NVIDIA AI Enterprise com NetApp e VMware

NetApp artificial intelligence solutions

NetApp
December 04, 2025

Índice

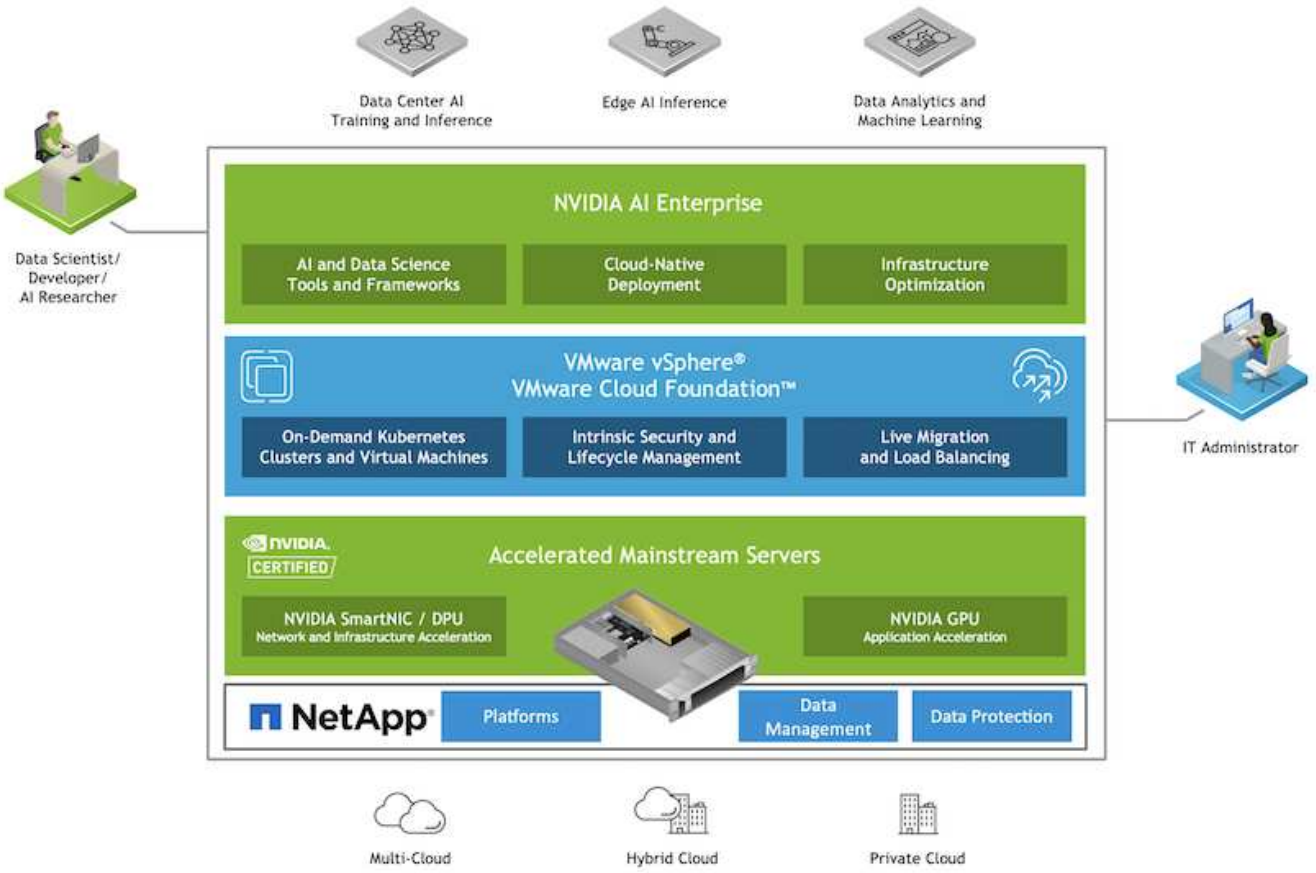
NVIDIA AI Enterprise com NetApp e VMware

NVIDIA AI Enterprise com NetApp e VMware

Mike Oglesby, NetApp

Para arquitetos e administradores de TI, as ferramentas de IA podem ser complicadas e desconhecidas. Além disso, muitas plataformas de IA não estão prontas para empresas. O NVIDIA AI Enterprise, com tecnologia NetApp e VMware, foi criado para oferecer uma arquitetura de IA simplificada e de nível empresarial.

