采用NetApp和VMware的NVIDIA AI Enterprise

NetApp Solutions

NetApp
April 12, 2024

目录

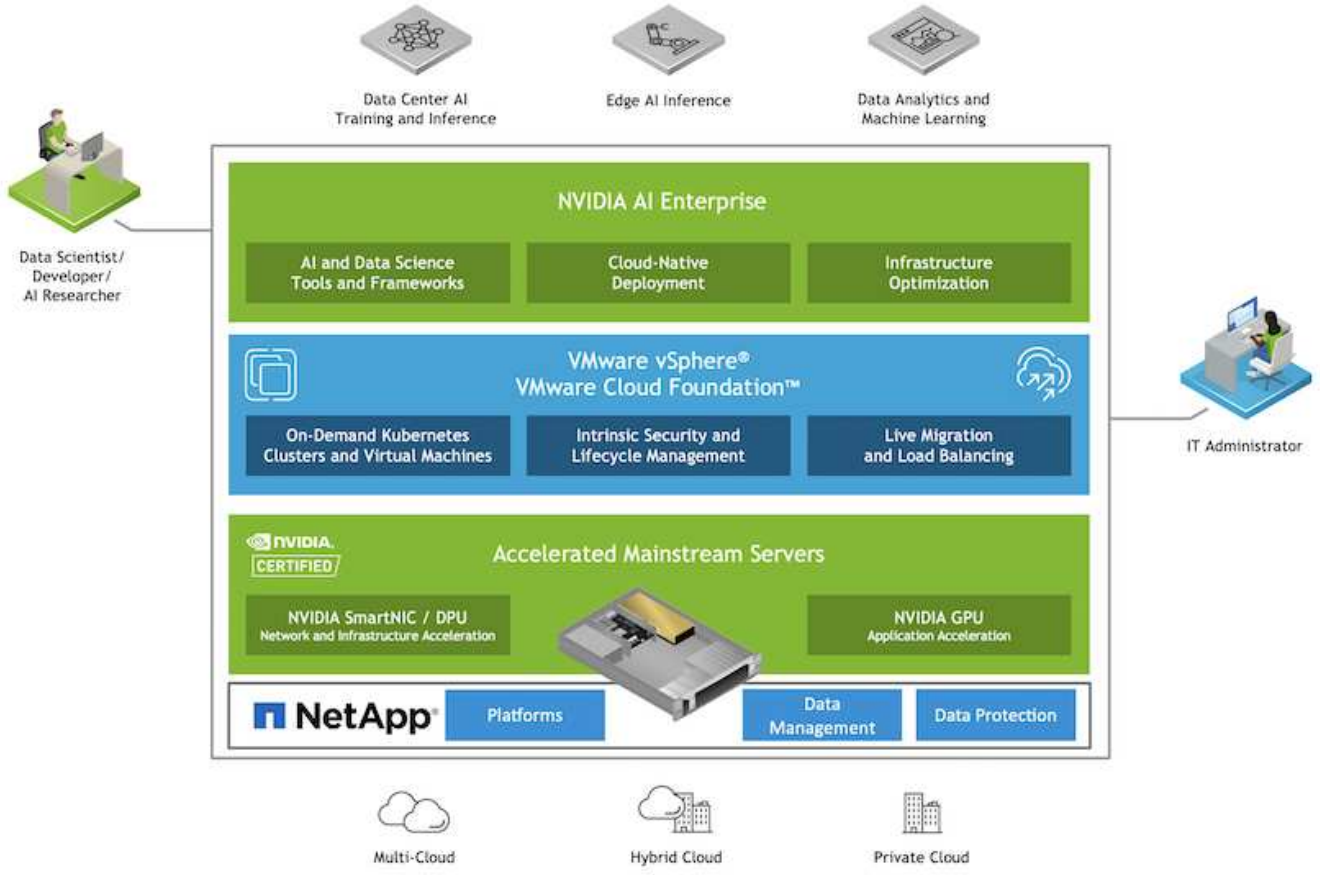
采用NetApp和VMware的NVIDIA AI Enterprise

采用NetApp和VMware的NVIDIA AI Enterprise

NetApp 公司 Mike Oglesby

对于IT架构师和管理员来说、AI工具可能非常复杂、而且不熟悉。此外、许多AI平台还没有为企业做好准备。由NetApp和VMware提供支持的NVIDIA AI Enterprise旨在提供简化的企业级AI架构。

