NVIDIA AI Enterprise avec NetApp et VMware

NetApp artificial intelligence solutions

NetApp
August 18, 2025

This PDF was generated from https://docs.netapp.com/fr-fr/netapp-solutions-ai/software/ai-nvaie-vmw-intro.html on August 18, 2025. Always check docs.netapp.com for the latest.

Sommaire

NVIDIA AI Enterprise avec NetApp et VMware . . . 1
NVIDIA AI Enterprise avec NetApp et VMware . . . 1
Aperçu de la technologie . . . 1
NVIDIA AI Enterprise . . . 2
NVIDIA GPU Cloud (NGC) . . . 2
VMware vSphere . . . 2
NetApp ONTAP . . . 2
Boîte à outils NetApp DataOps . . . 3
Architecture . . . 4
Configuration initiale . . . 5
Prérequis . . . 5
Installer le logiciel hôte NVIDIA AI Enterprise . . . 6
Utiliser le logiciel NVIDIA NGC . . . 6
Installation . . . 6
Exemple de cas d’utilisation : tâche d’entraînement TensorFlow . . . 8
Où trouver des informations supplémentaires . . . 10
Remerciements . . . 10

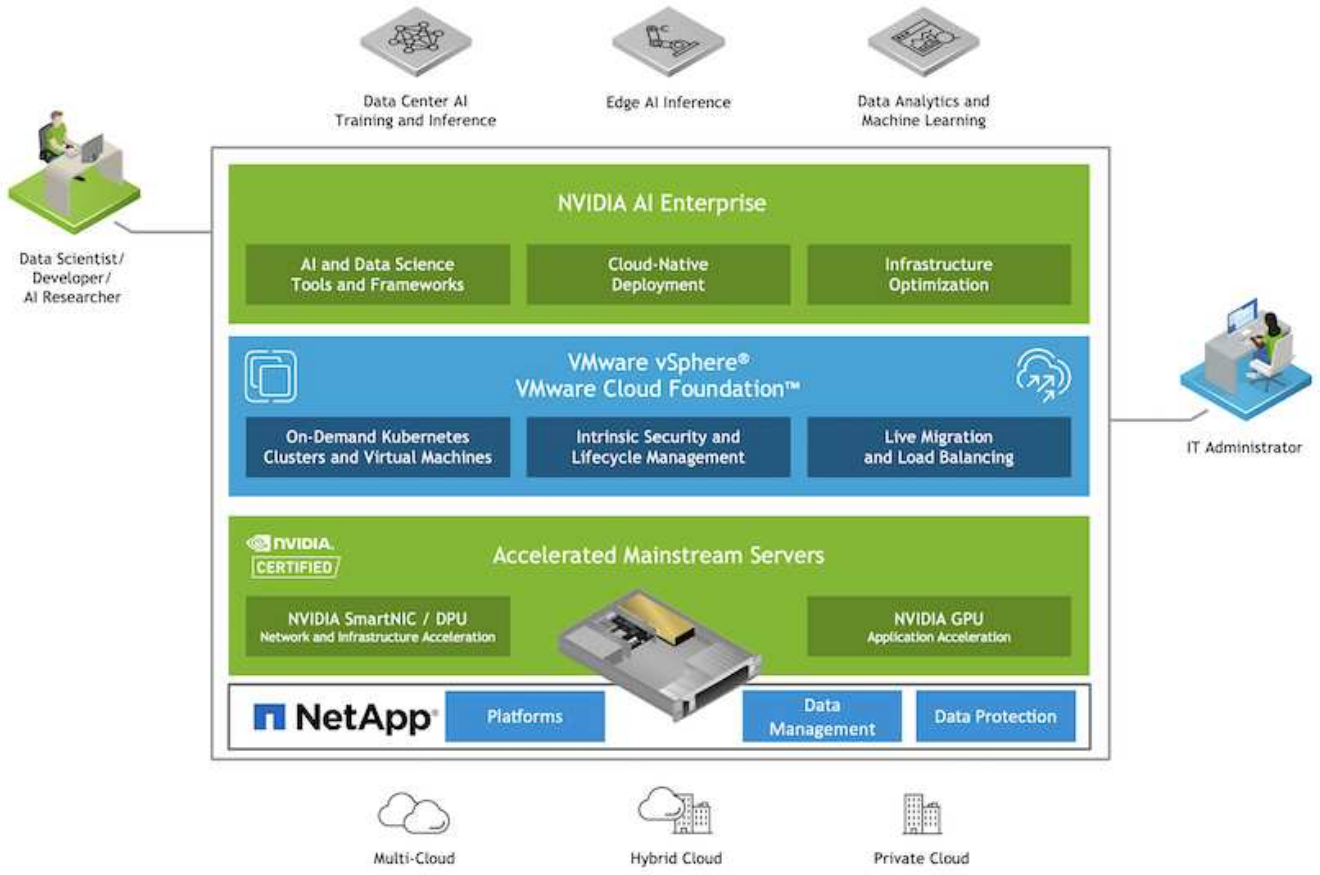
NVIDIA AI Enterprise avec NetApp et VMware