O NVIDIA AI Enterprise é um conjunto completo e nativo em nuvem de software de IA e análise de dados que é otimizado, certificado e suportado pela NVIDIA para ser executado no VMware vSphere com sistemas certificados pela NVIDIA. Este software facilita a implantação, o gerenciamento e o dimensionamento simples e rápidos de cargas de trabalho de IA no ambiente de nuvem híbrida moderno. O NVIDIA AI Enterprise, com tecnologia NetApp e VMware, oferece carga de trabalho de IA de nível empresarial e gerenciamento de dados em um pacote simplificado e familiar.



Visão geral da tecnologia

Esta seção fornece uma visão geral da tecnologia para NVIDIA AI Enterprise com NetApp e VMware.

NVIDIA AI Enterprise

O NVIDIA AI Enterprise é um conjunto completo e nativo em nuvem de software de IA e análise de dados que é otimizado, certificado e suportado pela NVIDIA para ser executado no VMware vSphere com sistemas certificados pela NVIDIA. Este software facilita a implantação, o gerenciamento e o dimensionamento simples e rápidos de cargas de trabalho de IA no ambiente de nuvem híbrida moderno.

Nuvem de GPU NVIDIA (NGC)

O NVIDIA NGC hospeda um catálogo de software otimizado para GPU para que profissionais de IA desenvolvam suas soluções de IA. Ele também fornece acesso a vários serviços de IA, incluindo o NVIDIA Base Command para treinamento de modelos, o NVIDIA Fleet Command para implantar e monitorar modelos e o NGC Private Registry para acessar e gerenciar com segurança software proprietário de IA. Além disso, os clientes do NVIDIA AI Enterprise podem solicitar suporte pelo portal NGC.

VMware vSphere

O VMware vSphere é a plataforma de virtualização da VMware, que transforma data centers em infraestruturas de computação agregadas que incluem recursos de CPU, armazenamento e rede. O vSphere gerencia essas infraestruturas como um ambiente operacional unificado e fornece aos administradores as ferramentas para gerenciar os data centers que participam desse ambiente.

Os dois principais componentes do vSphere são o ESXi e o vCenter Server. O ESXi é a plataforma de virtualização onde os administradores criam e executam máquinas virtuais e dispositivos virtuais. O vCenter Server é o serviço por meio do qual os administradores gerenciam vários hosts conectados em uma rede e agrupam recursos de host.

NetApp ONTAP

ONTAP 9, a última geração de software de gerenciamento de armazenamento da NetApp, permite que as empresas modernizem a infraestrutura e façam a transição para um data center pronto para a nuvem. Aproveitando os recursos de gerenciamento de dados líderes do setor, o ONTAP permite o gerenciamento e a proteção de dados com um único conjunto de ferramentas, independentemente de onde os dados residam. Você também pode mover dados livremente para onde for necessário: na borda, no núcleo ou na nuvem. O ONTAP 9 inclui vários recursos que simplificam o gerenciamento de dados, aceleram e protegem dados críticos e permitem recursos de infraestrutura de última geração em arquiteturas de nuvem híbrida.

Simplifique o gerenciamento de dados

O gerenciamento de dados é crucial para as operações de TI corporativas e cientistas de dados, para que recursos apropriados sejam usados para aplicativos de IA e treinamento de conjuntos de dados de IA/ML. As seguintes informações adicionais sobre as tecnologias NetApp estão fora do escopo desta validação, mas podem ser relevantes dependendo da sua implantação.

O software de gerenciamento de dados ONTAP inclui os seguintes recursos para otimizar e simplificar as operações e reduzir seu custo total de operação:

- Compactação de dados em linha e desduplicação expandida. A compactação de dados reduz o desperdício de espaço dentro dos blocos de armazenamento e a desduplicação aumenta significativamente a capacidade efetiva. Isso se aplica a dados armazenados localmente e dados em camadas na nuvem.
- Qualidade de serviço mínima, máxima e adaptável (AQoS). Controles granulares de qualidade de serviço (QoS) ajudam a manter os níveis de desempenho para aplicativos críticos em ambientes altamente

compartilhados.