NVIDIA AI Enterprise是一款端到端云原生AI和数据分析软件套件、经过NVIDIA优化、认证和支持、可在采用NVIDIA认证系统的VMware vSphere上运行。此软件有助于在现代混合云环境中轻松快速地部署、管理和扩展AI工作负载。由NetApp和VMware提供支持的NVIDIA AI Enterprise通过一个简单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

NVIDIA AI Enterprise

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

NVIDIA GPU Cloud （NGC）

NVIDIA NGC提供了一个GPU优化软件目录、供AI从业者开发其AI解决方案。此外、还可以访问各种AI服务、包括用于模型培训的NVIDIA Base Command、用于部署和监控模型的NVIDIA Baset Command以及用于安全访问和管理专有AI软件的NGC私有注册表。此外、NVIDIA AI Enterprise客户还可以通过NGC门户申请支持。

VMware vSphere

VMware vSphere是VMware的虚拟化平台、可将数据中心转变为包括CPU、存储和网络资源在内的聚合计算基础架构。vSphere将这些基础架构作为一个统一的操作环境进行管理、并为管理员提供用于管理参与该环境的数据中心的工具。

vSphere的两个核心组件是ESXi和vCenter Server。ESXi是一个虚拟化平台、管理员可以在此平台上创建和运行虚拟机和虚拟设备。vCenter Server是一项服务、管理员可以通过此服务管理连接到网络和池主机资源的多个主机。

NetApp ONTAP

ONTAP 9是NetApp推出的最新一代存储管理软件、可帮助企业打造现代化的基础架构并过渡到云就绪数据中心。借助行业领先的数据管理功能，无论数据位于何处， ONTAP 都可以通过一组工具来管理和保护数据。您还可以将数据自由移动到需要的任何位置：边缘，核心或云。ONTAP 9包含许多功能、可简化数据管理、加快和保护关键数据、并在混合云架构中实现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、这样才能将适当的资源用于AI应用程序和训练AI/ML数据集。以下有关NetApp技术的追加信息 不在此验证范围内、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：

- 实时数据缩减和扩展的重复数据删除。数据缩减可减少存储块中浪费的空间、重复数据删除可显著提高有效容量。此适用场景数据存储在本地，并分层到云。
- 最低、最高和自适应服务质量(AQoS)。精细的服务质量(QoS)控制有助于在高度共享的环境中保持关键应用程序的性能水平。
- NetApp FabricPool。可将冷数据自动分层到公有 和私有云存储选项、包括Amazon Web Services (AWS)、Azure和NetApp StorageGRID Storage解决方案。有关 FabricPool 的详细信息，请参见 "TR-4598：FabricPool 最佳实践"。

加速和保护数据

ONTAP 可提供卓越的性能和数据保护、并通过以下方式扩展这些功能：

- 性能和更低的延迟。ONTAP 可提供尽可能高的吞吐量和尽可能低的延迟。
- 数据保护ONTAP 可提供内置数据保护功能、并在所有平台之间进行通用管理。
- NetApp卷加密(NVE)。ONTAP 提供原生 卷级加密、并支持板载和外部密钥管理。
- 多租户和多因素身份验证。ONTAP 支持以最高的安全性级别共享基础架构资源。

Future-Proof 基础架构

ONTAP 可通过以下功能满足不断变化的苛刻业务需求：

- 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 NVMe 和 32 Gb FC 等最新技术，而无需进行成本高昂的数据迁移或中断。
- 云连接。ONTAP 是云互联程度最高的存储管理软件、可在所有公有 云中选择软件定义的存储(ONTAP Select)和云原生实例(NetApp Cloud Volumes Service)。
- 与新兴应用程序集成。ONTAP 通过使用支持现有企业应用程序的相同基础架构、为下一代平台和应用程序(例如自动驾驶汽车、智能城市和行业4.0)提供企业级数据服务。