NVIDIA AI Enterprise avec NetApp et VMware

Mike Oglesby, NetApp

Pour les architectes et les administrateurs informatiques, les outils d'IA peuvent être compliqués et peu familiers. De plus, de nombreuses plateformes d'IA ne sont pas adaptées aux entreprises. NVIDIA AI Enterprise, optimisé par NetApp et VMware, a été créé pour fournir une architecture d'IA rationalisée et de classe entreprise.

NVIDIA AI Enterprise est une suite de logiciels d'IA et d'analyse de données cloud native de bout en bout, optimisée, certifiée et prise en charge par NVIDIA pour s'exécuter sur VMware vSphere avec les systèmes certifiés NVIDIA. Ce logiciel facilite le déploiement, la gestion et la mise à l'échelle simples et rapides des charges de travail d'IA dans l'environnement cloud hybride moderne. NVIDIA AI Enterprise, optimisé par NetApp et VMware, offre une gestion des charges de travail et des données d'IA de niveau entreprise dans un package simplifié et familier.



Aperçu de la technologie

Cette section fournit un aperçu technologique de NVIDIA AI Enterprise avec NetApp et VMware.

NVIDIA AI Enterprise

NVIDIA AI Enterprise est une suite de logiciels d'IA et d'analyse de données cloud native de bout en bout, optimisée, certifiée et prise en charge par NVIDIA pour s'exécuter sur VMware vSphere avec les systèmes certifiés NVIDIA. Ce logiciel facilite le déploiement, la gestion et la mise à l'échelle simples et rapides des charges de travail d'IA dans l'environnement cloud hybride moderne.

NVIDIA GPU Cloud (NGC)

NVIDIA NGC héberge un catalogue de logiciels optimisés pour GPU permettant aux praticiens de l'IA de développer leurs solutions d'IA. Il donne également accès à divers services d'IA, notamment NVIDIA Base Command pour la formation des modèles, NVIDIA Fleet Command pour déployer et surveiller les modèles, et le registre privé NGC pour accéder et gérer en toute sécurité les logiciels d'IA propriétaires. Les clients NVIDIA AI Enterprise peuvent également demander de l'assistance via le portail NGC.

VMware vSphere

VMware vSphere est la plateforme de virtualisation de VMware, qui transforme les centres de données en infrastructures informatiques agrégées incluant des ressources CPU, de stockage et réseau. vSphere gère ces infrastructures comme un environnement d'exploitation unifié et fournit aux administrateurs les outils nécessaires pour gérer les centres de données qui y participent.

Les deux composants principaux de vSphere sont ESXi et vCenter Server. ESXi est la plateforme de virtualisation sur laquelle les administrateurs créent et exécutent des machines virtuelles et des appliances virtuelles. vCenter Server est le service qui leur permet de gérer plusieurs hôtes connectés à un réseau et de regrouper les ressources des hôtes.

