sur ses données historiques afin d'optimiser l'expérience qu'il fournit. Le kit NetApp DataOps facilite ces deux utilisations en clonant rapidement les données et en les distribuant dans tous les tableaux de bord.

Public visé

La solution cible plusieurs groupes d'utilisateurs :

- Responsables des employés
- Ingénieurs/data Scientists
- Administrateurs IT (sur site, dans le cloud ou hybride)

Le suivi des sentiments tout au long des conversations est un outil précieux pour évaluer les performances des employés. Grâce au tableau de bord ai, les responsables peuvent voir comment les employés et les appelants changent leurs sentiments en temps réel, ce qui permet des évaluations en direct et des sessions de guidage. De plus, les entreprises peuvent bénéficier de précieux avis de la part de clients engagés dans des conversations vocales, des chatbots de texte et des visioconférences. Cette analytique client exploite les fonctionnalités de traitement multimodal à grande échelle avec des modèles et des workflows d'IA modernes et à la pointe de la technologie.

En ce qui concerne les données, un grand nombre de fichiers audio sont traités quotidiennement par le centre d'assistance. Le kit NetApp DataOps facilite le traitement des données pour affiner régulièrement les modèles et les tableaux de bord d'analyse de sentiment.

Les administrateurs INFORMATIQUES profitent également du kit NetApp DataOps car il leur permet de déplacer rapidement des données entre les environnements de déploiement et de production. Les

environnements et les serveurs NVIDIA doivent également être gérés et distribués pour permettre l'inférence en temps réel.

Architecture

L'architecture de cette solution de centre de support s'articule autour des outils préfabriqués de NVIDIA et du kit NetApp DataOps. Les outils NVIDIA sont utilisés pour déployer rapidement des solutions d'IA haute performance à l'aide de modèles et de pipelines prédéfinis. Le kit NetApp DataOps simplifie les tâches de gestion des données pour accélérer le développement.

Technologie de la solution

"NVIDIA RIVA" Est un kit de développement accéléré par processeur graphique pour concevoir des applications d'IA multimodales qui fournissent des performances en temps réel sur les GPU. Le kit NVIDIA train, Adapt, and Optimize (TAO) offre un moyen plus rapide et plus simple d'accélérer l'entraînement et de créer rapidement des modèles d'IA spécifiques à un domaine, précis et performants.

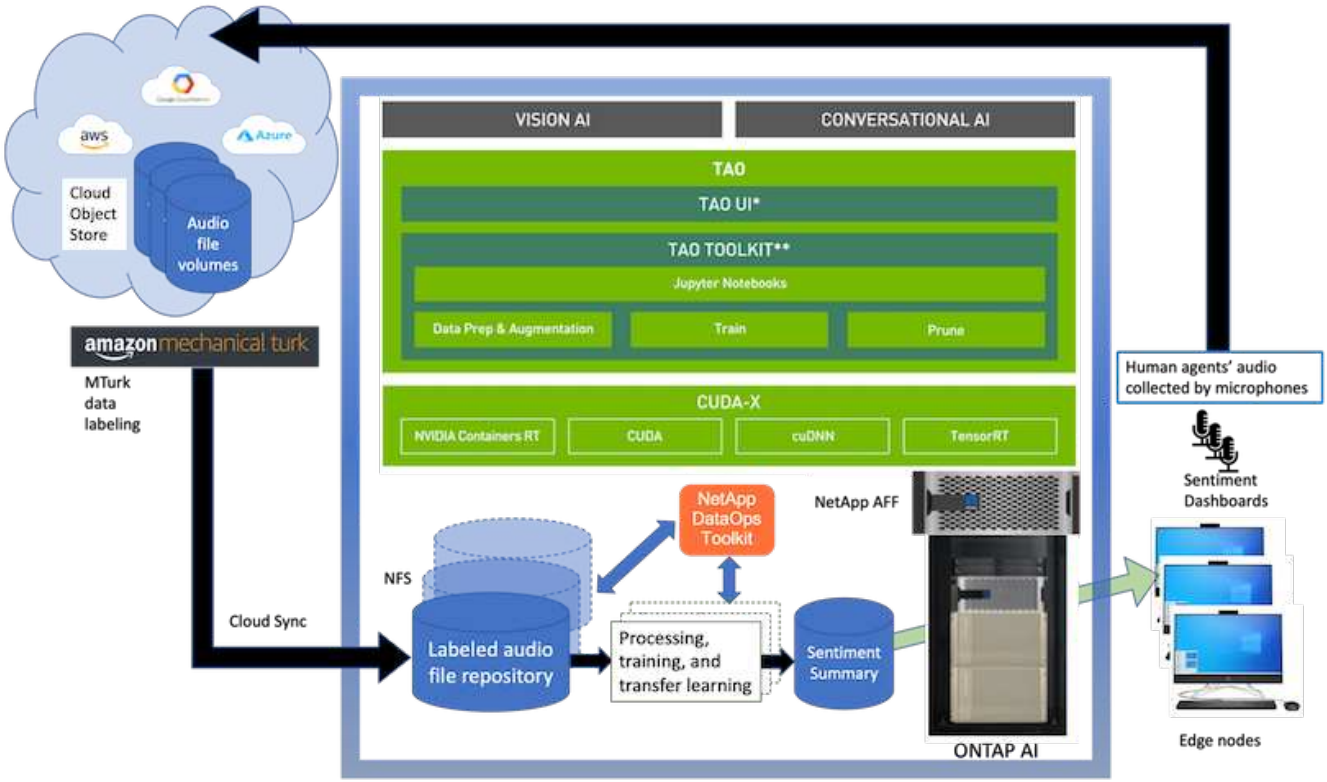
Le kit NetApp DataOps est une bibliothèque Python qui simplifie la tâche des développeurs, des data Scientists, des ingénieurs DevOps et des ingénieurs en gestion des données. Cela inclut le provisionnement quasi instantané d'un nouveau volume de données ou d'un nouvel espace de travail JupyterLab, le clonage quasi instantané d'un volume de données ou d'un espace de travail JupyterLab, ainsi que le clonage quasi instantané d'un volume de données ou d'un espace de travail JupyterLab à des fins de traçabilité et de base.

Diagramme architectural

Le schéma suivant illustre l'architecture de la solution. Il existe trois catégories d'environnement principales : le cloud, le cœur et la périphérie. Chacune des catégories peut être géographiquement dispersée. Par exemple, le cloud contient des magasins d'objets avec des fichiers audio dans des compartiments de différentes régions, tandis que le cœur peut contenir des data centers reliés par un réseau haut débit ou par copie et synchronisation NetApp BlueXP. Les nœuds périphériques indiquent les plates-formes de travail quotidiennes de chaque agent humain, où des microphones et des outils de tableau de bord interactifs sont disponibles pour visualiser le sentiment et collecter des données audio à partir de conversations avec les clients.

Dans les data centers accélérés par GPU, les entreprises peuvent utiliser NVIDIA "RIVA" Framework pour créer des applications d'IA conversationnelles, vers lesquelles le "Tao Toolkit" Se connecte pour le finetuning de modèles et le recyclage à l'aide de techniques d'apprentissage par transfert de L. Ces workflows et applications de calcul sont optimisés par le "Kit NetApp DataOps", Fournir les meilleures fonctionnalités de gestion des données qu'ONTAP offre. Ce kit permet aux équipes de gestion de données d'entreprise de créer rapidement des prototypes de modèles avec des données structurées et non structurées associées, à l'aide de snapshots et de clones pour la traçabilité, la gestion des versions, les tests A/B, ce qui assure la sécurité et la gouvernance et la conformité réglementaire. Voir la section "Design du stockage" pour en savoir plus.

Cette solution présente les étapes détaillées du traitement des fichiers audio, de l'entraînement des modèles NLP, de l'apprentissage par transfert et de la gestion des données. Le pipeline de bout en bout qui en résulte génère un résumé des sentiments qui s'affiche en temps réel sur les tableaux de bord des agents du support humain.



Configuration matérielle requise

Le tableau suivant répertorie les composants matériels requis pour implémenter la solution. Ils peuvent varier selon la mise en œuvre de la solution et les besoins du client.

Tests de latence de réponse	Temps (millisecondes)
Le traitement de données	10
Inférence	10

Ces tests de temps de réponse ont été exécutés sur plus de 50,000 fichiers audio pour 560 conversations. Chaque fichier audio a une taille de ~100 Ko en MP3 et de ~1 Mo lorsqu'il est converti en fichier WAV. L'étape de traitement des données convertit des fichiers MP3 en fichiers WAV. Les étapes d'inférence convertissent les fichiers audio en texte et extraient un sentiment du texte. Ces étapes sont toutes indépendantes les unes des autres et peuvent être parallélisées pour accélérer le processus.

En tenant compte de la latence du transfert des données entre les magasins, les responsables devraient être en mesure de voir les mises à jour de l'analyse des sentiments en temps réel dans une seconde de la fin de la phrase.

Matériel NVIDIA RIVA

Sous-jacent	De formation
OS	Linux x86_64
Mémoire GPU (ASR)	Modèles de diffusion en continu : ~5600 Mo modèles sans diffusion en continu : environ 3100 Mo
Mémoire GPU (NLP)	Environ 500 Mo par modèle BERT

Matériel du kit d'outils NVIDIA TAO

Sous-jacent	De formation
RAM système	32 GO
MÉMOIRE VIVE DU GPU	32 GO
CPU	8 cœurs
GPU	NVIDIA (A100, V100 et RTX 30x0)
SSD	100 GO

Système de stockage Flash

NetApp ONTAP 9

ONTAP 9.9, la dernière génération de logiciel de gestion du stockage de NetApp, permet aux entreprises de moderniser l'infrastructure et de passer à un data Center prêt pour le cloud. Avec des capacités de gestion des données à la pointe du secteur, ONTAP permet de gérer et de protéger les données avec un seul ensemble d'outils, quel que soit leur emplacement. Vous pouvez aussi déplacer vos données librement partout où elles sont nécessaires : la périphérie, le cœur ou le cloud. ONTAP 9.9 comprend de nombreuses fonctionnalités qui simplifient la gestion des données, accélèrent et protègent les données stratégiques, et permettent d'utiliser des fonctionnalités d'infrastructure nouvelle génération dans toutes les architectures de cloud hybride.

Copie et synchronisation NetApp BlueXP

"Copie et synchronisation BlueXP" Est un service NetApp permettant une synchronisation rapide et sécurisée des données, qui vous permet de transférer des fichiers entre des partages de fichiers NFS ou SMB sur site vers l'une des cibles suivantes :

- NetApp StorageGRID
- NetApp ONTAP S3
- NetApp Cloud Volumes Service
- Azure NetApp Files
- Amazon simple Storage Service (Amazon S3)
- Amazon Elastic File System (Amazon EFS)
- Blob d'Azure
- Google Cloud Storage
- IBM Cloud Object Storage

BlueXP Copy and Sync déplace les fichiers où vous le souhaitez, rapidement et de manière sécurisée. Une fois vos données transférées, elles peuvent être utilisées à la fois sur la source et sur la cible. BlueXP Copy and Sync synchronise en continu les données en fonction de votre planification prédéfinie, en déplaçant uniquement les données modifiées. Le temps et l'argent consacrés à la réplication des données sont ainsi réduits. BlueXP Copy and Sync est un outil SaaS (logiciel en tant que service) simple à configurer et à utiliser. Les transferts de données déclenchés par la copie et la synchronisation BlueXP sont effectués par les courtiers de données. Vous pouvez déployer des courtiers de données BlueXP Copy and Sync dans AWS, Azure, Google Cloud Platform ou sur site.

NetApp StorageGRID

La suite de stockage objet Software-defined de StorageGRID prend en charge de nombreux cas d'utilisation de manière transparente dans les environnements multiclouds publics, privés et hybrides. Grâce à des innovations de pointe, NetApp StorageGRID stocke, sécurise, protège et préserve les données non structurées à une utilisation polyvalente, y compris la gestion automatisée du cycle de vie sur de longues périodes. Pour plus d'informations, reportez-vous à la section "NetApp StorageGRID" le site.

Configuration logicielle requise

Les composants logiciels requis pour implémenter cette solution sont répertoriés dans le tableau suivant. Ils peuvent varier selon la mise en œuvre de la solution et les besoins du client.

Machine hôte	De formation
RIVA (anciennement JARVIS)	1.4.0
TAO Toolkit (anciennement Transfer Learning Toolkit)	3.0
ONTAP	9.9.1
SYSTÈME D'EXPLOITATION DGX	5.1
DOTK	2.0.0

Logiciel NVIDIA RIVA

Logiciel	De formation
Docker	Pour 19.02 (avec nvidia-docker installé)>=19.03 si vous n'utilisez pas DGX
Pilote NVIDIA	Plus de 465.19.01 418.40+, 440.33+, 450.51+ et 460.27+ pour les processeurs graphiques du data Center
OS de conteneur	Ubuntu 20.04
CUDA	11.3.0
CcuBLAS	11.5.1.101
Distance	8.2.0.41
NCCL	2.9.6
TensorRT	7.2.3.4
Serveur d'inférence Triton	2.9.0

Logiciel NVIDIA TAO Toolkit

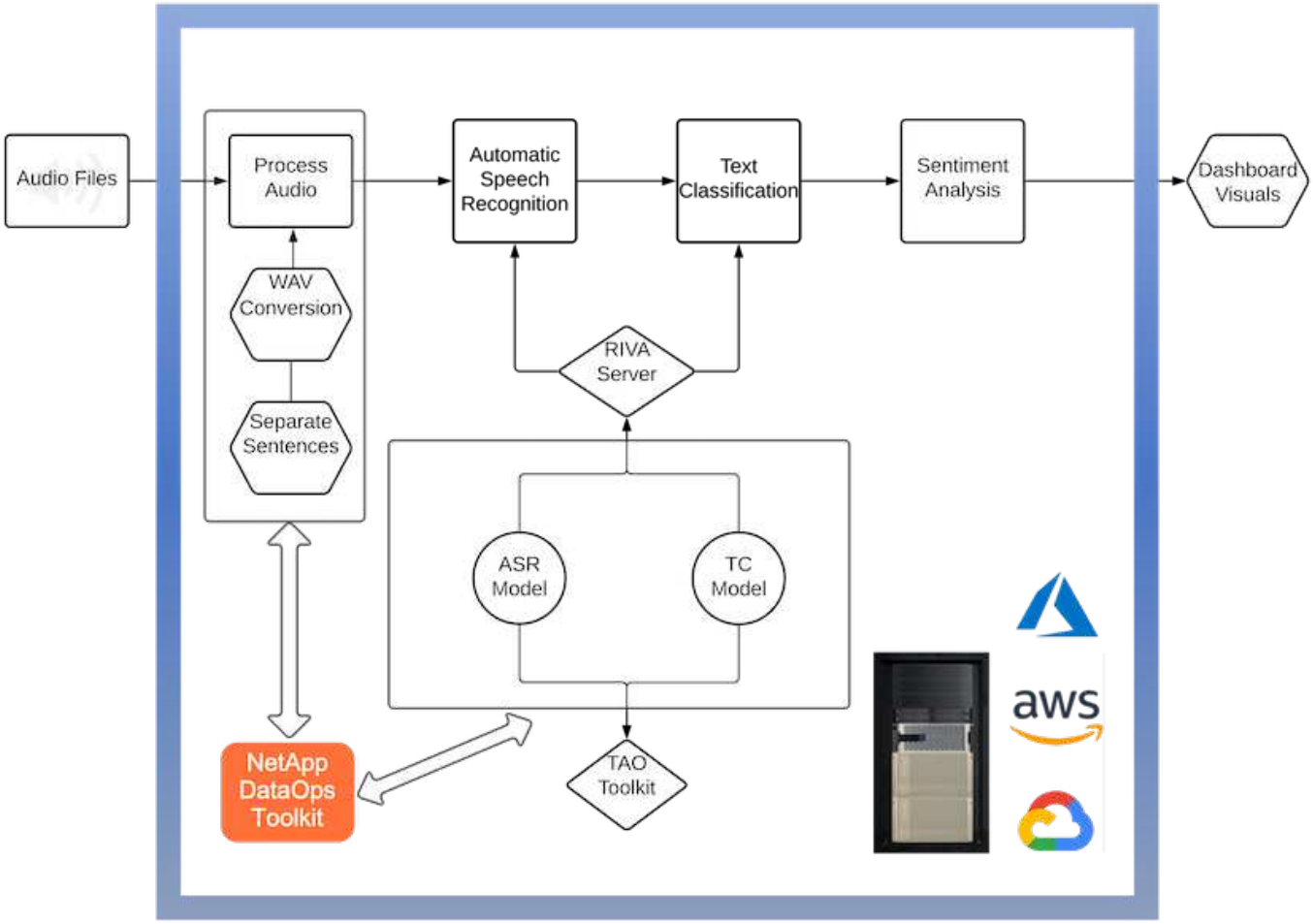
Logiciel	De formation
Ubuntu 18.04 LTS	18.04
python	>=3.6.9
docker-ce	Pour 19.03.5

Logiciel	De formation
docker-API	1.40
kit-conteneur-nvidia	Pour 1.3.0-1
exécution-conteneur-nvidia	3.4.0-1
nvidia-docker2	2.5.0-1
pilote nvidia	Pour 455
python-pip	Pour 21.06
nvidia-pyindex	Dernière version

Détails du cas d'utilisation

Cette solution s'applique aux cas d'utilisation suivants :

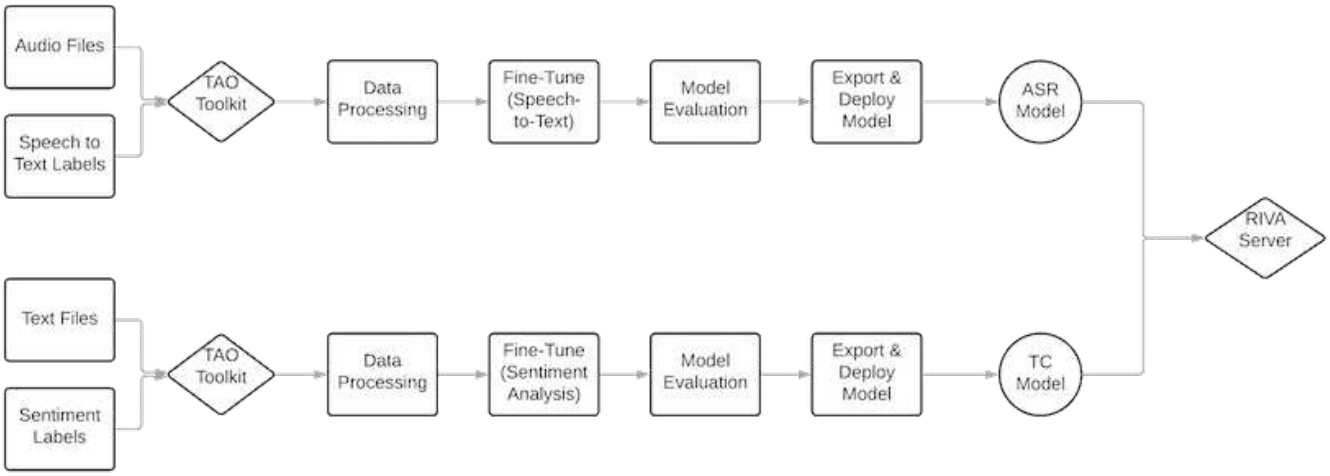
- Parole-à-texte
- Analyse des sentiments



L'utilisation de texte à texte commence par l'acquisition de fichiers audio pour les centres de support. Ce son est ensuite traité pour s'adapter à la structure requise par RIVA. Si les fichiers audio n'ont pas déjà été divisés en unités d'analyse, cela doit être fait avant de transmettre l'audio à RIVA. Une fois le fichier audio traité, il est transmis au serveur RIVA en tant qu'appel API. Le serveur utilise l'un des nombreux modèles qu'il héberge et

renvoie une réponse. Ce message vocal au texte (fait partie de la reconnaissance vocale automatique) renvoie une représentation textuelle de l'audio. À partir de là, le pipeline passe à la partie analyse de sentiment.

Pour l'analyse des sentiments, la sortie du texte de la reconnaissance vocale automatique sert d'entrée à la classification de texte. La classification de texte est le composant NVIDIA permettant de classer le texte dans n'importe quelle catégorie. Les catégories de sentiments varient de positif à négatif pour les conversations du centre de support. Les performances des modèles peuvent être évaluées à l'aide d'un jeu de retenue pour déterminer la réussite de l'étape de réglage précis.



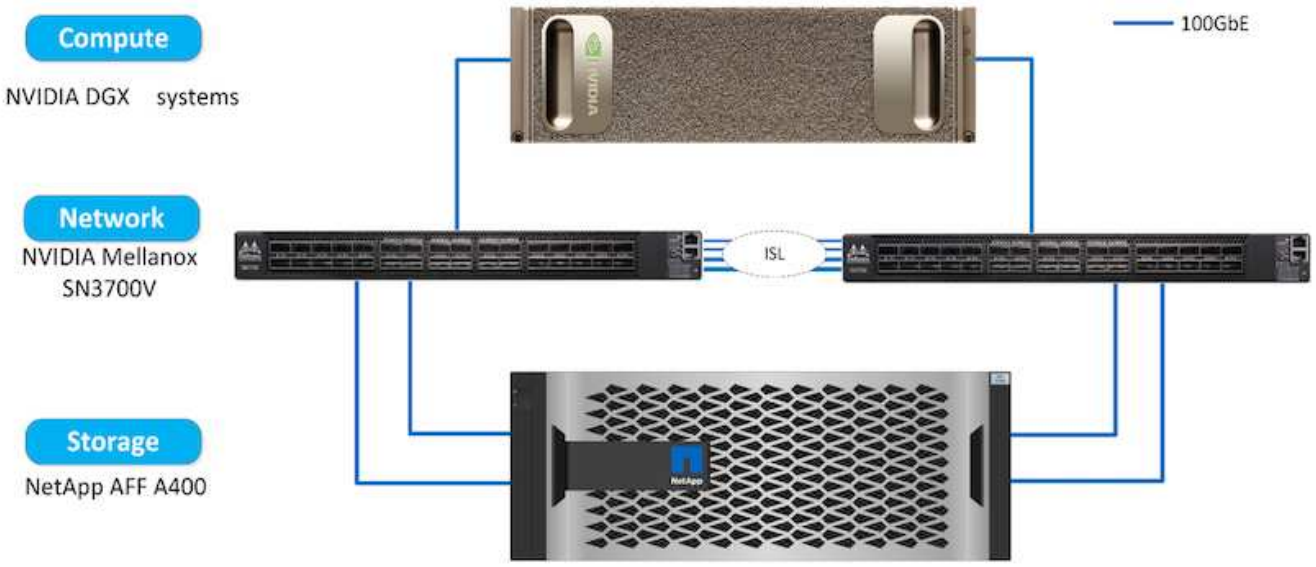
Un pipeline similaire est utilisé à la fois pour l'analyse de la parole au texte et de l'opinion dans la boîte à outils TAO. La différence majeure réside dans l'utilisation d'étiquettes nécessaires au réglage précis des modèles. Le pipeline TAT Toolkit commence par le traitement des fichiers de données. Puis les modèles pré-entraînés (provenant du "Catalogue NVIDIA NGC") sont affinée à l'aide des données du centre de support. Les modèles optimisés sont évalués en fonction des metrics de performance correspondants. S'ils sont plus performants que les modèles pré-entraînés, ils sont déployés sur le serveur RIVA.

Considérations relatives à la conception

Cette section décrit les différents composants de cette solution à prendre en compte lors de sa conception.

Conception du réseau et des ressources de calcul

En fonction des restrictions liées à la sécurité des données, toutes les données doivent rester dans l'infrastructure du client ou dans un environnement sécurisé.



Conception du stockage

Le kit NetApp DataOps sert de service principal pour la gestion des systèmes de stockage. Le kit DataOps est une bibliothèque Python qui facilite le clonage quasi instantané d'un volume de données ou d'un espace de travail JupyterLab par exemple, des développeurs, des data Scientists, et des ingénieurs DevOps, Et le snapshots quasi-instantané d'un volume de données ou d'un espace de travail JupyterLab à des fins de traçabilité ou de base. Cette bibliothèque Python peut fonctionner comme un utilitaire de ligne de commande ou une bibliothèque de fonctions pouvant être importées dans n'importe quel programme Python ou ordinateur portable Jupyter.

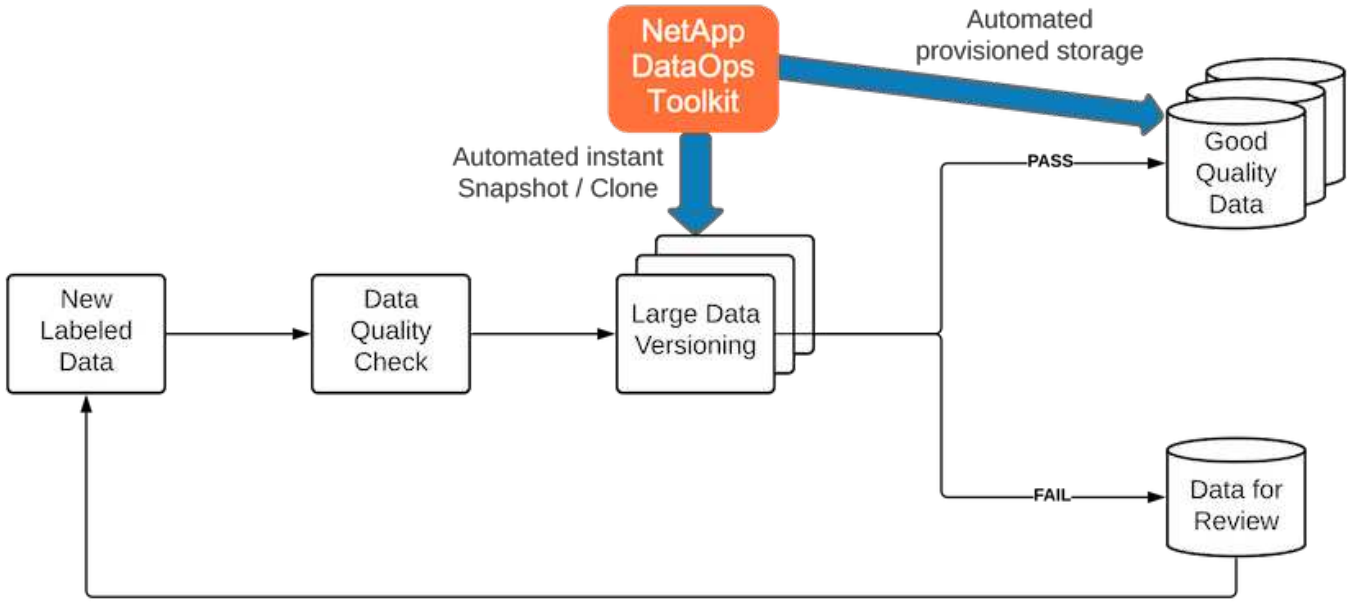
Meilleures pratiques RIVA

NVIDIA présente plusieurs avantages généraux "de meilleures pratiques sur les données" Pour utiliser RIVA :

- **Utilisez des formats audio sans perte si possible.** l'utilisation de codecs avec perte tels que MP3 peut réduire la qualité.
- **Augmenter les données de formation.** l'ajout de bruit de fond aux données de formation audio peut d'abord diminuer la précision tout en augmentant la robustesse.
- **Limiter la taille du vocabulaire en cas d'utilisation de texte gratté.** de nombreuses sources en ligne contiennent des fautes de frappe ou des pronoms auxiliaires et des mots peu communs. La suppression de ces éléments peut améliorer le modèle de langue.
- **Utilisez un taux d'échantillonnage minimum de 16 kHz si possible.** toutefois, essayez de ne pas rééchantillonner, car cela diminue la qualité audio.

Outre ces meilleures pratiques, il est nécessaire que les clients privilégient la collecte d'un dataset d'échantillon représentatif, avec des étiquettes précises pour chaque étape du pipeline. En d'autres termes, le dataset exemple doit refléter de manière proportionnelle les caractéristiques spécifiées présentées dans un dataset cible. De la même façon, les annotations du dataset ont la responsabilité d'équilibrer la précision et la vitesse d'étiquetage, de sorte que la qualité et la quantité des données soient optimisées. Par exemple, cette solution de centre de support nécessite des fichiers audio, du texte étiqueté et des étiquettes de sentiment. La nature séquentielle de cette solution signifie que les erreurs du début du pipeline sont propagées jusqu'à la fin Si les fichiers audio sont de mauvaise qualité, les transcriptions de texte et les libellés de sentiment seront aussi.

Cette propagation d'erreur s'applique également aux modèles entraînés sur ces données. Si les prévisions de sentiment sont exactes à 100 % mais que le modèle de parole à texte fonctionne mal, alors le pipeline final est limité par les transcriptions audio à texte initiales. Il est essentiel que les développeurs considèrent les performances de chaque modèle individuellement et comme un composant d'un pipeline plus vaste. Dans ce cas particulier, l'objectif final est de développer un pipeline capable de prédire avec précision le sentiment. Par conséquent, la mesure globale sur laquelle évaluer le pipeline est la précision des sentiments, que la transcription de la parole vers le texte affecte directement.



Le kit NetApp DataOps complète le pipeline de contrôle de la qualité des données grâce à la technologie de clonage quasi-instantané. Chaque fichier étiqueté doit être évalué et comparé aux fichiers étiquetés existants. La distribution de ces contrôles de qualité sur différents systèmes de stockage des données permet de réaliser ces vérifications rapidement et efficacement.

Déploiement de l'analyse des sentiments du centre de support

Le déploiement de la solution implique les composants suivants :

1. Kit NetApp DataOps
2. Configuration NGC
3. Serveur NVIDIA RIVA
4. Kit d'outils NVIDIA TAO
5. Exporter des modèles TAO vers RIVA

Pour effectuer le déploiement, procédez comme suit :

Kit NetApp DataOps : analyse des sentiments des centres de support

Pour utiliser le "Kit NetApp DataOps", effectuez les opérations suivantes :

1. PIP installer la boîte à outils.

```
python3 -m pip install netapp-dataops-traditional
```

2. Configuration de la gestion des données

```
netapp_dataops_cli.py config
```

Configuration NGC : analyse du sentiment du centre de support

Pour configurer "NVIDIA NGC", effectuez les opérations suivantes :

1. Télécharger le contrôleur NGC.

```
wget -O ngccli_linux.zip
https://ngc.nvidia.com/downloads/ngccli_linux.zip && unzip -o
ngccli_linux.zip && chmod u+x ngc
```

2. Ajoutez votre répertoire actuel au chemin d'accès.

```
echo "export PATH=\"\$PATH:$(pwd)\"" >> ~/.bash_profile && source
~/.bash_profile
```

3. Vous devez configurer l'interface de ligne de commandes NGC pour que vous puissiez exécuter les commandes. Entrez la commande suivante, y compris votre clé API lorsque vous y êtes invité.

```
ngc config set
```

Pour les systèmes d'exploitation qui ne sont pas basés sur Linux, rendez-vous sur "ici".

Serveur NVIDIA RIVA : analyse des sentiments des centres de support

Pour configurer "NVIDIA RIVA", effectuez les opérations suivantes :

1. Téléchargez les fichiers RIVA de NGC.

```
ngc registry resource download-version
nvidia/riva/riva_quickstart:1.4.0-beta
```

2. Initialiser LA configuration RIVA (`riva_init.sh`).
3. Démarrez LE serveur RIVA (`riva_start.sh`).
4. Démarrez le client RIVA (`riva_start_client.sh`).

5. Dans le client RIVA, installez la bibliothèque de traitement audio ("FFMPEG")

```
apt-get install ffmpeg
```

6. Démarrez le "Jupyter" serveur.
7. Exécutez le bloc-notes RIVA Inférence Pipeline.

Kit NVIDIA TAO : analyse des sentiments du centre de support

Pour configurer la boîte à outils NVIDIA TAO, procédez comme suit :

1. Préparez et activez un "virtualisé" Pour la boîte à outils TAO.
2. Installer le "packages requis".
3. Tirez manuellement l'image utilisée pendant l'entraînement et le réglage précis.

```
docker pull nvcr.io/nvidia/tao/tao-toolkit-pyt:v3.21.08-py3
```

4. Démarrez le "Jupyter" serveur.
5. Exécutez l'ordinateur portable TAT Fine Tuning.

Exporter des modèles TAO vers RIVA : prendre en charge l'analyse des sentiments des centres de support

À utiliser "Modèles TAO Toolkit à RIVA", effectuez les opérations suivantes :

1. Enregistrez des modèles dans le portable TAT Fine Tuning.
2. Copier des modèles entraînés TAO dans le répertoire de modèle RIVA.
3. Démarrez LE serveur RIVA (`riva_start.sh`).

Les obstacles au déploiement

Voici quelques points à garder à l'esprit lors du développement de votre propre solution :

- Le kit NetApp DataOps est installé en premier pour assurer le fonctionnement optimal du système de stockage des données.
- NVIDIA NGC doit être installé avant toute autre chose, car il authentifie le téléchargement des images et des modèles.
- RIVA doit être installé avant la boîte à outils TAO. L'installation DE RIVA configure le démon docker pour extraire les images si nécessaire.
- DGX et docker doivent avoir un accès à Internet pour télécharger les modèles.

Résultats de la validation

Comme mentionné dans la section précédente, les erreurs sont propagées dans l'ensemble du pipeline chaque fois qu'au moins deux modèles d'apprentissage machine sont exécutés dans l'ordre. Pour cette solution, le sentiment de la phrase est le facteur le

plus important dans la mesure du niveau de risque stock de l'entreprise. Le modèle de parole à texte, bien qu'essentiel au pipeline, sert d'unité de prétraitement avant que les sentiments ne puissent être prédits. Ce qui compte vraiment, c'est la différence de sentiment entre les phrases de vérité au sol et les phrases prévues. Cela sert de proxy pour le taux d'erreur de mot (WER). La précision de la parole au texte est importante, mais le WER n'est pas directement utilisé dans la mesure finale du pipeline.

```
PIPELINE_SENTIMENT_METRIC = MEAN(DIFF(GT_sentiment, ASR_sentiment))
```

Ces mesures de sentiment peuvent être calculées pour le score F1, le rappel et la précision de chaque phrase. Les résultats peuvent ensuite être agrégés et affichés dans une matrice de confusion, ainsi que les intervalles de confiance pour chaque mesure.

L'apprentissage par transfert améliore les performances des modèles pour une fraction des besoins de données, des délais d'entraînement et des coûts. Les modèles optimisés doivent également être comparés à leurs versions de base afin de garantir que l'apprentissage par transfert améliore les performances au lieu de l'associer. En d'autres termes, le modèle affinée doit optimiser les performances des données de centre de support par rapport au modèle pré-entraînés.

Évaluation du pipeline

Cas de test	Détails
Numéro de test	Mesure du sentiment du pipeline
Conditions préalables au test	Modèles optimisés pour les modèles d'analyse de la voix au texte et des sentiments
Résultat attendu	La mesure de sentiment du modèle affinée est plus performante que le modèle pré-pré-pré-entraînés d'origine.

Mesure du sentiment du pipeline

1. Calculez la mesure du sentiment pour le modèle de référence.
2. Calculez la mesure du sentiment pour le modèle affinée.
3. Calculez la différence entre ces mesures.
4. Faire la moyenne des différences entre toutes les phrases.

Vidéos et démonstrations

Deux ordinateurs portables contiennent le pipeline d'analyse des sentiments : ""“Support-Center-Model-Transfer-Learning-and-Fine-Tuning.ipynb”" et "« Support-Center-sentiment-analyse-pipeline.ipynb »". Ensemble, ces ordinateurs portables montrent comment développer un pipeline permettant d'ingérer les données du data Center et d'extraire les sentiments à partir de chaque phrase à l'aide de modèles de deep learning les plus sophistiqués sur les données de l'utilisateur.

Centre de soutien - Pipeline d'analyse de sentiment.ipynb

Cet ordinateur portable contient le pipeline RIVA d'inférence pour l'importation de fichiers audio, la conversion en texte et l'extraction de sentiments à utiliser dans un tableau de bord externe. Le jeu de données est automatiquement téléchargé et traité si ce n'est pas déjà fait. La première section du bloc-notes est la parole en texte qui gère la conversion de fichiers audio en texte. La section analyse de sentiment est suivie de la section analyse de sentiment qui extrait des sentiments pour chaque phrase de texte et affiche ces résultats dans un format similaire au tableau de bord proposé.

> Cet ordinateur portable doit être exécuté avant l'entraînement et le réglage précis du modèle car le jeu de données MP3 doit être téléchargé et converti au format approprié.

Call Center - Sentiment Analysis Pipeline

This notebook demonstrates how to build a pipeline for sentiment analysis of call center conversations. The goal of this pipeline is to develop sentiment analysis for use within an external dashboard.

This tutorial will guide you through the use of NVIDIA's RIVA for automatic speech recognition and text classification. This tutorial uses NetApp cloud storage for data storage and a pre-trained RIVA model.

Channels

These are the channels on which RIVA is hosting models.

- speech: `51051`
- voice: `61051`

These channels **must** be aligned with `riva_speech_api_port` and `riva_vision_api_port` within `config.sh`

In [4]:
```
speech_channel = "localhost:51051"
voice_channel = "localhost:61051"
```

Speech-To-Text

Automatic Speech Recognition (ASR) takes as input an audio stream or audio buffer and returns one or more text transcripts, along with additional optional metadata. ASR represents a full speech recognition pipeline that is GPU accelerated with optimized performance and accuracy. ASR supports synchronous and streaming recognition modes.

For more information on NVIDIA RIVA's Automatic Speech Recognition, visit here.

Constants

Use these constants to affect different aspects of this pipeline:

- `DATA_DIR` : base folder where data is stored
- `DATASET_NAME` : name of the call center dataset
- `COMPANY_DATE` : folder name identifying the particular call center conversation

Centre de support - modèle de formation et affinage.ipynb

L'environnement virtuel TAO Toolkit doit être configuré avant d'exécuter l'ordinateur portable (voir la section boîte à outils TAO dans la vue d'ensemble des commandes pour les instructions d'installation).

Cet ordinateur portable s'appuie sur le kit d'outils TAO pour affiner les modèles de deep learning sur les données clients. Comme pour le précédent bloc-notes, celui-ci est divisé en deux sections pour les composants parole-à-texte et analyse de sentiment. Chaque section passe par le traitement des données,

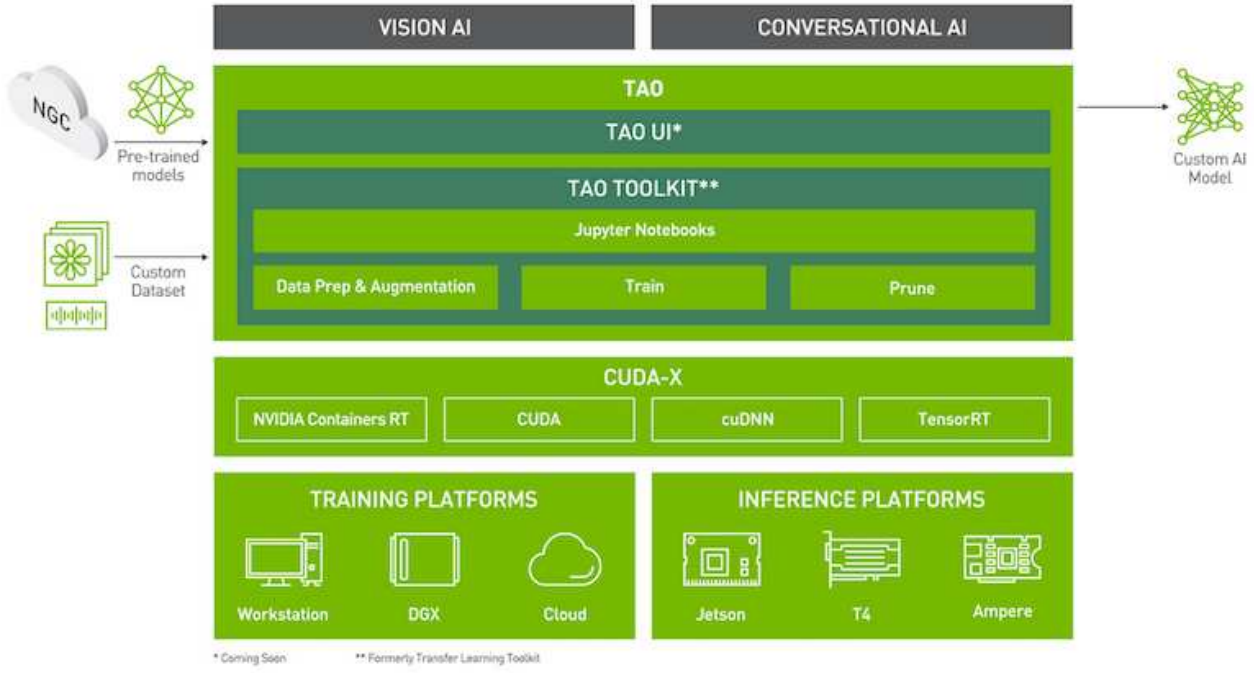
l'entraînement et le réglage des modèles, l'évaluation des résultats et l'exportation des modèles. Enfin, vous avez la fin du déploiement de vos deux modèles plus adaptés à RIVA.

Call Center - Model Transfer Learning and Fine-Tuning

TAO Toolkit is a python based AI toolkit for taking purpose-built pre-trained AI models and customizing them with your own data. Transfer learning extracts learned features from an existing neural network to a new one. Transfer learning is often used when creating a large training dataset is not feasible in order to enhance the base performance of state-of-the-art models.

For this call center solution, the speech-to-text and sentiment analysis models are fine-tuned on call center data to augment the model performance on business specific terminology.

For more information on the TAO Toolkit, please visit here.



Installing necessary dependencies

For ease of use, please install TAO Toolkit inside a python virtual environment. We recommend performing this step first and then launching the notebook from the virtual environment. Please refer to the README for these instructions.

Conclusion

L'expérience client étant de plus en plus considérée comme un champ de bataille essentiel de la concurrence, un centre de support mondial avec intelligence artificielle est devenu un composant essentiel que les entreprises de presque tous les secteurs ne peuvent se permettre de négliger. La solution proposée dans ce rapport technique a été démontrée pour soutenir la prestation de telles expériences client exceptionnelles. Le défi consiste désormais à s'assurer que les entreprises prennent des mesures pour moderniser leur infrastructure et leurs workflows d'IA.

Les meilleures implémentations d'IA dans le service client ne sont pas à remplacer des agents humains. L'IA leur permet plutôt de créer des expériences client exceptionnelles grâce à l'analyse des sentiments en temps réel, à l'escalade des conflits et à un calcul multimodal affectueux pour détecter des indices verbaux, non verbaux et faciaux avec lesquels des modèles d'IA complets peuvent formuler des recommandations à grande échelle et compléter les lacunes d'un agent humain. L'IA permet également une meilleure correspondance entre un client particulier et les agents disponibles actuellement. Grâce à l'IA, les entreprises peuvent extraire un sentiment précieux des clients concernant leurs pensées et impressions sur les produits, services et image de marque du fournisseur.

La solution peut également être utilisée pour construire des données de séries chronologiques destinées aux agents de support afin de servir de mesure objective d'évaluation des performances. Les enquêtes classiques sur la satisfaction des clients manquent souvent de réponses suffisantes. En recueillant des sentiments à long terme des employés et des clients, les employeurs peuvent prendre des décisions éclairées concernant le rendement des agents de soutien.

L'association de NetApp, des SFL Scientific, des frameworks d'orchestration open source et de NVIDIA rassemble les dernières technologies en tant que services gérés, tout en offrant une grande flexibilité pour accélérer l'adoption des technologies et améliorer le délai de mise sur le marché des nouvelles applications d'IA/ML. Ces services avancés sont fournis sur site, facilement ports pour les environnements cloud natifs, ainsi que pour les architectures de déploiement hybride.

Où trouver des informations complémentaires

Pour en savoir plus sur les informations données dans ce livre blanc, consultez ces documents et/ou sites web :

- Démonstrations interactives en 3D

  "www.netapp.com/ai"

- Entrez en contact directement avec un spécialiste de l'IA NetApp

  "https://www.netapp.com/artificial-intelligence/"

- Description de la solution NVDIA base Command Platform avec NetApp

  https://www.netapp.com/pdf.html?item=/media/32792-DS-4145-NVIDIA-Base-Command-Platform-with-NetApp.pdf

- Infographie NetApp pour l'IA 10 : bonnes raisons de choisir NetApp

  "https://www.netapp.com/us/media/netapp-ai-10-good-reasons.pdf"

- L'IA dans le domaine de la santé : l'apprentissage profond pour identifier les lésions de la COVID-19 dans les analyses CT pulmonaires

  https://www.netapp.com/pdf.html?item=/media/31240-WP-7342.pdf

- L'IA dans le domaine de la santé : surveillance de l'utilisation du masque facial dans les établissements de santé. Livre blanc

  https://www.netapp.com/pdf.html?item=/media/37490-NA-611-Monitoring-face-mask-usage-in-healthcare-settings.pdf

- Rapport technique sur l'IA dans le domaine de la santé : imagerie diagnostique

https://www.netapp.com/pdf.html?item=/media/7395-tr4811.pdf

- L'IA pour le Retail : NetApp transversaux à l'IA avec NVIDIA RIVA

  "https://docs.netapp.com/us-en/netapp-solutions/ai/cainvidia_executive_summary.html"

- Description de la solution NetApp ONTAP ai

  https://www.netapp.com/pdf.html?item=/media/6736-sb-3939.pdf

- Description de la solution du kit DataOps NetApp

  https://www.netapp.com/pdf.html?item=/media/21480-SB-4111-1220-NA-Data-Science-Toolkit.pdf

- Description de la solution NetApp ai Control plane

  https://www.netapp.com/pdf.html?item=/media/6737-sb-4055.pdf

- Transformation du secteur grâce à l'eBook Data Drive IA

  "https://www.netapp.com/us/media/na-337.pdf"

- Description de la solution d'IA NetApp EF-Series

  https://www.netapp.com/pdf.html?item=/media/26708-SB-4136-NetApp-AI-E-Series.pdf

- Description de la solution NetApp ai et Lenovo ThinkSystem for ai Inférence

  https://www.netapp.com/pdf.html?item=/media/25316-SB-4129.pdf

- Description des solutions NetApp pour l'IA et le ML Lenovo ThinkSystem pour les entreprises

  https://www.netapp.com/pdf.html?item=/media/25317-SB-4128.pdf

- NetApp et NVIDIA : redéfinir le contenu du possible avec la vidéo d'IA

  https://www.youtube.com/watch?v=38xw65SteUc

Formation distribuée dans Azure - prévision de taux par clic

Tr-4904 : formation distribuée dans Azure - prévision de taux par clic

Rick Huang, Verron Martina, Muneer Ahmad, NetApp

Le travail d'un data Scientist doit se concentrer sur l'entraînement et l'ajustement des modèles de machine learning (ML) et d'intelligence artificielle (IA). Toutefois, selon une étude menée par Google, les data Scientists consacrent environ 80 % de leur temps à comprendre comment exploiter leurs modèles pour les applications d'entreprise et leur fonctionnement à grande échelle.

Pour gérer des projets d'IA et DE ML de bout en bout, il faut avoir une vision plus large des composants de l'entreprise. Bien que le DevOps ait pris en charge la définition, l'intégration et le déploiement, ces types de composants ciblent LES opérations DE ML un flux similaire qui inclut les projets d'IA/ML. Pour découvrir ce

qu’un pipeline IA/ML de bout en bout touche dans l’entreprise, consultez la liste suivante de composants :

- Stockage
- Mise en réseau
- Les bases de données
- Systèmes de fichiers
- Conteneurs
- Intégration continue et pipeline de déploiement continu (ci/CD)
- Environnement de développement intégré (IDE)
- Sécurité
- Règles d’accès aux données
- Sous-jacent
- Le cloud
- Virtualisation
- Bibliothèques et jeux d’outils de science des données

Public visé

L’univers de la science des données touche plusieurs disciplines au NIVEAU DE L’INFORMATIQUE et des activités :

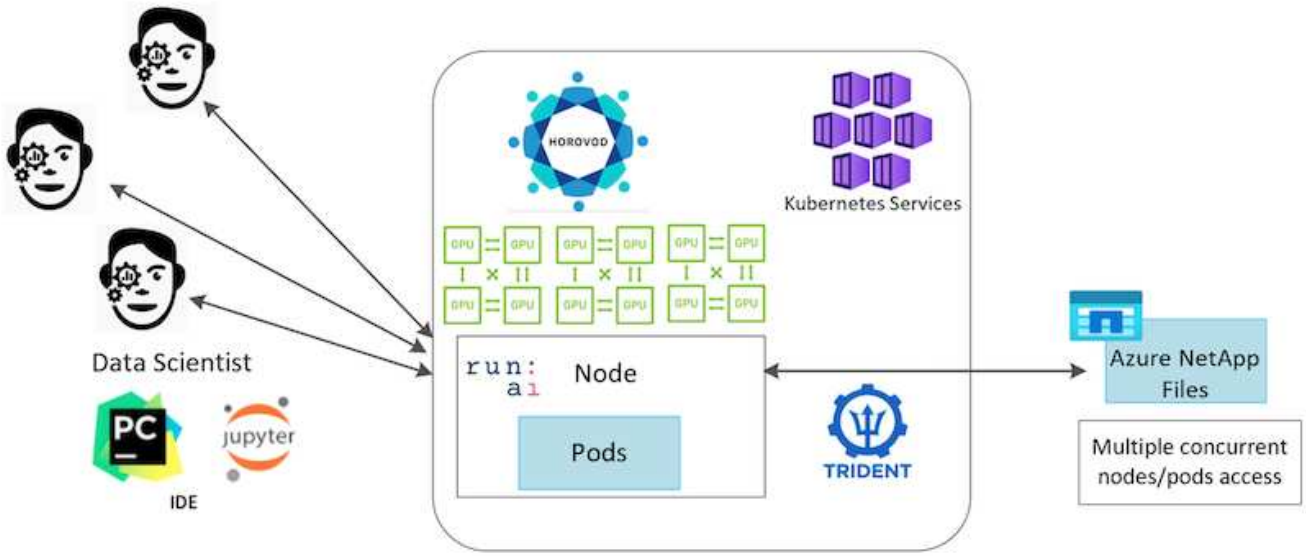
- Les data Scientists doivent donc pouvoir utiliser leurs outils et leurs bibliothèques de choix.
- L’ingénieur doit savoir comment les données circulent et où elles résident.
- Un ingénieur DevOps doit disposer des outils nécessaires pour intégrer les nouvelles applications d’IA et DE ML dans son pipeline ci/CD.
- Les administrateurs et architectes du cloud doivent pouvoir configurer et gérer les ressources Azure.
- Les utilisateurs professionnels veulent avoir accès aux applications d’IA et DE ML.

Dans ce rapport technique, nous décrirons comment Azure NetApp Files, RAPIDS IA, DASK et Azure aident chacun de ces rôles à générer de la valeur pour le business.

Présentation de la solution

Cette solution suit le cycle de vie d’une application d’IA/DE ML. Nous commençons par le travail des data Scientists pour définir les différentes étapes requises pour préparer les données et entraîner les modèles. Grâce À RAPIDS on DASK, nous procédons à une formation distribuée dans le cluster Azure Kubernetes Service (AKS) afin de réduire considérablement le temps d’entraînement par rapport à l’approche conventionnelle Python scikit d’apprentissage. Pour terminer le cycle complet, nous intégrons le pipeline à Azure NetApp Files.

Azure NetApp Files offre plusieurs tiers de performance. Les clients peuvent commencer avec un niveau standard, puis évoluer horizontalement et verticalement jusqu’à un niveau hautes performances sans interruption, sans déplacer aucune donnée. Ainsi, les data Scientists peuvent entraîner des modèles à grande échelle sans problèmes de performances, en évitant les silos de données sur le cluster, comme l’illustre la figure ci-dessous.



Présentation de la technologie

Cette page présente la technologie utilisée dans cette solution.

Microsoft et NetApp

Depuis mai 2019, Microsoft propose un service de portail Azure propriétaire natif pour les services de fichiers NFS et SMB d'entreprise basés sur la technologie NetApp ONTAP. Ce développement est régi par un partenariat stratégique entre Microsoft et NetApp, et étend encore davantage la portée des services de données ONTAP à Azure.

Azure NetApp Files

Le service Azure NetApp Files est un service de stockage de fichiers hautes performances et mesuré. Azure NetApp Files prend en charge tous les types de charges de travail et est extrêmement disponible par défaut. Vous pouvez sélectionner des niveaux de service et de performances et configurer des copies Snapshot via le service. Azure NetApp Files est un service Azure propriétaire pour la migration et l'exécution des workloads de fichiers d'entreprise les plus exigeants dans le cloud, y compris les bases de données, SAP et les applications de calcul haute performance sans modifier le code.

Et présente plusieurs avantages pour les SERVICES IT :

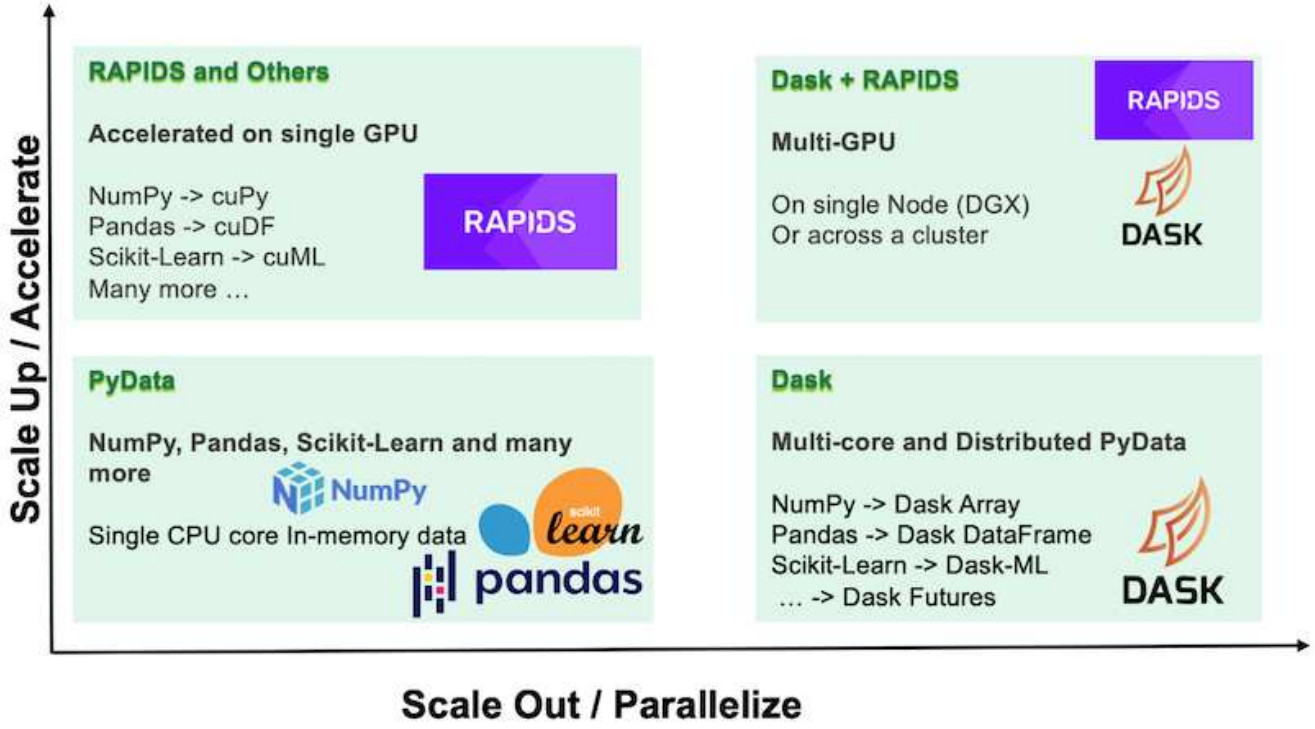
- Simplifie la conception
- Évolutivité indépendante des ressources de calcul et de stockage
- Possibilité de faire évoluer de manière fluide une infrastructure initiale de petite taille
- Propose plusieurs tiers de stockage pour répondre à des exigences variées de coûts et de performance

Présentation de DASK et NVIDIA RAPIDS

DASK est un outil informatique parallèle open source qui permet de faire évoluer les bibliothèques Python sur plusieurs machines et d'accélérer le traitement de grandes quantités de données. Il fournit une API similaire aux bibliothèques Python conventionnelles à thread unique, telles que Pandas, NumPy et scikit-Learn. Par conséquent, les utilisateurs Python natifs n'ont pas à modifier un grand nombre de leur code existant en vue d'utiliser les ressources sur l'ensemble du cluster.

NVIDIA RAPIDS est une suite de bibliothèques open source qui permet d'exécuter facilement les workflows D'AM et d'analytique de bout en bout sur les GPU. Avec DASK, il vous permet de passer facilement d'une station de travail GPU (évolutivité verticale) à des clusters multi-nœuds et multi-processeurs graphiques (évolutivité horizontale).

Pour le déploiement de DASK sur un cluster, vous pouvez utiliser Kubernetes pour l'orchestration des ressources. Vous pouvez également augmenter ou réduire les nœuds workers selon les besoins process, ce qui peut vous aider à optimiser la consommation des ressources du cluster, comme le montre la figure suivante.



Configuration logicielle requise

Le tableau suivant répertorie la configuration logicielle requise pour cette solution.

Logiciel	Version
Service Azure Kubernetes	1.18.14
Image conteneur RAPIDS et Dapose	Référentiel: "Rapidsai/rapidsai" tag: 0.17-cuda11.0-runtime-ubuntu18.04
NetApp Trident	20.01.1
Gouvernail	3.0.0

Ressources cloud nécessaires

Cette page décrit la configuration des ressources cloud pour Azure NetApp Files.

Configurez Azure NetApp Files

Configurez Azure NetApp Files comme décrit à la section "Démarrage rapide : configurez Azure NetApp Files et créez un volume NFS".

Vous pouvez passer outre la section « Créer un volume NFS pour Azure NetApp Files », car vous allez créer des volumes via Trident. Avant de continuer, procédez comme suit :

1. Inscrivez-vous à Azure NetApp Files et au fournisseur de ressources NetApp (via le shell Azure) ("lien").
2. Créer un compte dans Azure NetApp Files ("lien").
3. Configurez un pool de capacité (un minimum de 4 To Standard ou Premium, selon vos besoins) ("lien").le tableau suivant répertorie les configurations réseau requises pour la configuration dans le nuage. Le cluster DASK et Azure NetApp Files doivent se trouver sur le même réseau virtuel Azure (vnet) ou un réseau vnet pêche.

Ressources	Type/version
Service Azure Kubernetes	1.18.14
Nœud agent	3 x Standard_DS2_v2
Nœud de processeur graphique	3 x Standard_NC6s_v3
Azure NetApp Files	Pool de capacité standard
Capacité en To	4

Récapitulatif de l'utilisation des prédictions de taux par clic

Il s'agit d'un cas d'utilisation basé sur le cloud public "Fichiers de clic sur un téraoctet" jeu de données de "Criteo ai Lab". Avec les récentes avancées des plates-formes ET des applications DE ML, une grande attention se porte maintenant sur l'apprentissage à grande échelle. Le taux de clics (CTR) est défini comme le nombre moyen de clics par cent impressions de publicités en ligne (exprimé en pourcentage). Elle est largement adoptée comme indicateur clé dans différents secteurs d'activité et champs d'application, notamment le marketing digital, la vente au détail, l'e-commerce et les fournisseurs de services. Voici quelques exemples d'utilisation de CTR comme mesure importante pour le trafic potentiel des clients :

- **Marketing numérique:** in "Google Analytics", CTR peut être utilisé pour évaluer la manière dont bien un annonceur ou les mots-clés, annonces, et les listes libres sont en cours de réalisation. Un CTR élevé est une bonne indication que les utilisateurs trouvent vos annonces et listes utiles et pertinentes. CTR contribue également à votre mot clé CTR attendu, qui est un composant de "Rang d'annonce".
- * E-commerce:* en plus de tirer parti "Google Analytics", il y a au moins quelques statistiques de visiteurs dans un back-end de commerce électronique. Bien que ces statistiques ne semblent pas utiles à première vue, elles sont généralement faciles à lire et peuvent être plus précises que d'autres informations. Les ensembles de données de première partie composés de ces statistiques sont propriétaires et sont donc les plus pertinents pour les vendeurs, les acheteurs et les plates-formes de commerce électronique. Ces ensembles de données peuvent être utilisés pour établir des bancs d'essai, en comparant les résultats à l'année dernière et à la veille, en élaborant une série chronologique pour une analyse plus approfondie.
- **Retail:** les détaillants Brick-et-mortier peuvent mettre en corrélation le nombre de visiteurs et le nombre de clients avec le CTR. Le nombre de clients est visible depuis leur historique des points de vente. Le CTR

sur les sites Web des détaillants ou le trafic publicitaire peut entraîner les ventes susmentionnées. Les programmes de fidélité sont un autre cas d'utilisation, car les clients redirigés depuis des publicités en ligne ou d'autres sites Web peuvent se joindre pour gagner des récompenses. Les détaillants peuvent acquérir des clients par le biais de programmes de fidélité et enregistrer des comportements liés aux histoires de vente pour élaborer un système de recommandations qui non seulement prédit les comportements d'achat des consommateurs dans différentes catégories, mais également personnalise les bons de réduction et diminue le nombre de départs.

- **Fournisseurs de services:** les entreprises de télécommunications et les fournisseurs de services Internet disposent d'une abondance de données de télémétrie utilisateur de première partie pour des cas d'utilisation instructifs de l'IA, DU ML et de l'analytique. Par exemple, un opérateur télécom peut utiliser quotidiennement ses journaux d'historique de domaine de premier niveau pour affiner les modèles existants afin de générer une segmentation de l'audience à jour, d'anticiper le comportement des clients et de collaborer avec les annonceurs pour placer des publicités en temps réel et améliorer l'expérience en ligne. Dans ce flux de travail de marketing axé sur les données, CTR est une mesure importante pour refléter les conversions.

Dans le cadre du marketing numérique, "Criteo Terabyte cliquez sur journaux" Sont désormais le dataset de référence pour évaluer l'évolutivité des plateformes ET des algorithmes DE ML. En prédisant le taux de clics, un annonceur peut sélectionner les visiteurs qui sont les plus susceptibles de répondre aux annonces, analyser leur historique de navigation, et montrer les annonces les plus pertinentes en fonction des intérêts de l'utilisateur.

La solution proposée dans ce rapport technique présente les avantages suivants :

- Avantages de Azure NetApp Files dans le cadre d'une formation distribuée ou à grande échelle
- RAPIDES, traitement de données CUDA (cuDF, cuPy, etc.) et algorithmes DE ML (cuML)
- Le cadre informatique parallèle DASK pour la formation distribuée

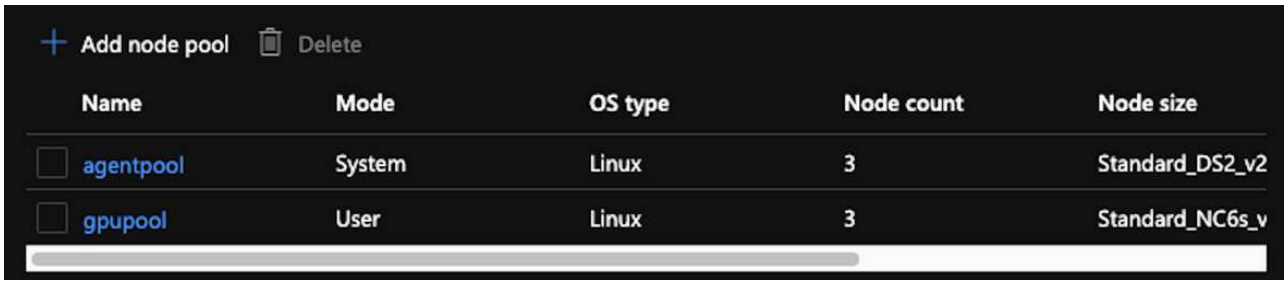
Un workflow complet basé SUR RAPIDS dans l'IA et Azure NetApp Files démontre l'amélioration drastique du temps d'entraînement des modèles forestiers aléatoires de deux ordres de grandeur. Cette amélioration est très significative par rapport à l'approche Pandas classique lors du traitement des journaux de clics du monde réel avec 45 Go de données tabulaires structurées (en moyenne) chaque jour. Cela équivaut à un DataFrame contenant environ vingt milliards de lignes. Dans ce rapport technique, nous allons présenter la configuration de l'environnement de cluster, l'installation de la structure et de la bibliothèque, le chargement et le traitement des données, les méthodes conventionnelles par rapport à la formation distribuée, la visualisation et la surveillance, et comparer les résultats critiques de l'exécution de bout en bout.

Configuration

Installez et configurez le cluster AKS

Pour installer et configurer le cluster AKS, reportez-vous à la page Web "Créer un cluster AKS" puis procédez comme suit :

1. Lors de la sélection du type de nœud (nœuds du système [CPU] ou du worker [GPU]), sélectionnez ce qui suit :
   a. Les nœuds du système primaire doivent être de type Standard DS2v2 (`agentpool` trois nœuds par défaut).
   b. Ajoutez ensuite le nœud de travail Standard_NC6s_v3 pool (trois nœuds minimum) pour le groupe d'utilisateurs (pour les nœuds GPU) nommé `gpupool`.