NetApp DataOps 工具包

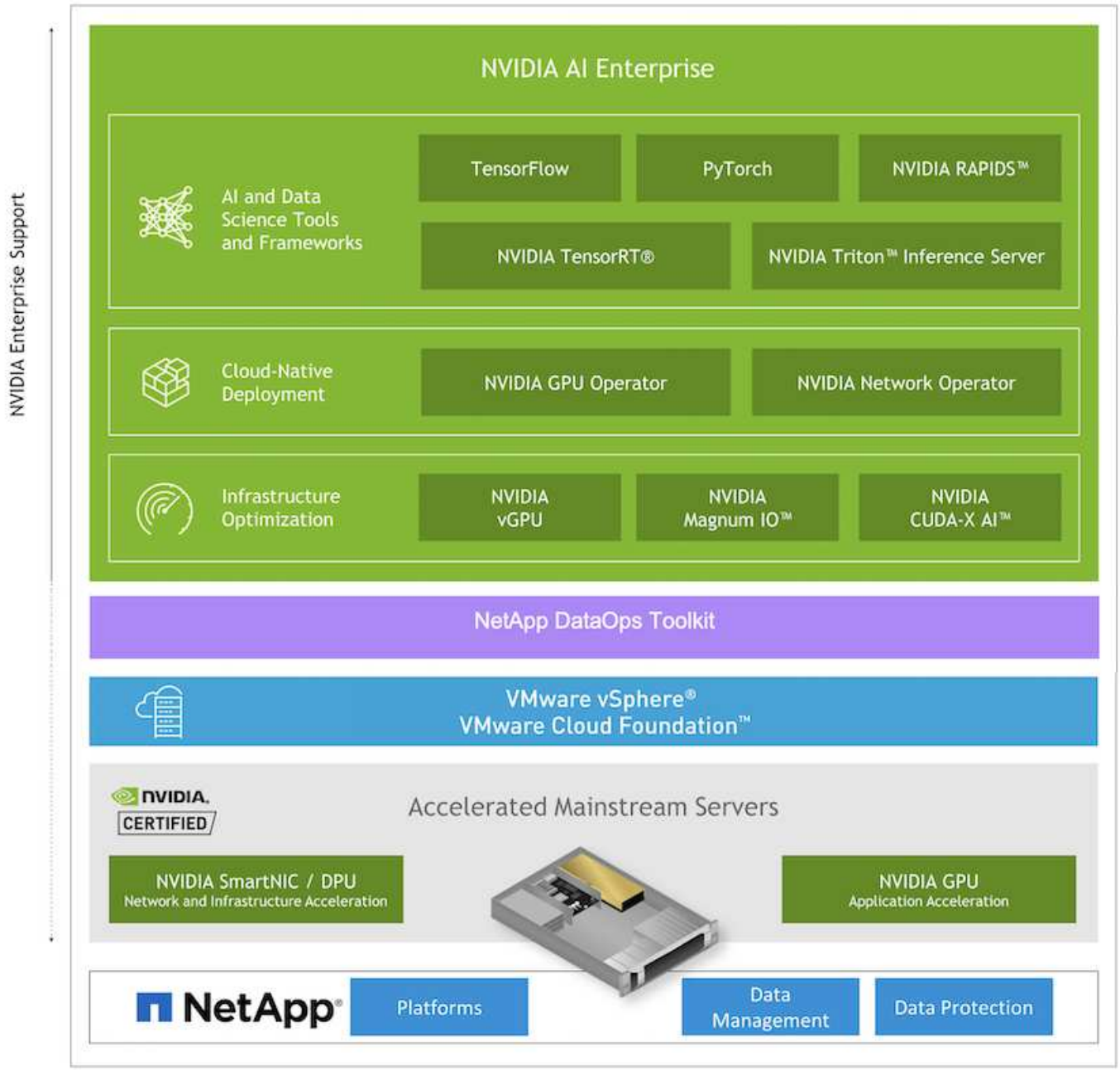
NetApp DataOps工具包是一款基于Python的工具、可简化开发/培训工作空间和推理服务器的管理、这些工作空间和服务器由高性能横向扩展NetApp存储提供支持。主要功能包括：