NetApp ONTAP

ONTAP 9, la dernière génération de logiciel de gestion du stockage de NetApp, permet aux entreprises de moderniser leur infrastructure et de passer à un centre de données prêt pour le cloud. En s'appuyant sur des capacités de gestion de données de pointe, ONTAP permet la gestion et la protection des données avec un seul ensemble d'outils, quel que soit l'endroit où résident ces données. Vous pouvez également déplacer librement les données là où elles sont nécessaires : vers la périphérie, le cœur ou le cloud. ONTAP 9 inclut de nombreuses fonctionnalités qui simplifient la gestion des données, accélèrent et protègent les données critiques et permettent des capacités d'infrastructure de nouvelle génération dans les architectures de cloud hybride.

Simplifier la gestion des données

La gestion des données est essentielle pour les opérations informatiques de l'entreprise et les scientifiques des données afin que les ressources appropriées soient utilisées pour les applications d'IA et la formation des ensembles de données d'IA/ML. Les informations supplémentaires suivantes sur les technologies NetApp ne sont pas couvertes par cette validation, mais 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 votre coût total d'exploitation :

- Compactage des données en ligne et déduplication étendue. La compaction des données réduit l'espace gaspillé à l'intérieur des blocs de stockage et la déduplication augmente considérablement la capacité effective. Cela s'applique aux données stockées localement et aux données hiérarchisées vers le cloud.
- Qualité de service minimale, maximale et adaptative (AQoS). Les contrôles granulaires de qualité de service (QoS) aident à maintenir les niveaux de performances des applications critiques dans les

environnements hautement partagés.

- FabricPool NetApp . Fournit une hiérarchisation automatique des données froides vers des options de stockage cloud publiques et privées, notamment Amazon Web Services (AWS), Azure et la solution de stockage NetApp StorageGRID . Pour plus d'informations sur FabricPool, voir "TR-4598 : Bonnes pratiques FabricPool" .

Accélérer et protéger les données

ONTAP offre des niveaux supérieurs de performance et de protection des données et étend ces capacités des manières suivantes :

- Performances et latence réduite. ONTAP offre le débit le plus élevé possible avec 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.
- Chiffrement de volume NetApp (NVE). ONTAP offre un cryptage natif au niveau du volume avec prise en charge de la gestion des clés intégrée et externe.
- Authentification multi-locataire et multifactorielle. ONTAP permet le partage des ressources d'infrastructure avec les plus hauts niveaux de sécurité.

Une infrastructure à l'épreuve du temps

ONTAP permet de répondre aux besoins commerciaux exigeants et en constante évolution grâce aux fonctionnalités suivantes :

- Mise à l'échelle transparente et opérations non perturbatrices. ONTAP prend en charge l'ajout non perturbateur de capacité aux contrôleurs existants et aux clusters évolutifs. Les clients peuvent passer aux dernières technologies, telles que NVMe et FC 32 Go, sans migrations de données ni pannes coûteuses.
- Connexion au Cloud. ONTAP est le logiciel de gestion de stockage le plus connecté au cloud, avec des options de stockage défini par logiciel (ONTAP Select) et des instances cloud natives (Google Cloud NetApp Volumes) dans tous les clouds publics.
- Intégration avec les applications émergentes. ONTAP propose des services de données de niveau 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.

Boîte à outils NetApp DataOps

NetApp DataOps Toolkit est un outil basé sur Python qui simplifie la gestion des espaces de travail de développement/formation et des serveurs d'inférence soutenus par un stockage NetApp hautes performances et évolutif. Les principales fonctionnalités comprennent :

