NVIDIA ai Enterprise con NetApp e VMware

NetApp Solutions

NetApp
April 26, 2024

Sommario

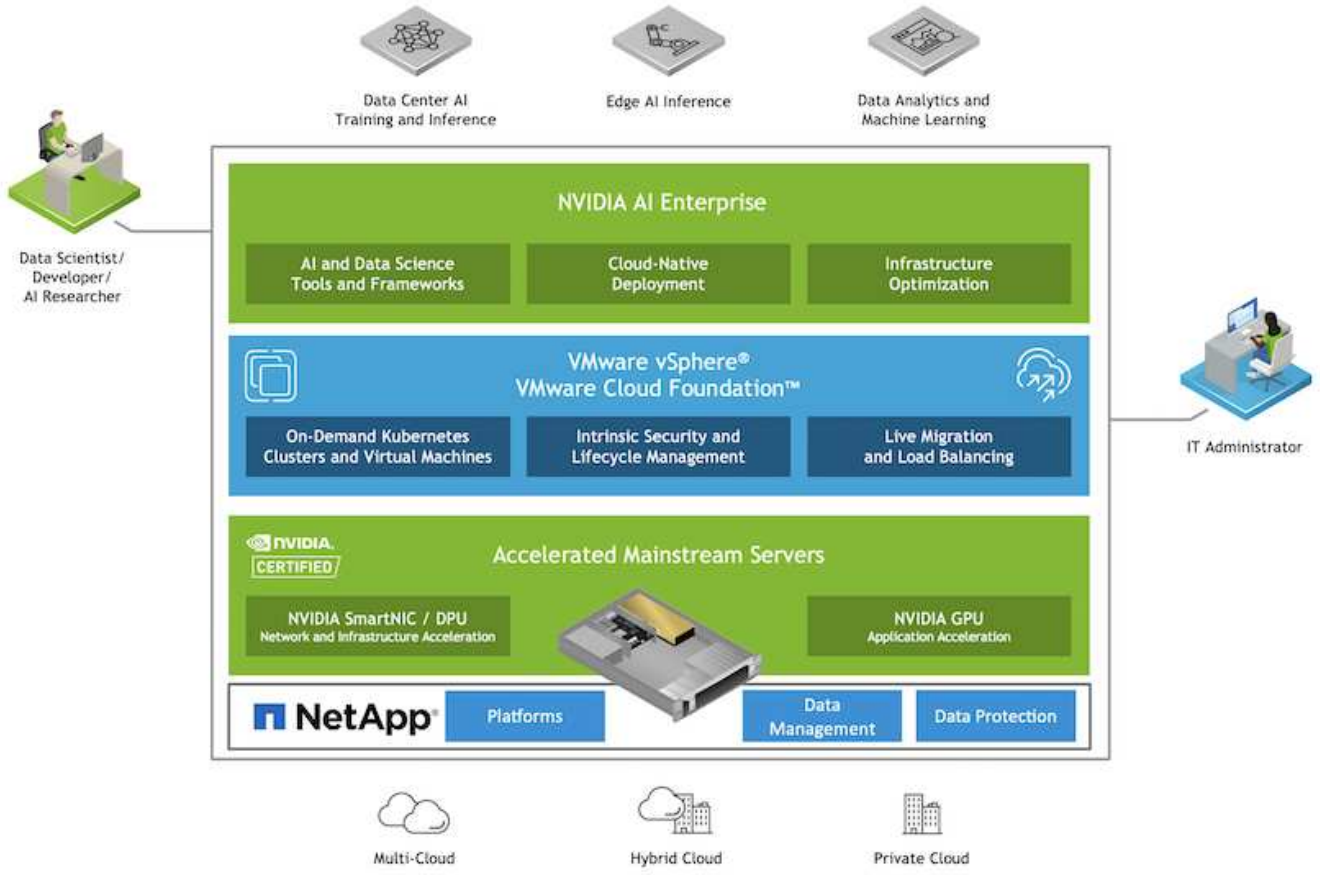
NVIDIA ai Enterprise con NetApp e VMware

NVIDIA ai Enterprise con NetApp e VMware

Mike Oglesby, NetApp

Per gli architetti E gli amministratori IT, gli strumenti di ai possono essere complicati e poco familiari. Inoltre, molte piattaforme ai non sono Enterprise-ready. NVIDIA ai Enterprise, basata su NetApp e VMware, è stata creata per offrire un'architettura ai di livello Enterprise ottimizzata.

