



NVIDIA AI Enterprise 与NetApp和 VMware 合作

NetApp artificial intelligence solutions

NetApp
February 12, 2026

目录

NVIDIA AI Enterprise 与NetApp和 VMware 合作	1
NVIDIA AI Enterprise 与NetApp和 VMware 合作	1
技术概述	1
NVIDIA AI 企业版	2
NVIDIA GPU 云 (NGC)	2
VMware vSphere	2
NetApp ONTAP	2
NetApp DataOps 工具包	3
架构	3
初始设置	4
前提条件	4
安装NVIDIA AI Enterprise Host 软件	5
使用NVIDIA NGC 软件	5
设置	5
示例用例 - TensorFlow 训练作业	7
如何查找更多信息	9
声明	9

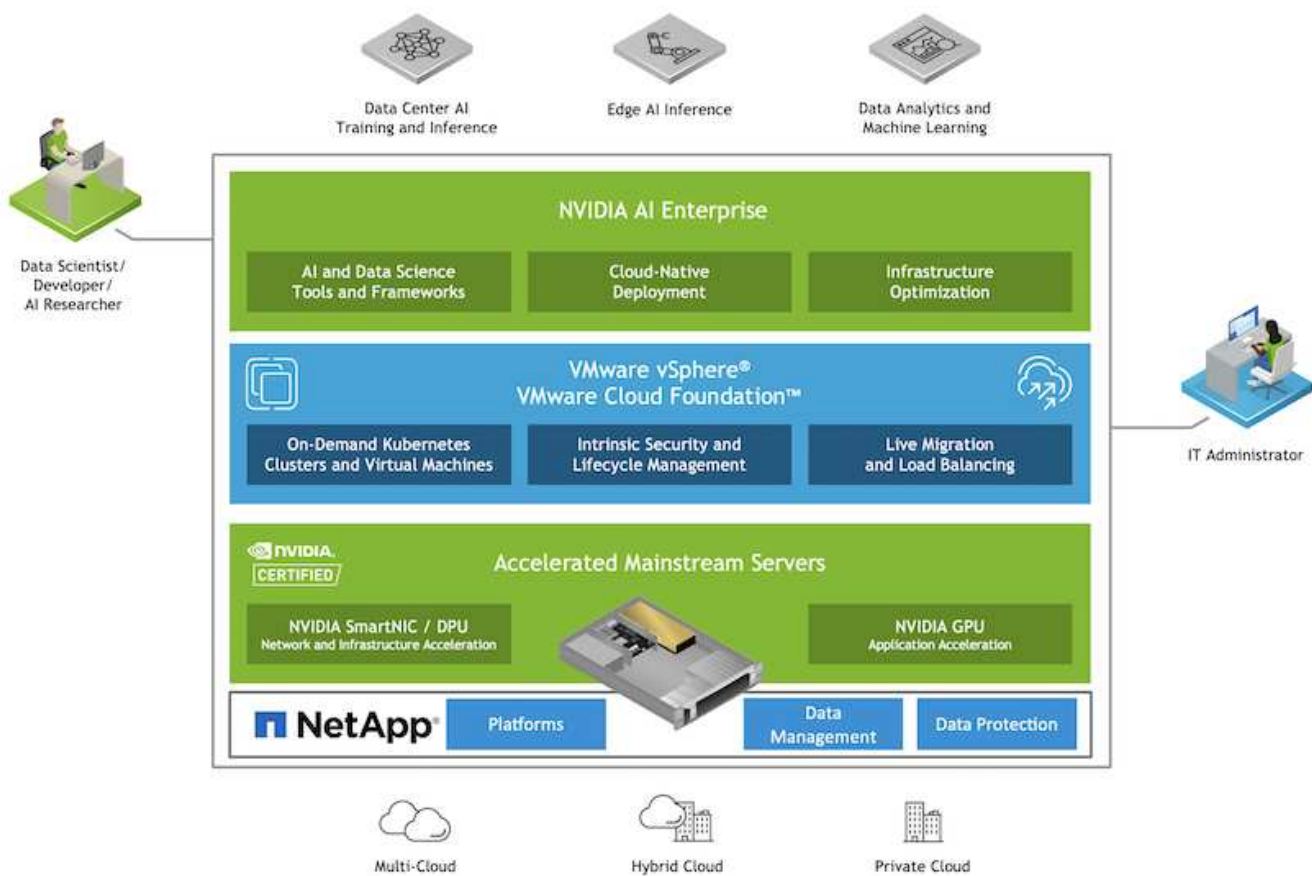
NVIDIA AI Enterprise 与NetApp和 VMware 合作

NVIDIA AI Enterprise 与NetApp和 VMware 合作

NetApp的 Mike Oglesby

对于 IT 架构师和管理员来说，AI 工具可能很复杂且陌生。此外，许多人工智能平台尚未为企业做好准备。NVIDIA AI Enterprise 由NetApp和 VMware 提供支持，旨在提供精简的企业级 AI 架构。