2. Le déploiement prend entre 5 et 10 minutes. Lorsque l'opération est terminée, cliquez sur Connect to Cluster.
3. Pour vous connecter au cluster AKS nouvellement créé, installez les éléments suivants à partir de votre environnement local (ordinateur portable/pc) :
   a. L'outil de ligne de commande Kubernetes à l'aide de "Instructions fournies pour votre système d'exploitation"
   b. L'interface de ligne de commandes Azure, comme décrit dans le document, "Installez l'interface de ligne de commande Azure"
4. Pour accéder au cluster AKS à partir du terminal, entrez `az login` et entrez les informations d'identification.
5. Exécutez les deux commandes suivantes :

```
az account set --subscription xxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxxx
aks get-credentials --resource-group resourcegroup --name aksclustername
```

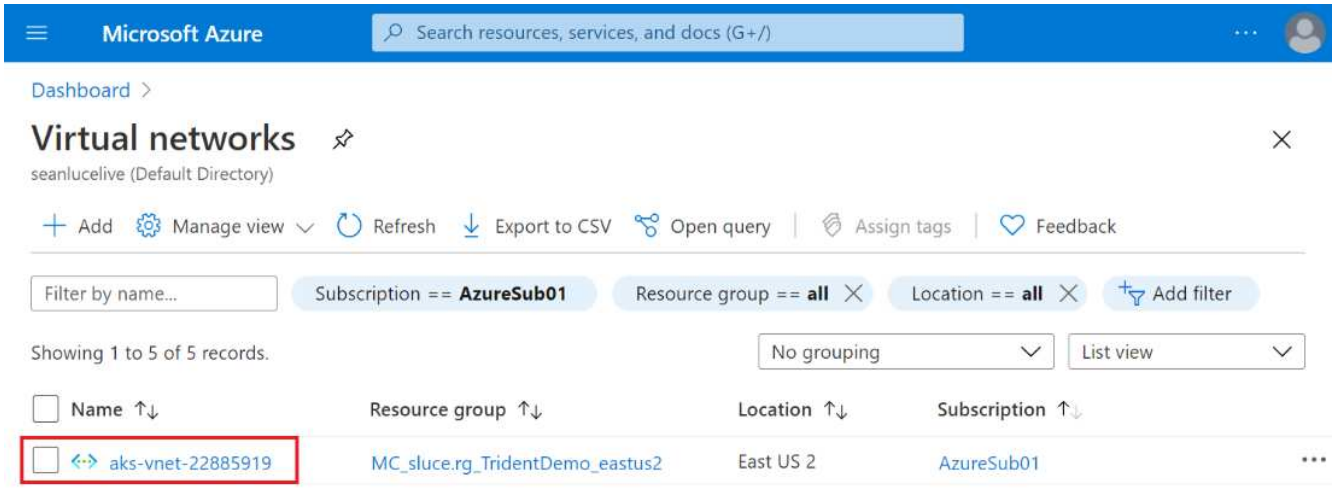
6. Entrez `Azure CLI: kubectl get nodes`.
7. Si les six nœuds sont actifs, comme illustré dans l'exemple suivant, votre cluster AKS est prêt et connecté à votre environnement local

```
verronmartina@verron-mac-0 ~ % kubectl get nodes
NAME                                STATUS   ROLES   AGE   VERSION
aks-agentpool-34613062-vmss000000   Ready    agent   22m   v1.18.14
aks-agentpool-34613062-vmss000001   Ready    agent   22m   v1.18.14
aks-agentpool-34613062-vmss000002   Ready    agent   22m   v1.18.14
aks-gpupool-34613062-vmss000000     Ready    agent   20m   v1.18.14
aks-gpupool-34613062-vmss000001     Ready    agent   20m   v1.18.14
aks-gpupool-34613062-vmss000002     Ready    agent   20m   v1.18.14
verronmartina@verron-mac-0 ~ %
```

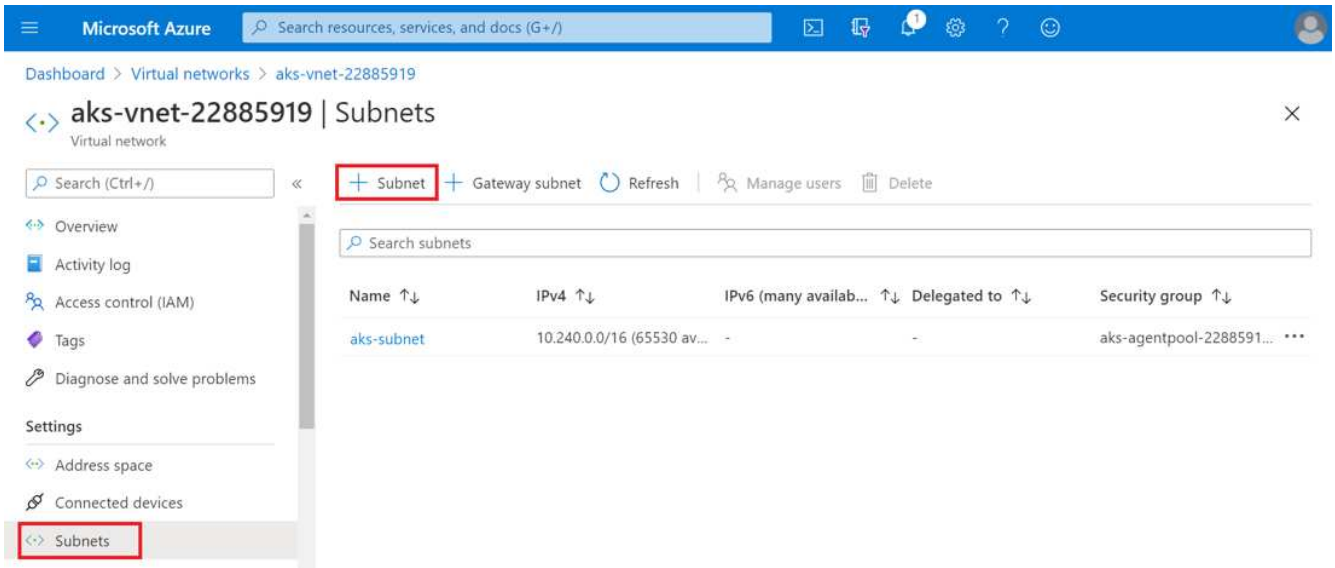
Créez un sous-réseau délégué pour Azure NetApp Files

Pour créer un sous-réseau délégué pour Azure NetApp Files, effectuez les opérations suivantes :

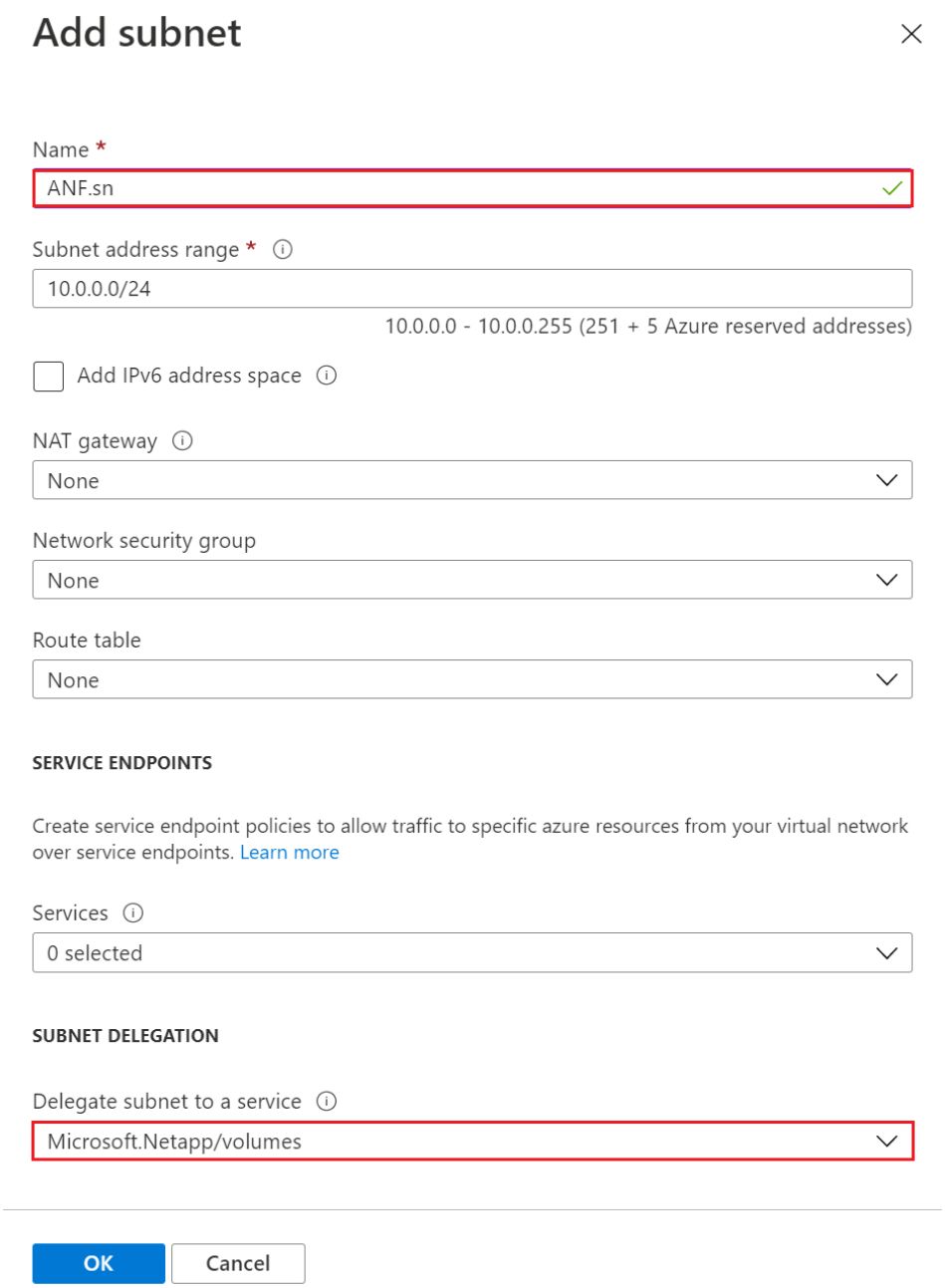
1. Accédez aux réseaux virtuels depuis le portail Azure. Trouvez votre nouveau réseau virtuel. Il doit avoir un préfixe tel que `aks-vnet`.
2. Cliquez sur le nom du vNet.



3. Cliquez sur sous-réseaux et sur +sous-réseau dans la barre d'outils supérieure.



4. Indiquez au sous-réseau un nom tel que `ANF.sn` Et, sous l'en-tête délégation de sous-réseau, sélectionnez `Microsoft.Netapp/volumes`. Ne rien changer. Cliquez sur OK.



Les volumes Azure NetApp Files sont alloués au cluster d'applications et utilisés en tant que demandes de volume persistant dans Kubernetes. Par conséquent, ce processus vous offre la possibilité de les associer à différents services, tels que les ordinateurs portables Jupyter, les fonctions sans serveur, etc.

Les utilisateurs des services peuvent consommer le stockage depuis la plateforme de différentes manières. Dans ce rapport technique, NFSS, les principaux avantages de Azure NetApp Files sont les suivants :

- Possibilité d'utiliser des copies Snapshot pour les utilisateurs.
- Possibilité pour les utilisateurs de stocker d'importants volumes de données sur des volumes Azure NetApp Files.
- Avantages des performances des volumes Azure NetApp Files lors de l'exécution de modèles sur de vastes ensembles de fichiers.

Peer AKS vnet et Azure NetApp Files vnet

Pour passer le réseau VNet AKS à Azure NetApp Files VNet, procédez comme suit :

1. Entrez les réseaux virtuels dans le champ de recherche.
2. Sélectionnez `vnet aks-vnet-name`. Cliquez dessus et saisissez Peerings dans le champ de recherche.
3. Cliquez sur +Ajouter.
4. Saisissez les descripteurs suivants :
   a. Le nom de la liaison de peering est `aks-vnet-name_to_anf`.
   b. Indice de connexion et Azure NetApp Files vnet en tant que partenaire de peering vnet.
   c. Laissez toutes les sections non astérisque avec les valeurs par défaut.
5. Cliquez sur Ajouter.

Pour plus d'informations, voir "Créez, modifiez ou supprimez un peering de réseau virtuel".

Installation de Trident

Pour installer Trident à l'aide de Helm, effectuez les opérations suivantes :

1. Installez Helm (pour obtenir des instructions d'installation, consultez le "source").
2. Téléchargez et extrayez le programme d'installation de Trident 20.01.1.

```
$wget
$tar -xf trident-installer-21.01.1.tar.gz
```

3. Définissez le répertoire sur `trident-installer`.

```
$cd trident-installer
```

4. Copier `tridentctl` dans un répertoire de votre système `$PATH`.

```
$sudo cp ./tridentctl /usr/local/bin
```

5. Installez Trident sur le cluster Kubernetes (K8s) avec Helm ("source") :
   a. Changez le répertoire en `helm` répertoire.

   ```
   $cd helm
   ```

   b. Installation de Trident.

```
$helm install trident trident-operator-21.01.1.tgz --namespace
trident --create-namespace
```

c. Vérifier l'état des pods Trident.

```
$kubectl -n trident get pods
```

Si tous les modules sont opérationnels, Trident est installé et vous pouvez passer à l'étape supérieure.

6. Configurer le système Azure NetApp Files backend et la classe de stockage pour AKS.

a. Créer un principe de service Azure

Le principal service est de la façon dont Trident communique avec Azure pour manipuler vos ressources Azure NetApp Files.

```
$az ad sp create-for-rbac --name ""
```

Le résultat de la commande doit ressembler à l'exemple suivant :

```
{
"appId": "xxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx",
"displayName": "netapptrident",
"name": "",
"password": "xxxxxxxxxxxxxxx.xxxxxxxxxxxxxx",
"tenant": "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxx"
}
```

7. Créez un fichier json backend Trident, par exemple un nom `anf-backend.json`.

8. À l'aide de votre éditeur de texte préféré, renseignez les champs suivants à l'intérieur du `anf-backend.json` fichier :

```
{
    "version": 1,
    "storageDriverName": "azure-netapp-files",
    "subscriptionID": "fakec765-4774-fake-ae98-a721add4fake",
    "tenantID": "fakef836-edc1-fake-bff9-b2d865eefake",
    "clientID": "fake0f63-bf8e-fake-8076-8de91e57fake",
    "clientSecret": "SECRET",
    "location": "westeurope",
    "serviceLevel": "Standard",
    "virtualNetwork": "anf-vnet",
    "subnet": "default",
    "nfsMountOptions": "vers=3,proto=tcp",
    "limitVolumeSize": "500Gi",
    "defaults": {
    "exportRule": "0.0.0.0/0",
    "size": "200Gi"
}
```

9. Remplacez les champs suivants :
   - `subscriptionID`. Votre ID d'abonnement Azure.
   - `tenantID`. Votre ID de locataire Azure à partir des résultats de `az ad sp` à l'étape précédente.
   - `clientID`. Votre AppID à partir de la sortie de `az ad sp` à l'étape précédente.
   - `clientSecret`. Votre mot de passe à partir de la sortie de `az ad sp` à l'étape précédente.
10. Demandez à Trident de créer le back-end Azure NetApp Files dans le système `trident` espace de noms avec `anf-backend.json` comme fichier de configuration :

```
$tridentctl create backend -f anf-backend.json -n trident
```

```
+------------------------+--------------------+--------------------------------------+--------+---------+
|          NAME          |   STORAGE DRIVER   |                 UUID                 | STATE  | VOLUMES |
+------------------------+--------------------+--------------------------------------+--------+---------+
| azurenetappfiles_86181 | azure-netapp-files | 2ca85462-59ac-4946-be05-c03f5575a2ad | online |       0 |
+------------------------+--------------------+--------------------------------------+--------+---------+
```

11. Créer une classe de stockage. Les utilisateurs Kubernetes provisionnent les volumes à l'aide des demandes de volume qui spécifient une classe de stockage par nom. Demandez à K8s de créer une classe de stockage `azurenetappfiles` Qui fait référence au système back-end Trident créé dans l'étape précédente.
12. Créez un YAML (`anf-storage-class.yaml`) fichier pour la classe de stockage et la copie.

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
name: azurenetappfiles
provisioner: netapp.io/trident
parameters:
backendType: "azure-netapp-files"
$kubectl create -f anf-storage-class.yaml
```

13. Vérifiez que la classe de stockage a été créée.

```
kubectl get sc azurenetappfiles
```

```
NAME              PROVISIONER            RECLAIMPOLICY   VOLUMEBINDINGMODE   ALLOWVOLUMEEXPANSION   AGE
azurenetappfiles  csi.trident.netapp.io  Delete          Immediate           false                  98s
```

Mise en place de DASK avec RAPIDS sur AKS à l'aide de Helm

Pour configurer le déploiement de DASK avec RAPIDS sur AKS à l'aide de Helm, procédez comme suit :

1. Créer un espace de noms pour l'installation de DASK avec RAPIDS.

```
kubectl create namespace rapids-dask
```

2. Création d'une demande de volume persistant pour stocker le dataset de taux de clics :
   a. Enregistrez le contenu YAML suivant dans un fichier pour créer une demande de volume persistant.

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: pvc-criteo-data
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 1000Gi
  storageClassName: azurenetappfiles
```

   b. Appliquez le fichier YAML sur votre cluster Kubernetes.

```
kubectl -n rapids-dask apply -f <your yaml file>
```

3. Cloner le `rapidsai git` référentiel ("https://github.com/rapidsai/helm-chart").

```
git clone https://github.com/rapidsai/helm-chart helm-chart
```

4. Modifier `values.yaml` Et inclure le PVC créé précédemment pour les travailleurs et l'espace de travail Jupyter.

   a. Accédez au `rapidsai` répertoire du référentiel.

```
cd helm-chart/rapidsai
```

   b. Mettez à jour le `values.yaml` Fichier et montage du volume à l'aide de la demande de volume persistant.

```
dask:
  …
  worker:
    name: worker
    …
    mounts:
      volumes:
        - name: data
          persistentVolumeClaim:
            claimName: pvc-criteo-data
      volumeMounts:
        - name: data
          mountPath: /data
    …
  jupyter:
    name: jupyter
    …
    mounts:
      volumes:
        - name: data
          persistentVolumeClaim:
            claimName: pvc-criteo-data
      volumeMounts:
        - name: data
          mountPath: /data
    …
```

5. Accédez au répertoire personnel du référentiel et déployez DASK avec trois noeuds de travail sur AKS en utilisant Helm.

```
cd ..
helm dep update rapidsai
helm install rapids-dask --namespace rapids-dask rapidsai
```

Tiers de performance Azure NetApp Files

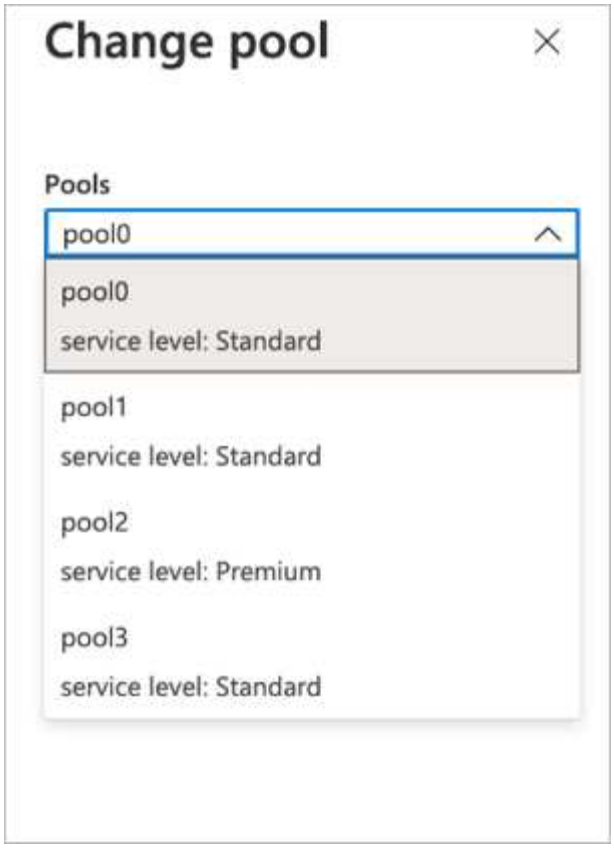
Vous pouvez modifier le niveau de service d'un volume existant en déplaçant ce volume vers un autre pool de capacité qui utilise le niveau de service souhaité pour le volume. Elle permet de commencer avec un petit dataset et un petit nombre de GPU dans le Tier standard, puis de monter en charge horizontalement ou verticalement jusqu'à Premium Tier en fonction du volume de données et du nombre de GPU. Le niveau Premium offre un débit par téraoctet avec un niveau standard, et il est nécessaire de faire évoluer l'infrastructure sans déplacer de données pour modifier le niveau de service d'un volume.

Modification dynamique du niveau de service d'un volume

Pour modifier de manière dynamique le niveau de service d'un volume, procédez comme suit :

1. Sur la page volumes, cliquez avec le bouton droit de la souris sur le volume dont vous souhaitez modifier le niveau de service. Sélectionnez Modifier le pool.

2. Dans la fenêtre change Pool, sélectionnez le pool de capacité vers lequel vous souhaitez déplacer le volume.



3. Cliquez sur OK.

Automatisez les changements au niveau des tiers de performance

Plusieurs options sont disponibles pour automatiser les modifications des tiers de performance :

- Le changement de niveau de service dynamique est toujours dans l'aperçu public à ce moment et n'est pas activé par défaut. Pour activer cette fonctionnalité sur l'abonnement Azure, consultez cette documentation sur la procédure à suivre "Modification dynamique du niveau de service d'un volume".
- Les commandes de modification du pool de volumes de l'interface de ligne de commande Azure sont fournies dans le "documentation sur les modifications apportées au pool de volumes" et dans l'exemple suivant :

```
az netappfiles volume pool-change -g mygroup --account-name myaccname
--pool-name mypoolname --name myvolname --new-pool-resource-id
mynewresourceid
```

- PowerShell : le "Cmdlet Set-AzNetAppFilesVolumePool" Modifie le pool d'un volume Azure NetApp Files et est présenté dans l'exemple suivant :

```
Set-AzNetAppFilesVolumePool
-ResourceGroupName "MyRG"
-AccountName "MyAnfAccount"
-PoolName "MyAnfPool"
-Name "MyAnfVolume"
-NewPoolResourceId 7d6e4069-6c78-6c61-7bf6-c60968e45fbf
```

Cliquez par le biais du traitement des données de prédiction de taux et de l'entraînement des modèles

Bibliothèques de traitement de données et d'entraînement des modèles

Le tableau suivant répertorie les bibliothèques et les structures utilisées pour créer cette tâche. Tous ces composants ont été entièrement intégrés aux contrôles d'accès et de sécurité basés sur les rôles d'Azure.

Bibliothèques/structures	Description
DMASK cuML	Pour QUE LE ML fonctionne sur le GPU, le "Bibliothèque cuML" Donne accès au forfait RAPIDS cuML avec Dask. RAPIDS cuML implémente des algorithmes DE ML courants, notamment la mise en cluster, la réduction de la dimensionnalité et les approches de régression, avec des implémentations basées sur GPU haute performance, offrant des vitesses jusqu'à 100 fois supérieures aux approches basées sur les processeurs.
DASK cuDF	CuDF inclut diverses autres fonctions prenant en charge l'extraction accélérée par GPU, la transformation, la charge (ETL), telles que la sous-définition de données, les transformations, l'encodage à chaud unique, etc. L'équipe DE RAPIDS a "bibliothèque dask-cudf" Cela inclut des méthodes d'aide pour utiliser DASK et cuDF.
Apprendre Scikit	Scikit-Learn fournit des dizaines d'algorithmes et de modèles d'apprentissage machine intégrés, appelés estimateurs. Chacun "estimateur" peut être installé sur certaines données à l'aide de son "ajustement" méthode.

Nous avons utilisé deux ordinateurs portables pour construire les pipelines DE ML à des fins de comparaison. L'une est l'approche classique Pandas scikit d'apprentissage et l'autre, une formation distribuée à RAPIDS et DASK. Chaque ordinateur portable peut être testé individuellement pour connaître les performances en termes de temps et d'échelle. Nous recouvrons chaque ordinateur portable individuellement pour démontrer les avantages de la formation distribuée à L'aide DE RAPIDS et DASK.

Charger Criteo Click Logs Day 15 dans Pandas et former un modèle de forêt aléatoire de scikit

Cette section décrit comment nous avons utilisé Pandas et DASK DataFrames pour charger les données de journaux de clic du dataset Criteo Terabyte. Le cas d'utilisation est pertinent dans la publicité numérique pour les échanges publicitaires afin de créer les profils des utilisateurs en prédisant si les annonces seront cliqué ou si l'échange n'utilise pas un modèle précis dans un pipeline automatisé.

Nous avons chargé les données du jour 15 à partir du jeu de données Click Logs, soit un total de 45 Go. Exécution de la cellule suivante dans le bloc-notes Jupyter `CTR-PandasRF-collated.ipynb` Crée un Pandas DataFrame contenant les 50 premiers millions de lignes et génère un modèle aléatoire d'apprentissage de forêt de scikit.

```
%%time
import pandas as pd
import numpy as np
header = ['col'+str(i) for i in range (1,41)] #note that according to
criteo, the first column in the dataset is Click Through (CT). Consist of
40 columns
first_row_taken = 50_000_000 # use this in pd.read_csv() if your compute
resource is limited.
# total number of rows in day15 is 20B
# take 50M rows
"""
Read data & display the following metrics:
1. Total number of rows per day
2. df loading time in the cluster
3. Train a random forest model
"""
df = pd.read_csv(file, nrows=first_row_taken, delimiter='\t',
names=header)
# take numerical columns
df_sliced = df.iloc[:, 0:14]
# split data into training and Y
Y = df_sliced.pop('col1') # first column is binary (click or not)
# change df_sliced data types & fillna
df_sliced = df_sliced.astype(np.float32).fillna(0)
from sklearn.ensemble import RandomForestClassifier
# Random Forest building parameters
# n_streams = 8 # optimization
max_depth = 10
n_bins = 16
n_trees = 10
rf_model = RandomForestClassifier(max_depth=max_depth,
n_estimators=n_trees)
rf_model.fit(df_sliced, Y)
```

Pour effectuer une prédiction à l'aide d'un modèle de forêt aléatoire entraîné, exécutez le paragraphe suivant dans ce bloc-notes. Nous avons effectué le dernier million de lignes à partir du jour 15 comme jeu de tests pour éviter toute duplication. La cellule calcule également la précision de la prévision, définie comme le pourcentage d'occurrences que le modèle prédit avec précision si un utilisateur clique ou non sur une annonce. Pour passer en revue tous les composants inconnus de cet ordinateur portable, reportez-vous au "documentation officielle d'apprentissage du kit de science".

```
# testing data, last 1M rows in day15
test_file = '/data/day_15_test'
with open(test_file) as g:
    print(g.readline())

# dataFrame processing for test data
test_df = pd.read_csv(test_file, delimiter='\t', names=header)
test_df_sliced = test_df.iloc[:, 0:14]
test_Y = test_df_sliced.pop('col1')
test_df_sliced = test_df_sliced.astype(np.float32).fillna(0)
# prediction & calculating error
pred_df = rf_model.predict(test_df_sliced)
from sklearn import metrics
# Model Accuracy
print("Accuracy:",metrics.accuracy_score(test_Y, pred_df))
```

Charger le jour 15 à DASK et former un modèle forestier aléatoire de DASk cuML

D'une manière similaire à la section précédente, chargez Criteo Click Logs Day 15 dans Pandas et entraînez un modèle de forêt aléatoire d'apprentissage de scikit. Dans cet exemple, nous avons effectué le chargement de DataFrame avec DASK cuDF et formé un modèle de forêt aléatoire dans DASk cuML. Nous avons comparé les différences en termes de temps et d'échelle d'entraînement dans la section ""Comparaison du temps de formation.""

criteo_dAsk_RF.ipynb

Cet ordinateur portable importe `numpy`, `cuml`, et le nécessaire `dask` bibliothèques, comme illustré dans l'exemple suivant :

```
import cuml
from dask.distributed import Client, progress, wait
import dask_cudf
import numpy as np
import cudf
from cuml.dask.ensemble import RandomForestClassifier as cumlDaskRF
from cuml.dask.common import utils as dask_utils
```

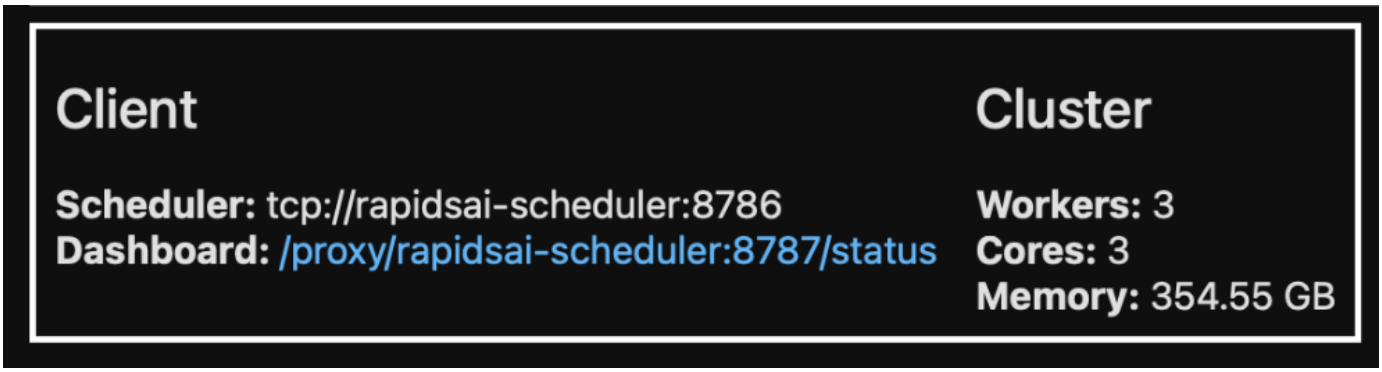
Lancez DASK client().

```
client = Client()
```

Si le cluster est configuré correctement, vous pouvez afficher l'état des nœuds workers.

```
client
workers = client.has_what().keys()
n_workers = len(workers)
n_streams = 8 # Performance optimization
```

Dans notre cluster AKS, l'état suivant s'affiche :



Notez que DASK utilise le paradigme d'exécution paresseux : plutôt que d'exécuter le code de traitement instantanément, DASK construit plutôt un graphe acyclique dirigé (DAG) d'exécution. DAG contient un ensemble de tâches et leurs interactions dont chaque employé a besoin pour s'exécuter. Cette disposition signifie que les tâches ne sont pas exécutées tant que l'utilisateur n'a pas dit à DASK de les exécuter d'une manière ou d'une autre. Avec DASK, vous disposez de trois options principales :

- **Call Compute() sur un DataFrame.** cet appel traite toutes les partitions, puis renvoie les résultats au planificateur pour l'agrégation finale et la conversion en cuDF DataFrame. Cette option doit être utilisée avec parcimonie et avec des résultats fortement réduits, à moins que le nœud du planificateur ne manque de mémoire.
- **Call persistent() sur un DataFrame.** cet appel exécute le graphique, mais, au lieu de renvoyer les résultats au nœud du planificateur, il les maintient dans le cluster en mémoire afin que l'utilisateur puisse réutiliser ces résultats intermédiaires dans le pipeline sans avoir besoin de réexécuter le même traitement.
- **Call head() sur un DataFrame.** tout comme avec cuDF, cet appel renvoie 10 enregistrements au nœud du planificateur. Cette option permet de vérifier rapidement si votre DataFrame contient le format de sortie souhaité ou si les enregistrements eux-mêmes ont un sens, selon votre traitement et votre calcul.

Par conséquent, à moins que l'utilisateur n'appelle l'une ou l'autre de ces actions, les travailleurs restent inactifs en attendant que le planificateur lance le traitement. Ce modèle d'exécution paresseux est courant dans les infrastructures informatiques modernes parallèles et distribuées telles qu'Apache Spark.

Le paragraphe suivant forme un modèle de forêt aléatoire en utilisant DASK cuML pour le calcul accéléré par GPU distribué et calcule la précision de prévision des modèles.

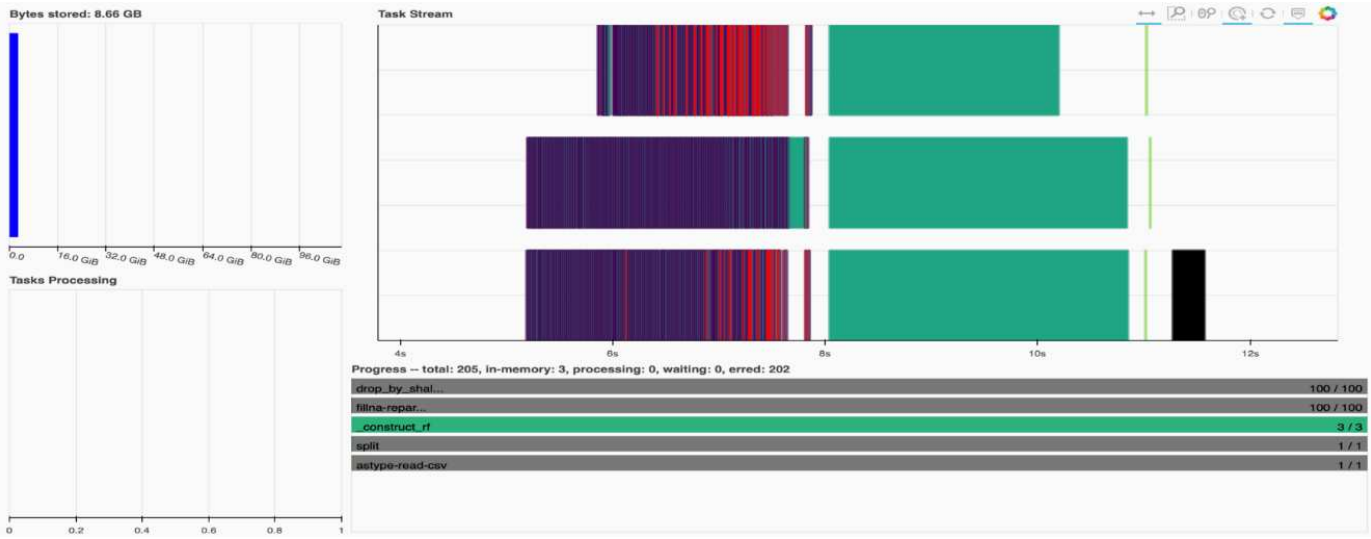
```
Adsf
# Random Forest building parameters
n_streams = 8 # optimization
max_depth = 10
n_bins = 16
n_trees = 10
cuml_model = cumlDaskRF(max_depth=max_depth, n_estimators=n_trees,
n_bins=n_bins, n_streams=n_streams, verbose=True, client=client)
cuml_model.fit(gdf_sliced_small, Y)
# Model prediction
pred_df = cuml_model.predict(gdf_test)
# calculate accuracy
cu_score = cuml.metrics.accuracy_score( test_y, pred_df )
```

Surveiller le DASK à l'aide du tableau de bord des flux de tâches natifs

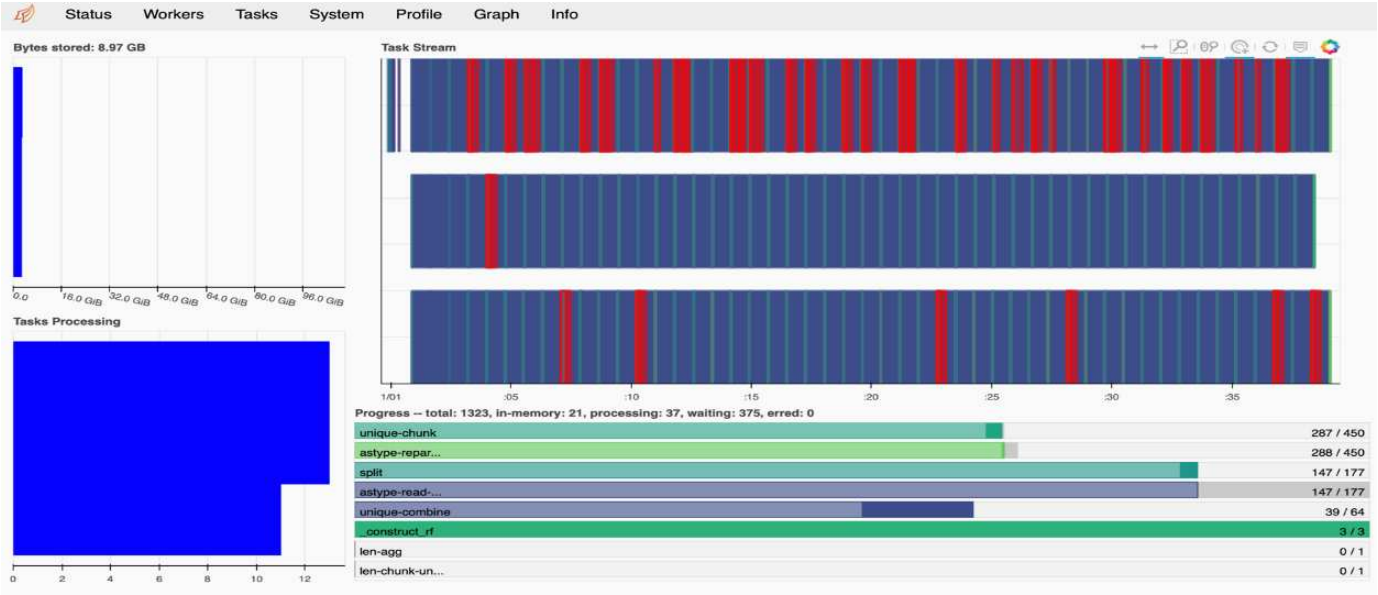
Le "Planificateur distribué DASK" fournit des commentaires en direct sous deux formes :

- Un tableau de bord interactif contenant de nombreux tracés et tableaux avec des informations en direct
- Barre de progression adaptée à une utilisation interactive dans les consoles ou les ordinateurs portables

Dans notre cas, la figure suivante montre comment surveiller la progression de la tâche, y compris les octets stockés, le flux de tâches avec une répartition détaillée du nombre de flux et la progression par nom de tâche avec les fonctions associées exécutées. Dans notre cas, étant donné que nous avons trois nœuds workers, il y a trois principaux segments de flux et les codes de couleurs indiquent des tâches différentes dans chaque flux.



Vous avez la possibilité d'analyser des tâches individuelles et d'examiner le temps d'exécution en millisecondes ou d'identifier tout obstacle ou obstacle. Par exemple, la figure suivante montre les flux de tâches pour le stade d'ajustement du modèle forestier aléatoire. Il y a beaucoup plus de fonctions exécutées, y compris le bloc unique pour le traitement de DataFrame, _construct_rf pour l'adaptation de la forêt aléatoire, et ainsi de suite. La plupart du temps a été consacré à des opérations DataFrame en raison de la grande taille (45 Go) des données d'une journée provenant des journaux Criteo Click.



Comparaison de la durée de formation

Cette section compare le temps d'entraînement du modèle à l'aide de Pandas conventionnels par rapport à DASK. Pour Pandas, nous avons chargé une plus petite quantité de données en raison de la nature du temps de traitement plus lent pour éviter le débordement de mémoire. Par conséquent, nous avons interpolé les résultats pour offrir une comparaison équitable.

Le tableau suivant montre la comparaison du temps d'entraînement brut lorsque le modèle Pandas utilise beaucoup moins de données pour sa forêt aléatoire (50 millions de lignes sur une base de 20 milliards par jour et 15 du dataset). Cet échantillon utilise uniquement moins de 0.25 % de toutes les données disponibles. Alors que pour DASK-cuML nous avons formé le modèle forestier aléatoire sur les 20 milliards de lignes disponibles. Les deux approches ont donné un temps de formation comparable.

Approche	Temps de formation
Scikit-Learn: L'utilisation de seulement 50M rangées en jour 15 comme données d'entraînement	47 minutes et 21 secondes
RAPIDS-DASK : utiliser les 20B rangées en jour 15 comme données d'entraînement	1 heure, 12 minutes et 11 secondes

Si nous interpoler les résultats de temps d'entraînement de façon linéaire, comme le montre le tableau suivant, il y a un avantage significatif à utiliser la formation distribuée avec DASK. L'approche classique Pandas scikit-Learn prendra 13 jours pour traiter et entraîner 45 Go de données pour une seule journée de journaux, tandis QUE L'approche RAPIDS-DASK traite la même quantité de données 262.39 fois plus vite.

Approche	Temps de formation
Scikit-Learn: Utiliser les données d'entraînement de toutes les lignes 20B en jour 15	13 jours, 3 heures, 40 minutes et 11 secondes
RAPIDS-DASK : utiliser les 20B rangées en jour 15 comme données d'entraînement	1 heure, 12 minutes et 11 secondes

Le tableau précédent montre que, GRÂCE À RAPIDS et DASK, le traitement des données et l'entraînement

des modèles sur plusieurs instances GPU, la durée d'exécution est considérablement plus courte que le traitement classique Pandas DataFrame avec entraînement des modèles scikit. Cette structure permet une évolutivité verticale et horizontale dans le cloud et sur site au sein d'un cluster multinœud et multiprocesseur graphique.

Surveillez DAsk et RAPIDS avec Prometheus et Grafana

Une fois que tout est déployé, exécutez des inférences sur les nouvelles données. Les modèles prédisent si un utilisateur clique sur une publicité en fonction des activités de navigation. Les résultats de la prédiction sont stockés dans un DASK cuDF. Vous pouvez surveiller les résultats avec Prometheus et visualiser dans les tableaux de bord Grafana.

Pour plus d'informations, reportez-vous à ce document "RAPIDS ai moyen post".

Gestion des versions de datasets et de modèles avec le kit NetApp DataOps

Le kit NetApp DataOps pour Kubernetes extrait les ressources de stockage et les workloads Kubernetes jusqu'à l'espace de travail data-science. Ces fonctionnalités sont packagées dans une interface simple et conviviale conçue pour les data Scientists et les ingénieurs de données. Ce kit, qui se présente sous la forme familière d'un programme Python, permet aux scientifiques et aux ingénieurs des données de provisionner et de déployer des espaces de travail JupyterLab en quelques secondes. Ces espaces de travail peuvent contenir des téraoctets, voire des pétaoctets, de capacité de stockage, ce qui permet aux data Scientists de stocker tous leurs datasets d'entraînement directement dans leur espace de travail de projet. Il ne fait plus que gérer séparément les espaces de travail et les volumes de données.

Pour en savoir plus, visitez la boîte à outils "Référentiel GitHub".

Ordinateurs portables Jupyter à titre de référence

Deux ordinateurs portables Jupyter sont associés à ce rapport technique :

- **"CTR-PandasRF-Assemblé.ipynb."** Cet ordinateur portable charge le jour 15 à partir du jeu de données de journaux Criteo Terabyte Click, traite et formate les données dans un Pandas DataFrame, forme un modèle de forêt aléatoire d'apprentissage Scikit, effectue des prédictions et calcule la précision.
- **"criteo_dAsk_RF.ipynb."** Cet ordinateur portable charge le jour 15 à partir du jeu de données de journal Criteo Terabun Click Logs, traite et formate les données dans un DAsk cuDF, forme un modèle forestier aléatoire DASk cuML, effectue des prévisions et calcule la précision. L'exploitation de plusieurs nœuds workers avec des processeurs graphiques permet de tirer parti de cette approche de traitement et d'entraînement des modèles distribués pour les données. Plus vous traitez de données, plus vous gagnez de temps par rapport à une approche ML classique. Vous pouvez déployer cet ordinateur portable dans le cloud, sur site ou dans un environnement hybride dans lequel le cluster Kubernetes contient des ressources de calcul et de stockage à différents emplacements, à condition que la configuration réseau permet la libre circulation des données et de la distribution des modèles.

Conclusion

Azure NetApp Files, RAPIDS et DASK accélèrent et simplifient le déploiement du traitement DE ML et de la formation à grande échelle en intégrant les outils

d'orchestration tels que Docker et Kubernetes. En unifiant le pipeline de données de bout en bout, cette solution réduit la latence et la complexité inhérentes à de nombreuses charges de travail informatiques avancées, afin de combler efficacement l'écart entre le développement et les opérations. Les data Scientists peuvent exécuter des requêtes sur d'importants jeux de données et partager en toute sécurité les données et les modèles algorithmiques avec d'autres utilisateurs au cours de la phase d'entraînement.

Lors de la création de vos propres pipelines d'IA et DE ML, la configuration de l'intégration, de la gestion, de la sécurité et de l'accessibilité des composants d'une architecture représente un véritable défi. Le fait de donner aux développeurs accès à leur environnement et de le contrôler constitue un autre ensemble de défis.

En créant un modèle d'entraînement distribué complet et un pipeline de données dans le cloud, nous avons mis en avant deux ordres d'amélioration considérable au niveau du temps d'exécution total des workflows, par rapport à une approche open source classique qui n'exploite pas les frameworks de calcul et de traitement des données accélérés par GPU.

L'association de NetApp, Microsoft, des frameworks d'orchestration open source et NVIDIA rassemble les dernières technologies en tant que services gérés, tout en offrant une grande flexibilité pour accélérer l'adoption de technologies et améliorer le délai de mise sur le marché de nouvelles applications d'IA et DE ML. Ces services avancés sont fournis dans un environnement cloud natif facilement porté pour les architectures de déploiement sur site ainsi que pour les architectures de déploiement hybrides.

Où trouver des informations complémentaires

Pour en savoir plus sur les informations données dans ce document, consultez les ressources suivantes :

- Azure NetApp Files :
  - Architecture des solutions pour Azure NetApp Files

    "https://docs.microsoft.com/azure/azure-netapp-files/azure-netapp-files-solution-architectures"
- Trident pour le stockage persistant pour les conteneurs :
  - Azure NetApp Files et Trident

    "https://netapptrident.readthedocs.io/en/stablev20.07/kubernetes/operations/tasks/backends/anf.html"
- DASK and RAPIDS :
  - DASK

    "https://docs.dask.org/en/latest/"
  - Installer DASK

    "https://docs.dask.org/en/latest/install.html"
  - API DASK

    "https://docs.dask.org/en/latest/api.html"
  - DASK machine Learning

"https://examples.dask.org/machine-learning.html"

  - Diagnostic distribué DASK

    "https://docs.dask.org/en/latest/diagnostics-distributed.html"

- Structure ET outils ML :
  - TensorFlow : un framework d'apprentissage machine open source pour tous

    "https://www.tensorflow.org/"

  - Docker

    "https://docs.docker.com"

  - Kubernetes

    "https://kubernetes.io/docs/home/"

  - Kubeflow

    "http://www.kubeflow.org/"

  - Jupyter Notebook Server

    "http://www.jupyter.org/"

Tr-4896: Formation distribuée dans Azure: Détection de voie - conception de la solution

Muneer Ahmad et Verron Martina, NetApp Ronen Dar, EXÉCUTEZ :ai

Depuis mai 2019, Microsoft propose un service de portail Azure propriétaire natif pour les services de fichiers NFS et SMB d'entreprise basés sur la technologie NetApp ONTAP. Ce développement est régi par un partenariat stratégique entre Microsoft et NetApp, et étend encore davantage la portée des services de données ONTAP à Azure.

NetApp, l'un des principaux fournisseurs de services de données cloud, s'est associé À CETTE SOLUTION : l'IA, une entreprise qui virtualise l'infrastructure d'IA pour accélérer les tests d'IA en utilisant pleinement les GPU. Ce partenariat permet aux équipes d'accélérer l'IA en exécutant de nombreuses expériences en parallèle, avec un accès rapide aux données et en exploitant des ressources de calcul sans limites. EXÉCUTION : l'IA permet l'utilisation totale des GPU en automatisant l'allocation des ressources. L'architecture éprouvée du Azure NetApp Files permet également à tous les tests de s'exécuter à la vitesse maximale en éliminant les obstacles au pipeline de données.

NetApp et LA GESTION : l'IA s'est associée afin de proposer aux clients une plateforme pérenne pour leur transition vers l'IA dans Azure. De l'analytique et du calcul haute performance aux décisions autonomes (où les clients peuvent optimiser leurs investissements IT en ne payant que ce dont ils ont besoin, quand ils en ont besoin), la alliance entre NetApp ET EXÉCUTION : l'IA offre une expérience unifiée dans le cloud Azure.

Présentation de la solution

Dans cette architecture, l'accent est mis sur la partie la plus gourmande en capacité de calcul du processus d'IA ou de machine learning (ML) distribué de détection de voies. La détection de voie est l'une des tâches les plus importantes en conduite autonome, ce qui aide à guider les véhicules en localisant les marquages de voie. Des composants statiques comme les marquages de voie guident le véhicule pour conduire sur la route de manière interactive et sûre.

Les approches basées sur le réseau neuronal convolutif (CNN) ont poussé la compréhension et la segmentation des scènes à un nouveau niveau. Bien qu'il ne fonctionne pas bien pour les objets avec de longues structures et régions qui pourraient être obstruées (par exemple, poteaux, ombre sur la voie, etc.). Le réseau neuronal spatial convolutif (SCNN) généralise la CNN à un niveau spatial riche. Elle permet la propagation d'informations entre les neurones dans la même couche, ce qui le rend mieux adapté aux objets structurés tels que les voies, les pôles ou le chariot avec occlusions. Cette compatibilité est due au fait que les informations spatiales peuvent être renforcées et qu'elles préservent la fluidité et la continuité.

Des milliers d'images de scène doivent être injectées dans le système pour permettre au modèle d'apprendre et de distinguer les différents composants du jeu de données. Ces images comprennent la météo, la journée ou la nuit, les routes à plusieurs voies et d'autres conditions de circulation.

Pour la formation, il est nécessaire de disposer d'une bonne qualité et d'une bonne quantité de données. L'entraînement peut prendre plusieurs jours ou plusieurs semaines avec un seul GPU. L'entraînement distribué par les données peut accélérer le processus en utilisant plusieurs processeurs graphiques ou plusieurs nœuds. Horovod est un cadre de ce type qui permet des formations distribuées, mais la lecture des données dans l'ensemble des clusters des GPU peut jouer un rôle obstacle. Les systèmes Azure NetApp Files assurent des débits élevés et une faible latence continue pour fournir des fonctionnalités scale-up et scale-up que les GPU utilisent afin d'exploiter au mieux leur capacité de calcul. Nos expériences ont démontré que tous les GPU du cluster étaient utilisés à plus de 96 % en moyenne pour l'entraînement de la détection de voie à l'aide de SCNN.

Public visé

La science des données inclut plusieurs disciplines INFORMATIQUES et commerciales. Par conséquent, plusieurs personnages font partie du public ciblé :

- Les data Scientists doivent donc pouvoir utiliser les outils et les bibliothèques de leur choix.
- Les ingénieurs de données doivent savoir comment elles circulent et où elles résident.
- Des experts en matière de conduite autonome.
- Administrateurs et architectes cloud pour configurer et gérer les ressources cloud (Azure).
- Un ingénieur DevOps a besoin des outils pour intégrer les nouvelles applications d'IA et DE ML dans son pipeline d'intégration et de déploiement continus.
- Les utilisateurs professionnels veulent avoir accès aux applications d'IA et DE ML.

Dans ce document, nous vous présentons comment Azure NetApp Files, L'EXÉCUTION : l'IA et Microsoft Azure aident chacun de ces rôles à créer de la valeur commerciale.

Technologie de la solution

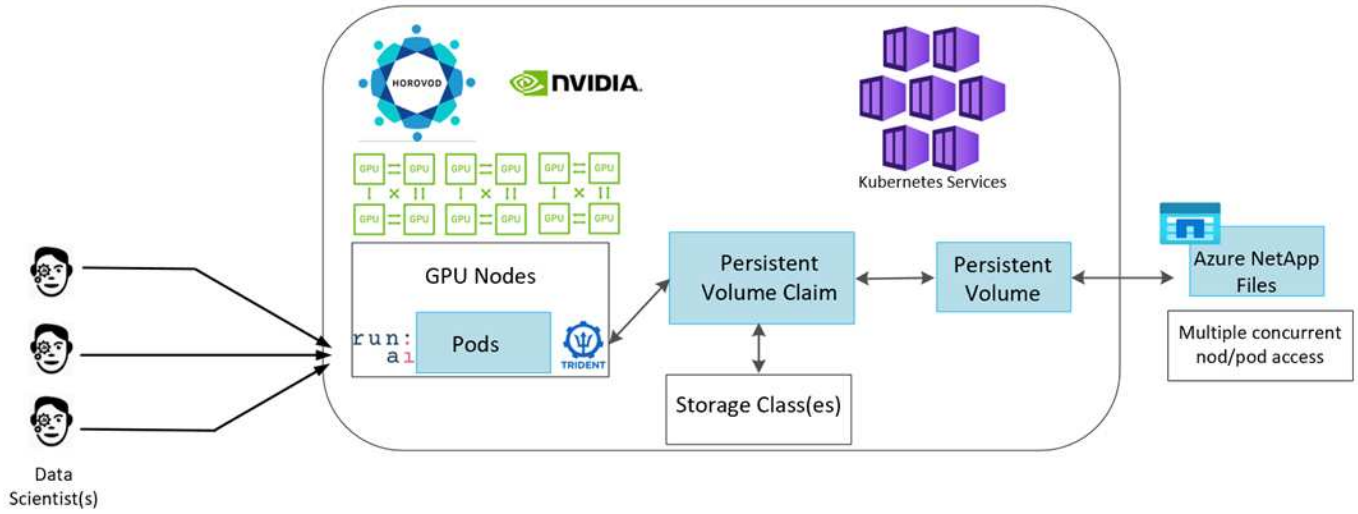
Cette section traite des exigences technologiques pour la détection de voie en mettant en œuvre une solution de formation distribuée à grande échelle qui s'exécute entièrement dans le cloud Azure. La figure ci-dessous

présente l'architecture de la solution.

Les éléments utilisés dans cette solution sont les suivants :

- Azure Kubernetes Service (AKS)
- Références de calcul Azure avec processeurs graphiques NVIDIA
- Azure NetApp Files
- EXÉCUTEZ : L'IA
- NetApp Trident

Les liens vers tous les éléments mentionnés ici sont répertoriés dans le "Informations supplémentaires" section.



Ressources et conditions des services clouds

Le tableau suivant répertorie les composants matériels requis pour implémenter la solution. Les composants cloud utilisés dans toute implémentation de cette solution peuvent varier en fonction des besoins du client.

Le cloud	Quantité
AKS	Au moins trois nœuds système et trois nœuds workers GPU
Nœuds système SKU de machine virtuelle (VM)	Trois Standard_DS2_v2
Nœuds worker GPU référence VM	Trois Standard_NC6s_v3
Azure NetApp Files	Niveau standard 4 To

Configuration logicielle requise

Le tableau suivant répertorie les composants logiciels requis pour implémenter la solution. Ils peuvent varier selon l'implémentation de la solution et les besoins du client.

Logiciel	Version ou autres informations
AKS - version Kubernetes	1.18.14

Logiciel	Version ou autres informations
EXÉCUTEZ :CLI AI	v2.2.25
EXÉCUTION : version de l'opérateur Kubernetes d'orchestration d'IA	1.0.109
Horovod	0.21.2
NetApp Trident	20.01.1
Gouvernail	3.0.0

Détection de voie – entraînement distribué avec RUN:ai

Cette section fournit des informations détaillées sur la configuration de la plateforme pour exécuter l'entraînement à grande échelle avec détection de voie, à l'aide de L'OUTIL RUN : ai orchestrator. Nous discutons de l'installation de tous les éléments de la solution et de l'exécution du travail de formation distribuée sur ladite plate-forme. La gestion des versions DE ML est réalisée à l'aide de NetApp SnapshotTM lié À L'EXÉCUTION : des expériences d'IA pour atteindre la reproductibilité des données et des modèles. La gestion des versions DE ML joue un rôle crucial dans le suivi des modèles, le partage du travail entre les membres de l'équipe, la reproductibilité des résultats, la rotation de nouvelles versions de modèles vers la production et la provenance des données. Le contrôle des versions DE NetApp ML (Snapshot) permet de capturer des versions instantanées des données, des modèles entraînés et des journaux associés à chaque expérience. Grâce à sa prise en charge avancée des API, il est facile à intégrer AVEC la plateforme D'IA POUR L'EXÉCUTION : il vous suffit de déclencher un événement basé sur l'état d'entraînement. Par ailleurs, il est nécessaire de capturer l'état de l'expérience dans son intégralité, sans modifier quoi que ce soit dans le code ou les conteneurs exécutés sur Kubernetes (K8s).

Enfin, ce rapport technique se conclut par une évaluation des performances sur plusieurs nœuds compatibles avec les processeurs graphiques dans AKS.

Formation distribuée pour la détection de voie utilisation du boîtier à l'aide du jeu de données Tusimple

Dans le présent rapport technique, une formation distribuée est réalisée sur le jeu de données Tusimple pour la détection des voies. Horovod est utilisé dans le code d'entraînement pour l'entraînement distribué des données sur plusieurs nœuds GPU simultanément dans le cluster Kubernetes via AKS. Le code est emballé sous forme d'images conteneur pour le téléchargement et le traitement de données Tusimple. Les données traitées sont stockées sur des volumes persistants alloués par le plug-in NetApp Trident. Pour l'entraînement, une autre image de conteneur est créée, et elle utilise les données stockées sur les volumes persistants créés lors du téléchargement des données.

Pour envoyer les données et la tâche d'entraînement, utilisez L'INTELLIGENCE artificielle pour orchestrer l'allocation et la gestion des ressources. EXÉCUTION : ai vous permet d'effectuer des opérations MPI (message Passing interface) nécessaires à Horovod. Cette disposition permet à plusieurs nœuds GPU de communiquer les uns avec les autres afin de mettre à jour le poids de l'entraînement après chaque mini-lot d'entraînement. Il permet également de surveiller la formation via l'interface et l'interface de ligne de commande, ce qui facilite le suivi de la progression des expériences.

NetApp Snapshot est intégré au code d'entraînement et capture l'état des données et le modèle d'entraînement pour chaque expérience. Cette fonctionnalité vous permet de suivre la version des données et du code utilisé, ainsi que le modèle entraîné associé généré.

Installation et installation d'AKS

Pour la configuration et l'installation du cluster AKS, accédez à "Créer un cluster AKS". Ensuite, suivez ces étapes :

1. Lors de la sélection du type de nœuds (qu'il s'agisse de nœuds système (CPU) ou de nœuds worker (GPU), sélectionnez les éléments suivants :

   a. Ajouter le nœud système principal nommé `agentpool` au `Standard_DS2_v2` taille. Utilisez les trois nœuds par défaut.

   b. Ajouter un nœud de travail `gpupool` avec `the Standard_NC6s_v3` de taille de pool. Utilisez trois nœuds au minimum pour les nœuds GPU.

   + Add node pool Delete

Name	Mode	OS type	Node count	Node size
agentpool	System	Linux	3	Standard_DS2_v2
gpupool	User	Linux	3	Standard_NC6s_v

   ⓘ Le déploiement prend de 5 à 10 minutes.

2. Une fois le déploiement terminé, cliquez sur connexion au cluster. Pour vous connecter au cluster AKS récemment créé, installez l'outil de ligne de commandes Kubernetes à partir de votre environnement local (ordinateur portable/PC). Visitez "Installation des outils" Pour l'installer conformément à votre système d'exploitation.
3. "Installez l'interface de ligne de commandes Azure dans votre environnement local".
4. Pour accéder au cluster AKS à partir du terminal, entrez tout d'abord `az login` et de saisir les informations d'identification.
5. Exécutez les deux commandes suivantes :

```
az account set --subscription xxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxxx
aks get-credentials --resource-group resourcegroup --name aksclustername
```

6. Entrez cette commande dans l'interface de ligne de commande Azure :

```
kubectl get nodes
```

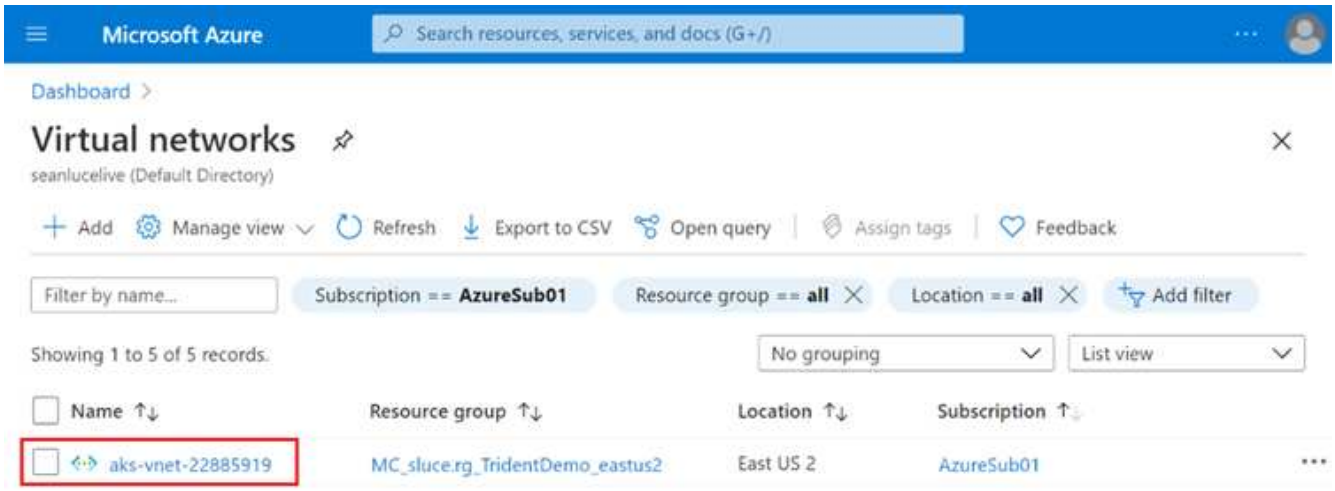
   ⓘ Si les six nœuds sont opérationnels comme indiqué ici, votre cluster AKS est prêt et connecté à votre environnement local.

```
verronmartina@verron-mac-0 ~ % kubectl get nodes
NAME                                STATUS   ROLES    AGE   VERSION
aks-agentpool-34613062-vmss000000   Ready    agent    22m   v1.18.14
aks-agentpool-34613062-vmss000001   Ready    agent    22m   v1.18.14
aks-agentpool-34613062-vmss000002   Ready    agent    22m   v1.18.14
aks-gpupool-34613062-vmss000000     Ready    agent    20m   v1.18.14
aks-gpupool-34613062-vmss000001     Ready    agent    20m   v1.18.14
aks-gpupool-34613062-vmss000002     Ready    agent    20m   v1.18.14
verronmartina@verron-mac-0 ~ %
```

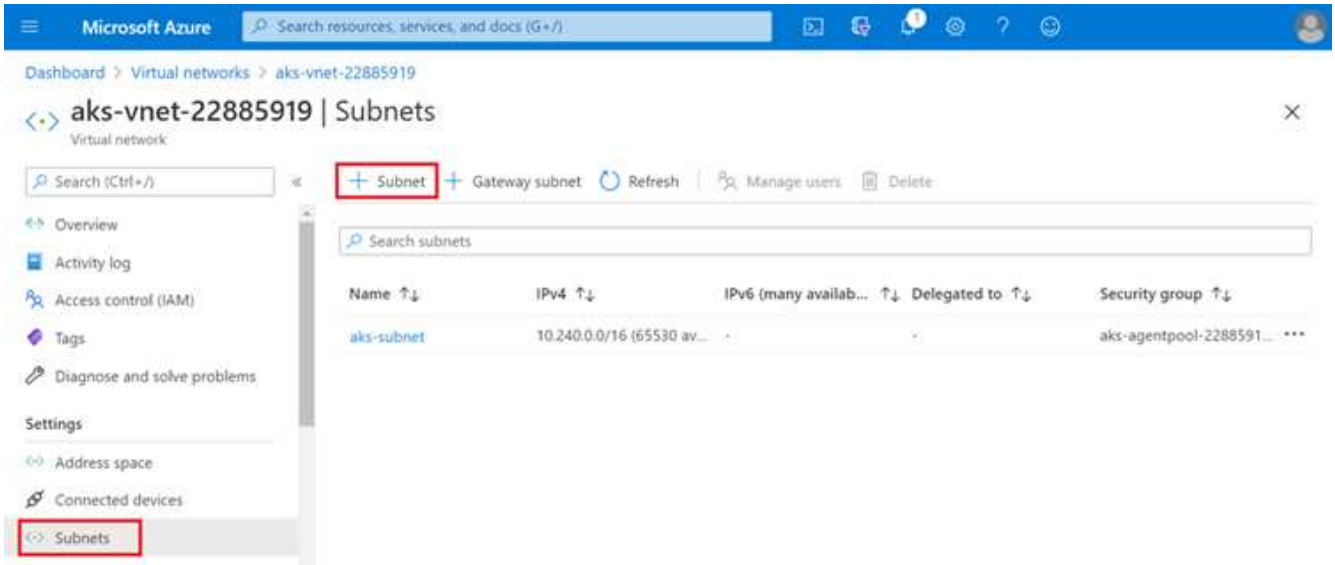
Créez un sous-réseau délégué pour Azure NetApp Files

Pour créer un sous-réseau délégué pour Azure NetApp Files, suivez cette série d'étapes :

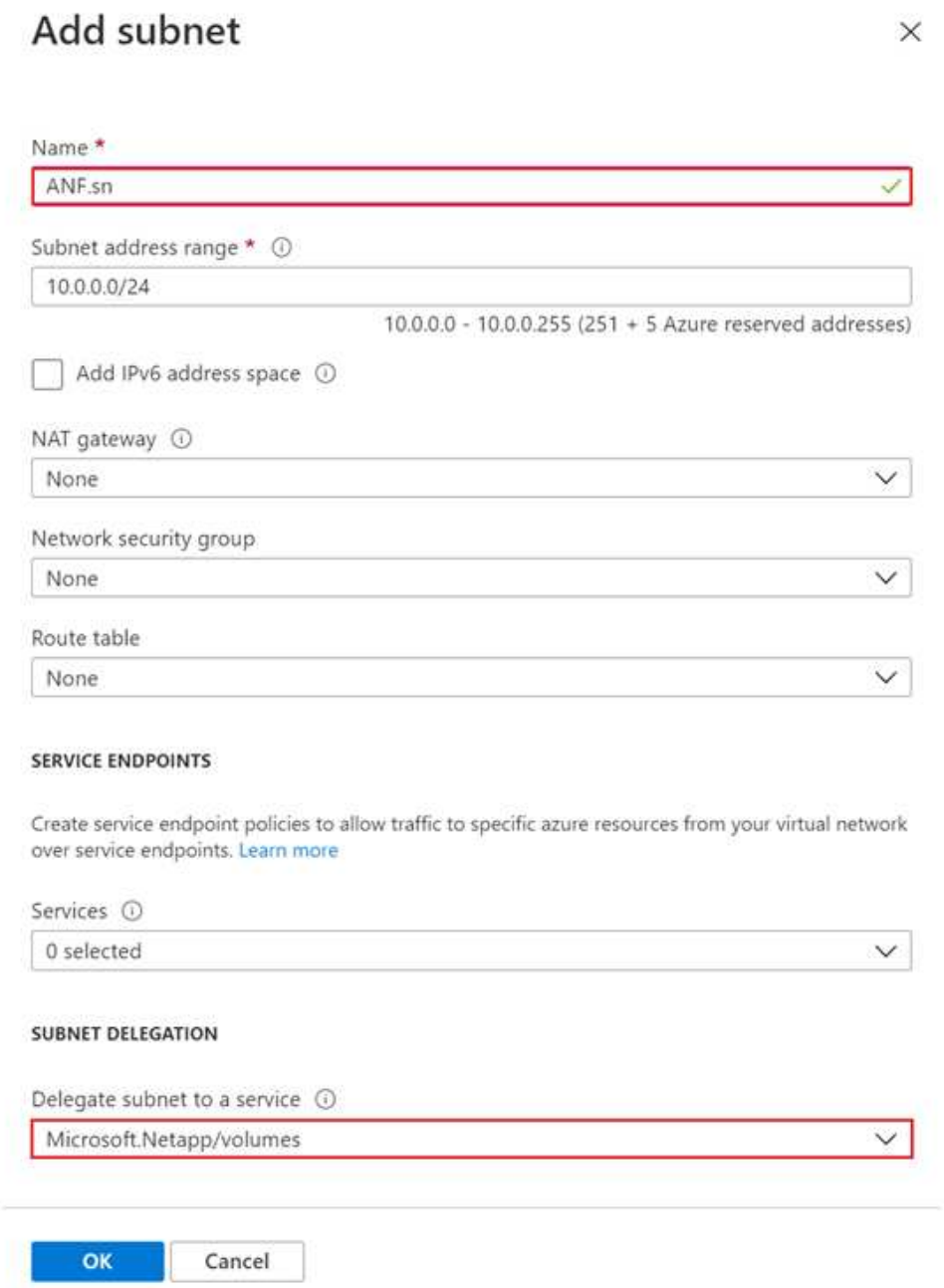
1. Accédez aux réseaux virtuels depuis le portail Azure. Trouvez votre nouveau réseau virtuel. Il devrait avoir un préfixe tel que aks-vnet, comme indiqué ici. Cliquez sur le nom du réseau virtuel.