NVIDIA ai Enterprise è una suite end-to-end nativa del cloud di software di ai e data analytics ottimizzato, certificato e supportato da NVIDIA per l'esecuzione su VMware vSphere con sistemi certificati NVIDIA. Questo software facilita l'implementazione, la gestione e la scalabilità semplici e rapide dei carichi di lavoro ai nel moderno ambiente di cloud ibrido. NVIDIA ai Enterprise, basata su NetApp e VMware, offre gestione dei dati e dei workload ai di livello Enterprise in un pacchetto semplice e familiare.



Panoramica sulla tecnologia

NVIDIA ai Enterprise

NVIDIA ai Enterprise è una suite end-to-end nativa del cloud di software di ai e data analytics ottimizzato, certificato e supportato da NVIDIA per l'esecuzione su VMware vSphere con sistemi certificati NVIDIA. Questo

software facilita l'implementazione, la gestione e la scalabilità semplici e rapide dei carichi di lavoro ai nel moderno ambiente di cloud ibrido.

NVIDIA GPU CLOUD (NGC)

NVIDIA NGC ospita un catalogo di software ottimizzato per GPU per i professionisti dell'ai per sviluppare le proprie soluzioni di ai. Fornisce inoltre accesso a vari servizi di ai, tra cui NVIDIA base Command per la formazione sui modelli, NVIDIA Fleet Command per l'implementazione e il monitoraggio dei modelli e NGC Private Registry per l'accesso e la gestione sicuri del software proprietario di ai. Inoltre, i clienti di NVIDIA ai Enterprise possono richiedere il supporto tramite il portale NGC.

VMware vSphere

VMware vSphere è la piattaforma di virtualizzazione di VMware, che trasforma i data center in infrastrutture di calcolo aggregate che includono CPU, storage e risorse di rete. VSphere gestisce queste infrastrutture come un ambiente operativo unificato e fornisce agli amministratori gli strumenti per gestire i data center che partecipano a tale ambiente.

I due componenti principali di vSphere sono ESXi e vCenter Server. ESXi è la piattaforma di virtualizzazione in cui gli amministratori creano ed eseguono macchine virtuali e appliance virtuali. VCenter Server è il servizio attraverso il quale gli amministratori gestiscono più host connessi in una rete e pool di risorse host.

NetApp ONTAP

ONTAP 9, l'ultima generazione di software per la gestione dello storage NetApp, consente alle aziende di modernizzare l'infrastruttura e passare a un data center predisposto per il cloud. Sfruttando le funzionalità di gestione dei dati leader del settore, ONTAP consente la gestione e la protezione dei dati con un singolo set di strumenti, indipendentemente dalla posizione dei dati. Puoi anche spostare liberamente i dati ovunque siano necessari: Edge, core o cloud. ONTAP 9 include numerose funzionalità che semplificano la gestione dei dati, accelerano e proteggono i dati critici e abilitano le funzionalità dell'infrastruttura di nuova generazione nelle architetture di cloud ibrido.

Semplifica la gestione dei dati

La gestione dei dati è fondamentale per le operazioni IT aziendali e per i data scientist, in modo che le risorse appropriate vengano utilizzate per le applicazioni ai e per la formazione dei set di dati ai/ML. Le seguenti informazioni aggiuntive sulle tecnologie NetApp non rientrano nell'ambito di questa convalida, ma potrebbero essere rilevanti a seconda dell'implementazione.

Il software per la gestione dei dati ONTAP include le seguenti funzionalità per ottimizzare e semplificare le operazioni e ridurre il costo totale delle operazioni:

- Compaction dei dati inline e deduplica estesa. La compattazione dei dati riduce lo spazio sprecato all'interno dei blocchi di storage e la deduplica aumenta significativamente la capacità effettiva. Ciò vale per i dati memorizzati localmente e per i dati a più livelli nel cloud.
- Qualità del servizio (AQoS) minima, massima e adattativa. I controlli granulari della qualità del servizio (QoS) aiutano a mantenere i livelli di performance per le applicazioni critiche in ambienti altamente condivisi.
- NetApp FabricPool. Offre il tiering automatico dei dati cold per le opzioni di cloud storage pubblico e privato, tra cui Amazon Web Services (AWS), Azure e la soluzione di storage NetApp StorageGRID. Per ulteriori informazioni su FabricPool, vedere "TR-4598: Best practice FabricPool".