NVIDIA AI Enterprise 是一款端到端、云原生的 AI 和数据分析软件套件，经过NVIDIA优化、认证和支持，可在具有NVIDIA认证系统的 VMware vSphere 上运行。该软件有助于在现代混合云环境中简单、快速地部署、管理和扩展 AI 工作负载。由NetApp和 VMware 提供支持的NVIDIA AI Enterprise 以简化、熟悉的软件包提供企业级 AI 工作负载和数据管理。



技术概述

本节概述了NVIDIA AI Enterprise 与NetApp和 VMware 的技术。

NVIDIA AI 企业版

NVIDIA AI Enterprise 是一款端到端、云原生的 AI 和数据分析软件套件，经过NVIDIA优化、认证和支持，可在具有NVIDIA认证系统的 VMware vSphere 上运行。该软件有助于在现代混合云环境中简单、快速地部署、管理和扩展 AI 工作负载。

NVIDIA GPU 云 (NGC)

NVIDIA NGC 为 AI 从业者提供了一系列针对 GPU 优化的软件，以开发他们的 AI 解决方案。它还提供对各种 AI 服务的访问，包括用于模型训练的NVIDIA Base Command、用于部署和监控模型的NVIDIA Fleet Command 以及用于安全访问和管理专有 AI 软件的 NGC Private Registry。此外，NVIDIA AI Enterprise 客户可以通过 NGC 门户请求支持。

VMware vSphere

VMware vSphere 是 VMware 的虚拟化平台，它将数据中心转变为包含 CPU、存储和网络资源的聚合计算基础架构。vSphere 将这些基础架构作为统一的操作环境进行管理，并为管理员提供管理该环境中数据中心的工具。

vSphere 的两个核心组件是 ESXi 和 vCenter Server。ESXi 是管理员创建和运行虚拟机及虚拟设备的虚拟化平台。vCenter Server 是管理员管理网络中连接的多台主机并池化主机资源的服务。

NetApp ONTAP

ONTAP 9 是NetApp最新一代存储管理软件，它支持企业实现基础架构现代化并过渡到云就绪数据中心。ONTAP利用业界领先的数据管理功能，只需一套工具即可管理和保护数据，无论数据位于何处。您还可以将数据自由移动到任何需要的地方：边缘、核心或云端。ONTAP 9 包含众多功能，可简化数据管理、加速和保护关键数据，并支持跨混合云架构的下一代基础架构功能。

简化数据管理

数据管理对于企业 IT 运营和数据科学家至关重要，以便将适当的资源用于 AI 应用程序和训练 AI/ML 数据集。以下有关NetApp技术的附加信息超出了本次验证的范围，但可能与您的部署相关。

ONTAP数据管理软件包括以下功能，可简化操作并降低总运营成本：

- 内联数据压缩和扩展重复数据删除。数据压缩减少了存储块内部浪费的空间，重复数据删除显著增加了有效容量。这适用于本地存储的数据和分层到云的数据。
- 最小、最大和自适应服务质量 (AQoS)。细粒度的服务质量 (QoS) 控制有助于维持高度共享环境中关键应用程序的性能水平。
- NetApp FabricPool。提供冷数据自动分层到公共和私有云存储选项，包括 Amazon Web Services (AWS)、Azure 和NetApp StorageGRID存储解决方案。有关FabricPool的更多信息，请参阅 ["TR-4598：FabricPool最佳实践"](#)。

加速并保护数据

ONTAP提供卓越级别的性能和数据保护，并通过以下方式扩展这些功能：

- 性能和更低的延迟。ONTAP以尽可能低的延迟提供尽可能高的吞吐量。
- 数据保护。ONTAP提供内置数据保护功能，并在所有平台上提供通用管理。
- NetApp卷加密 (NVE)。ONTAP提供原生卷级加密，同时支持板载和外部密钥管理。

