NVIDIA AI Enterprise con NetApp y VMware

NetApp Solutions

NetApp
April 12, 2024

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

Tabla de contenidos

NVIDIA AI Enterprise con NetApp y VMware . . . 1
- NVIDIA AI Enterprise con NetApp y VMware . . . 1
- Visión general de la tecnología . . . 1
- Arquitectura . . . 4
- Configuración inicial . . . 5
- Utilice el software NVIDIA NGC . . . 5
- Dónde encontrar información adicional . . . 11

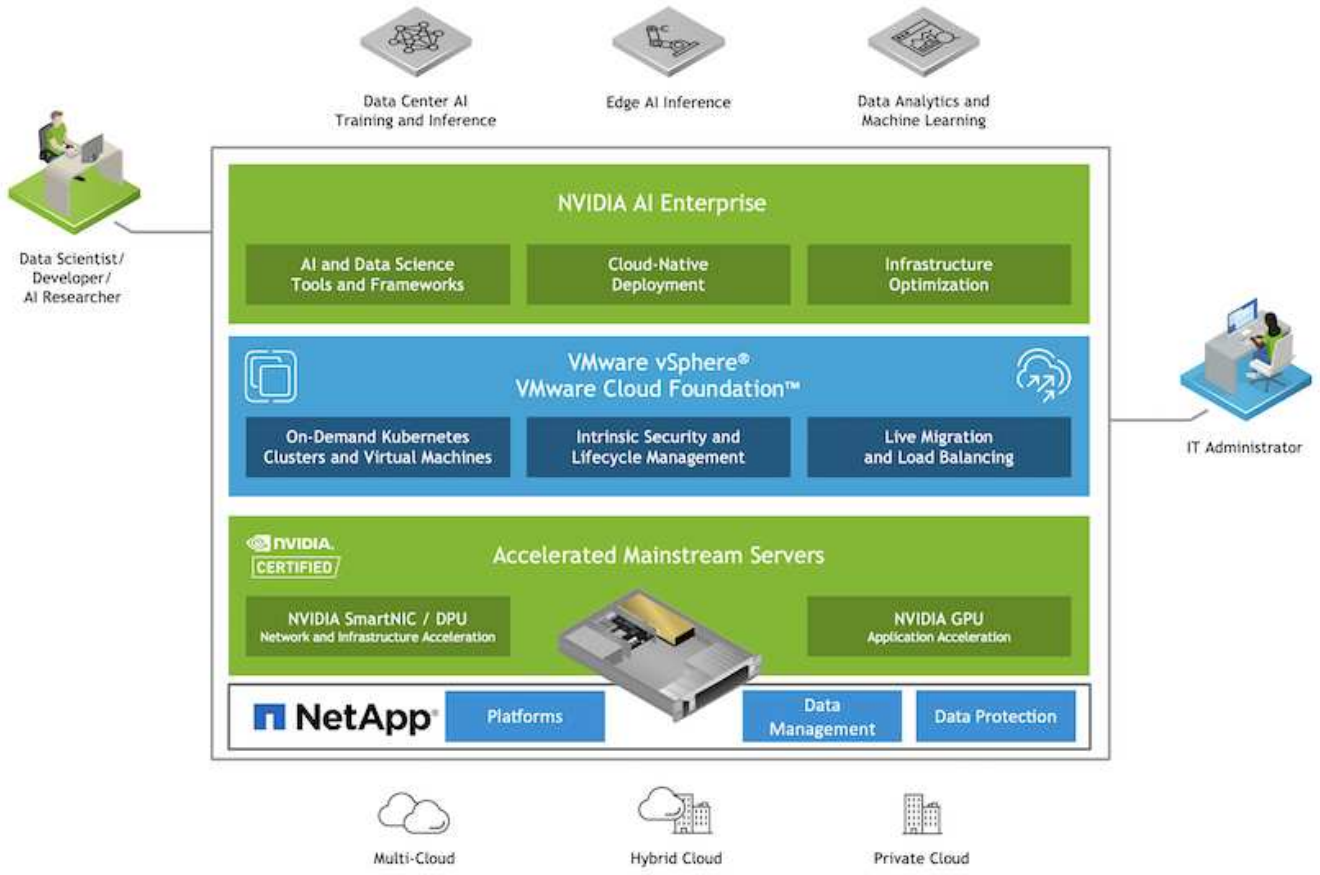
NVIDIA AI Enterprise con NetApp y VMware