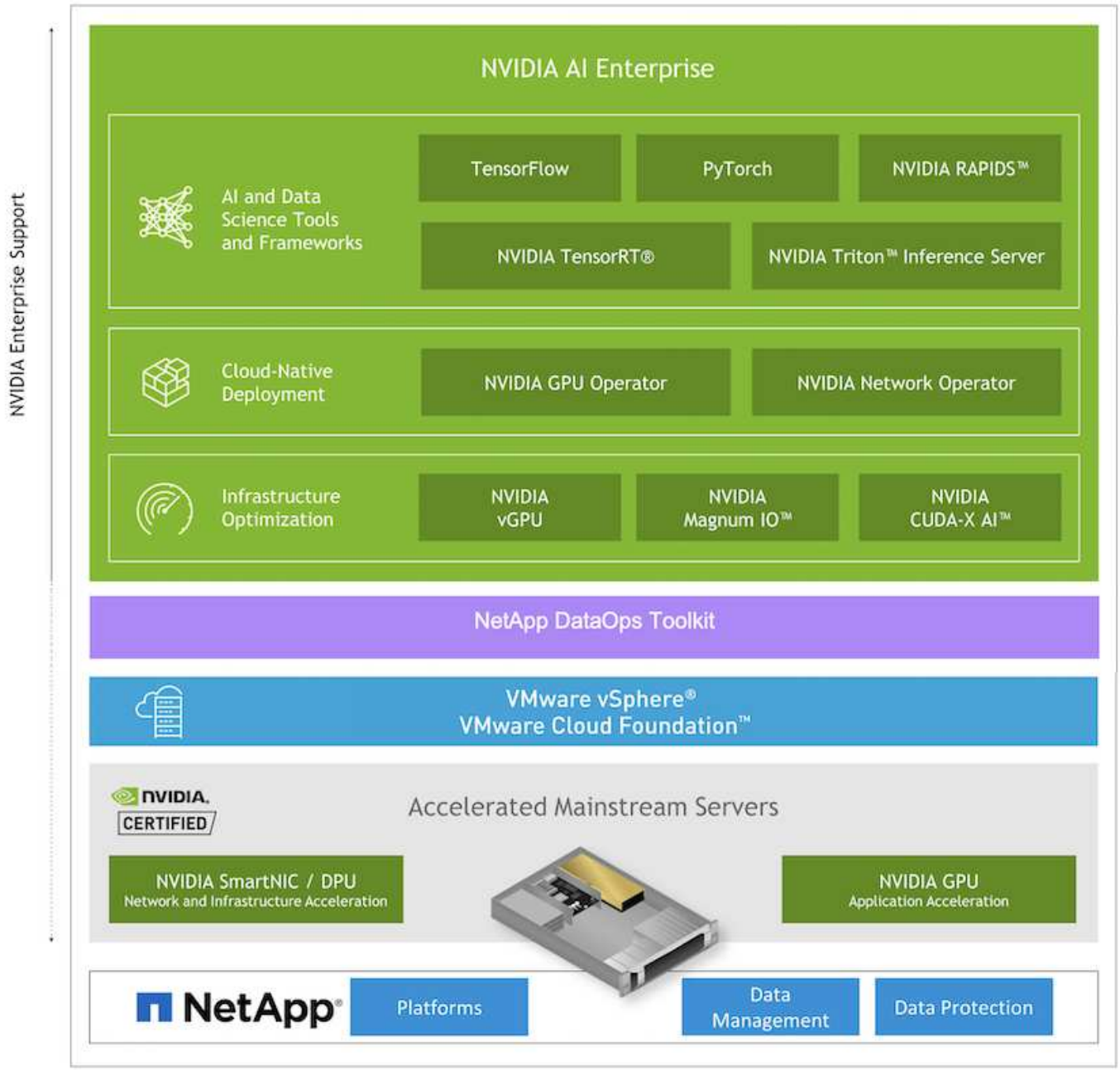
- Provisionnez rapidement de nouveaux espaces de travail JupyterLab haute capacité soutenus par un stockage NetApp hautes performances et évolutif.
- Provisionnez rapidement de nouvelles instances de NVIDIA Triton Inference Server soutenues par un stockage NetApp de classe entreprise.
- Clonez presque instantanément des espaces de travail JupyterLab de grande capacité afin de permettre l'expérimentation ou l'itération rapide.
- Enregistrez presque instantanément des instantanés d'espaces de travail JupyterLab haute capacité pour la sauvegarde et/ou la traçabilité/l'établissement de référence.

- Provisionnez, clonez et capturez des instantanés de volumes de données haute capacité et hautes performances de manière quasi instantanée.