Accelera e proteggi i dati

ONTAP offre livelli superiori di performance e protezione dei dati ed estende queste funzionalità nei seguenti modi:

- Performance e latenza ridotta. ONTAP offre il throughput più elevato possibile con la latenza più bassa possibile.
- Protezione dei dati. ONTAP offre funzionalità di protezione dei dati integrate con gestione comune su tutte le piattaforme.
- NetApp Volume Encryption (NVE). ONTAP offre crittografia nativa a livello di volume con supporto per la gestione delle chiavi sia integrata che esterna.
- Multi-tenancy e autenticazione a più fattori. ONTAP consente la condivisione delle risorse dell'infrastruttura con i massimi livelli di sicurezza.

Infrastruttura a prova di futuro

ONTAP aiuta a soddisfare le esigenze di business esigenti e in continua evoluzione con le seguenti funzionalità:

- Scalabilità perfetta e operazioni senza interruzioni. ONTAP supporta l'aggiunta senza interruzioni di capacità ai controller esistenti e ai cluster scale-out. I clienti possono eseguire l'upgrade alle tecnologie più recenti, come NVMe e 32GB FC, senza costose migrazioni dei dati o interruzioni.
- Connessione al cloud. ONTAP è il software per la gestione dello storage più connesso al cloud, con opzioni per lo storage software-defined (ONTAP Select) e le istanze native del cloud (NetApp Cloud Volumes Service) in tutti i cloud pubblici.
- Integrazione con le applicazioni emergenti. ONTAP offre servizi dati di livello Enterprise per piattaforme e applicazioni di prossima generazione, come veicoli autonomi, città intelligenti e industria 4.0, utilizzando la stessa infrastruttura che supporta le applicazioni aziendali esistenti.

NetApp DataOps Toolkit

Il NetApp DataOps Toolkit è uno strumento basato su Python che semplifica la gestione degli spazi di lavoro di sviluppo/formazione e dei server di inferenza supportati dallo storage NetApp scale-out dalle performance elevate. Le funzionalità principali includono:

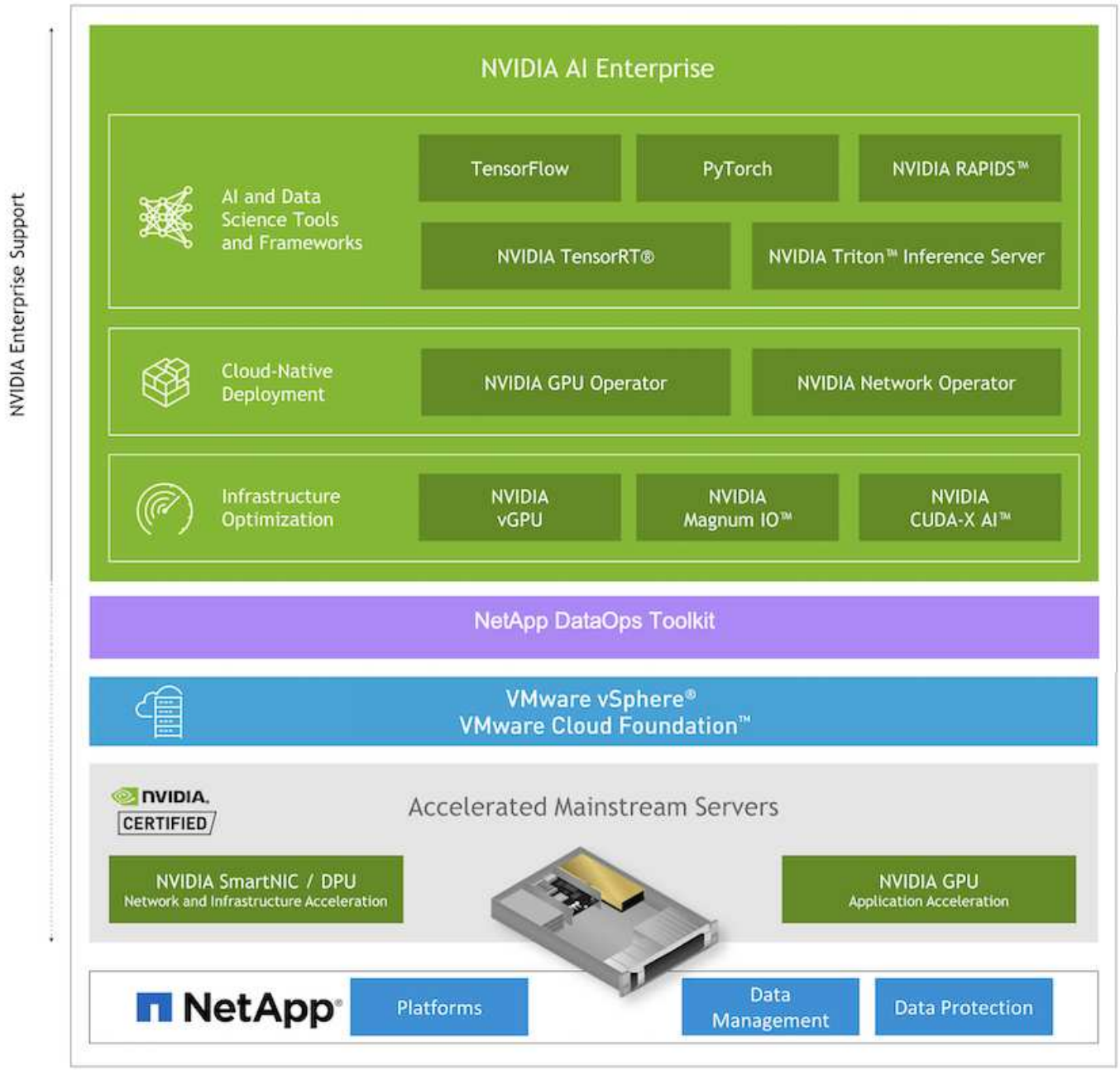
- Provisioning rapido di nuove aree di lavoro JupyterLab ad alta capacità supportate da storage NetApp scale-out dalle performance elevate.
- Provisioning rapido delle nuove istanze di NVIDIA Triton Inference Server supportate dallo storage NetApp di livello Enterprise.
- Clonare quasi instantaneamente le aree di lavoro JupyterLab ad alta capacità per consentire la sperimentazione o l'iterazione rapida.
- Salvataggio quasi istantaneo di snapshot di aree di lavoro JupyterLab ad alta capacità per backup e/o tracciabilità/baselining.
- Provisioning, cloning e snapshot near-instataneamente di volumi di dati ad alta capacità e performance elevate.

Architettura

Questa soluzione si basa su un'architettura collaudata e familiare con sistemi certificati

NetApp, VMware e NVIDIA. Per ulteriori informazioni, consultare la tabella seguente.

Componente	Dettagli
Software ai e Data Analytics	"NVIDIA ai Enterprise per VMware"
Piattaforma di virtualizzazione	"VMware vSphere"
Piattaforma di calcolo	"Sistemi certificati NVIDIA"
Piattaforma per la gestione dei dati	"NetApp ONTAP"



Setup iniziale

In questa sezione vengono descritte le attività di configurazione iniziali che devono

essere eseguite per utilizzare NVIDIA ai Enterprise con NetApp e VMware.

Prerequisiti

Prima di eseguire i passaggi descritti in questa sezione, si presuppone che siano già state implementate VMware vSphere e NetApp ONTAP. Fare riferimento a. "Matrice di supporto dei prodotti NVIDIA ai Enterprise" Per ulteriori informazioni sulle versioni di vSphere supportate. Fare riferimento a. "Documentazione sulle soluzioni NetApp e VMware" Per informazioni dettagliate sull'implementazione di VMware vSphere con NetApp ONTAP.

Installare il software host NVIDIA ai Enterprise

Per installare il software host NVIDIA ai EnEnterprise, seguire le istruzioni riportate nelle sezioni 1-4 del "NVIDIA ai Enterprise Quick Start Guide".

Utilizzare il software NVIDIA NGC

Questa sezione descrive le attività da eseguire per utilizzare il software NVIDIA NGC Enterprise in un ambiente NVIDIA ai Enterprise.

Setup (Configurazione)

In questa sezione vengono descritte le operazioni di configurazione iniziali da eseguire per utilizzare il software NVIDIA NGC Enterprise in un ambiente NVIDIA ai Enterprise.