NVIDIA AI Enterprise con NetApp y VMware

Mike Oglesby, NetApp

Para los arquitectos y administradores DE TECNOLOGÍA, las herramientas de IA pueden ser complicadas y no conocidas. Además, muchas plataformas de IA no están listas para las empresas. NVIDIA AI Enterprise, con la tecnología de NetApp y VMware, se ha creado para proporcionar una arquitectura de IA optimizada para la gran empresa.

NVIDIA AI Enterprise es una suite integral y nativa en el cloud de software de IA y análisis de datos optimizado, certificado y compatible con NVIDIA para ejecutarse en VMware vSphere con sistemas certificados por NVIDIA. Este software facilita la puesta en marcha, la gestión y el escalado simples y rápidos de las cargas de trabajo de IA en el entorno de cloud híbrido moderno. NVIDIA AI Enterprise, con la tecnología de NetApp y VMware, ofrece una excelente carga de trabajo de IA y gestión de datos en un paquete simplificado y conocido.



Visión general de la tecnología

IA Enterprise de NVIDIA

NVIDIA AI Enterprise es una suite integral y nativa en el cloud de software de IA y análisis de datos

optimizado, certificado y compatible con NVIDIA para ejecutarse en VMware vSphere con sistemas certificados por NVIDIA. Este software facilita la puesta en marcha, la gestión y el escalado simples y rápidos de las cargas de trabajo de IA en el entorno de cloud híbrido moderno.

GPU CLOUD DE NVIDIA (NGC)

NVIDIA NGC aloja un catálogo de software optimizado para GPU para que los profesionales de la IA puedan desarrollar sus soluciones de IA. También proporciona acceso a diversos servicios de IA, incluido NVIDIA base Command para el entrenamiento de modelos, NVIDIA Fleet Command para implementar y supervisar modelos, y el registro privado NGC para acceder y gestionar de forma segura el software de IA propio. Además, los clientes de NVIDIA AI Enterprise pueden solicitar soporte a través del portal NGC.

VSphere de VMware

VMware vSphere es la plataforma de virtualización de VMware, que transforma los centros de datos en infraestructuras informáticas globales que incluyen recursos de CPU, almacenamiento y red. VSphere gestiona estas infraestructuras como un entorno operativo unificado y proporciona a los administradores las herramientas necesarias para gestionar los centros de datos que participan en ese entorno.

Los dos componentes principales de vSphere son ESXi y vCenter Server. ESXi es la plataforma de virtualización donde los administradores crean y ejecutan máquinas virtuales y dispositivos virtuales. VCenter Server es el servicio a través del cual los administradores gestionan varios hosts conectados en una red y un pool de recursos de host.

ONTAP de NetApp

ONTAP 9, la última generación del software de gestión del almacenamiento de NetApp, permite a las empresas modernizar su infraestructura y realizar la transición a un centro de datos preparado para el cloud. ONTAP ofrece las mejores capacidades de gestión de datos y permite la gestión y protección de los datos con un solo conjunto de herramientas, sin importar dónde residan. También puede mover los datos libremente a donde sea necesario: El perímetro, el núcleo o el cloud. ONTAP 9 incluye numerosas funciones que simplifican la gestión de datos, aceleran y protegen los datos esenciales y permiten disfrutar de funcionalidades de infraestructura de nueva generación en arquitecturas de cloud híbrido.

