Casi di utilizzo

NetApp Solutions

NetApp
September 10, 2024

Sommario

Casi di utilizzo

Ai responsabile e deduzione riservata - NetApp ai con Protopia Image Transformation

TR-4928: Ai responsabile e deduzione riservata - NetApp ai con Protopia Image e Data Transformation

Sathish Thyagarajan, Michael Oglesby, NetApp Byung Hoon Ahn, Jennifer Cwagenberg, Protopia

Le interpretazioni visive sono diventate parte integrante della comunicazione con l'emergere dell'acquisizione e dell'elaborazione delle immagini. L'intelligenza artificiale (ai) nell'elaborazione di immagini digitali offre nuove opportunità di business, come nel campo medico per l'identificazione di tumori e altre malattie, nell'analisi visiva geospaziale per lo studio dei rischi ambientali, nel riconoscimento dei modelli, nell'elaborazione video per la lotta alla criminalità e così via. Tuttavia, questa opportunità comporta anche responsabilità straordinarie.

Più le organizzazioni prendono decisioni in mano all'ai, più accettano rischi legati alla privacy e alla sicurezza dei dati e a questioni legali, etiche e normative. L'intelligenza artificiale responsabile consente a aziende e organizzazioni governative di creare fiducia e governance che sono fondamentali per l'intelligenza artificiale su larga scala nelle grandi imprese. Questo documento descrive una soluzione di inferenza ai convalidata da NetApp in tre scenari diversi utilizzando le tecnologie di gestione dei dati NetApp con il software di offuscamento dei dati Protopia per privatizzare i dati sensibili e ridurre rischi e preoccupazioni etiche.

Milioni di immagini vengono generate ogni giorno con diversi dispositivi digitali sia da consumatori che da aziende. La conseguente massiccia esplosione dei dati e del carico di lavoro di calcolo fa sì che le aziende si rivolgono alle piattaforme di cloud computing per scalabilità ed efficienza. Nel frattempo, i problemi di privacy relativi alle informazioni sensibili contenute nei dati delle immagini sorgono con il trasferimento a un cloud pubblico. La mancanza di garanzie di sicurezza e privacy diventa la principale barriera all'implementazione dei sistemi ai di elaborazione delle immagini.

Inoltre, è disponibile la "diritto di cancellazione" Dal GDPR, il diritto di un individuo di richiedere che un'organizzazione cancella tutti i propri dati personali. C'è anche il "Legge sulla privacy", che stabilisce un codice di pratiche di informazione eque. Le immagini digitali come le fotografie possono costituire dati personali ai sensi del GDPR, che regola le modalità di raccolta, elaborazione e cancellazione dei dati. In caso contrario, la mancata conformità al GDPR potrebbe comportare multe elevate per la violazione delle conformità che possono danneggiare seriamente le organizzazioni. I principi di privacy sono tra le fondamenta dell'implementazione dell'ai responsabile che garantisce l'equità nelle previsioni del modello di apprendimento automatico (ML) e di apprendimento approfondito (DL) e riduce i rischi associati alla violazione della privacy o della conformità alle normative.

Questo documento descrive una soluzione di progettazione validata in tre scenari diversi, con e senza offuscamento delle immagini, per preservare la privacy e implementare una soluzione di ai responsabile:

- **Scenario 1.** deduzione on-demand nel notebook Jupyter.
- **Scenario 2.** deduzione batch su Kubernetes.
- **Scenario 3.** Server di inferenza NVIDIA Triton.

Per questa soluzione, utilizziamo Face Detection Data Set and Benchmark (FDDB), un set di dati delle regioni

dei volti progettato per studiare il problema del rilevamento dei volti senza vincoli, in combinazione con il framework di apprendimento automatico PyTorch per l'implementazione di FaceBoxes. Questo set di dati contiene le annotazioni per 5171 volti in un set di 2845 immagini di varie risoluzioni. Inoltre, questo report tecnico presenta alcune delle aree di soluzione e i casi di utilizzo rilevanti raccolti dai clienti NetApp e dai tecnici sul campo nelle situazioni in cui questa soluzione è applicabile.