- NetApp FabricPool. Fornece hierarquização automática de dados frios para opções de armazenamento em nuvem pública e privada, incluindo Amazon Web Services (AWS), Azure e solução de armazenamento NetApp StorageGRID . Para obter mais informações sobre FabricPool, consulte "TR-4598: Melhores práticas do FabricPool" .

Acelere e proteja os dados

O ONTAP oferece níveis superiores de desempenho e proteção de dados e estende esses recursos das seguintes maneiras:

- Desempenho e menor latência. ONTAP oferece o maior rendimento possível com a menor latência possível.
- Proteção de dados. O ONTAP fornece recursos integrados de proteção de dados com gerenciamento comum em todas as plataformas.
- Criptografia de volume NetApp (NVE). O ONTAP oferece criptografia nativa em nível de volume com suporte para gerenciamento de chaves externo e integrado.
- Multilocação e autenticação multifator. O ONTAP permite o compartilhamento de recursos de infraestrutura com os mais altos níveis de segurança.

Infraestrutura à prova do futuro

O ONTAP ajuda a atender às necessidades empresariais exigentes e em constante mudança com os seguintes recursos:

- Escalabilidade perfeita e operações não disruptivas. O ONTAP oferece suporte à adição não disruptiva de capacidade aos controladores existentes e aos clusters escaláveis. Os clientes podem atualizar para as tecnologias mais recentes, como NVMe e 32Gb FC, sem migrações de dados dispendiosas ou interrupções.
- Conexão em nuvem. ONTAP é o software de gerenciamento de armazenamento mais conectado à nuvem, com opções para armazenamento definido por software (ONTAP Select) e instâncias nativas da nuvem (Google Cloud NetApp Volumes) em todas as nuvens públicas.
- Integração com aplicações emergentes. A ONTAP oferece serviços de dados de nível empresarial para plataformas e aplicativos de última geração, como veículos autônomos, cidades inteligentes e Indústria 4.0, usando a mesma infraestrutura que dá suporte aos aplicativos empresariais existentes.

Kit de ferramentas NetApp DataOps

O NetApp DataOps Toolkit é uma ferramenta baseada em Python que simplifica o gerenciamento de espaços de trabalho de desenvolvimento/treinamento e servidores de inferência apoiados por armazenamento NetApp de alto desempenho e escalonável. Os principais recursos incluem:

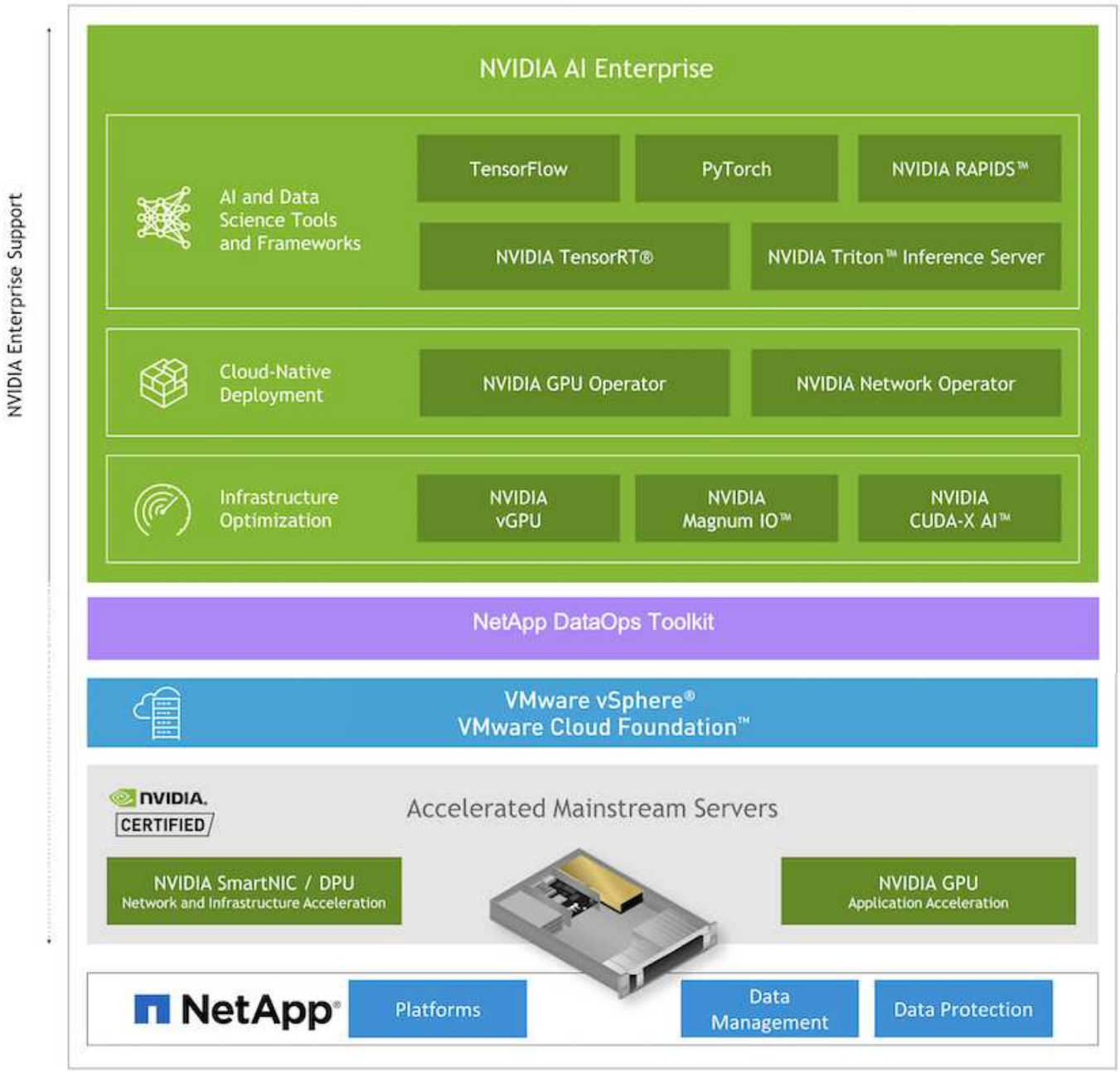
- Provisione rapidamente novos espaços de trabalho JupyterLab de alta capacidade, apoiados por armazenamento NetApp escalável e de alto desempenho.
- Provisione rapidamente novas instâncias do NVIDIA Triton Inference Server com suporte de armazenamento NetApp de nível empresarial.
- Clone quase instantaneamente espaços de trabalho de alta capacidade do JupyterLab para permitir experimentação ou iteração rápida.
- Salve quase instantaneamente snapshots de espaços de trabalho de alta capacidade do JupyterLab para backup e/ou rastreabilidade/linha de base.

- Provisione, clone e crie snapshots de volumes de dados de alta capacidade e alto desempenho quase instantaneamente.

Arquitetura

Esta solução se baseia em uma arquitetura comprovada e familiar com sistemas certificados pela NetApp, VMware e NVIDIA. Veja a tabela a seguir para mais detalhes.

Componente	Detalhes
Software de IA e análise de dados	"NVIDIA AI Enterprise para VMware"
Plataforma de Virtualização	"VMware vSphere"
Plataforma de computação	"Sistemas certificados pela NVIDIA"
Plataforma de Gerenciamento de Dados	"NetApp ONTAP"



Configuração inicial

Esta seção descreve as tarefas de configuração inicial que precisam ser executadas para utilizar o NVIDIA AI Enterprise com NetApp e VMware.

Pré-requisitos

Antes de executar as etapas descritas nesta seção, presumimos que você já tenha implantado o VMware vSphere e o NetApp ONTAP. Consulte o"Matriz de suporte a produtos empresariais NVIDIA AI" para obter detalhes sobre as versões suportadas do vSphere. Consulte o"Documentação da solução NetApp e VMware" para obter detalhes sobre a implantação do VMware vSphere com o NetApp ONTAP.