Simplificar la gestión de los datos

La gestión de los datos es crucial para las operaciones TECNOLÓGICAS empresariales y los científicos de datos, para que se utilicen recursos apropiados para las aplicaciones de IA y para entrenar conjuntos de datos de IA/ML. La siguiente información adicional sobre las tecnologías de NetApp no está disponible para esta validación, pero puede ser relevante en función de su puesta en marcha.

El software para la gestión de datos ONTAP incluye las siguientes funciones para mejorar y simplificar las operaciones, y reducir el coste total de funcionamiento:

- Compactación de datos inline y deduplicación expandida. La compactación de datos reduce el espacio perdido dentro de los bloques de almacenamiento, mientras que la deduplicación aumenta la capacidad efectiva de forma significativa. Esto es aplicable a los datos almacenados localmente y a los datos organizados en niveles en el cloud.
- Calidad de servicio (AQoS) mínima, máxima y adaptativa. Los controles granulares de calidad de servicio (QoS) ayudan a mantener los niveles de rendimiento para aplicaciones críticas en entornos altamente compartidos.
- FabricPool de NetApp. Proporciona la organización automática en niveles de datos fríos en opciones de almacenamiento en cloud privado como Amazon Web Services (AWS), Azure y la solución de

almacenamiento StorageGRID de NetApp. Para obtener más información sobre FabricPool, consulte "TR-4598: Prácticas recomendadas de FabricPool".

Acelere y proteja sus datos

ONTAP no solo ofrece niveles de rendimiento y protección de datos superiores, sino que amplía estas capacidades de las siguientes maneras:

- Rendimiento y menor latencia. ONTAP ofrece la salida más alta posible con la menor latencia posible.
- Protección de datos. ONTAP ofrece capacidades integradas de protección de datos, con una administración común entre todas las plataformas.
- Cifrado de volúmenes de NetApp (NVE). ONTAP ofrece cifrado nativo en el nivel de volumen y permite la gestión de claves incorporada o externa.
- Multi-tenancy y autenticación multifactor. ONTAP permite compartir recursos de infraestructura con los niveles más altos de seguridad.

Infraestructura preparada para futuros retos

ONTAP ayuda a satisfacer las exigentes y siempre cambiantes necesidades de su empresa con las siguientes funciones:

- Escalado sencillo y operaciones no disruptivas. ONTAP admite la adición no disruptiva de capacidad a las controladoras existentes y a clústeres de escalado horizontal. Los clientes pueden empezar a utilizar tecnologías punteras como NVMe y FC 32 GB, sin necesidad de realizar costosas migraciones de datos y sin cortes.
- Conexión de cloud. ONTAP es el software de gestión de almacenamiento con mejor conexión de cloud e incluye opciones de almacenamiento definido por software (ONTAP Select) e instancias nativas del cloud (NetApp Cloud Volumes Service) en todos los clouds públicos.
- Integración con aplicaciones emergentes. ONTAP ofrece servicios de datos de clase empresarial para plataformas y aplicaciones de última generación, como vehículos autónomos, ciudades inteligentes e Industria 4.0, utilizando la misma infraestructura que da soporte a las aplicaciones empresariales existentes.