2. Cliquez sur sous-réseaux et sélectionnez +sous-réseau dans la barre d'outils supérieure.



3. Indiquez au sous-réseau un nom tel que `ANF.sn` Et sous l'en-tête délégation de sous-réseau, sélectionnez Microsoft.NetApp/volumes. Ne rien changer. Cliquez sur OK.



Les volumes Azure NetApp Files sont alloués au cluster d'applications et utilisés en tant que demandes de volume persistant dans Kubernetes. Par conséquent, cette allocation nous offre la flexibilité nécessaire pour mapper les volumes à différents services, qu'il s'agit d'ordinateurs portables Jupyter, de fonctions sans serveur, etc

Les utilisateurs des services peuvent consommer le stockage depuis la plateforme de différentes manières. Les principaux avantages de Azure NetApp Files sont les suivants :

- Permet aux utilisateurs d'utiliser les snapshots.
- Permet aux utilisateurs de stocker de grandes quantités de données sur des volumes Azure NetApp Files.
- Optimisez les performances des volumes Azure NetApp Files lorsque vous exécutez leurs modèles sur de

vastes ensembles de fichiers.

Définition Azure NetApp Files

Pour terminer la configuration de Azure NetApp Files, vous devez d'abord la configurer comme décrit à la section "Démarrage rapide : configurez Azure NetApp Files et créez un volume NFS".

Toutefois, vous pouvez ignorer les étapes de création d'un volume NFS pour Azure NetApp Files lorsque vous créez des volumes via Trident. Avant de continuer, assurez-vous d'avoir :

1. "Enregistré pour Azure NetApp Files et le fournisseur de ressources NetApp (via le shell cloud Azure)".
2. "Créé un compte dans Azure NetApp Files".
3. "Configurez un pool de capacité" (Minimum 4 Tio Standard ou Premium selon vos besoins).

Peering de réseau virtuel AKS et de réseau virtuel Azure NetApp Files

Ensuite, Peer the AKS Virtual Network (vnet) with the Azure NetApp Files vNet en suivant les étapes suivantes :

1. Dans la zone de recherche située en haut du portail Azure, saisissez les réseaux virtuels.
2. Cliquez sur VNet aks- vnet-name, puis entrez Peerings dans le champ de recherche.
3. Cliquez sur +Add et entrez les informations fournies dans le tableau ci-dessous :

Champ	Valeur ou description
Nom de la liaison de peering	aks-vnet-name_to_anf
ID d'abonnement	Abonnement au réseau VNet Azure NetApp Files auquel vous vous trouvez
Partenaire de peering vnet	Azure NetApp Files vnet

> Laissez toutes les sections non astérisque par défaut

4. Cliquez SUR AJOUTER ou sur OK pour ajouter le peering au réseau virtuel.

Pour plus d'informations, rendez-vous sur "Créez, modifiez ou supprimez un peering de réseau virtuel".

Trident

Trident est un projet open source piloté par NetApp, conçu pour répondre aux demandes de stockage persistant des applications conteneurisées. Trident a été implémenté en tant que contrôleur de provisionnement externe. Fonctionnant comme un pod autonome, il contrôle les volumes et automatise entièrement le provisionnement.

NetApp Trident facilite l'intégration avec K8s en créant et en connectant des volumes persistants pour le stockage des datasets d'entraînement et des modèles entraînés. Grâce à cette fonctionnalité, les data Scientists et les ingénieurs de données peuvent utiliser K8s en toute simplicité, sans avoir à gérer et à stocker manuellement les datasets. Avec Trident, les data Scientists n'ont plus besoin d'apprendre à gérer de nouvelles plateformes de données, puisqu'il intègre les tâches liées à la gestion des données via l'intégration d'API logiques.

Installation de Trident

Pour installer le logiciel Trident, procédez comme suit :

1. "Installez tout d'abord le gouvernail".
2. Téléchargez et extrayez le programme d'installation de Trident 21.01.1.

```
wget
https://github.com/NetApp/trident/releases/download/v21.01.1/trident-
installer-21.01.1.tar.gz
tar -xf trident-installer-21.01.1.tar.gz
```

3. Définissez le répertoire sur `trident-installer`.

```
cd trident-installer
```

4. Copier `tridentctl` dans un répertoire de votre système `$PATH`.

```
cp ./tridentctl /usr/local/bin
```

5. Installation de Trident sur le cluster K8s avec Helm :
   a. Changez le répertoire en répertoire Helm.

   ```
   cd helm
   ```

   b. Installation de Trident.

   ```
   helm install trident trident-operator-21.01.1.tgz --namespace trident
   --create-namespace
   ```

   c. Vérifiez l'état des modules Trident de la façon habituelle de K8s :

   ```
   kubectl -n trident get pods
   ```

   d. Si tous les modules sont opérationnels, Trident est installé et vous pouvez passer à l'étape supérieure.

Configurer le back-end et la classe de stockage Azure NetApp Files

Pour configurer la back-end et la classe de stockage Azure NetApp Files, procédez comme suit :

1. Revenir au répertoire de base.

```
cd ~
```

2. Cloner le "référentiel de projet" `lane-detection-SCNN-horovod`.
3. Accédez au `trident-config` répertoire.

```
cd ./lane-detection-SCNN-horovod/trident-config
```

4. Créez un principe de service Azure (le principe du service est celui de la façon dont Trident communique avec Azure pour accéder à vos ressources Azure NetApp Files).

```
az ad sp create-for-rbac --name
```

Le résultat de la commande doit ressembler à l'exemple suivant :

```
{
  "appId": "xxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx",
   "displayName": "netapptrident",
    "name": "http://netapptrident",
    "password": "xxxxxxxxxxxxxx.xxxxxxxxxxxxxx",
    "tenant": "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxx"
 }
```

5. Création de Trident `backend json` fichier.
6. À l'aide de votre éditeur de texte préféré, renseignez les champs suivants du tableau ci-dessous à l'intérieur du `anf-backend.json` fichier.

Champ	Valeur
ID d'abonnement	Votre ID d'abonnement Azure
ID de tenantID	Votre ID de locataire Azure (à partir de la sortie d'az ad sp à l'étape précédente)
ID client	Votre AppID (à partir de la sortie d'az ad sp à l'étape précédente)
ClientSecret	Votre mot de passe (à partir de la sortie d'az ad sp à l'étape précédente)

Le fichier doit ressembler à l'exemple suivant :

```
{
    "version": 1,
    "storageDriverName": "azure-netapp-files",
    "subscriptionID": "fakec765-4774-fake-ae98-a721add4fake",
    "tenantID": "fakef836-edc1-fake-bff9-b2d865eefake",
    "clientID": "fake0f63-bf8e-fake-8076-8de91e57fake",
    "clientSecret": "SECRET",
    "location": "westeurope",
    "serviceLevel": "Standard",
    "virtualNetwork": "anf-vnet",
    "subnet": "default",
    "nfsMountOptions": "vers=3,proto=tcp",
    "limitVolumeSize": "500Gi",
    "defaults": {
    "exportRule": "0.0.0.0/0",
    "size": "200Gi"
}
```

7. Demandez à Trident de créer le back-end Azure NetApp Files dans le `trident` espace de noms, utilisation `anf-backend.json` comme le fichier de configuration :

```
tridentctl create backend -f anf-backend.json -n trident
```

8. Créer la classe de stockage :

   a. Les utilisateurs de K8 peuvent provisionner des volumes à l'aide des ESV qui spécifient une classe de stockage par nom. Demandez à K8s de créer une classe de stockage `azurenetappfiles` Cela fera référence au back-end Azure NetApp Files créé à l'étape précédente en utilisant les éléments suivants :

   ```
   kubectl create -f anf-storage-class.yaml
   ```

   b. Vérifiez que la classe de stockage est créée à l'aide de la commande suivante :

   ```
   kubectl get sc azurenetappfiles
   ```

   Le résultat de la commande doit ressembler à l'exemple suivant :

```
NAME                PROVISIONER             RECLAIMPOLICY   VOLUMEBINDINGMODE   ALLOWVOLUMEEXPANSION   AGE
azurenetappfiles    csi.trident.netapp.io   Delete          Immediate           false                  98s
```

Déploiement et configuration des composants de snapshot de volume sur AKS

Si votre cluster n'est pas préinstallé avec les composants de snapshot de volume appropriés, vous pouvez

installer ces composants manuellement en exécutant les étapes suivantes :

> AKS 1.18.14 n'a pas de contrôleur Snapshot préinstallé.

1. Installez les CRD bêta de Snapshot à l'aide des commandes suivantes :

```
kubectl create -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-
3.0/client/config/crd/snapshot.storage.k8s.io_volumesnapshotclasses.yaml
kubectl create -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-
3.0/client/config/crd/snapshot.storage.k8s.io_volumesnapshotcontents.yam
l
kubectl create -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-
3.0/client/config/crd/snapshot.storage.k8s.io_volumesnapshots.yaml
```

2. Installez le contrôleur Snapshot à l'aide des documents suivants de GitHub :

```
kubectl apply -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-3.0/deploy/kubernetes/snapshot-
controller/rbac-snapshot-controller.yaml
kubectl apply -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-3.0/deploy/kubernetes/snapshot-
controller/setup-snapshot-controller.yaml
```

3. Configuration de K8s `volumesnapshotclass`: Avant de créer un snapshot de volume, a "classe de snapshot de volume" doit être configuré. Créez une classe de snapshot de volumes pour Azure NetApp Files et utilisez-la pour gérer les versions DE MACHINE LEARNING avec la technologie NetApp Snapshot. Création `volumesnapshotclass netapp-csi-snapclass` et définissez-le sur `volumesnapshotclass `par défaut tels que :

```
kubectl create -f netapp-volume-snapshot-class.yaml
```

   Le résultat de la commande doit ressembler à l'exemple suivant :

```
volumesnapshotclass.snapshot.storage.k8s.io/netapp-csi-snapclass created
```

4. Vérifier que la classe de copie Snapshot du volume a été créée à l'aide de la commande suivante :

```
kubectl get volumesnapshotclass
```

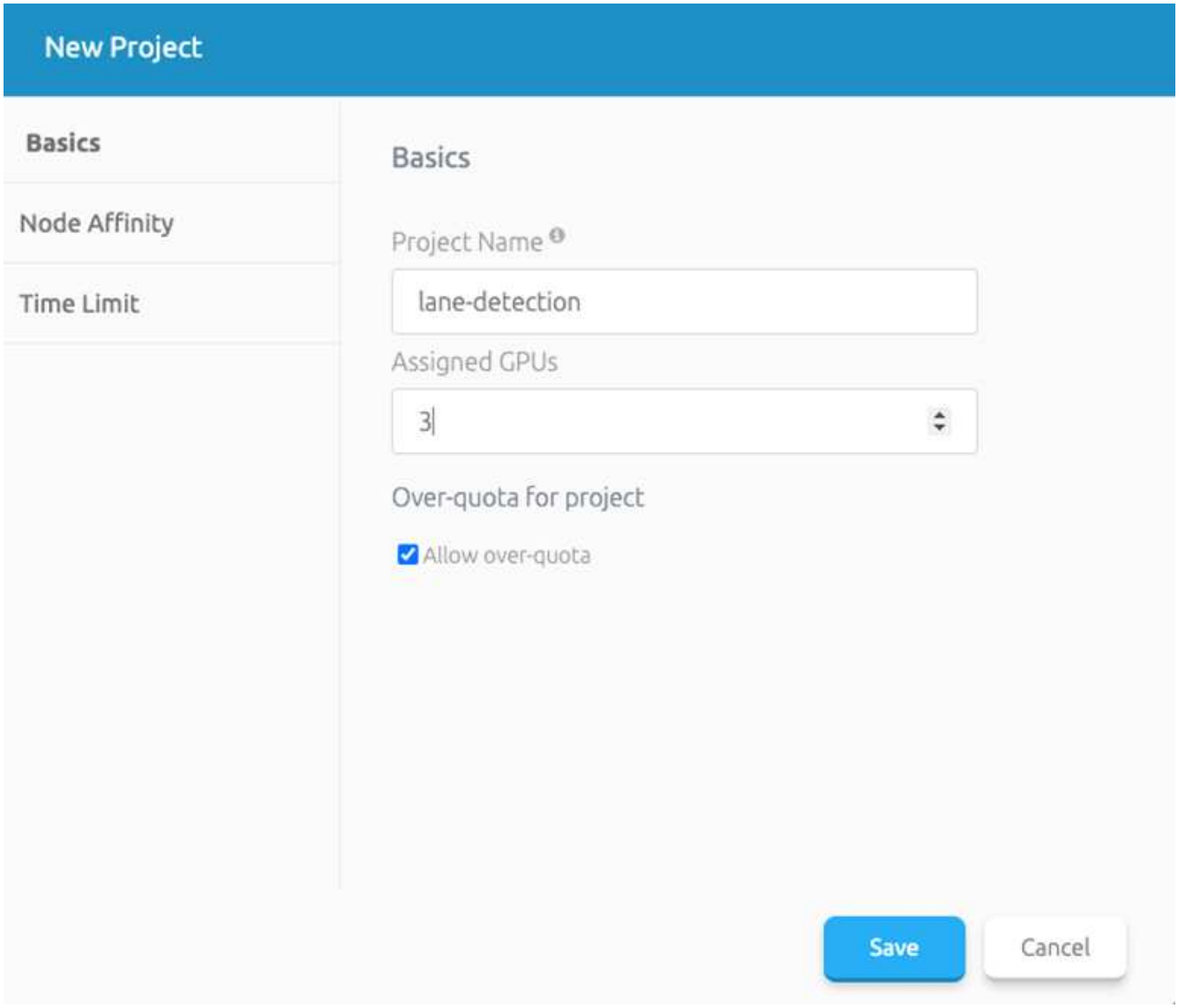
   Le résultat de la commande doit ressembler à l'exemple suivant :

```
NAME                    DRIVER                  DELETIONPOLICY   AGE
netapp-csi-snapclass    csi.trident.netapp.io   Delete           63s
```

EXÉCUTEZ :installation d'ai

Pour installer RUN:ai, procédez comme suit :

1. "Installez le cluster RUN:ai sur AKS".
2. Accédez à app.runai.ai, cliquez sur Créer un nouveau projet et nommez-le détection de voie. Un namespace est créé sur un cluster K8s à partir de `runai-` suivi du nom du projet. Dans ce cas, l'espace de noms créé serait runai-Lane-détection.



3. "INSTALLER RUN:AI CLI".
4. Sur votre terminal, définissez la détection de voie comme EXÉCUTION par défaut : projet ai à l'aide de la commande suivante :

```
`runai config project lane-detection`
```

Le résultat de la commande doit ressembler à l'exemple suivant :

```
Project lane-detection has been set as default project
```

5. Créer ClusterRole et ClusterRoleBinding pour l'espace de noms du projet (par exemple, `lane-detection)` donc le compte de service par défaut appartenant à `runai-lane-detection` l'espace de noms est autorisé à effectuer `volumesnapshot` opérations durant l'exécution du travail :

   a. Indiquez les espaces de noms pour vérifier cela `runai-lane-detection` existe à l'aide de cette commande :

   ```
   kubectl get namespaces
   ```

   La sortie doit apparaître comme dans l'exemple suivant :

```
NAME                   STATUS   AGE
default                Active   130m
kube-node-lease        Active   130m
kube-public            Active   130m
kube-system            Active   130m
runai                  Active   4m44s
runai-lane-detection   Active   13s
trident                Active   102m
```

6. Créer ClusterRole `netappsnapshot` Et ClusterRoleBinding `netappsnapshot` à l'aide des commandes suivantes :

```
`kubectl create -f runai-project-snap-role.yaml`
`kubectl create -f runai-project-snap-role-binding.yaml`
```

Téléchargez et traitez le jeu de données Tusimple AS RUN:ai

Le processus de téléchargement et de traitement de TuDataset simple en TANT QU'EXÉCUTION : travail ai est facultatif. Elle comprend les étapes suivantes :

1. Créez et poussez l'image docker ou omettez cette étape si vous souhaitez utiliser une image docker existante (par exemple, `muneer7589/download-tusimple:1.0)`

   a. Basculer vers le home Directory :

   ```
   cd ~
   ```

b. Accédez au répertoire des données du projet `lane-detection-SCNN-horovod`:

```
cd ./lane-detection-SCNN-horovod/data
```

c. Modifier `build_image.sh` script shell et remplacez le référentiel docker par le vôtre. Par exemple, remplacer `muneer7589` avec le nom de votre référentiel docker. Vous pouvez également modifier le nom et LA BALISE de l'image docker (par exemple, `download-tusimple` et `1.0`) :

```
#!/bin/bash
#
# A simple script to build the Docker image.
#
# $ build_image.sh
set -ex

IMAGE=muneer7589/download-tusimple
TAG=1.0

# Build image
echo "Building image: "$IMAGE
docker build . -f Dockerfile \
 --tag "${IMAGE}:${TAG}"
echo "Finished building image: "$IMAGE

# Push image
echo "Pushing image: "$IMAGE
docker push "${IMAGE}:${TAG}"
echo "Finished pushing image: "$IMAGE
```

d. Exécutez le script pour créer l'image docker et l'envoyer dans le référentiel docker à l'aide des commandes suivantes :

```
chmod +x build_image.sh
./build_image.sh
```

2. Soumettez la tâche RUN : ai pour télécharger, extraire, pré-traiter et stocker le jeu de données Tusimple Lane Detection dans un `pvc`, Qui est créé de manière dynamique par NetApp Trident :

a. Utiliser les commandes suivantes pour envoyer l'EXÉCUTION du travail ai :

```
runai submit
--name download-tusimple-data
--pvc azurenetappfiles:100Gi:/mnt
--image muneer7589/download-tusimple:1.0
```

b. Saisissez les informations du tableau ci-dessous pour soumettre le travail RUN:ai :

Champ	Valeur ou description
-nom	Nom du travail
-pvc	PVC du format [StorageClassName]:Size:ContainerMountPath dans la soumission de tâche ci-dessus, vous créez un PVC basé sur la demande à l'aide de Trident avec la classe de stockage azurenetappfiles. La capacité de volume persistant est ici de 100Gi et elle est montée sur le chemin /mnt.
-image	Image Docker à utiliser lors de la création du conteneur pour cette tâche

Le résultat de la commande doit ressembler à l'exemple suivant :

```
The job 'download-tusimple-data' has been submitted successfully
You can run `runai describe job download-tusimple-data -p lane-detection` to check the job status
```

c. Répertorier les travaux RUN:ai soumis.

```
runai list jobs
```

```
Showing jobs for project lane-detection
NAME                  STATUS            AGE  NODE                          IMAGE                          TYPE  PROJECT        USER          GPUs Allocated (Requested)
PODs Running (Pending)  SERVICE URL(S)
download-tusimple-data  ContainerCreating  1m   aks-agentpool-34613062-vmss00000a  muneer7589/download-tusimple:1.0  Train  lane-detection  verronmartina  0 (0)
1 (0)
```

d. Vérifiez les journaux des travaux soumis.

```
runai logs download-tusimple-data -t 10
```

```
751150K .......... .......... .......... .......... ..........  6% 16.2M 20m37s
751200K .......... .......... .......... .......... ..........  6% 11.1M 20m37s
751250K .......... .......... .......... .......... ..........  6% 12.5M 20m36s
751300K .......... .......... .......... .......... ..........  6% 11.3M 20m36s
751350K .......... .......... .......... .......... ..........  6% 15.2M 20m36s
751400K .......... .......... .......... .......... ..........  6% 10.5M 20m36s
751450K .......... .......... .......... .......... ..........  6% 15.2M 20m36s
751500K .......... .......... .......... .......... ..........  6% 14.1M 20m36s
751550K .......... .......... .......... .......... ..........  6% 24.3M 20m36s
751600K .......... .......... .......... .......... ..........  6% 26.3M 20m36s
```

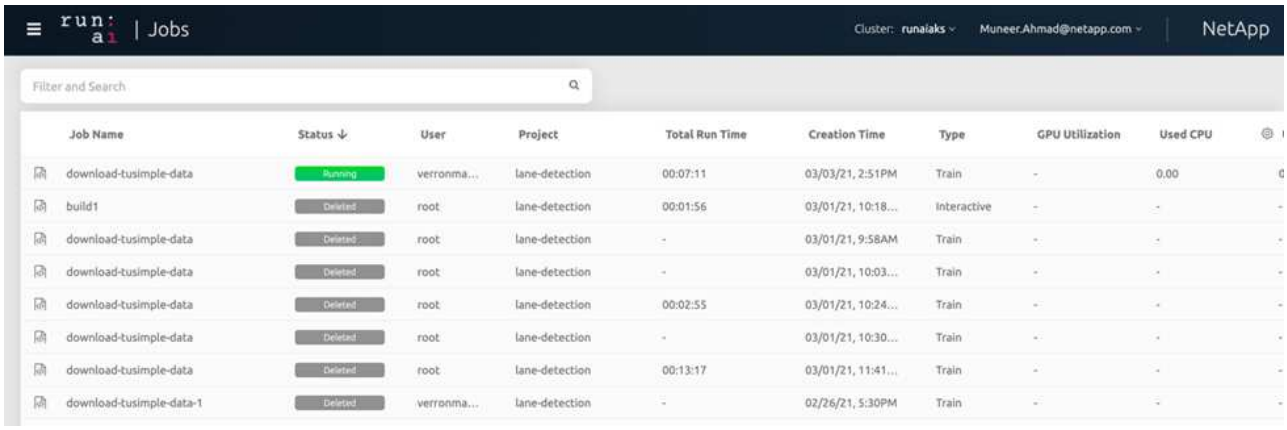
e. Énumérez le `pvc` créé. Utilisez-le `pvc` commande pour la formation à l'étape suivante.

```
kubectl get pvc | grep download-tusimple-data
```

Le résultat de la commande doit ressembler à l'exemple suivant :

```
pvc-download-tusimple-data-0   Bound    pvc-bb03b74d-2c17-40c4-a445-79f3de8d16d5   100Gi      RWO            azurenetappfiles   4m47s
```

a. Vérifiez le travail EN COURS D'EXÉCUTION : ai UI (ou `app.run.ai`).



Effectuer une formation sur la détection de voie distribuée à l'aide de Horovod

La formation sur la détection de voie distribuée à l'aide de Horovod est un processus facultatif. Notez toutefois que voici les étapes impliquées :

1. Créez et poussez l'image docker ou ignorez cette étape pour utiliser l'image docker existante (par exemple, `muneer7589/dist-lane-detection:3.1)`:

   a. Basculez vers le répertoire de base.

   ```
   cd ~
   ```

   b. Accédez au répertoire du projet `lane-detection-SCNN-horovod.`

   ```
   cd ./lane-detection-SCNN-horovod
   ```

   c. Modifiez le `build_image.sh` script shell et remplacez le référentiel docker par le vôtre (par exemple, remplacez-le `muneer7589` avec le nom de votre référentiel docker). Vous pouvez également modifier le nom et LA BALISE de l'image docker (`dist-lane-detection` et `3.1, for example)`.

```
#!/bin/bash
#
# A simple script to build the distributed Docker image.
#
# $ build_image.sh
set -ex

IMAGE=muneer7589/dist-lane-detection
TAG=3.0

# Build image
echo "Building image: "$IMAGE
docker build . -f Dockerfile \
 --tag "${IMAGE}:${TAG}"
echo "Finished building image: "$IMAGE

# Push image
echo "Pushing image: "$IMAGE
docker push "${IMAGE}:${TAG}"
echo "Finished pushing image: "$IMAGE
```

d. Exécutez le script pour créer l'image docker et l'envoyer dans le référentiel docker.

```
chmod +x build_image.sh
./build_image.sh
```

2. Soumettre le COURSE : travail d'IA pour la formation distribuée (MPI) :

a. Utilisation de l'option Submit of RUN : l'IA pour la création automatique de volume persistant à l'étape précédente (pour le téléchargement des données) vous permet uniquement d'avoir un accès RWO, qui permet non à plusieurs pods ou nœuds d'accéder au même volume persistant pour l'entraînement distribué. Mettez à jour le mode d'accès sur ReadWriteMany et utilisez le patch Kubernetes pour le faire.

b. Commencez par obtenir le nom du volume de la demande de volume persistant en exécutant la commande suivante :

```
kubectl get pvc | grep download-tusimple-data
```

```
root@ai-w-gpu-2:/mnt/ai_data/anf_runai/lane-detection-SCNN-horovod# kubectl get pvc | grep download-tusimple-data
pvc-download-tusimple-data-0   Bound    pvc-bb03b74d-2c17-40c4-a445-79f3de8d16d5   100Gi      RWX            azurenetappfiles   2d4h
```

c. Corriger le volume et mettre à jour le mode d'accès à ReadWriteMany (remplacer le nom du volume par le vôtre dans la commande suivante) :

```
kubectl patch pv pvc-bb03b74d-2c17-40c4-a445-79f3de8d16d5 -p
'{"spec":{"accessModes":["ReadWriteMany"]}}'
```

d. Soumettre le STAGE : ai MPI pour l'exécution du travail de formation répartie en utilisant les informations du tableau ci-dessous :

```
runai submit-mpi
--name dist-lane-detection-training
--large-shm
--processes=3
--gpu 1
--pvc pvc-download-tusimple-data-0:/mnt
--image muneer7589/dist-lane-detection:3.1
-e USE_WORKERS="true"
-e NUM_WORKERS=4
-e BATCH_SIZE=33
-e USE_VAL="false"
-e VAL_BATCH_SIZE=99
-e ENABLE_SNAPSHOT="true"
-e PVC_NAME="pvc-download-tusimple-data-0"
```

Champ	Valeur ou description
nom	Nom du travail de formation distribué
grand shm	Montage d'un périphérique grand /dev/shm il s'agit d'un système de fichiers partagé monté sur la RAM et fournit suffisamment de mémoire partagée pour que plusieurs collaborateurs du processeur puissent traiter et charger des lots dans la RAM du CPU.
processus	Nombre de processus de formation distribués
gpu	Nombre de GPU/processus à allouer pour le travail, trois processus utilisateur sont nécessaires (--processus=3), chacun étant alloué avec un seul GPU (--gpu 1).
pvc	Utilisez le volume persistant existant (pvc-download-tusimple-data-0) créé par le travail précédent (download-tusimple-data) et monté sur le chemin /mnt
image	Image Docker à utiliser lors de la création du conteneur pour cette tâche
Définissez les variables d'environnement à définir dans le conteneur	
EMPLOYÉS_UTILISÉS	Le fait de définir l'argument sur true active le chargement de données multi-processus
NOMBRE_D'EMPLOYÉS	Nombre de processus de travail du chargeur de données
TAILLE_LOT	Taille des lots d'entraînement
USE_VAL	Le fait de définir l'argument sur vrai permet la validation
VAL_BATCH_SIZE	Taille du lot de validation

Champ	Valeur ou description
ACTIVER_SNAPSHOT	La définition de l'argument sur true permet de prendre des données et des snapshots de modèles entraînés à des fins de gestion des versions DU ML
NOM_PVC	Nom de la demande de volume persistant pour créer un snapshot de. Dans la soumission des travaux ci-dessus, vous prenez un snapshot de pvc-download-tsimple-Data-0, composé d'un dataset et de modèles entraînés

Le résultat de la commande doit ressembler à l'exemple suivant :

```
The job 'dist-lane-detection-training' has been submitted successfully
You can run `runai describe job dist-lane-detection-training -p lane-detection` to check the job status
```

e. Répertorier le travail soumis.

```
runai list jobs
```

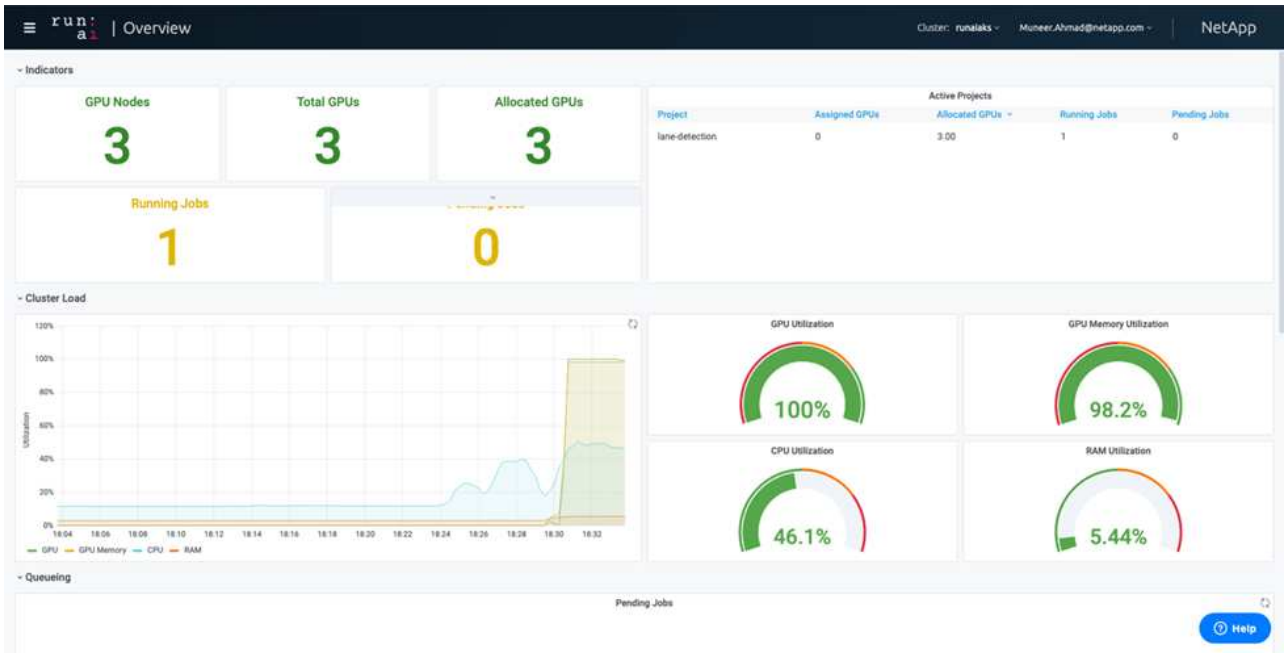
```
NAME                          STATUS     AGE  NODE        IMAGE                              TYPE   PROJECT         USER           GPUs Allocated (Requested)  PODs
 SERVICE URL(S)
download-tusimple-data        Succeeded  1d               muneer7589/download-tusimple:1.0   Train  lane-detection  verronmartina  - (0)                       0 (0)

dist-lane-detection-training  Init:0/1   2m   <multiple>  muneer7589/dist-lane-detection:3.1 Train  lane-detection  root           3 (3)                       4 (0)
```

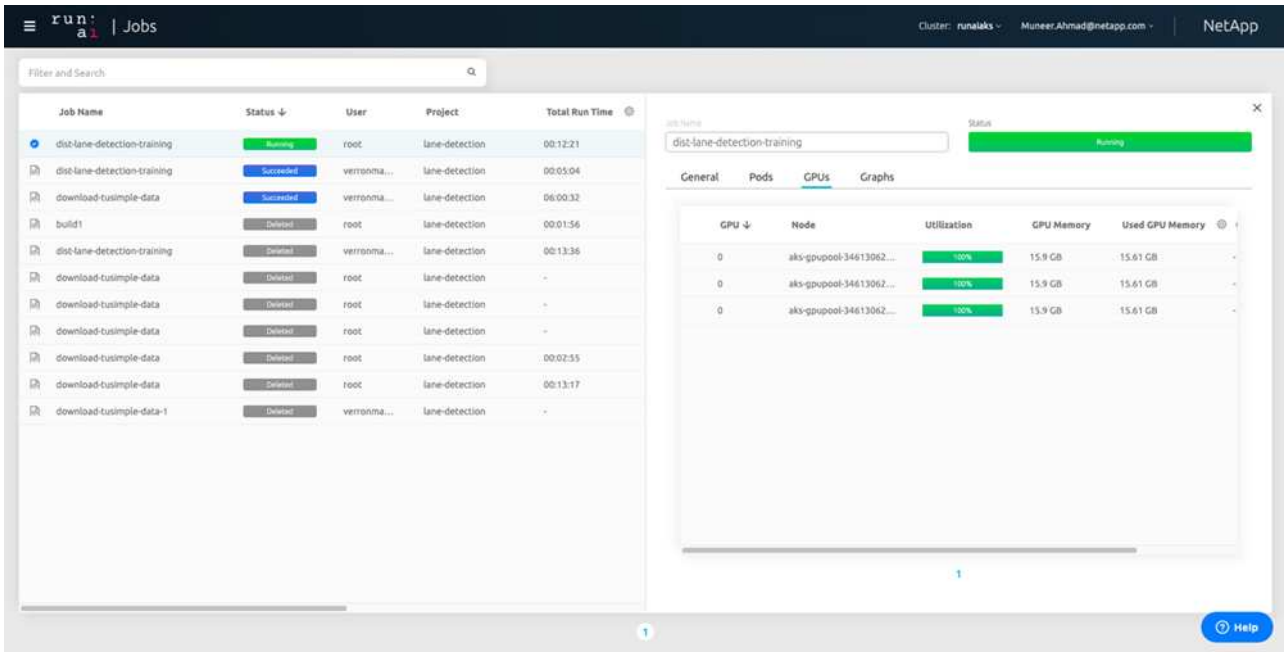
f. Journaux des travaux soumis :

```
runai logs dist-lane-detection-training
```

```
root@ai-w-gpu-2:~/runai# runai logs dist-lane-detection-training
Running with 3 workers
2021-03-04 17:29:23.158449: I tensorflow/stream_executor/platform/default/dso_loader.cc:48] Successfully opened dynamic library libcudart.so.10.1
+ POD_NAME=dist-lane-detection-training-worker-0
+ [ d = - ]
+ shift
+ /opt/kube/kubectl cp /opt/kube/hosts dist-lane-detection-training-worker-0:/etc/hosts_of_nodes
+ POD_NAME=dist-lane-detection-training-worker-2
+ [ d = - ]
+ shift
+ /opt/kube/kubectl cp /opt/kube/hosts dist-lane-detection-training-worker-2:/etc/hosts_of_nodes
+ POD_NAME=dist-lane-detection-training-worker-1
```

g. Consulter la tâche d'entraînement EXÉCUTÉE : GUI d'IA (ou app.runai.ai): SESSIONS : tableau de bord d'IA, comme le montre les figures ci-dessous). La première figure présente trois processeurs graphiques alloués à la tâche d'entraînement distribuée sur trois nœuds sur AKS, puis la seconde SESSION :ai Jobs :





h. Une fois l’entraînement terminé, vérifiez la copie NetApp Snapshot créée et associée à L’EXÉCUTION du travail : IA.

```
runai logs dist-lane-detection-training --tail 1
```

```
[1,0]<stdout>:Snapshot snap-pvc-download-tusimple-data-0-dist-lane-detection-training-launcher-2021-03-05-16-23-42 created in namespace runai-lane-detection
```

```
kubectl get volumesnapshots | grep download-tusimple-data-0
```

Restaurez les données à partir de la copie NetApp Snapshot

Pour restaurer les données à partir de la copie NetApp Snapshot, effectuez la procédure suivante :

1. Basculez vers le répertoire de base.

```
cd ~
```

2. Accédez au répertoire du projet `lane-detection-SCNN-horovod`.

```
cd ./lane-detection-SCNN-horovod
```

3. Modifier `restore-snaphot-pvc.yaml` et mettre à jour `dataSource` `name` Champ de la copie Snapshot à partir duquel vous souhaitez restaurer les données. Vous pouvez également modifier le nom du volume persistant dans lequel les données seront restaurées, dans cet exemple son `restored-tusimple`.

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: restored-tusimple
spec:
  storageClassName: azurenetappfiles
  dataSource:
    name: snap-pvc-download-tusimple-data-0-dist-lane-detection-training-launcher-2021-03-05-16-23-42
    kind: VolumeSnapshot
    apiGroup: snapshot.storage.k8s.io
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 100Gi
```

4. Créez un nouveau PVC à l'aide de `restore-snapshot-pvc.yaml`.

```
kubectl create -f restore-snapshot-pvc.yaml
```

Le résultat de la commande doit ressembler à l'exemple suivant :

```
persistentvolumeclaim/restored-tusimple created
```

5. Si vous souhaitez utiliser les données restaurées pour l'entraînement, la candidature reste la même qu'auparavant. Remplacez uniquement la `PVC_NAME` avec le restauré `PVC_NAME` lors de la soumission du travail de formation, comme l'indique les commandes suivantes :

```
runai submit-mpi
--name dist-lane-detection-training
--large-shm
--processes=3
--gpu 1
--pvc restored-tusimple:/mnt
--image muneer7589/dist-lane-detection:3.1
-e USE_WORKERS="true"
-e NUM_WORKERS=4
-e BATCH_SIZE=33
-e USE_VAL="false"
-e VAL_BATCH_SIZE=99
-e ENABLE_SNAPSHOT="true"
-e PVC_NAME="restored-tusimple"
```

Évaluation des performances

Pour montrer l'évolutivité linéaire de la solution, des tests de performance ont été réalisés dans deux scénarios : un GPU et trois GPU. L'allocation du GPU, l'utilisation du GPU et de la mémoire, différents metrics à un ou trois nœuds ont été capturés lors de l'entraînement sur le dataset de détection Tulane simple. Les données sont multiplié par cinq dans le seul but d'analyser l'utilisation des ressources au cours des processus d'entraînement.

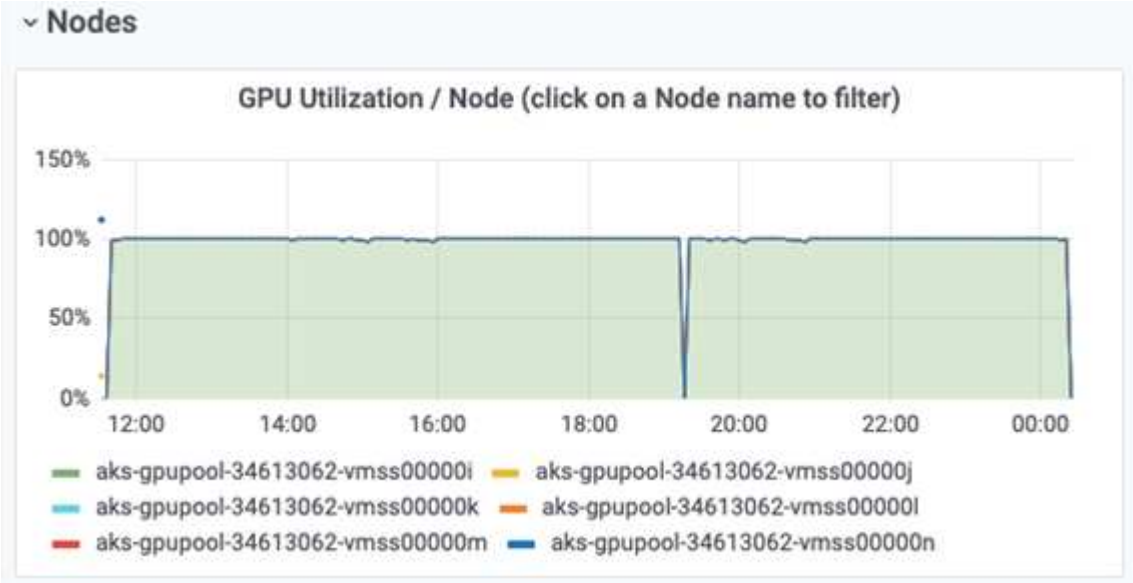
La solution permet de commencer avec un petit dataset et quelques GPU. Lorsque le volume de données et la demande de GPU augmentent, les clients peuvent faire évoluer horizontalement dynamiquement les téraoctets dans le niveau standard et monter jusqu'au niveau Premium pour obtenir un débit par téraoctet sans déplacer de données. Ce processus est expliqué plus en détail dans la section, "Niveaux de service Azure NetApp Files".

Le temps de traitement d'un GPU était de 12 heures et 45 minutes. Le temps de traitement sur trois GPU sur trois nœuds était d'environ 4 heures et 30 minutes.

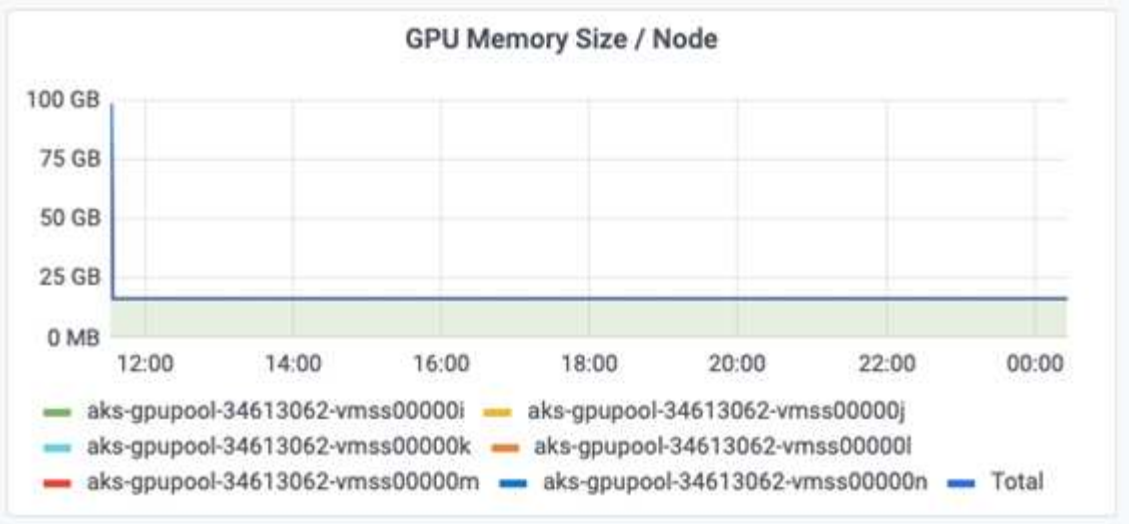
Les chiffres présentés dans la suite de ce document illustrent des exemples de performances et d'évolutivité en fonction des besoins spécifiques de l'entreprise.

La figure ci-dessous illustre l'allocation de 1 GPU et l'utilisation de la mémoire.

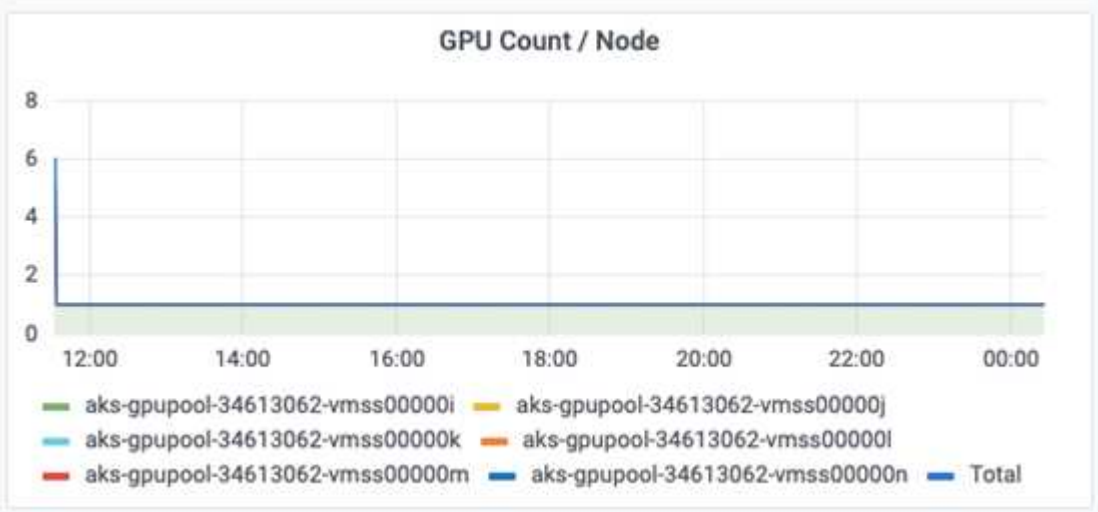
La figure ci-dessous illustre l'utilisation des GPU d'un nœud.



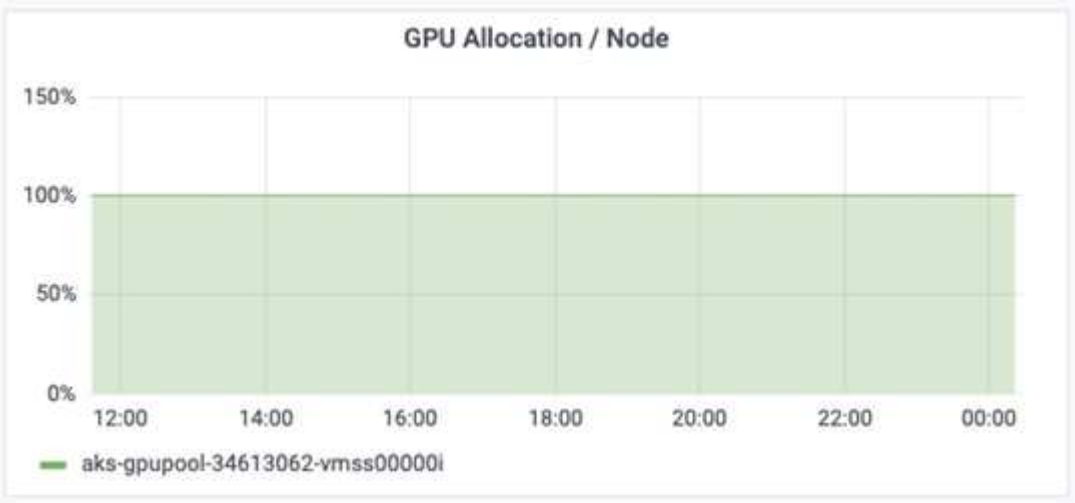
La figure ci-dessous illustre la taille de la mémoire d'un nœud unique (16 Go).



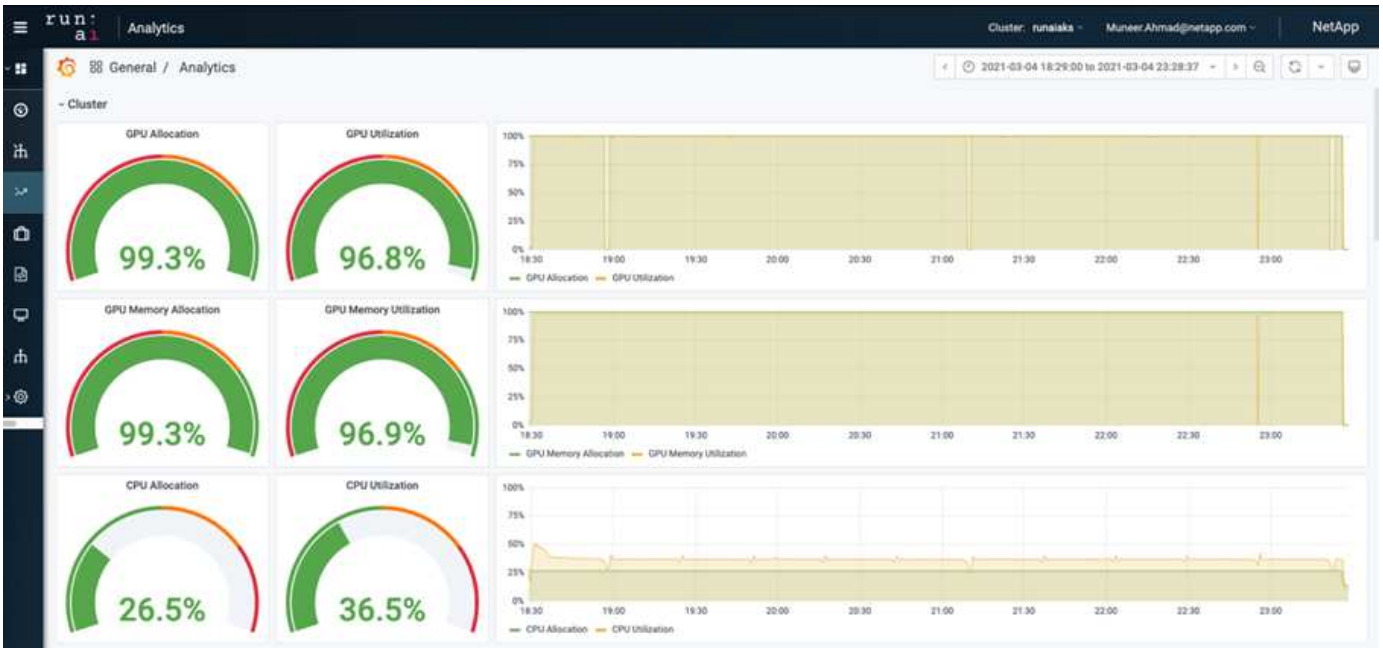
La figure ci-dessous illustre le nombre de GPU d'un nœud (1).



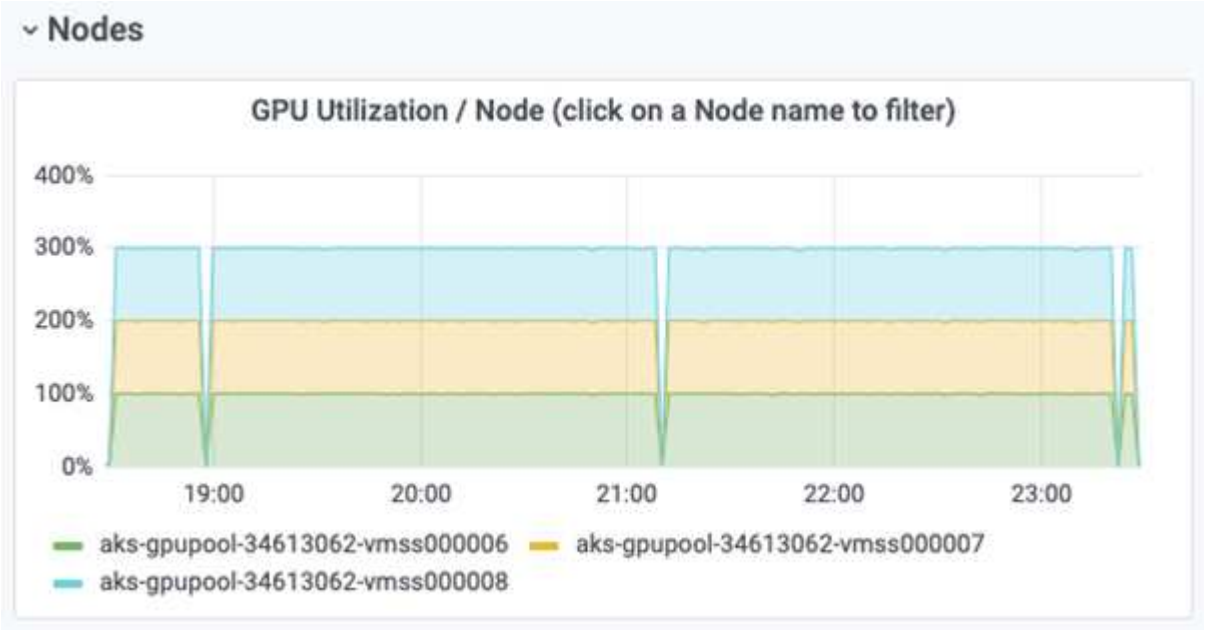
La figure ci-dessous illustre l'allocation de GPU d'un nœud (%).



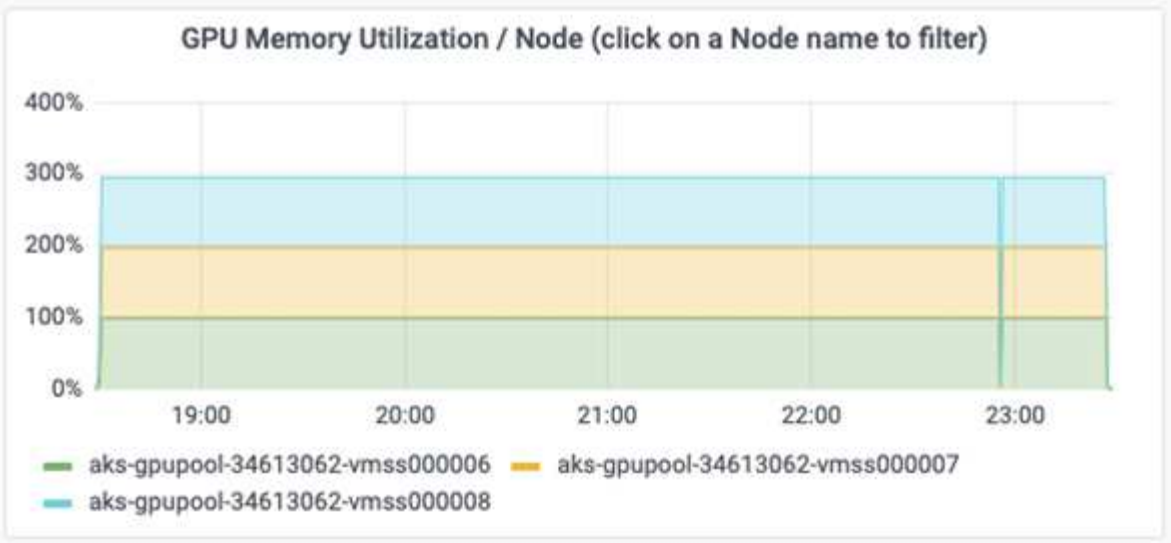
La figure ci-dessous illustre trois GPU répartis sur trois nœuds : l'allocation des GPU et la mémoire.



La figure ci-dessous illustre le taux d'utilisation de trois GPU sur trois nœuds (%).



La figure ci-dessous illustre l'utilisation de la mémoire de trois nœuds sur trois GPU (%).



Niveaux de service Azure NetApp Files

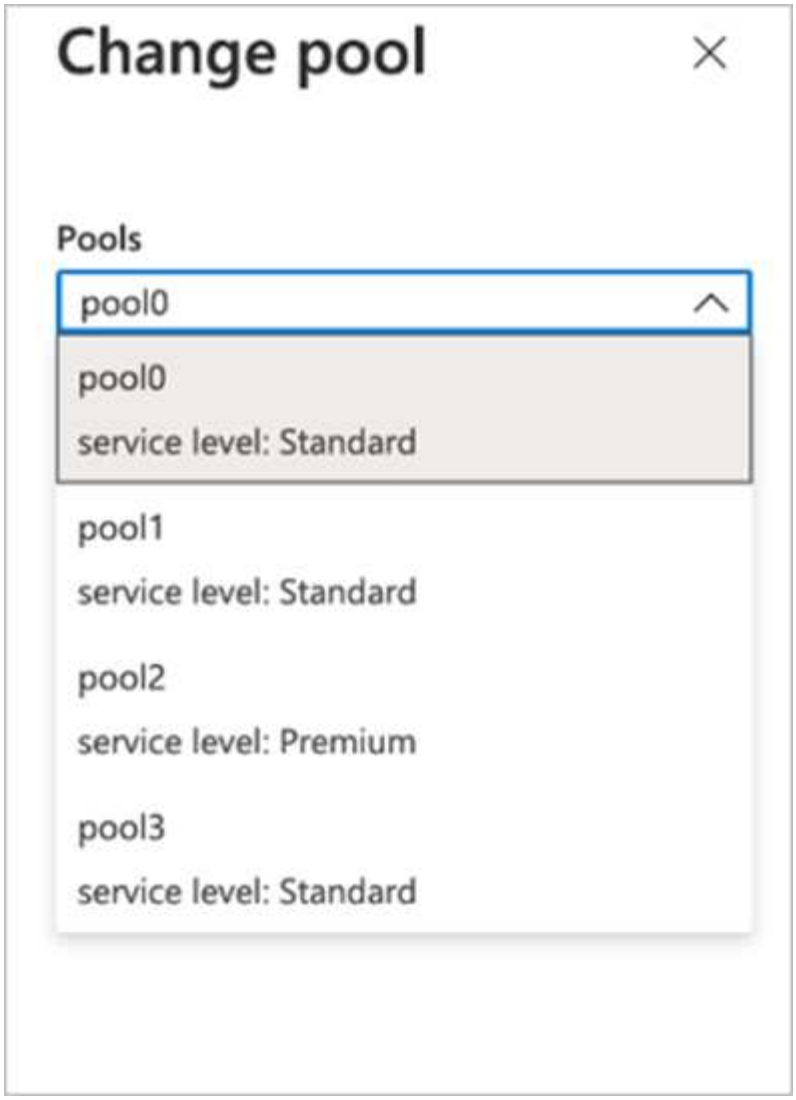
Vous pouvez modifier le niveau de service d'un volume existant en déplaçant ce volume vers un autre pool de capacité qui utilise le "niveau de service" vous voulez le volume. Cette modification de niveau de service existante pour le volume n'exige pas la migration des données. Elle n'affecte pas non plus l'accès au volume.

Modification dynamique du niveau de service d'un volume

Pour modifier le niveau de service d'un volume, procédez comme suit :

1. Sur la page volumes, cliquez avec le bouton droit de la souris sur le volume dont vous souhaitez modifier le niveau de service. Sélectionnez Modifier le pool.

2. Dans la fenêtre change Pool, sélectionnez le pool de capacité vers lequel vous souhaitez déplacer le volume. Cliquez ensuite sur OK.



Automatisez les changements de niveau de services

Le changement de niveau de service dynamique est actuellement dans l'aperçu public, mais il n'est pas activé par défaut. Pour activer cette fonction sur l'abonnement Azure, suivez les étapes indiquées dans le document « "Modification dynamique du niveau de service d'un volume". »

- Vous pouvez également utiliser les commandes suivantes pour Azure : interface de ligne de commandes. Pour plus d'informations sur la modification de la taille du pool de Azure NetApp Files, rendez-vous sur "Volume az netappfiles : gestion des ressources de volume Azure NetApp Files (ANF)".

```
az netappfiles volume pool-change -g mygroup
--account-name myaccname
-pool-name mypoolname
--name myvolname
--new-pool-resource-id mynewresourceid
```

- Le `set- aznetappfilesvolumepool` Cmdlet affichée ici peut modifier le pool d'un volume Azure NetApp Files. Pour plus d'informations sur la modification de la taille du pool de volumes et d'Azure PowerShell, rendez-vous sur "Modification du pool d'un volume Azure NetApp Files".