- 快速配置新的高容量JupyterLab工作空间、这些工作空间以高性能横向扩展NetApp存储为后盾。
- 快速配置由企业级NetApp存储提供支持的新NVIDIA Triton推理服务器实例。
- 可近乎即时地克隆高容量JupyterLab工作空间、以便进行实验或快速迭代。
- 可近乎即时地保存高容量JupyterLab工作空间的快照、以实现备份和/或可追溯性/基线化。
- 近乎即时地配置、克隆和快照高容量、高性能数据卷。

架构

此解决方案 基于经验证且熟悉的架构构建、该架构采用NetApp、VMware和NVIDIA认证系统。有关详细信息、请参见下表。

组件	详细信息
AI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

本节介绍在NetApp和VMware中使用NVIDIA AI Enterprise时需要执行的初始设置任务。

前提条件

在执行本节所述的步骤之前、我们假定您已部署VMware vSphere和NetApp ONTAP。请参见 "NVIDIA AI企业产品支持表" 有关受支持的vSphere版本的详细信息。请参见 "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 Entrprise主机软件、请按照中第1-4节所述的说明进行操作 "NVIDIA AI Enterprise快速入门指南"。

使用NVIDIA NGC软件

本节介绍在NVIDIA AI Enterprise环境中使用NVIDIA NGC企业软件需要执行的任务。

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

在执行本节所述的步骤之前、我们假定您已按照中所述的说明部署NVIDIA AI Entrprise主机软件 "初始设置" 页面。

使用vGPU创建Ubuntu子虚拟机

首先、您必须使用vGPU创建Ubuntu 20.04子虚拟机。要使用vGPU创建Ubuntu 20.04子虚拟机、请按照中概述的说明进行操作 "NVIDIA AI Enterprise部署指南"。

下载并安装NVIDIA子软件

接下来、您必须在上一步创建的子虚拟机中安装所需的NVIDIA子系统软件。要在子虚拟机中下载并安装所需的NVIDIA子软件、请按照中第5.1-5.4节所述的说明进行操作 "NVIDIA AI Enterprise快速入门指南"。

> 在执行第5.4节所述的验证任务时、您可能需要使用不同的CUDA容器映像版本标记、因为自编写本指南以来、CUDA容器映像已进行了更新。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、以便它们可以在子虚拟机中使用。要在子虚拟机中下载框架容器、请按照中所述的说明进行操作 "NVIDIA AI Enterprise部署指南"。

安装和配置NetApp DataOps工具包

接下来、您必须在子虚拟机中安装适用于传统环境的NetApp DataOps工具包。NetApp DataOps工具包可用于直接从子虚拟机中的终端管理ONTAP 系统上的横向扩展数据卷。要在子虚拟机中安装NetApp DataOps工具包、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. 从子虚拟机终端中注销、然后重新登录。
3. 配置NetApp DataOps工具包。要完成此步骤、您需要有关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。