- 多租户和多因素身份验证。 ONTAP支持以最高级别的安全性共享基础设施资源。

面向未来的基础设施

ONTAP具有以下功能，可帮助满足苛刻且不断变化的业务需求：

- 无缝扩展和无中断运行。 ONTAP支持无中断地向现有控制器和横向扩展集群添加容量。客户可以升级到最新技术，例如 NVMe 和 32Gb FC，而无需昂贵的数据迁移或中断。
- 云连接。 ONTAP是与云连接最紧密的存储管理软件，在所有公共云中均提供软件定义存储（ONTAP Select）和云原生实例（Google Cloud NetApp Volumes）的选项。
- 与新兴应用程序的集成。 ONTAP使用支持现有企业应用的相同基础架构，为下一代平台和应用（如自动驾驶汽车、智能城市和工业 4.0）提供企业级数据服务。

NetApp DataOps 工具包

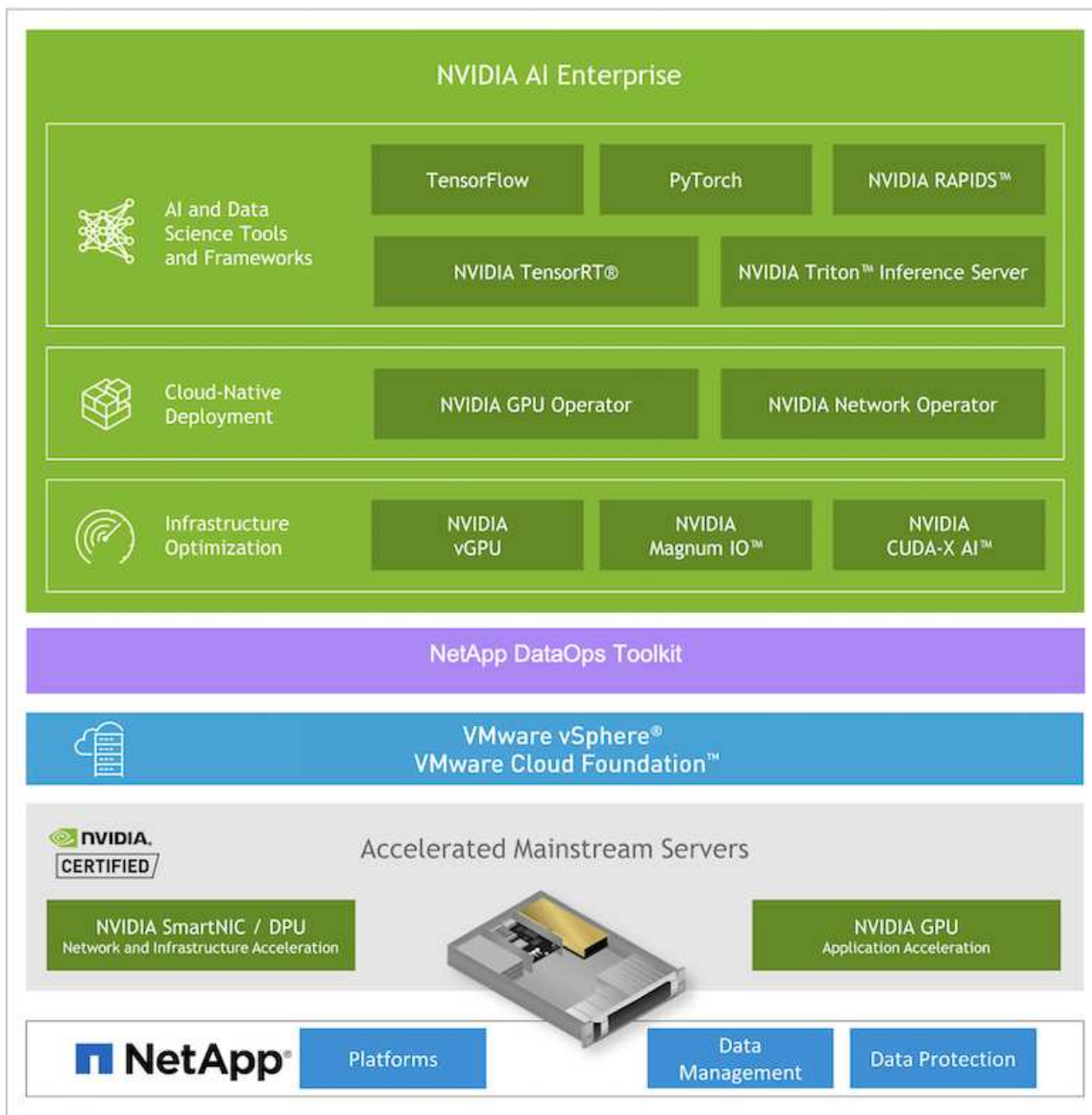
NetApp DataOps Toolkit 是一款基于 Python 的工具，可简化由高性能、横向扩展NetApp存储支持的开发/培训工作区和推理服务器的管理。主要功能包括：

