NVIDIA AI Enterprise con NetApp e VMware

NetApp artificial intelligence solutions

NetApp
December 04, 2025

Sommario

NVIDIA AI Enterprise con NetApp e VMware . . . 1
- NVIDIA AI Enterprise con NetApp e VMware . . . 1
- Panoramica della tecnologia . . . 1
  - NVIDIA AI Enterprise . . . 2
  - Cloud GPU NVIDIA (NGC) . . . 2
  - VMware vSphere . . . 2
  - NetApp ONTAP . . . 2
  - Kit di strumenti NetApp DataOps . . . 3
- Architettura . . . 4
- Configurazione iniziale . . . 5
  - Prerequisiti . . . 5
  - Installa il software NVIDIA AI Enterprise Host . . . 5
- Utilizzare il software NVIDIA NGC . . . 5
  - Impostare . . . 5
  - Esempio di caso d’uso: lavoro di formazione TensorFlow . . . 8
- Dove trovare ulteriori informazioni . . . 10
  - Ringraziamenti . . . 11

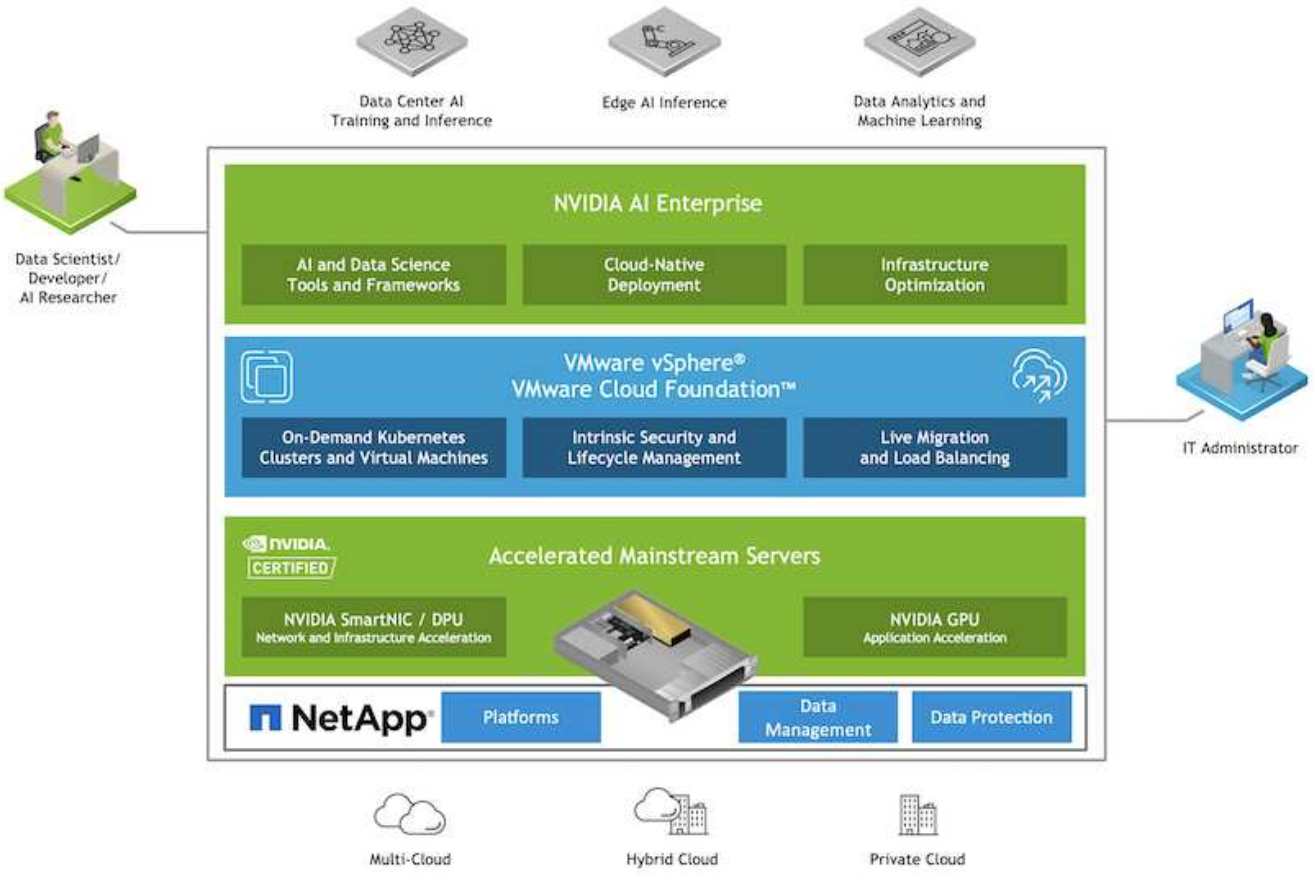
NVIDIA AI Enterprise con NetApp e VMware