Kit de herramientas de operaciones de datos de NetApp

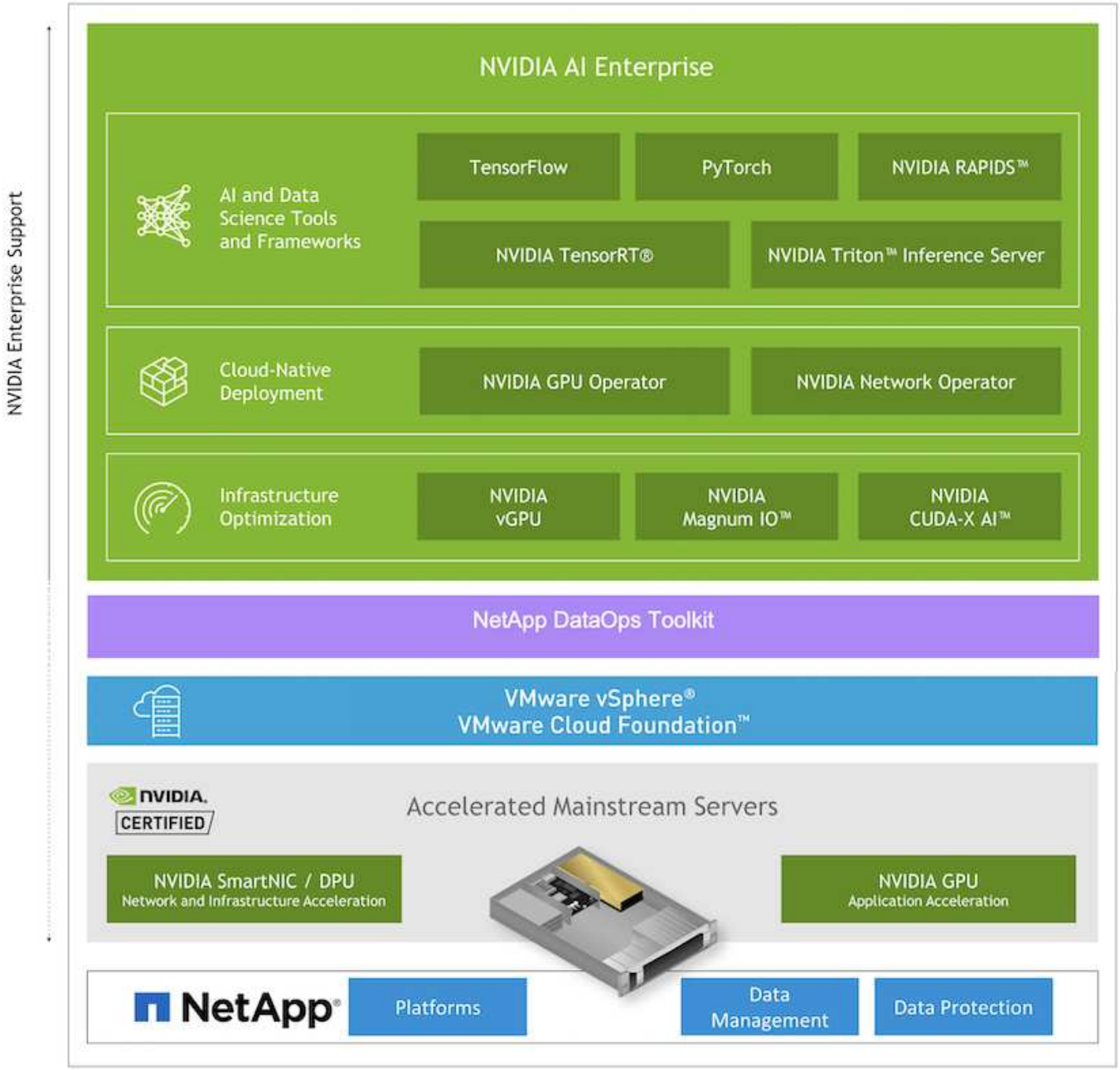
El kit de herramientas DataOps de NetApp es una herramienta basada en Python que simplifica la gestión de espacios de trabajo de desarrollo/formación y servidores de inferencia respaldados por un almacenamiento de NetApp de escalado horizontal y de alto rendimiento. Estas son algunas funcionalidades clave:

- Aprovisione con rapidez nuevos espacios de trabajo de JuppyterLab de alta capacidad respaldados por el almacenamiento de NetApp de escalado horizontal y de alto rendimiento.
- Aprovisione rápidamente nuevas instancias del servidor de inferencia de NVIDIA Triton, respaldadas por un almacenamiento empresarial de NetApp.
- Clone por vía casi instreutilizada los espacios de trabajo de JuppyterLab de gran capacidad para permitir la experimentación o la iteración rápida.
- Guarde copias Snapshot casi por vía inutilizada de espacios de trabajo JuppyterLab de gran capacidad para backup o seguimiento o creación de bases.
- Aprovisionamiento, clonado y copias Snapshot de alta capacidad y alto rendimiento casi por vía casi instusand.