- 快速配置由高性能、横向扩展NetApp存储支持的新的容量 JupyterLab 工作区。
- 快速配置由企业级NetApp存储支持的全新NVIDIA Triton 推理服务器实例。
- 近乎即时地克隆容量的 JupyterLab 工作区，以实现实验或快速迭代。
- 近乎即时地保存容量 JupyterLab 工作区的快照，以用于备份和/或可追溯性/基准测试。
- 近乎即时地提供、克隆和快照大容量、高性能数据卷。

架构

该解决方案建立在经过验证且熟悉的架构之上，具有NetApp、VMware 和NVIDIA认证的系系统。请参阅下表以了解详情。

组件	详细信息
人工智能和数据分析软件	"适用于 VMware 的NVIDIA AI Enterprise"
虚拟化平台	"VMware vSphere"
计算平台	"NVIDIA认证系统"
数据管理平台	"NetApp ONTAP"



初始设置

本节介绍了将NVIDIA AI Enterprise 与NetApp和 VMware 结合使用所需执行的初始设置任务。

前提条件

在执行本节概述的步骤之前，我们假设您已经部署了 VMware vSphere 和NetApp ONTAP。请参阅["NVIDIA AI Enterprise 产品支持矩阵"](#)有关受支持的 vSphere 版本的详细信息。请参阅["NetApp和 VMware 解决方案文档"](#)有关使用NetApp ONTAP部署 VMware vSphere 的详细信息。

安装NVIDIA AI Enterprise Host 软件

要安装NVIDIA AI Enterprise 主机软件，请按照["NVIDIA AI Enterprise 快速入门指南"](#)。

使用NVIDIA NGC 软件

设置

本节介绍在NVIDIA AI Enterprise 环境中使用NVIDIA NGC 企业软件所需执行的初始设置任务。

前提条件

在执行本节中概述的步骤之前，NVIDIA假设您已经按照["初始设置"](#)页。

使用 vGPU 创建 Ubuntu Guest VM

首先，您必须创建一个带有 vGPU 的 Ubuntu 20.04 客户虚拟机。要创建带有 vGPU 的 Ubuntu 20.04 来宾虚拟机，请按照["NVIDIA AI 企业部署指南"](#)。

下载并安装NVIDIA客户软件

接下来，您必须在上一步创建的客户虚拟机中安装所需的NVIDIA客户软件。要在客户虚拟机中下载并安装所需的NVIDIA客户软件，请按照["NVIDIA AI Enterprise 快速入门指南"](#)。



执行第 5.4 节中概述的验证任务时，您可能需要使用不同的 CUDA 容器映像版本标签，因为自编写指南以来 CUDA 容器映像已更新。在我们的验证中，我们使用了“nvidia/cuda:11.0.3-base-ubuntu20.04”。

下载 AI/分析框架容器

接下来，您必须从NVIDIA NGC 下载所需的 AI 或分析框架容器映像，以便它们可以在您的客户虚拟机中使用。要在客户虚拟机中下载框架容器，请按照["NVIDIA AI 企业部署指南"](#)。

安装和配置NetApp DataOps 工具包

接下来，您必须在来宾虚拟机中安装适用于传统环境的NetApp DataOps Toolkit。NetApp DataOps Toolkit 可用于直接从客户虚拟机内的终端管理ONTAP系统上的横向扩展数据卷。要在客户虚拟机中安装NetApp DataOps Toolkit，请执行以下任务。

1. 安装 pip。

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

2. 注销客户 VM 终端，然后重新登录。

3. 配置NetApp DataOps 工具包。为了完成此步骤，您将需要ONTAP系统的 API 访问详细信息。您可能需要从存储管理员处获取这些信息。

```
$ 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
```