Architecture

Cette solution s'appuie sur une architecture éprouvée et familière comprenant des systèmes certifiés NetApp, VMware et NVIDIA. Voir le tableau suivant pour plus de détails.

Composant	Détails
Logiciels d'IA et d'analyse de données	"NVIDIA AI Enterprise pour VMware"
Plateforme de virtualisation	"VMware vSphere"
Plateforme de calcul	"Systèmes certifiés NVIDIA"
Plateforme de gestion des données	"NetApp ONTAP"



Configuration initiale

Cette section décrit les tâches de configuration initiales qui doivent être effectuées pour utiliser NVIDIA AI Enterprise avec NetApp et VMware.

Prérequis

Avant d'effectuer les étapes décrites dans cette section, nous supposons que vous avez déjà déployé VMware vSphere et NetApp ONTAP. Se référer à la"Matrice de support des produits NVIDIA AI Enterprise" pour plus de détails sur les versions de vSphere prises en charge. Se référer à la"Documentation des solutions NetApp et VMware" pour plus de détails sur le déploiement de VMware vSphere avec NetApp ONTAP.

Installer le logiciel hôte NVIDIA AI Enterprise

Pour installer le logiciel hôte NVIDIA AI Entrprise, suivez les instructions décrites dans les sections 1 à 4 du"Guide de démarrage rapide de NVIDIA AI Enterprise" .

Utiliser le logiciel NVIDIA NGC

Installation

Cette section décrit les tâches de configuration initiales qui doivent être effectuées pour utiliser le logiciel d'entreprise NVIDIA NGC dans un environnement NVIDIA AI Enterprise.

Prérequis

Avant d'effectuer les étapes décrites dans cette section, nous supposons que vous avez déjà déployé le logiciel hôte NVIDIA AI Entrprise en suivant les instructions décrites sur le"Configuration initiale" page.

Créer une machine virtuelle invitée Ubuntu avec vGPU

Tout d'abord, vous devez créer une machine virtuelle invitée Ubuntu 20.04 avec vGPU. Pour créer une machine virtuelle invitée Ubuntu 20.04 avec vGPU, suivez les instructions décrites dans le"Guide de déploiement de NVIDIA AI Enterprise" .

Téléchargez et installez le logiciel invité NVIDIA

Ensuite, vous devez installer le logiciel invité NVIDIA requis dans la machine virtuelle invitée que vous avez créée à l'étape précédente. Pour télécharger et installer le logiciel invité NVIDIA requis dans la machine virtuelle invitée, suivez les instructions décrites dans les sections 5.1 à 5.4 du"Guide de démarrage rapide de NVIDIA AI Enterprise" .

> Lors de l'exécution des tâches de vérification décrites dans la section 5.4, vous devrez peut-être utiliser une balise de version d'image de conteneur CUDA différente, car l'image de conteneur CUDA a été mise à jour depuis la rédaction du guide. Dans notre validation, nous avons utilisé « nvidia/cuda:11.0.3-base-ubuntu20.04 ».

Télécharger le(s) conteneur(s) du framework IA/Analytics

Ensuite, vous devez télécharger les images de conteneur d'IA ou de framework d'analyse nécessaires à partir de NVIDIA NGC afin qu'elles soient disponibles dans votre machine virtuelle invitée. Pour télécharger les conteneurs de framework dans la machine virtuelle invitée, suivez les instructions décrites dans le"Guide de déploiement de NVIDIA AI Enterprise" .

Installer et configurer la boîte à outils NetApp DataOps

Ensuite, vous devez installer NetApp DataOps Toolkit pour les environnements traditionnels dans la machine virtuelle invitée. La boîte à outils NetApp DataOps peut être utilisée pour gérer les volumes de données évolutifs sur votre système ONTAP directement à partir du terminal dans la machine virtuelle invitée. Pour installer NetApp DataOps Toolkit dans la machine virtuelle invitée, effectuez les tâches suivantes.