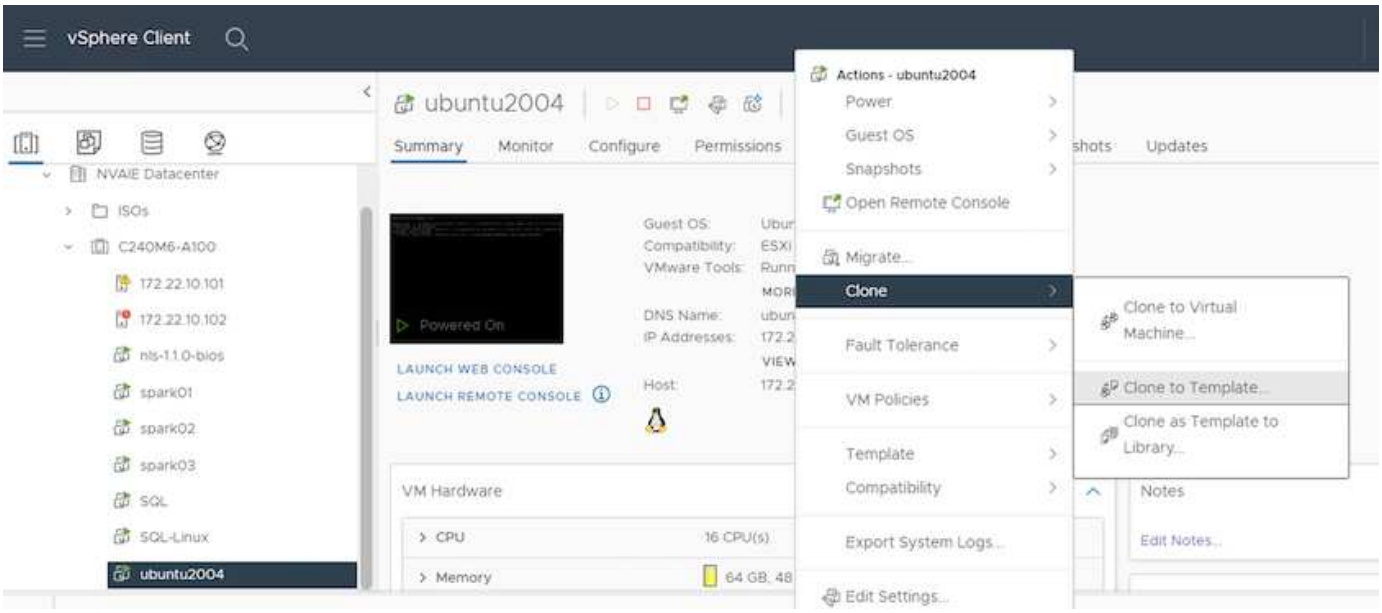
```
$ netapp_dataops_cli.py config

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'.
```

创建子虚拟机模板

最后、您必须根据子虚拟机创建VM模板。您可以使用此模板快速创建子虚拟机、以便使用NVIDIA NGC软件。

要基于来宾VM创建VM模板、请登录到VMware vSphere、然后右键单击来宾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培训作业

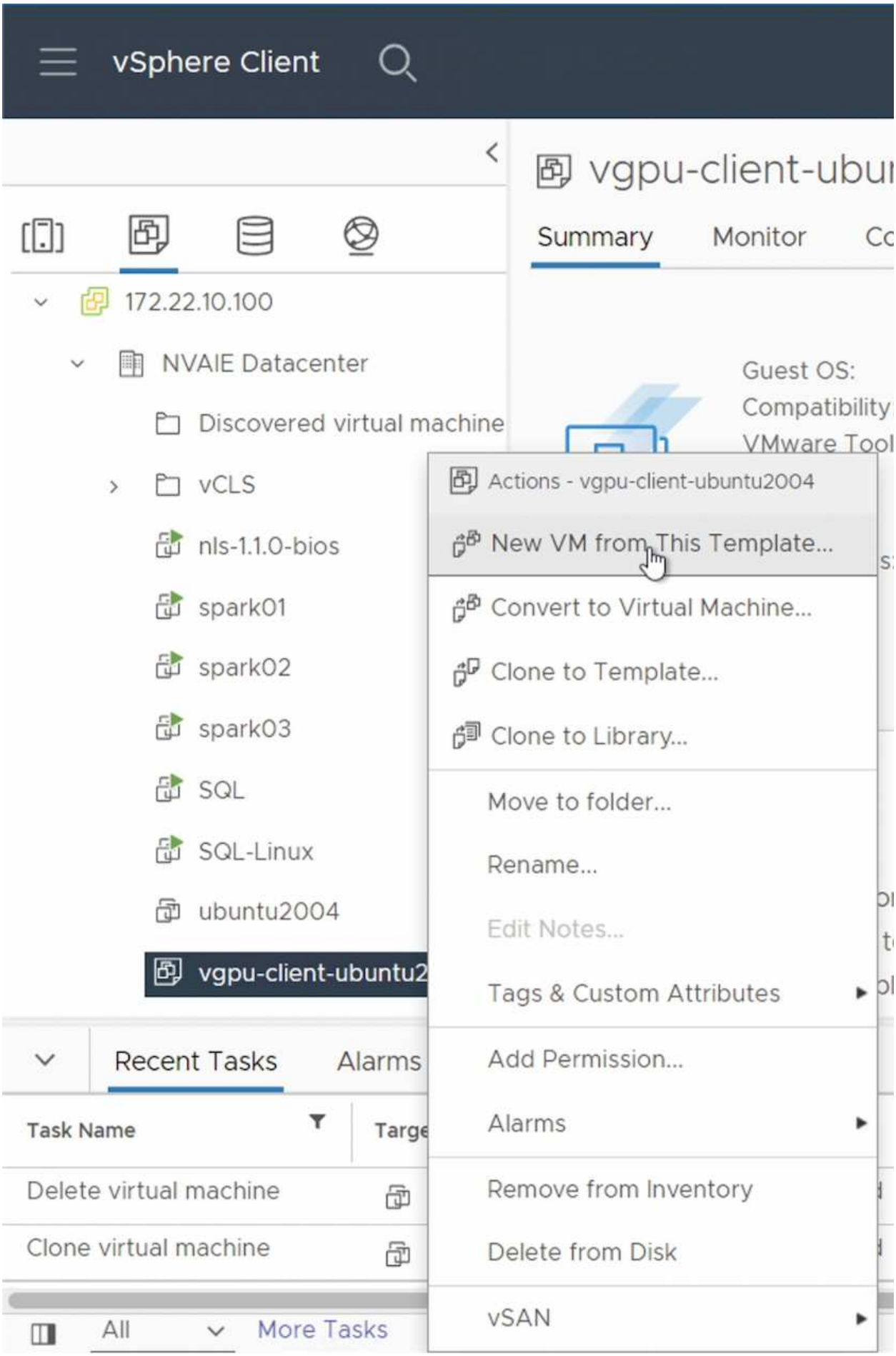
本节介绍在NVIDIA AI Enterprise环境中执行TensorFlow培训作业所需执行的任务。