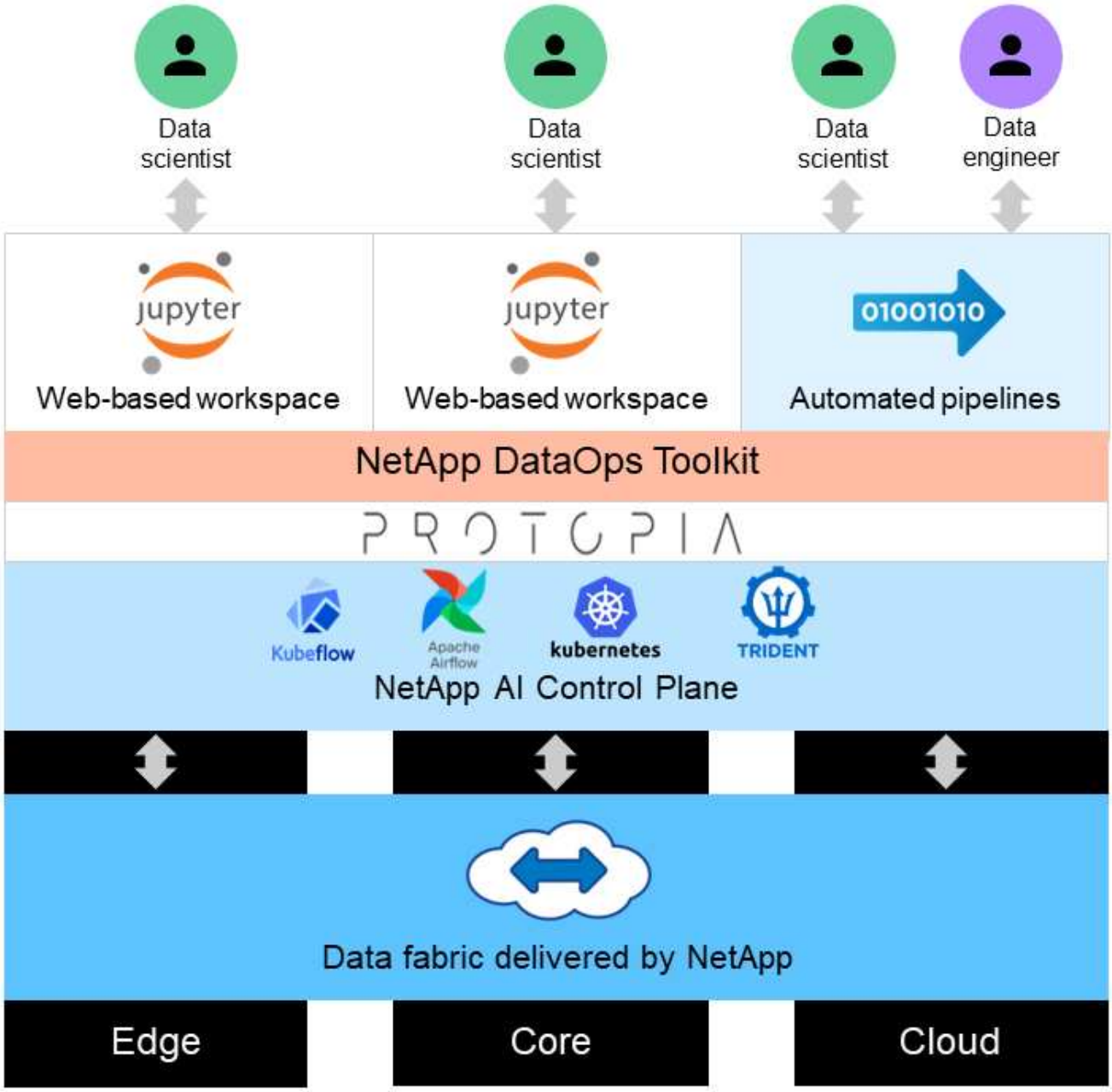
Pubblico di riferimento

Questo report tecnico è destinato ai seguenti destinatari:

- Business leader e architetti aziendali che desiderano progettare e implementare un'ai responsabile e affrontare i problemi di protezione dei dati e privacy relativi all'elaborazione delle immagini facciali negli spazi pubblici.
- Data scientist, data engineer, ricercatori ai/ machine learning (ML) e sviluppatori di sistemi ai/ML che mirano a proteggere e preservare la privacy.
- Architetti aziendali che progettano soluzioni di offuscamento dei dati per modelli e applicazioni ai/ML conformi agli standard normativi come GDPR, CCPA o il Privacy Act del Department of Defense (DoD) e organizzazioni governative.
- Data scientist e ingegneri ai alla ricerca di modi efficienti per implementare modelli di deduzione ai/ML/DL e deep learning (DL) in grado di proteggere le informazioni sensibili.
- Edge Device Manager e amministratori degli edge server responsabili dell'implementazione e della gestione dei modelli di inferenza edge.