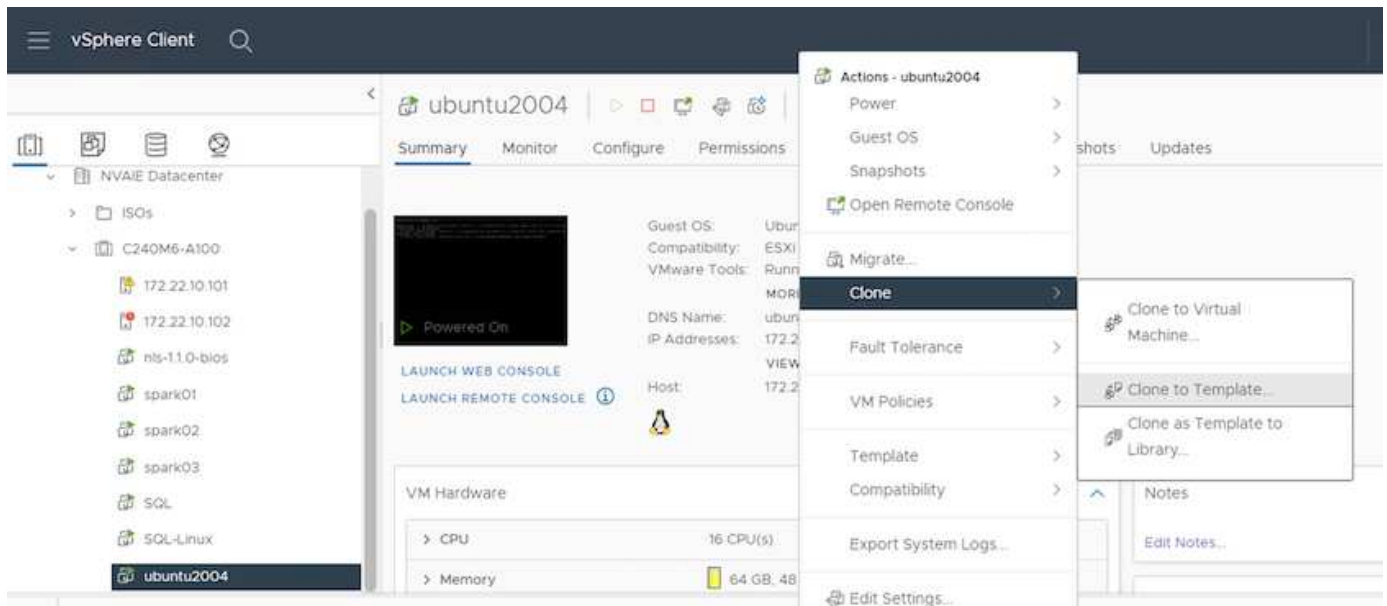
```
Do you intend to use this toolkit to push/pull from S3? (yes/no): no
```

```
Created config file: '/home/user/.netapp_dataops/config.json'.
```

创建来宾虚拟机模板

最后，您必须根据您的客户虚拟机创建一个虚拟机模板。您将能够使用此模板快速创建用于利用NVIDIA NGC 软件的客户虚拟机。

要根据您的来宾 VM 创建 VM 模板，请登录 VMware vSphere，右键单击来宾 VM 名称，选择“克隆”，选择“克隆到模板...”，然后按照向导进行操作。