1. Installer pip.

```
$ sudo apt update
$ sudo apt install python3-pip
$ python3 -m pip install netapp-dataops-traditional
```

2. Déconnectez-vous du terminal VM invité, puis reconnectez-vous.
3. Configurez la boîte à outils NetApp DataOps. Pour terminer cette étape, vous aurez besoin des détails d'accès API pour votre système ONTAP . Vous devrez peut-être les obtenir auprès de votre administrateur de stockage.

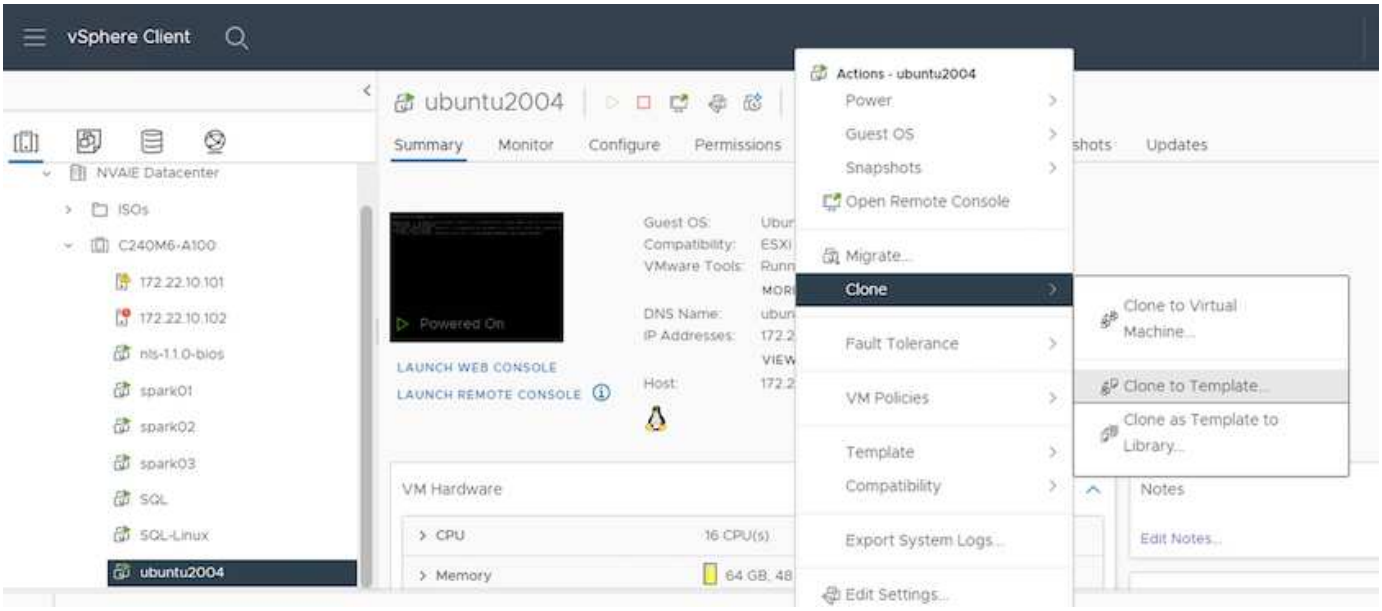
```
$ netapp_dataops_cli.py config

Enter ONTAP management LIF hostname or IP address (Recommendation: Use
SVM management interface): 172.22.10.10
Enter SVM (Storage VM) name: NVAIE-client
Enter SVM NFS data LIF hostname or IP address: 172.22.13.151
Enter default volume type to use when creating new volumes
(flexgroup/flexvol) [flexgroup]:
Enter export policy to use by default when creating new volumes
[default]:
Enter snapshot policy to use by default when creating new volumes
[none]:
Enter unix filesystem user id (uid) to apply by default when creating
new volumes (ex. '0' for root user) [0]:
Enter unix filesystem group id (gid) to apply by default when creating
new volumes (ex. '0' for root group) [0]:
Enter unix filesystem permissions to apply by default when creating new
volumes (ex. '0777' for full read/write permissions for all users and
groups) [0777]:
Enter aggregate to use by default when creating new FlexVol volumes:
aff_a400_01_NVME_SSD_1
Enter ONTAP API username (Recommendation: Use SVM account): admin
Enter ONTAP API password (Recommendation: Use SVM account):
Verify SSL certificate when calling ONTAP API (true/false): false
Do you intend to use this toolkit to trigger BlueXP Copy and Sync
operations? (yes/no): no
Do you intend to use this toolkit to push/pull from S3? (yes/no): no
Created config file: '/home/user/.netapp_dataops/config.json'.
```