Architettura della soluzione

Questa soluzione è progettata per gestire carichi di lavoro ai di deduzione in batch e in tempo reale su grandi set di dati utilizzando la potenza di elaborazione delle GPU insieme alle CPU tradizionali. Questa convalida dimostra l'inferenza di conservazione della privacy per ML e la gestione ottimale dei dati richiesta per le organizzazioni che cercano implementazioni ai responsabili. Questa soluzione offre un'architettura adatta per una piattaforma Kubernetes a nodo singolo o multiplo per il cloud computing e l'edge interconnesso con NetApp ONTAP ai al core on-premise, il toolkit NetApp DataOps e il software di offuscamento Protopia utilizzando le interfacce Jupyter Lab e CLI. La figura seguente mostra una panoramica dell'architettura logica del data fabric basato su NetApp con DataOps Toolkit e Protopia.



Il software di offuscamento Protopia viene eseguito perfettamente sul NetApp DataOps Toolkit e trasforma i dati prima di lasciare il server di storage.

Aree di soluzione

L'elaborazione delle immagini digitali offre numerosi vantaggi, consentendo a molte organizzazioni di sfruttare al meglio i dati associati alle rappresentazioni visive. Questa soluzione NetApp e Protopia offre un design di inferenza ai unico per proteggere e privatizzare i dati ai/ML nel ciclo di vita ML/DL. Consente ai clienti di mantenere la proprietà dei dati sensibili, utilizzare modelli di implementazione del cloud pubblico o ibrido per scalabilità ed efficienza, alleviando i problemi relativi alla privacy e implementando l'inferenza ai ai ai ai edge.

Intelligence ambientale

Esistono diversi modi in cui i settori possono sfruttare le analisi geospaziali nelle aree dei rischi ambientali. I governi e il dipartimento delle opere pubbliche possono trarre utili informazioni sulla salute pubblica e sulle condizioni meteorologiche per consigliare meglio il pubblico durante una pandemia o un disastro naturale come gli incendi. Ad esempio, è possibile identificare un paziente COVID-positivo in spazi pubblici, come aeroporti o ospedali, senza compromettere la privacy della persona interessata e avvisare le rispettive autorità e il pubblico nelle vicinanze per le misure di sicurezza necessarie.

Dispositivi indossabili edge

Nel settore militare e nei campi di battaglia, è possibile utilizzare l'inferenza ai all'edge come dispositivi indossabili per monitorare la salute dei soldati, monitorare il comportamento dei conducenti e avvisare le autorità sulla sicurezza e i rischi associati all'avvicinamento ai veicoli militari, preservando e proteggendo la privacy dei soldati. Il futuro dei militari sta diventando high-tech con Internet of Battlefield Things (IoBT) e Internet of Military Things (IoMT) per attrezzature da combattimento indossabili che aiutano i soldati a identificare i nemici e a migliorare le performance in battaglia utilizzando il calcolo rapido edge. La protezione e la conservazione dei dati visivi raccolti da dispositivi edge come droni e dispositivi indossabili è fondamentale per tenere a bada hacker e nemici.

Operazioni di evacuazione non combattente

Le operazioni di evacuazione dei non combattenti (NEO) sono condotte dal DOD per aiutare a evacuare cittadini e cittadini degli Stati Uniti, personale civile del DOD e persone designate (nazione ospitante (HN) e cittadini di paesi terzi (TCN)) la cui vita è in pericolo per un adeguato rifugio sicuro. I controlli amministrativi in uso utilizzano processi di screening per l'evacuazione in gran parte manuali. Tuttavia, l'accuratezza, la sicurezza e la velocità dell'identificazione degli evacuati, del monitoraggio degli evacuati e dello screening delle minacce potrebbero potenzialmente essere migliorate utilizzando strumenti ai/ML altamente automatizzati combinati con tecnologie di offuscamento video ai/ML.