NVIDIA AI Enterprise con NetApp e VMware

Mike Oglesby, NetApp

Per gli architetti e gli amministratori IT, gli strumenti di intelligenza artificiale possono risultare complessi e poco familiari. Inoltre, molte piattaforme di intelligenza artificiale non sono pronte per l'uso aziendale. NVIDIA AI Enterprise, basato su NetApp e VMware, è stato creato per offrire un'architettura di intelligenza artificiale semplificata e di livello aziendale.

NVIDIA AI Enterprise è una suite end-to-end e cloud-native di software di analisi dei dati e intelligenza artificiale, ottimizzata, certificata e supportata da NVIDIA per l'esecuzione su VMware vSphere con sistemi certificati NVIDIA. Questo software semplifica e velocizza l'implementazione, la gestione e il ridimensionamento dei carichi di lavoro di intelligenza artificiale nel moderno ambiente cloud ibrido. NVIDIA AI Enterprise, basato su NetApp e VMware, offre una gestione dei dati e dei carichi di lavoro di intelligenza artificiale di livello aziendale in un pacchetto semplificato e familiare.



Panoramica della tecnologia

Questa sezione fornisce una panoramica della tecnologia per NVIDIA AI Enterprise con NetApp e VMware.

NVIDIA AI Enterprise

NVIDIA AI Enterprise è una suite end-to-end e cloud-native di software di analisi dei dati e intelligenza artificiale, ottimizzata, certificata e supportata da NVIDIA per l'esecuzione su VMware vSphere con sistemi certificati NVIDIA. Questo software semplifica e velocizza l'implementazione, la gestione e il ridimensionamento dei carichi di lavoro di intelligenza artificiale nel moderno ambiente cloud ibrido.

Cloud GPU NVIDIA (NGC)

NVIDIA NGC ospita un catalogo di software ottimizzati per GPU che consente ai professionisti dell'intelligenza artificiale di sviluppare le proprie soluzioni di intelligenza artificiale. Fornisce inoltre accesso a vari servizi di intelligenza artificiale, tra cui NVIDIA Base Command per l'addestramento dei modelli, NVIDIA Fleet Command per distribuire e monitorare i modelli e NGC Private Registry per l'accesso e la gestione sicuri del software di intelligenza artificiale proprietario. Inoltre, i clienti NVIDIA AI Enterprise possono richiedere supporto tramite il portale NGC.

VMware vSphere

VMware vSphere è la piattaforma di virtualizzazione di VMware che trasforma i data center in infrastrutture di elaborazione aggregate che includono risorse di CPU, storage e 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 tramite il quale gli amministratori gestiscono più host connessi in rete e mettono in pool le risorse host.

NetApp ONTAP

ONTAP 9, l'ultima generazione di software di gestione dello storage di NetApp, consente alle aziende di modernizzare l'infrastruttura e passare a un data center pronto per il cloud. Sfruttando le funzionalità di gestione dei dati leader del settore, ONTAP consente la gestione e la protezione dei dati con un unico set di strumenti, indipendentemente da dove risiedano. È inoltre possibile 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 infrastrutturali di nuova generazione nelle architetture cloud ibride.

Semplificare la gestione dei dati

La gestione dei dati è fondamentale per le operazioni IT aziendali e per gli scienziati dei dati, in modo che vengano utilizzate risorse appropriate per le applicazioni di intelligenza artificiale e per la formazione di set di dati di intelligenza artificiale/apprendimento automatico. Le seguenti informazioni aggiuntive sulle tecnologie NetApp esulano dall'ambito di questa convalida, ma potrebbero essere rilevanti a seconda della distribuzione.