Instalar o software host empresarial NVIDIA AI

Para instalar o software host NVIDIA AI Enterprise, siga as instruções descritas nas seções 1 a 4 do"Guia de início rápido do NVIDIA AI Enterprise" .

Use o software NVIDIA NGC

Configurar

Esta seção descreve as tarefas de configuração inicial que precisam ser executadas para utilizar o software empresarial NVIDIA NGC em um ambiente NVIDIA AI Enterprise.

Pré-requisitos

Antes de executar as etapas descritas nesta seção, presumimos que você já tenha implantado o software host NVIDIA AI Enterprise seguindo as instruções descritas na"Configuração inicial" página.

Crie uma VM convidada do Ubuntu com vGPU

Primeiro, você deve criar uma VM convidada do Ubuntu 20.04 com vGPU. Para criar uma VM convidada do Ubuntu 20.04 com vGPU, siga as instruções descritas no"Guia de implantação empresarial do NVIDIA AI" .

Baixe e instale o software convidado NVIDIA

Em seguida, você deve instalar o software convidado NVIDIA necessário na VM convidada que você criou na etapa anterior. Para baixar e instalar o software convidado NVIDIA necessário na VM convidada, siga as instruções descritas nas seções 5.1-5.4 do"Guia de início rápido do NVIDIA AI Enterprise" .

> Ao executar as tarefas de verificação descritas na seção 5.4, talvez seja necessário usar uma tag de versão de imagem de contêiner CUDA diferente, pois a imagem de contêiner CUDA foi atualizada desde a redação do guia. Em nossa validação, usamos 'nvidia/cuda:11.0.3-base-ubuntu20.04'.

Baixar contêiner(es) do AI/Analytics Framework

Em seguida, você deve baixar as imagens de contêiner da estrutura de IA ou análise necessárias do NVIDIA NGC para que elas fiquem disponíveis na sua VM convidada. Para baixar contêineres de estrutura na VM convidada, siga as instruções descritas no"Guia de implantação empresarial do NVIDIA AI" .

Instalar e configurar o NetApp DataOps Toolkit

Em seguida, você deve instalar o NetApp DataOps Toolkit para ambientes tradicionais na VM convidada. O NetApp DataOps Toolkit pode ser usado para gerenciar volumes de dados escaláveis no seu sistema ONTAP diretamente do terminal na VM convidada. Para instalar o NetApp DataOps Toolkit na VM convidada, execute as seguintes tarefas.

1. Instalar pip.

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

2. Saia do terminal da VM convidada e depois efetue login novamente.
3. Configure o NetApp DataOps Toolkit. Para concluir esta etapa, você precisará de detalhes de acesso à API para seu sistema ONTAP . Talvez seja necessário obtê-los com seu administrador de armazenamento.