```
Set-AzNetAppFilesVolumePool
-ResourceGroupName "MyRG"
-AccountName "MyAnfAccount"
-PoolName "MyAnfPool"
-Name "MyAnfVolume"
-NewPoolResourceId 7d6e4069-6c78-6c61-7bf6-c60968e45fbf
```

Conclusion

NetApp et EXÉCUTION : l'IA se sont associées à la création de ce rapport technique pour présenter les fonctionnalités uniques de Azure NetApp Files et DE L'EXÉCUTION : une plateforme d'IA qui simplifie l'orchestration des workloads d'IA. Ce rapport technique fournit une architecture de référence pour rationaliser le processus des pipelines de données et de l'orchestration de la charge de travail pour la formation sur la détection de voies distribuées.

En conclusion, en ce qui concerne la formation distribuée à grande échelle (notamment dans un environnement de cloud public), le composant d'orchestration et de stockage des ressources est essentiel à la solution. S'assurer que la gestion des données n'entrave pas le traitement plusieurs GPU, ce qui optimise l'utilisation des cycles GPU. Ainsi, le système est aussi rentable que possible à des fins de formation distribuée à grande échelle.

La Data Fabric fournie par NetApp relève le défi en permettant aux data Scientists et aux ingénieurs de données de se connecter sur site et dans le cloud à des données synchrones, sans aucune intervention manuelle. En d'autres termes, la Data Fabric opère en douceur pour gérer les workflows d'IA répartis sur plusieurs sites. Cette solution facilite également la disponibilité des données à la demande en rapprochant les données de calcul et en exécutant des opérations d'analyse, d'entraînement et de validation, quand et où vous en avez besoin. Cette fonctionnalité permet non seulement l'intégration des données, mais aussi la protection et la sécurité de l'ensemble du pipeline de données.

Informations supplémentaires

Pour en savoir plus sur les informations données dans ce livre blanc, consultez ces documents et/ou sites web :

- Jeu de données : simple

  "https://github.com/TuSimple/tusimple-benchmark/tree/master/doc/lane_detection"

- Architecture de réseau d'apprentissage profond : réseau neuronal spatial

  "https://arxiv.org/abs/1712.06080"

- Framework de formation d'apprentissage profond distribué : Horovod

  "https://horovod.ai/"

- EXÉCUTION : solution d'orchestration de conteneurs d'IA : EXÉCUTION : présentation d'un produit d'IA

"https://docs.run.ai/home/components/"

- EXÉCUTION : documentation d'installation d'IA

  "https://docs.run.ai/Administrator/Cluster-Setup/cluster-install/#step-3-install-runai"
  "https://docs.run.ai/Administrator/Researcher-Setup/cli-install/#runai-cli-installation"

- Soumission des tâches EN COURS D'EXÉCUTION : interface de ligne de commande d'IA

  "https://docs.run.ai/Researcher/cli-reference/runai-submit/"

  "https://docs.run.ai/Researcher/cli-reference/runai-submit-mpi/"

- Ressources cloud Azure : Azure NetApp Files

  "https://docs.microsoft.com/azure/azure-netapp-files/"

- Service Azure Kubernetes

  "https://azure.microsoft.com/services/kubernetes-service/-features"

- Références de machines virtuelles Azure

  "https://azure.microsoft.com/services/virtual-machines/"

- Machine virtuelle Azure avec références de processeurs graphiques

  "https://docs.microsoft.com/azure/virtual-machines/sizes-gpu"

- NetApp Trident

  "https://github.com/NetApp/trident/releases"

- Data Fabric optimisée par NetApp

  "https://www.netapp.com/data-fabric/what-is-data-fabric/"

- Documentation produit NetApp

  "https://www.netapp.com/support-and-training/documentation/"

Tr-4841 : système d'exploitation pour l'IA dans le cloud hybride avec mise en cache des données

Rick Huang, David Arnette, NetApp Yochay Ettun, cnvrg.io

La croissance exponentielle des données et la croissance exponentielle DU ML et de l'IA ont convergé pour créer une économie de zettaoctet dotée de défis uniques en matière de développement et d'implémentation.

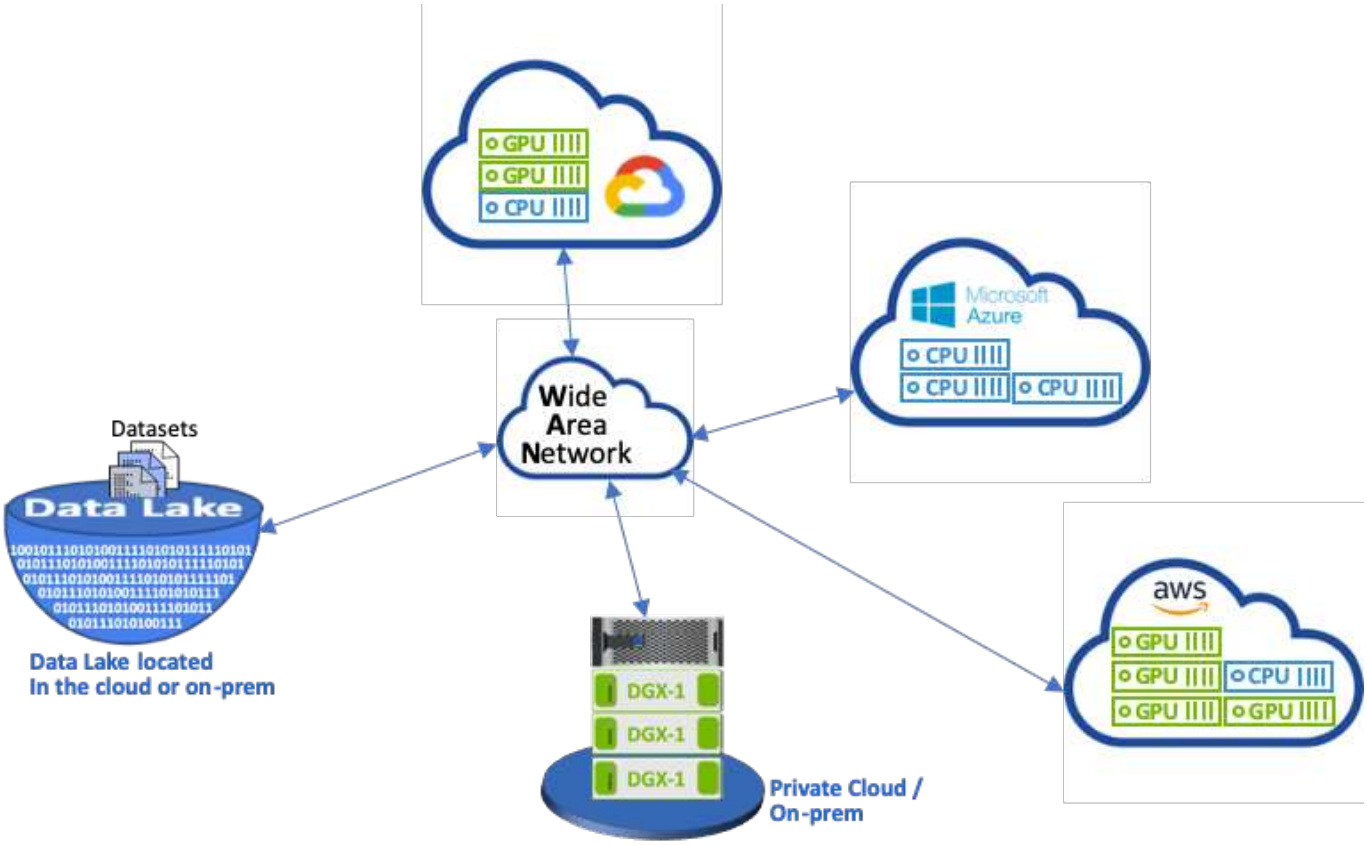
Bien que tous les modèles DE ML aient besoin de données gourmandes en ressources de stockage haute performance, l'implémentation de ce modèle n'est pas beaucoup plus directe, en particulier avec le cloud hybride et les instances de calcul flexibles. Des quantités massives de données sont généralement stockées dans des data Lakes de faible coût, où les ressources de calcul d'IA haute performance telles que les GPU ne

peuvent pas y accéder de manière efficace. Ce problème est aggravé dans l'infrastructure du cloud hybride où certaines charges de travail s'exécutent dans le cloud et certaines se trouvent sur site ou dans un environnement HPC différent.

Dans ce document, nous présentons une solution inédite qui permet aux professionnels DE L'IT et aux ingénieurs de données de créer une plateforme d'IA de cloud hybride véritablement équipée d'un hub de données capable de créer instantanément et automatiquement un cache de leurs jeux de données à proximité de leurs ressources de calcul. où qu'ils soient situés. Par conséquent, l'entraînement des modèles haute performance est non seulement possible, mais de nouveaux avantages sont également créés, notamment la collaboration de plusieurs professionnels de l'IA, qui disposent d'un accès immédiat aux caches, aux versions et aux lignées de dataset dans un hub de version du dataset.

Présentation du cas d'utilisation et déclaration du problème

Les datasets et les versions des datasets se trouvent généralement dans un data Lake, tel que le stockage objet StorageGRID de NetApp, ce qui permet de réduire les coûts et d'autres avantages opérationnels. Les data Scientists extraient ces datasets et les préparent en plusieurs étapes pour l'entraînement d'un modèle spécifique, créant souvent plusieurs versions tout au long du processus. À l'étape suivante, le data Scientist doit sélectionner des ressources de calcul optimisées (GPU, instances de processeurs haut de gamme, cluster sur site, etc.) pour exécuter le modèle. La figure suivante illustre l'absence de proximité du dataset dans un environnement de calcul DE ML.



Toutefois, plusieurs expériences d'entraînement doivent être menées en parallèle dans différents environnements de calcul. Chacune d'entre elles nécessite le téléchargement du dataset depuis le data Lake, ce qui est un processus coûteux et chronophage. La proximité du dataset avec l'environnement de calcul (notamment dans le cloud hybride) n'est pas garantie. De plus, les autres membres de l'équipe qui exécutent

leurs propres expériences avec le même dataset doivent suivre le même processus fastidieux. Au-delà de la lenteur évidente de l'accès aux données, le suivi des versions de dataset, le partage de dataset, la collaboration et la reproductibilité sont des difficultés autres.

Besoins des clients

Les exigences des clients peuvent varier afin d'atteindre les performances d'EXÉCUTION D'APPRENTISSAGE MACHINE tout en utilisant efficacement les ressources. Par exemple, les clients peuvent exiger les éléments suivants :

- L'accès rapide aux datasets à partir de chaque instance de calcul exécutant le modèle d'entraînement, sans complexité coûteuse liée à l'accès aux données ni à des téléchargements
- L'utilisation d'une instance de calcul (GPU ou processeur) dans le cloud ou sur site sans craindre l'emplacement des datasets
- Efficacité et productivité accrues grâce aux tests d'entraînement exécutés en parallèle avec les différentes ressources de calcul sur le même dataset, sans délais supplémentaires et sans latence des données
- Réduction des coûts des instances de calcul
- Meilleure reproductibilité grâce à des outils permettant de conserver les enregistrements des ensembles de données, leur traçabilité, leurs versions et d'autres détails de métadonnées
- Partage et collaboration améliorés pour que les membres de l'équipe autorisés puissent accéder aux données et réaliser des expériences

Pour implémenter la mise en cache du dataset avec le logiciel de gestion des données NetApp ONTAP, les clients doivent effectuer les tâches suivantes :

- Configurez et définissez le stockage NFS le plus proche des ressources de calcul.
- Détermination du jeu de données et de la version à mettre en cache
- Contrôlez la mémoire totale allouée aux datasets en cache et la quantité de stockage NFS disponible pour les validations de cache supplémentaires (par exemple, la gestion du cache).
- L'âge des datasets dans le cache s'ils n'ont pas été utilisés pendant un certain temps. La valeur par défaut est un jour ; d'autres options de configuration sont disponibles.

Présentation de la solution

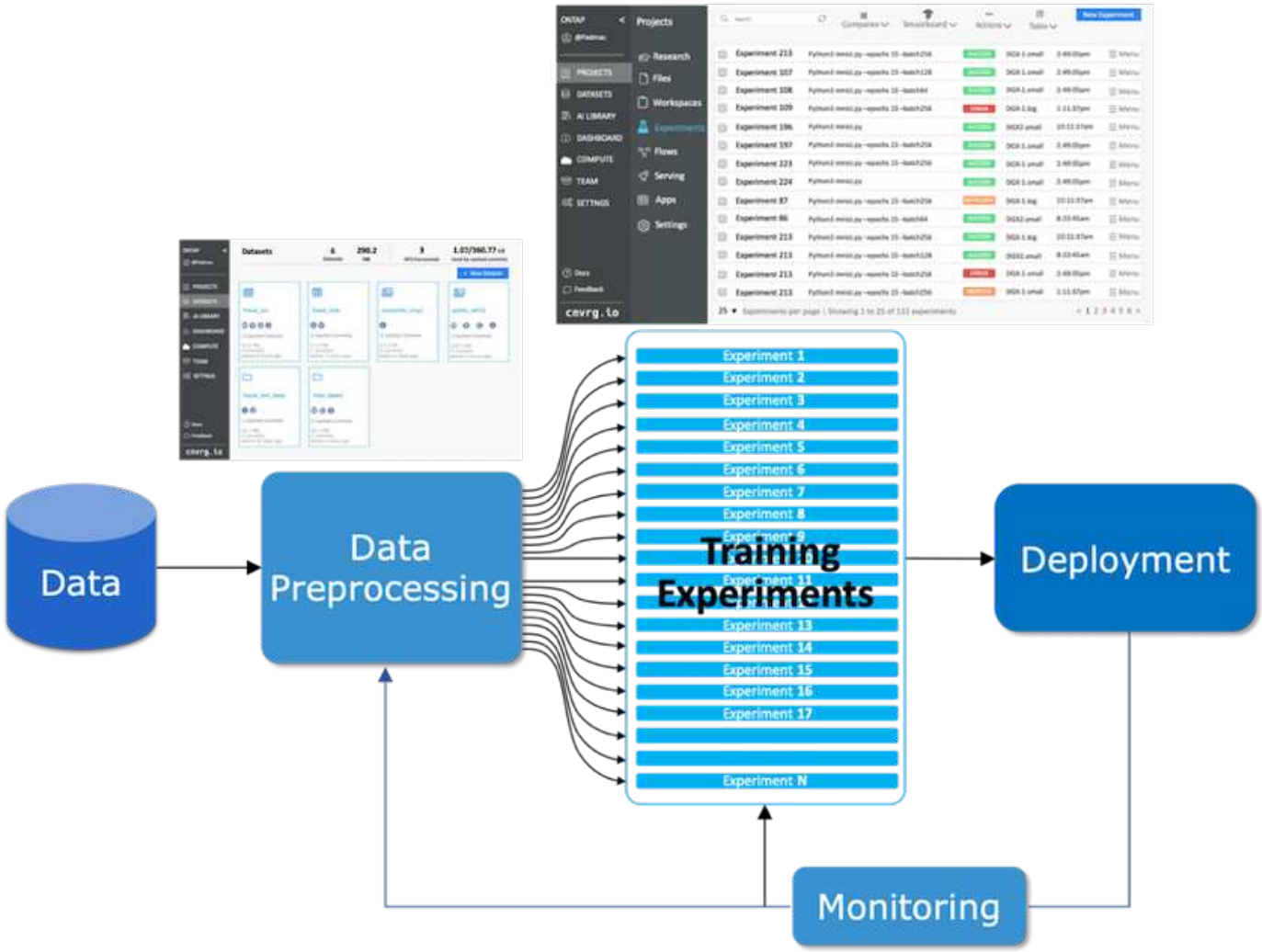
Cette section examine un pipeline conventionnel de data science et ses inconvénients. Il présente également l'architecture de la solution de mise en cache de jeux de données proposée.

Pipeline de traitement de données conventionnel et inconvénients

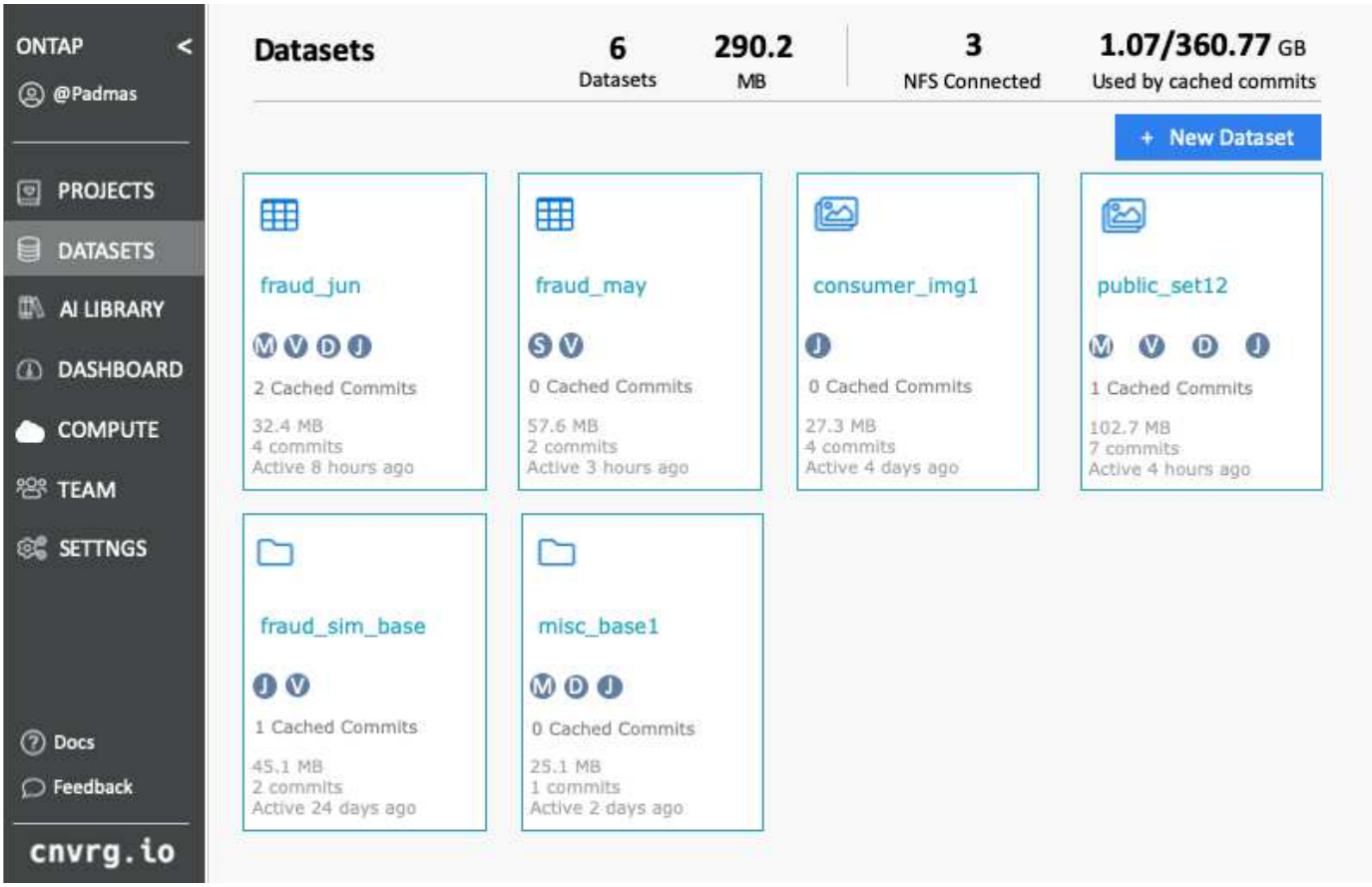
Une séquence standard de développement et de déploiement de modèles DE ML implique des étapes itératives qui incluent :

- Ingestion des données
- Prétraitement des données (création de plusieurs versions des datasets)
- Exécution de plusieurs expériences impliquant l'optimisation des hyperparamètres, différents modèles, etc
- Déploiement
- Monitoringcnvrg.io a développé une plateforme complète pour automatiser toutes les tâches, de la

recherche au déploiement. Un petit exemple de captures d'écran du tableau de bord concernant le pipeline est présenté dans la figure suivante.



Il est très courant d'avoir plusieurs jeux de données dans les référentiels publics et les données privées. En outre, plusieurs versions sont possibles pour chaque jeu de données suite à un nettoyage de jeux de données ou à l'ingénierie des fonctionnalités. Un tableau de bord fournissant un hub de jeux de données et une version Hub est nécessaire pour garantir la disponibilité des outils de collaboration et de cohérence à l'équipe, comme l'illustre la figure suivante.



L'étape suivante du pipeline est l'entraînement, qui nécessite plusieurs instances parallèles de modèles d'entraînement, chacun associé à un dataset et à une instance de calcul spécifique. La liaison d'un dataset à une certaine instance de calcul est un défi, car il est possible que certaines expériences soient réalisées par des instances GPU d'Amazon Web Services (AWS), tandis que d'autres expériences sont réalisées par des instances DGX-1 ou DGX-2 sur site. D'autres expériences peuvent être exécutées sur des serveurs CPU dans GCP, tandis que l'emplacement du dataset n'est pas à proximité des ressources de calcul lors de l'entraînement. Une proximité raisonnable aurait une connectivité complète 10 GbE ou plus faible latence entre le stockage du dataset et l'instance de calcul.

Les data Scientists doivent en effet télécharger le dataset sur l'instance de calcul pour effectuer l'entraînement et l'expérience. Toutefois, cette approche peut rencontrer plusieurs problèmes :

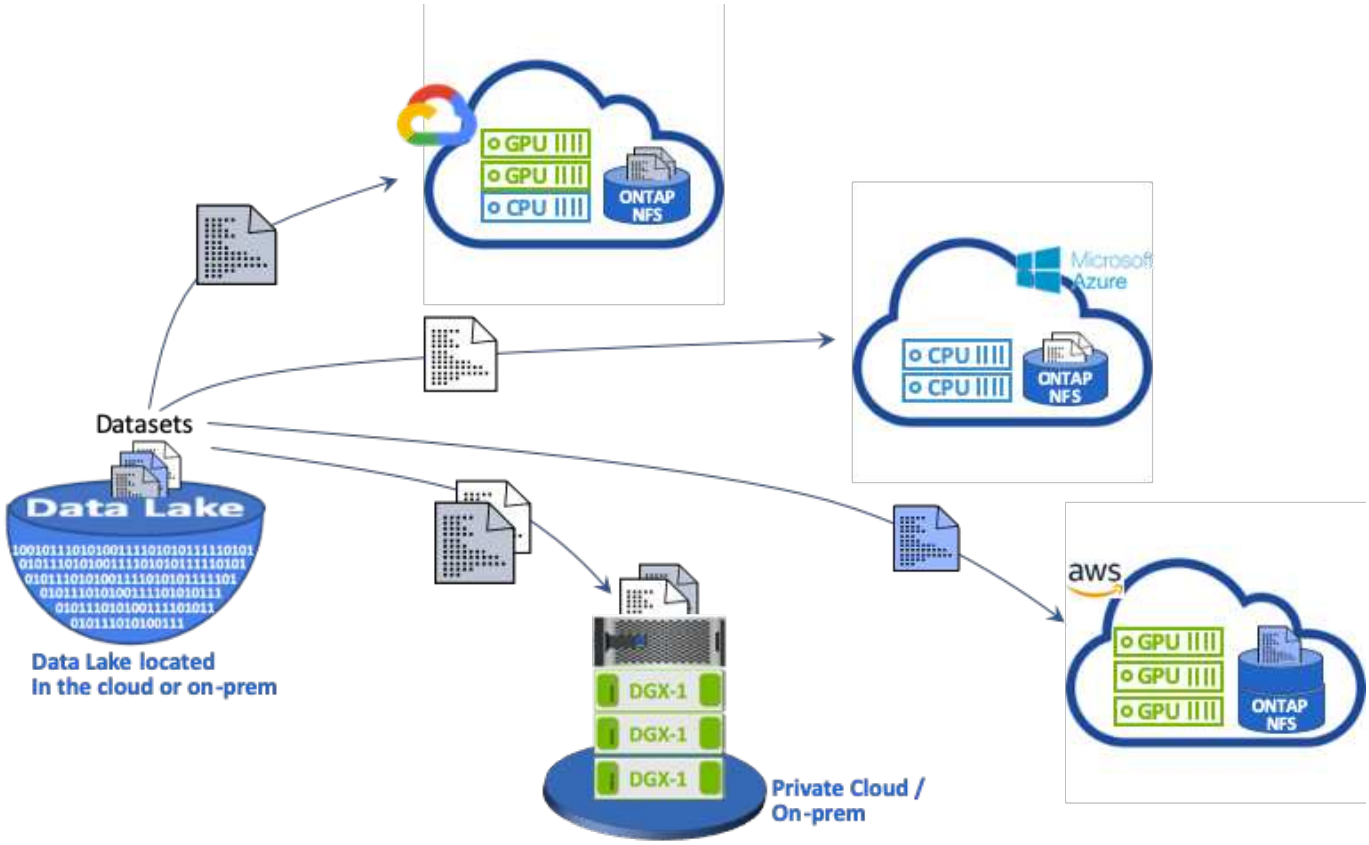
- Lorsque l'analyste de données télécharge le dataset sur une instance de calcul, aucune garantie ne garantit que le stockage de calcul intégré est performant (un exemple de système haute performance serait la solution NVMe de ONTAP AFF A800).
- Lorsque le dataset téléchargé réside dans un nœud de calcul, le stockage peut former un goulot d'étranglement lorsque les modèles distribués sont exécutés sur plusieurs nœuds (contrairement au stockage distribué haute performance NetApp ONTAP).
- La prochaine itération de l'expérience d'entraînement peut être réalisée dans une autre instance de calcul en raison de conflits ou de priorités de file d'attente, créant ainsi une distance réseau considérable entre le dataset et l'emplacement de calcul.
- Les autres membres de l'équipe effectuant des expériences de formation sur le même cluster de calcul ne peuvent pas partager ce dataset ; chacun effectue le téléchargement (coûteux) du dataset à partir d'un emplacement arbitraire.
- Si d'autres datasets ou versions du même dataset sont nécessaires pour les tâches d'entraînement ultérieures, les data Scientists doivent de nouveau effectuer le téléchargement (coûteux) du dataset sur

l'instance de calcul exécutant les fichiers training.NetApp et cnvrg.io ont créé une nouvelle solution de mise en cache des datasets qui élimine ces obstacles. La solution accélère l'exécution du pipeline DE ML en mettant en cache les datasets fortement sollicités sur le système de stockage hautes performances ONTAP. Avec ONTAP NFS, les datasets sont mis en cache une seule fois dans une structure de données optimisée par NetApp (comme AFF A800), qui est située dans un environnement de calcul. Le stockage ultra-rapide NetApp ONTAP NFS peut servir plusieurs nœuds de calcul DE ML. Résultat : les performances des modèles de formation sont optimisées, ce qui permet à l'entreprise de réaliser des économies, d'améliorer la productivité et d'optimiser l'efficacité opérationnelle.

Architecture de la solution

Cette solution de NetApp et cnvrg.io fournit une mise en cache des datasets, comme l'illustre la figure suivante. La mise en cache des datasets permet aux data Scientists de choisir la version souhaitée du jeu de données ou du dataset et de la déplacer vers le cache NFS ONTAP, qui se trouve à proximité du cluster de calcul DU ML. Le data Scientist peut désormais réaliser plusieurs expériences sans devoir engendrer de retards ou de téléchargements supplémentaires. En outre, tous les ingénieurs collaborant peuvent utiliser le même dataset avec le cluster de calcul associé (avec la liberté de choisir un nœud) sans téléchargements supplémentaires depuis le data Lake. Ils ont également accès à un tableau de bord qui assure le suivi et le contrôle de tous les datasets et versions, et offre une vue des datasets mis en cache.

La plateforme cnvrg.io détecte automatiquement les datasets obsolètes qui n'ont pas été utilisés depuis un certain temps et les supprime du cache, ce qui conserve l'espace de cache NFS libre pour les datasets plus fréquemment utilisés. Il est important de noter que la mise en cache du dataset avec ONTAP fonctionne dans le cloud et sur site, pour une flexibilité maximale.



Concepts et composants

Cette section aborde les concepts et les composants associés à la mise en cache de

données dans un flux DE TRAVAIL DE ML.

Apprentissage machine

L'APPRENTISSAGE MACHINE est en passe de devenir essentiel pour de nombreuses entreprises et organisations à travers le monde. Par conséquent, LES équipes IT et DevOps doivent aujourd'hui relever les défis liés à la standardisation des workloads DE ML et au provisionnement des ressources cloud, sur site et de calcul hybride qui prennent en charge les workflows dynamiques et intensifs requis par les pipelines de ML et les pipelines.

Apprentissage machine basé sur un conteneur et Kubernetes

Les conteneurs sont des instances isolées de l'espace utilisateur qui s'exécutent sur un noyau de système d'exploitation hôte partagé. Les conteneurs se sont de plus en plus rapidement adoption. Les conteneurs offrent bon nombre des avantages de la boxe applicative offerts par les machines virtuelles. Cependant, les couches de l'hyperviseur et du système d'exploitation invité sur lesquelles reposent les machines virtuelles ont été éliminées, les conteneurs sont beaucoup plus légers.

Les conteneurs permettent également de packaging efficace des dépendances entre applications, des durées d'exécution, etc. Directement avec une application. Le format de conditionnement de conteneurs le plus utilisé est le container Docker. Une application conteneurisée dans le format de conteneur Docker peut être exécutée sur n'importe quel ordinateur capable d'exécuter des conteneurs Docker. Cela est vrai même si les dépendances de l'application ne sont pas présentes sur la machine, car toutes les dépendances sont emballées dans le conteneur lui-même. Pour plus d'informations, consultez la "Site Web de Docker".

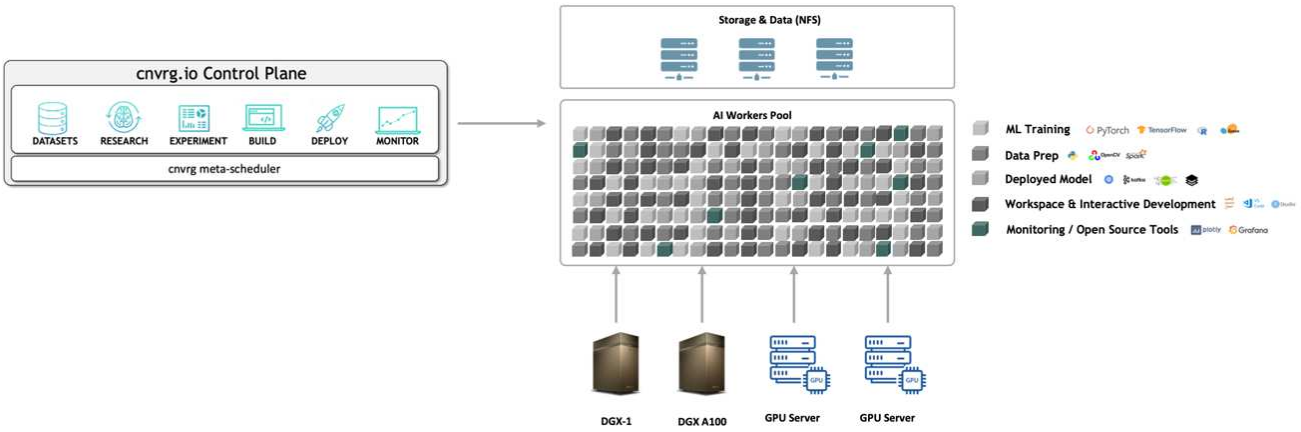
Kubernetes, l'orchestrateur de conteneur le plus populaire, permet aux data Scientists de lancer des travaux et des pipelines flexibles basés sur des conteneurs. Elle permet également aux équipes chargées de l'infrastructure de gérer et de contrôler les workloads DE ML dans un environnement cloud et géré unique. Pour plus d'informations, consultez la "Site Web de Kubernetes".

cnvrg.io

Cnvrg.io est un système d'exploitation d'IA qui transforme la façon dont les entreprises gèrent, évoluent et accélèrent le développement de l'IA et de la data science dans la recherche jusqu'à la production. Conçue par les data Scientists, la plateforme basée sur le code First offre aux data Scientists une grande flexibilité pour une exécution sur site ou dans le cloud. Avec la gestion des modèles, la division MLOps et les solutions continues DE ML, cnvrg.io propose aux équipes de science des données des technologies de pointe afin qu'elles puissent consacrer moins de temps au DevOps et se concentrer sur les algorithmes réels. Depuis l'utilisation de cnvrg.io, les équipes de différents secteurs ont obtenu davantage de modèles en production, ce qui a eu pour conséquence une plus grande valeur commerciale.

Méta-planificateur cnvrg.io

cnvrg. l'e/s dispose d'une architecture unique qui permet AUX INGÉNIEURS ET AUX DÉPARTEMENTS INFORMATIQUES d'associer différentes ressources de calcul au même plan de contrôle et de permettre à cnvrg.io de gérer les tâches DE ML dans l'ensemble des ressources. Cela signifie qu'ELLE peut connecter plusieurs clusters Kubernetes sur site, des serveurs de machine virtuelle et des comptes cloud, et exécuter des workloads DE ML sur toutes les ressources, comme l'illustre la figure suivante.



Cache des données cnvrg.io

grâce à sa technologie de mise en cache des données, cnvrg.io permet aux data scientists de définir des versions de datasets actives et inactives. Par défaut, les jeux de données sont stockés dans une base de données de stockage objet centralisée. Les data Scientists peuvent ensuite mettre en cache une version de données spécifique sur la ressource de calcul sélectionnée pour gagner du temps lors du téléchargement et ainsi augmenter LE DÉVELOPPEMENT DU ML et la productivité. Les datasets qui sont mis en cache et ne sont pas utilisés pendant quelques jours sont automatiquement supprimés du protocole NFS sélectionné. La mise en cache et la suppression du cache peuvent être réalisées en un clic. Aucun codage ni AUCUNE tâche n'est requise POUR LE DevOps.

Flux cnvrg.io et pipelines DE ML

Cnvrg.io écoulements est un outil de construction de pipelines DE ML de production. Chaque composant d'un flux est un script/code exécuté sur un calcul sélectionné avec une image docker de base. Cette conception permet aux data Scientists et aux ingénieurs de créer un pipeline unique capable d'exécuter à la fois sur site et dans le cloud. cnvrg.io garantit le déplacement des données, des paramètres et des artefacts entre les différents composants. En outre, chaque flux est surveillé et suivi pour une science des données 100 % reproductible.

CŒUR cnvrg.io

Le CŒUR cnvrg.io est une plateforme gratuite pour la communauté des sciences des données afin d'aider les data Scientists à se concentrer davantage sur l'informatique et moins sur le DevOps. L'infrastructure flexible DU CŒUR permet aux data Scientists d'utiliser n'importe quel langage, framework d'IA ou environnement de calcul, qu'ils se trouvent sur site ou dans le cloud pour qu'ils puissent faire ce qu'ils font de mieux en créant des algorithmes. Le CŒUR cnvrg.io peut être facilement installé à l'aide d'une seule commande sur tout cluster Kubernetes.

NetApp ONTAP ai

ONTAP ai est une architecture de référence de data Center pour les workloads D'AM et d'AP qui utilisent les systèmes de stockage NetApp AFF et les systèmes NVIDIA DGX avec des GPU Tesla V100. ONTAP ai repose sur le protocole de fichiers NFS standard plus de 100 Gb Ethernet, offrant ainsi aux clients une infrastructure DE ML/DL haute performance qui utilise des technologies de data Center standard afin de réduire les frais d'implémentation et d'administration. L'utilisation d'un réseau et de protocoles standardisés permet l'intégration d'ONTAP ai aux environnements de cloud hybride tout en maintenant la cohérence et la simplicité opérationnelles. En tant que solution d'infrastructure prévalidée, ONTAP ai réduit les délais de déploiement et les risques, et réduit considérablement la surcharge administrative, ce qui permet aux clients

d'accélérer le retour sur investissement.

NVIDIA DeepOps

DeepOps est un projet open source de NVIDIA qui, grâce à Ansible, automatise le déploiement de clusters de serveurs GPU conformément aux bonnes pratiques. DeepOps est modulaire et peut être utilisé pour diverses tâches de déploiement. Pour ce document et l'exercice de validation décrit, DeepOps est utilisé pour déployer un cluster Kubernetes composé de nœuds workers de serveurs GPU. Pour plus d'informations, consultez la "Site Web de DeepOps".

NetApp Trident

Trident est un orchestrateur de stockage open source développé et géré par NetApp qui simplifie considérablement la création, la gestion et la consommation du stockage persistant pour les workloads Kubernetes. Trident est une application native de Kubernetes qui s'exécute directement dans un cluster Kubernetes. Avec Trident, les utilisateurs de Kubernetes (développeurs, data Scientists, administrateurs Kubernetes, etc.) peuvent créer, gérer et interagir avec les volumes de stockage persistant dans le format Kubernetes standard qu'ils connaissent déjà. Ils peuvent également bénéficier des fonctionnalités avancées de gestion des données de NetApp et d'un environnement Data Fabric optimisé par la technologie NetApp. Trident élimine les complexités du stockage persistant et facilite la consommation. Pour plus d'informations, consultez la "Site Web Trident".

NetApp StorageGRID

NetApp StorageGRID est une plateforme de stockage objet Software-defined conçue pour répondre à ces besoins avec un stockage simple et similaire à celui du cloud, accessible via le protocole S3. StorageGRID est un système scale-out conçu pour prendre en charge plusieurs nœuds sur plusieurs sites connectés par Internet, quelle que soit la distance. Grâce au moteur de règles intelligent de StorageGRID, les utilisateurs ont la possibilité de choisir des objets de codage d'effacement sur l'ensemble des sites pour assurer la résilience géographique ou la réplication d'objets entre les sites distants, afin de minimiser la latence d'accès au réseau WAN. Avec cette solution, StorageGRID propose un excellent data Lake de stockage objet primaire de cloud privé.

NetApp Cloud Volumes ONTAP

Le logiciel de gestion des données NetApp Cloud Volumes ONTAP permet de contrôler et de protéger les données utilisateur et d'optimiser l'efficacité du stockage, tout en bénéficiant de la flexibilité des fournisseurs de cloud public tels que AWS, Google Cloud Platform et Microsoft Azure. Cloud Volumes ONTAP est un logiciel de gestion des données natif cloud basé sur le logiciel de stockage NetApp ONTAP. Il offre aux utilisateurs une plateforme de stockage universelle haute performance qui répond à leurs besoins cloud. En utilisant le même logiciel de gestion du stockage dans le cloud et sur site, les utilisateurs bénéficient de l'avantage d'un environnement Data Fabric sans devoir former les équipes IT à des méthodes entièrement nouvelles de gestion des données.

Si le client s'intéresse aux modèles de déploiement de cloud hybride, Cloud Volumes ONTAP offre des fonctionnalités et des performances exceptionnelles dans la plupart des clouds publics, afin d'offrir à ses utilisateurs une expérience cohérente et transparente, quel que soit l'environnement.

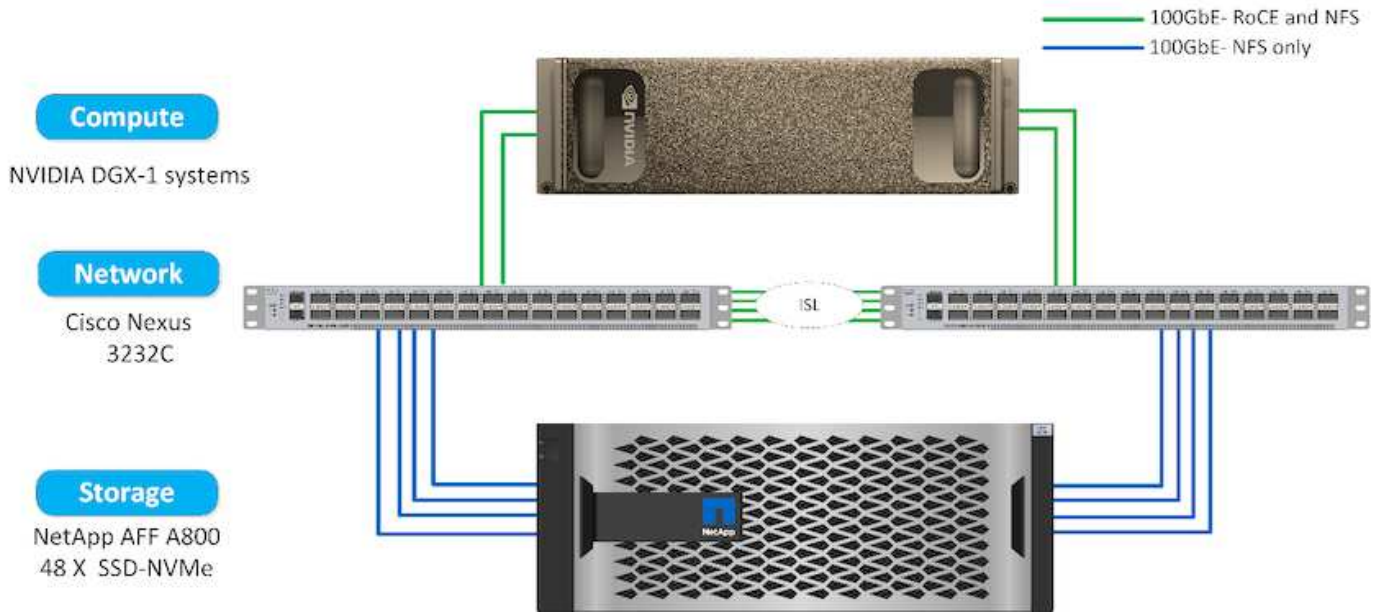
Configuration matérielle et logicielle requise

Cette section présente les exigences technologiques de la solution ONTAP d'IA.

Configuration matérielle requise

Bien que les exigences matérielles dépendent de workloads spécifiques, ONTAP ai peut être déployé à n'importe quelle échelle pour l'ingénierie des données, l'entraînement des modèles et l'inférence de production à partir d'un seul GPU jusqu'à des configurations en rack pour les opérations D'AM/AP à grande échelle. Pour en savoir plus sur ONTAP ai, rendez-vous sur le "Site Web ONTAP ai".

Cette solution a été validée à l'aide d'un système DGX-1 pour les ressources de calcul, d'un système de stockage NetApp AFF A800 et de Cisco Nexus 3232C pour la connectivité réseau. L'outil AFF A800 utilisé dans cette validation peut prendre en charge jusqu'à 10 systèmes DGX-1 pour la plupart des workloads D'APPRENTISSAGE PROFOND/D'APPRENTISSAGE PROFOND. La figure suivante présente la topologie ONTAP ai utilisée pour l'entraînement des modèles dans cette validation.



Pour étendre cette solution à un cloud public, Cloud Volumes ONTAP peut être déployé avec des ressources de calcul GPU cloud et intégré dans un environnement Data Fabric de cloud hybride qui permet aux clients d'utiliser les ressources appropriées pour chaque charge de travail.

Configuration logicielle requise

Le tableau suivant indique les versions spécifiques utilisées pour la validation de cette solution.

Composant	Version
Ubuntu	18.04.4 LTS
SYSTÈME D'EXPLOITATION NVIDIA DGX	4.4.0
NVIDIA DeepOps	20.02.1
Kubernetes	1.15
Gouvernail	3.1.0
cnvrg.io	3.0.0
NetApp ONTAP	9.6P4

Pour la validation de cette solution, Kubernetes a été déployé en tant que cluster à un seul nœud sur le

système DGX-1. Pour les déploiements à grande échelle, des nœuds maîtres Kubernetes indépendants doivent être déployés pour assurer la haute disponibilité des services de gestion et réserver des ressources DGX précieuses pour les workloads DE ML et de DL.

Détails sur le déploiement et la validation de la solution

Les sections suivantes abordent les détails du déploiement et de la validation de la solution.

Avec le déploiement d'ONTAP ai

Le déploiement de ONTAP ai nécessite l'installation et la configuration du matériel de réseau, de calcul et de stockage. Ce rapport n'a pas pour objectif d'obtenir des instructions spécifiques pour le déploiement de l'infrastructure d'IA d'ONTAP, Pour des informations détaillées sur le déploiement, reportez-vous à la section "NVA-1121-DEPLOY : NetApp ONTAP ai, optimisé par NVIDIA".

Pour la validation de cette solution, un volume unique a été créé et monté sur le système DGX-1. Ce point de montage a ensuite été monté sur les conteneurs pour faciliter l'accès aux données pour l'entraînement. Pour les déploiements à grande échelle, NetApp Trident automatise la création et le montage des volumes de manière à éliminer les surcharges administratives et à permettre la gestion des ressources des utilisateurs finaux.

Déploiement Kubernetes

Pour déployer et configurer votre cluster Kubernetes avec NVIDIA DeepOps, effectuez les tâches suivantes à partir d'un hôte de démarrage du déploiement :

1. Téléchargez NVIDIA DeepOps en suivant les instructions sur le "Page de démarrage" Sur le site GitHub NVIDIA DeepOps.
2. Déployez Kubernetes dans votre cluster en suivant les instructions du "Guide de déploiement Kubernetes" Sur le site GitHub NVIDIA DeepOps.

> Pour que le déploiement Kubernetes de DeepOps fonctionne, le même utilisateur doit exister sur tous les nœuds maîtres et workers Kubernetes.

Si le déploiement échoue, modifiez la valeur de `kubectl_localhost` à faux dans `deepops/config/group_vars/k8s-cluster.yml` et répétez l'étape 2. Le `Copy kubectl binary to ansible host` tâche, qui s'exécute uniquement lorsque la valeur de `kubectl_localhost` Est vrai, dépend du module récupérer Ansible, qui présente des problèmes connus d'utilisation de la mémoire. Ces problèmes d'utilisation de la mémoire peuvent parfois entraîner l'échec de la tâche. Si la tâche échoue en raison d'un problème de mémoire, le reste de l'opération de déploiement ne s'effectue pas correctement.

Si le déploiement se termine correctement une fois que vous avez modifié la valeur de `kubectl_localhost` à `false`, vous devez ensuite copier manuellement `kubectl binary` D'un nœud maître Kubernetes vers l'hôte de démarrage du déploiement Vous pouvez trouver l'emplacement du `kubectl binary` sur un nœud maître spécifique en exécutant le `which kubectl` commandez directement sur ce nœud.

Déploiement cnvrg.io

Déploiement DU CŒUR cnvrg à l'aide de Helm

Helm est le moyen le plus simple de déployer rapidement cnvrg dans n'importe quel cluster, sur site, Minikube ou dans n'importe quel cluster cloud (AKS, EKS et GKE, par exemple). Cette section décrit comment cnvrg a été installé sur une instance locale (DGX-1) avec Kubernetes installé.

Prérequis

Avant de terminer l'installation, vous devez installer et préparer les dépendances suivantes sur votre ordinateur local :

- Kubectl
- Helm 3.x
- Cluster Kubernetes 1.15+

Déploiement à l'aide de Helm

1. Pour télécharger les graphiques cnvrg Helm les plus mis à jour, exécutez la commande suivante :

```
helm repo add cnvrg https://helm.cnvrg.io
helm repo update
```

2. Avant de déployer cnvrg, vous devez disposer de l'adresse IP externe du cluster et du nom du nœud sur lequel vous allez déployer cnvrg. Pour déployer cnvrg sur un cluster Kubernetes sur site, exécutez la commande suivante :

```
helm install cnvrg cnvrg/cnvrg --timeout 1500s  --wait \ --set
global.external_ip=<ip_of_cluster> \ --set global.node=<name_of_node>
```

3. Exécutez le `helm install` commande. L'ensemble des services et systèmes s'installent automatiquement sur le cluster. Ce processus peut prendre jusqu'à 15 minutes.
4. Le `helm install` la commande peut prendre jusqu'à 10 minutes. Une fois le déploiement terminé, accédez à l'URL de votre nouveau cluster cnvrg déployé ou ajoutez-le en tant que ressource au sein de votre organisation. Le `helm` Commande vous informe de l'URL correcte.

```
Thank you for installing cnvrg.io!
Your installation of cnvrg.io is now available, and can be reached via:
Talk to our team via email at
```

5. Lorsque l'état de tous les conteneurs est en cours d'exécution ou terminé, cnvrg a été déployé avec succès. Similaire à la sortie d'exemple suivante :