Ricerca sanitaria e biomedica

L'elaborazione delle immagini viene utilizzata per diagnosticare patologie per la pianificazione chirurgica da immagini 3D ottenute dalla tomografia computerizzata (TC) o dall'imaging a risonanza magnetica (MRI). Le norme sulla privacy di HIPAA regolano le modalità di raccolta, elaborazione e cancellazione dei dati da parte delle organizzazioni per tutte le informazioni personali e le immagini digitali come le fotografie. Affinché i dati possano qualificarsi come condivisibili in base alle normative HIPAA Safe Harbor, è necessario rimuovere le immagini fotografiche a pieno facciale e le immagini comparabili. Tecniche automatizzate come la deidentificazione o il cranio-gli algoritmi di spelatura utilizzati per oscurare le caratteristiche facciali di un individuo dalle immagini strutturali TC/RM sono diventati una parte essenziale del processo di condivisione dei dati per gli istituti di ricerca biomedica.

Migrazione nel cloud degli analytics ai/ML

I clienti aziendali hanno tradizionalmente addestrato e implementato modelli ai/ML on-premise. Per motivi di efficienza e scalabilità, questi clienti si stanno espandendo per spostare le funzioni ai/ML in implementazioni di cloud pubblico, ibrido o multi-cloud. Tuttavia, sono vincolati da quali dati possono essere esposti ad altre infrastrutture. Le soluzioni NetApp affrontano una gamma completa di minacce alla cybersicurezza richieste per "protezione dei dati" E la valutazione della sicurezza e, se combinata con la trasformazione dei dati di Protopia, riducono al minimo i rischi associati alla migrazione dei carichi di lavoro ai/ML di elaborazione delle immagini nel cloud.

Per ulteriori casi di utilizzo per l'edge computing e l'inferenza ai in altri settori, vedere "TR-4886 ai Inferencing at the Edge" E il blog NetApp ai, "Intelligence e privacy".

Panoramica della tecnologia

Questa sezione fornisce una panoramica dei vari componenti tecnici necessari per completare questa soluzione.

Protopia

Protopia ai offre una soluzione software-only senza invadenze per l'inferenza riservata nel mercato odierno. La soluzione Protopia offre una protezione senza pari per i servizi di inferenza riducendo al minimo l'esposizione delle informazioni sensibili. L'intelligenza artificiale viene alimentata solo nelle informazioni contenute nel record di dati che sono veramente essenziali per eseguire l'attività in corso e niente di più. La maggior parte delle attività di inferenza non utilizza tutte le informazioni presenti in ogni record di dati. Indipendentemente dal fatto che l'ai stia utilizzando immagini, voce, video o persino dati tabulari strutturati, Protopia offre solo ciò di cui ha bisogno il servizio di inferenza. La tecnologia brevettata core utilizza il rumore matematicamente curato per trasformare in modo stocoso i dati e raccogliere le informazioni non necessarie per un determinato servizio ML. Questa soluzione non maschera i dati, ma modifica la rappresentazione dei dati utilizzando un rumore casuale a cura.

La soluzione Protopia formula il problema di modificare la rappresentazione come metodo di massimizzazione delle perturbazioni basato su gradiente che conserva ancora le informazioni pertinenti nello spazio delle funzionalità di input rispetto alla funzionalità del modello. Questo processo di rilevamento viene eseguito come un passo di messa a punto al termine della formazione sul modello ML. Dopo che il pass genera automaticamente un set di distribuzioni di probabilità, una trasformazione dei dati a basso overhead applica campioni di rumore da queste distribuzioni ai dati, offuscandoli prima di passarli al modello per l'inferenza.