Arquitectura

Esta solución se basa en una arquitectura probada y familiar que incluye NetApp, VMware y sistemas con certificación NVIDIA. Consulte la siguiente tabla para obtener más información.

Componente	Detalles
Software de IA y análisis de datos	"NVIDIA AI Enterprise para VMware"
Plataforma de virtualización	"VSphere de VMware"
Plataforma informática	"Sistemas con certificación NVIDIA"
Plataforma de gestión de datos	"ONTAP de NetApp"



Configuración inicial

En esta sección se describen las tareas de configuración inicial que se deben realizar para utilizar NVIDIA AI Enterprise con NetApp y VMware.

Requisitos previos

Antes de realizar los pasos descritos en esta sección, asumimos que ya ha implementado VMware vSphere y ONTAP de NetApp. Consulte la "Matriz de compatibilidad de productos empresariales de NVIDIA AI" Para obtener más detalles sobre las versiones de vSphere compatibles. Consulte la "Documentación sobre las soluciones de NetApp y VMware" Para obtener más información sobre cómo implementar VMware vSphere con ONTAP de NetApp.

Instale el software de host de NVIDIA AI Enterprise

Para instalar el software host NVIDIA AI Entrprise, siga las instrucciones que se describen en las secciones 1-4 de la "Guía de inicio rápido de NVIDIA AI Enterprise".

Utilice el software NVIDIA NGC

En esta sección, se describen las tareas que se deben realizar para utilizar el software empresarial NVIDIA NGC en un entorno de NVIDIA AI Enterprise.

Configuración

En esta sección se describen las tareas de configuración inicial que se deben realizar para utilizar el software empresarial NVIDIA NGC en un entorno de NVIDIA AI Enterprise.

Requisitos previos

Antes de realizar los pasos descritos en esta sección, asumimos que ya ha implementado el software host NVIDIA AI Entrprise siguiendo las instrucciones que se describen en la "Configuración inicial" página.

Cree una máquina virtual Ubuntu Guest con vGPU

En primer lugar, debe crear una máquina virtual invitada Ubuntu 20.04 con vGPU. Para crear una VM huésped de Ubuntu 20.04 con vGPU, siga las instrucciones descritas en el "Guía de puesta en marcha de NVIDIA AI Enterprise".

Descargue e instale el software invitado de NVIDIA

A continuación, debe instalar el software invitado NVIDIA requerido en la máquina virtual invitada que creó en el paso anterior. Para descargar e instalar el software invitado NVIDIA necesario en el equipo virtual invitado, siga las instrucciones que se describen en las secciones 5.1-5.4 de la "Guía de inicio rápido de NVIDIA AI Enterprise".

> Al realizar las tareas de verificación descritas en la sección 5.4, es posible que necesite utilizar una etiqueta de versión de imagen de contenedor CUDA diferente ya que la imagen de contenedor CUDA se ha actualizado desde la escritura de la guía. En nuestra validación, utilizamos 'nvidia/cuda:11.0.3-base-ubuntu20.04'.

Descargue contenedores de marcos de análisis/IA

A continuación, debe descargar las imágenes de contenedor de IA o marco de análisis necesarias de NVIDIA NGC para que estén disponibles en la máquina virtual invitada. Para descargar los contenedores de marco en el equipo virtual invitado, siga las instrucciones que se describen en "Guía de puesta en marcha de NVIDIA AI Enterprise".