Créer un modèle de machine virtuelle invitée

Enfin, vous devez créer un modèle de VM basé sur votre VM invitée. Vous pourrez utiliser ce modèle pour créer rapidement des machines virtuelles invitées pour utiliser le logiciel NVIDIA NGC.

Pour créer un modèle de machine virtuelle basé sur votre machine virtuelle invitée, connectez-vous à VMware vSphere, cliquez avec le bouton droit sur le nom de la machine virtuelle invitée, choisissez « Cloner »,

choisissez « Cloner vers un modèle… », puis suivez l'assistant.



Exemple de cas d'utilisation : tâche d'entraînement TensorFlow

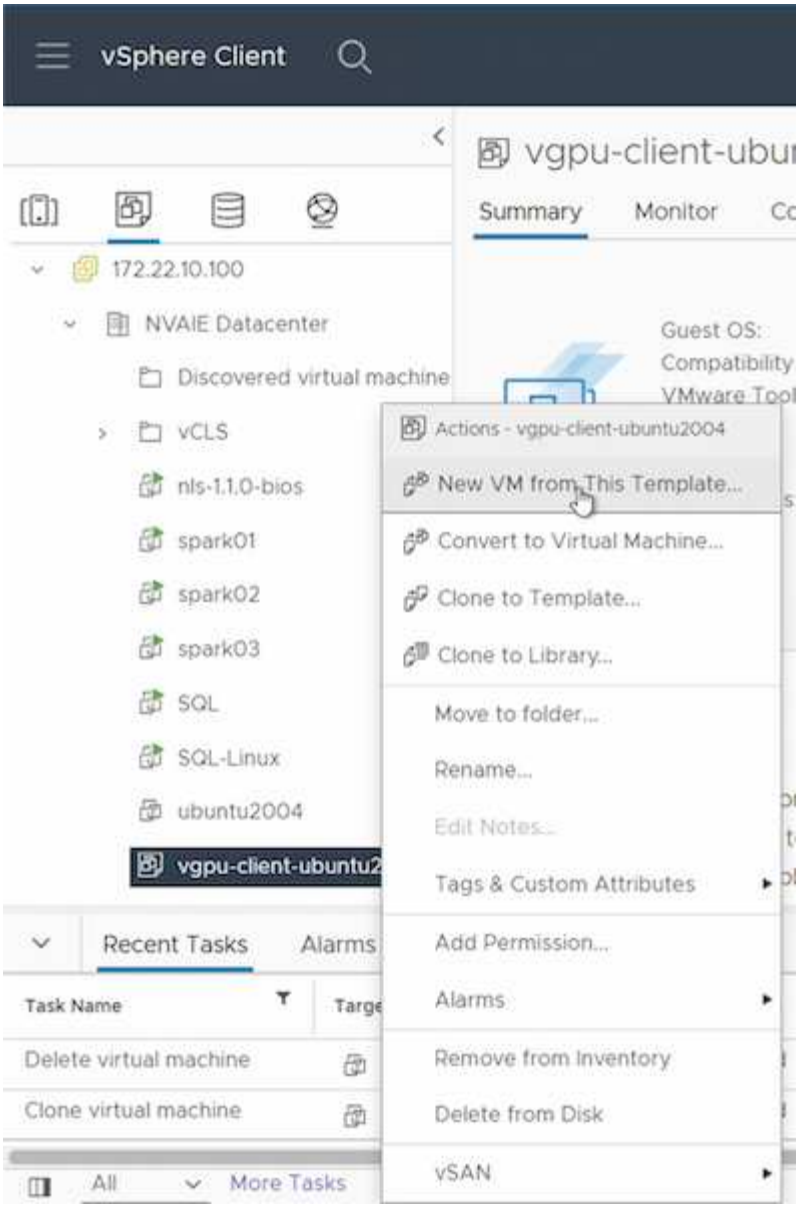
Cette section décrit les tâches à effectuer pour exécuter une tâche de formation TensorFlow dans un environnement NVIDIA AI Enterprise.