前提条件

在执行本节所述的步骤之前、我们假定您已按照中所述的说明创建了子虚拟机模板 "设置" 页面。

使用模板创建子虚拟机

首先、您必须使用上一节中创建的模板创建新的子虚拟机。要使用模板创建新的子虚拟机、请登录到VMware vSphere、然后右键单击模板名称、选择"从此模板新建虚拟机…"、然后按照向导进行操作。


vSphere Client
vgpu-client-ubu
Summary
Monitor
172.22.10.100
NVAIE Datacenter
Discovered virtual machine
vCLS
nls-1.1.0-bios
spark01
spark02
spark03
SQL
SQL-Linux
ubuntu2004
vgpu-client-ubuntu2
Guest OS:
Compatibility
VMware Tool
Actions - vgpu-client-ubuntu2004
New VM from This Template...
Convert to Virtual Machine...
Clone to Template...
Clone to Library...
Move to folder...
Rename...
Edit Notes...
Tags & Custom Attributes
Add Permission...
Alarms
Remove from Inventory
Delete from Disk
vSAN
Recent Tasks
Alarms
Task Name
Delete virtual machine
Clone virtual machine
All
More Tasks

创建和挂载数据卷

接下来、您必须创建一个新的数据卷、用于存储培训数据集。您可以使用NetApp DataOps工具包快速创建新的数据卷。以下命令示例显示了如何创建容量为2 TB的名为"imagenet"的卷。

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

在为数据卷填充数据之前、必须先将其挂载到子虚拟机中。您可以使用NetApp DataOps工具包快速挂载数据卷。下面的示例命令显示了上一步创建的卷的布线。

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

填充数据卷

配置并挂载新卷后、可以从源位置检索培训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**培训作业

现在、您已准备好执行TensorFlow培训作业。要执行TensorFlow培训作业、请执行以下任务。

1. 提取NVIDIA NGC企业TensorFlow容器映像。

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

2. 启动NVIDIA NGC企业版TensorFlow容器的实例。使用"-v"选项将数据卷连接到容器。

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

3. 在容器中执行TensorFlow培训计划。下面的示例命令显示了容器映像中包含的示例RESNET-50培训计划的执行情况。

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

从何处查找追加信息

要了解有关本文档中所述信息的更多信息，请参见以下文档和 / 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"

- 采用VMware的NVIDIA AI Enterprise

  https://www.nvidia.com/en-us/data-center/products/ai-enterprise/vmware/^]

致谢

- Bobby Oommen、高级NetApp经理
- NetApp系统管理员Ramesh Isaac
- NetApp技术营销工程师Roney Daniel