Instalación y configuración del kit de herramientas de operaciones de datos de NetApp

A continuación, debe instalar el kit de herramientas DataOPS de NetApp para entornos tradicionales en el equipo virtual invitado. El kit de herramientas DataOPS de NetApp se puede usar para gestionar los volúmenes de datos de escalado horizontal en su sistema ONTAP directamente desde el terminal dentro del equipo virtual «guest». Para instalar el kit de herramientas DataOPS de NetApp en el equipo virtual invitado, realice las siguientes tareas.

1. Instalar la tubería.

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

2. Cierre la sesión en el terminal de la máquina virtual invitada y vuelva a iniciarla.
3. Configurar el kit de herramientas de operaciones de datos de NetApp. Para completar este paso, necesitará detalles de acceso a la API para su sistema ONTAP. Es posible que tenga que obtenerlos del administrador de almacenamiento.

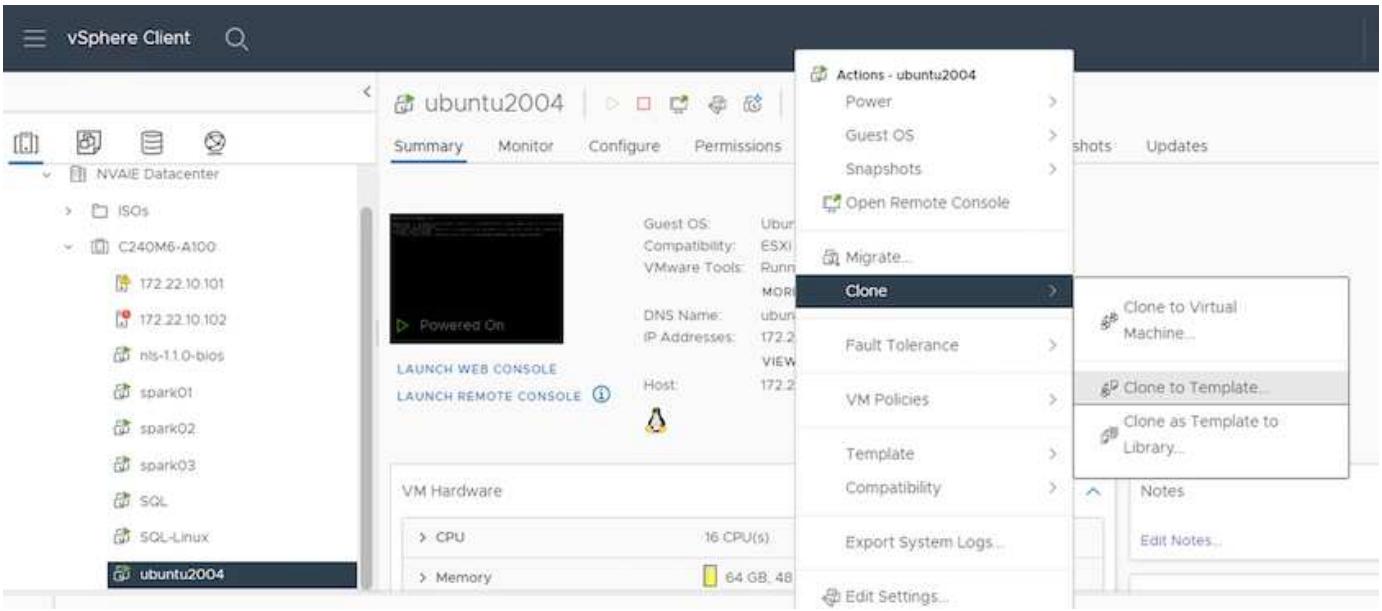
```
$ netapp_dataops_cli.py config

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'.
```

Cree una plantilla de máquina virtual invitada

Por último, debe crear una plantilla de equipo virtual basada en el equipo virtual «guest». Podrá utilizar esta plantilla para crear rápidamente equipos virtuales invitados con el software NVIDIA NGC.