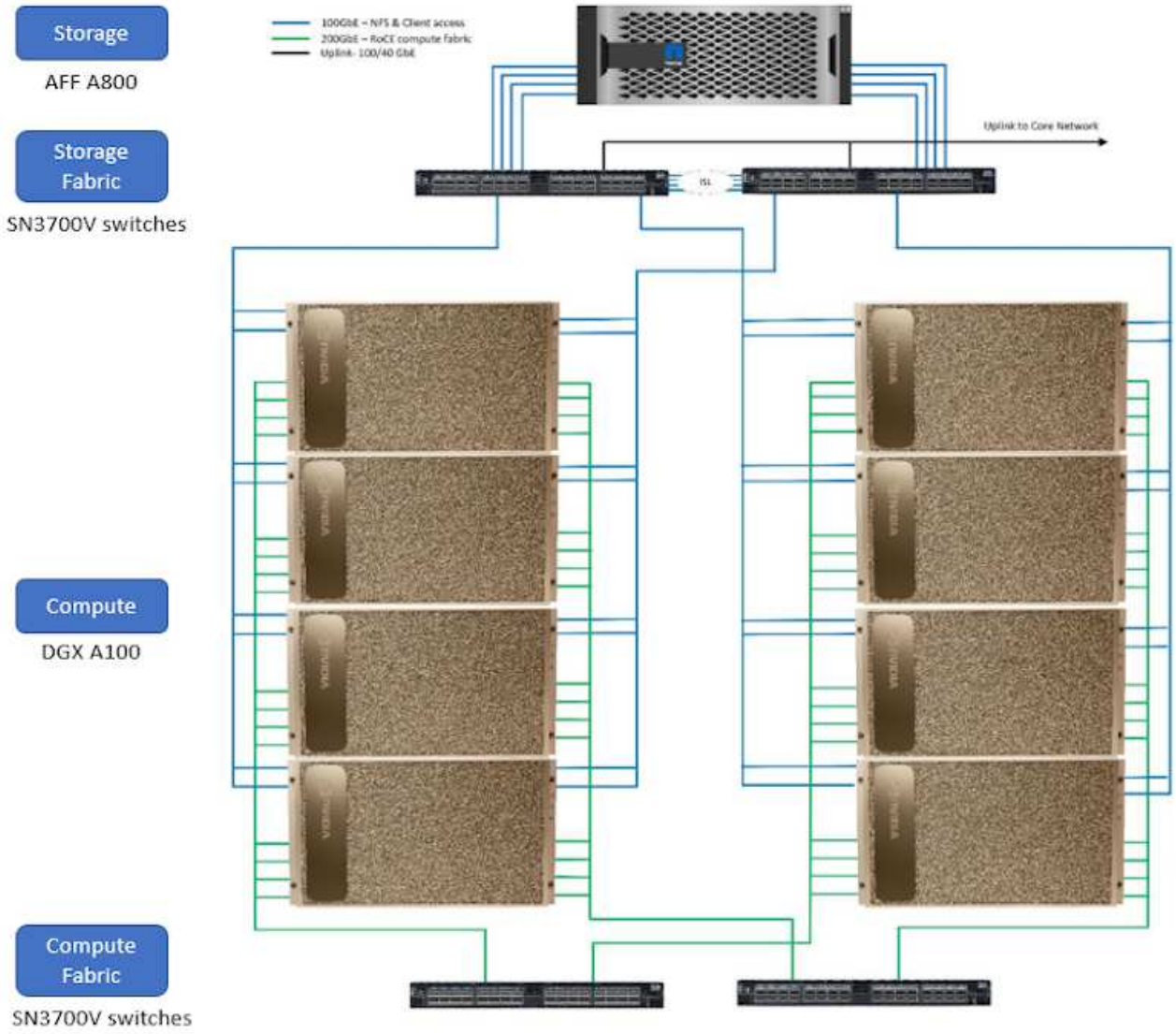
NetApp ONTAP ai

L'architettura di riferimento NetApp ONTAP ai, basata su sistemi DGX A100 e sistemi storage connessi al cloud NetApp, è stata sviluppata e verificata da NetApp e NVIDIA. Offre alle organizzazioni IT un'architettura che offre i seguenti vantaggi:

- Elimina le complessità di progettazione
- Consente una scalabilità indipendente di calcolo e storage
- Consente ai clienti di partire da piccoli e scalare perfettamente
- Offre una gamma di opzioni di storage per diverse performance e costi

ONTAP ai integra perfettamente i sistemi DGX A100 e i sistemi storage NetApp AFF A800 con reti all'avanguardia. ONTAP ai semplifica le implementazioni di ai eliminando la complessità e le congetture di progettazione. I clienti possono iniziare con poco e crescere senza interruzioni, gestendo in modo intelligente i dati dall'edge al core, fino al cloud e viceversa.