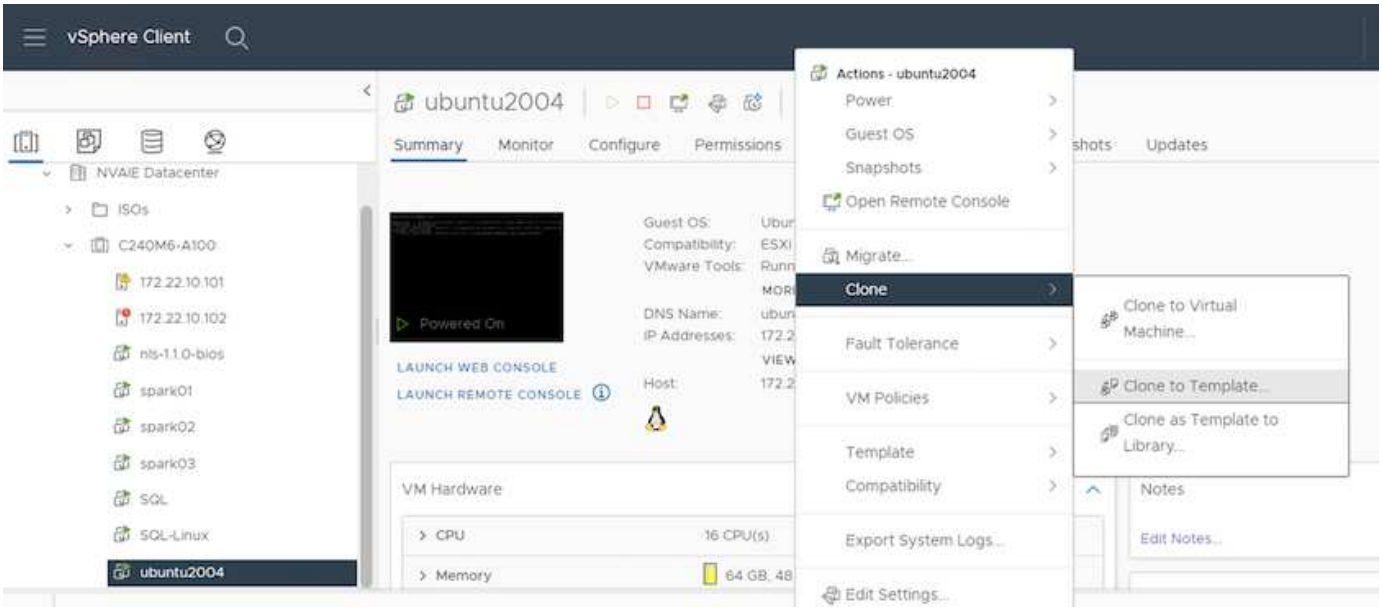
```
$ netapp_dataops_cli.py config

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'.
```

Criar um modelo de VM convidada

Por fim, você deve criar um modelo de VM com base na sua VM convidada. Você poderá usar este modelo para criar rapidamente VMs convidadas para utilizar o software NVIDIA NGC.

Para criar um modelo de VM com base na sua VM convidada, faça login no VMware vSphere, clique com o botão direito do mouse no nome da VM convidada, escolha "Clonar", escolha "Clonar para modelo…" e siga o

assistente.



Exemplo de caso de uso - Tarefa de treinamento do TensorFlow

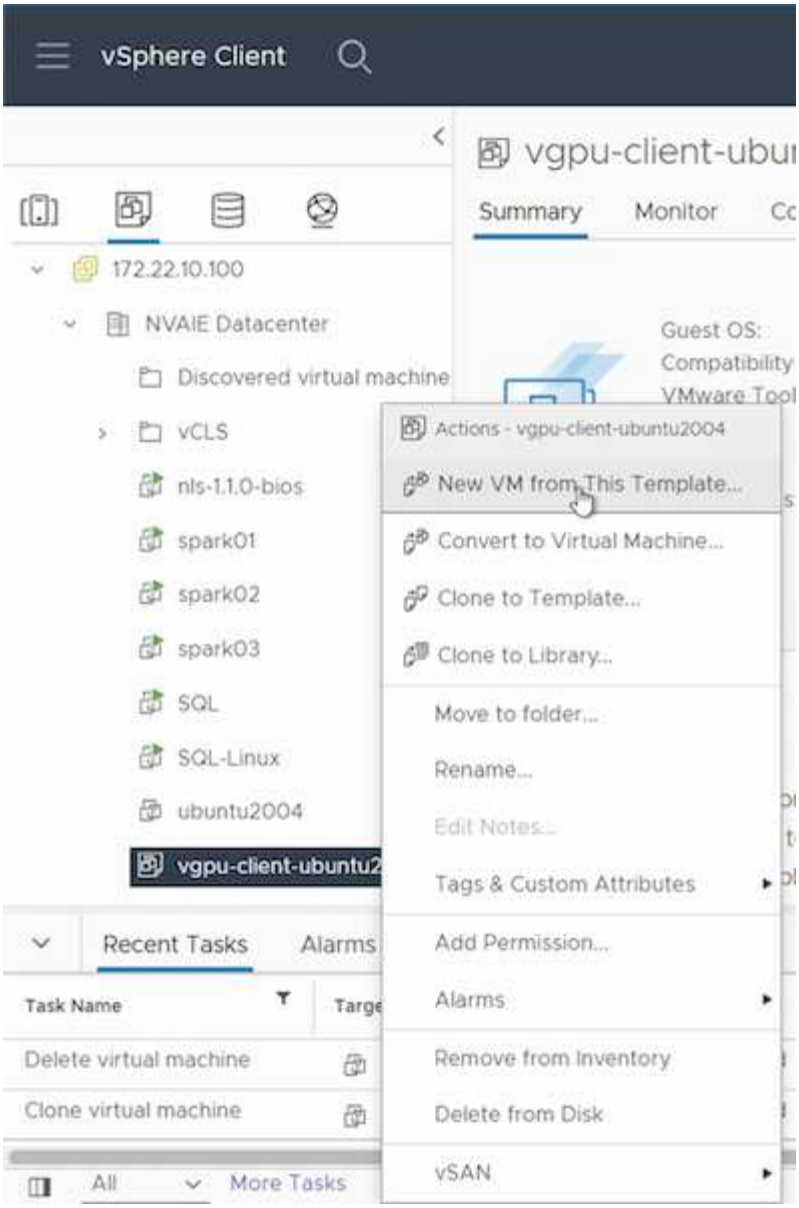
Esta seção descreve as tarefas que precisam ser executadas para executar um trabalho de treinamento do TensorFlow em um ambiente NVIDIA AI Enterprise.