```
NAME                                READY   STATUS      RESTARTS   AGE
cnvrg-app-69fbb9df98-6xrgf                  1/1     Running     0          2m
cnvrg-sidekiq-b9d54d889-5x4fc               1/1     Running     0          2m
controller-65895b47d4-s96v6                 1/1     Running     0          2m
init-app-vs-config-wv9c4                    0/1     Completed   0          9m
init-gateway-vs-config-2zbpp                0/1     Completed   0          9m
init-minio-vs-config-cd2rg                  0/1     Completed   0          9m
minio-0                                     1/1     Running     0          2m
postgres-0                                  1/1     Running     0          2m
redis-695c49c986-kcbt9                      1/1     Running     0          2m
seeder-wh655                                0/1     Completed   0          2m
speaker-5sghr                               1/1     Running     0          2m
```

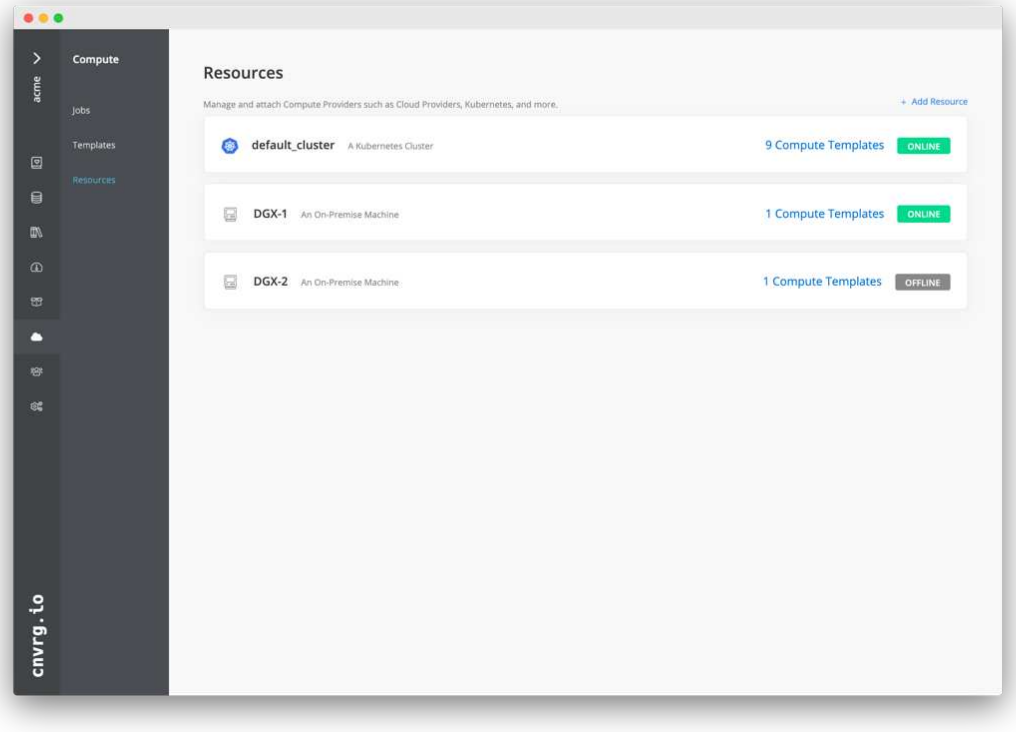
Computer Vision Model Training with ResNet50 and the thorax X-ray Dataset

Le système d'exploitation cesnvrg.io ai a été déployé sur une configuration Kubernetes sur une architecture NetApp ONTAP ai optimisée par le système NVIDIA DGX. Pour la validation, nous avons utilisé le jeu de données de radiographie du thorax NIH composé d'images déidentifiées des radiographies pulmonaires. Les images étaient au format PNG. Les données ont été fournies par le Centre clinique des NIH et sont disponibles par l'intermédiaire du "Site de téléchargement NIH". Nous avons utilisé un échantillon de 250 Go de données comportant 627, 615 images dans 15 classes.

Le dataset a été téléchargé sur la plateforme cnvrg et a été mis en cache sur une exportation NFS depuis le système de stockage NetApp AFF A800.

Configurez les ressources de calcul

L'architecture cnvrg et la fonctionnalité de méta-planification permettent aux ingénieurs et aux professionnels DE L'IT d'associer différentes ressources de calcul à une seule plateforme. Dans notre configuration, nous avons utilisé le même cluster cnvrg qui a été déployé pour exécuter les workloads d'apprentissage profond. Si vous devez ajouter des clusters, utilisez l'interface graphique, comme illustré ci-dessous.



Charger les données

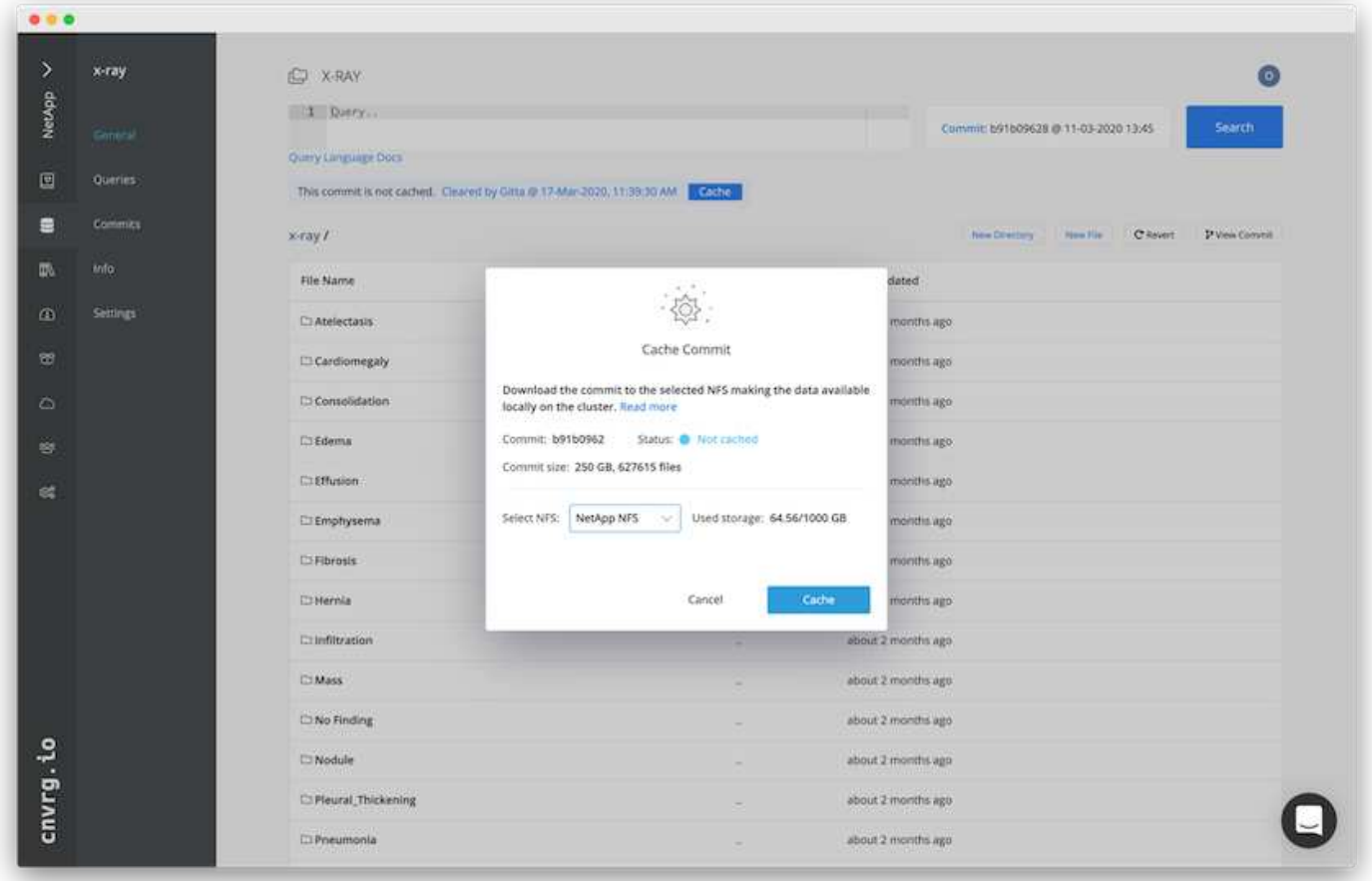
Pour charger des données sur la plateforme cnvrg, vous pouvez utiliser l'interface graphique ou l'interface de ligne de commande cnvrg. Pour les jeux de données volumineux, NetApp recommande l'utilisation de l'interface de ligne de commande, car il s'agit d'un outil robuste, évolutif et fiable capable de gérer un grand nombre de fichiers.

Pour télécharger des données, procédez comme suit :

1. Téléchargez le "CLI cnvrg".
2. accédez au répertoire des rayons x.
3. Initialisez le dataset dans la plate-forme à l'aide du `cnvrg data init` commande.
4. Télécharger tout le contenu du répertoire vers le data Lake central avec le `cnvrg data sync` Commande.une fois les données chargées dans le magasin d'objets central (StorageGRID, S3 ou autres), vous pouvez naviguer à l'aide de l'interface graphique. La figure suivante montre un fichier PNG d'image de fibrose radiologique thoracique chargé. De plus, cnvrg version les données afin que tout modèle que vous construisez puisse être reproduit à la version des données.

Données de compte-tours

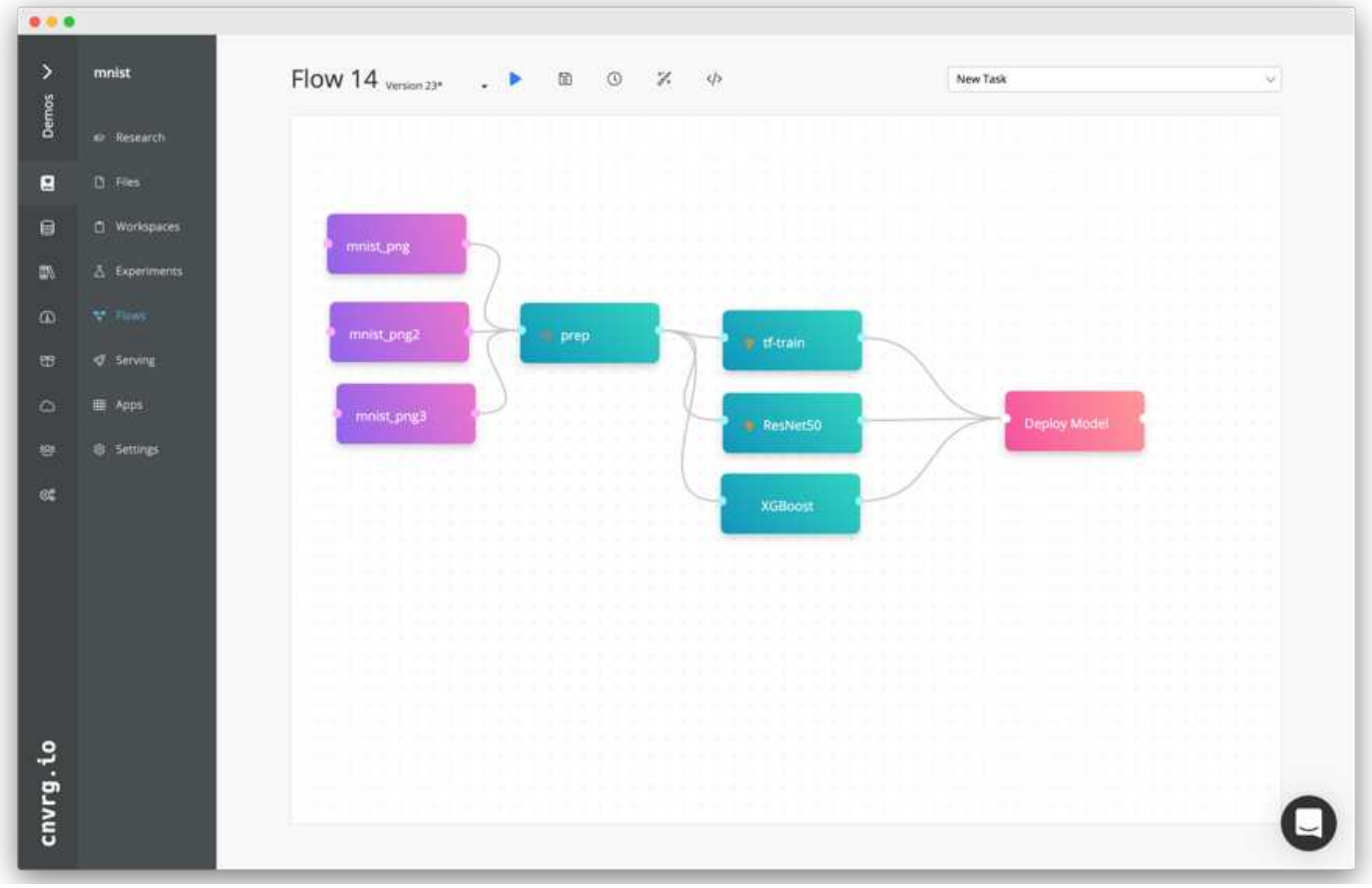
Pour accélérer l'entraînement et éviter de télécharger plus de 600 000 fichiers pour chaque essai et entraînement du modèle, nous avons utilisé la fonctionnalité de mise en cache des données après la première mise en ligne des données dans le magasin d'objets central du data Lake.



Lorsque les utilisateurs cliquent sur cache, cnvrg télécharge les données de leur validation spécifique depuis le magasin d'objets distant et les met en cache sur le volume NFS ONTAP. Les données sont ensuite disponibles pour un entraînement instantané. De plus, si les données ne sont pas utilisées pendant quelques jours (pour l'entraînement ou l'exploration des modèles, par exemple), cnvrg efface automatiquement le cache.

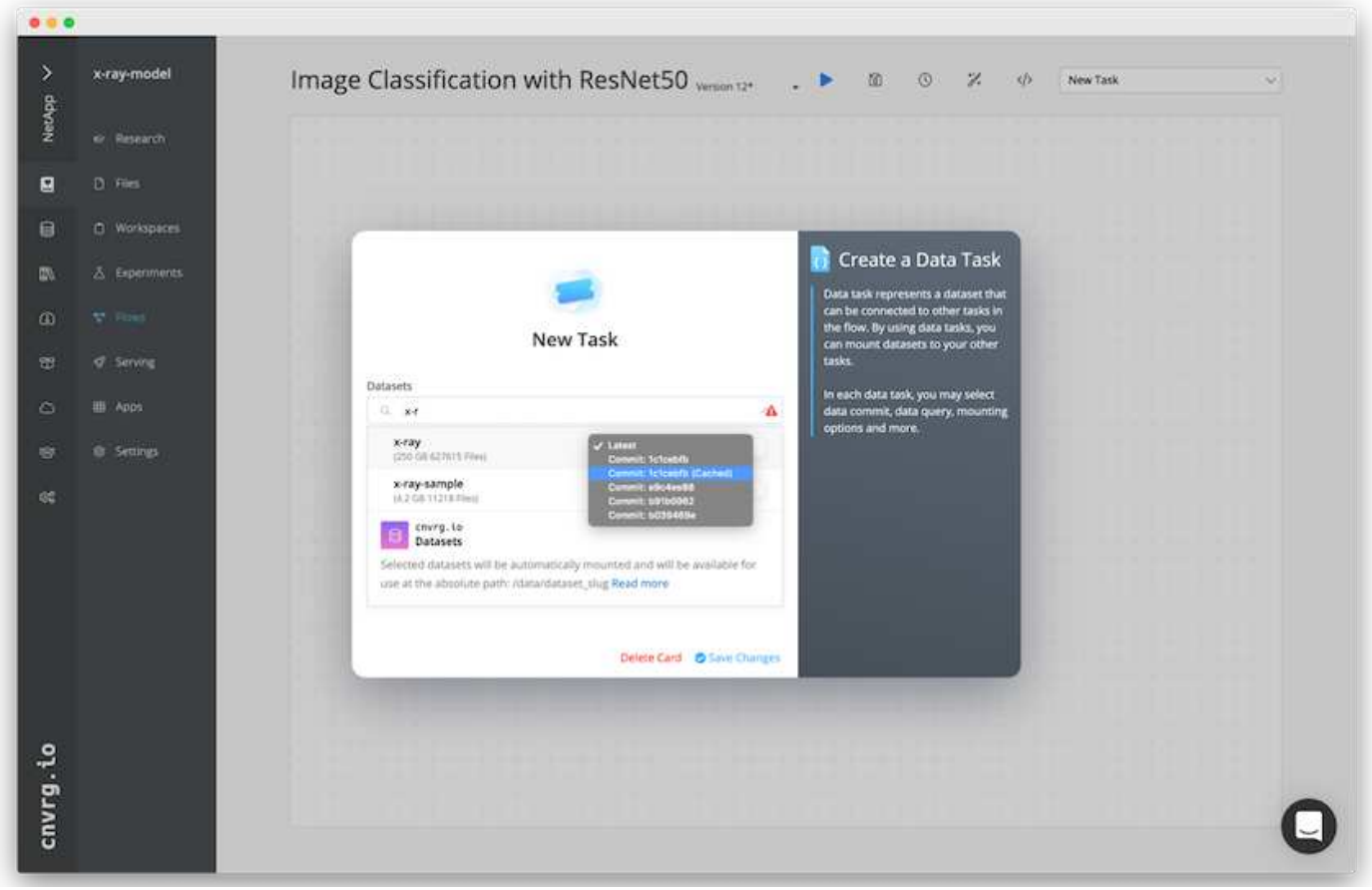
Créez un pipeline DE ML avec les données en cache

Les flux cnvrg vous permettent de construire facilement des pipelines DE ML de production. Les flux sont flexibles, peuvent fonctionner pour tous types d'utilisation DU ML et peuvent être créés via l'interface utilisateur graphique ou le code. Chaque composant d'un flux peut s'exécuter sur une ressource de calcul différente avec une image Docker différente, ce qui permet de créer un cloud hybride et des pipelines DE ML optimisés.



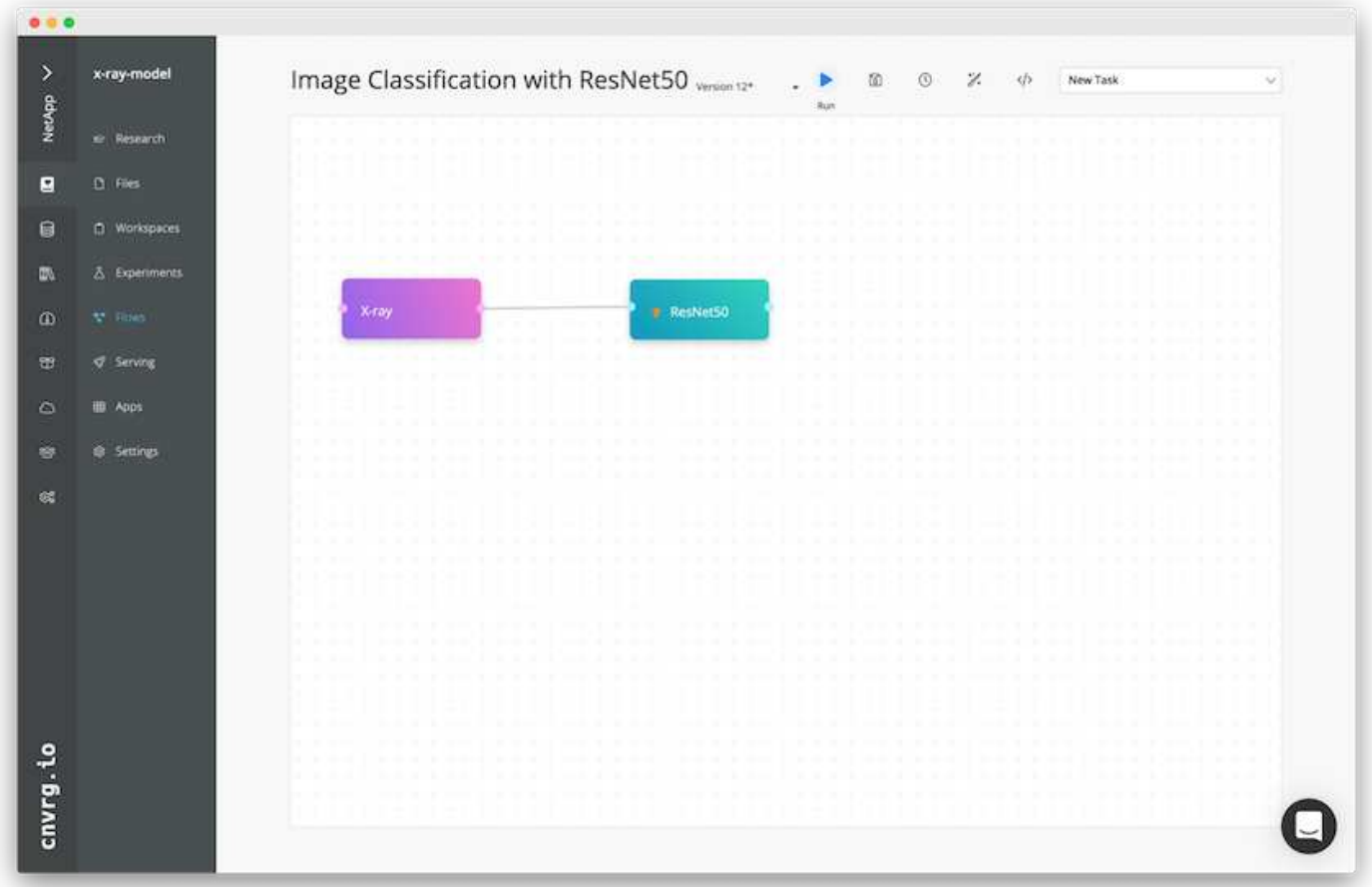
Création du flux de rayons X thorax : réglage des données

Nous avons ajouté notre dataset à un nouveau flux créé. Lors de l'ajout du dataset, vous pouvez sélectionner la version spécifique (commit) et indiquer si vous voulez la version en cache. Dans cet exemple, nous avons sélectionné la validation en cache.



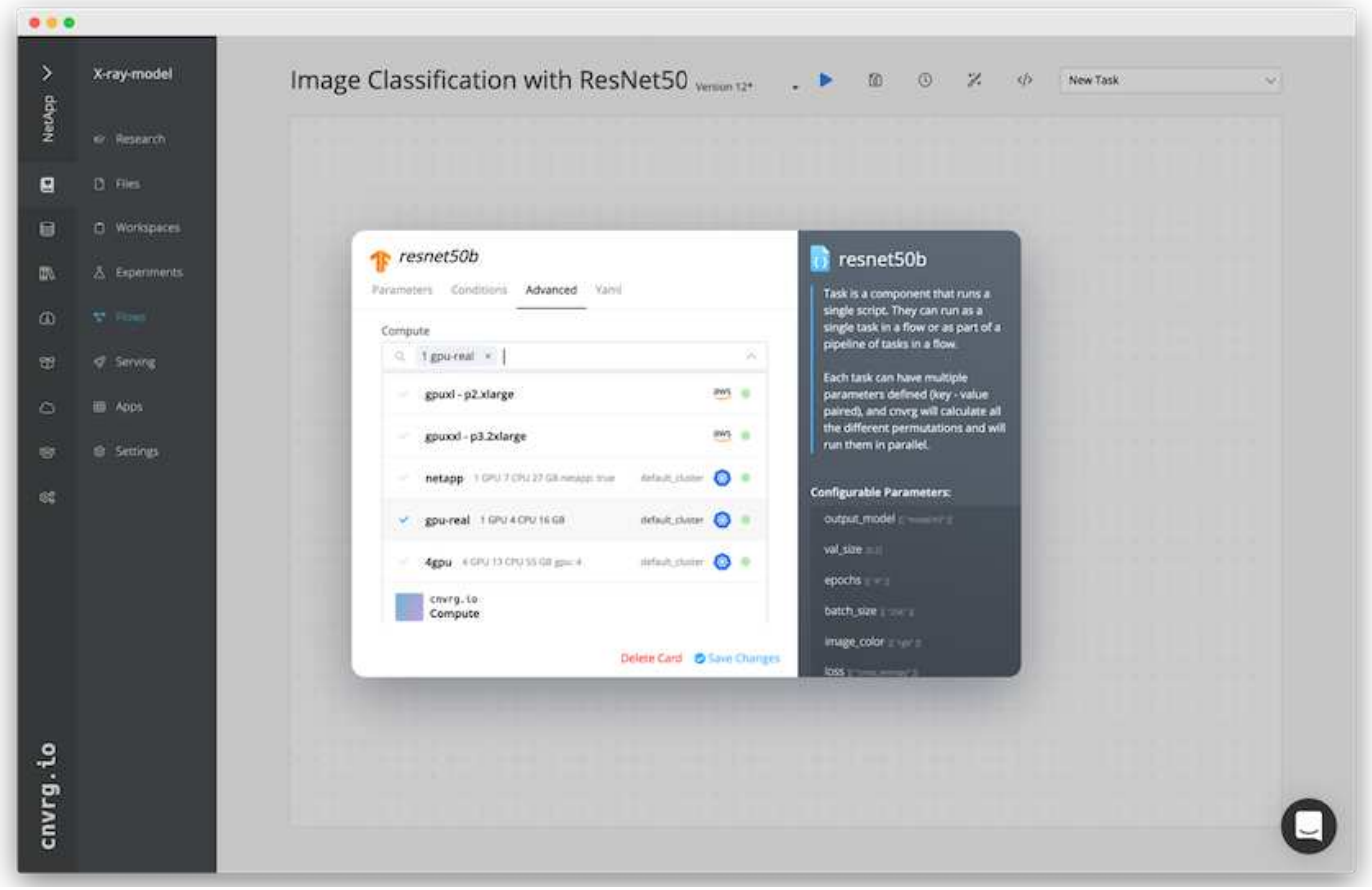
Construction du flux de rayons X thorax : réglage du modèle d'entraînement : ResNet50

Dans le pipeline, vous pouvez ajouter n'importe quel type de code personnalisé. Dans cnvrg, il y a aussi la bibliothèque ai, une collection de composants DE ML réutilisables. La bibliothèque d'IA comprend des algorithmes, des scripts, des sources de données et d'autres solutions qui peuvent être utilisés dans n'importe quel flux DE ML ou de deep learning. Dans cet exemple, nous avons sélectionné le module ResNet50 prédéfini. Nous avons utilisé des paramètres par défaut tels que batch_size:128, des séries de tests:10, etc. Ces paramètres peuvent être affichés dans la documentation de la bibliothèque d'IA. La capture d'écran suivante montre le nouveau flux avec le jeu de données de rayons X connecté à ResNet50.



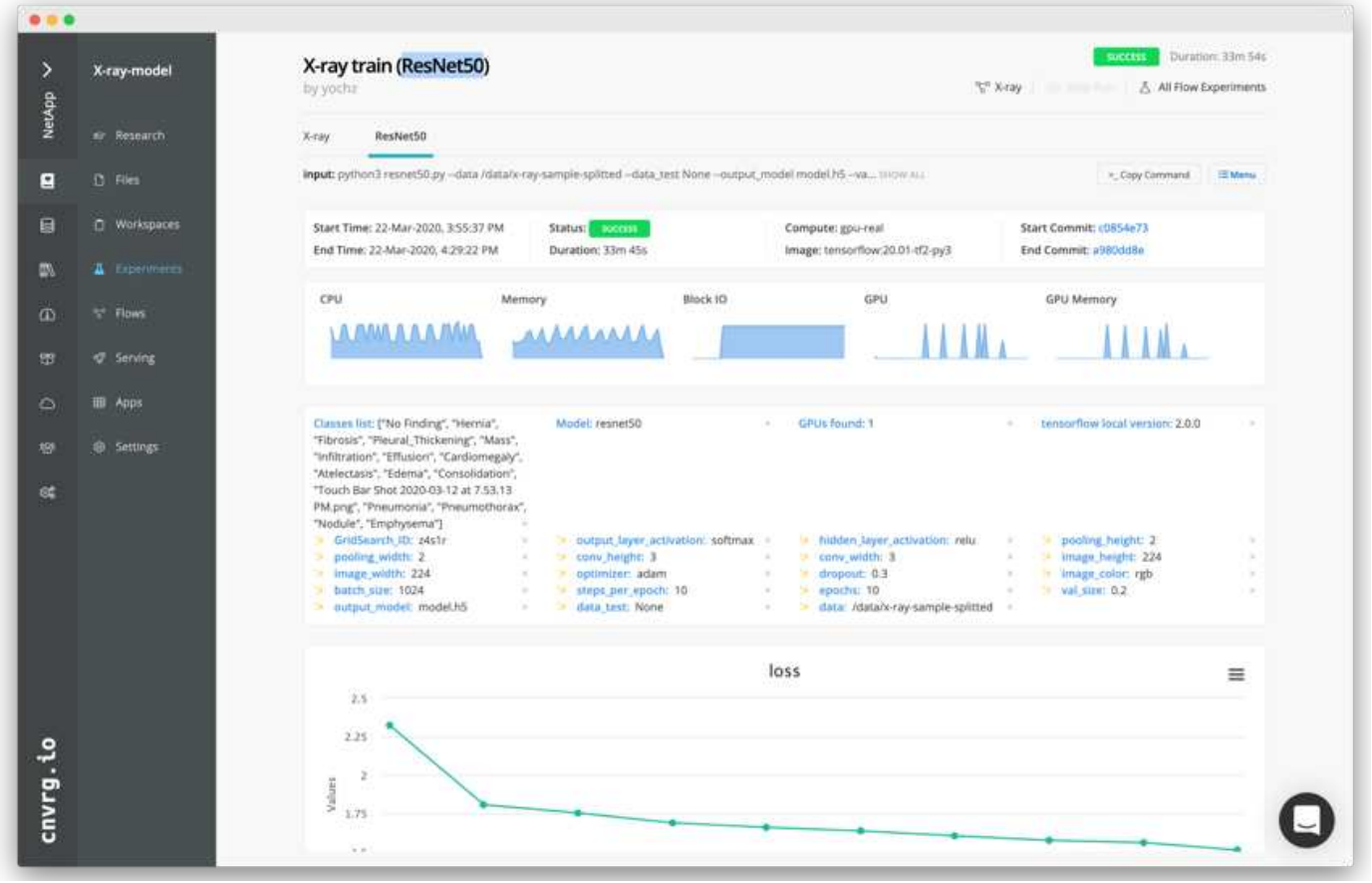
Définissez la ressource de calcul pour ResNet50

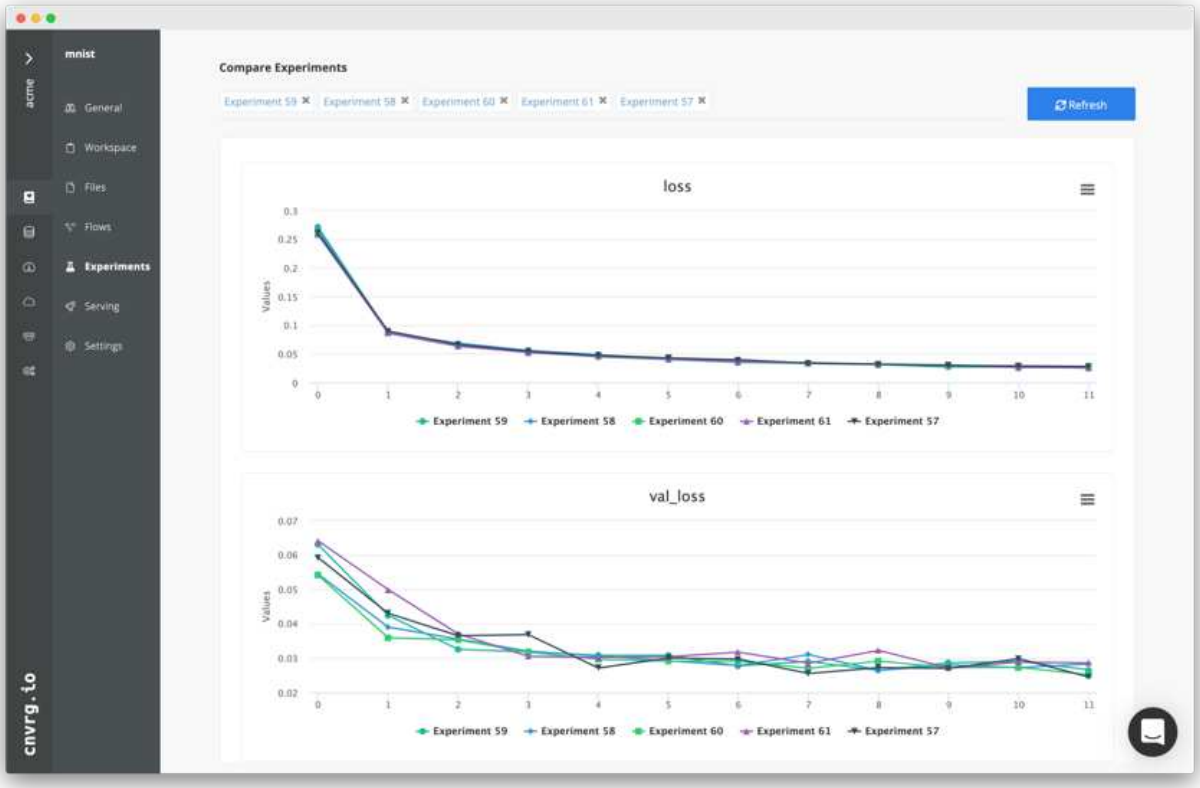
Chaque algorithme ou composant des flux cnvrg peut être exécuté sur une instance de calcul différente, avec une image Docker différente. Dans notre configuration, nous voulions exécuter l'algorithme d'entraînement sur les systèmes NVIDIA DGX avec l'architecture NetApp ONTAP ai. Dans la figure suivante, nous avons sélectionné `gpu-real`, qui est un modèle de calcul et une spécification pour notre cluster sur site. Nous avons également créé une file d'attente de modèles et sélectionné plusieurs modèles. De cette façon, si le `gpu-real` les ressources ne peuvent pas être allouées (si, par exemple, d'autres data scientists l'utilisent), vous pouvez activer la bursting en ajoutant un modèle de fournisseur de cloud. La capture d'écran suivante montre l'utilisation de gpu-Real comme nœud de calcul pour ResNet50.



Suivi et surveillance des résultats

Après l'exécution d'un flux, cnvrg déclenche le moteur de suivi et de surveillance. Chaque cycle d'un flux est automatiquement documenté et mis à jour en temps réel. Hyperparamètres, mesures, utilisation des ressources (utilisation des GPU, etc.), version de code, artéfacts, journaux Et ainsi de suite sont automatiquement disponibles dans la section expériences, comme indiqué dans les deux captures d'écran suivantes.


X-ray-model
X-ray train (ResNet50)
by yochz
SUCCESS
Duration: 33m 54s
X-ray
All Flow Experiments
X-ray
ResNet50
Research
Files
Workspaces
Experiments
Flows
Serving
Apps
Settings
Start Time: 22-Mar-2020, 3:55:37 PM
End Time: 22-Mar-2020, 4:29:22 PM
Status: SUCCESS
Duration: 33m 45s
Compute: gpu-real
Image: tensorflow:20.01-tf2-py3
Start Commit: c0854e73
End Commit: a980dd8e
CPU
Memory
Block IO
GPU
GPU Memory
Model: resnet50
GPUs found: 1
tensorflow local version: 2.0.0
loss
Values
cnvrg.io


mnist
Compare Experiments
Refresh
General
Workspace
Files
Flows
Experiments
Serving
Settings
loss
val_loss
Values
Experiment 59
Experiment 58
Experiment 60
Experiment 61
Experiment 57
cnvrg.io

Conclusion

NetApp et cnvrg.io se sont associés pour offrir aux clients une solution complète de gestion des données pour le développement de logiciels DE ML et de DL. ONTAP ai fournit des ressources de calcul et de stockage haute performance à n'importe quelle échelle d'opérations. Le logiciel nvrg.io rationalise les workflows en data science et améliore l'utilisation des ressources.

Remerciements

- Mike Oglesby, Ingénieur marketing et technique, NetApp
- Santosh Rao, Directeur technique senior, NetApp

Sources d'informations complémentaires

Pour en savoir plus sur les informations données dans ce document, consultez les ressources suivantes :

- Cnvrg.io ("https://cnvrg.io") :
  - CŒUR Cnvrg (plateforme DE ML gratuite)

    https://cnvrg.io/platform/core
  - Documents Cnvrg

    "https://app.cnvrg.io/docs"
- Serveurs NVIDIA DGX-1 :
  - Serveurs NVIDIA DGX-1

    https://www.nvidia.com/en-us/data-center/dgx-1/
  - GPU NVIDIA Tesla V100 à cœurs Tensor

    https://www.nvidia.com/en-us/data-center/tesla-v100/
  - NVIDIA GPU CLOUD (NGC)

    https://www.nvidia.com/en-us/gpu-cloud/
- Systèmes NetApp AFF :
  - Fiche technique AFF

    https://www.netapp.com/us/media/d-3582.pdf
  - NetApp Flash Advantage pour AFF

    https://www.netapp.com/us/media/ds-3733.pdf
  - Documentation ONTAP 9.x.

    http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- Rapport technique NetApp FlexGroup

  https://www.netapp.com/us/media/tr-4557.pdf

- Le stockage persistant NetApp pour les conteneurs :
  - NetApp Trident

    https://netapp.io/persistent-storage-provisioner-for-kubernetes/

- Matrice d'interopérabilité NetApp :
  - Matrice d'interopérabilité NetApp

    http://support.netapp.com/matrix

- Connectivité réseau d'ONTAP ai :
  - Switchs Cisco Nexus 3232C

    https://www.cisco.com/c/en/us/products/switches/nexus-3232c-switch/index.html

  - Switchs Mellanox Spectrum 2000

    http://www.mellanox.com/page/products_dyn?product_family=251&mtag=sn2000

- Structure ET outils ML :
  - DALI

    https://github.com/NVIDIA/DALI

  - TensorFlow : un framework d'apprentissage machine open source pour tous

    https://www.tensorflow.org/

  - Horovod : le cadre d'apprentissage profond Open source distribué d'Uber pour TensorFlow

    https://eng.uber.com/horovod/

  - Optimisation des GPU dans l'écosystème d'exécution du conteneur

    https://devblogs.nvidia.com/gpu-containers-runtime/

  - Docker

    https://docs.docker.com

  - Kubernetes

    https://kubernetes.io/docs/home/

  - NVIDIA DeepOps

    https://github.com/NVIDIA/deepops

  - Kubeflow

http://www.kubeflow.org/

- Jupyter Notebook Server

  http://www.jupyter.org/

- Dataset et bancs d'essai :
  - Ensemble de données de radiographie du thorax NIH

    https://nihcc.app.box.com/v/ChestXray-NIHCC

  - Xiaotong Wang, Yifan Peng, le lu, Zhiyong lu, Mohammadadi Bagheri, Ronald Summers, ChestX-ray8: Base de données à rayons X dans les hôpitaux et bancs d'essai sur la classification faiblement supervisée et la localisation des maladies du thorax commun, IEEE CVPR, pp 3462-3471, 2017 TR-4841-0620

Inférence d'IA en périphérie - NetApp avec Lenovo ThinkSystem - conception de la solution

Tr-4886 : Inférence d'IA en périphérie - NetApp avec Lenovo ThinkSystem - conception de la solution

Satish Thyagarajan, NetApp Miroslav Hodak, Lenovo

Récapitulatif

Plusieurs scénarios d'applications émergents, tels que les systèmes d'assistance à la conduite automobile (ADAS), le concept d'industrie 4.0, les Smart cities et l'Internet des objets (IoT), requièrent le traitement des flux de données en continu avec une latence quasi nulle. Ce document présente une architecture de calcul et de stockage permettant de déployer l'inférence d'intelligence artificielle (IA) basée sur les processeurs graphiques sur des contrôleurs de stockage NetApp et des serveurs Lenovo ThinkSystem dans un environnement en périphérie qui répond à ces exigences. Ce document fournit également des données de performance pour le banc d'essai MLPerf Inférence standard, en évaluant diverses tâches d'inférence sur des serveurs de périphérie équipés de processeurs graphiques NVIDIA T4. Nous examinons les performances des scénarios d'inférence hors ligne, à flux unique et à flux multiples. Nous montrons également que l'architecture associée à un système de stockage en réseau partagé économique est hautement performante et qu'elle constitue un point central pour la gestion des données et des modèles pour les serveurs de périphérie multiples.

Introduction

De plus en plus d'entreprises génèrent des volumes massifs de données à la périphérie du réseau. Pour tirer le meilleur parti des capteurs intelligents et des données de l'IoT, les entreprises recherchent une solution de streaming en temps réel qui met à profit l'informatique en périphérie. De ce fait, les tâches de calcul exigeantes sont de plus en plus effectuées à la périphérie, en dehors des data centers. L'inférence d'IA est un moteur de cette tendance. Les serveurs de périphérie offrent une puissance de calcul suffisante pour ces charges de travail, en particulier lorsqu'ils sont utilisés des accélérateurs, mais le stockage est souvent limité, en particulier dans les environnements multiserveurs. Dans ce document, nous vous montrons comment déployer un système de stockage partagé dans l'environnement de périphérie et comment il tire parti des charges de travail d'inférence d'IA sans imposer de baisse des performances.

Ce document décrit une architecture de référence pour l'inférence d'IA à la périphérie. Elle regroupe plusieurs serveurs Lenovo ThinkSystem Edge avec un système de stockage NetApp pour créer une solution facile à

déployer et à gérer. Il est destiné à servir de guide de référence pour les déploiements pratiques dans diverses situations, comme le sol en usine avec plusieurs caméras et capteurs industriels, les systèmes de point de vente dans les transactions au détail ou les systèmes de conduite autonome (FSD) qui identifient les anomalies visuelles des véhicules autonomes.

Ce document comprend le test et la validation d'une configuration de calcul et de stockage composée du serveur Lenovo ThinkSystem SE350 Edge Server et d'un système de stockage NetApp AFF et EF-Series d'entrée de gamme. Ces architectures de référence constituent une solution efficace et économique pour les déploiements d'IA. Elles fournissent également des services de données complets, une protection des données intégrée, une évolutivité transparente et un stockage de données connecté au cloud avec NetApp ONTAP et le logiciel de gestion des données NetApp SANtricity.

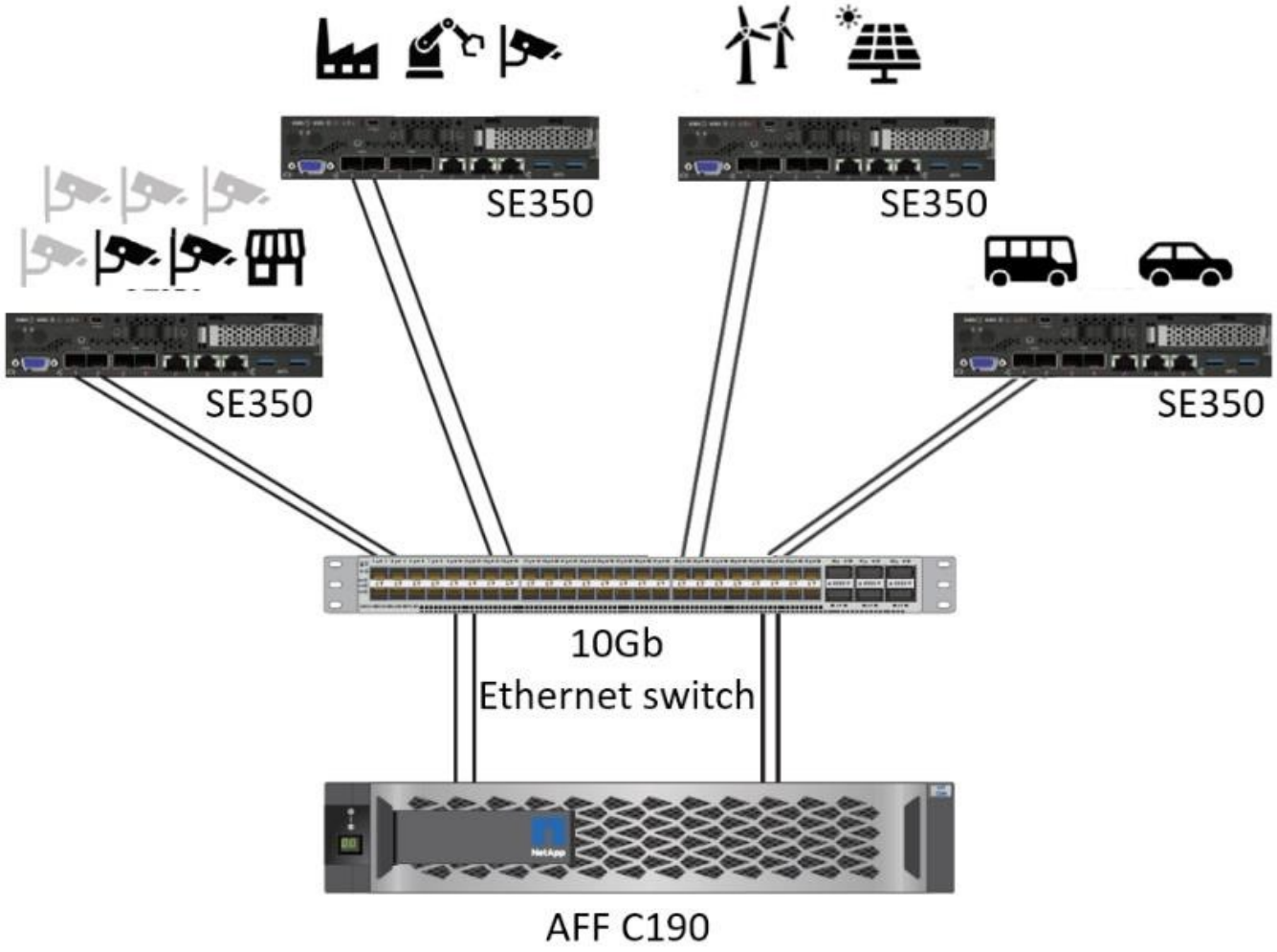
Public visé

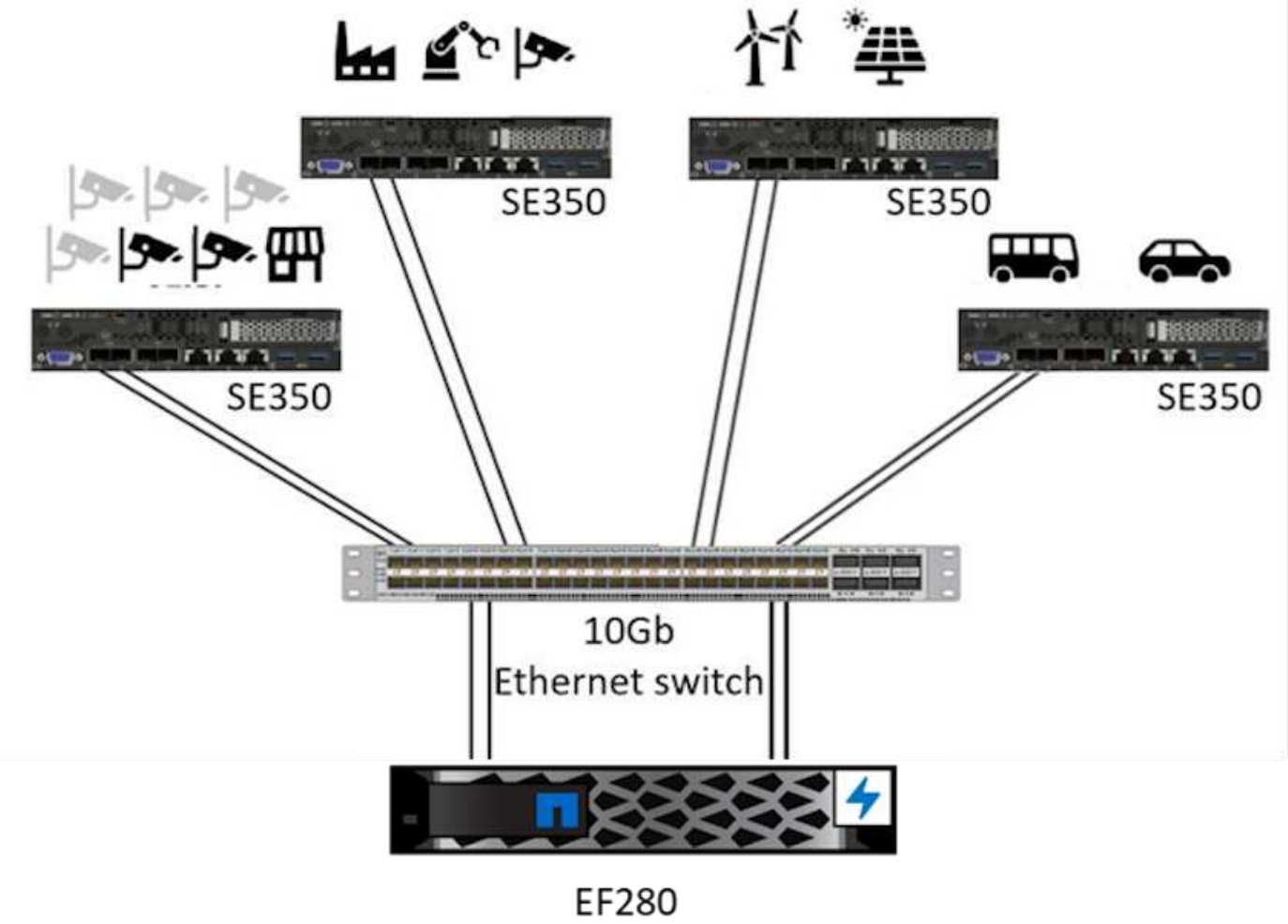
Ce document est destiné aux publics suivants :

- Les dirigeants et les architectes d'entreprise qui souhaitent proposer l'IA de manière adéquate à la périphérie.
- Les data Scientists, les ingénieurs de données, les chercheurs d'IA et de machine learning (ML) et les développeurs de systèmes d'IA.
- Les architectes d'entreprise qui conçoivent des solutions pour le développement de modèles et d'applications d'IA/ML.
- Les data Scientists et les ingénieurs d'IA qui recherchent des méthodes efficaces pour déployer les modèles de deep learning (DL) et DE MACHINE LEARNING (ML).
- Des gestionnaires de terminaux et des administrateurs de serveurs en périphérie sont chargés du déploiement et de la gestion des modèles d'inférence à la périphérie.

Architecture de la solution

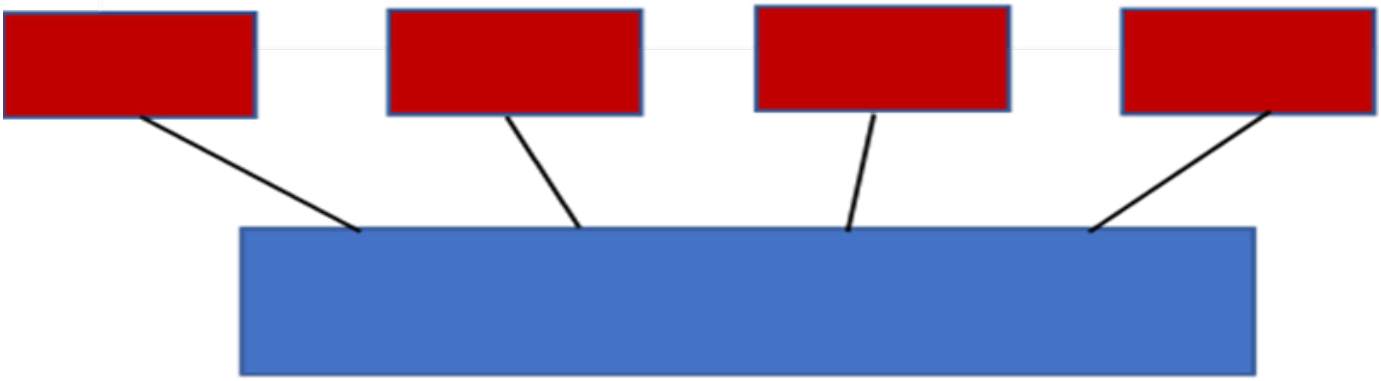
Ce serveur Lenovo ThinkSystem, avec NetApp ONTAP ou NetApp SANtricity, est conçu pour gérer l'inférence d'IA sur de grands datasets grâce à la puissance de traitement des GPU avec les processeurs classiques. Cette validation démontre les performances élevées et une gestion optimale des données grâce à une architecture qui utilise des serveurs Lenovo de périphérie SR350 uniques ou multiples interconnectés à un seul système de stockage AFF NetApp, comme l'illustre les deux figures suivantes.


SE350
SE350
SE350
SE350
10Gb
Ethernet switch
AFF C190



La présentation de l'architecture logique de la figure suivante montre le rôle des éléments de calcul et de stockage dans cette architecture. Elle indique en particulier les éléments suivants :

- Les dispositifs de calcul de périphérie exécutant l'inférence sur les données qu'ils reçoivent des caméras, des capteurs, etc.
- Élément de stockage partagé qui répond à plusieurs besoins :
  - Fournit un emplacement central pour les modèles d'inférence et les autres données requises pour effectuer l'inférence. Les serveurs de calcul accèdent directement au stockage et utilisent les modèles d'inférence à l'échelle du réseau sans avoir à les copier localement.
  - Les modèles mis à jour sont fournis ici.
  - Archive les données d'entrée que les serveurs de périphérie reçoivent pour une analyse ultérieure. Par exemple, si les périphériques sont connectés à des caméras, l'élément de stockage conserve les vidéos capturées par les caméras.



rouge	bleu
Système de calcul Lenovo	Système de stockage NetApp AFF
Périphériques exécutant l'inférence sur les entrées des caméras, des capteurs, etc.	Stockage partagé qui contient les modèles d'inférence et les données de terminaux en périphérie à des fins d'analyse ultérieure.

Cette solution NetApp et Lenovo offre les avantages suivants :

- Les GPU ont accéléré le calcul en périphérie.
- Déploiement de plusieurs serveurs de périphérie sauvegardés et gérés à partir d'un stockage partagé.
- Protection fiable des données pour respecter les objectifs de point de récupération (RPO) et de délai de restauration (RTO) faibles, sans perte de données.
- Gestion des données optimisée grâce aux copies et clones NetApp Snapshot pour rationaliser les workflows de développement.

Utilisation de cette architecture

Ce document valide la conception et les performances de l'architecture proposée. Toutefois, nous n'avons pas testé certains éléments logiciels, tels que les conteneurs, la charge de travail ou la synchronisation des données avec le cloud ou le data Center sur site, car ils sont spécifiques à un scénario de déploiement. Il existe ici plusieurs choix.

Au niveau de la gestion des conteneurs, la gestion des conteneurs Kubernetes est un bon choix et est bien prise en charge dans une version entièrement en amont (Canonical) ou dans une version modifiée adaptée aux déploiements d'entreprise (Red Hat). Le "Plan de contrôle d'IA NetApp" Qui utilise NetApp Trident et le nouveau système ajouté "Kit NetApp DataOps" Traçabilité intégrée, fonctions de gestion des données, interfaces et outils pour les data Scientists et les ingénieurs de données à intégrer au stockage NetApp. Kubeflow, le kit DE ML pour Kubernetes, fournit des fonctionnalités d'IA supplémentaires, ainsi qu'une prise en charge des versions de modèles et KFServing sur plusieurs plateformes telles que TensorFlow ou NVIDIA Triton Inférence Server. Autre option : la plateforme NVIDIA EGX, qui fournit la gestion des charges de travail et l'accès à un catalogue de conteneurs d'inférence d'IA compatibles avec les GPU. Toutefois, il peut être nécessaire de mettre en production des efforts et une expertise considérables. Pour ce faire, il faut parfois l'assistance d'un éditeur indépendant de logiciels ou d'un consultant.

Domaines de solutions

Le principal avantage de l'inférence d'IA et du calcul en périphérie est la capacité des terminaux à calculer, à traiter et à analyser les données avec un niveau élevé de qualité et sans latence. Ce document présente trop d'exemples d'utilisation de Edge Computing, mais voici quelques exemples bien connus :

Automobiles : véhicules autonomes

Le classique illustration se trouve dans les systèmes d'assistance à la conduite automobile (ADAS) pour les véhicules autonomes (AV). Pour être un conducteur sûr, l'IA dans les voitures sans chauffeur doit rapidement traiter un grand nombre de données provenant des caméras et des capteurs. Prendre trop de temps pour interpréter un objet et un être humain peut signifier la vie ou la mort, c'est pourquoi il est essentiel de pouvoir traiter ces données aussi près que possible du véhicule. Dans ce cas, un ou plusieurs serveurs de calcul de périphérie gèrent les entrées des caméras, DES RADARS, lidar et autres capteurs, tandis que le stockage partagé contient les modèles d'inférence et stocke les données d'entrée des capteurs.

Soins de santé : surveillance des patients

L'un des plus grands impacts de l'IA et de l'Edge Computing est sa capacité à améliorer la surveillance continue des patients atteints de maladies chroniques dans les établissements de soins à domicile et les unités de soins intensifs. Les données provenant de dispositifs périphériques qui surveillent les niveaux d'insuline, la respiration, l'activité neurologique, le rythme cardiaque et les fonctions gastro-intestinales nécessitent une analyse instantanée des données qui doivent être traitées immédiatement parce qu'il y a peu de temps pour agir afin de sauver la vie d'une personne.

Vente au détail : paiement sans caissier

L'informatique en périphérie peut alimenter l'IA et LE ML afin d'aider les détaillants à réduire les délais d'exécution et à accroître le trafic au pied. Les systèmes sans caissier prennent en charge différents composants, tels que :

- Authentification et accès. Connecter le client-revendeur physique à un compte validé et autoriser l'accès à l'espace de vente.
- Surveillance des stocks. Utilisation de capteurs, de tags RFID et de systèmes de vision informatique pour confirmer la sélection ou la désélection d'articles par les acheteurs.

  Dans ce cas, chacun des serveurs périphériques gère chaque compteur de réservation et le système de stockage partagé sert de point de synchronisation central.

Services financiers : sécurité humaine dans les kiosques et prévention de la fraude

Les services bancaires exploitent l'IA et l'informatique en périphérie pour innover et créer des expériences bancaires personnalisées. Les kiosques interactifs qui utilisent l'analytique en temps réel et l'inférence d'IA permettent désormais aux clients d'utiliser les guichets automatiques pour se retirer de l'argent, mais ils surveillent de façon proactive les kiosques visant à identifier les risques de sécurité humaine ou de comportements frauduleux. Dans ce scénario, les serveurs de calcul en périphérie et les systèmes de stockage partagé sont connectés à des kiosques et caméras interactifs pour aider les banques à collecter et à traiter des données à l'aide de modèles d'inférence IA.

Fabrication : industrie 4.0

La quatrième révolution industrielle (Industrie 4.0) a commencé, ainsi que les nouvelles tendances comme Smart Factory et l'impression 3D. Pour se préparer à un avenir piloté par les données, une communication M2M (machine-to-machine) à grande échelle et l'Internet des objets sont intégrés pour permettre une automatisation accrue sans intervention humaine. Le secteur de la fabrication est déjà hautement automatisé et l'ajout de fonctionnalités d'IA représente une continuation naturelle de la tendance à long terme. L'IA permet d'automatiser les opérations qui peuvent être automatisées avec la vision par ordinateur et d'autres fonctionnalités d'IA. Vous pouvez automatiser le contrôle de qualité ou les tâches qui reposent sur la vision humaine ou la prise de décision pour réaliser des analyses plus rapides des matériaux sur les lignes d'assemblage dans les usines de fabrication afin d'aider les usines à respecter les normes ISO requises en

matière de sécurité et de gestion de la qualité. Ici, chaque serveur Edge de calcul est connecté à une matrice de capteurs qui surveille le processus de fabrication et les modèles d'inférence mis à jour sont déplacés vers le stockage partagé, si nécessaire.

Télécommunications : détection de la rouille, inspection des tours et optimisation du réseau

Le secteur des télécommunications utilise la vision par ordinateur et des techniques d'IA pour traiter les images qui détectent automatiquement la rouille et identifient les tours cellulaires qui contiennent de la corrosion et requièrent donc une inspection plus poussée. L'utilisation d'images de drone et de modèles d'IA pour identifier des régions distinctes d'une tour pour analyser la rouille, les fissures de surface et la corrosion a augmenté au cours des dernières années. La demande continue d'augmenter pour les technologies d'IA, qui permettent aux infrastructures de télécommunications et aux tours cellulaires d'être inspectées efficacement, évaluées régulièrement pour vérifier leur dégradation et réparées rapidement si nécessaire.

En outre, dans le secteur des télécommunications, l'utilisation d'algorithmes d'IA et DE ML permet de prévoir les modèles de trafic de données, de détecter des périphériques compatibles 5G et d'automatiser et d'augmenter la gestion de l'énergie à entrées multiples et à sorties multiples (MIMO). Le matériel MIMO est utilisé dans les tours radio pour augmenter la capacité du réseau, mais cela est fourni avec des coûts d'énergie supplémentaires. Les modèles ML pour le « mode veille MIMO » déployés sur les sites cellulaires permettent de prévoir l'utilisation efficace des radios et de réduire les coûts de consommation d'énergie pour les opérateurs de réseaux mobiles (MNO). Les solutions d'inférence et de calcul en périphérie d'IA permettent aux entreprises non seulement de réduire la quantité de données transmises aux data centers, mais aussi de réduire leur coût total de possession, d'optimiser les opérations réseau et d'améliorer les performances globales des utilisateurs.

Présentation de la technologie

Cette section décrit la base technologique de cette solution d'IA.

Systèmes NetApp AFF

Les systèmes de stockage NetApp AFF de pointe permettent les déploiements d'inférence d'IA à la périphérie de répondre aux besoins de stockage des entreprises avec les meilleures performances du secteur, la flexibilité supérieure, l'intégration au cloud et une gestion de données optimale. Conçues spécifiquement pour les systèmes Flash, les AFF NetApp contribuent à accélérer, gérer et protéger les données stratégiques.

- Les systèmes de stockage NetApp AFF d'entrée de gamme sont basés sur du matériel FAS2750 et sur les disques SSD Flash
- Deux contrôleurs en configuration haute disponibilité

Les fonctionnalités des systèmes de stockage AFF C190 d'entrée de gamme de NetApp sont les suivantes :

- Le nombre maximal de disques SSD 24 x 960 Go
- Deux configurations possibles :
  - Ethernet (10GbE) : 4 ports 10GBASE-T (RJ-45)
  - Unifié (FC 16 Gb ou 10 GbE) : 4 ports UTA2 (adaptateur « Unified Target » 2)
- Un maximum de 50,5 To de capacité effective

> Pour les workloads NAS, un seul système AFF C190 d'entrée de gamme prend en charge un débit de 4,4 Gbit/s pour les lectures séquentielles, et 230 000 IOPS pour les lectures aléatoires de petite taille à des latences inférieures ou inférieures à 1 ms.

Avec AFF A220

NetApp propose également d'autres systèmes de stockage d'entrée de gamme qui offrent des performances et une évolutivité supérieures pour les déploiements à plus grande échelle. Pour les charges de travail NAS, un système AFF A220 unique d'entrée de gamme prend en charge :

- Débit de 6,2 Gbit/s pour les lectures séquentielles
- 375 000 IOPS pour les lectures aléatoires de petite taille à des latences inférieures ou inférieures à 1 ms.
- Le nombre maximum de disques de 144 SSD de 960 Go, de 3,8 To ou 7,6 To
- AFF A220 évolue jusqu'à plus de 1 po de capacité effective

NetApp AFF A250

- La capacité effective maximale est de 35 po avec une évolutivité scale-out maximale de 2-24 nœuds (12 paires HA)
- Offre une augmentation ≥ 45 % des performances par rapport à AFF A220
- 440 000 IOPS en lectures aléatoires à 1 ms.
- Basé sur la dernière version de NetApp ONTAP : ONTAP 9.8

- Exploite deux Ethernet 25 Gb pour la haute disponibilité et l'interconnexion de clusters

Baies NetApp E-Series EF

La gamme EF-Series est une gamme de baies de stockage SAN d'entrée de gamme et de milieu de gamme. Elles accélèrent l'accès à vos données et vous permettent de les valoriser plus rapidement grâce au logiciel NetApp SANtricity. Proposant un stockage Flash NVMe et SAS, ces systèmes incluent des IOPS exceptionnelles, des temps de réponse inférieurs à 100 microsecondes et 44 Gbit/s de bande passante. Ils sont ainsi parfaitement adaptés aux workloads mixtes et aux applications exigeantes telles que l'inférence d'IA et le calcul haute performance (HPC).

La figure suivante montre la baie de stockage NetApp EF280.

NetApp EF280

- Prise en charge des protocoles FC 32 Gb/16 Gb, iSCSI 25 Gb/10 Gb et SAS 12 Gb
- La capacité effective maximale est de 96 disques totalisant 1,5 po
- Débit de 10 Gbit/s (lectures séquentielles)
- 300 000 IOPS (lectures aléatoires)
- La baie NetApp EF280 est la baie 100 % Flash la moins chère du portefeuille NetApp

NetApp EF300

- 24 disques SSD NVMe pour une capacité totale de 367 To
- Vous pouvez étendre des options d'extension au total 240 disques durs NL-SAS, 96 disques SSD SAS, ou une combinaison
- NVMe/IB 100 Gb, NVMe/RoCE, iser/IB et SRP/IB
- NVME/FC 32 GB, FCP
- ISCSI 25 Gb
- 20 Gbit/s (lectures séquentielles)
- 670 000 IOPS (lectures aléatoires)

Pour plus d'informations, reportez-vous à la section "Fiche technique des baies 100 % Flash NetApp EF-Series EF600, F300, EF570 et EF280".

NetApp ONTAP 9

ONTAP 9.8.1, la dernière génération de logiciel de gestion du stockage de NetApp, permet aux entreprises de moderniser l'infrastructure et de passer à un data Center prêt pour le cloud. Avec des capacités de gestion des données à la pointe du secteur, ONTAP permet de gérer et de protéger les données avec un seul ensemble d'outils, quel que soit leur emplacement. Vous pouvez aussi déplacer vos données librement partout où elles sont nécessaires : la périphérie, le cœur ou le cloud. ONTAP 9.8.1 comprend de nombreuses fonctionnalités qui simplifient la gestion des données, accélèrent et protègent les données stratégiques, et permettent d'utiliser des fonctionnalités d'infrastructure nouvelle génération dans toutes les architectures de cloud hybride.

Gestion simplifiée

La gestion des données est essentielle pour les opérations IT, car elle permet d'utiliser les ressources appropriées pour les applications et les jeux de données. ONTAP inclut les fonctionnalités suivantes pour rationaliser et simplifier les opérations et réduire le coût total d'exploitation :

- **La compaction des données à la volée et la déduplication étendue** la compaction des données réduit le gaspillage d'espace dans les blocs de stockage, et la déduplication augmente considérablement la capacité effective. Cela s'applique aux données stockées localement et à leur placement dans le cloud.
- **Les contrôles de qualité de service (AQoS) granulaires (minimum, maximum et adaptative).** les contrôles de qualité de service (QoS) aident à maintenir les niveaux de performance des applications critiques dans des environnements hautement partagés.
- **NetApp FabricPool** cette fonctionnalité permet une hiérarchisation automatique des données inactives vers des options de stockage en cloud public et privé, notamment Amazon Web Services (AWS), Azure et NetApp StorageGRID. Pour plus d'informations sur FabricPool, voir "TR-4598".

Accélération et protection des données

ONTAP 9 offre des niveaux supérieurs de performances et de protection des données et étend ces fonctionnalités aux méthodes suivantes :

- * Performances et latence plus faible.* ONTAP offre le débit le plus élevé possible à la latence la plus faible possible.
- **Protection des données.** ONTAP fournit des fonctionnalités de protection des données intégrées avec une gestion commune sur toutes les plates-formes.
- **NetApp Volume Encryption (NVE).** ONTAP offre le chiffrement natif au niveau du volume, avec prise en charge de la gestion des clés à la fois intégrée et externe.
- **Colocation et authentification multifactorielle.** ONTAP permet de partager les ressources de l'infrastructure avec les niveaux de sécurité les plus élevés.

Une infrastructure pérenne

ONTAP 9 propose les fonctionnalités suivantes pour répondre aux besoins métier en constante évolution :

- **Évolutivité transparente et continuité de l'activité.** ONTAP prend en charge l'ajout non disruptif de capacité aux contrôleurs et l'évolution scale-out des clusters. Les clients peuvent effectuer la mise à niveau vers les technologies les plus récentes, telles que NVMe et FC 32 Gb, sans migration des données ni panne coûteuse.
- **Connexion au cloud.** ONTAP est le logiciel de gestion de stockage le plus connecté au cloud, avec des options de stockage SDS (ONTAP Select) et des instances natives du cloud (NetApp Cloud Volumes Service) dans tous les clouds publics.
- **Intégration avec les applications émergentes** ONTAP offre des services de données d'entreprise pour

les plates-formes et applications de nouvelle génération, telles que les véhicules autonomes, les villes intelligentes et l'industrie 4.0, en utilisant la même infrastructure qui prend en charge les applications d'entreprise existantes.

NetApp SANtricity

Les systèmes SANtricity de NetApp offrent les meilleures performances, la fiabilité et la simplicité des baies 100 % Flash hybrides E-Series et EF-Series. Optimisez les performances et l'utilisation de vos baies 100 % Flash hybrides E-Series et EF-Series pour les applications nécessitant des charges de travail importantes, notamment l'analytique des données, la vidéosurveillance, et la sauvegarde et la restauration. Avec SANtricity, les tâches de configuration, de maintenance et d'extension de la capacité peuvent être effectuées en garantissant la disponibilité du système de stockage. SANtricity offre d'excellentes fonctionnalités de protection des données et de surveillance proactive, ainsi qu'une sécurité certifiée. System Manager, son interface intégrée, est facile d'emploi. Pour en savoir plus, consultez le "Logiciel SANtricity NetApp E-Series : Fiche technique".

Optimisation des performances

Le logiciel SANtricity combine d'excellentes performances, des IOPS élevées, un haut débit et une faible latence, pour l'analytique, la vidéosurveillance et les applications de sauvegarde. Accélérez les performances des applications à débit d'IOPS élevé et à faible latence, et celles des applications à large bande passante et à haut débit.

Disponibilité optimisée

Réalisez toutes vos tâches de gestion pendant que le stockage reste en ligne. Modifiez la configuration, effectuez la maintenance ou étendez la capacité de stockage sans interrompre les E/S. Bénéficiez d'une fiabilité exceptionnelle avec les fonctionnalités automatisées, la configuration en ligne, la technologie DPP (Dynamic Disk pools), et bien plus encore.

Travaillez en toute sérénité

Le logiciel SANtricity, qui respecte les normes de sécurité les plus strictes, offre d'excellentes fonctionnalités de protection des données et assure une surveillance proactive. System Manager, son interface intégrée, est facile d'emploi. Simplifiez les tâches courantes de gestion du stockage. Obtenez la flexibilité dont vous avez besoin pour un réglage ultra-précis de tous les systèmes de stockage E-Series. Gérez votre système NetApp E-Series grâce à Une interface web intégrée conçue pour simplifier vos workflows de gestion.

NetApp Trident

"Trident" À partir de NetApp, est un orchestrateur de stockage dynamique open source pour Docker et Kubernetes qui simplifie la création, la gestion et la consommation du stockage persistant. Trident, une application Kubernetes native, s'exécute directement dans un cluster Kubernetes. Trident permet de déployer de manière transparente des images de conteneur d'apprentissage profond sur un système de stockage NetApp et offre une expérience haute performance pour les déploiements de conteneurs d'IA. Les utilisateurs de Kubernetes (développeurs DE ML et data Scientists, par exemple) peuvent créer, gérer et automatiser l'orchestration et le clonage pour exploiter les fonctionnalités avancées de gestion de données de NetApp optimisées par la technologie NetApp.

Copie et synchronisation NetApp BlueXP

"Copie et synchronisation BlueXP" Est un service NetApp qui permet une synchronisation sûre et rapide des données. Que vous ayez besoin de transférer des fichiers entre des partages de fichiers NFS ou SMB sur site, NetApp StorageGRID, NetApp ONTAP S3, NetApp Cloud Volumes Service, Azure NetApp Files, Amazon simple Storage Service (Amazon S3), Amazon Elastic File System (Amazon EFS), Azure Blob, Google Cloud

Storage, ou IBM Cloud Object Storage, BlueXP Copy and Sync déplace les fichiers là où vous en avez besoin rapidement et en toute sécurité. Une fois vos données transférées, elles peuvent être utilisées à la source et à la cible. BlueXP Copy and Sync synchronise en continu les données en fonction de votre calendrier prédéfini, en déplaçant uniquement les données modifiées. Le temps et l'argent consacrés à la réplication des données sont ainsi réduits. BlueXP Copy and Sync est un outil SaaS extrêmement simple à configurer et à utiliser. Les transferts de données déclenchés par la copie et la synchronisation BlueXP sont effectués par les courtiers de données. Vous pouvez déployer des courtiers de données BlueXP Copy and Sync dans AWS, Azure, Google Cloud Platform ou sur site.

Serveurs Lenovo ThinkSystem

Les serveurs Lenovo ThinkSystem sont dotés de matériel, de logiciels et de services innovants qui répondent aux défis actuels des clients et offrent une approche évolutive, adaptée et modulaire pour répondre aux défis de demain. Ces serveurs exploitent les meilleures technologies standard du secteur, associées à des innovations Lenovo différenciées, pour offrir la plus grande flexibilité possible aux serveurs x86.

Les principaux avantages du déploiement des serveurs Lenovo ThinkSystem sont les suivants :

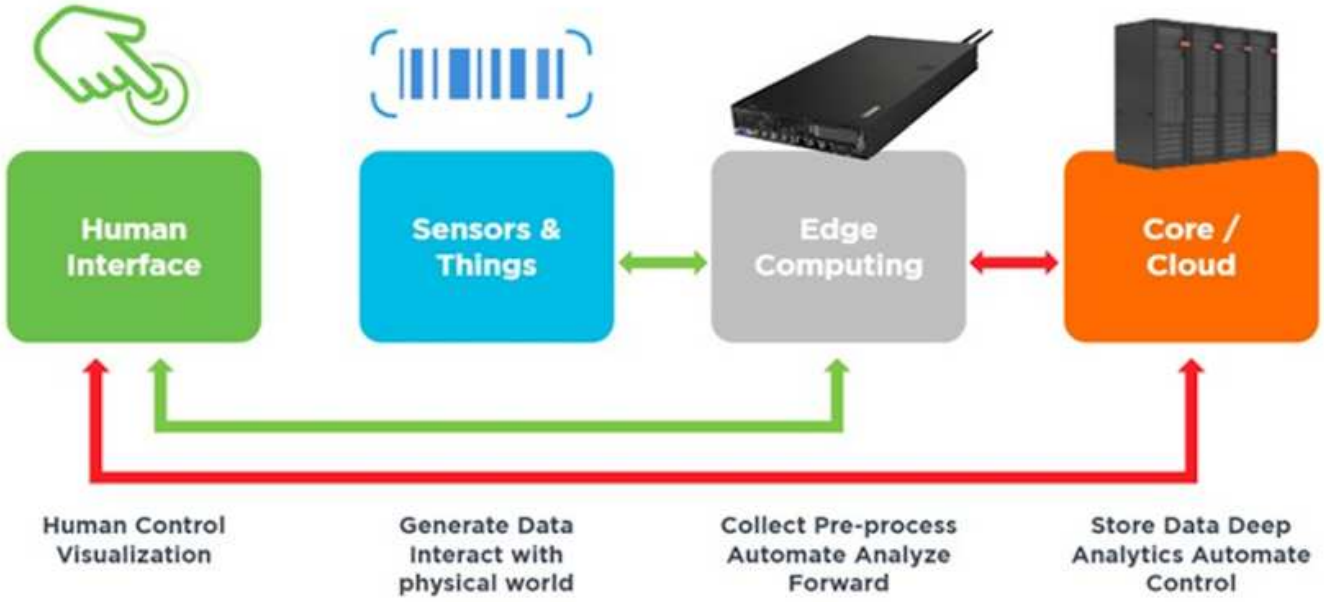
- Des conceptions modulaires extrêmement évolutives qui s'étendent à votre business
- La résilience optimale du secteur pour économiser des heures de temps d'arrêt imprévus coûteux
- Des technologies Flash rapides pour des latences plus faibles, des temps de réponse plus rapides et une gestion intelligente des données en temps réel

Dans le domaine de l'IA, Lenovo propose une approche pratique pour aider les entreprises à comprendre et à exploiter les avantages DU ML et de l'IA pour leurs workloads. Les clients Lenovo peuvent explorer et évaluer les offres d'IA de Lenovo dans les centres d'innovation d'IA de Lenovo afin de connaître pleinement la valeur de leur utilisation. Pour améliorer le retour sur investissement, cette approche axée sur le client permet aux clients de réaliser des démonstrations de faisabilité pour les plateformes de développement de solutions prêtes à l'emploi et optimisées pour l'IA.

Serveur Lenovo ThinkSystem SE350 Edge

Le Edge Computing permet aux données des terminaux IoT d'être analysées à la périphérie du réseau avant d'être envoyées vers le data Center ou le cloud. Le Lenovo ThinkSystem SE350, tel qu'illustré dans la figure ci-dessous, est conçu pour répondre aux exigences uniques de déploiement en périphérie, avec un accent sur la flexibilité, la connectivité, la sécurité et la téléadministration dans un format compact renforcé et résistant à l'environnement.

Doté d'un processeur Intel Xeon D avec la flexibilité nécessaire pour prendre en charge l'accélération des charges de travail Edge ai, le SE350 est conçu pour relever les défis de déploiement de serveurs dans divers environnements en dehors du centre de données.



Diminution des

MLPerf est une suite de banc d'essai leader du secteur pour évaluer les performances de l'IA. Il couvre de nombreux domaines de l'IA appliquée, notamment le classement des images, la détection des objets, l'imagerie médicale et le traitement du langage naturel (NLP). Dans cette validation, nous avons utilisé des charges de travail Inférence v0.7, qui est la dernière itération de l'Inférence MLPerf à la fin de cette validation. Le "Inférence MLPerf v0.7" la suite comprend quatre nouveaux bancs d'essai pour les systèmes de data center et de périphérie :

- **BERT.** Encoder bidirectionnel représentation des transformateurs (BERT) affinée pour répondre aux questions en utilisant le jeu de données de l'équipe.
- **DLRM.** le modèle de recommandation en apprentissage profond (DLRM) est un modèle de personnalisation et de recommandation qui est formé pour optimiser les taux de clics (CTR).
- **3D U-Net.** l'architecture 3D U-Net est formée sur le dataset de segmentation de la tumeur cérébrale (brats).

- **RNN-T.** transducteur de réseau neuronal récurrent (RNN-T) est un modèle de reconnaissance vocale automatique (ASR) qui est entraîné sur un sous-ensemble de LibriSpeech. Les résultats et le code MLPerf Inférence sont accessibles au public et publiés sous licence Apache. MLPerf Inférence possède une division Edge qui prend en charge les scénarios suivants :
- **Single stream.** ce scénario imite les systèmes où la réactivité est un facteur critique, comme les requêtes ai hors ligne effectuées sur les smartphones. Les requêtes individuelles sont envoyées au système et les temps de réponse sont enregistrés. le résultat est une latence du 90e centile de toutes les réponses.
- **Multistream.** ce banc d'essai est destiné aux systèmes qui traitent l'entrée à partir de plusieurs capteurs. Pendant le test, les requêtes sont envoyées à un intervalle de temps fixe. Une contrainte de QoS (latence maximale autorisée) est imposée. Le test indique le nombre de flux que le système peut traiter tout en respectant la contrainte QoS.
- **Hors ligne.** c'est le scénario le plus simple pour les applications de traitement par lots et la mesure est le débit en échantillons par seconde. Toutes les données sont disponibles pour le système et le banc d'essai mesure le temps nécessaire pour traiter tous les échantillons.

Lenovo a publié les scores d'inférence MLPerf pour SE350 avec T4, le serveur utilisé dans ce document. Voir les résultats à "https://mlperf.org/inference-results-0-7/" Dans la section "Edge, Closed Division" de l'entrée 0.7-145.

Plan de test

Ce document suit l'Inférence MLPerf v0.7 "code", MLPerf Inférence v1.1 "code", et "règles". Nous avons exécuté des bancs d'essai MLPerf conçus pour l'inférence à la périphérie, comme défini dans le tableau suivant.

De service	Tâche	Modèle	Jeu de données	Taille QSL	Qualité	La contrainte de latence multiflux
Vision	Classification des images	Reset50v1.5	ImageNET (224 x 224)	1024	99 % de FP32	50 ms.
Vision	Détection d'objet (grande taille)	SSD-ResNet34	COCO (1200 x 1200)	64	99 % de FP32	66 ms.
Vision	Détection d'objet (petite taille)	SSD : MobileNetsv1	COCO (300 x 300)	256	99 % de FP32	50 ms.
Vision	Segmentation des images médicales	NON-ET 3D	Brages 2019 (224 x 224 x 160)	16	99 % et 99.9 % du FP32	s/o
Voix	Parole au texte	RNNT	Lirispeech dev-Clean	2513	99 % de FP32	s/o
Langue	Traitement de la langue	BERT	Squad v1.1	10833	99 % de FP32	s/o

Le tableau suivant présente des scénarios de banc d'essai pour Edge.

De service	Tâche	Scénarios
Vision	Classification des images	Flux unique, hors ligne, flux multiples
Vision	Détection d'objet (grande taille)	Flux unique, hors ligne, flux multiples
Vision	Détection d'objet (petite taille)	Flux unique, hors ligne, flux multiples
Vision	Segmentation des images médicales	Flux unique, hors ligne
Voix	Parole-à-texte	Flux unique, hors ligne
Langue	Traitement de la langue	Flux unique, hors ligne

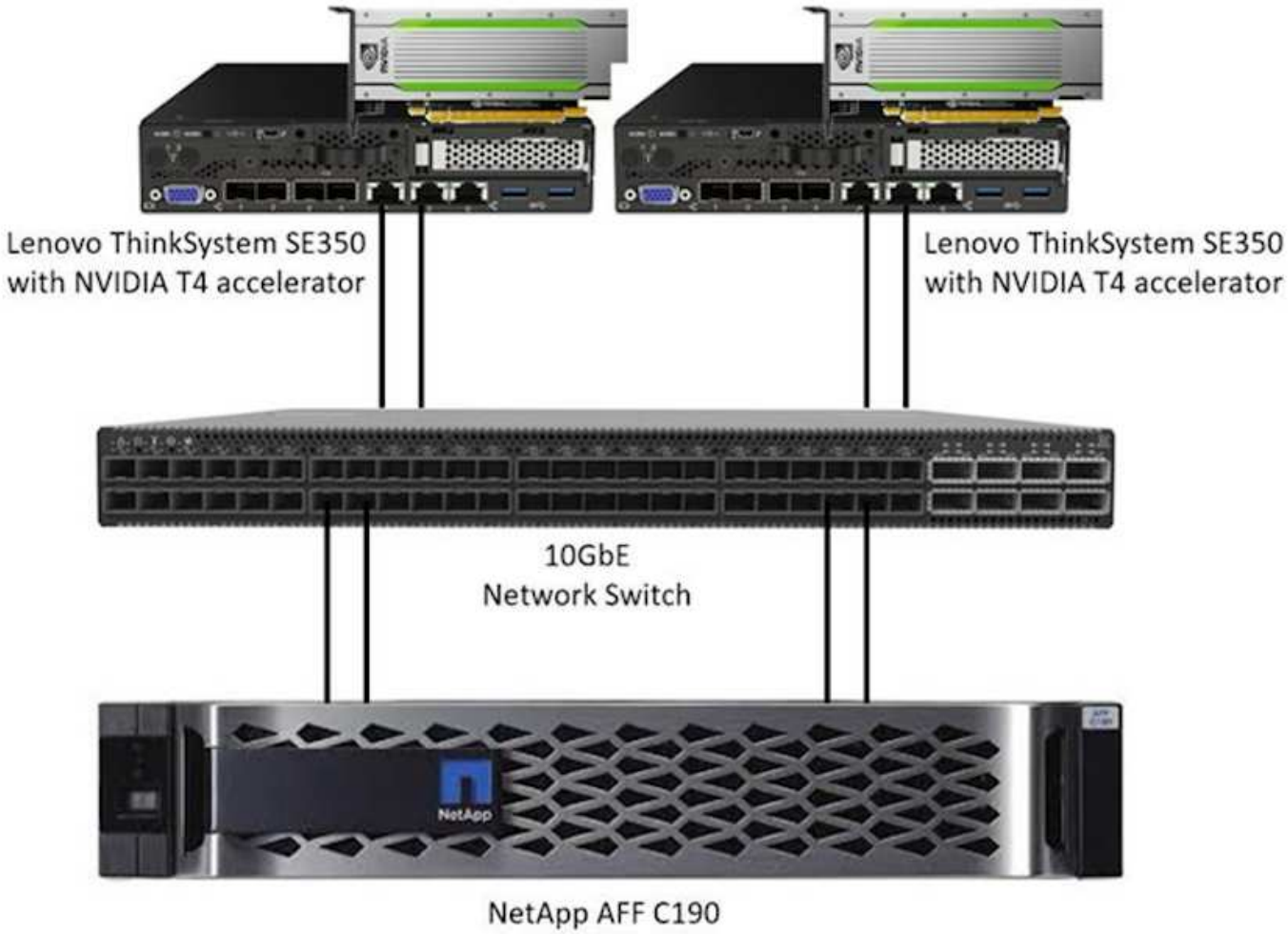
Nous avons réalisé ces bancs d'essai à l'aide de l'architecture de stockage en réseau développée lors de cette validation et avons comparé les résultats à ceux des exécutions locales sur les serveurs de périphérie préalablement soumis à MLPerf. La comparaison consiste à déterminer l'impact du stockage partagé sur les performances d'inférence.

Configuration du test

La figure suivante présente la configuration de test. Nous avons utilisé le système de stockage NetApp AFF C190 et deux serveurs Lenovo ThinkSystem SE350 (chacun avec un accélérateur NVIDIA T4). Ces composants sont connectés via un commutateur réseau 10GbE. Le stockage réseau contient des datasets de validation/test et des modèles pré-entraînés. Les serveurs fournissent des fonctionnalités de calcul, et le stockage est accessible via le protocole NFS.

Cette section décrit les configurations testées, l'infrastructure réseau, le serveur SE350 et les détails du provisionnement de stockage. Le tableau suivant répertorie les composants de base de l'architecture de la solution.

Composants de la solution	Détails
Serveurs Lenovo ThinkSystem	• 2 serveurs SE350 chacun avec une carte graphique NVIDIA T4
	• Chaque serveur contient un processeur Intel Xeon D-2123IT avec quatre cœurs physiques fonctionnant à 2,20 GHz et 128 Go de RAM
Système de stockage NetApp AFF d'entrée de gamme (paire HA)	• Le logiciel NetApp ONTAP 9 • SSD 24x 960 Go • Protocole NFS • Un groupe d'interface par contrôleur, avec quatre adresses IP logiques pour les points de montage



Le tableau suivant répertorie la configuration du stockage : AFF C190 avec 24 emplacements de disque 2RU.

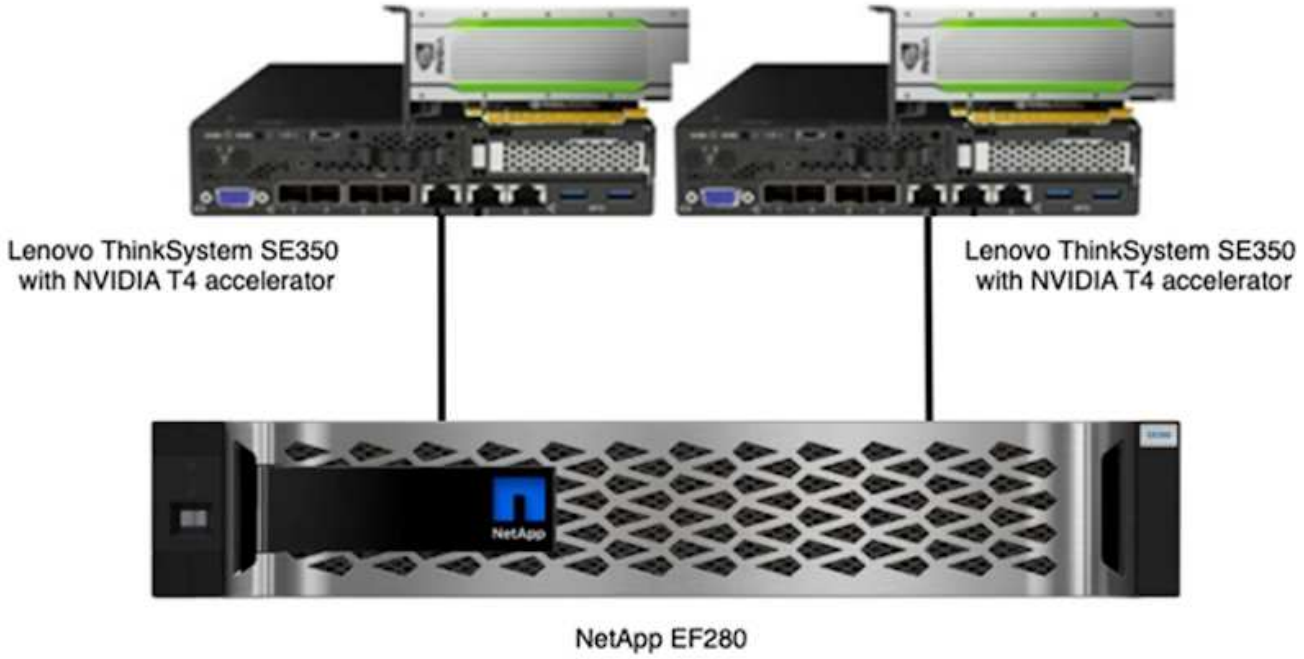
Contrôleur	Agrégat	Volume FlexGroup	Agrégez-les	Volumétrique	Point de montage du système d'exploitation
Contrôleur 1	Agr1	/Netapplenovo_ai_fg	8,42 Tio	15 TO	/netapp_lenovo_fg
Contrôleur 2	Agr2		8,42 Tio		

Le dossier /netappLenovo_ai_fg contient les ensembles de données utilisés pour la validation du modèle.

La figure ci-dessous présente la configuration de test. Nous avons utilisé le système de stockage NetApp EF280 et deux serveurs Lenovo ThinkSystem SE350 (chacun avec un accélérateur NVIDIA T4). Ces composants sont connectés via un commutateur réseau 10GbE. Le stockage réseau contient des datasets de validation/test et des modèles pré-entraînés. Les serveurs fournissent des fonctionnalités de calcul, et le stockage est accessible via le protocole NFS.

Le tableau ci-dessous répertorie la configuration de stockage de la baie EF280.

Contrôleur	Groupe de volumes	Volumétrie	Volumétrique	DDPsize	Méthode de connexion
Contrôleur 1	DDP1	Volume 1	8,42 Tio	16 TO	SE350-1 vers LUN iSCSI 0
Contrôleur 2		Volume 2	8,42 Tio		SE350-2 vers la LUN iSCSI 1



Procédure de test

Cette section décrit les procédures de test utilisées pour valider cette solution.

Installation du système d'exploitation et de l'inférence d'IA

Pour la baie AFF C190, nous avons utilisé Ubuntu 18.04 avec les pilotes NVIDIA et docker avec la prise en charge des GPU NVIDIA et MLPerf "code" Disponible dans le cadre de la soumission Lenovo à MLPerf Inférence v0.7.

Pour l'EF280, nous avons utilisé Ubuntu 20.04 avec les pilotes NVIDIA et docker avec prise en charge des GPU NVIDIA et MLPerf "code" Disponible dans le cadre de la soumission Lenovo à MLPerf Inférence v1.1.

Pour configurer l'inférence d'IA, procédez comme suit :

1. Téléchargez les datasets qui nécessitent un enregistrement, le jeu de validation ImageNet 2012, le jeu de données Criteo Terabyte et brats 2019 Training Set, puis décompressez les fichiers.
2. Créez un répertoire de travail d'au moins 1 To et définissez une variable d'environnement `MLPERF_SCRATCH_PATH` se référant au répertoire.

   Vous devez partager ce répertoire sur le stockage partagé pour le cas d'utilisation du stockage réseau ou sur le disque local lors des tests avec des données locales.

3. Exécuter la marque `prebuild` commande, qui crée et lance le conteneur docker pour les tâches d'inférence requises.

> Les commandes suivantes sont toutes exécutées depuis le conteneur Docker en cours d'exécution :

- Télécharger les modèles d'IA pré-entraînés pour les tâches d'inférence MLPerf : `make download_model`
- Téléchargez des datasets supplémentaires qui peuvent être téléchargés librement : `make download_data`
- Prétraiter les données : marque `preprocess_data`
- Exécuter : `make build`.
- Création de moteurs d'inférence optimisés pour les GPU dans les serveurs de calcul : `make generate_engines`
- Pour exécuter des workloads Inférence, exécutez les éléments suivants (une commande) :