Il software di gestione dati ONTAP include le seguenti funzionalità per semplificare e snellire le operazioni e ridurre i costi operativi totali:

- Compattazione dei dati in linea e deduplicazione estesa. La compattazione dei dati riduce lo spazio sprecato all'interno dei blocchi di archiviazione, mentre la deduplicazione aumenta significativamente la capacità effettiva. Ciò vale sia per i dati archiviati localmente sia per i dati archiviati a livelli nel cloud.
- Qualità del servizio minima, massima e adattiva (AQoS). I controlli granulari della qualità del servizio (QoS) aiutano a mantenere i livelli di prestazioni per le applicazioni critiche in ambienti altamente condivisi.
- NetApp FabricPool. Fornisce la suddivisione automatica dei dati inattivi in opzioni di archiviazione cloud

pubbliche e private, tra cui Amazon Web Services (AWS), Azure e la soluzione di archiviazione NetApp StorageGRID . Per ulteriori informazioni su FabricPool, vedere "TR-4598: Buone pratiche FabricPool" .

Accelerare e proteggere i dati

ONTAP garantisce livelli superiori di prestazioni e protezione dei dati ed estende queste capacità nei seguenti modi:

- Prestazioni e latenza più bassa. ONTAP offre la massima capacità di trasmissione possibile con la minima latenza possibile.
- Protezione dei dati. ONTAP offre funzionalità integrate di protezione dei dati con gestione comune su tutte le piattaforme.
- Crittografia del volume NetApp (NVE). ONTAP offre la crittografia nativa a livello di volume con supporto sia per la gestione delle chiavi integrate che per quella esterna.
- Multitenancy e autenticazione multifattore. ONTAP consente la condivisione delle risorse infrastrutturali con i massimi livelli di sicurezza.

Infrastruttura a prova di futuro

ONTAP aiuta a soddisfare le esigenze aziendali più esigenti e in continua evoluzione grazie alle seguenti funzionalità:

- Scalabilità fluida e operazioni senza interruzioni. ONTAP supporta l'aggiunta non distruttiva di capacità ai controller esistenti e ai cluster scalabili. I clienti possono effettuare l'aggiornamento alle tecnologie più recenti, come NVMe e FC da 32 Gb, senza costose migrazioni di dati o interruzioni.
- Connessione cloud. ONTAP è il software di gestione dello storage più connesso al cloud, con opzioni per lo storage definito dal software (ONTAP Select) e istanze cloud-native (Google Cloud NetApp Volumes) in tutti i cloud pubblici.
- Integrazione con applicazioni emergenti. ONTAP offre servizi dati di livello aziendale per piattaforme e applicazioni di nuova generazione, come veicoli autonomi, città intelligenti e Industria 4.0, utilizzando la stessa infrastruttura che supporta le app aziendali esistenti.