Si desea crear una plantilla de máquina virtual basada en su equipo virtual invitado, inicie sesión en VMware vSphere, haga clic con el botón derecho en el nombre de la máquina virtual invitada, seleccione 'Clonar', elija 'Clonar en plantilla…' y, a continuación, siga el asistente.



Ejemplo de caso práctico: Trabajo de formación de TensorFlow

En esta sección se describen las tareas que se deben realizar para ejecutar una tarea de entrenamiento de TensorFlow dentro de un entorno NVIDIA AI Enterprise.

Requisitos previos

Antes de realizar los pasos descritos en esta sección, asumimos que ya ha creado una plantilla de equipo virtual invitado siguiendo las instrucciones que se describen en la "Configuración" página.

Crear máquina virtual invitada a partir de la plantilla

En primer lugar, debe crear un nuevo equipo virtual invitado a partir de la plantilla que ha creado en la sección anterior. Para crear una máquina virtual invitada nueva desde la plantilla, inicie sesión en VMware vSphere, haga clic en el nombre de la plantilla, seleccione "New VM from this Template…" y, a continuación, siga el asistente.

vSphere Client

vgpu-client-ubu

Summary Monitor Co

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 Targe

Delete virtual machine

Clone virtual machine

All More Tasks

Cree y monte el volumen de datos

A continuación, debe crear un nuevo volumen de datos en el que almacenar el conjunto de datos de entrenamiento. Puede crear rápidamente un nuevo volumen de datos con el kit de herramientas de operaciones de datos de NetApp. El siguiente comando de ejemplo muestra la creación de un volumen llamado 'imagenet' con una capacidad de 2 TB.

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

Antes de poder completar el volumen de datos con datos, debe montarlo en la máquina virtual invitada. Puede montar rápidamente un volumen de datos con el kit de herramientas Data OPS de NetApp. El comando de ejemplo siguiente muestra el bigote del volumen que se creó en el paso anterior.

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

Completar volumen de datos

Una vez aprovisionado y montado el volumen nuevo, el conjunto de datos de entrenamiento puede recuperarse de la ubicación de origen y colocarse en el volumen nuevo. Normalmente, esto implica extraer los datos de un lago de datos de S3 o Hadoop y a veces contará con la ayuda de un ingeniero de datos.

Ejecutar la tarea de formación de TensorFlow

Ahora, estará listo para ejecutar su trabajo de formación de TensorFlow. Para ejecutar el trabajo de entrenamiento de TensorFlow, realice las siguientes tareas.

1. Tire de la imagen del contenedor NVIDIA NGC para TensorFlow empresarial.

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

2. Inicie una instancia del contenedor NVIDIA NGC para TensorFlow empresarial. Utilice la opción '-v' para adjuntar el volumen de datos al contenedor.

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

3. Ejecute su programa de formación TensorFlow en el contenedor. El comando de ejemplo siguiente muestra la ejecución de un programa de entrenamiento ResNet-50 de ejemplo que se incluye en la imagen contenedora.

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

Dónde encontrar información adicional

Para obtener más información sobre la información descrita en este documento, consulte los siguientes documentos y/o sitios web:

- Software de gestión de datos ONTAP de NetApp: Biblioteca de información de ONTAP

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

- Kit de herramientas de operaciones de datos de NetApp

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

- NVIDIA AI Enterprise con VMware

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

Reconocimientos

- Bobby Oommen, Sr. Sénior, NetApp
- Ramesh Isaac, Administrador de sistemas, NetApp
- Roney Daniel, Ingeniero Técnico de Marketing de NetApp

Información de copyright

Copyright © 2024 NetApp, Inc. Todos los derechos reservados. Imprimido en EE. UU. No se puede reproducir este documento protegido por copyright ni parte del mismo de ninguna forma ni por ningún medio (gráfico, electrónico o mecánico, incluidas fotocopias, grabaciones o almacenamiento en un sistema de recuperación electrónico) sin la autorización previa y por escrito del propietario del copyright.