```
make run_harness RUN_ARGS="--benchmarks=<BENCHMARKS>
--scenarios=<SCENARIOS>"
```

L'inférence d'IA s'exécute

Trois types de passages ont été exécutés :

- Inférence d'IA à un seul serveur avec le stockage local
- Inférence d'IA à un serveur unique avec le stockage réseau
- Inférence d'IA à plusieurs serveurs à l'aide d'un stockage réseau

Résultats du test

Une multitude de tests ont été effectués afin d'évaluer les performances de l'architecture proposée.

Il existe six charges de travail différentes (classification des images, détection des objets [petite taille], détection des objets [grande], imagerie médicale, synthèse vocale, Et le traitement du langage naturel [NLP]), que vous pouvez exécuter dans trois scénarios différents : hors ligne, flux unique et flux multiples.

> Le dernier scénario est implémenté uniquement pour la classification des images et la détection des objets.

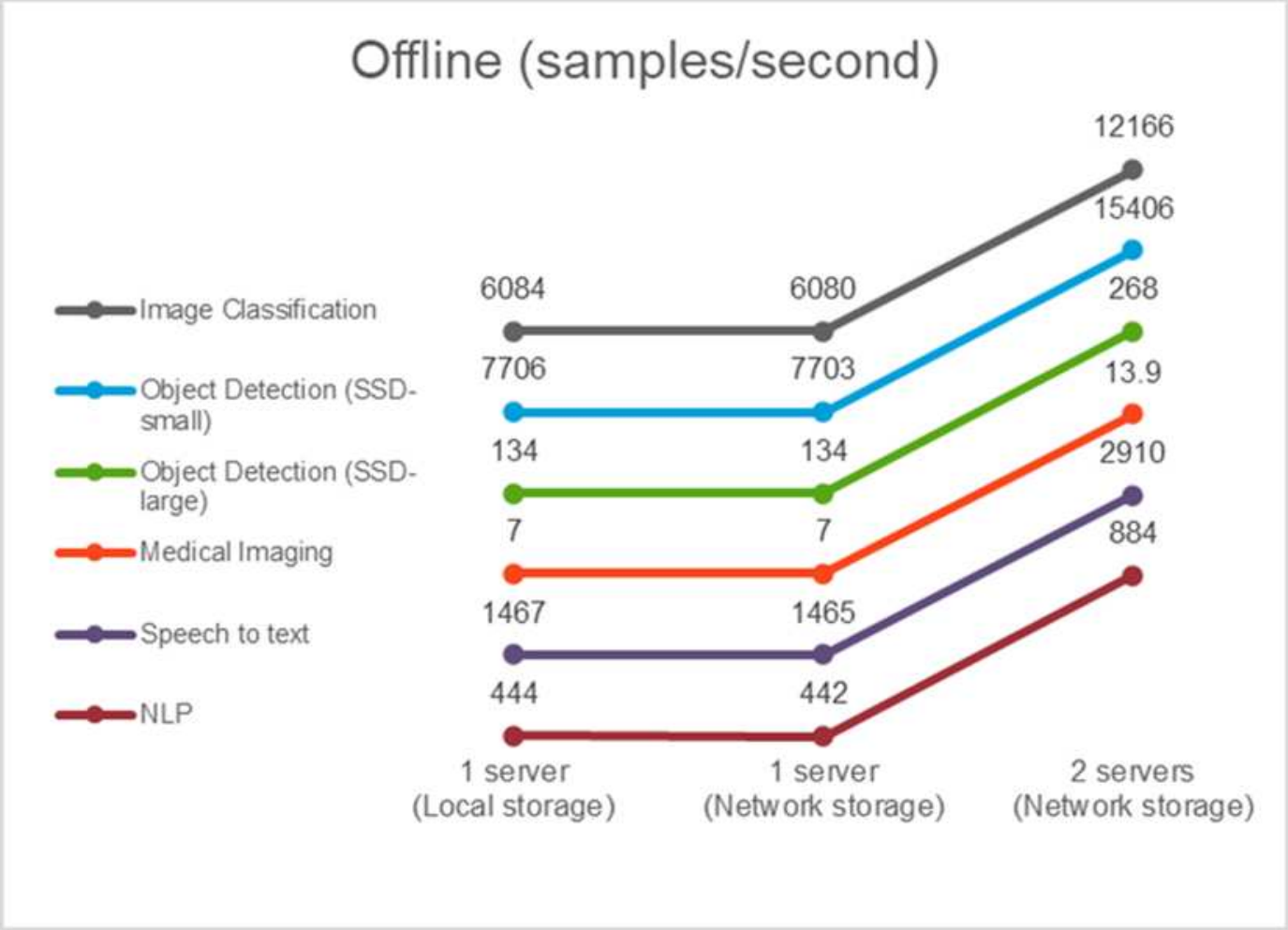
Ainsi, 15 charges de travail ont été testées sous trois configurations différentes :

- Stockage local/serveur unique
- Stockage à serveur unique/réseau
- Stockage multiserveur/réseau

Les résultats sont décrits dans les sections suivantes.

Inférence d'IA dans un scénario hors ligne pour AFF

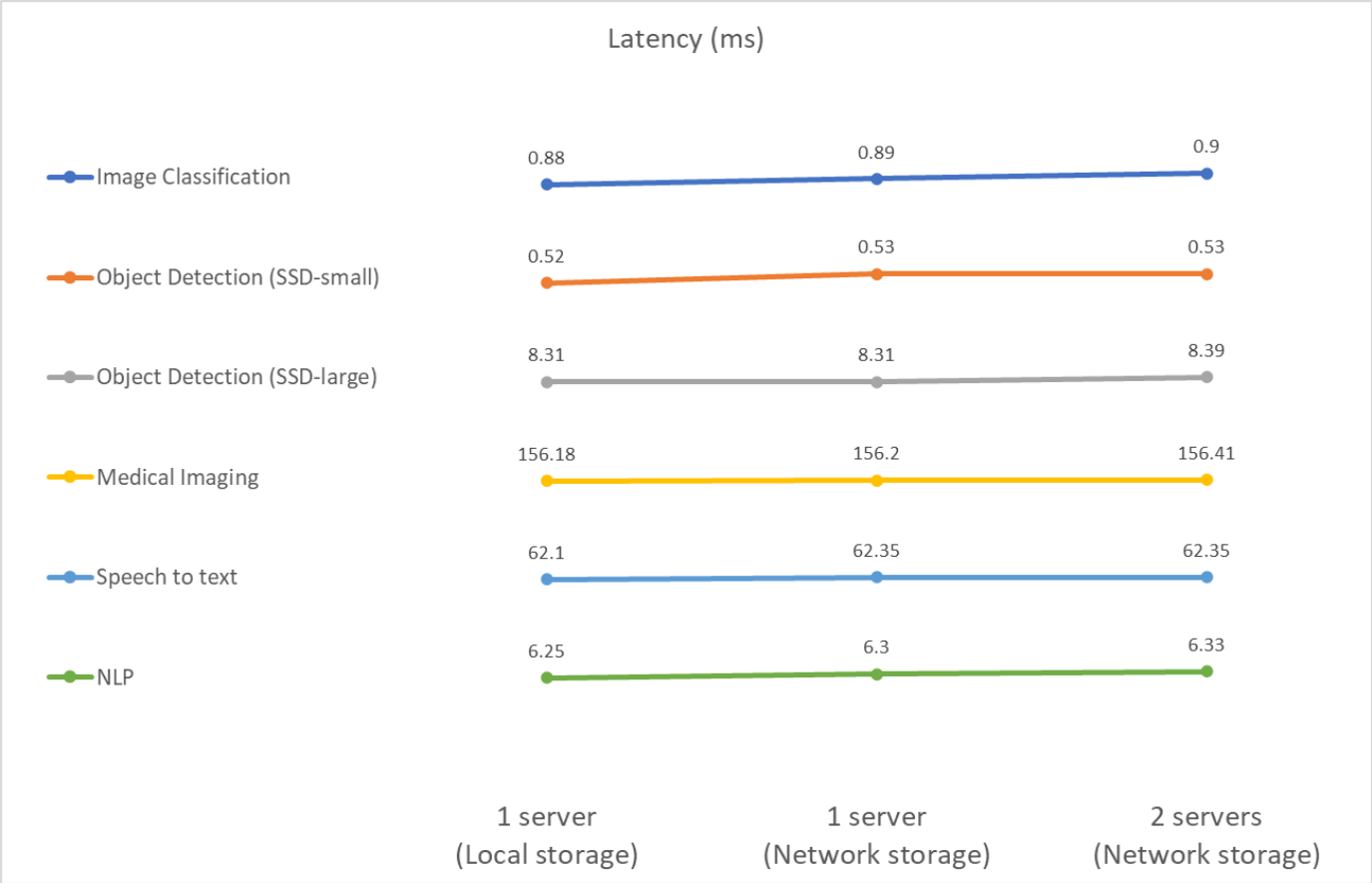
Dans ce scénario, toutes les données étaient disponibles pour le serveur et le temps nécessaire pour traiter tous les échantillons a été mesuré. Nous avons signalé des largeurs de bande dans les échantillons par seconde comme résultats des tests. Lorsque plus d'un serveur de calcul était utilisé, nous faisons état de la bande passante totale additionnée sur tous les serveurs. Les résultats des trois cas d'utilisation sont présentés dans la figure ci-dessous. Pour le cas des deux serveurs, nous avons signalé la bande passante combinée des deux serveurs.



Les résultats montrent que le stockage en réseau n'affecte pas les performances : le changement est minime et aucune n'a été trouvée pour certaines tâches. Lorsque vous ajoutez le second serveur, la bande passante totale double exactement ou, dans le pire des cas, cette modification est inférieure à 1 %.

L'inférence d'IA dans un scénario de flux unique pour le AFF

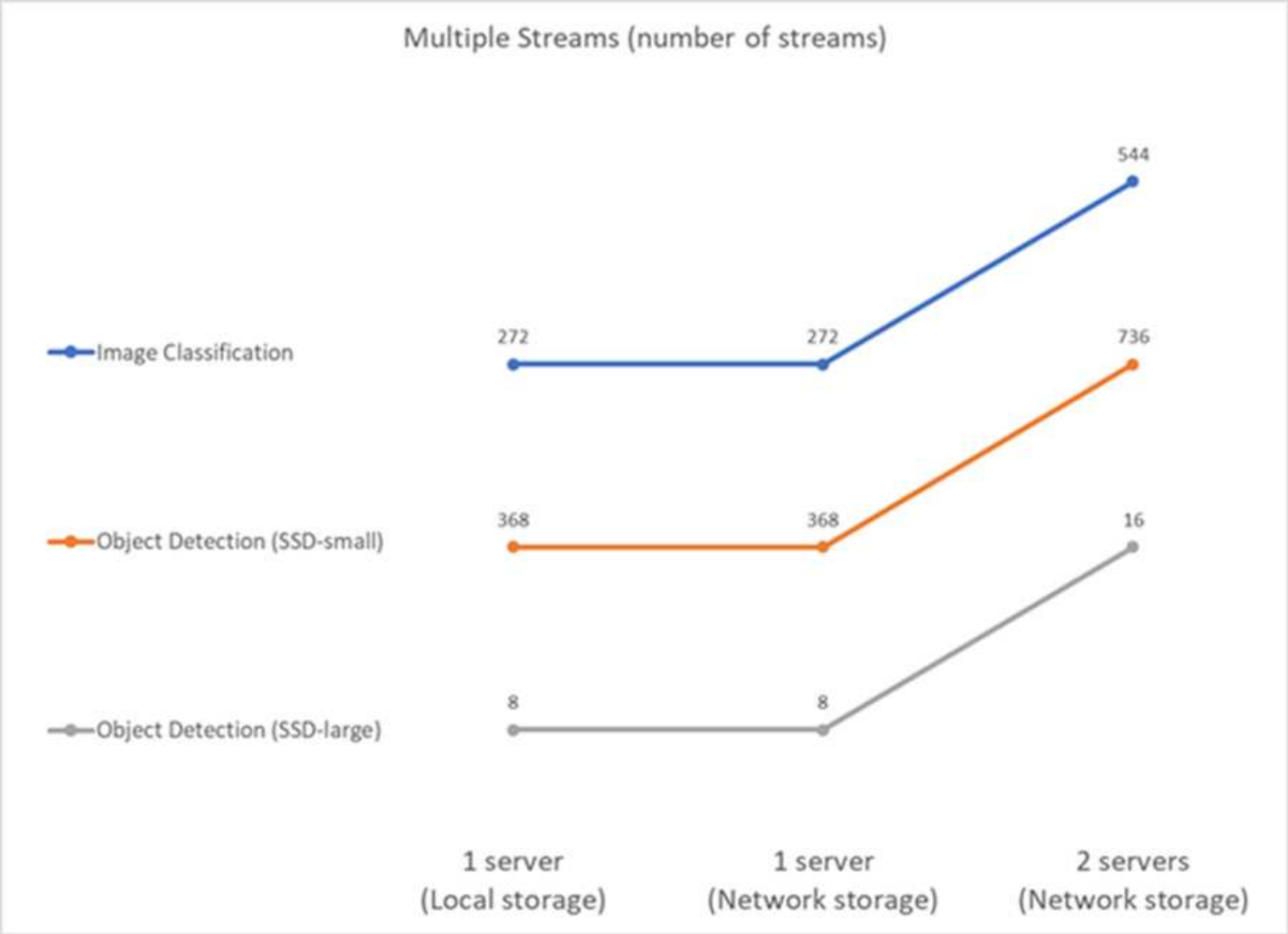
Ce banc d'essai mesure la latence. Dans le cas d'un serveur de calcul multiple, nous faisons état de la latence moyenne. Les résultats de la suite de tâches sont indiqués dans la figure ci-dessous. Pour le cas des deux serveurs, nous avons signalé la latence moyenne des deux serveurs.



Les résultats, encore une fois, montrent que le stockage réseau suffit à gérer les tâches. La différence entre le stockage local et le stockage réseau dans un même cas de serveur est minime, voire aucune. De même, lorsque deux serveurs utilisent le même stockage, la latence sur les deux serveurs reste inchangée ou les modifications par une très petite quantité.

Inférence IA dans le scénario à flux multiples pour le AFF

Dans ce cas, le résultat est le nombre de flux que le système peut traiter tout en satisfaisant la contrainte QoS. Ainsi, le résultat est toujours un entier. Pour plus d'un serveur, nous faisons état du nombre total de flux sur tous les serveurs. Toutes les charges de travail ne prennent pas en charge ce scénario, mais nous en avons fait toutes. Les résultats de nos tests sont résumés dans la figure ci-dessous. Dans le cas des deux serveurs, nous faisons état du nombre combiné de flux provenant des deux serveurs.



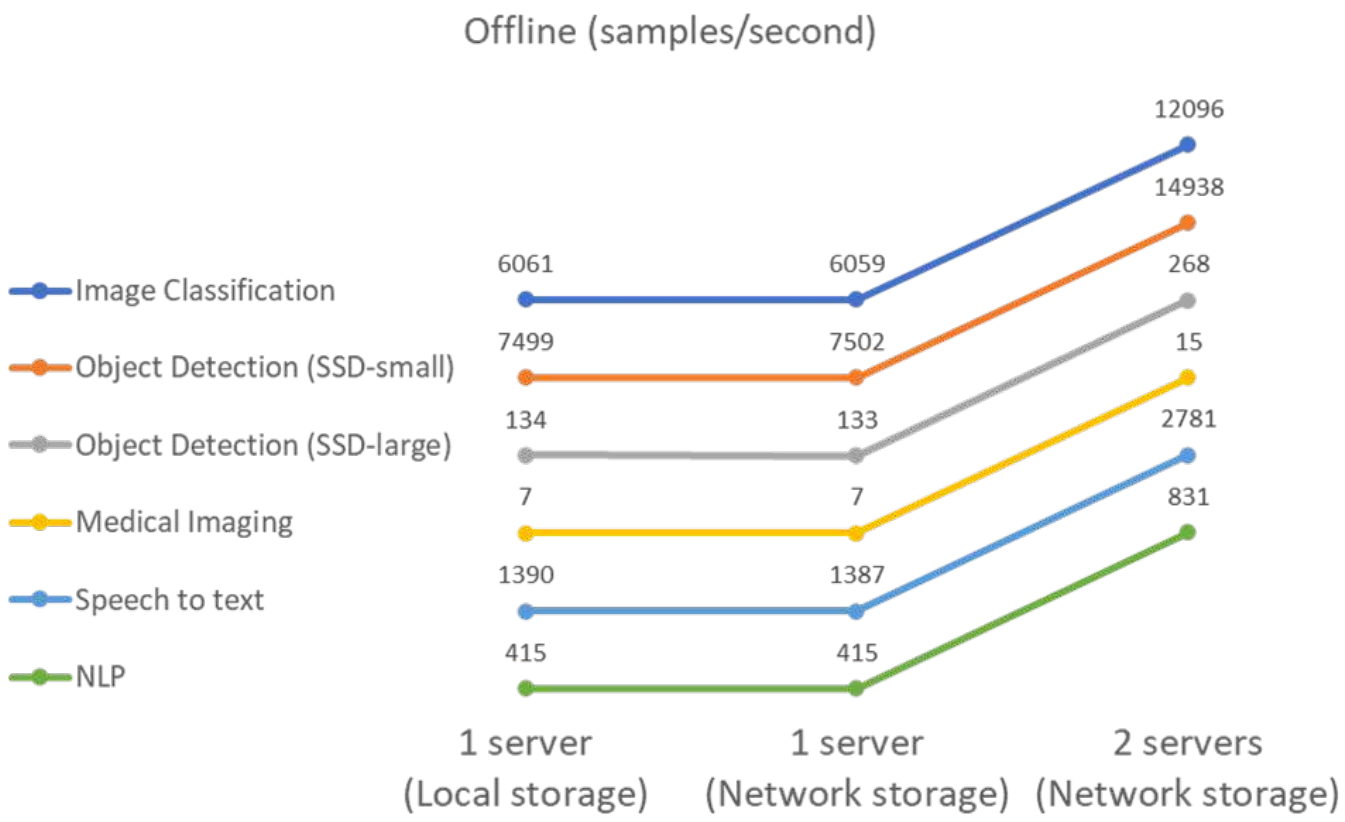
Les résultats montrent des performances parfaites de la configuration : le stockage local et réseau donne les mêmes résultats et l'ajout du second serveur double le nombre de flux que la configuration proposée peut gérer.

Résultats des tests pour EF

Une multitude de tests ont été effectués afin d'évaluer les performances de l'architecture proposée. Il existe six charges de travail différentes (classification des images, détection des objets [petite taille], détection des objets [grande], imagerie médicale, synthèse vocale, Et traitement du langage naturel [NLP]), qui ont été exécutés dans deux scénarios différents : hors ligne et flux unique. Les résultats sont décrits dans les sections suivantes.

Inférence d'IA dans un scénario hors ligne pour la baie EF

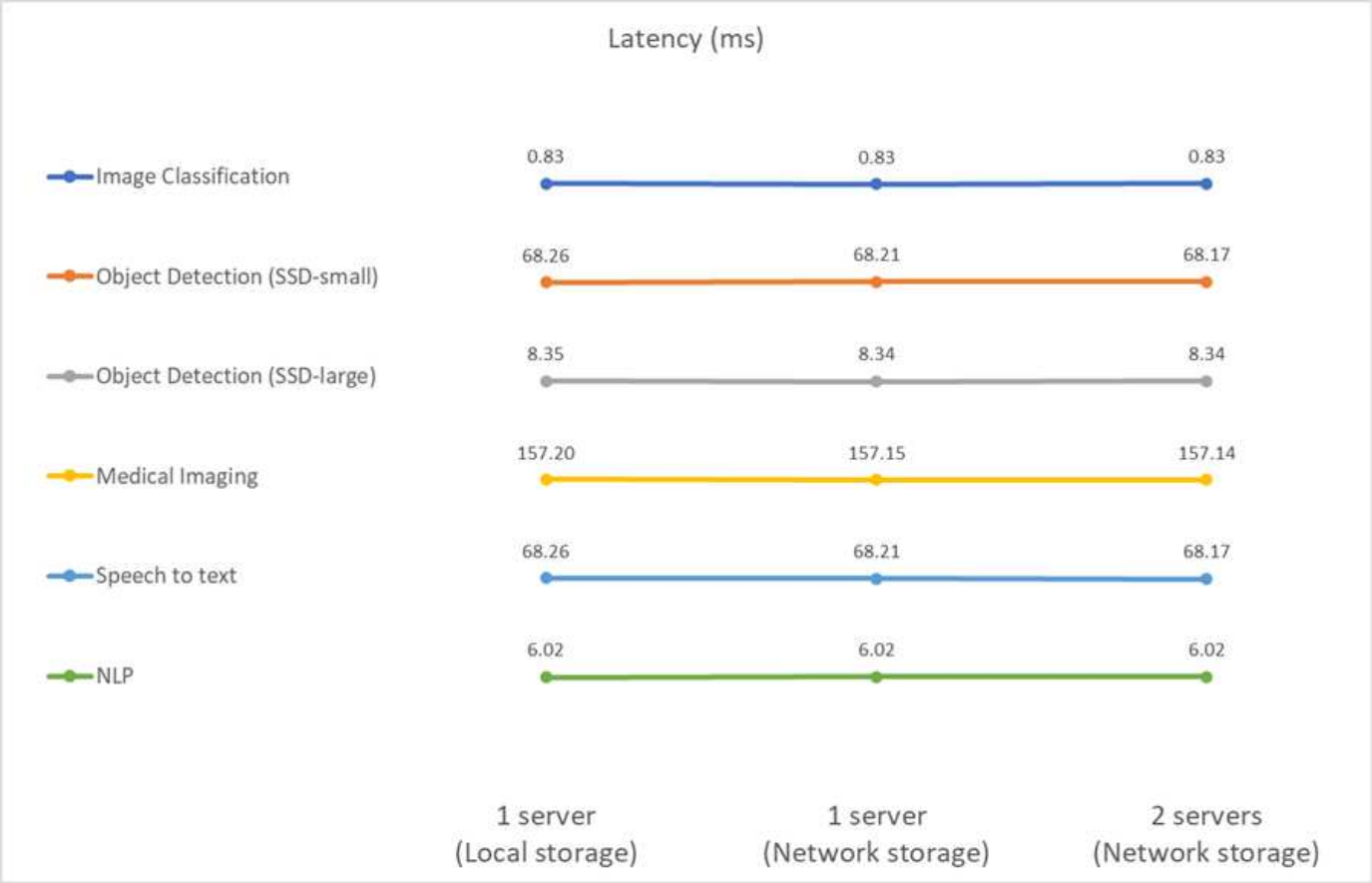
Dans ce scénario, toutes les données étaient disponibles pour le serveur et le temps nécessaire pour traiter tous les échantillons a été mesuré. Nous avons signalé des largeurs de bande dans les échantillons par seconde comme résultats des tests. Pour les exécutions d'un nœud, nous faisons état d'une moyenne des deux serveurs, tandis que pour deux exécutions de serveur, nous avons signalé que la bande passante totale a été additionnée sur l'ensemble des serveurs. Les résultats pour des cas d'utilisation sont présentés dans la figure ci-dessous.



Les résultats montrent que le stockage en réseau n'affecte pas les performances : le changement est minime et aucune n'a été trouvée pour certaines tâches. Lorsque vous ajoutez le second serveur, la bande passante totale double exactement ou, dans le pire des cas, cette modification est inférieure à 1 %.

Scénario d'inférence d'IA dans un flux unique pour la gamme EF

Ce banc d'essai mesure la latence. Dans tous les cas, nous faisons état d'une latence moyenne sur tous les serveurs en cours d'exécution. Les résultats de la suite de tâches sont donnés.



Les résultats indiquent à nouveau que le stockage réseau suffit à gérer les tâches. La différence entre le stockage local et réseau dans un cas de serveur est minime ou aucun. De même, lorsque deux serveurs utilisent le même stockage, la latence sur les deux serveurs reste inchangée ou les modifications par une très petite quantité.

Options de dimensionnement de l'architecture

Vous pouvez ajuster la configuration utilisée pour la validation afin qu'elle s'adapte à d'autres cas d'utilisation.

Serveur de calcul

Nous avons utilisé un processeur Intel Xeon D-2123IT, qui est le niveau de CPU le plus bas pris en charge dans SE350, avec quatre cœurs physiques et 60 W TDP. Bien que le serveur ne prend pas en charge le remplacement des processeurs, il peut être commandé avec un processeur plus puissant. Le processeur le plus pris en charge est le processeur Intel Xeon D-2183IT avec 16 cœurs, 100 W fonctionnant à 2,20 GHz. Cela augmente considérablement la capacité de calcul du processeur. Bien que le processeur ne constitue pas un goulot d'étranglement pour l'exécution des workloads d'inférence eux-mêmes, il facilite le traitement des données et d'autres tâches liées à l'inférence. Actuellement, NVIDIA T4 est le seul GPU disponible pour les cas d'utilisation en périphérie. Il n'est donc pas possible de mettre à niveau ou de rétrograder le GPU.

Stockage partagé

À des fins de test et de validation, le système NetApp AFF C190, qui possède une capacité de stockage maximale de 50,5 To, un débit de 4,4 Gbit/s pour les lectures séquentielles et 230 000 IOPS pour les lectures aléatoires de petite taille, a été utilisé pour ce document. Il est prouvé qu'il est parfaitement adapté aux workloads d'inférence à la périphérie.

Toutefois, si vous avez besoin de plus de capacité de stockage ou de vitesses réseau plus rapides, vous devriez utiliser NetApp AFF A220 ou bien "NetApp AFF A250" systèmes de stockage netapp fas. Elle a également permis d'utiliser le système NetApp EF280, dont la capacité maximale est de 1,5 po et 10 Gbit/s de bande passante, pour cette validation. Si vous préférez augmenter votre capacité de stockage avec une bande passante plus élevée, "NetApp EF300" peut être utilisé.

Conclusion

L'automatisation basée sur l'IA et l'Edge Computing constituent une approche majeure pour aider les entreprises à réussir la transformation digitale et à optimiser l'efficacité opérationnelle et la sécurité. Avec l'Edge Computing, le traitement des données est beaucoup plus rapide car il n'est pas nécessaire de se rendre au data Center ni de les déplacer. C'est pourquoi les coûts associés à l'envoi de données vers des data centers ou le cloud sont réduits. Une latence plus faible et une vitesse plus élevée peuvent être utiles pour les entreprises qui doivent prendre des décisions en temps quasi réel à l'aide de modèles d'inférence d'IA déployés à la périphérie.

Les systèmes de stockage NetApp offrent les performances identiques ou supérieures au stockage SSD local et offrent les avantages suivants aux data Scientists, aux ingénieurs du data Center, aux développeurs d'IA/ML et aux décideurs IT ou business :

- Partage sans effort des données entre des systèmes d'IA, l'analytique et d'autres systèmes métier stratégiques. Ce partage des données permet de réduire la surcharge de l'infrastructure, d'améliorer les performances et de rationaliser la gestion des données à l'échelle de l'entreprise.
- Faites évoluer indépendamment les ressources de calcul et de stockage pour réduire les coûts et améliorer l'utilisation des ressources.
- Rationalisation des workflows de développement et de déploiement à l'aide de copies et de clones Snapshot intégrés pour des espaces de travail utilisateur instantanés et compacts, un contrôle intégré des versions et un déploiement automatisé.
- Une protection des données d'entreprise pour la reprise après incident et la continuité de l'activité. La solution NetApp et Lenovo présentée dans ce document est une architecture scale-out flexible, idéale pour les déploiements d'inférence d'IA en périphérie.

Remerciements

- J.J. Falkanger Directeur des solutions HPC et d'IA, Lenovo
- Dave Arnette, Ingénieur marketing et technique, NetApp
- Joey Parnell, Tech Lead E-Series ai Solutions, NetApp
- Cody Harryman, ingénieur QA, NetApp

Où trouver des informations complémentaires

Pour en savoir plus sur les informations fournies dans ce document, consultez ces documents et/ou sites web :

- Page produit sur les baies NetApp AFF A-Series

  "https://www.netapp.com/data-storage/aff-a-series/"

- Le logiciel de gestion des données NetApp ONTAP – bibliothèque d'informations ONTAP 9

http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- Tr-4727 : introduction à la gamme EF-Series de NetApp

  https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf

- Logiciel SANtricity NetApp E-Series : Fiche technique

  https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf

- NetApp Trident pour le stockage persistant pour les conteneurs

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- Diminution des
  - "https://mlcommons.org/en/"
  - "http://www.image-net.org/"
  - "https://mlcommons.org/en/news/mlperf-inference-v11/"
- Copie et synchronisation NetApp BlueXP

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"

- Banc d'essai TensorFlow

  "https://github.com/tensorflow/benchmarks"

- Serveur Lenovo ThinkSystem SE350 Edge

  "https://lenovopress.com/lp1168"

- Baie de stockage Flash unifiée Lenovo ThinkSystem DM5100F

  "https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"

WP-7328 : la stratégie d'IA de NetApp basée sur NVIDIA Jarvis

Rick Huang, Sung-Han Lin, NetApp Davide Onofrio, NVIDIA

La gamme NVIDIA DGX comprend les premiers systèmes basés sur l'intelligence artificielle intégrée (IA) conçus pour l'IA d'entreprise. Les systèmes de stockage NetApp AFF combinent performances optimales et fonctionnalités de pointe pour la gestion des données dans le cloud hybride. NetApp et NVIDIA se sont associés pour créer l'architecture de référence NetApp ONTAP ai, une solution clé en main qui prend en charge les workloads d'IA et de machine learning (ML), tout en offrant des performances et une fiabilité exceptionnelles.

Ce livre blanc fournit des recommandations aux clients qui conçoivent des systèmes d'IA pour divers champs d'application dans différents secteurs d'activité. Elle inclut des informations sur le déploiement du système à l'aide de NVIDIA Jarvis. Les tests ont été réalisés avec une station NVIDIA DGX et un système de stockage NetApp AFF A220.

La solution cible plusieurs groupes d'utilisateurs :

- Les architectes d'entreprise qui conçoivent des solutions pour le développement de modèles et de logiciels d'IA dans le cadre de l'élaboration d'outils d'IA, tels que l'assistant de détail virtuel
- Les data Scientists à la recherche de moyens efficaces pour atteindre les objectifs de développement de la modélisation linguistique
- Les ingénieurs de données sont chargés de maintenir et de traiter les données texte telles que les questions des clients et les transcriptions des dialogues
- Dirigeants, décideurs IT et responsables qui veulent transformer les échanges d'expérience d'IA et accélérer le time-to-market des initiatives d'IA

Présentation de la solution

NetApp ONTAP ai et copie et synchronisation BlueXP

L'architecture NetApp ONTAP ai, optimisée par les systèmes NVIDIA DGX et les systèmes de stockage NetApp connectés au cloud, a été développée et vérifiée par NetApp et NVIDIA. Et présente plusieurs avantages pour les SERVICES IT :

- Simplifie la conception
- Évolutivité indépendante des ressources de calcul et de stockage
- Possibilité de faire évoluer de manière fluide une infrastructure initiale de petite taille
- Propose toute une gamme d'options de stockage pour répondre à différents besoins de performance et de pointsNetApp ONTAP ai intègre étroitement les systèmes DGX et de stockage NetApp AFF A220 avec une connectivité réseau optimale. Les systèmes NetApp ONTAP ai et DGX simplifient les déploiements d'IA en éliminant la complexité et les approximations. Les clients peuvent commencer avec un déploiement de petite taille, puis évoluer sans interruption d'activité, tout en gérant intelligemment leurs données de la périphérie au cœur, et jusqu'au cloud, et inversement.

La copie et la synchronisation NetApp BlueXP vous permettent de déplacer facilement des données selon divers protocoles, que ce soit entre deux partages NFS, deux partages CIFS ou un partage de fichiers et Amazon S3, Amazon Elastic File System (EFS) ou Azure Blob Storage. Avec le mode actif-actif, vous pouvez continuer à travailler avec la source et la cible en même temps, en synchronisant les modifications de données de manière incrémentielle lorsque nécessaire. Avec BlueXP Copy and Sync, vous pouvez déplacer et synchroniser de manière incrémentielle les données entre tous les systèmes source et de destination, que ce soit sur site ou dans le cloud. Il offre de nombreuses nouvelles façons d'utiliser les données. La migration des données entre des systèmes sur site, l'intégration dans le cloud et la migration vers le cloud, ou encore la collaboration et l'analytique, sont toutes des opérations qui deviennent facilement réalisables. La figure ci-dessous montre les sources et destinations disponibles.

Dans les systèmes d'IA conversationnels, les développeurs peuvent utiliser BlueXP Copy and Sync pour archiver l'historique des conversations du cloud aux data centers afin de permettre l'entraînement hors ligne de modèles de traitement du langage naturel (NLP). En formant des modèles permettant de reconnaître des intentions plus nombreuses, le système d'IA conversationnel sera mieux équipé pour gérer des questions plus complexes de la part des utilisateurs finaux.

Structure multimodale NVIDIA Jarvis

"NVIDIA Jarvis" Est un framework de bout en bout pour la création de services d'IA conversationnels. Elle comprend les services optimisés pour les processeurs graphiques suivants :

- Reconnaissance vocale automatique (ASR)
- Compréhension du langage naturel (NLU)
- Intégration avec des services d'exécution spécifiques à un domaine
- Texte à parole (TTS)
- Vision par ordinateur (CV)les services basés à Jarvis font appel à des modèles de deep learning sophistiqués pour gérer la tâche complexe et complexe qui consiste à gérer l'IA conversationnelle en temps réel. Pour permettre une interaction naturelle en temps réel avec un utilisateur final, les modèles doivent réaliser le calcul en moins de 300 millisecondes. Les interactions naturelles sont difficiles, ce qui nécessite une intégration sensorielle multimodale. Les pipelines modèles sont également complexes et nécessitent une coordination entre les services susmentionnés.

Jarvis est un framework d'application entièrement accéléré qui permet de créer des services d'IA multimodaux et d'exploiter un pipeline d'apprentissage profond de bout en bout. Le cadre de Jarvis inclut des modèles, des outils et des services d'IA pré-entraînés, optimisés pour la parole, la vision et les tâches de la NLU. Outre les services d'IA, Jarvis vous permet de fusionner simultanément la vision, le son et d'autres entrées de capteur afin de fournir des fonctionnalités telles que les conversations multi-utilisateurs et multi-contexte dans des applications telles que les assistants virtuels, la diarisation multi-utilisateurs et les assistants du centre d'appels.

NVIDIA Nemo

"NVIDIA Nemo" Est un kit Python open source permettant de créer, d'entraîner et d'affiner les modèles d'IA à accélération par processeur graphique, en utilisant des API (interfaces de programmation des applications) simples d'utilisation. Nemo exécute des ressources de calcul de précision mixte avec des cœurs Tensor dans les processeurs graphiques NVIDIA. Il peut facilement évoluer jusqu'à plusieurs GPU pour offrir les meilleures performances d'entraînement possibles. Nemo est utilisé pour concevoir des modèles pour les applications ASR, NLP et TTS en temps réel, telles que les transcriptions d'appels vidéo, les assistants vidéo intelligents et le support automatisé des centres d'appels dans différents secteurs industriels, y compris la santé, les

finances, le commerce de détail et les télécommunications.

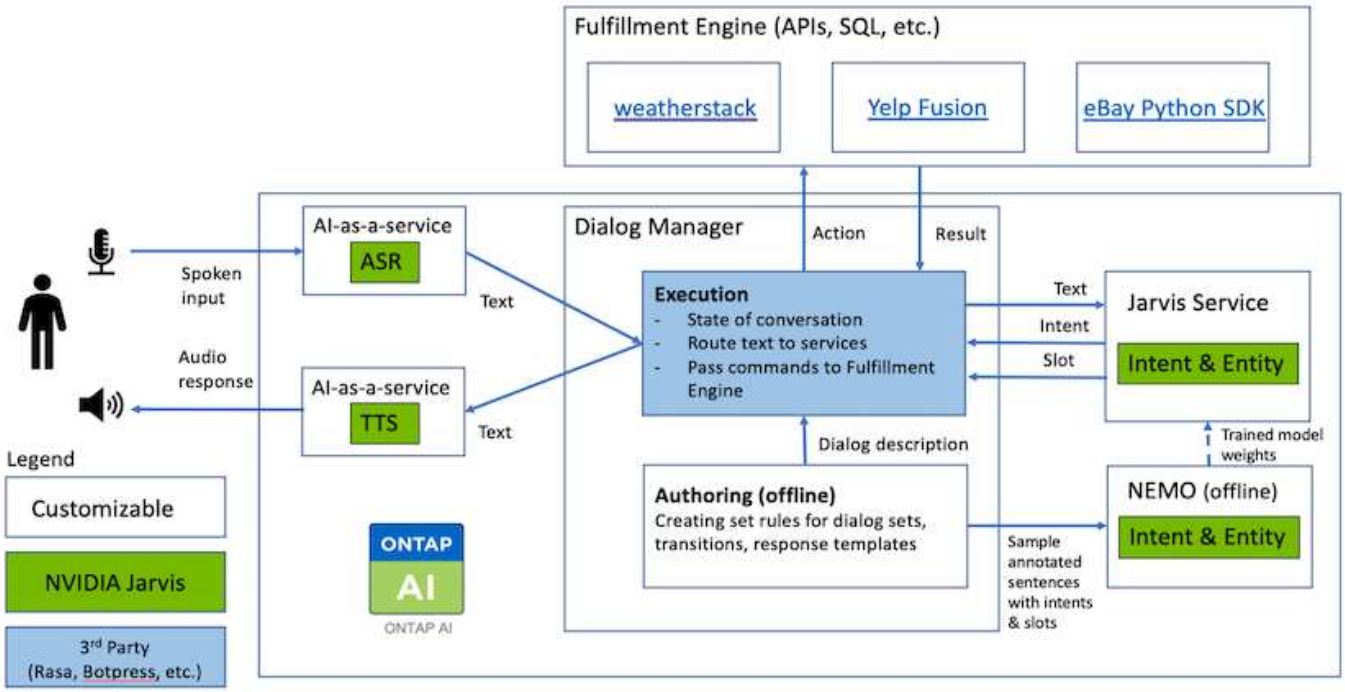
Nous avons utilisé Nemo pour former des modèles qui reconnaissent les intentions complexes des questions des utilisateurs dans l'historique des conversations archivées. Cette formation étend les capacités de l'assistant virtuel au détail au-delà de ce que Jarvis soutient comme prestation.

Sommaire de l'utilisation du commerce de détail

Avec NVIDIA Jarvis, nous avons construit un assistant de vente au détail virtuel qui accepte les commentaires vocaux ou texte et répond aux questions relatives à la météo, aux points d'intérêt et aux prix des stocks. Le système d'IA sur laquelle vous discutez peut mémoriser le flux de conversation, par exemple, posez une question de suivi si l'utilisateur n'a pas précisé le lieu par météo ou les points d'intérêt. Le système reconnaît également des entités complexes telles que « la nourriture thaïlandaise » ou « la mémoire des ordinateurs portables ». Il comprend des questions de langage naturel comme « la semaine prochaine à Los Angeles sera-t-elle pluviale ? » Vous trouverez une démonstration de l'assistant virtuel de vente au détail dans "Personnalisez les États et les flux pour le cas d'utilisation Vente au détail".

Technologie de la solution

La figure suivante illustre l'architecture du système d'IA proposé. Vous pouvez interagir avec le système avec le signal vocal ou la saisie de texte. Si l'entrée vocale est détectée, Jarvis ai-as-service (AIaaS) exécute ASR pour produire du texte pour Dialog Manager. Dialogue Manager mémorise les États de conversation, achemine le texte vers les services correspondants et transmet les commandes au moteur de traitement. Jarvis NLP Service prend du texte, reconnaît les intentions et les entités, et renvoie ces intentions et slots d'entité au Dialog Manager, qui envoie ensuite action au moteur de traitement. Le moteur d'exécution se compose d'API tierces ou de bases de données SQL qui répondent aux requêtes des utilisateurs. Après réception du résultat du moteur de traitement, le Gestionnaire de dialogue achemine le texte vers Jarvis TTS AIaaS afin de générer une réponse audio pour l'utilisateur final. Nous pouvons archiver l'historique des conversations, annoter des phrases avec des intentions et des slots pour la formation de Nemo, de sorte que le service NLP s'améliore à mesure que davantage d'utilisateurs interagissent avec le système.



Configuration matérielle requise

Cette solution a été validée à l'aide d'une station DGX et d'un système de stockage AFF A220. Jarvis requiert un processeur graphique T4 ou V100 pour effectuer des calculs profonds des réseaux neuronaux.

Le tableau suivant répertorie les composants matériels requis pour implémenter la solution testée.

Sous-jacent	Quantité
GPU T4 OU V100	1
Station NVIDIA DGX	1

Configuration logicielle requise

Le tableau suivant répertorie les composants logiciels requis pour implémenter la solution testée.

Logiciel	Version ou autres informations
Le logiciel de gestion des données NetApp ONTAP	9.6
Firmware du switch Cisco NX-OS	7.0(3)I6(1)
SYSTÈME D'EXPLOITATION NVIDIA DGX	4.0.4 - Ubuntu 18.04 LTS
Structure de NVIDIA Jarvis	EA v0.2
NVIDIA Nemo	nvcr.io/nvidia/nemo:v0.10
Plateforme de conteneurisation Docker	18.06.1 ce [e68fc7a]

Présentation

Cette section fournit des détails sur la mise en œuvre de l'assistant de vente au détail

virtuel.

Jarvis déploiement

Vous pouvez vous inscrire à "Jarvis accès précoce" Pour accéder aux conteneurs Jarvis sur NVIDIA GPU Cloud (NGC). Après avoir reçu les informations d'identification de NVIDIA, vous pouvez déployer Jarvis en procédant comme suit :

1. Connexion au contrôleur NGC.
2. Définissez votre organisation sur NGC : `ea-2-jarvis`.
3. Localiser Jarvis EA v0.2 biens: Les conteneurs Jarvis sont dans `Private Registry` > `Organization Containers`.
4. Sélectionnez Jarvis : accédez à `Model Scripts` et cliquez sur `Jarvis Quick Start`
5. Vérifiez que toutes les ressources fonctionnent correctement.
6. Trouvez la documentation pour créer vos propres applications : les PDF se trouvent dans `Model Scripts` > `Jarvis Documentation` > `File Browser`.

Personnalisez les États et les flux pour le cas d'utilisation Vente au détail

Vous pouvez personnaliser les États et les flux de Dialog Manager pour vos cas d'utilisation spécifiques. Dans notre exemple de vente au détail, nous avons les quatre fichiers yaml suivants pour diriger la conversation selon différentes intentions.

Se la liste suivante des noms de fichier et la description de chaque fichier :

- `main_flow.yml`: Définit les principaux flux et États de conversation et dirige le flux vers les trois autres fichiers yaml si nécessaire.
- `retail_flow.yml`: Contient des États liés à des questions de détail ou de points d'intérêt. Le système fournit soit les informations du magasin le plus proche, soit le prix d'un article donné.
- `weather_flow.yml`: Contient des États relatifs aux questions météorologiques. Si l'emplacement ne peut pas être déterminé, le système pose une question de suivi pour clarifier.
- `error_flow.yml`: Gère les cas où les intentions des utilisateurs ne tombent pas dans les trois fichiers yaml ci-dessus. Après l'affichage d'un message d'erreur, le système revient à accepter les questions de l'utilisateur. Les sections suivantes contiennent les définitions détaillées de ces fichiers yaml.

main_flow.yml

```
name: JarvisRetail
intent_transitions:
  jarvis_error: error
  price_check: retail_price_check
  inventory_check: retail_inventory_check
  store_location: retail_store_location
  weather.weather: weather
  weather.temperature: temperature
  weather.sunny: sunny
```

```
  weather.cloudy: cloudy
  weather.snow: snow
  weather.rainfall: rain
  weather.snow_yes_no: snowfall
  weather.rainfall_yes_no: rainfall
  weather.temperature_yes_no: tempyesno
  weather.humidity: humidity
  weather.humidity_yes_no: humidity
  navigation.startnavigationpoi: retail # Transitions should be context
and slot based. Redirecting for now.
  navigation.geteta: retail
  navigation.showdirection: retail
  navigation.showmappoi: idk_what_you_talkin_about
  nomatch.none: idk_what_you_talkin_about
states:
  init:
    type: message_text
    properties:
      text: "Hi, welcome to NARA retail and weather service. How can I
help you?"
  input_intent:
    type: input_context
    properties:
      nlp_type: jarvis
      entities:
        intent: dontcare
# This state is executed if the intent was not understood
  dont_get_the intent:
    type: message_text_random
    properties:
      responses:
        - "Sorry I didn't get that! Please come again."
        - "I beg your pardon! Say that again?"
        - "Are we talking about weather? What would you like to know?"
        - "Sorry I know only about the weather"
        - "You can ask me about the weather, the rainfall, the
temperature, I don't know much more"
      delay: 0
    transitions:
      next_state: input_intent
  idk_what_you_talkin_about:
    type: message_text_random
    properties:
      responses:
        - "Sorry I didn't get that! Please come again."
        - "I beg your pardon! Say that again?"
```

```
        - "Are we talking about retail or weather? What would you like to
know?"
        - "Sorry I know only about retail and the weather"
        - "You can ask me about retail information or the weather, the
rainfall, the temperature. I don't know much more."
      delay: 0
    transitions:
      next_state: input_intent
  error:
    type: change_context
    properties:
        update_keys:
           intent: 'error'
    transitions:
        flow: error_flow
  retail_inventory_check:
    type: change_context
    properties:
        update_keys:
           intent: 'retail_inventory_check'
    transitions:
        flow: retail_flow
  retail_price_check:
    type: change_context
    properties:
        update_keys:
           intent: 'check_item_price'
    transitions:
        flow: retail_flow
  retail_store_location:
    type: change_context
    properties:
        update_keys:
           intent: 'find_the_store'
    transitions:
        flow: retail_flow
  weather:
    type: change_context
    properties:
        update_keys:
           intent: 'weather'
    transitions:
        flow: weather_flow
  temperature:
    type: change_context
    properties:
```

```
        update_keys:
           intent: 'temperature'
    transitions:
        flow: weather_flow
  rainfall:
    type: change_context
    properties:
        update_keys:
           intent: 'rainfall'
    transitions:
        flow: weather_flow
  sunny:
    type: change_context
    properties:
        update_keys:
           intent: 'sunny'
    transitions:
        flow: weather_flow
  cloudy:
    type: change_context
    properties:
        update_keys:
           intent: 'cloudy'
    transitions:
        flow: weather_flow
  snow:
    type: change_context
    properties:
        update_keys:
           intent: 'snow'
    transitions:
        flow: weather_flow
  rain:
    type: change_context
    properties:
        update_keys:
           intent: 'rain'
    transitions:
        flow: weather_flow
  snowfall:
      type: change_context
      properties:
          update_keys:
             intent: 'snowfall'
      transitions:
          flow: weather_flow
```

```
  tempyesno:
      type: change_context
      properties:
          update_keys:
             intent: 'tempyesno'
      transitions:
          flow: weather_flow
  humidity:
      type: change_context
      properties:
          update_keys:
             intent: 'humidity'
      transitions:
          flow: weather_flow
  end_state:
    type: reset
    transitions:
      next_state: init
```

retail_flow.yml

```
name: retail_flow
states:
  store_location:
    type: conditional_exists
    properties:
      key: '{{location}}'
    transitions:
      exists: retail_state
      notexists: ask_retail_location
  retail_state:
    type: Retail
    properties:
    transitions:
      next_state: output_retail
  output_retail:
      type: message_text
      properties:
        text: '{{retail_status}}'
      transitions:
        next_state: input_intent
  ask_retail_location:
    type: message_text
    properties:
      text: "For which location? I can find the closest store near you."
```

```
    transitions:
      next_state: input_retail_location
  input_retail_location:
    type: input_user
    properties:
      nlp_type: jarvis
      entities:
        slot: location
      require_match: true
    transitions:
      match: retail_state
      notmatch: check_retail_jarvis_error
  output_retail_acknowledge:
    type: message_text_random
    properties:
      responses:
        - 'ok in {{location}}'
        - 'the store in {{location}}'
        - 'I always wanted to shop in {{location}}'
      delay: 0
    transitions:
      next_state: retail_state
  output_retail_notlocation:
    type: message_text
    properties:
      text: "I did not understand the location. Can you please repeat?"
    transitions:
      next_state: input_intent
  check_rerail_jarvis_error:
    type: conditional_exists
    properties:
      key: '{{jarvis_error}}'
    transitions:
      exists: show_retail_jarvis_api_error
      notexists: output_retail_notlocation
  show_retail_jarvis_api_error:
    type: message_text
    properties:
      text: "I am having troubled understanding right now. Come again on
that?"
    transitions:
      next_state: input_intent
```

météo_flow.yml

```
```

```
name: weather_flow
states:
  check_weather_location:
    type: conditional_exists
    properties:
      key: '{{location}}'
    transitions:
      exists: weather_state
      notexists: ask_weather_location
  weather_state:
    type: Weather
    properties:
    transitions:
      next_state: output_weather
  output_weather:
      type: message_text
      properties:
        text: '{{weather_status}}'
      transitions:
        next_state: input_intent
  ask_weather_location:
    type: message_text
    properties:
      text: "For which location?"
    transitions:
      next_state: input_weather_location
  input_weather_location:
    type: input_user
    properties:
      nlp_type: jarvis
      entities:
        slot: location
      require_match: true
    transitions:
      match: weather_state
      notmatch: check_jarvis_error
  output_weather_acknowledge:
    type: message_text_random
    properties:
      responses:
        - 'ok in {{location}}'
        - 'the weather in {{location}}'
        - 'I always wanted to go in {{location}}'
      delay: 0
    transitions:
      next_state: weather_state
```

```
  output_weather_notlocation:
    type: message_text
    properties:
      text: "I did not understand the location, can you please repeat?"
    transitions:
      next_state: input_intent
  check_jarvis_error:
    type: conditional_exists
    properties:
      key: '{{jarvis_error}}'
    transitions:
      exists: show_jarvis_api_error
      notexists: output_weather_notlocation
  show_jarvis_api_error:
    type: message_text
    properties:
      text: "I am having troubled understanding right now. Come again on
that, else check jarvis services?"
    transitions:
      next_state: input_intent
```

error_flow.yml

```
name: error_flow
states:
  error_state:
    type: message_text_random
    properties:
      responses:
        - "Sorry I didn't get that!"
        - "Are we talking about retail or weather? What would you like to
know?"
        - "Sorry I know only about retail information or the weather"
        - "You can ask me about retail information or the weather, the
rainfall, the temperature. I don't know much more"
        - "Let's talk about retail or the weather!"
      delay: 0
    transitions:
      next_state: input_intent
```

Connectez-vous à des API tierces en tant que moteur de traitement

Nous avons connecté les API tierces suivantes en tant que moteur de traitement pour répondre aux questions :

- "API WeatherStack": renvoie la météo, la température, la pluie et la neige dans un endroit donné.
- "API Yelp Fusion": renvoie les informations de magasin les plus proches dans un emplacement donné.
- "Kit de développement logiciel eBay Python": renvoie le prix d'un article donné.

Démonstration de NetApp Retail Assistant

Nous avons enregistré une vidéo de démonstration de l'Assistant vente au détail NetApp (NARA).

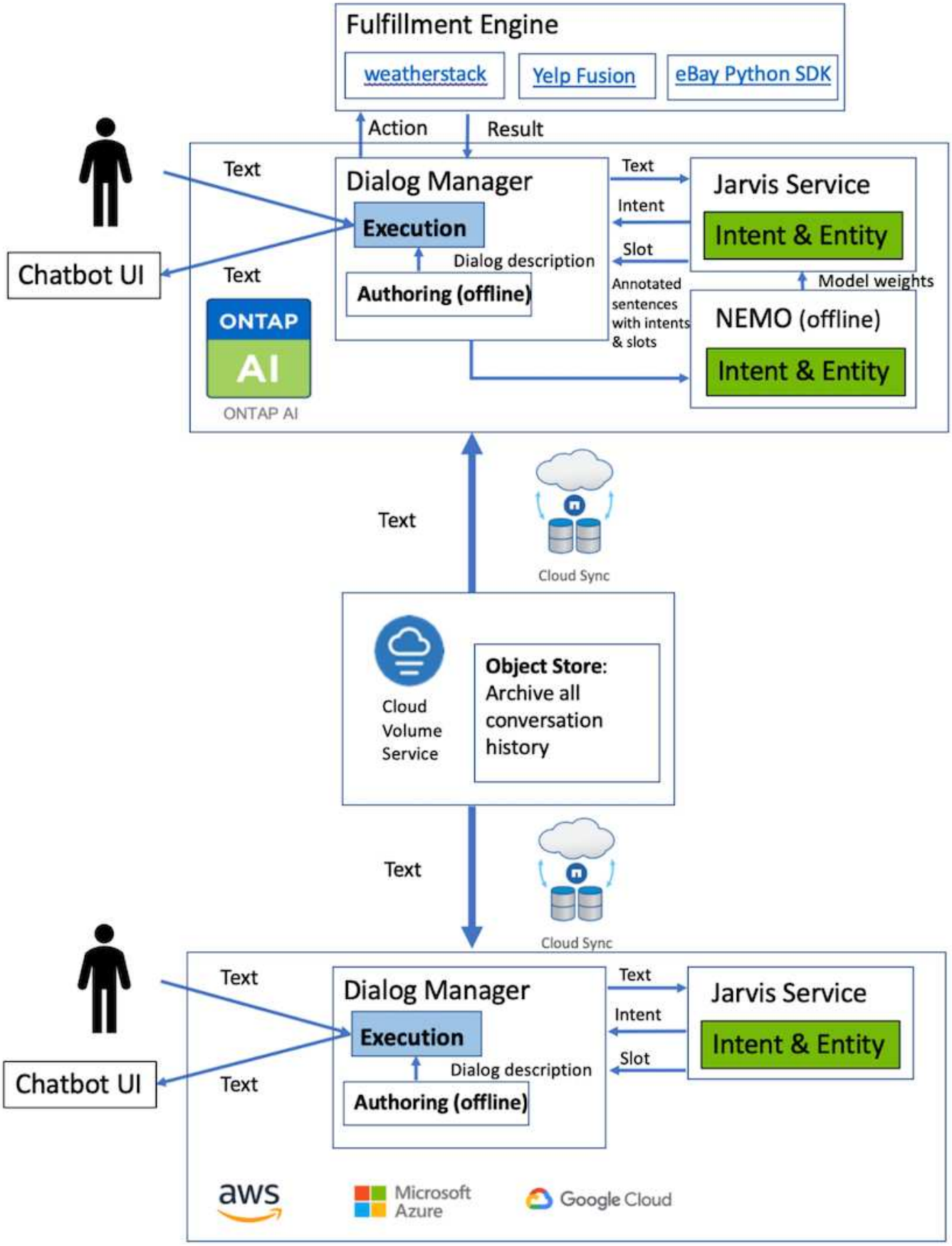
Vidéo de démonstration de NARA

Vidéo de démonstration de NARA

Utilisez la copie et la synchronisation NetApp BlueXP pour archiver l'historique des conversations

En vidant une fois par jour l'historique des conversations dans un fichier CSV, nous pouvons ensuite utiliser BlueXP Copy and Sync pour télécharger les fichiers journaux dans un système de stockage local. La figure suivante montre l'architecture du déploiement de Jarvis sur site et dans les clouds publics, tout en utilisant BlueXP Copy

and Sync pour envoyer l’historique des conversations pour la formation Nemo. Les détails de la formation de Nemo sont disponibles dans la section "Développez les modèles d’intention à l’aide de la formation Nemo".


Fulfillment Engine
weatherstack
Yelp Fusion
eBay Python SDK
Action
Result
Text
Chatbot UI
Text
ONTAP
AI
ONTAP AI
Dialog Manager
Execution
Dialog description
Authoring (offline)
Text
Intent
Slot
Annotated sentences with intents & slots
Jarvis Service
Intent & Entity
Model weights
NEMO (offline)
Intent & Entity
Text
Cloud Sync
Cloud Volume Service
Object Store: Archive all conversation history
Text
Cloud Sync
Text
Chatbot UI
Text
Dialog Manager
Execution
Dialog description
Authoring (offline)
Text
Intent
Slot
Jarvis Service
Intent & Entity
aws
Microsoft Azure
Google Cloud

Développez les modèles d'intention à l'aide de la formation Nemo

NVIDIA Nemo est un kit conçu par NVIDIA pour créer des applications d'IA conversationnelles. Ce kit comprend des ensembles de modules pré-entraînés pour ASR, NLP et TTS, ce qui permet aux chercheurs et aux scientifiques des données de composer facilement des architectures de réseaux neuronaux complexes et de se concentrer davantage sur la conception de leurs propres applications.

Comme le montre l'exemple précédent, NARA ne peut traiter qu'un type limité de question. En effet, le modèle NLP pré-formé ne s'entraîne que sur ces types de questions. Si nous voulons permettre À NARA de gérer un plus large éventail de questions, nous devons le réentraîner avec nos propres jeux de données. Ainsi, ici, nous démontrons comment nous pouvons utiliser Nemo pour étendre le modèle NLP pour satisfaire les exigences. Nous commençons par convertir le journal collecté à partir DE NARA dans le format pour Nemo, puis nous entraînons avec le dataset pour améliorer le modèle NLP.

Modèle

Notre objectif est de permettre À NARA de trier les éléments en fonction des préférences de l'utilisateur. Par exemple, nous pourrions demander À NARA de proposer le restaurant de sushis le mieux noté ou pourrait vouloir NARA chercher les jeans avec le prix le plus bas. À cette fin, nous utilisons le modèle de détection d'intention et de remplissage de fente fourni dans Nemo comme modèle d'entraînement. Ce modèle permet À NARA de comprendre l'intention de la préférence de recherche.

Préparation des données

Pour entraîner le modèle, nous collectons l'ensemble de données pour ce type de question et le convertissons au format Nemo. Ici, nous avons répertorié les fichiers que nous utilisons pour entraîner le modèle.

dict.intents.csv

Ce fichier liste tous les éléments que nous voulons que le Nemo comprenne. Ici, nous avons deux intentions principales et une intention seulement utilisée pour classer les questions qui ne correspondent à aucune des intentions principales.

```
price_check
find_the_store
unknown
```

dict.slots.csv

Ce fichier répertorie tous les emplacements que nous pouvons étiqueter sur nos questions de formation.

```
B-store.type
B-store.name
B-store.status
B-store.hour.start
B-store.hour.end
B-store.hour.day
B-item.type
```

```
B-item.name
B-item.color
B-item.size
B-item.quantity
B-location
B-cost.high
B-cost.average
B-cost.low
B-time.period_of_time
B-rating.high
B-rating.average
B-rating.low
B-interrogative.location
B-interrogative.manner
B-interrogative.time
B-interrogative.personal
B-interrogative
B-verb
B-article
I-store.type
I-store.name
I-store.status
I-store.hour.start
I-store.hour.end
I-store.hour.day
I-item.type
I-item.name
I-item.color
I-item.size
I-item.quantity
I-location
I-cost.high
I-cost.average
I-cost.low
I-time.period_of_time
I-rating.high
I-rating.average
I-rating.low
I-interrogative.location
I-interrogative.manner
I-interrogative.time
I-interrogative.personal
I-interrogative
I-verb
I-article
O
```

train.tsv

Il s'agit du dataset d'entraînement principal. Chaque ligne commence par la question qui suit la liste des catégories d'intention dans le fichier dict.intent.csv. L'étiquette est énumérée à partir de zéro.

train_slots.tsv

```
20 46 24 25 6 32 6
52 52 24 6
23 52 14 40 52 25 6 32 6
…
```

Entraîner le modèle

```
docker pull nvcr.io/nvidia/nemo:v0.10
```

Nous utilisons ensuite la commande suivante pour lancer le conteneur. Dans cette commande, nous limitons le conteneur à utiliser un seul GPU (ID de processeur graphique = 1), car il s'agit d'un exercice d'entraînement léger. Nous mappons également notre espace de travail local /Workspace/nemo/ vers le dossier à l'intérieur du conteneur /nemo.

```
NV_GPU='1' docker run --runtime=nvidia -it --shm-size=16g \
                        --network=host --ulimit memlock=-1 --ulimit
stack=67108864 \
                        -v /workspace/nemo:/nemo\
                        --rm nvcr.io/nvidia/nemo:v0.10
```

Dans le conteneur, si nous voulons commencer par le modèle original de BERT pré-formé, nous pouvons utiliser la commande suivante pour démarrer la procédure de formation. data_dir est l'argument pour définir le chemin des données d'entraînement. work_dir vous permet de configurer l'emplacement où vous souhaitez stocker les fichiers de point de contrôle.

```
cd examples/nlp/intent_detection_slot_tagging/
python joint_intent_slot_with_bert.py \
    --data_dir /nemo/training_data\
    --work_dir /nemo/log
```

Si nous avons de nouveaux datasets d'entraînement et que nous souhaitons améliorer le modèle précédent, nous pouvons utiliser la commande suivante pour continuer à partir du point que nous avons arrêté. checkpoint_dir indique le chemin d'accès au dossier points de contrôle précédent.