Prérequis

Avant d'effectuer les étapes décrites dans cette section, nous supposons que vous avez déjà créé un modèle de machine virtuelle invitée en suivant les instructions décrites dans le"Installation" page.

Créer une machine virtuelle invitée à partir d'un modèle

Tout d'abord, vous devez créer une nouvelle machine virtuelle invitée à partir du modèle que vous avez créé dans la section précédente. Pour créer une nouvelle machine virtuelle invitée à partir de votre modèle, connectez-vous à VMware vSphere, cliquez avec le bouton droit sur le nom du modèle, choisissez « Nouvelle machine virtuelle à partir de ce modèle… », puis suivez l'assistant.



Créer et monter un volume de données

Ensuite, vous devez créer un nouveau volume de données sur lequel stocker votre ensemble de données d'entraînement. Vous pouvez créer rapidement un nouveau volume de données à l'aide de NetApp DataOps Toolkit. L'exemple de commande qui suit montre la création d'un volume nommé « imagenet » d'une capacité de 2 To.

```
$ netapp_dataops_cli.py create vol -n imagenet -s 2TB
```

Avant de pouvoir remplir votre volume de données avec des données, vous devez le monter dans la machine virtuelle invitée. Vous pouvez monter rapidement un volume de données à l'aide de NetApp DataOps Toolkit. L'exemple de commande qui suit montre le montage du volume qui a été créé à l'étape précédente.

```
$ sudo -E netapp_dataops_cli.py mount vol -n imagenet -m ~/imagenet
```

Remplir le volume de données

Une fois le nouveau volume provisionné et monté, l'ensemble de données de formation peut être récupéré à partir de l'emplacement source et placé sur le nouveau volume. Cela impliquera généralement d'extraire les données d'un lac de données S3 ou Hadoop et nécessitera parfois l'aide d'un ingénieur de données.

Exécuter la tâche d'entraînement TensorFlow

Vous êtes maintenant prêt à exécuter votre tâche de formation TensorFlow. Pour exécuter votre tâche de formation TensorFlow, effectuez les tâches suivantes.

1. Extrayez l'image du conteneur NVIDIA NGC Enterprise TensorFlow.

```
$ sudo docker pull nvcr.io/nvaie/tensorflow-2-1:22.05-tf1-nvaie-2.1-py3
```

2. Lancez une instance du conteneur TensorFlow d'entreprise NVIDIA NGC. Utilisez l'option « -v » pour attacher votre volume de données au conteneur.

```
$ sudo docker run --gpus all -v ~/imagenet:/imagenet -it --rm
nvcr.io/nvaie/tensorflow-2-1:22.05-tf1-nvaie-2.1-py3
```

3. Exécutez votre programme de formation TensorFlow dans le conteneur. L'exemple de commande qui suit montre l'exécution d'un exemple de programme de formation ResNet-50 inclus dans l'image du conteneur.

```
$ python ./nvidia-examples/cnn/resnet.py --layers 50 -b 64 -i 200 -u
batch --precision fp16 --data_dir /imagenet/data
```

Où trouver des informations supplémentaires

Pour en savoir plus sur les informations décrites dans ce document, reportez-vous aux documents et/ou sites Web suivants :

- Logiciel de gestion de données NetApp ONTAP — Bibliothèque d'informations ONTAP

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

- Boîte à outils NetApp DataOps

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

- NVIDIA AI Enterprise avec VMware

  "https://www.nvidia.com/en-us/data-center/products/ai-enterprise/"

Remerciements

- Bobby Oommen, directeur principal, NetApp

- Ramesh Isaac, administrateur système, NetApp
- Roney Daniel, ingénieur marketing technique, NetApp

Informations sur le copyright

Copyright © 2025 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.