El software derivado del material de NetApp con copyright está sujeto a la siguiente licencia y exención de responsabilidad:

ESTE SOFTWARE LO PROPORCIONA NETAPP «TAL CUAL» Y SIN NINGUNA GARANTÍA EXPRESA O IMPLÍCITA, INCLUYENDO, SIN LIMITAR, LAS GARANTÍAS IMPLÍCITAS DE COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN CONCRETO, CUYA RESPONSABILIDAD QUEDA EXIMIDA POR EL PRESENTE DOCUMENTO. EN NINGÚN CASO NETAPP SERÁ RESPONSABLE DE NINGÚN DAÑO DIRECTO, INDIRECTO, ESPECIAL, EJEMPLAR O RESULTANTE (INCLUYENDO, ENTRE OTROS, LA OBTENCIÓN DE BIENES O SERVICIOS SUSTITUTIVOS, PÉRDIDA DE USO, DE DATOS O DE BENEFICIOS, O INTERRUPCIÓN DE LA ACTIVIDAD EMPRESARIAL) CUALQUIERA SEA EL MODO EN EL QUE SE PRODUJERON Y LA TEORÍA DE RESPONSABILIDAD QUE SE APLIQUE, YA SEA EN CONTRATO, RESPONSABILIDAD OBJETIVA O AGRAVIO (INCLUIDA LA NEGLIGENCIA U OTRO TIPO), QUE SURJAN DE ALGÚN MODO DEL USO DE ESTE SOFTWARE, INCLUSO SI HUBIEREN SIDO ADVERTIDOS DE LA POSIBILIDAD DE TALES DAÑOS.

NetApp se reserva el derecho de modificar cualquiera de los productos aquí descritos en cualquier momento y sin aviso previo. NetApp no asume ningún tipo de responsabilidad que surja del uso de los productos aquí descritos, excepto aquello expresamente acordado por escrito por parte de NetApp. El uso o adquisición de este producto no lleva implícita ninguna licencia con derechos de patente, de marcas comerciales o cualquier otro derecho de propiedad intelectual de NetApp.

Es posible que el producto que se describe en este manual esté protegido por una o más patentes de EE. UU., patentes extranjeras o solicitudes pendientes.

LEYENDA DE DERECHOS LIMITADOS: el uso, la copia o la divulgación por parte del gobierno están sujetos a las restricciones establecidas en el subpárrafo (b)(3) de los derechos de datos técnicos y productos no comerciales de DFARS 252.227-7013 (FEB de 2014) y FAR 52.227-19 (DIC de 2007).

Los datos aquí contenidos pertenecen a un producto comercial o servicio comercial (como se define en FAR 2.101) y son propiedad de NetApp, Inc. Todos los datos técnicos y el software informático de NetApp que se proporcionan en este Acuerdo tienen una naturaleza comercial y se han desarrollado exclusivamente con fondos privados. El Gobierno de EE. UU. tiene una licencia limitada, irrevocable, no exclusiva, no transferible, no sublicenciable y de alcance mundial para utilizar los Datos en relación con el contrato del Gobierno de los Estados Unidos bajo el cual se proporcionaron los Datos. Excepto que aquí se disponga lo contrario, los Datos no se pueden utilizar, desvelar, reproducir, modificar, interpretar o mostrar sin la previa aprobación por escrito de NetApp, Inc. Los derechos de licencia del Gobierno de los Estados Unidos de América y su Departamento de Defensa se limitan a los derechos identificados en la cláusula 252.227-7015(b) de la sección DFARS (FEB de 2014).

Información de la marca comercial

NETAPP, el logotipo de NETAPP y las marcas que constan en http://www.netapp.com/TM son marcas comerciales de NetApp, Inc. El resto de nombres de empresa y de producto pueden ser marcas comerciales de sus respectivos propietarios.