Kit di strumenti NetApp DataOps

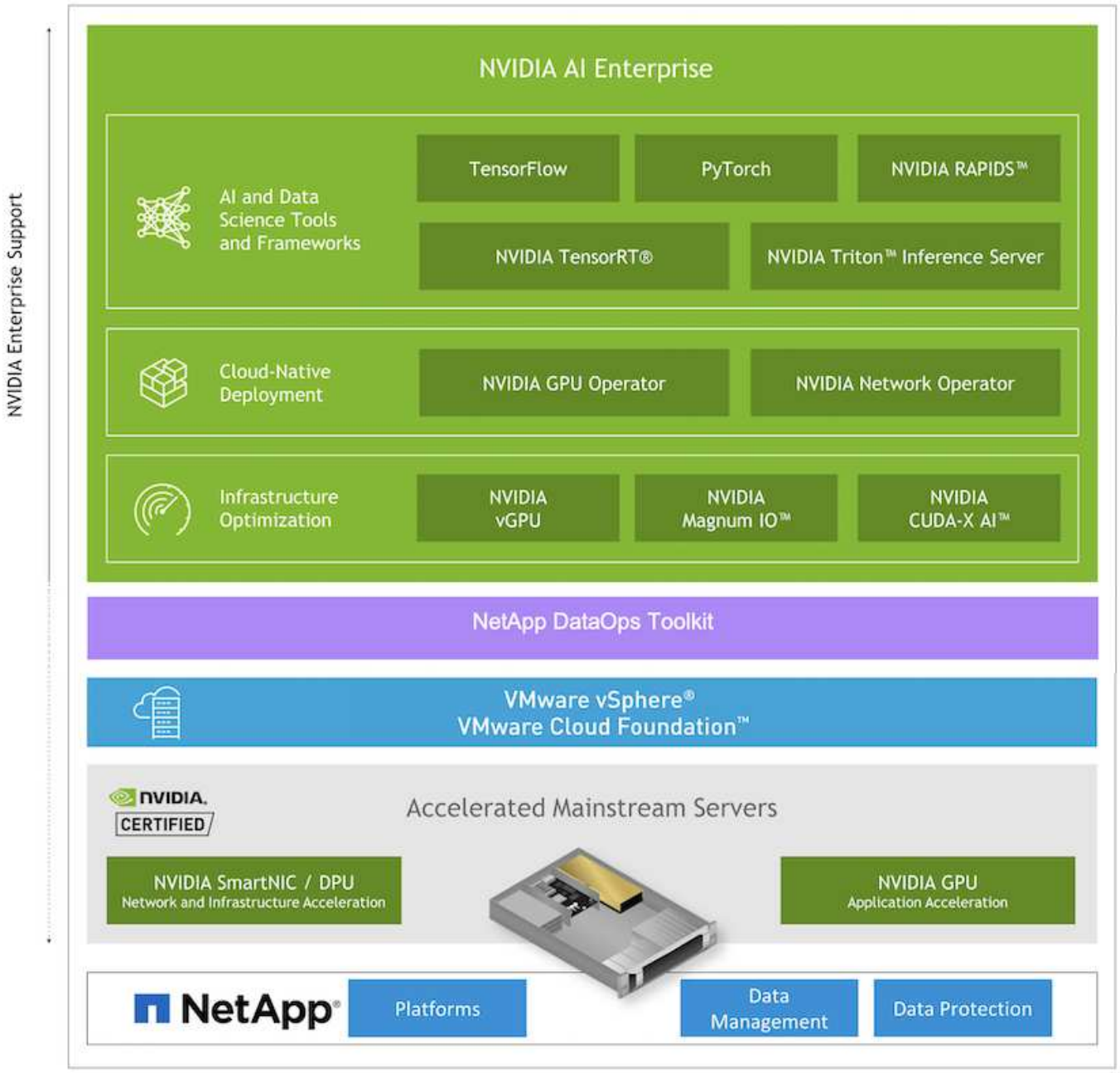
NetApp DataOps Toolkit è uno strumento basato su Python che semplifica la gestione degli spazi di lavoro di sviluppo/formazione e dei server di inferenza supportati da storage NetApp ad alte prestazioni e scalabile. Le principali funzionalità includono:

- Fornisci rapidamente nuovi spazi di lavoro JupyterLab ad alta capacità supportati da storage NetApp scalabile e ad alte prestazioni.
- Fornisci rapidamente nuove istanze di NVIDIA Triton Inference Server supportate da storage NetApp di classe enterprise.
- Clonare quasi istantaneamente gli spazi di lavoro JupyterLab ad alta capacità per consentire la sperimentazione o l'iterazione rapida.
- Salvataggio quasi istantaneo di snapshot di spazi di lavoro JupyterLab ad alta capacità per backup e/o tracciabilità/baselining.
- Esegui il provisioning, la clonazione e l'istantanea di volumi di dati ad alta capacità e ad alte prestazioni in modo quasi istantaneo.

Architettura

Questa soluzione si basa su un'architettura collaudata e familiare che comprende sistemi certificati NetApp, VMware e NVIDIA. Per i dettagli, vedere la tabella seguente.