示例用例 - TensorFlow 训练作业

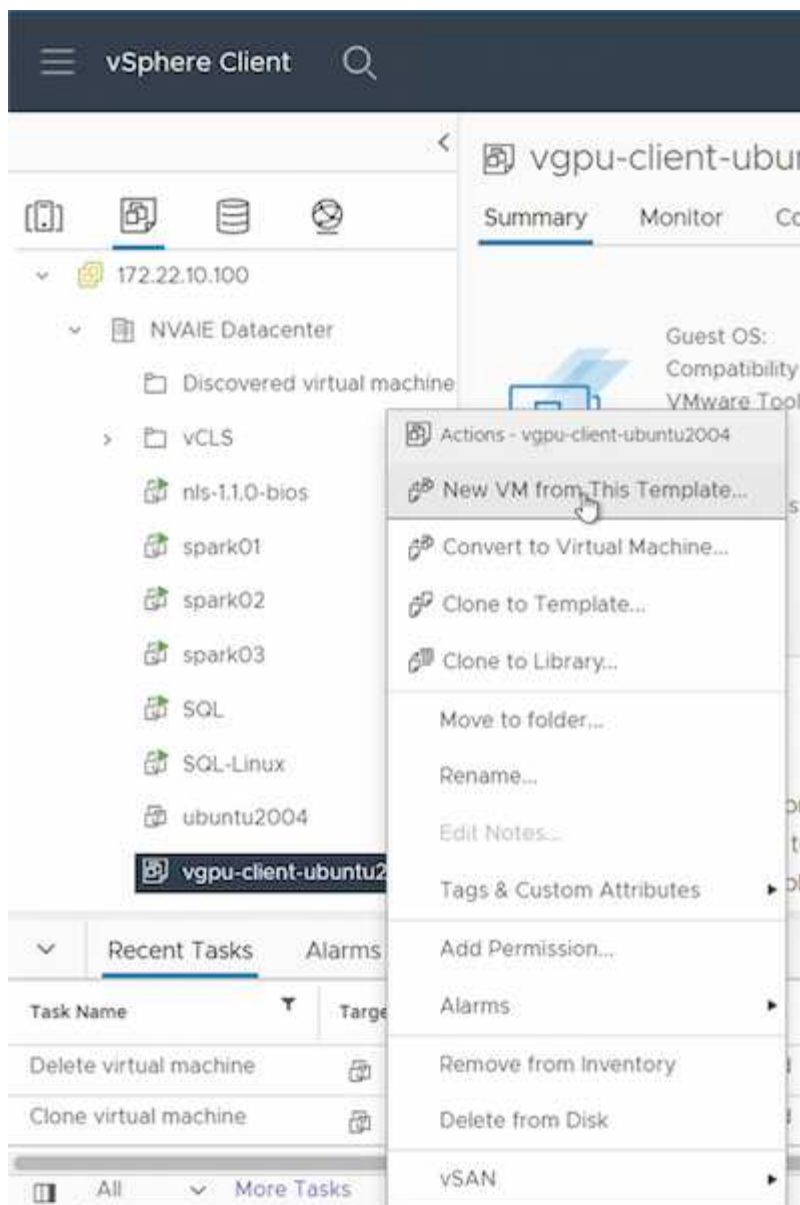
本节介绍在NVIDIA AI Enterprise 环境中执行 TensorFlow 训练作业需要执行的任务。

前提条件

在执行本节中概述的步骤之前，我们假设您已经按照["设置"](#)页。

从模板创建来宾虚拟机

首先，您必须根据上一节中创建的模板创建一个新的客户虚拟机。要从模板创建新的客户虚拟机，请登录 VMware vSphere，右键单击模板名称，选择“从此模板新建虚拟机...”，然后按照向导操作。



创建并挂载数据卷

接下来，您必须创建一个新的数据卷来存储您的训练数据集。您可以使用NetApp DataOps Toolkit 快速创建新的数据卷。下面的示例命令显示创建一个名为“imagenet”、容量为 2 TB 的卷。

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

在您用数据填充数据卷之前，您必须将其安装在来宾虚拟机中。您可以使用NetApp DataOps Toolkit 快速安装数据卷。下面的示例命令显示了上一步中创建的卷的安装。

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

填充数据量

在配置并安装新卷后，可以从源位置检索训练数据集并将其放置在新卷上。这通常涉及从 S3 或 Hadoop 数据湖中提取数据，有时还需要数据工程师的帮助。

执行 TensorFlow 训练作业

现在，您已准备好执行 TensorFlow 训练作业。要执行 TensorFlow 训练作业，请执行以下任务。

1. 拉取NVIDIA NGC 企业 TensorFlow 容器镜像。

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

2. 启动NVIDIA NGC 企业 TensorFlow 容器的实例。使用“-v”选项将数据卷附加到容器。

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

3. 在容器内执行您的 TensorFlow 训练程序。下面的示例命令展示了容器镜像中包含的示例 ResNet-50 训练程序的执行。

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

如何查找更多信息

要了解有关本文档中描述的信息的更多信息，请参阅以下文档和/或网站：

- NetApp ONTAP数据管理软件 — ONTAP信息库

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

- NetApp DataOps 工具包

<https://github.com/NetApp/netapp-dataops-toolkit>

- NVIDIA AI Enterprise 与 VMware

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

声明

- Bobby Oommen, NetApp高级经理
- Ramesh Isaac, NetApp系统管理员

- NetApp技术营销工程师 Roney Daniel

版权信息

版权所有 © 2026 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 <http://www.netapp.com/TM> 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。