La figura seguente mostra diverse varianti della famiglia di soluzioni ai ONTAP con sistemi DGX A100. Le prestazioni del sistema AFF A800 sono verificate con un massimo di otto sistemi DGX A100. Aggiungendo coppie di controller storage al cluster ONTAP, l'architettura può scalare su più rack per supportare molti sistemi DGX A100 e petabyte di capacità storage con performance lineari. Questo approccio offre la flessibilità di modificare i rapporti calcolo-storage in modo indipendente in base alle dimensioni dei modelli DL utilizzati e alle metriche di performance richieste.



Per ulteriori informazioni su ONTAP ai, vedere "NVA-1153: NetApp ONTAP ai con sistemi NVIDIA DGX A100 e switch Ethernet Mellanox Spectrum."

NetApp ONTAP

ONTAP 9.11, 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 9.11 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.

NetApp DataOps Toolkit

NetApp DataOps Toolkit è una libreria Python che consente a sviluppatori, data scientist, ingegneri DevOps e data engineer di eseguire facilmente varie attività di gestione dei dati, come il provisioning quasi istantaneo di un nuovo volume di dati o di un'area di lavoro JupyterLab, la clonazione quasi istantanea di un volume di dati o di un'area di lavoro JupyterLab, Snapshot quasi istantanee di un volume di dati o di uno spazio di lavoro JupyterLab per la tracciabilità o il baselining. Questa libreria Python può funzionare come un'utility a riga di

comando o una libreria di funzioni che è possibile importare in qualsiasi programma Python o notebook Jupyter.

Server di inferenza NVIDIA Triton

NVIDIA Triton Inference Server è un software open-source per l'inferenza che aiuta a standardizzare l'implementazione e l'esecuzione del modello per offrire ai in produzione rapida e scalabile. Triton Inference Server ottimizza l'inferenza ai consentendo ai team di implementare, eseguire e scalare modelli di ai addestrati da qualsiasi framework su qualsiasi infrastruttura basata su GPU o CPU. Triton Inference Server supporta tutti i framework principali, come TensorFlow, NVIDIA TensorRT, PyTorch, MXNet, OpenVINO e così via. Triton si integra con Kubernetes per l'orchestrazione e la scalabilità che puoi utilizzare in tutte le principali piattaforme ai e Kubernetes del cloud pubblico. È inoltre integrato con molte soluzioni software MLOPS.

PyTorch

"PyTorch" È un framework ML open-source. Si tratta di una libreria di tensore ottimizzata per il deep learning che utilizza GPU e CPU. Il pacchetto PyTorch contiene strutture di dati per i tensionatori multidimensionali che forniscono molte utility per la serializzazione efficiente dei tensionatori tra altre utili utility. Dispone inoltre di una controparte CUDA che consente di eseguire i calcoli del tensore su una GPU NVIDIA con funzionalità di calcolo. In questa convalida, utilizziamo la libreria OpenCV-Python (cv2) per validare il nostro modello, sfruttando al contempo i concetti di computer vision più intuitivi di Python.

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 Astra Control

La famiglia di prodotti NetApp Astra offre servizi di storage e gestione dei dati applicativa per le applicazioni Kubernetes on-premise e nel cloud pubblico, basati sulle tecnologie di storage e gestione dei dati di NetApp. Consente di eseguire facilmente il backup delle applicazioni Kubernetes, migrare i dati in un cluster diverso e creare istantaneamente cloni applicativi funzionanti. Se è necessario gestire le applicazioni Kubernetes in esecuzione in un cloud pubblico, consultare la documentazione per "Servizio di controllo Astra". Astra Control Service è un servizio gestito da NetApp che fornisce la gestione dei dati applicativa dei cluster Kubernetes in Google Kubernetes Engine (GKE) e Azure Kubernetes Service (AKS).

NetApp Astra Trident

Astra "Trident" NetApp è uno storage dinamico open-source orchestrator per Docker e Kubernetes che semplifica la creazione, la gestione e il consumo dello storage persistente. Trident, un'applicazione nativa di Kubernetes, viene eseguita direttamente all'interno di un cluster Kubernetes. Trident consente ai clienti di implementare senza problemi le immagini dei container DL sullo storage NetApp e offre un'esperienza di livello Enterprise per le implementazioni dei container ai. Gli utenti di Kubernetes (sviluppatori ML, data scientist e così via) possono creare, gestire e automatizzare orchestrazione e cloning per sfruttare le funzionalità avanzate di gestione dei dati basate sulla tecnologia NetApp.