Componente	Dettagli
Software di intelligenza artificiale e analisi dei dati	"NVIDIA AI Enterprise per VMware"
Piattaforma di virtualizzazione	"VMware vSphere"
Piattaforma di calcolo	"Sistemi certificati NVIDIA"
Piattaforma di gestione dei dati	"NetApp ONTAP"



Configurazione iniziale

Questa sezione descrive le attività di configurazione iniziale che devono essere eseguite per utilizzare NVIDIA AI Enterprise con NetApp e VMware.

Prerequisiti

Prima di eseguire i passaggi descritti in questa sezione, presumiamo che tu abbia già distribuito VMware vSphere e NetApp ONTAP. Fare riferimento al"Matrice di supporto dei prodotti NVIDIA AI Enterprise" per i dettagli sulle versioni vSphere supportate. Fare riferimento al"Documentazione delle soluzioni NetApp e VMware" per i dettagli sulla distribuzione di VMware vSphere con NetApp ONTAP.

Installa il software NVIDIA AI Enterprise Host

Per installare il software host NVIDIA AI Entrprise, seguire le istruzioni descritte nelle sezioni 1-4 del"Guida rapida all'avvio di NVIDIA AI Enterprise" .

Utilizzare il software NVIDIA NGC

Impostare

In questa sezione vengono descritte le attività di configurazione iniziale che devono essere eseguite per utilizzare il software aziendale NVIDIA NGC in un ambiente NVIDIA AI Enterprise.

Prerequisiti

Prima di eseguire i passaggi descritti in questa sezione, si presume che sia già stato distribuito il software host NVIDIA AI Entrprise seguendo le istruzioni descritte nel"Configurazione iniziale" pagina.

Creare una VM guest Ubuntu con vGPU

Per prima cosa, devi creare una VM guest Ubuntu 20.04 con vGPU. Per creare una VM guest Ubuntu 20.04 con vGPU, seguire le istruzioni descritte in"Guida all'implementazione aziendale NVIDIA AI" .

Scarica e installa il software guest NVIDIA

Successivamente, è necessario installare il software guest NVIDIA richiesto nella VM guest creata nel passaggio precedente. Per scaricare e installare il software guest NVIDIA richiesto all'interno della VM guest, seguire le istruzioni descritte nelle sezioni 5.1-5.4 nella"Guida rapida all'avvio di NVIDIA AI Enterprise" .

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

Scarica i contenitori del framework AI/Analytics

Successivamente, è necessario scaricare le immagini dei container del framework di analisi o di intelligenza artificiale necessarie da NVIDIA NGC, in modo che siano disponibili nella VM guest. Per scaricare i contenitori del framework all'interno della VM guest, seguire le istruzioni descritte in"Guida all'implementazione aziendale

NVIDIA AI" .

Installa e configura NetApp DataOps Toolkit

Successivamente, è necessario installare NetApp DataOps Toolkit for Traditional Environemnts nella VM guest. NetApp DataOps Toolkit può essere utilizzato per gestire volumi di dati scalabili sul sistema ONTAP direttamente dal terminale all'interno della VM guest. Per installare NetApp DataOps Toolkit nella VM guest, eseguire le seguenti attività.

1. Installa pip.

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

2. Disconnettersi dal terminale della VM guest e quindi effettuare nuovamente l'accesso.
3. Configurare NetApp DataOps Toolkit. Per completare questo passaggio, avrai bisogno dei dettagli di accesso API per il tuo sistema ONTAP . Potrebbe essere necessario richiederli all'amministratore del tuo archivio.