```
cd examples/nlp/intent_detection_slot_tagging/
python joint_intent_slot_infer.py \
    --data_dir /nemo/training_data \
    --checkpoint_dir /nemo/log/2020-05-04_18-34-20/checkpoints/ \
    --eval_file_prefix test
```

Inférence du modèle

Nous devons valider la performance du modèle entraîné après un certain nombre de tests. La commande suivante nous permet de tester la requête un par un. Par exemple, dans cette commande, nous voulons vérifier si notre modèle peut correctement identifier l'intention de la requête `where can I get the best pasta`.

```
cd examples/nlp/intent_detection_slot_tagging/
python joint_intent_slot_infer_b1.py \
--checkpoint_dir /nemo/log/2020-05-29_23-50-58/checkpoints/ \
--query "where can i get the best pasta" \
--data_dir /nemo/training_data/ \
--num_epochs=50
```

Ensuite, le résultat suivant est le résultat de l'inférence. Dans ce résultat, nous pouvons constater que notre modèle entraîné peut prévoir correctement l'intention Find_the_store et renvoyer les mots-clés qui nous intéressent. Avec ces mots-clés, nous permettons à NARA de rechercher ce que les utilisateurs veulent et de faire une recherche plus précise.

```
[NeMo I 2020-05-30 00:06:54 actions:728] Evaluating batch 0 out of 1
[NeMo I 2020-05-30 00:06:55 inference_utils:34] Query: where can i get the
best pasta
[NeMo I 2020-05-30 00:06:55 inference_utils:36] Predicted intent:       1
find_the_store
[NeMo I 2020-05-30 00:06:55 inference_utils:50] where   B-
interrogative.location
[NeMo I 2020-05-30 00:06:55 inference_utils:50] can     O
[NeMo I 2020-05-30 00:06:55 inference_utils:50] i       O
[NeMo I 2020-05-30 00:06:55 inference_utils:50] get     B-verb
[NeMo I 2020-05-30 00:06:55 inference_utils:50] the     B-article
[NeMo I 2020-05-30 00:06:55 inference_utils:50] best    B-rating.high
[NeMo I 2020-05-30 00:06:55 inference_utils:50] pasta   B-item.type
```

Conclusion

Un véritable système d'IA et d'échange s'engage dans un dialogue similaire à celui de l'homme, comprend le contexte et propose des réponses intelligentes. Ces modèles d'IA sont souvent énormes et extrêmement complexes. Avec les processeurs graphiques

NVIDIA et le stockage NetApp, des modèles linguistiques de pointe peuvent être entraînés et optimisés pour exécuter rapidement l'inférence. C'est une tendance majeure vers la fin du compromis entre un modèle d'IA qui est rapide contre un modèle important et complexe. Les modèles de compréhension du langage optimisés par les GPU peuvent être intégrés aux applications d'IA dans des secteurs tels que la santé, le commerce de détail et les services financiers, permettant ainsi aux assistants numériques avancés dans les enceintes intelligentes et les lignes de services client. Ces systèmes d'IA conversationnels de haute qualité permettent aux entreprises de tous les marchés verticaux de fournir des services personnalisés jusqu'alors inaccessibles lors de l'entretien avec les clients.

Jarvis permet le déploiement de cas d'utilisation tels que des assistants virtuels, des avatars numériques, des Fusion de capteur multimodal (CV fusionné avec ASR/NLP/TTS), ou tout cas d'utilisation autonome ASR/NLP/TTS/CV, comme la transcription. Nous avons créé un assistant de vente au détail virtuel qui répond à vos questions concernant la météo, les points d'intérêt et les prix des stocks. Nous avons également démontré comment améliorer les fonctionnalités de compréhension du langage naturel du système d'IA conversationnel en archivant l'historique des conversations à l'aide de BlueXP Copy and Sync et en formant les modèles Nemo aux nouvelles données.

Remerciements

Les auteurs reconnaissent les contributions qui ont été faites à ce livre blanc par nos collègues estimés de NVIDIA : Davide Onofrio, Alex Qi, Sitong Ji, Marty Jain et Robert Sohigian. Les auteurs apprécient également les contributions des principaux membres de l'équipe NetApp : Santosh Rao, David Arnette, Michael Oglesby, Brent Davis, Andy Sayare, Erik Mulder et Mike McNamara.

Nous tenons à remercier sincèrement tous ces contributeurs dont les connaissances et l'expertise ont étayé l'élaboration de ce document.

Sources d'informations complémentaires

Pour en savoir plus sur les informations données dans ce document, consultez les ressources suivantes :

- NVIDIA DGX Station, V100 GPU, cloud GPU
  - Station NVIDIA DGXhttps://www.nvidia.com/en-us/data-center/dgx-station/["https://www.nvidia.com/en-us/data-center/dgx-station/"^]
  - GPU NVIDIA V100 à cœurs Tensorhttps://www.nvidia.com/en-us/data-center/tesla-v100/["https://www.nvidia.com/en-us/data-center/tesla-v100/"^]
  - NVIDIA NGChttps://www.nvidia.com/en-us/gpu-cloud/["https://www.nvidia.com/en-us/gpu-cloud/"^]
- Structure multimodale NVIDIA Jarvis
  - NVIDIA Jarvishttps://developer.nvidia.com/nvidia-jarvis["https://developer.nvidia.com/nvidia-jarvis"^]
  - NVIDIA Jarvis accès précocehttps://developer.nvidia.com/nvidia-jarvis-early-access["https://developer.nvidia.com/nvidia-jarvis-early-access"^]
- NVIDIA Nemo

    - NVIDIA Nemohttps://developer.nvidia.com/nvidia-nemo["https://developer.nvidia.com/nvidia-nemo"^]
    - Guide du développeurhttps://nvidia.github.io/NeMo/["https://nvidia.github.io/NeMo/"^]
- Systèmes NetApp AFF
    - Fiche technique NetApp AFF A-Serieshttps://www.netapp.com/us/media/ds-3582.pdf["https://www.netapp.com/us/media/ds-3582.pdf"^]
    - Avantages du Flash NetApp pour les systèmes FAS 100 % Flashhttps://www.netapp.com/us/media/ds-3733.pdf["https://www.netapp.com/us/media/ds-3733.pdf"^]
    - Bibliothèque d'informations ONTAP 9http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286["http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286"^]
    - Rapport technique sur NetApp ONTAP FlexGroup volumeshttps://www.netapp.com/us/media/tr-4557.pdf["https://www.netapp.com/us/media/tr-4557.pdf"^]
- NetApp ONTAP ai
    - Guide de conception d'ONTAP ai avec DGX-1 et connectivité réseau Ciscohttps://www.netapp.com/us/media/nva-1121-design.pdf["https://www.netapp.com/us/media/nva-1121-design.pdf"^]
    - Guide de déploiement d'ONTAP ai avec DGX-1 et réseau Ciscohttps://www.netapp.com/us/media/nva-1121-deploy.pdf["https://www.netapp.com/us/media/nva-1121-deploy.pdf"^]
    - Guide de conception d'ONTAP ai avec DGX-1 et réseau Mellanoxhttp://www.netapp.com/us/media/nva-1138-design.pdf["http://www.netapp.com/us/media/nva-1138-design.pdf"^]
    - Guide de conception de ONTAP ai avec DGX-2https://www.netapp.com/us/media/nva-1135-design.pdf["https://www.netapp.com/us/media/nva-1135-design.pdf"^]

Tr-4858 : solution d'orchestration NetApp avec Run:ai

Rick Huang, David Arnette, Sung-Han Lin, NetApp Yaron Goldberg, Run:ai

Les systèmes de stockage NetApp AFF combinent performances optimales et fonctionnalités de pointe pour la gestion des données dans le cloud hybride. NetApp et Run:ai se sont associés pour présenter les fonctionnalités uniques de la solution NetApp ONTAP ai pour les workloads d'intelligence artificielle (IA) et de machine learning (ML), qui offre des performances, une fiabilité et un support haute performance. Exécution :l'orchestration ai des workloads d'IA ajoute une plateforme Kubernetes de planification et d'utilisation des ressources afin d'aider les chercheurs à gérer et à optimiser l'utilisation des GPU. Associée aux systèmes NVIDIA DGX, la solution combinée de NetApp, NVIDIA et Run:ai fournit une pile d'infrastructure dédiée aux workloads d'IA d'entreprise. Ce rapport technique fournit des recommandations aux clients qui conçoivent des systèmes d'IA pour divers champs d'application et dans plusieurs secteurs d'activité. Vous y trouverez des informations sur le déploiement d'Run:ai et un système de stockage NetApp AFF A800, et sert d'architecture de référence pour déployer rapidement des initiatives d'IA de manière simple.

La solution cible plusieurs groupes d'utilisateurs :

- Les architectes d'entreprise qui conçoivent des solutions pour le développement de modèles et de logiciels d'IA pour des cas d'utilisation basés sur Kubernetes, tels que les microservices conteneurisés
- Data Scientists qui recherchent des méthodes efficaces pour atteindre leurs objectifs de développement de modèles dans un environnement en cluster comprenant plusieurs équipes et projets
- Les ingénieurs de données sont chargés de maintenir et d'exécuter des modèles de production
- Dirigeants, décideurs IT et responsables qui veulent créer une expérience optimale en matière d'utilisation des ressources des clusters Kubernetes et accélérer le time-to-market des initiatives d'IA

Présentation de la solution

NetApp ONTAP ai et la solution ai Control plane

L'architecture NetApp ONTAP ai, développée et vérifiée par NetApp et NVIDIA, est optimisée par les systèmes NVIDIA DGX et les systèmes de stockage NetApp connectés au cloud. Et présente plusieurs avantages pour les SERVICES IT :

- Simplifie la conception
- Évolutivité indépendante des ressources de calcul et de stockage
- Possibilité de faire évoluer de manière fluide une infrastructure initiale de petite taille
- Propose plusieurs options de stockage pour répondre à des exigences variées de coûts et de performance

NetApp ONTAP ai intègre étroitement les systèmes DGX et de stockage NetApp AFF A800 avec une connectivité réseau optimale. Les systèmes NetApp ONTAP ai et DGX simplifient les déploiements d'IA en éliminant la complexité et les approximations. Les clients peuvent commencer avec un déploiement de petite taille, puis évoluer sans interruption d'activité, tout en gérant intelligemment leurs données de la périphérie au cœur, et jusqu'au cloud, et inversement.

NetApp ai Control plane est une solution complète d'IA, DE MACHINE LEARNING et de deep learning (DL) et de gestion des tests pour les data Scientists et les ingénieurs de données. Alors que les entreprises ont de plus en plus recours à l'IA, elles sont confrontées à de nombreux défis, notamment l'évolutivité des workloads et la disponibilité des données. NetApp ai Control plane répond à ces challenges grâce à des fonctionnalités telles que le clonage rapide d'un namespace de données comme Git repo. Il définit et met en œuvre des workflows d'entraînement d'IA qui incluent la création quasi instantanée de données et de références de modèles pour la traçabilité et la gestion des versions. NetApp ai Control plane vous permet de répliquer de manière transparente des données entre plusieurs sites et régions, et de provisionner rapidement des espaces de travail Jupyter Notebook avec un accès à des datasets volumineux.

Exécutez :plateforme d'IA pour l'orchestration des workloads d'IA

Run :l'IA a conçu la première plateforme mondiale d'orchestration et de virtualisation pour les infrastructures d'IA. En retirant les charges de travail du matériel sous-jacent, Run:ai crée un pool partagé de ressources GPU qui peut être provisionné de manière dynamique. Il est ainsi possible d'orchestrer efficacement les workloads d'IA et d'optimiser l'utilisation des GPU. Les data Scientists peuvent utiliser de façon transparente des quantités massives de puissance GPU pour améliorer et accélérer leurs recherches. Les équipes IT conservent un contrôle inter-site centralisé et une visibilité en temps réel sur le provisionnement des ressources, la mise en file d'attente et l'utilisation. La plateforme Run:IA repose sur Kubernetes, pour une intégration simple avec les workflows IT et de data science.

La plateforme Run:ai offre plusieurs avantages :

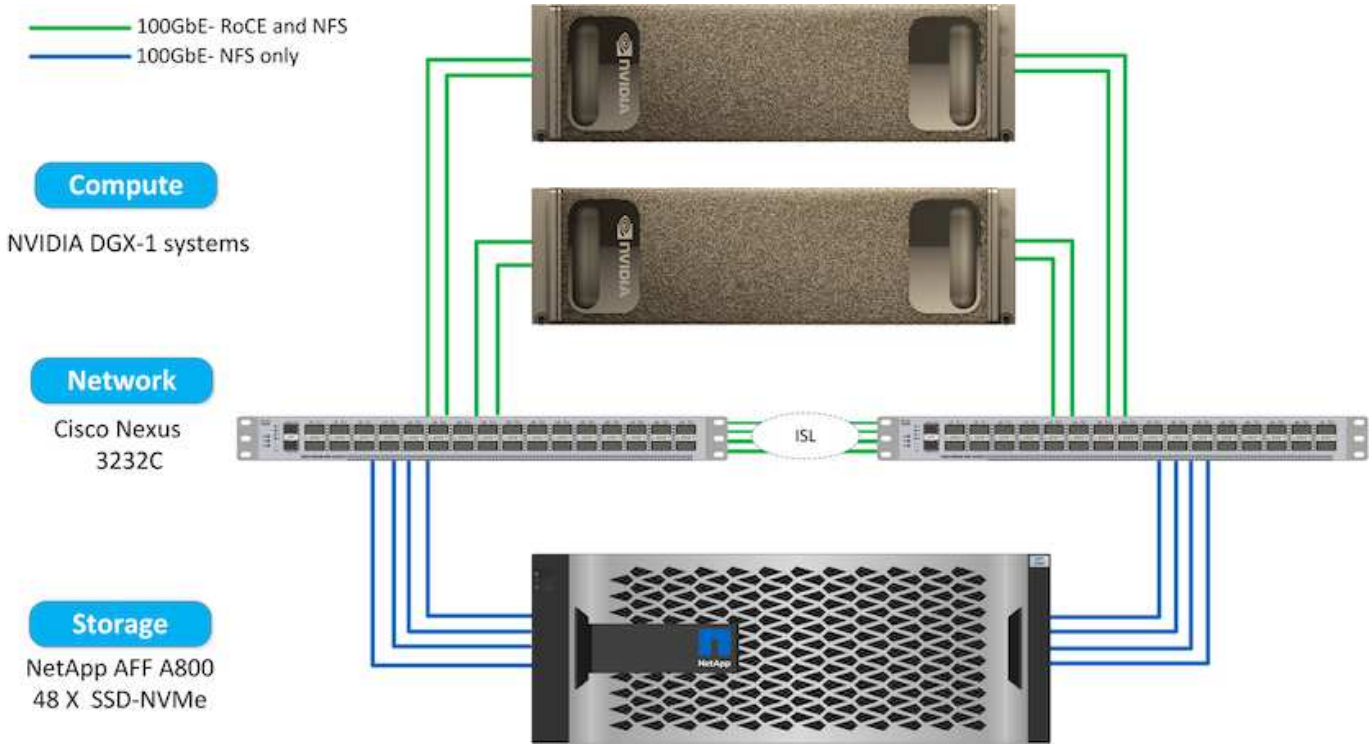
- **Accélération du temps vers l'innovation** en utilisant le pool de ressources Run:ai, la mise en file d'attente et les mécanismes de priorisation avec un système de stockage NetApp, les chercheurs ne sont

plus aux problèmes de gestion de l'infrastructure et peuvent se concentrer exclusivement sur la science des données. Exécution :les clients qui possèdent des solutions d'IA et NetApp augmentent la productivité en exécutant autant de workloads que nécessaire, sans goulot d'étranglement au niveau du calcul ou du pipeline de données.

- **Productivité accrue de l'équipe.** les algorithmes Run:ai Equiéquité garantissent que tous les utilisateurs et équipes obtiennent leur juste part de ressources. Les stratégies relatives aux projets prioritaires peuvent être prédéfinies et la plateforme permet une allocation dynamique des ressources d'un utilisateur ou d'une équipe à l'autre, ce qui permet aux utilisateurs d'accéder rapidement aux ressources GPU convoitées.
- **Amélioration de l'utilisation des GPU.** le planificateur Run:ai permet aux utilisateurs d'utiliser facilement des GPU fractionnaires, des GPU entiers et plusieurs nœuds de GPU pour l'entraînement distribué sur Kubernetes. De cette façon, les workloads d'IA s'exécutent en fonction de vos besoins, pas de la capacité. Les équipes de data science sont en mesure d'exécuter davantage d'expériences d'IA sur la même infrastructure.

Technologie de la solution

Cette solution a été implémentée avec un système NetApp AFF A800, deux serveurs DGX-1 et deux switchs Cisco Nexus 3232C de 100 GbE. Chaque serveur DGX-1 est relié aux switchs Nexus par des liaisons de 100 GbE utilisées pour les communications entre les GPU via le protocole RoCE (RDMA over Converged Ethernet). Les communications IP classiques pour l'accès au stockage NFS s'effectuent également sur ces liaisons. Chaque contrôleur de stockage est relié aux switchs réseau par quatre liaisons de 100 GbE. La figure suivante montre l'architecture de la solution ONTAP ai utilisée dans ce rapport technique pour tous les scénarios de test.



Matériel utilisé dans cette solution

Cette solution a été validée à l'aide de l'architecture de référence ONTAP ai, deux nœuds DGX-1 et un système de stockage AFF A800. Voir "NVA-1121" pour plus de détails sur l'infrastructure utilisée dans cette

validation.

Le tableau suivant répertorie les composants matériels requis pour implémenter la solution testée.

Sous-jacent	Quantité
Systèmes DGX-1	2
AFF A800	1
Switchs Nexus 3232C	2

Configuration logicielle requise

Cette solution a été validée à l'aide d'un déploiement Kubernetes de base avec l'opérateur Run:ai installé. Kubernetes a été déployé à l'aide du "NVIDIA DeepOps" le moteur de déploiement déploie tous les composants nécessaires pour un environnement prêt à la production. Déploiement automatique de DeepOps "NetApp Trident" Pour l'intégration du stockage persistant avec l'environnement k8s et des classes de stockage par défaut ont été créées afin que les conteneurs exploitent le stockage du système de stockage AFF A800. Pour en savoir plus sur Trident avec Kubernetes sur ONTAP ai, rendez-vous sur "TR-4798".

Le tableau suivant répertorie les composants logiciels requis pour implémenter la solution testée.

Logiciel	Version ou autres informations
Le logiciel de gestion des données NetApp ONTAP	9.6p4
Firmware du switch Cisco NX-OS	7.0(3)I6(1)
SYSTÈME D'EXPLOITATION NVIDIA DGX	4.0.4 - Ubuntu 18.04 LTS
Version Kubernetes	1.17
Version Trident	20.04.0
Exécutez :CLI ai	v2.1.13
Exécution : version de l'opérateur Kubernetes d'orchestration d'IA	1.0.39
Plateforme de conteneurisation Docker	18.06.1 ce [e68fc7a]

D'autres exigences logicielles supplémentaires pour Run:ai sont disponibles à l'adresse "Pré-requis pour le cluster de GPU Run:ai".

Optimisation de l'utilisation des clusters et des GPU avec Run:ai

Les sections suivantes décrivent l'installation de Run:ai, les scénarios de test et les résultats obtenus lors de cette validation.

Le fonctionnement et les performances de ce système ont été validées à l'aide d'outils de test standard du secteur, dont des bancs d'essai TensorFlow. Le dataset ImageNet a été utilisé pour entraîner le modèle ResNet-50, qui est un modèle d'apprentissage profond connu des réseaux neuronaux convolutifs (CNN) pour la classification d'images. RESNET-50 offre des résultats d'entraînement précis avec un temps de traitement plus rapide, ce qui nous a permis d'obtenir une demande suffisante en matière de stockage.

Exécutez : installation ai

Pour installer Run:ai, procédez comme suit :

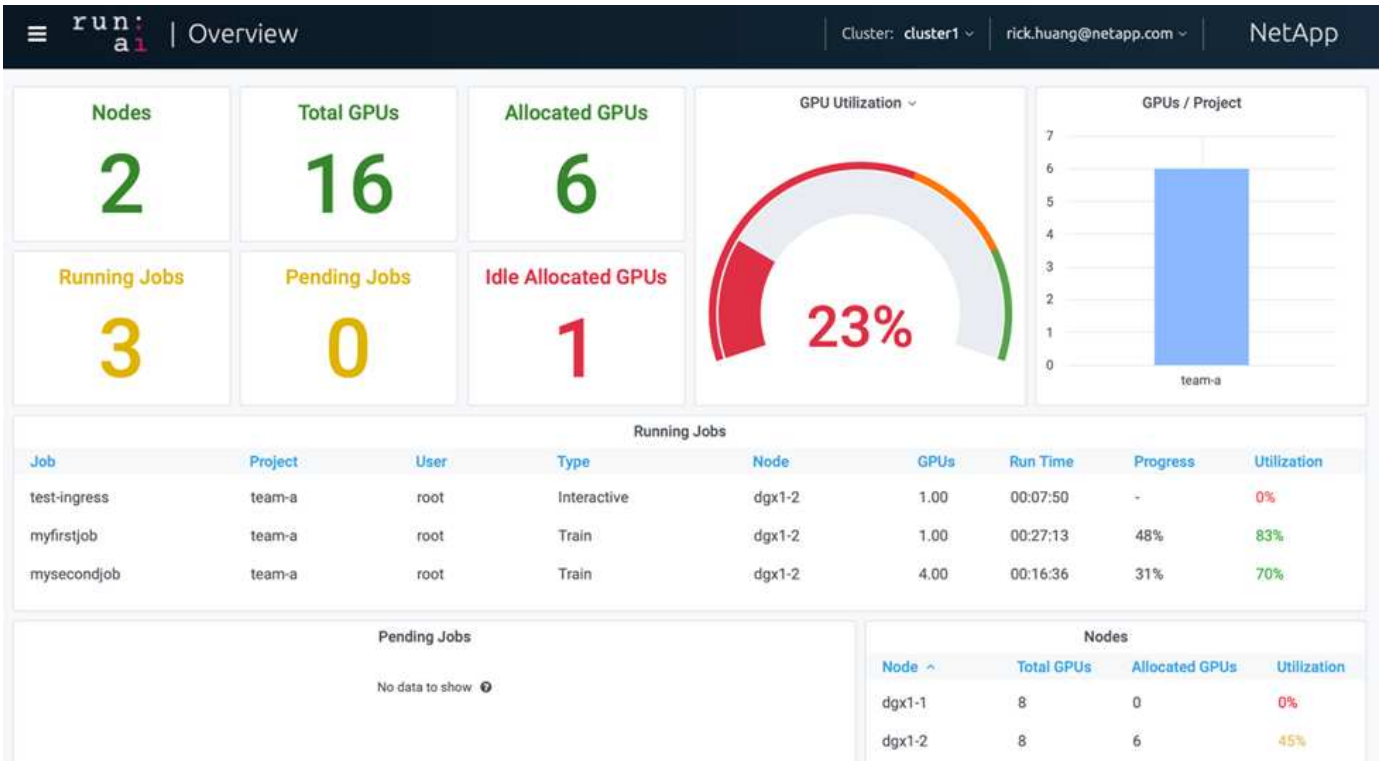
1. Installez le cluster Kubernetes à l'aide de DeepOps et configurez la classe de stockage par défaut de NetApp.
2. Préparez les nœuds GPU :
   a. Vérifiez que les pilotes NVIDIA sont installés sur les nœuds GPU.
   b. Vérifiez-le `nvidia-docker` est installé et configuré comme exécution docker par défaut.
3. Installer Run:ai :
   a. Connectez-vous au "Exécution : interface d'administration d'IA" pour créer le cluster.
   b. Téléchargez le créé `runai-operator-<clustername>.yaml` fichier.
   c. Appliquer la configuration de l'opérateur au cluster Kubernetes.

```
kubectl apply -f runai-operator-<clustername>.yaml
```

4. Vérifiez l'installation :
   a. Accédez à "https://app.run.ai/".
   b. Accédez au tableau de bord Présentation.
   c. Vérifiez que le nombre de GPU situé en haut à droite indique le nombre attendu de GPU et de nœuds GPU dans la liste des serveurs.pour plus d'informations sur le déploiement Run:ai, consultez "Installation de Run:ai sur un cluster Kubernetes sur site" et "Installation de la CLI Run:ai".

Exécution :tableaux de bord et vues d'IA

Une fois l'installation de Run:ai sur votre cluster Kubernetes et la configuration correcte des conteneurs, vous voyez les tableaux de bord et les vues suivants sur "https://app.run.ai" dans votre navigateur, comme illustré dans la figure suivante.



Le cluster fournit au total 16 GPU par deux nœuds DGX-1. Vous pouvez voir le nombre de nœuds, le nombre total de GPU disponibles, les GPU alloués avec des charges de travail, le nombre total de tâches en cours d'exécution, les tâches en attente et les GPU alloués inactifs. Sur le côté droit, le diagramme à barres affiche les GPU par projet, qui récapitule la manière dont les différentes équipes utilisent les ressources du cluster. Au milieu se trouve la liste des travaux en cours d'exécution avec des détails sur le travail, y compris le nom du travail, le projet, l'utilisateur, le type de travail, Le nœud sur lequel chaque travail est exécuté, le nombre de GPU alloués pour ce travail, la durée d'exécution actuelle du travail, la progression du travail en pourcentage et l'utilisation du GPU pour ce travail. Notez que le cluster est sous-utilisé (taux d'utilisation des GPU à 23 %) car seules trois tâches en cours d'exécution sont soumises par une seule équipe (`team-a`).

Dans la section suivante, nous allons voir comment créer plusieurs équipes dans l'onglet projets et allouer des GPU à chaque équipe pour optimiser l'utilisation du cluster et gérer les ressources lorsqu'il y a de nombreux utilisateurs par cluster. Les scénarios de test reproduisent les environnements d'entreprise dans lesquels la mémoire et les ressources GPU sont partagées entre les charges de travail d'entraînement, d'inférence et interactives.

Création de projets pour les équipes de data science et allocation de GPU

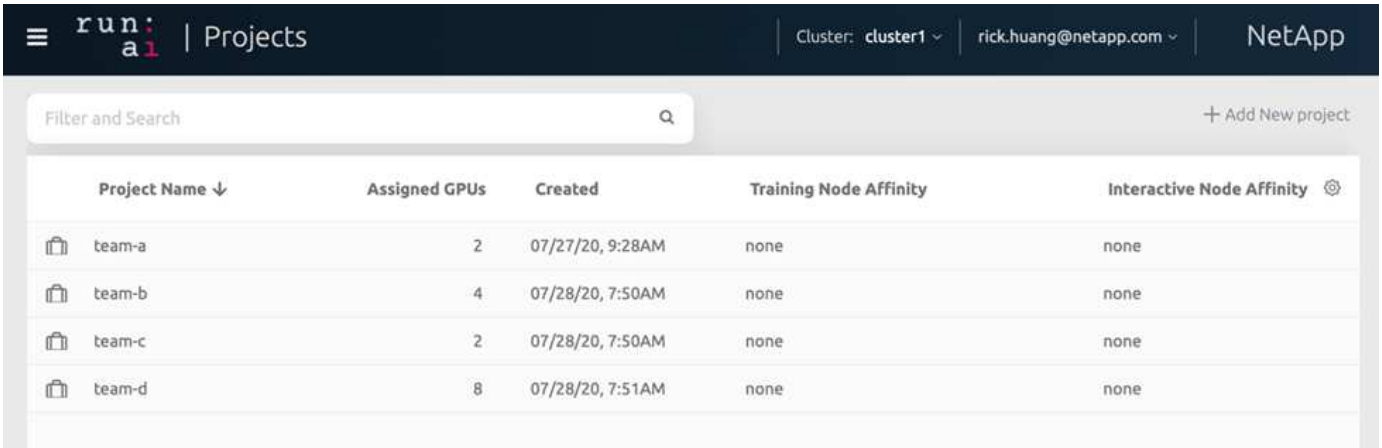
Les chercheurs peuvent envoyer des workloads via les processus Run:ai CLI, Kubeflow ou similaires. Afin de rationaliser l'allocation des ressources et de créer des priorités, Run:ai introduit le concept de projets. Les projets sont des entités de quotas qui associent un nom de projet à l'allocation et aux préférences GPU. Cette méthode simple et pratique vous permet de gérer plusieurs équipes de data science.

Un chercheur soumettant une charge de travail doit associer un projet à une demande de charge de travail. Le planificateur Run:ai compare la demande aux allocations en cours et au projet et détermine si la charge de travail peut être allouée à des ressources ou si elle doit rester dans un état en attente.

En tant qu'administrateur système, vous pouvez définir les paramètres suivants dans l'onglet exécution:projets ai :

- **Modèles de projets.** définissez un projet par utilisateur, définissez un projet par équipe d'utilisateurs et définissez un projet par projet d'organisation réel.
- **Quotas de projet.** chaque projet est associé à un quota de GPU qui peut être alloué pour ce projet en même temps. C'est un quota garanti. Les chercheurs qui utilisent ce projet sont garantis d'obtenir ce nombre de GPU, quel que soit l'état du cluster. En règle générale, la somme de l'allocation du projet doit être égale au nombre de GPU dans le cluster. Au-delà de cela, un utilisateur de ce projet peut recevoir un surquota. Tant que les GPU ne sont pas utilisés, un chercheur qui utilise ce projet peut obtenir davantage de processeurs graphiques. Nous démontrons des scénarios de test sur-quota et des considérations d'équité dans "Une utilisation élevée des clusters grâce à une allocation GPU sur-quota", "Équité de l'allocation des ressources de base", et "Équité excessive".
- Créez un nouveau projet, mettez à jour un projet existant et supprimez un projet existant.
- **Limiter l'exécution des travaux sur des groupes de nœuds spécifiques**. Vous pouvez affecter des projets spécifiques à exécuter uniquement sur des nœuds spécifiques. Ceci est utile lorsque l'équipe de projet a besoin de matériel spécialisé, par exemple, avec suffisamment de mémoire. Une équipe de projet peut également être propriétaire de matériel spécifique acquis avec un budget spécialisé, ou lorsqu'il peut être nécessaire de concevoir directement des charges de travail interactives ou de travailler sur du matériel plus faible et d'entraîner plus longtemps ou des charges de travail non surveillées sur des nœuds plus rapides. Pour des commandes permettant de regrouper des nœuds et de définir l'affinité pour un projet spécifique, reportez-vous à la "Exécuter:Documentation ai".
- **Limiter la durée des emplois interactifs**. Les chercheurs oublient souvent de fermer des emplois interactifs. Cela peut conduire à un gaspillage de ressources. Certaines organisations préfèrent limiter la durée des emplois interactifs et les fermer automatiquement.

La figure suivante montre la vue projets avec quatre équipes créées. Un nombre différent de GPU est attribué à chaque équipe pour prendre en compte différentes charges de travail, avec un nombre total de GPU égal à celui du nombre total de GPU disponibles dans un cluster composé de deux DGX-1.



Soumission des travaux dans l'interface de ligne de commande Run:ai

Cette section fournit des informations détaillées sur les commandes Basic Run:ai que vous pouvez utiliser pour exécuter n'importe quel travail Kubernetes. Elle est divisée en trois parties selon le type de charge de travail. Les workloads d'IA/AM/AP peuvent être divisés en deux types génériques :

- * Sessions de formation sans surveillance*. Lorsque ces types de charges de travail sont évoqués, le data Scientist prépare une charge de travail autonome et l'envoie pour exécution. Lors de l'exécution, le client peut examiner les résultats. Ce type de charge de travail est souvent utilisé en production ou lorsqu'il s'agit du développement du modèle, auquel aucune intervention humaine n'est requise.

- **Sessions de construction interactives**. Avec ce type de charges de travail, un data Scientist ouvre une session interactive avec Bash, Jupyter Notebook, PyChARM distant ou des IDE similaires. Ce dernier accède directement aux ressources GPU. Nous incluons un troisième scénario pour l'exécution de charges de travail interactives avec des ports connectés pour révéler un port interne pour l'utilisateur du conteneur.

Charges de travail d'entraînement sans assistance

Une fois les projets et l'allocation de GPU, vous pouvez exécuter n'importe quel workload Kubernetes à l'aide de la commande suivante sur la ligne de commande :

```
$ runai project set team-a runai submit hyper1 -i gcr.io/run-ai-
demo/quickstart -g 1
```

Cette commande lance une tâche d'entraînement automatique pour TEAM-a avec une allocation d'un GPU unique. La tâche est basée sur un exemple d'image docker, `gcr.io/run-ai-demo/quickstart`. Nous avons nommé le travail `hyper1`. Vous pouvez ensuite contrôler la progression du travail en exécutant la commande suivante :

```
$ runai list
```

La figure suivante montre le résultat du `runai list` commande. Les États typiques que vous pouvez voir incluent les suivants :

- `ContainerCreating`. Le conteneur docker est téléchargé depuis le référentiel cloud.
- `Pending`. Le travail est en attente d'être programmé.
- `Running`. La tâche est en cours d'exécution.

```
~> runai list
Showing jobs for project team-a
NAME    STATUS   AGE  NODE                                    IMAGE                        TYPE        PROJECT  USER   GPUs
hyper1  Running  11s  gke-dev-yaron1-gpu-4-pool-154f511d-5nk5  gcr.io/run-ai-demo/quickstart  Train       team-a   yaron  1
```

Pour obtenir un état supplémentaire de votre travail, exécutez la commande suivante :

```
$ runai get hyper1
```

Pour afficher les journaux du travail, exécutez le `runai logs <job-name>` commande :

```
$ runai logs hyper1
```

Dans cet exemple, vous devriez voir le journal d'une session d'apprentissage profond en cours d'exécution, y compris la période d'entraînement actuelle, la date d'arrivée prévue, la valeur de la fonction de perte, la précision et le temps écoulé pour chaque étape.

Vous pouvez afficher l'état du cluster dans l'interface utilisateur Run:ai à "https://app.run.ai/". Sous tableaux de bord > Présentation, vous pouvez surveiller l'utilisation des GPU.

Pour arrêter cette charge de travail, exécutez la commande suivante :

```
$ runai delte hyper1
```

Cette commande arrête la charge de travail d'entraînement. Vous pouvez vérifier cette action en exécutant `runai list` encore. Pour plus de détails, voir "lancement de charges de travail d'entraînement non surveillées".

Workloads de construction interactifs

Après avoir configuré des projets et affecté des GPU, vous pouvez exécuter un workload de construction interactif à l'aide de la commande suivante sur la ligne de commande :

```
$ runai submit build1 -i python -g 1 --interactive --command sleep --args
infinity
```

Le travail est basé sur un exemple de python d'image docker. Nous avons nommé le chantier 1.

> Le `-- interactive` drapeau signifie que le travail n'a pas de début ou de fin Il incombe au chercheur de fermer le poste. L'administrateur peut définir une limite de temps pour les travaux interactifs au terme desquels ils sont résiliés par le système.

Le `--g 1` Flag alloue un GPU unique à ce travail. La commande et l'argument fournis sont `--command sleep—args infinity`. Vous devez fournir une commande, ou le conteneur démarre et se ferme immédiatement.

Les commandes suivantes fonctionnent de la même manière que les commandes décrites dans la Charges de travail d'entraînement sans assistance:

- `runai list`: Affiche le nom, l'état, l'âge, le nœud, l'image, Le projet, l'utilisateur et les processeurs graphiques pour la réalisation de tâches.
- `runai get build1`: Affiche l'état supplémentaire sur le build de job 1.
- `runai delete build1`: Arrête la charge de travail interactive build1.pour obtenir un shell bash au conteneur, la commande suivante :

```
$ runai bash build1
```

Ceci fournit un shell direct dans l'ordinateur. Les data Scientists peuvent ensuite développer ou affiner leurs modèles dans le conteneur.

Vous pouvez afficher l'état du cluster dans l'interface utilisateur Run:ai à "https://app.run.ai". Pour plus de détails, voir "démarrage et utilisation de workloads de construction interactifs".

Charges de travail interactives avec des ports connectés

Comme extension des charges de travail de construction interactives, vous pouvez révéler les ports internes de l'utilisateur du conteneur lors du démarrage d'un conteneur à l'aide de l'interface de ligne de commande

Run:ai. Ce procédé est utile pour les environnements cloud, avec des ordinateurs portables Jupyter ou pour la connexion à d'autres microservices. "Entrée" Permet l'accès aux services Kubernetes depuis l'extérieur du cluster Kubernetes. Vous pouvez configurer l'accès en créant un ensemble de règles qui définissent les connexions entrantes qui atteignent quels services.

Pour une meilleure gestion de l'accès externe aux services d'un cluster, nous vous recommandons d'installer les administrateurs du cluster "Entrée" Et configurer LoadBalancer.

Pour utiliser Ingress as a service type, exécutez la commande suivante pour définir le type de méthode et les ports lors de l'envoi de votre workload :

```
$ runai submit test-ingress -i jupyter/base-notebook -g 1 \
  --interactive --service-type=ingress --port 8888 \
  --args="--NotebookApp.base_url=test-ingress" --command=start-notebook.sh
```

Une fois le conteneur démarré avec succès, exécutez `runai list` pour voir le `SERVICE URL(S)` Pour accéder au Jupyter Notebook. L'URL se compose du noeud final d'entrée, du nom du travail et du port. Par exemple, voir https://10.255.174.13/test-ingress-8888.

Pour plus de détails, voir "lancement d'une charge de travail de construction interactive avec des ports connectés".

Optimiser l'utilisation des clusters

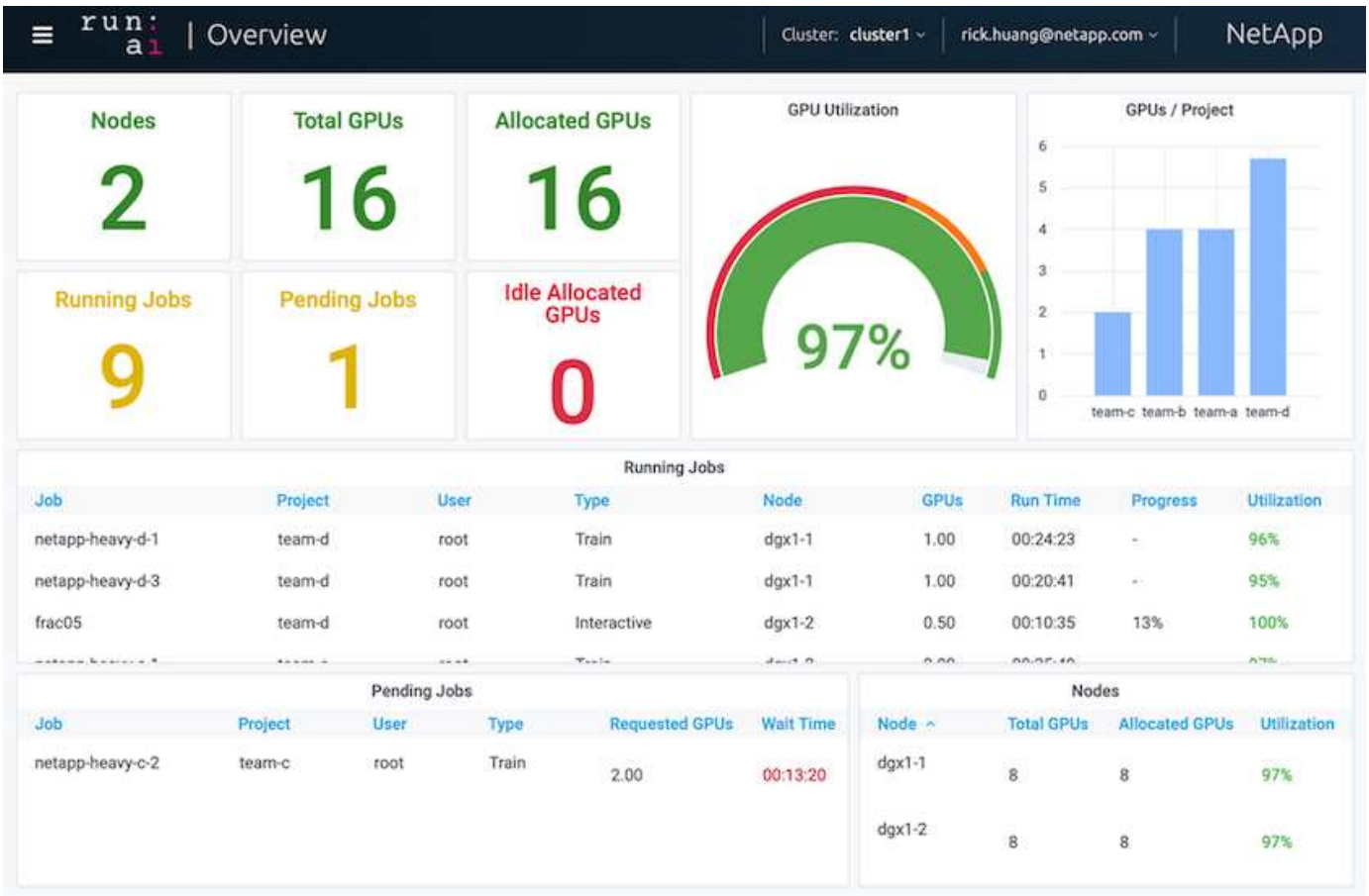
Dans cette section, nous émulons un scénario réaliste dans lequel quatre équipes de data science soumettent leurs propres charges de travail pour présenter la solution d'orchestration Run:ai qui optimise l'utilisation des clusters, tout en maintenant les niveaux de priorité et l'équilibrage des ressources GPU. Nous commençons par le banc d'essai ResNet-50 décrit dans la section "RESNET-50 avec résumé du banc d'essai ImageNet sur le dataset":

```
$ runai submit netapp1 -i netapp/tensorflow-tf1-py3:20.01.0 --local-image
--large-shm  -v /mnt:/mnt -v /tmp:/tmp --command python --args
"/netapp/scripts/run.py" --args "--
dataset_dir=/mnt/mount_0/dataset/imagenet/imagenet_original/" --args "--
num_mounts=2"  --args "--dgx_version=dgx1" --args "--num_devices=1" -g 1
```

Nous avons exécuté le même banc d'essai ResNet-50 que dans "NVA-1121". Nous avons utilisé le drapeau `--local-image` pour les conteneurs qui ne résident pas dans le référentiel docker public. Nous avons monté les répertoires `/mnt` et `/tmp` Sur le nœud hôte DGX-1 vers `/mnt` et `/tmp` vers le conteneur, respectivement. Le dataset est disponible dans le système NetApp AFFA800 avec le `dataset_dir` argument pointant vers le répertoire. Les deux `--num_devices=1` et `-g 1` Cela signifie que nous allouez un GPU pour ce travail. Le premier est un argument pour le `run.py` script, tandis que ce dernier est un indicateur pour le `runai submit` commande.

La figure ci-dessous montre un tableau de bord de présentation du système avec un taux d'utilisation des GPU de 97 % et les seize GPU disponibles alloués. Vous pouvez consulter le nombre de GPU alloués à chaque équipe dans le graphique à barres projet/GPU. Le volet travaux en cours affiche les noms des travaux en cours d'exécution, le projet, l'utilisateur, le type, le nœud, GPU utilisés, temps d'exécution, progression et

informations d'utilisation. Une liste des charges de travail en file d'attente avec leur temps d'attente est affichée dans tâches en attente. Enfin, la zone nœuds fournit les nombres de GPU et l'utilisation de nœuds DGX-1 individuels dans le cluster.



Allocation fractionnelle de GPU pour des charges de travail moins exigeantes ou interactives

Lorsque les chercheurs et développeurs travaillent sur leurs modèles, que ce soit au stade de développement, de réglage des hyperparamètres ou de débogage, ces charges de travail nécessitent généralement moins de ressources de calcul. Il est donc plus efficace de provisionner des GPU et de la mémoire fractionnaires afin que les mêmes GPU puissent être alloués simultanément à d'autres charges de travail. La solution d'orchestration Run:ai propose un système de partage GPU fractionnaire pour les workloads conteneurisés sur Kubernetes. Le système prend en charge les charges de travail exécutant des programmes CUDA et est particulièrement adapté aux tâches d'IA légères telles que l'inférence et la création de modèles. Le système GPU fractionnel permet aux équipes d'ingénierie d'IA et de data science d'exécuter plusieurs charges de travail simultanément sur un seul GPU. Les entreprises peuvent ainsi exécuter davantage de charges de travail, comme la vision informatique, la reconnaissance vocale et le traitement du langage naturel sur le même matériel, ce qui réduit les coûts.

Run :le système GPU fractionnel de l'IA crée efficacement des GPU logiques virtualisés avec leur propre mémoire et espace de calcul que les conteneurs peuvent utiliser et accéder comme s'il s'agissait de processeurs autonomes. Ce qui permet d'exécuter plusieurs charges de travail côte à côte dans des conteneurs sur le même GPU sans interférer entre les deux. La solution est transparente, simple et portable.

Elle ne nécessite aucune modification des conteneurs eux-mêmes.

Une upecase type peut voir deux à huit tâches s'exécutant sur le même GPU, ce qui signifie que vous pouvez faire huit fois le travail avec le même matériel.

Pour le travail `frac05` appartenant au projet `team-d` La figure suivante montre que le nombre de GPU alloués était de 0.50. Ceci est vérifié par le `nvidia-smi` La commande, qui montre que la mémoire GPU disponible pour le conteneur était de 16 255 Mo : la moitié de 32 Go par GPU V100 dans le nœud DGX-1.

```
root@run-deploy:~# runai bash frac05 -p team-d
root@frac05-0:/workload# nvidia-smi
Tue Jul 28 15:17:03 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.05    Driver Version: 450.51.05    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:07:00.0 Off |                    0 |
| N/A   57C    P0   240W / 300W |  15525MiB / 16255MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+


+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A       156      C   python3                       15525MiB |
+-----------------------------------------------------------------------------+
```

Une utilisation élevée des clusters grâce à une allocation GPU sur-quota

Dans cette section et dans les sections "Équité de l'allocation des ressources de base", et "Équité excessive", Nous avons conçu des scénarios de test avancés pour démontrer les capacités d'orchestration Run:ai pour une gestion de charges de travail complexe, une planification préventive automatique et un provisionnement GPU sur-quota. Nous avons ainsi pu atteindre une utilisation élevée des ressources en cluster et optimiser la productivité de nos équipes de data Scientists dans un environnement ONTAP d'IA.

Pour ces trois sections, définissez les projets et quotas suivants :

Projet	Quota
équipe a	4
équipe-b	2
equipe-c	2

Projet	Quota
équipe-d	8

En outre, nous utilisons les conteneurs suivants pour ces trois sections :

- Ordinateur portable Jupyter : `jupyter/base-notebook`
- Run:ai quickstart : `gcr.io/run-ai-demo/quickstart`

Nous avons défini les objectifs suivants dans ce scénario de test :

- Présenter la simplicité du provisionnement des ressources et le mode d'abstraction des ressources des utilisateurs
- Montrer comment les utilisateurs peuvent facilement provisionner des fractions d'un GPU et d'un nombre entier de GPU
- Montrer comment le système élimine les goulots d'étranglement de calcul en permettant aux équipes ou aux utilisateurs de dépasser leur quota de ressources en cas de GPU gratuits dans le cluster
- Montrez comment éliminer les goulots d'étranglement dans les pipelines de données avec la solution NetApp lors de l'exécution de tâches de calcul intensives, telles que le conteneur NetApp
- Montre comment plusieurs types de conteneurs sont exécutés à l'aide du système
  - Ordinateur portable Jupyter
  - Exécutez :conteneur d'IA
- Affiche une utilisation élevée lorsque le cluster est plein

Pour plus de détails sur la séquence de commande réelle exécutée pendant le test, reportez-vous à la section "Détails des tests pour la section 4.8".

Une fois les 13 workloads envoyés, vous voyez la liste des noms de conteneurs et des GPU alloués, comme illustré ci-dessous. Nous avons sept formations et six tâches interactives qui simulent quatre équipes de data science, chacune ayant ses propres modèles en cours de développement ou en cours d'exécution. Pour les travaux interactifs, les développeurs individuels utilisent Jupyter Notebooks pour écrire ou déboguer leur code. Il convient donc de provisionner des fractions GPU sans utiliser trop de ressources du cluster.

```
root@run-deploy:~# runai list -A
NAME           STATUS   AGE  NODE    IMAGE                          TYPE        PROJECT  USER  GPUs  CREATED BY CLI  SERVICE URL(S)
b-4-gg         Running  2m   dgx1-2  gcr.io/run-ai-demo/quickstart  Train       team-b   root  2     true
c-5-g          Running  2m   dgx1-2  gcr.io/run-ai-demo/quickstart  Train       team-c   root  1     true
c-4-gg         Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-c   root  2     true
b-3-g          Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-b   root  1     true
c-3-g02        Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Interactive team-c   root  0.2   true
d-1-gggg       Running  2m   dgx1-2  gcr.io/run-ai-demo/quickstart  Train       team-d   root  4     true
c-2-g03        Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Interactive team-c   root  0.3   true
c-1-g05        Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Interactive team-c   root  0.5   true
a-2-gg         Running  3m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-a   root  2     true
b-2-g04        Running  3m   dgx1-2  gcr.io/run-ai-demo/quickstart  Interactive team-b   root  0.4   true
a-1-g          Running  3m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-a   root  1     true
b-1-g06        Running  3m   dgx1-2  gcr.io/run-ai-demo/quickstart  Interactive team-b   root  0.6   true
a-1-1-jupyter  Running  3m   dgx1-1  jupyter/base-notebook          Interactive team-a   root  1     true            http://10.61.218.134/a-1-1-jupyter,
https://10.61.218.134/a-1-1-jupyter
```

Les résultats de ce scénario de test montrent les éléments suivants :

- Le cluster doit être complet : 16/16 GPU sont utilisés.
- Forte utilisation du cluster.
- Plus de tests que les GPU en raison de l'allocation fractionnaire.

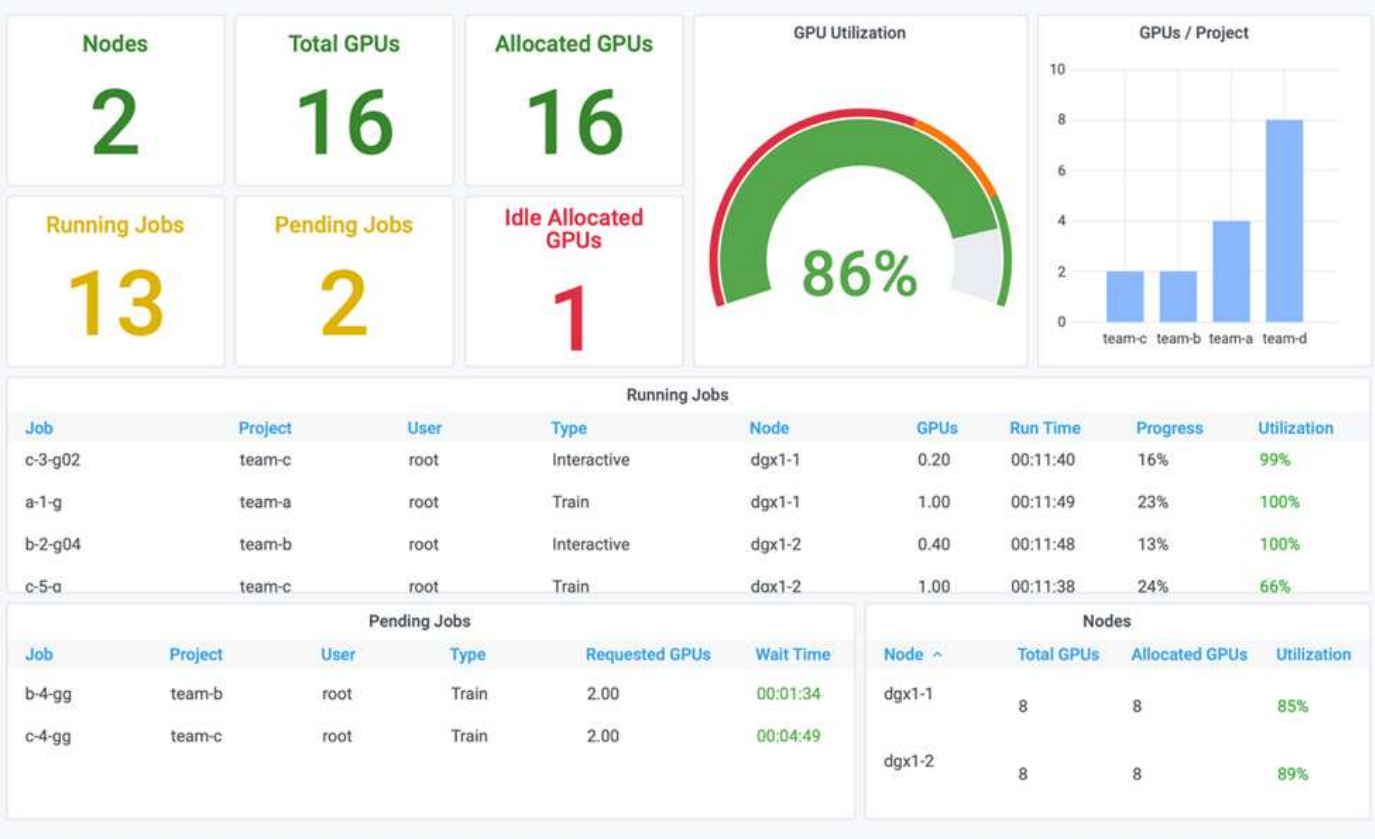
- `team-d` n'utilise pas tous leurs quotas ; par conséquent, `team-b` et `team-c` Peuvent exploiter des GPU supplémentaires pour leurs expérimentations, ce qui permet d'accélérer l'innovation.

Équité de l'allocation des ressources de base

Dans cette section, quand nous le montrons `team-d` Lorsque plusieurs GPU sont en attente (ils sont sous leur quota), le système interrompt les charges de travail de `team-b` et `team-c` et les déplace dans un état en attente de façon équitable.

Pour plus de détails, y compris les soumissions de travaux, les images de conteneur utilisées et les séquences de commandes exécutées, reportez-vous à la section "Détails des tests pour la section 4.9".

La figure suivante montre l'utilisation du cluster résultant, les GPU alloués par équipe et les tâches en attente dus à l'équilibrage de la charge automatique et à la planification préventive. Nous pouvons observer que lorsque le nombre total de GPU requis par toutes les charges de travail d'équipe dépasse le nombre total de GPU disponibles dans le cluster, l'algorithme d'équité interne de Run:ai interrompt chaque tâche pour `team-b` et `team-c` parce qu'ils ont atteint leur quota de projet. Cela assure, de manière générale, l'utilisation élevée des clusters et les équipes de data Scientists restent contraintes pour les ressources définies par un administrateur.



Les résultats de ce scénario de test démontrent ce qui suit :

- **Équilibrage automatique de la charge.** le système équilibre automatiquement le quota des GPU, de sorte que chaque équipe utilise maintenant son quota. Les charges de travail suspendues appartiennent à des équipes qui ont terminé leur quota.
- **Pause de partage équitable.** le système choisit d'arrêter la charge de travail d'une équipe qui a surpassé leur quota et de mettre fin à la charge de travail de l'autre équipe. Exécutez :l'IA dispose d'algorithmes

d’équité internes.

Équité excessive

Dans cette section, nous étudions le scénario dans lequel plusieurs équipes soumettent des charges de travail pour dépasser leurs quotas. De cette façon, nous démontrons comment l’algorithme Run:ai d’équité alloue des ressources de cluster en fonction du rapport des quotas prédéfinis.

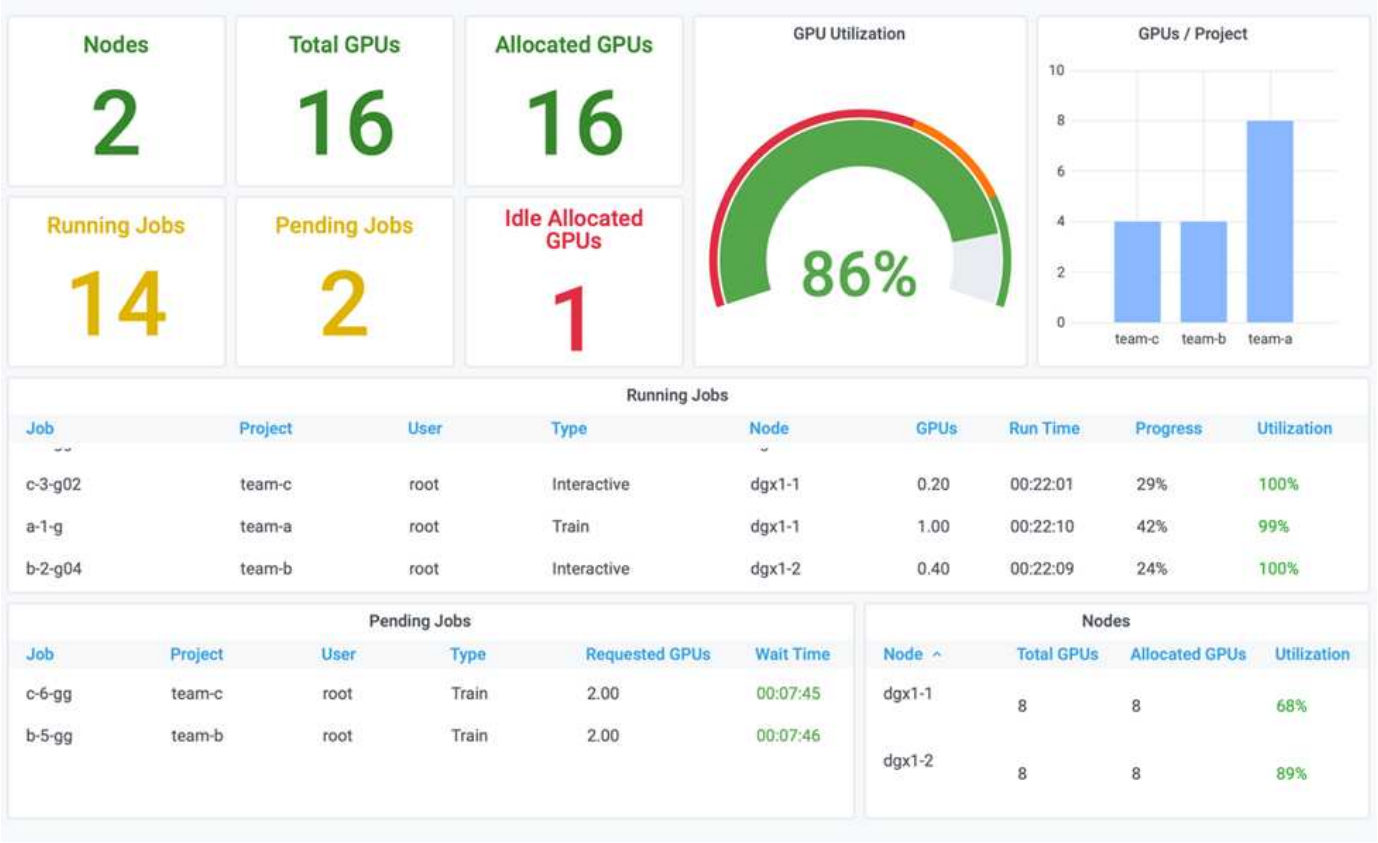
Objectifs de ce scénario de test :

- Montrer le mécanisme de mise en file d’attente lorsque plusieurs équipes demandent des GPU sur leur quota.
- Montrer comment le système distribue une juste part du cluster entre plusieurs équipes qui surpartagent leur quota d’après le rapport entre leurs quotas, de sorte que l’équipe disposant du quota supérieur puisse obtenir une plus grande part de la capacité disponible.

À la fin de "Équité de l’allocation des ressources de base", deux charges de travail sont mises en file d’attente : une pour `team-b` et un pour `team-c`. Cette section décrit les workloads supplémentaires mis en file d’attente.

Pour plus de détails, y compris les soumissions de travaux, les images de conteneur utilisées et les séquences de commandes exécutées, voir "Détails des tests pour la section 4.10".

Lorsque tous les travaux sont soumis conformément à la section "Détails des tests pour la section 4.10", le tableau de bord du système le montre `team-a`, `team-b`, et `team-c` Tous ont plus de GPU que leur quota prédéfini. `team-a` Occupe quatre GPU de plus que son soft quota (quatre) prédéfini, alors que `team-b` et `team-c` Chacun d’eux occupent deux GPU de plus que leur soft quota (deux). Le rapport des GPU sur-quotas alloués est égal à celui de leur quota prédéfini. En effet, le système a utilisé le quota prédéfini comme référence en priorité et provisionné en conséquence lorsque plusieurs équipes demandent plus de GPU, ce qui dépasse leur quota. Un tel équilibrage automatique de la charge garantit équité et priorisation lorsque les équipes de data science d’entreprise sont activement engagées dans le développement et la production de modèles d’IA.



Les résultats de ce scénario de test montrent les éléments suivants :

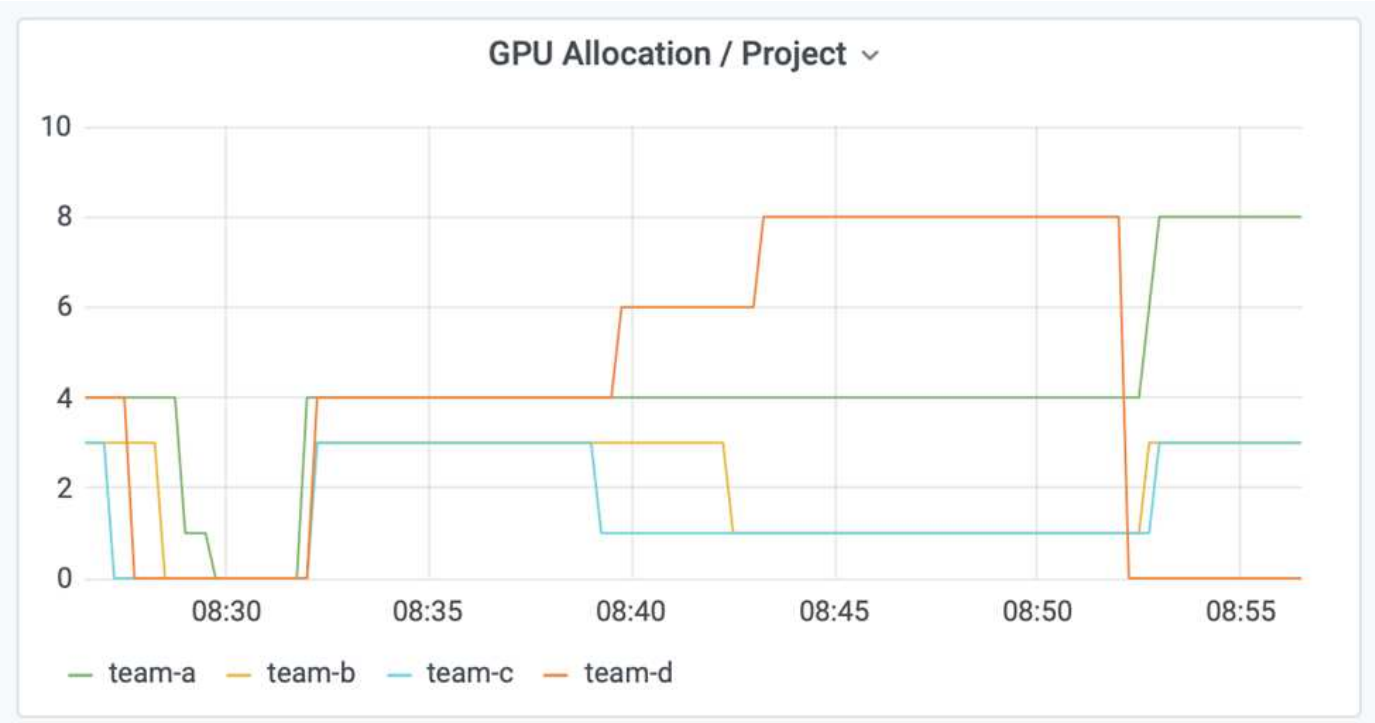
- Le système commence à mettre hors file d'attente les charges de travail des autres équipes.
- L'ordre de la démise en file d'attente est déterminé selon des algorithmes d'équité, de telle sorte que `team-b` et `team-c` Obtenir la même quantité de GPU sur-quota (puisqu'ils ont un quota similaire), et `team-a` Double quantité de GPU dans la mesure où le quota est deux fois plus élevé que celui de `team-b` et `team-c`.
- Toute l'allocation est effectuée automatiquement.

Par conséquent, le système doit se stabiliser sur les États suivants :

Projet	GPU alloués	Commentaire
équipe a	8/4	Quatre GPU par rapport au quota. File d'attente vide.
équipe-b	4/2	Deux GPU sur le quota. Une charge de travail mise en file d'attente
equipe-c	4/2	Deux GPU sur le quota. Une charge de travail mise en file d'attente
équipe-d	0/8	N'utilisez pas du tout des GPU, aucun workload mis en attente.

La figure suivante montre l'allocation des GPU par projet dans le temps dans le tableau de bord Run:ai Analytics "Une utilisation élevée des clusters grâce à une allocation GPU sur-quota", "Équité de l'allocation des

ressources de base", et "Équité excessive". Chaque ligne de la figure indique le nombre de GPU provisionnés pour une équipe de data science donnée à tout moment. Le système obtient ainsi l'allocation dynamique de GPU en fonction des charges de travail envoyées. Cette approche permet aux équipes de dépasser les quotas en présence de GPU disponibles dans le cluster, puis d'anticiper les tâches par équité avant d'atteindre la stabilité des quatre équipes.



Sauvegarde des données vers un volume persistant provisionné par Trident

NetApp Trident est un projet open source entièrement pris en charge et conçu pour vous aider à répondre aux exigences de persistance sophistiquées de vos applications conteneurisées. Vous pouvez lire et écrire des données sur un volume persistant Kubernetes persistant provisionné par Trident avec le Tiering des données, le chiffrement, la technologie NetApp Snapshot, la conformité et les performances élevées proposés par le logiciel de gestion des données NetApp ONTAP.

Réutilisation des ESV dans un espace de noms existant

Pour les projets d'IA plus volumineux, il est possible que les conteneurs soient plus efficaces pour lire et écrire les données sur le même volume persistant Kubernetes. Pour réutiliser une demande de volume persistant Kubernetes, l'utilisateur doit avoir déjà créé une demande de volume persistant. Voir la "Documentation NetApp Trident" Pour plus de détails sur la création d'une demande de volume persistant. Voici un exemple de réutilisation d'un PVC existant :