Prerequisiti

Prima di eseguire i passaggi descritti in questa sezione, si presuppone che il software host NVIDIA ai EnEnterprise sia già stato implementato seguendo le istruzioni riportate nella "Setup iniziale" pagina.

Creare una VM ospite Ubuntu con vGPU

Innanzitutto, è necessario creare una macchina virtuale guest Ubuntu 20.04 con vGPU. Per creare una macchina virtuale guest Ubuntu 20.04 con vGPU, seguire le istruzioni riportate nella "Guida all'implementazione di NVIDIA ai Enterprise".

Scaricare e installare il software NVIDIA Guest

Quindi, è necessario installare il software guest NVIDIA richiesto nella macchina virtuale guest creata al passaggio precedente. Per scaricare e installare il software guest NVIDIA richiesto nella macchina virtuale guest, seguire le istruzioni riportate nelle sezioni 5.1-5.4 della "NVIDIA ai Enterprise Quick Start Guide".

> Quando si eseguono le attività di verifica descritte nella sezione 5.4, potrebbe essere necessario utilizzare un tag di versione dell'immagine contenitore CUDA diverso poiché l'immagine contenitore CUDA è stata aggiornata dopo la scrittura della guida. Nella nostra convalida, abbiamo utilizzato "nvidia/cuda:11.0.3-base-ubuntu20.04".

Scarica i container ai/Analytics Framework

Quindi, devi scaricare le immagini container di ai o del framework di analisi necessarie da NVIDIA NGC in modo che siano disponibili all'interno della tua macchina virtuale guest. Per scaricare i container del framework

all’interno della macchina virtuale guest, seguire le istruzioni riportate nella "Guida all’implementazione di NVIDIA ai Enterprise".

Installare e configurare il NetApp DataOps Toolkit

Quindi, è necessario installare il NetApp DataOps Toolkit per ambienti tradizionali all’interno della macchina virtuale guest. Il toolkit NetApp DataOps può essere utilizzato per gestire volumi di dati scale-out sul sistema ONTAP direttamente dal terminale all’interno della macchina virtuale guest. Per installare il NetApp DataOps Toolkit nella macchina virtuale guest, eseguire le seguenti operazioni.

1. Installare il pip.

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

2. Disconnettersi dal terminale della macchina virtuale guest e quindi effettuare nuovamente l’accesso.
3. Configurare il NetApp DataOps Toolkit. Per completare questo passaggio, sono necessari i dettagli di accesso API per il sistema ONTAP. Potrebbe essere necessario ottenerli dall’amministratore dello storage.

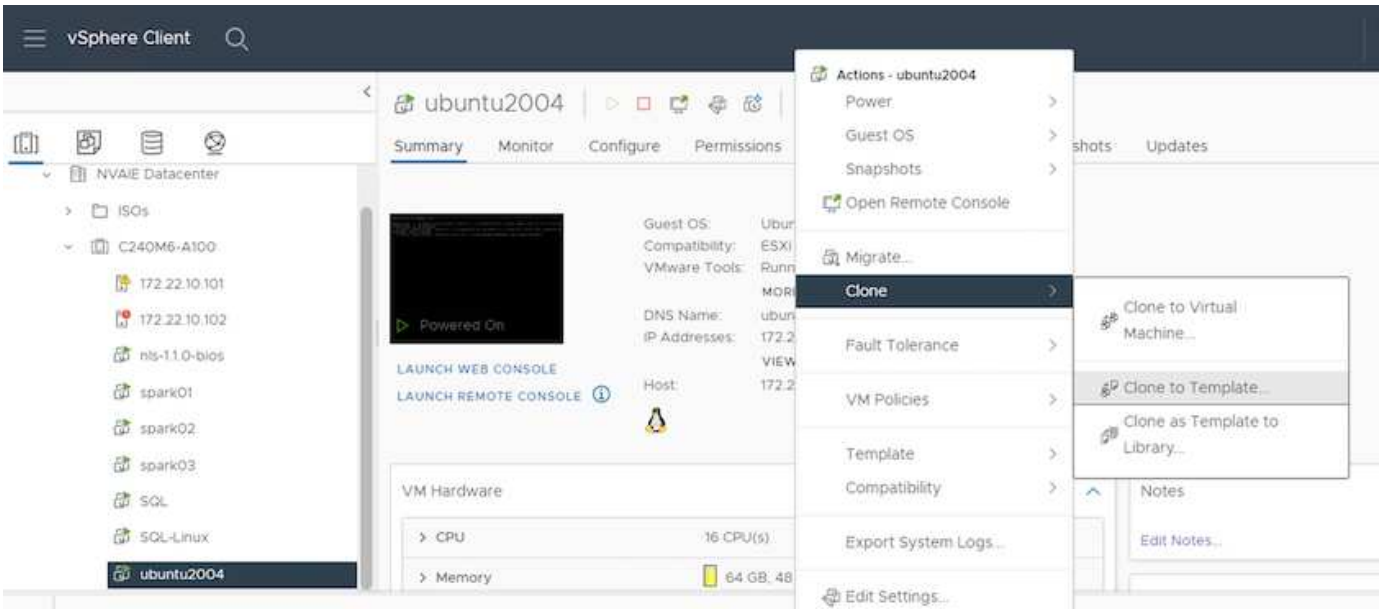
```
$ netapp_dataops_cli.py config

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'.
```