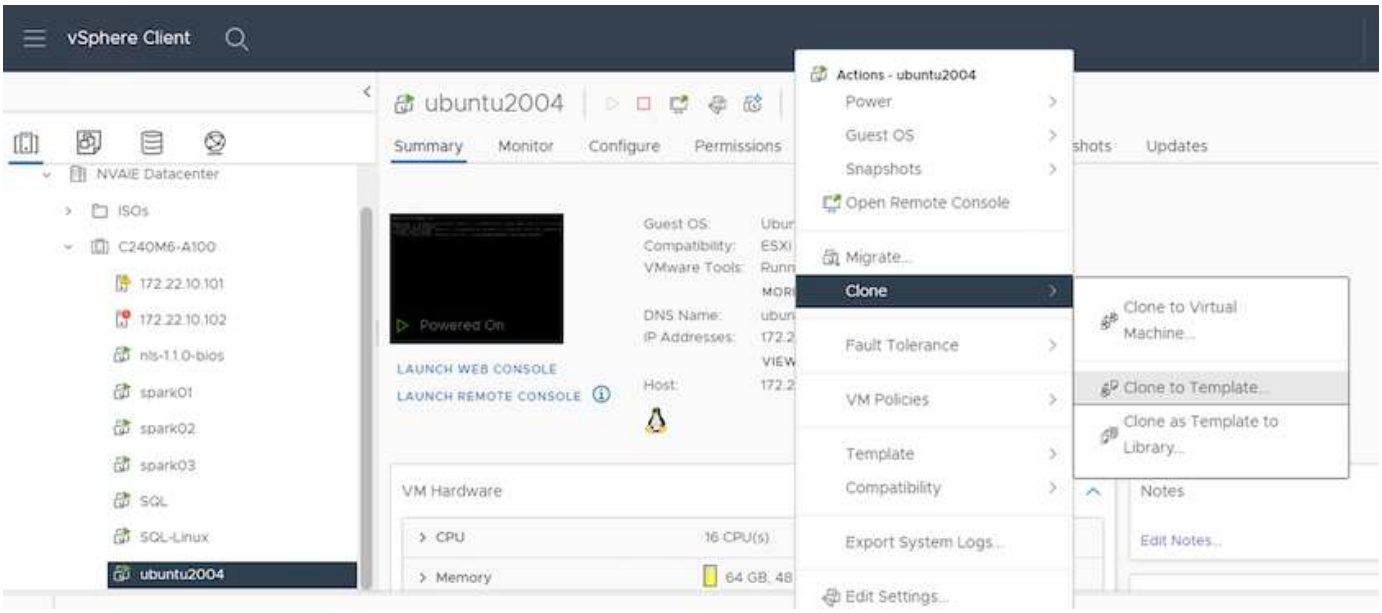
```
$ netapp_dataops_cli.py config

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 VM guest

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

Per creare un modello di VM basato sulla VM guest, accedi a VMware vSphere, fai clic con il pulsante destro del mouse sul nome della VM guest, seleziona "Clona", quindi "Clona su modello…" e segui la procedura guidata.



Esempio di caso d'uso: lavoro di formazione TensorFlow

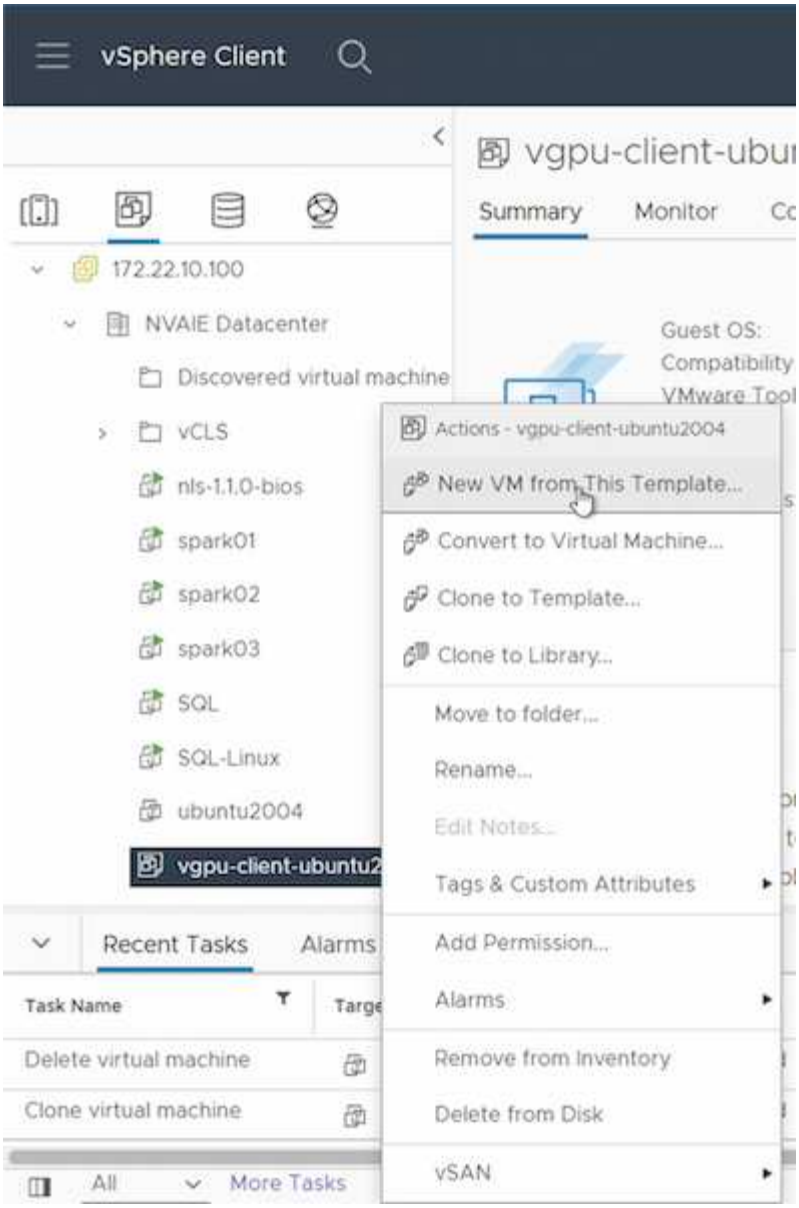
Questa sezione descrive le attività che devono essere eseguite per eseguire un processo di training TensorFlow in un ambiente NVIDIA AI Enterprise.

Prerequisiti

Prima di eseguire i passaggi descritti in questa sezione, si presume che sia già stato creato un modello di VM guest seguendo le istruzioni descritte in"Impostare" pagina.

Crea una VM guest dal modello

Per prima cosa, devi creare una nuova VM guest dal modello creato nella sezione precedente. Per creare una nuova VM guest dal tuo modello, accedi a VMware vSphere, fai clic con il pulsante destro del mouse sul nome del modello, scegli "Nuova VM da questo modello…" e segui la procedura guidata.



Crea e monta il volume dati

Successivamente, è necessario creare un nuovo volume di dati su cui archiviare il set di dati di addestramento. È possibile creare rapidamente un nuovo volume di dati utilizzando 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 il volume dati con i dati, è necessario montarlo nella VM guest. È possibile montare rapidamente un volume di dati utilizzando NetApp DataOps Toolkit. Il comando di esempio che segue mostra il montaggio del volume creato nel passaggio precedente.

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