```
$ runai submit pvc-test -p team-a --pvc test:/tmp/pvc1mount -i gcr.io/run-ai-demo/quickstart -g 1
```

Exécutez la commande suivante pour afficher le statut de cette tâche `pvc-test` pour projet `team-a`:

```
$ runai get pvc-test -p team-a
```

Vous devriez voir le PV /tmp/pvc1mount monté sur `team-a` travail `pvc-test`. Ainsi, plusieurs conteneurs peuvent être lus depuis le même volume, ce qui est utile lorsque plusieurs modèles concurrents sont en cours de développement ou en production. Les data Scientists peuvent élaborer un ensemble de modèles, puis combiner les résultats de prédiction par vote à la majorité ou d'autres techniques.

Pour accéder au shell du conteneur, procédez comme suit :

```
$ runai bash pvc-test -p team-a
```

Vous pouvez ensuite vérifier le volume monté et accéder à vos données dans le conteneur.

Cette fonctionnalité de réutilisation des volumes virtuels fonctionne avec des volumes NetApp FlexVol et des volumes NetApp ONTAP FlexGroup. Les ingénieurs de données peuvent ainsi utiliser des options de gestion des données plus flexibles et plus robustes pour exploiter votre Data Fabric optimisée par NetApp.

Conclusion

NetApp et Run:ai se sont associés dans ce rapport technique pour présenter les fonctionnalités uniques de la solution NetApp ONTAP ai avec la plateforme Run:ai pour simplifier l'orchestration des workloads d'IA. Les étapes précédentes procurent une architecture de référence afin de rationaliser le processus de pipelines de données et d'orchestration des workloads pour le deep learning. Les clients qui souhaitent implémenter ces solutions sont invités à contacter NetApp et Run:ai pour en savoir plus.

Détails des tests pour la section 4.8

Cette section contient les détails des tests de la section "Une utilisation élevée des clusters grâce à une allocation GPU sur-quota".

Soumettre les travaux dans l'ordre suivant :

Projet	Image	NB de GPU	Total	Commentaire
équipe a	Jupyter	1	1/4	–
équipe a	NetApp	1	2/4	–
équipe a	Exécutez :ai	2	4/4	Utiliser tous leurs quotas
équipe-b	Exécutez :ai	0.6	0.6/2	GPU fractionnaires
équipe-b	Exécutez :ai	0.4	1/2	GPU fractionnaires
équipe-b	NetApp	1	2/2	–
équipe-b	NetApp	2	4/2	Deux sur le quota
equipe-c	Exécutez :ai	0.5	0.5/2	GPU fractionnaires

Projet	Image	NB de GPU	Total	Commentaire
equipe-c	Exécutez :ai	0.3	0.8/2	GPU fractionnaires
equipe-c	Exécutez :ai	0.2	1/2	GPU fractionnaires
equipe-c	NetApp	2	3/2	Un au-delà des quotas
equipe-c	NetApp	1	4/2	Deux sur le quota
équipe-d	NetApp	4	4/8	Utilisant la moitié de leur quota

Structure de commande :

```
$ runai submit <job-name> -p <project-name> -g <#GPUs> -i <image-name>
```

Séquence de commande réelle utilisée lors du test :

```
$ runai submit a-1-1-jupyter -i jupyter/base-notebook -g 1 \
  --interactive --service-type=ingress --port 8888 \
  --args="--NotebookApp.base_url=team-a-test-ingress" --command=start
-notebook.sh -p team-a
$ runai submit a-1-g -i gcr.io/run-ai-demo/quickstart -g 1 -p team-a
$ runai submit a-2-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-a
$ runai submit b-1-g06 -i gcr.io/run-ai-demo/quickstart -g 0.6
--interactive -p team-b
$ runai submit b-2-g04 -i gcr.io/run-ai-demo/quickstart -g 0.4
--interactive -p team-b
$ runai submit b-3-g -i gcr.io/run-ai-demo/quickstart -g 1 -p team-b
$ runai submit b-4-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-b
$ runai submit c-1-g05 -i gcr.io/run-ai-demo/quickstart -g 0.5
--interactive -p team-c
$ runai submit c-2-g03 -i gcr.io/run-ai-demo/quickstart -g 0.3
--interactive -p team-c
$ runai submit c-3-g02 -i gcr.io/run-ai-demo/quickstart -g 0.2
--interactive -p team-c
$ runai submit c-4-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-c
$ runai submit c-5-g -i gcr.io/run-ai-demo/quickstart -g 1 -p team-c
$ runai submit d-1-gggg -i gcr.io/run-ai-demo/quickstart -g 4 -p team-d
```

À ce stade, vous devez avoir les États suivants :

Projet	GPU alloués	Charges de travail en file d’attente
équipe a	4/4 (quota mou/allocation réelle)	Aucune

Projet	GPU alloués	Charges de travail en file d'attente
équipe-b	4/2	Aucune
equipe-c	4/2	Aucune
équipe-d	4/8	Aucune

Voir la section "Optimiser l'utilisation des clusters grâce à l'allocation GPU sur-organismes" pour les discussions sur le scénario de test en cours.

Détails des tests pour la section 4.9

Cette section contient des détails sur les tests pour la section "Équité de l'allocation des ressources de base".

Soumettre les travaux dans l'ordre suivant :

Projet	NB de GPU	Total	Commentaire
équipe-d	2	6/8	La charge de travail TEAM-b/c s'interrompt et passe à `pending`.
équipe-d	2	8/8	Les autres charges de travail d'équipe (b/c) sont suspendues et transférées vers `pending`.

Voir la séquence de commande exécutée suivante :

```
$ runai submit d-2-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-d$
runai submit d-3-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-d
```

À ce stade, vous devez avoir les États suivants :

Projet	GPU alloués	Charges de travail en file d'attente
équipe a	4/4	Aucune
équipe-b	2/2	Aucune
equipe-c	2/2	Aucune
équipe-d	8/8	Aucune

Voir la section "Équité de l'allocation des ressources de base" pour une discussion sur le scénario de test en cours.

Détails des tests pour la section 4.10

Cette section contient des détails sur les tests pour la section "Équité excessive".

Soumettre les travaux dans l'ordre suivant pour `team-a`, `team-b`, et `team-c`:

Projet	NB de GPU	Total	Commentaire
équipe a	2	4/4	1 charge de travail mise en attente
équipe a	2	4/4	2 workloads mis en file d'attente
équipe-b	2	2/2	2 workloads mis en file d'attente
equipe-c	2	2/2	2 workloads mis en file d'attente

Voir la séquence de commande exécutée suivante :

```
$ runai submit a-3-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-a$
runai submit a-4-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-a$ runai
submit b-5-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-b$ runai
submit c-6-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-c
```

À ce stade, vous devez avoir les États suivants :

Projet	GPU alloués	Charges de travail en file d'attente
équipe a	4/4	Deux workloads demandent deux GPU
équipe-b	2/2	Deux workloads demandent chacun deux GPU
equipe-c	2/2	Deux workloads demandent chacun deux GPU
équipe-d	8/8	Aucune

Supprimez ensuite toutes les charges de travail de `team-d`:

```
$ runai delete -p team-d d-1-gggg d-2-gg d-3-gg
```

Voir la section "Équité excessive", pour les discussions sur le scénario de test en cours.

Sources d'informations complémentaires

Pour en savoir plus sur les informations données dans ce document, consultez les

ressources suivantes :

- Systèmes NVIDIA DGX
  - Système NVIDIA DGX-1https://www.nvidia.com/en-us/data-center/dgx-1/[]
  - GPU NVIDIA V100 à cœurs Tensorhttps://www.nvidia.com/en-us/data-center/tesla-v100/[]
  - NVIDIA NGChttps://www.nvidia.com/en-us/gpu-cloud/[]
- Exécutez :solution d'orchestration de conteneurs d'IA
  - Run :présentation des produits d'IAhttps://docs.run.ai/home/components/[]
  - Run:documentation d'installation aihttps://docs.run.ai/Administrator/Cluster-Setup/Installing-Run-AI-on-an-on-premise-Kubernetes-Cluster/[] https://docs.run.ai/Administrator/Researcher-Setup/Installing-the-Run-AI-Command-Line-Interface/
  - Soumission de travaux dans l'interface de ligne de commande Run:aihttps://docs.run.ai/Researcher/Walkthroughs/Walkthrough-Launch-Unattended-Training-Workloads-/[] https://docs.run.ai/Researcher/Walkthroughs/Walkthrough-Start-and-Use-Interactive-Build-Workloads-/
  - Allocation de fractions GPU dans l'interface de ligne de commande Run:aihttps://docs.run.ai/Researcher/Walkthroughs/Walkthrough-Using-GPU-Fractions/[]
- Plan de contrôle d'IA NetApp
  - Rapport techniquehttps://www.netapp.com/us/media/tr-4798.pdf[]
  - Démonstration sous forme abrégéehttps://youtu.be/gfr_sO27Rvo[]
  - Référentiel GitHubhttps://github.com/NetApp/kubeflow_jupyter_pipeline[]
- Systèmes NetApp AFF
  - Fiche technique NetApp AFF A-Serieshttps://www.netapp.com/us/media/ds-3582.pdf[]
  - Avantages du Flash NetApp pour les systèmes FAS 100 % Flashhttps://www.netapp.com/us/media/ds-3733.pdf[]
  - Bibliothèque d'informations ONTAP 9http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286[]
  - Rapport technique sur NetApp ONTAP FlexGroup volumeshttps://www.netapp.com/us/media/tr-4557.pdf[]
- NetApp ONTAP ai
  - Guide de conception d'ONTAP ai avec DGX-1 et connectivité réseau Ciscohttps://www.netapp.com/us/media/nva-1121-design.pdf[]
  - Guide de déploiement d'ONTAP ai avec DGX-1 et réseau Ciscohttps://www.netapp.com/us/media/nva-1121-deploy.pdf[]
  - Guide de conception d'ONTAP ai avec DGX-1 et réseau Mellanoxhttp://www.netapp.com/us/media/nva-1138-design.pdf[]
  - Guide de conception de ONTAP ai avec DGX-2https://www.netapp.com/us/media/nva-1135-design.pdf[]

Tr-4799 : architecture de référence NetApp ONTAP ai pour les workloads de conduite autonome

David Arnette et Sung-Han Lin, NetApp

La gamme NVIDIA DGX est la première plateforme d'intelligence artificielle intégrée au monde conçue spécialement pour l'IA d'entreprise. Les systèmes de stockage NetApp AFF combinent performances optimales et fonctionnalités de pointe pour la gestion des données dans le cloud hybride. NetApp et NVIDIA se sont associés pour créer l'architecture de référence NetApp ONTAP ai, afin de proposer une solution clé en main qui prend en charge les workloads d'IA et de machine learning (ML), tout en offrant des performances et une fiabilité exceptionnelles.

"Tr-4799 : architecture de référence NetApp ONTAP ai pour les workloads de conduite autonome"

Tr-4811 : Architecture de référence NetApp ONTAP ai pour le secteur de la santé : imagerie diagnostique - conception de la solution

Rick Huang, Sung-Han Lin, Satish Thyagarajan, NetApp Jacci Cencii, NVIDIA

Cette architecture de référence apporte des lignes directrices aux clients du secteur de la santé qui créent des infrastructures d'intelligence artificielle (IA) reposant sur des systèmes NVIDIA DGX-2 et de stockage NetApp AFF. Vous y trouverez des informations sur les workflows généraux utilisés dans le développement de modèles de deep learning (DL) pour l'imagerie diagnostique médicale, des scénarios et des résultats de tests de validation Il inclut également des recommandations de dimensionnement pour les déploiements client.

"Tr-4811 : Architecture de référence NetApp ONTAP ai pour le secteur de la santé : imagerie diagnostique - conception de la solution"

Tr-4807 : architecture de référence NetApp ONTAP ai pour les workloads de services financiers - conception de la solution

Karthikeyan Nagalingam, Sung-Han Lin, NetApp Jacci Cenci, NVIDIA

Cette architecture de référence apporte des lignes directrices aux clients qui créent une infrastructure d'intelligence artificielle reposant sur des systèmes NVIDIA DGX-1 et de stockage NetApp AFF pour les champs d'application du secteur financier. Vous y trouverez des informations sur les workflows généraux utilisés dans le développement de modèles de deep learning pour les scénarios de test et les résultats des services financiers. Il inclut également des recommandations de dimensionnement pour les déploiements client.

"Tr-4807 : architecture de référence NetApp ONTAP ai pour les workloads de services financiers - conception de la solution"

Valeur de l’IA générative et de l’NetApp

Auteur: Sathyish Thyagarajan, NetApp

Résumé

La demande en intelligence artificielle générative (IA) génère des perturbations dans tous les secteurs, ce qui favorise la créativité commerciale et l’innovation produit. De nombreuses entreprises utilisent l’IA générative pour créer de nouvelles fonctionnalités, améliorer la productivité de l’ingénierie et créer des prototypes d’applications basés sur l’IA qui offrent de meilleurs résultats et expériences client. L’IA générative, telle que Generative Pre-entrainés Transformers (GPT), utilise des réseaux neuronaux pour créer de nouveaux contenus aussi divers que du texte, de l’audio et de la vidéo. Étant donné l’ampleur extrême et les jeux de données massifs impliqués dans les grands modèles de langage (LLMS), il est essentiel de concevoir une infrastructure d’IA robuste qui tire parti des puissantes fonctionnalités de stockage des données des options de déploiement sur site, hybride et multicloud et réduit les risques associés à la mobilité des données. protection et gouvernance des données avant que les entreprises ne puissent concevoir des solutions d’IA. Ce livre blanc présente ces considérations et les fonctionnalités NetApp® ai correspondantes qui permettent une gestion transparente des données et un déplacement fluide des données dans le pipeline de traitement de données d’IA pour l’entraînement, le recyclage, le réglage fin et l’inférence des modèles d’IA génératifs.

Synthèse

Plus récemment, après le lancement de ChatGPT, un spin-off de GPT-3 en novembre 2022, de nouveaux outils d’IA utilisés pour générer du texte, du code, de l’image, ou même des protéines thérapeutiques en réponse aux invites de l’utilisateur ont acquis une renommée significative. Cela indique que les utilisateurs peuvent faire une demande en utilisant le langage naturel. L’IA interprète et génère du texte, par exemple des articles d’actualité ou des descriptions de produits qui reflètent les demandes de l’utilisateur ou produisent du code, de la musique, de la parole, des effets visuels et des ressources 3D à l’aide d’algorithmes entraînés sur des données déjà existantes. En conséquence, des expressions telles que diffusion stable, hallucinations, ingénierie rapide et alignement de la valeur apparaissent rapidement dans la conception des systèmes d’IA. Ces modèles de machine learning (ML) autosupervisés ou semi-supervisés deviennent largement disponibles sous forme de modèles de base pré-entraînés (FM) via des fournisseurs de services cloud et d’autres fournisseurs de firmwares d’IA. Ces modèles sont adoptés par divers établissements d’entreprises de tous secteurs pour un large éventail de tâches de traitement du langage naturel (TLN) en aval. Comme l’a affirmé des sociétés d’analyse de recherche telles que McKinsey, « l’impact de l’IA générative sur la productivité pourrait ajouter des milliards de dollars à l’économie mondiale. » Si les entreprises réinventent l’IA en tant que partenaires d’opinion pour l’homme et que le FMS s’élargit simultanément à ce que les entreprises et les institutions peuvent faire avec l’IA générative, les opportunités de gestion de volumes massifs de données vont continuer de croître. Nous présentons dans ce document des informations d’introduction à l’IA générative et les concepts de conception relatifs aux fonctionnalités NetApp qui apportent de la valeur aux clients NetApp, dans les environnements sur site et hybrides ou multicloud.

Que contient NetApp dans leurs environnements d’IA ? NetApp aide les entreprises à faire face aux complexités liées à la croissance rapide des données et du cloud, à la gestion multicloud et à l’adoption de technologies nouvelle génération, telles que l’IA. NetApp a combiné diverses fonctionnalités à un logiciel de gestion intelligente des données et à une infrastructure de stockage parfaitement équilibrées. Les solutions d’IA génératives telles que les LLMS doivent à plusieurs reprises lire et traiter leurs datasets sources depuis le stockage vers la mémoire afin de favoriser l’intelligence. NetApp a été l’un des leaders des technologies de mobilité, de gouvernance et de sécurité des données dans l’écosystème périphérie/cœur/cloud, pour aider les entreprises à créer des solutions d’IA à grande échelle. Avec un solide réseau de partenaires, NetApp aide les Chief Data Officer, les ingénieurs d’IA, les architectes d’entreprise et les data Scientists à concevoir un pipeline de données fluide pour la préparation et la protection des données, et des responsabilités stratégiques en matière de gestion des données pour l’entraînement des modèles d’IA et l’inférence, afin d’optimiser les

performances et l'évolutivité du cycle de vie d'IA et de ML. Les technologies et fonctionnalités de données NetApp telles que le pipeline de données NetApp® ONTAP ai® pour le deep learning, NetApp® SnapMirror® pour le transport fluide et efficace des données entre les terminaux de stockage, et NetApp® FlexCache® pour un rendu en temps réel lorsque le flux de données passe du mode batch au mode temps réel et que l'ingénierie des données se produit rapidement, apportez une valeur ajoutée au déploiement de modèles d'IA génératifs en temps réel. Alors que les entreprises de tous types adoptent de nouveaux outils d'IA, elles sont confrontées aux défis liés aux données, de la périphérie au data Center et jusqu'au cloud, qui exigent des solutions d'IA évolutives, responsables et explicables. En tant que référence en matière de gestion des données dans le cloud hybride et le multicloud, NetApp s'engage à créer un réseau de partenaires et de solutions conjointes pouvant aider à tous les aspects de la construction d'un pipeline de données et de data Lakes pour l'entraînement de modèles d'IA génératifs (pré-entraînement), le réglage précis, l'inférence basée sur le contexte et la surveillance de la désintégration de modèles de LLMS.

Qu'est-ce que l'IA générative ?

L'IA générative révolutionne la façon dont nous créons du contenu, générons de nouveaux concepts et explorons de nouvelles compositions. Il illustre les réseaux neuronaux comme le réseau antagoniste génératif (GAN), les codeurs automatiques variationnels (VAE) et les transformateurs pré-entraînés génératifs (GPT), qui peuvent générer un nouveau contenu comme le texte, le code, les images, l'audio, la vidéo, et des données synthétiques. Des modèles basés sur des transformateurs comme le Chat-GPT d'OpenAI, le Bard de Google, la FLORAISON DE la face de se tasser, et le Llama de Meta sont apparus comme la technologie de base qui sous-tend de nombreuses avancées dans les grands modèles linguistiques. De même, Dall-E d'OpenAI, CM3leon de Meta et Imagen de Google sont des exemples de modèles de diffusion texte-image qui offrent aux clients un degré sans précédent de photoréalisme pour créer de nouvelles images complexes à partir de zéro ou modifier des images existantes afin de générer des images de haute qualité prenant en compte le contexte en utilisant l'augmentation des jeux de données et la synthèse texte-image reliant la sémantique textuelle et visuelle. Les artistes numériques commencent à appliquer une combinaison de technologies de rendu comme Nerf (neural Radiance Field) avec l'IA générative pour convertir des images 2D statiques en scènes 3D immersives. En général, les LLMS se caractérisent par quatre paramètres : (1) taille du modèle (généralement en milliards de paramètres) ; (2) taille du dataset d'entraînement ; (3) coût de l'entraînement et (4) performances du modèle après l'entraînement. Les LLMS se répartissent également principalement dans trois architectures de transformateur. (i) modèles avec encodeur uniquement. Par exemple, BERT (Google, 2018) ; (ii) Encoder-décodeur, par exemple BART (Meta, 2020) et (iii) Decoder-Only modèles. Par exemple, Llama (Meta, 2023), Palm-E (Google, 2023). Selon les besoins de l'entreprise, quelle que soit l'architecture choisie, le nombre de paramètres du modèle (N) et le nombre de jetons (D) dans le dataset d'entraînement déterminent généralement le coût de base de l'entraînement (pré-entraînement) ou le réglage d'un LLM.

Cas d'utilisation en entreprise et tâches de NLP en aval

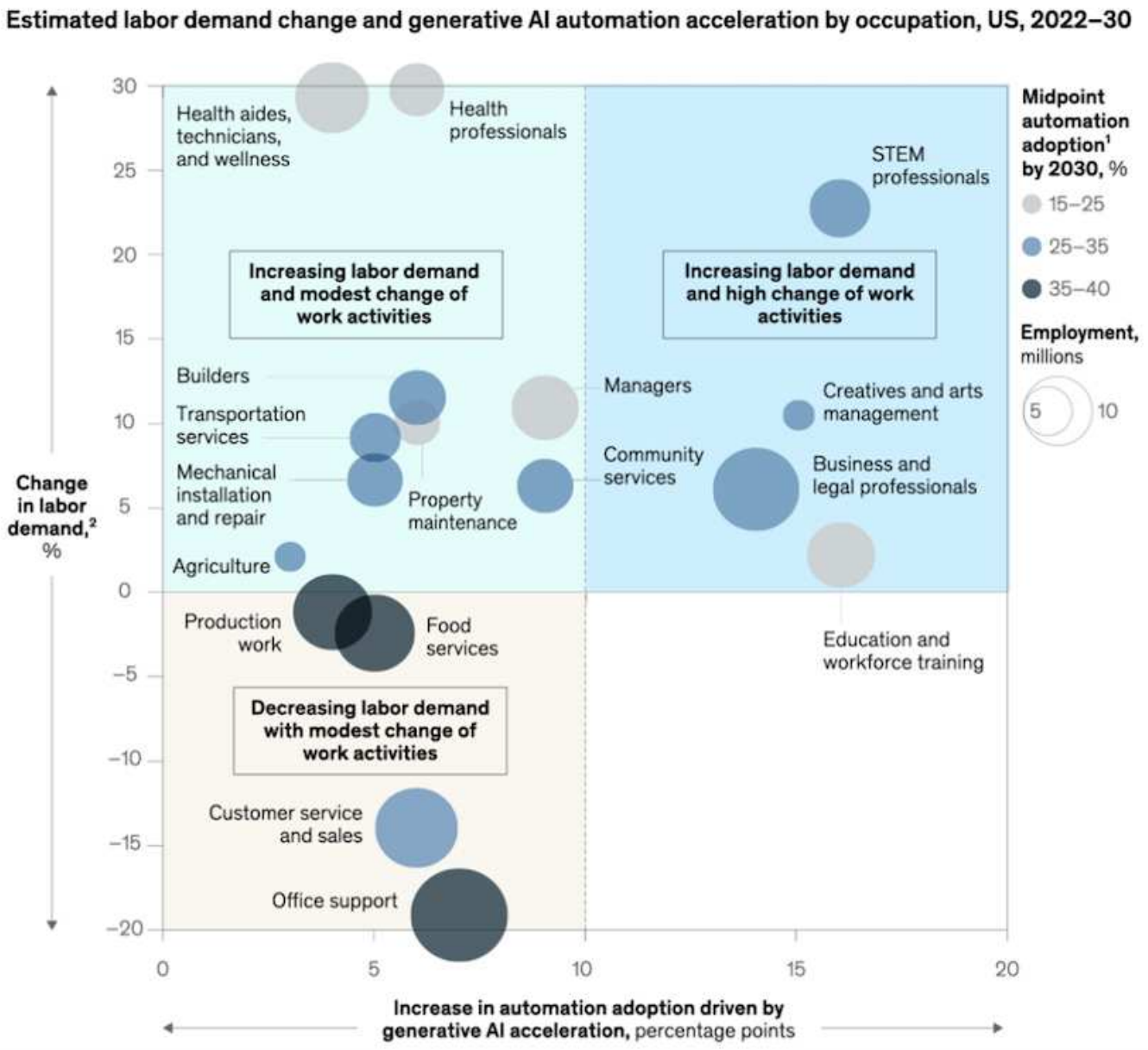
Les entreprises de tous les secteurs d'activité découvrent qu'elles ont de plus en plus de potentiel pour l'IA afin d'extraire et de produire de nouvelles formes de valeur à partir des données existantes pour les opérations métier, les ventes, le marketing et les services juridiques. Selon l'étude de marché IDC (International Data Corporation) sur les cas d'utilisation et les investissements mondiaux d'IA générative, la gestion des connaissances dans le développement de logiciels et la conception de produits doit être la plus impactée, suivie par la création de scénarios pour le marketing et la génération de code pour les développeurs. Dans le domaine de la santé, les organismes de recherche clinique font une percée dans la médecine. Des modèles préformés comme ProteinBERT intègrent des annotations d'ontologie génique (GO) pour concevoir rapidement des structures protéiques pour les médicaments médicaux, ce qui représente une étape importante dans la découverte de médicaments, la bioinformatique et la biologie moléculaire. Les entreprises de biotechnologie ont lancé des essais humains pour la médecine générative découverte par l'IA, qui vise à traiter des maladies comme la fibrose pulmonaire (FPI), une maladie pulmonaire qui provoque une cicatrisation irréversible des tissus pulmonaires.

Figure 1 : cas d’utilisation de l’IA générative

Chatbots	Drug discovery
Text generation	Genome model expression
Image generation	Classification
Code generation	Speech-to-Text

L’adoption croissante de l’automatisation, stimulée par l’IA générative, modifie également l’offre et la demande d’activités professionnelles pour de nombreuses professions. Comme le cabinet de conseil McKinsey, le marché du travail américain (diagramme ci-dessous) a connu une transition rapide, qui ne peut se poursuivre qu’en tenant compte de l’impact de l’IA.

Source : McKinsey & Company



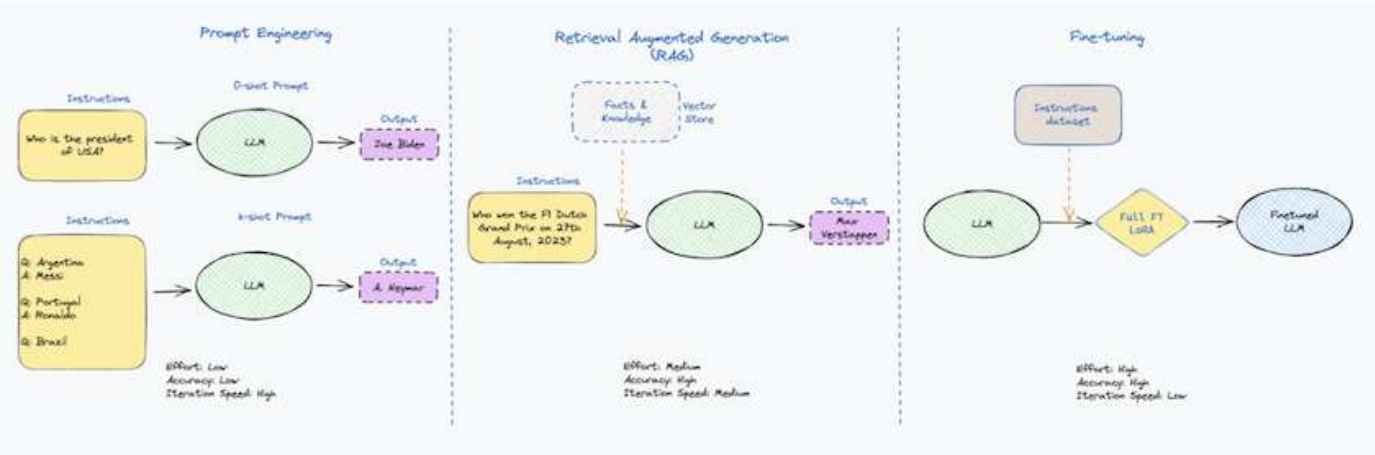
Rôle du stockage dans l'IA générative

Les LLMS reposent essentiellement sur le deep learning, les processeurs graphiques et les ressources de calcul. Cependant, lorsque le tampon GPU est plein, les données doivent être écrites rapidement dans le stockage. Alors que certains modèles d'IA sont suffisamment petits pour s'exécuter en mémoire, les LLMS nécessitent des IOPS élevées et un stockage à haut débit pour fournir un accès rapide à des jeux de données volumineux, en particulier s'ils impliquent des milliards de jetons ou des millions d'images. Pour un besoin en mémoire GPU type d'un LLM, la mémoire nécessaire pour entraîner un modèle avec 1 milliard de paramètres peut atteindre 80 Go avec une précision totale de 32 bits. Dans ce cas, le Llama 2 de Meta, une famille de LLM allant de 7 milliards à 70 milliards de paramètres, peut nécessiter 70x80, environ 5600 Go ou 5,6 To de RAM GPU. En outre, la quantité de mémoire dont vous avez besoin est directement proportionnelle au nombre maximal de jetons que vous souhaitez générer. Par exemple, si vous voulez générer des sorties allant jusqu'à 512 jetons (environ 380 mots), vous avez besoin "512 MO". Cela peut sembler sans conséquence, mais si vous voulez exécuter des lots plus importants, il commence à s'additionner. Par conséquent, le coût de la formation ou du réglage précis des LLM en mémoire est donc très élevé, ce qui fait du stockage la pierre angulaire de l'IA générative.

Trois approches principales des LLMS

Pour la plupart des entreprises, en fonction des tendances actuelles, l'approche de déploiement des LLMS peut être condensée en 3 scénarios de base. Comme décrit dans un récent "« Harvard Business Review »" article : (1) formation (pré-formation) un LLM à partir de zéro – coûteux et nécessitant des compétences spécialisées en IA/ML ; (2) affinement d'un modèle de base avec des données d'entreprise – complexe mais réalisable ; (3) utilisation de la génération augmentée de récupération (RAG) pour interroger des référentiels de documents, des API et des bases de données vectorielles qui contiennent des données d'entreprise. Chacun de ces éléments comporte des compromis entre l'effort, la vitesse d'itération, la rentabilité et la précision du modèle dans sa mise en œuvre, utilisés pour résoudre différents types de problèmes (diagramme ci-dessous).

Figure 3 : types de problèmes



Modèles de base

Un modèle de base également connu sous le nom de modèle de base est un grand modèle d'IA (LLM) entraîné sur de grandes quantités de données non étiquetées, grâce à l'autosurveillance à grande échelle, généralement adapté à un large éventail de tâches NLP en aval. Puisque les données d'entraînement ne sont pas étiquetées par les humains, le modèle émerge plutôt que d'être explicitement encodé. Cela signifie que le modèle peut générer des histoires ou un récit de son propre sans être explicitement programmé pour le faire. Par conséquent, une caractéristique importante de la FM est l'homogénéisation, ce qui signifie que la même méthode est utilisée dans de nombreux domaines. Cependant, avec des techniques de personnalisation et de réglage précis, FMS intégré dans les produits apparaissant ces jours-ci ne sont pas seulement bon pour la génération de texte, texte-images, et texte-à-code, mais aussi pour expliquer des tâches spécifiques de domaine ou de code de débogage. Par exemple, FMS comme le Codex d'OpenAI ou le Code de méta Llama peut générer du code dans plusieurs langages de programmation basés sur les descriptions en langage naturel d'une tâche de programmation. Ces modèles sont compétents dans plus d'une douzaine de langages de programmation, notamment Python, C#, JavaScript, Perl, Ruby, et SQL. Ils comprennent l'intention de l'utilisateur et génèrent du code spécifique qui effectue la tâche souhaitée, utile pour le développement de logiciels, l'optimisation du code et l'automatisation des tâches de programmation.

Affinage, spécificité de domaine et recyclage

L'une des pratiques courantes avec le déploiement LLM après la préparation des données et le pré-traitement des données consiste à sélectionner un modèle pré-entraîné qui a été entraîné sur un dataset volumineux et diversifié. Dans le contexte du réglage précis, il peut s'agir d'un modèle de langage large open source, par exemple "Meta's Llama 2" entraînement sur 70 milliards de paramètres et 2 billions de jetons. Une fois le modèle pré-entraîné sélectionné, l'étape suivante consiste à l'ajuster sur les données spécifiques au domaine. Pour ce faire, il faut ajuster les paramètres du modèle et l'entraîner à l'égard des nouvelles données pour qu'elles s'adaptent à un domaine et à une tâche spécifiques. Par exemple, BloombergGPT, un LLM

propriétaire formé à un large éventail de données financières au service de l'industrie financière. Les modèles spécifiques à un domaine conçus et entraînés pour une tâche spécifique ont généralement une précision et des performances supérieures dans leur portée, mais une faible transférabilité entre les autres tâches ou domaines. Lorsque l'environnement commercial et les données évoluent sur une période donnée, la précision de prévision de la FM pourrait commencer à diminuer par rapport à ses performances pendant les tests. C'est lorsque le recyclage ou l'ajustement du modèle devient crucial. Le recyclage des modèles dans l'IA/ML classique consiste à mettre à jour un modèle de ML déployé avec de nouvelles données généralement réalisées pour éliminer deux types de dérives. (1) dérive du concept – lorsque le lien entre les variables d'entrée et les variables cibles change au fil du temps, puisque la description de ce que nous voulons prédire les changements, le modèle peut produire des prédictions inexactes. (2) dérive des données : survient lorsque les caractéristiques des données d'entrée changent, comme des changements dans les habitudes ou le comportement des clients au fil du temps, et donc l'incapacité du modèle à répondre à de tels changements. De la même manière, le recyclage s'applique aux FMS/LLMS. Cependant, il peut être beaucoup plus coûteux (en millions de dollars), donc pas quelque chose que la plupart des organisations pourraient envisager. Il fait l'objet de recherches actives, qui émergent encore dans le domaine du LLMOps. Ainsi, au lieu de réentraîner, lorsque la dégradation des modèles se produit dans un FMS ajusté, les entreprises peuvent choisir de procéder à des ajustements (beaucoup moins chers) avec un nouveau dataset. Pour ce qui est des coûts, vous trouverez ci-dessous un exemple de tableau des prix des modèles des services Azure-OpenAI. Les clients peuvent ajuster et évaluer les modèles de chaque catégorie de tâche à partir de datasets spécifiques.

Source : Microsoft Azure

Model	**Per 1000 token**
Text-Ada	$0.0001
GPT-3.5 Turbo	$0.003
GPT-4	$0.06
Text-Davinci	$0.02
Model	**Per 100 images**
Dall-E	$2

Ingénierie et inférence rapides

L'ingénierie rapide fait référence aux méthodes efficaces de communication avec les LLMS pour effectuer les tâches souhaitées sans mettre à jour les pondérations du modèle. Si l'entraînement et le réglage précis des modèles d'IA sont essentiels pour les applications NLP, l'inférence est tout aussi importante, là où les modèles entraînés répondent aux invites de l'utilisateur. La configuration système requise pour l'inférence est généralement bien plus importante sur les performances de lecture du système de stockage d'IA qui alimente les données des LLMS vers les GPU lorsqu'il doit pouvoir appliquer des milliards de paramètres de modèle stockés afin de fournir la meilleure réponse possible.

LLMOps, Model Monitoring et Vectorstores

À l'instar des opérations classiques de machine learning (MLOps), les opérations de modèle linguistique grand format (LLMOps) nécessitent également la collaboration de data Scientists et d'ingénieurs DevOps, avec des outils et des bonnes pratiques de gestion des LLM dans des environnements de production. Cependant, le flux de travail et la pile technique des LLMS peuvent varier d'une manière ou d'une autre. Par exemple, les pipelines LLM construits à l'aide de structures comme LangChain String ensemble plusieurs appels API LLM

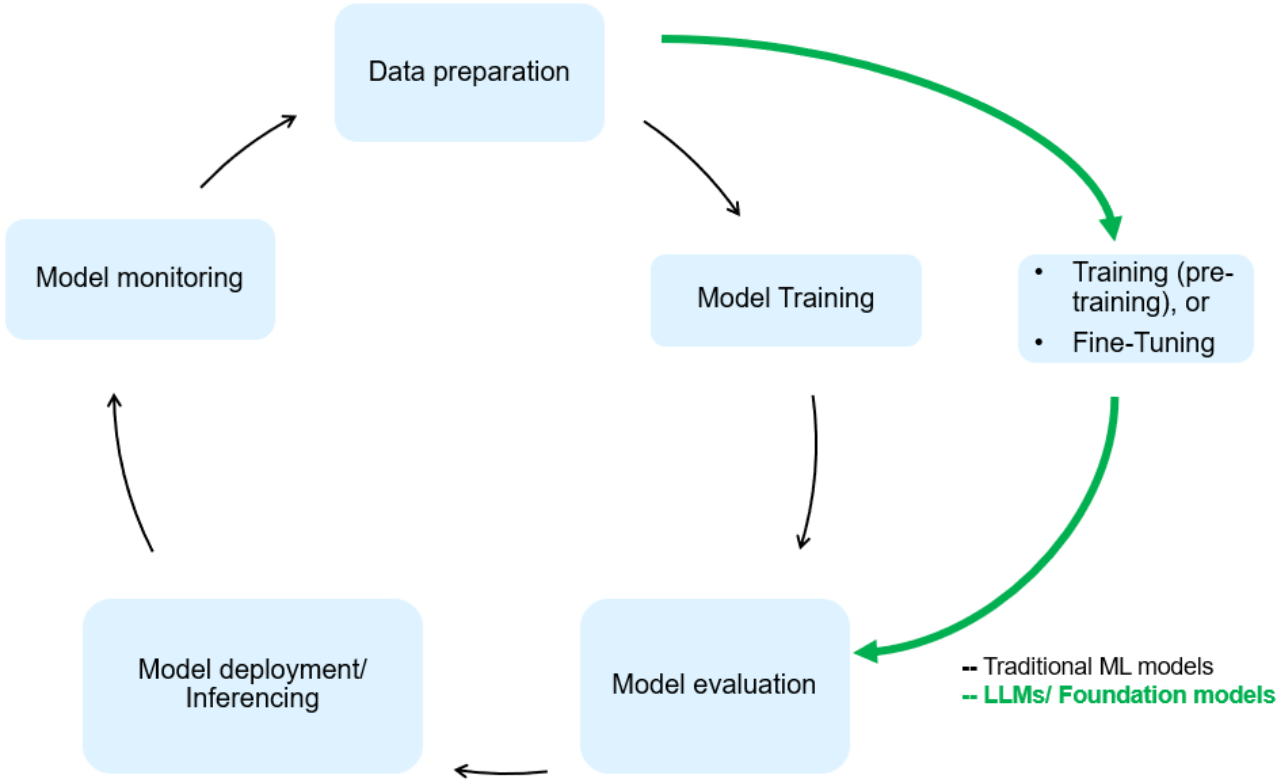
vers des points finaux d'intégration externes tels que des vectorstores ou des bases de données vectorielles. L'utilisation d'un point de terminaison d'intégration et d'un vectorstore pour les connecteurs en aval (comme une base de données vectorielle) représente un développement significatif dans la façon dont les données sont stockées et consultées. Contrairement aux modèles de ML traditionnels développés à partir de zéro, les LLMS s'appuient souvent sur l'apprentissage par transfert puisque ces modèles commencent par FMS qui sont ajustés avec de nouvelles données afin d'améliorer les performances dans un domaine plus spécifique. Par conséquent, il est crucial que les LLMOps offrent les capacités de gestion des risques et de surveillance de la désintégration des modèles.

Risques et éthique à l'ère de l'IA générative

« ChatGPT – c'est un peu ridicule, mais il ne fait pas de sens. »– MIT Tech Review. Le traitement des déchets a toujours été un véritable défi pour l'informatique. La seule différence avec l'IA générative est qu'elle excelle dans la crédibilité des déchets, ce qui conduit à des résultats inexacts. Les LLMS sont enclins à inventer des faits qui correspondent au discours qu'ils construisent. Par conséquent, les entreprises qui considèrent l'IA générative comme une grande opportunité de réduire leurs coûts avec des équivalents en IA doivent détecter efficacement les faux-pas, réduire les préjugés et réduire les risques pour que les systèmes restent honnêtes et éthiques. Un pipeline de données fluide, doté d'une infrastructure d'IA robuste, qui prend en charge la mobilité des données, la qualité, la gouvernance et la protection des données via un chiffrement de bout en bout et des rambardes d'IA, est un éminent dans la conception de modèles d'IA génératifs responsables et explicables.

Scénario client et NetApp

Figure 3 : workflow de modèle d'apprentissage machine/langage large



Sommes-nous en train de former ou de peaufiner? la question de savoir si (a) former un modèle LLM à partir de zéro, peaufiner un FM pré-entraîné, ou utiliser RAG pour extraire des données des référentiels de documents en dehors d'un modèle de base et augmenter les invites, et (b) soit en exploitant des LLMS open source (par exemple, Llama 2), soit en utilisant des FMS propriétaires (par exemple, ChatGPT, Bard, AWS

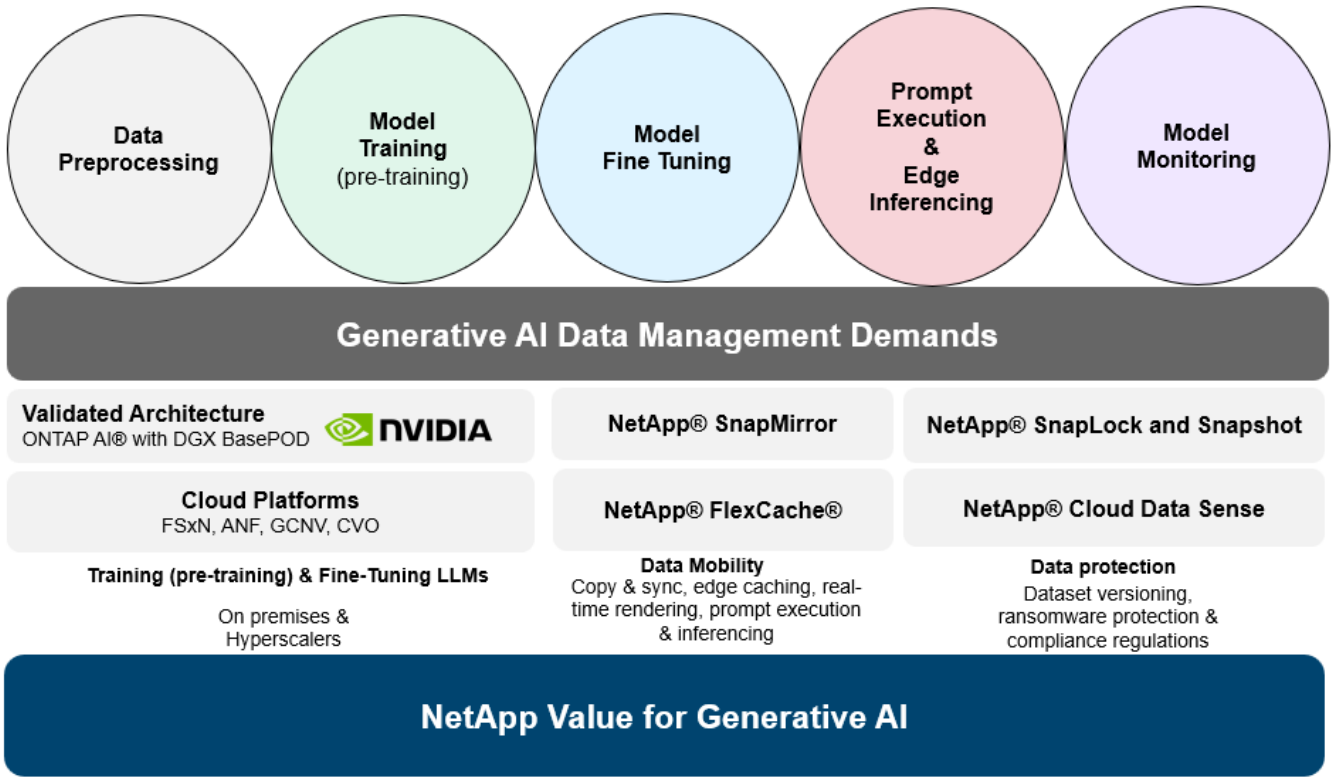
Bedrock) est une décision stratégique pour les entreprises. Chaque approche implique un compromis entre la rentabilité, la gravité des données, les opérations, la précision des modèles et la gestion des LLMS.

En tant qu'entreprise, NetApp adopte l'IA en interne dans sa culture du travail et dans son approche de la conception de produits et de ses efforts d'ingénierie. Par exemple, la protection anti-ransomware autonome de NetApp est conçue à l'aide de l'IA et du machine learning. Il assure une détection précoce des anomalies du système de fichiers pour aider à identifier les menaces avant qu'elles n'affectent les opérations. Deuxièmement, NetApp utilise l'IA prédictive pour ses activités commerciales, telles que la prévision des ventes et des stocks et les chatbots, pour aider ses clients dans les services de support produit des centres d'appels, les spécifications techniques, la garantie, les manuels de service, etc. Troisièmement, NetApp apporte une valeur ajoutée au pipeline de données d'IA et au workflow ML/LLM via des produits et des solutions destinés aux clients qui conçoivent des solutions d'IA prédictives telles que la prévision de la demande, l'imagerie médicale, l'analyse des sentiments, et des solutions d'IA générative telles que Gans pour la détection des anomalies d'images industrielles dans le secteur manufacturier et la lutte contre le blanchiment d'argent et la détection des fraudes dans les services bancaires et financiers avec des produits et des fonctionnalités NetApp tels que NetApp® ONTAP ai®, NetApp® SnapMirror® et NetApp® FlexCache®.

Fonctionnalités NetApp

Le déplacement et la gestion des données dans les applications d'IA générative telles que le chatbot, la génération de code, la génération d'images ou l'expression de modèle génomique peuvent s'étendre de la périphérie, au data Center privé et à l'écosystème multicloud hybride. Par exemple, un robot d'IA en temps réel qui aide un passager à mettre à niveau son billet d'avion pour passer à une classe affaires à partir d'une application utilisateur exposée via des API de modèles pré-entraînés tels que ChatGPT ne peut pas accomplir cette tâche en lui-même puisque les informations sur les passagers ne sont pas accessibles au public sur Internet. L'API requiert l'accès aux informations personnelles du passager et aux informations sur le billet de la compagnie aérienne, qui peuvent exister dans un écosystème hybride ou multicloud. Un scénario similaire peut s'appliquer aux scientifiques partageant une molécule de médicament et les données des patients via une application utilisateur qui utilise les LLM pour effectuer des essais cliniques dans le cadre de la découverte de médicaments impliquant un à plusieurs établissements de recherche biomédicale. Les données sensibles transmises au FMS ou au LLMS peuvent inclure des PII, des informations financières, des informations de santé, des données biométriques, des données de localisation, données de communication, comportement en ligne et informations juridiques. Dans ce cas, le rendu en temps réel, l'exécution rapide et l'inférence de périphérie font passer les données de l'application utilisateur final aux terminaux de stockage via des modèles LLM propriétaires ou open source vers un data Center sur des plateformes de cloud public ou sur site. Dans tous ces scénarios, la mobilité et la protection des données sont essentielles pour les opérations d'IA impliquant des LLMS qui s'appuient sur d'importants datasets d'entraînement et le déplacement de ces données.

Figure 4 : pipeline de traitement de données LLM - IA générative



Le portefeuille NetApp d'infrastructures de stockage, de données et de services cloud est optimisé par un logiciel de gestion intelligente des données.

Préparation des données : le premier pilier de la pile technologique LLM est largement intact de la pile traditionnelle ML plus ancienne. Dans le pipeline d'IA, un prétraitement des données est nécessaire pour les normaliser et les nettoyer avant de procéder à l'entraînement ou aux réglages fins. Cette étape comprend des connecteurs qui permettent d'acquérir les données où qu'elles résident sous la forme d'un Tier Amazon S3 ou dans des systèmes de stockage sur site tels qu'un magasin de fichiers ou d'objets comme NetApp StorageGRID.

NetApp® ONTAP est la technologie de base sur laquelle reposent les solutions de stockage stratégiques de NetApp dans le data Center et le cloud. ONTAP comprend plusieurs fonctionnalités et fonctions de gestion et de protection des données, notamment la protection automatique contre les ransomware contre les cyberattaques, le transport intégré des données et l'efficacité du stockage pour toute une gamme d'architectures : sur site, hybride, multiclouds dans NAS, SAN, objet, et SDS (Software Defined Storage) dans les déploiements LLM.

NetApp® ONTAP ai® pour l'entraînement des modèles de deep learning. NetApp® ONTAP® prend en charge NVIDIA GPU Direct Storage™ en utilisant NFS over RDMA pour les clients NetApp dotés d'un cluster de stockage ONTAP et de nœuds de calcul NVIDIA DGX . Elle offre des performances rentables pour lire et traiter plusieurs fois les jeux de données source du stockage dans la mémoire afin de favoriser l'intelligence, ce qui permet aux entreprises de bénéficier d'une formation, d'un réglage précis et d'un accès évolutif aux LLMS.

NetApp® FlexCache® est une capacité de mise en cache à distance qui simplifie la distribution des fichiers et met uniquement en cache les données lues activement. Cela peut s'avérer utile pour l'entraînement, la réentraînement et le réglage précis du LLM, ce qui apporte de la valeur aux clients ayant des exigences métier telles que le rendu en temps réel et l'inférence LLM.

NetApp® SnapMirror est une fonctionnalité ONTAP qui réplique les snapshots de volumes entre deux systèmes ONTAP. Cette fonctionnalité transfère de manière optimale les données de la périphérie vers votre data Center sur site ou dans le cloud. SnapMirror permet de déplacer des données efficacement et en toute

sécurité entre les clouds des hyperscalers et sur site lorsque les clients souhaitent développer une IA générative dans des clouds avec RAG contenant des données d'entreprise. Il ne transfère efficacement que les modifications, ce qui permet d'économiser de la bande passante et d'accélérer la réplication, ce qui apporte des fonctionnalités essentielles de mobilité des données lors des opérations de formation, de réentraînement et de réglage précis des FMS ou des LLMS.

NetApp® SnapLock apporte une capacité de disque immuable sur les systèmes de stockage ONTAP pour la gestion des versions de Datasets. L'architecture des microcœurs est conçue pour protéger les données des clients à l'aide du moteur « zéro confiance » FPolicy™. NetApp garantit la disponibilité des données client en résistant aux attaques par déni de service (dos) lorsqu'un attaquant interagit avec un LLM de manière particulièrement consommatrice de ressources.

NetApp® Cloud Data Sense permet d'identifier, de cartographier et de classer les informations personnelles présentes dans les jeux de données d'entreprise, d'adopter des politiques, de répondre aux exigences de confidentialité sur site ou dans le cloud, d'améliorer la sécurité et de se conformer aux réglementations.

Classification NetApp® BlueXP™, optimisée par Cloud Data Sense. Les clients peuvent automatiquement analyser, analyser, catégoriser et agir sur les données dans l'intégralité de données, détecter les risques pour la sécurité, optimiser le stockage et accélérer les déploiements cloud. Cette solution combine les services de stockage et de données via son plan de contrôle unifié, et permet d'utiliser les instances GPU pour le calcul, ainsi que les environnements multiclouds hybrides pour le Tiering de stockage à froid, l'archivage et les sauvegardes.

NetApp Dualité fichier-objet. NetApp ONTAP permet un accès à double protocole pour NFS et S3. Avec cette solution, les clients peuvent accéder aux données NFS à partir des ordinateurs portables Amazon AWS SageMaker via des compartiments S3 de NetApp Cloud Volumes ONTAP. Les clients ont besoin d'un accès simplifié aux sources de données hétérogènes et de la possibilité de partager des données à partir de NFS et S3. Par exemple, affinage des modèles de génération de texte FMS tels que Llama 2 de Meta sur SageMaker avec accès aux compartiments fichier-objet.

Le service NetApp® Cloud Sync offre un moyen simple et sécurisé de migrer des données vers n'importe quelle cible, dans le cloud ou sur site. Cloud Sync transfère et synchronise de manière transparente les données entre le stockage sur site ou dans le cloud, le stockage NAS et les magasins d'objets.

NetApp XCP est un logiciel client qui permet des migrations de données any-to-NetApp et NetApp-to-NetApp rapides et fiables. XCP permet également de déplacer efficacement des données en bloc depuis des systèmes de fichiers Hadoop HDFS vers ONTAP NFS, S3 ou StorageGRID et l'analytique de fichiers XCP pour assurer une meilleure visibilité sur le système de fichiers.

NetApp® DataOps Toolkit est une bibliothèque Python qui permet aux data Scientists, aux équipes DevOps et aux ingénieurs de données d'effectuer facilement diverses tâches de gestion des données, telles que le provisionnement quasi instantané, le clonage ou la création de snapshots de volumes de données ou d'espaces de travail JupyterLab bénéficiant d'un stockage NetApp hautes performances à évolutivité horizontale.

Sécurité des produits NetApp. Les LLMS peuvent par inadvertance révéler des données confidentielles dans leurs réponses, ce qui constitue un problème pour les RSSI qui étudient les vulnérabilités associées aux applications d'IA exploitant les LLMS. Comme le souligne le projet OWASP (Open Worldwide application Security Project), les problèmes de sécurité tels que l'empoisonnement de données, la fuite de données, le déni de service et les injections rapides au sein des LLMS peuvent avoir un impact sur les entreprises, de l'exposition aux données à des accès non autorisés aux attaquants. Les exigences en matière de stockage des données doivent inclure des contrôles d'intégrité et des snapshots immuables pour les données structurées, semi-structurées et non structurées. Les copies Snapshot NetApp et SnapLock sont utilisées pour la gestion des versions des datasets. Elle inclut un contrôle d'accès basé sur des rôles (RBAC) strict, ainsi que des protocoles sécurisés et un chiffrement standard pour sécuriser les données au repos et en transit.

Ensemble, Cloud Insights et Cloud Data Sense offrent des fonctionnalités qui vous aident à identifier la source de la menace d'un point de vue judiciaire et à hiérarchiser les données à restaurer.

ONTAP ai avec DGX BasePOD

L'architecture de référence NetApp® ONTAP® ai avec NVIDIA DGX BasePOD est une architecture évolutive pour les workloads de machine learning (ML) et d'intelligence artificielle (IA). Pour la phase d'entraînement critique des LLMS, les données sont généralement copiées à intervalles réguliers du stockage de données vers le cluster d'apprentissage. Les serveurs utilisés lors de cette phase ont recours à des GPU pour paralléliser les calculs, dont l'appétit en données est gigantesque. Il est essentiel de répondre aux besoins en bande passante d'E/S brute pour maintenir un taux élevé d'utilisation des GPU.

ONTAP ai avec NVIDIA ai Enterprise

NVIDIA ai Enterprise est une suite logicielle cloud complète d'IA et d'analytique de données optimisée, certifiée et prise en charge par NVIDIA pour s'exécuter sur VMware vSphere avec les systèmes NVIDIA certifiés. Ce logiciel facilite le déploiement, la gestion et l'évolutivité simples et rapides des workloads d'IA dans un environnement de cloud hybride moderne. La solution NVIDIA ai Enterprise, optimisée par NetApp et VMware, fournit un workload d'IA haute performance et une gestion des données dans un pack simplifié et familier.

1P plateformes cloud

Les offres de stockage cloud entièrement gérées sont disponibles de manière native sur Microsoft Azure en tant qu'Azure NetApp Files (ANF), sur AWS en tant qu'Amazon FSX pour NetApp ONTAP (FSxN) et sur Google Cloud NetApp volumes (GNCV). 1P est un système de fichiers géré haute performance qui permet aux clients d'exécuter des workloads d'IA hautement disponibles avec une meilleure sécurité des données dans les clouds publics. Il permet d'affiner les LLMS/FMS avec des plateformes de ML cloud natives comme AWS SageMaker, Azure-OpenAI Services et Google Vertex ai.

Suite de solutions partenaires NetApp

Outre ses produits, ses technologies et ses fonctionnalités de base pour les données, NetApp collabore étroitement avec un solide réseau de partenaires d'IA afin d'apporter une valeur ajoutée aux clients.

Les rails NVIDIA Guardrails dans les systèmes d'IA servent de protection pour assurer l'utilisation éthique et responsable des technologies d'IA. Les développeurs d'IA peuvent choisir de définir le comportement des applications LLM sur des sujets spécifiques et l'empêcher de s'engager dans des discussions sur des sujets indésirables. Guardrails, un kit d'outils open-source, permet de connecter un LLM à d'autres services de manière transparente et sécurisée pour créer des systèmes conversationnels LLM fiables, sûrs et sécurisés.

Domino Data Lab fournit des outils polyvalents, professionnels pour la construction et la production de Generative ai - rapide, sûr et économique, où que vous soyez dans votre parcours d'IA. Avec la plateforme Enterprise MLOps de Domino, les data Scientists peuvent utiliser les outils préférés et toutes leurs données, entraîner et déployer facilement des modèles en tout lieu et gérer les risques et les coûts de manière rentable, le tout à partir d'un seul centre de contrôle.

Modzy pour Edge ai. NetApp® et Modzy se sont associés pour fournir l'IA à grande échelle à tous les types de données, y compris les images, l'audio, le texte et les tableaux. Modzy est une plateforme MLOps pour le déploiement, l'intégration et l'exécution de modèles d'IA. Elle offre aux data Scientists des fonctionnalités de surveillance des modèles, de détection des écarts et d'explication, avec une solution intégrée pour l'inférence LLM transparente.

Run:ai et NetApp se sont associés pour démontrer les fonctionnalités uniques de la solution NetApp ONTAP ai avec la plateforme de gestion des clusters Run:ai afin de simplifier l'orchestration des workloads d'IA. Cette

solution divise et joint automatiquement les ressources GPU, afin de faire évoluer vos pipelines de traitement de données vers des centaines de machines dotées de structures d'intégration intégrées pour Spark, Ray, Dask et Rapids.

Conclusion

L'IA générative ne peut produire des résultats efficaces que lorsque le modèle est entraîné sur des volumes de données de qualité. Bien que les LLMS aient atteint des jalons remarquables, il est essentiel de reconnaître ses limites, ses défis en matière de conception et les risques associés à la mobilité et à la qualité des données. Les LLMS s'appuient sur des datasets d'entraînement volumineux et disparates provenant de sources de données hétérogènes. Des résultats inexacts ou des résultats biaisés générés par les modèles peuvent mettre en péril à la fois les entreprises et les consommateurs. Ces risques peuvent correspondre à des contraintes pour les LLMS qui émergent potentiellement des défis liés à la gestion des données associés à la qualité, à la sécurité des données et à la mobilité des données. NetApp aide les entreprises à relever les défis liés à la croissance rapide des données, à la mobilité des données, à la gestion multicloud et à l'adoption de l'IA. L'utilisation d'une infrastructure d'IA à grande échelle et d'une gestion efficace des données sont essentielles à la réussite des applications d'IA telles que l'IA générative. Il est essentiel que nos clients couvrent tous les scénarios de déploiement sans compromettre la capacité à se développer en fonction des besoins de l'entreprise tout en contrôlant les coûts, la gouvernance des données et les pratiques éthiques liées à l'IA. NetApp travaille en permanence pour aider les clients à simplifier et à accélérer leurs déploiements d'IA.

Tr-4785 : déploiement de l'IA avec NetApp E-Series et BeeGFS

Nagalakshmi Radu, Daniel Landes, Nathan Swartz, Amine Bennani, NetApp

Les applications d'intelligence artificielle (IA), de machine learning (ML) et de deep learning (DL) impliquent des datasets volumineux et des calculs importants. Pour exécuter correctement ces charges de travail, vous avez besoin d'une infrastructure agile qui vous permet de faire évoluer horizontalement les nœuds de calcul et de stockage de manière transparente. Ce rapport comprend des étapes pour exécuter un modèle d'entraînement d'IA dans un mode distribué, qui permet une évolutivité scale-out transparente des nœuds de calcul et de stockage. Il inclut également plusieurs metrics de performance pour démontrer comment une solution alliant le stockage NetApp E-Series et le système de fichiers parallèle BeeGFS fournit une solution simple, flexible et économique pour les workloads d'IA.

"Tr-4785 : déploiement de l'IA avec NetApp E-Series et BeeGFS"

NVA-1150-DESIGN : guide de design Quantum StorNext avec les systèmes NetApp E-Series

Ryan rodine, NetApp

Ce document explique en détail comment concevoir une solution de systèmes de fichiers parallèles StorNext avec les systèmes de stockage NetApp E-Series. Cette solution présente les baies 100 % Flash NetApp EF280, les baies NVMe 100 % Flash NetApp EF300, les baies NVMe 100 % Flash EF600 et les systèmes hybrides NetApp E5760. Il offre une caractérisation des performances basée sur le banc d'essai Frametest, un outil

largement utilisé pour les tests dans l'industrie des médias et du divertissement.

"NVA-1150-DESIGN : guide de design Quantum StorNext avec les systèmes NetApp E-Series"

NVA-1150-DEPLOY : guide de déploiement Quantum StorNext avec les systèmes NetApp E-Series

Ryan rodine, NetApp

Ce document explique en détail comment déployer une solution de systèmes de fichiers parallèles StorNext avec les systèmes de stockage NetApp E-Series. Cette solution englobe les baies 100 % Flash NetApp EF280, la baie NVMe 100 % Flash NetApp EF300, la baie NVMe 100 % Flash EF600 de NetApp et le système hybride NetApp E5760. Il offre une caractérisation des performances basée sur le banc d'essai Frametest, un outil largement utilisé pour les tests dans l'industrie des médias et du divertissement.

"NVA-1150-DEPLOY : guide de déploiement Quantum StorNext avec les systèmes NetApp E-Series"

Informations sur le copyright

Copyright © 2024 NetApp, Inc. Tous droits réservés. Imprimé aux États-Unis. Aucune partie de ce document protégé par copyright ne peut être reproduite sous quelque forme que ce soit ou selon quelque méthode que ce soit (graphique, électronique ou mécanique, notamment par photocopie, enregistrement ou stockage dans un système de récupération électronique) sans l'autorisation écrite préalable du détenteur du droit de copyright.

Les logiciels dérivés des éléments NetApp protégés par copyright sont soumis à la licence et à l'avis de non-responsabilité suivants :

CE LOGICIEL EST FOURNI PAR NETAPP « EN L'ÉTAT » ET SANS GARANTIES EXPRESSES OU TACITES, Y COMPRIS LES GARANTIES TACITES DE QUALITÉ MARCHANDE ET D'ADÉQUATION À UN USAGE PARTICULIER, QUI SONT EXCLUES PAR LES PRÉSENTES. EN AUCUN CAS NETAPP NE SERA TENU POUR RESPONSABLE DE DOMMAGES DIRECTS, INDIRECTS, ACCESSOIRES, PARTICULIERS OU EXEMPLAIRES (Y COMPRIS L'ACHAT DE BIENS ET DE SERVICES DE SUBSTITUTION, LA PERTE DE JOUISSANCE, DE DONNÉES OU DE PROFITS, OU L'INTERRUPTION D'ACTIVITÉ), QUELLES QU'EN SOIENT LA CAUSE ET LA DOCTRINE DE RESPONSABILITÉ, QU'IL S'AGISSE DE RESPONSABILITÉ CONTRACTUELLE, STRICTE OU DÉLICTUELLE (Y COMPRIS LA NÉGLIGENCE OU AUTRE) DÉCOULANT DE L'UTILISATION DE CE LOGICIEL, MÊME SI LA SOCIÉTÉ A ÉTÉ INFORMÉE DE LA POSSIBILITÉ DE TELS DOMMAGES.

NetApp se réserve le droit de modifier les produits décrits dans le présent document à tout moment et sans préavis. NetApp décline toute responsabilité découlant de l'utilisation des produits décrits dans le présent document, sauf accord explicite écrit de NetApp. L'utilisation ou l'achat de ce produit ne concède pas de licence dans le cadre de droits de brevet, de droits de marque commerciale ou de tout autre droit de propriété intellectuelle de NetApp.

Le produit décrit dans ce manuel peut être protégé par un ou plusieurs brevets américains, étrangers ou par une demande en attente.

LÉGENDE DE RESTRICTION DES DROITS : L'utilisation, la duplication ou la divulgation par le gouvernement sont sujettes aux restrictions énoncées dans le sous-paragraphe (b)(3) de la clause Rights in Technical Data-Noncommercial Items du DFARS 252.227-7013 (février 2014) et du FAR 52.227-19 (décembre 2007).

Les données contenues dans les présentes se rapportent à un produit et/ou service commercial (tel que défini par la clause FAR 2.101). Il s'agit de données propriétaires de NetApp, Inc. Toutes les données techniques et tous les logiciels fournis par NetApp en vertu du présent Accord sont à caractère commercial et ont été exclusivement développés à l'aide de fonds privés. Le gouvernement des États-Unis dispose d'une licence limitée irrévocable, non exclusive, non cessible, non transférable et mondiale. Cette licence lui permet d'utiliser uniquement les données relatives au contrat du gouvernement des États-Unis d'après lequel les données lui ont été fournies ou celles qui sont nécessaires à son exécution. Sauf dispositions contraires énoncées dans les présentes, l'utilisation, la divulgation, la reproduction, la modification, l'exécution, l'affichage des données sont interdits sans avoir obtenu le consentement écrit préalable de NetApp, Inc. Les droits de licences du Département de la Défense du gouvernement des États-Unis se limitent aux droits identifiés par la clause 252.227-7015(b) du DFARS (février 2014).

Informations sur les marques commerciales

NETAPP, le logo NETAPP et les marques citées sur le site http://www.netapp.com/TM sont des marques déposées ou des marques commerciales de NetApp, Inc. Les autres noms de marques et de produits sont des marques commerciales de leurs propriétaires respectifs.