Pré-requisitos

Antes de executar as etapas descritas nesta seção, presumimos que você já criou um modelo de VM convidada seguindo as instruções descritas na"Configurar" página.

Criar VM convidada a partir do modelo

Primeiro, você deve criar uma nova VM convidada a partir do modelo criado na seção anterior. Para criar uma nova VM convidada a partir do seu modelo, faça login no VMware vSphere, clique com o botão direito do mouse no nome do modelo, escolha "Nova VM deste modelo…" e siga o assistente.



Criar e montar volume de dados

Em seguida, você deve criar um novo volume de dados para armazenar seu conjunto de dados de treinamento. Você pode criar rapidamente um novo volume de dados usando o NetApp DataOps Toolkit. O comando de exemplo a seguir mostra a criação de um volume chamado 'imagenet' com capacidade de 2 TB.

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

Antes de preencher seu volume de dados com dados, você deve montá-lo na VM convidada. Você pode montar rapidamente um volume de dados usando o NetApp DataOps Toolkit. O comando de exemplo a seguir mostra a montagem do volume que foi criado na etapa anterior.

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

Preencher volume de dados

Depois que o novo volume for provisionado e montado, o conjunto de dados de treinamento poderá ser recuperado do local de origem e colocado no novo volume. Isso normalmente envolverá extrair dados de um data lake S3 ou Hadoop e, às vezes, envolverá a ajuda de um engenheiro de dados.

Executar tarefa de treinamento do TensorFlow

Agora, você está pronto para executar seu trabalho de treinamento do TensorFlow. Para executar seu trabalho de treinamento do TensorFlow, execute as seguintes tarefas.

1. Puxe a imagem do contêiner NVIDIA NGC Enterprise TensorFlow.

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

2. Inicie uma instância do contêiner NVIDIA NGC Enterprise TensorFlow. Use a opção '-v' para anexar seu volume de dados ao contêiner.

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

3. Execute seu programa de treinamento do TensorFlow dentro do contêiner. O comando de exemplo a seguir mostra a execução de um programa de treinamento ResNet-50 de exemplo que está incluído na imagem do contêiner.

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

Onde encontrar informações adicionais

Para saber mais sobre as informações descritas neste documento, consulte os seguintes documentos e/ou sites:

- Software de gerenciamento de dados NetApp ONTAP — biblioteca de informações ONTAP

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

- Kit de ferramentas NetApp DataOps

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

- NVIDIA AI Enterprise com VMware

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

Agradecimentos

- Bobby Oommen, gerente sênior, NetApp
- Ramesh Isaac, Administrador de Sistemas, NetApp
- Roney Daniel, Engenheiro de Marketing Técnico, NetApp

Informações sobre direitos autorais

Informações sobre marcas comerciais