Popola il volume dei dati

Dopo aver eseguito il provisioning e il montaggio del nuovo volume, il set di dati di training può essere recuperato dalla posizione di origine e posizionato sul nuovo volume. In genere, ciò comporta l'estrazione dei dati da un data lake S3 o Hadoop e talvolta richiede l'aiuto di un data engineer.

Esegui il job di training di TensorFlow

Ora sei pronto per eseguire il tuo job di training TensorFlow. Per eseguire il processo di addestramento TensorFlow, eseguire le seguenti attività.

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

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

2. Avvia un'istanza del contenitore NVIDIA NGC Enterprise TensorFlow. Utilizzare l'opzione '-v' per allegare il volume di dati al contenitore.

```
$ 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 programma di addestramento TensorFlow all'interno del contenitore. Il comando di esempio seguente mostra l'esecuzione di un programma di formazione ResNet-50 di esempio incluso nell'immagine del contenitore.

```
$ 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 saperne di più sulle informazioni descritte nel presente documento, fare riferimento ai seguenti documenti e/o siti web:

- Software di gestione dati NetApp ONTAP — Libreria di informazioni ONTAP

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

- Kit di strumenti NetApp DataOps

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

- NVIDIA AI Enterprise con VMware

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

Ringraziamenti

- Bobby Oommen, Direttore senior, NetApp
- Ramesh Isaac, amministratore di sistema, NetApp
- Roney Daniel, ingegnere tecnico di marketing, NetApp