Creare un modello di macchina virtuale guest

Infine, è necessario creare un modello di macchina virtuale basato sulla macchina virtuale guest. Sarà possibile utilizzare questo modello per creare rapidamente macchine virtuali guest per l'utilizzo del software NVIDIA NGC.

Per creare un modello di macchina virtuale in base alla macchina virtuale guest, accedere a VMware vSphere, fare clic sul nome della macchina virtuale guest, scegliere "Clone", "Clone to Template…", quindi seguire la procedura guidata.



Esempio di caso d’utilizzo - lavoro di training TensorFlow

Questa sezione descrive le attività da eseguire per eseguire un lavoro di training TensorFlow in un ambiente NVIDIA ai Enterprise.

Prerequisiti

Prima di eseguire i passaggi descritti in questa sezione, si presuppone che sia già stato creato un modello di macchina virtuale guest seguendo le istruzioni riportate nella "Setup (Configurazione)" pagina.

Creare una macchina virtuale guest dal modello

Innanzitutto, è necessario creare una nuova macchina virtuale guest dal modello creato nella sezione precedente. Per creare una nuova macchina virtuale guest dal modello, accedere a VMware vSphere, fare clic sul nome del modello, scegliere 'Nuova macchina virtuale da questo modello…', quindi seguire la procedura guidata.

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

Creare e montare un volume di dati

Quindi, è necessario creare un nuovo volume di dati su cui memorizzare il set di dati di training. È possibile creare rapidamente un nuovo volume di dati utilizzando il NetApp DataOps Toolkit. Il comando di esempio che segue mostra la creazione di un volume denominato 'imagenet' con una capacità di 2 TB.

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

Prima di poter popolare i dati nel volume di dati, è necessario montarli all'interno della macchina virtuale guest. È possibile montare rapidamente un volume di dati utilizzando il NetApp DataOps Toolkit. Il comando di esempio che segue mostra il mouse del volume creato nel passaggio precedente.

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

Popolare il volume di dati

Una volta eseguito il provisioning e il montaggio del nuovo volume, è possibile recuperare il set di dati di training dalla posizione di origine e posizionarlo sul nuovo volume. In genere, ciò comporta il prelievo dei dati da un data Lake S3 o Hadoop e talvolta comporta l'aiuto di un data engineer.

Eseguire il lavoro di training TensorFlow

Ora, sei pronto per eseguire il tuo lavoro di training TensorFlow. Per eseguire il tuo lavoro di training TensorFlow, esegui le seguenti attività.

1. Estrarre l'immagine del container NVIDIA NGC Enterprise TensorFlow.

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

2. Avviare un'istanza di NVIDIA NGC Enterprise TensorFlow Container. Utilizzare l'opzione '-v' per collegare il volume di dati al container.

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

3. Esegui il tuo programma di training TensorFlow all'interno del container. Il comando di esempio che segue mostra l'esecuzione di un programma di training ResNet-50 di esempio incluso nell'immagine container.

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

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i

seguenti documenti e/o siti Web:

- Software per la gestione dei dati NetApp ONTAP: Libreria di informazioni ONTAP

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

- NetApp DataOps Toolkit

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

- NVIDIA ai Enterprise con VMware

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

Ringraziamenti

- Bobby Oommen, Sr Manager, NetApp
- Ramesh Isaac, System Administrator, NetApp
- Roney Daniel, Technical Marketing Engineer, NetApp