Copia e sincronizzazione di NetApp BlueXP

"Copia e sincronizzazione di BlueXP" È un servizio NetApp per una sincronizzazione dei dati rapida e sicura. Sia che tu debba trasferire file tra condivisioni di file SMB o NFS on-premise, NetApp StorageGRID, NetApp ONTAP S3, NetApp Cloud Volumes Service, Azure NetApp Files, Amazon Simple Storage Service (Amazon S3), Amazon Elastic file System (Amazon EFS), BLOB di Azure, Google Cloud Storage, o IBM Cloud Object Storage, BlueXP Copy and Sync sposta i file dove ne hai bisogno in modo rapido e sicuro. Una volta trasferiti, i dati sono completamente disponibili per l'utilizzo sia sull'origine che sulla destinazione. BlueXP Copy e Syncc sincronizzano costantemente i dati in base alla pianificazione predefinita, spostando solo i delta, in modo da ridurre al minimo tempo e denaro necessari per la replica dei dati. BlueXP Copy and Sync è un tool software-as-a-service (SaaS) estremamente semplice da configurare e utilizzare. I trasferimenti dei dati attivati da BlueXP Copy e Sync sono effettuati dai broker di dati. Puoi implementare i broker di dati BlueXP Copy e Sync in AWS, Azure, Google Cloud Platform o on-premise.

Classificazione NetApp BlueXP

Basato su potenti algoritmi ai, "Classificazione NetApp BlueXP" offre controlli automatizzati e governance dei dati nell'intero data estate. Puoi individuare facilmente i risparmi sui costi, identificare i problemi di conformità e privacy e trovare opportunità di ottimizzazione. La dashboard di classificazione BlueXP ti fornisce le informazioni utili per identificare i dati duplicati per eliminare la ridondanza, mappare i dati personali, non personali e sensibili e attivare gli avvisi per i dati sensibili e le anomalie.

Piano di test e validazione

Per questa progettazione della soluzione, sono stati validati i seguenti tre scenari:

- Un'attività di inferenza, con e senza offuscamento di Protopia, all'interno di uno spazio di lavoro JupyterLab orchestrato utilizzando il NetApp DataOps Toolkit per Kubernetes.
- Un processo di deduzione in batch, con e senza offuscamento di Protopia, su Kubernetes con un volume di dati orchestrato utilizzando NetApp DataOps Toolkit per Kubernetes.
- Un'attività di deduzione che utilizza un'istanza di NVIDIA Triton Inference Server orchestrata utilizzando NetApp DataOps Toolkit per Kubernetes. Abbiamo applicato l'offuscamento protopico all'immagine prima di invocare l'API di inferenza Triton per simulare il requisito comune che prevede l'offuscamento dei dati trasmessi sulla rete. Questo flusso di lavoro è applicabile ai casi di utilizzo in cui i dati vengono raccolti all'interno di una zona attendibile ma devono essere trasferiti all'esterno di tale zona attendibile per l'deduzione. Senza l'offuscamento di Protopia, non è possibile implementare questo tipo di workflow senza che i dati sensibili abbandonino la zona attendibile.

Eseguire il test della configurazione

La seguente tabella descrive l'ambiente di convalida della progettazione della soluzione.

Componente	Versione
Kubernetes	1.21.6
Driver NetApp Astra Trident CSI	22.01.0
NetApp DataOps Toolkit per Kubernetes	2.3.0
Server di inferenza NVIDIA Triton	21.11-py3

Procedura di test

Questa sezione descrive le attività necessarie per completare la convalida.

Prerequisiti

Per eseguire le attività descritte in questa sezione, è necessario avere accesso a un host Linux o macOS con i seguenti strumenti installati e configurati:

- Kubernetl (configurato per l'accesso a un cluster Kubernetes esistente)
  - È possibile trovare le istruzioni di installazione e configurazione "qui".
- NetApp DataOps Toolkit per Kubernetes
  - È possibile trovare le istruzioni per l'installazione "qui".

Scenario 1 – deduzione on-demand in JupyterLab

1. Creare uno spazio dei nomi Kubernetes per i carichi di lavoro di inferenza ai/ML.

```
$ kubectl create namespace inference
namespace/inference created
```

2. Utilizza il NetApp DataOps Toolkit per eseguire il provisioning di un volume persistente per l'archiviazione dei dati su cui eseguire l'inferenza.

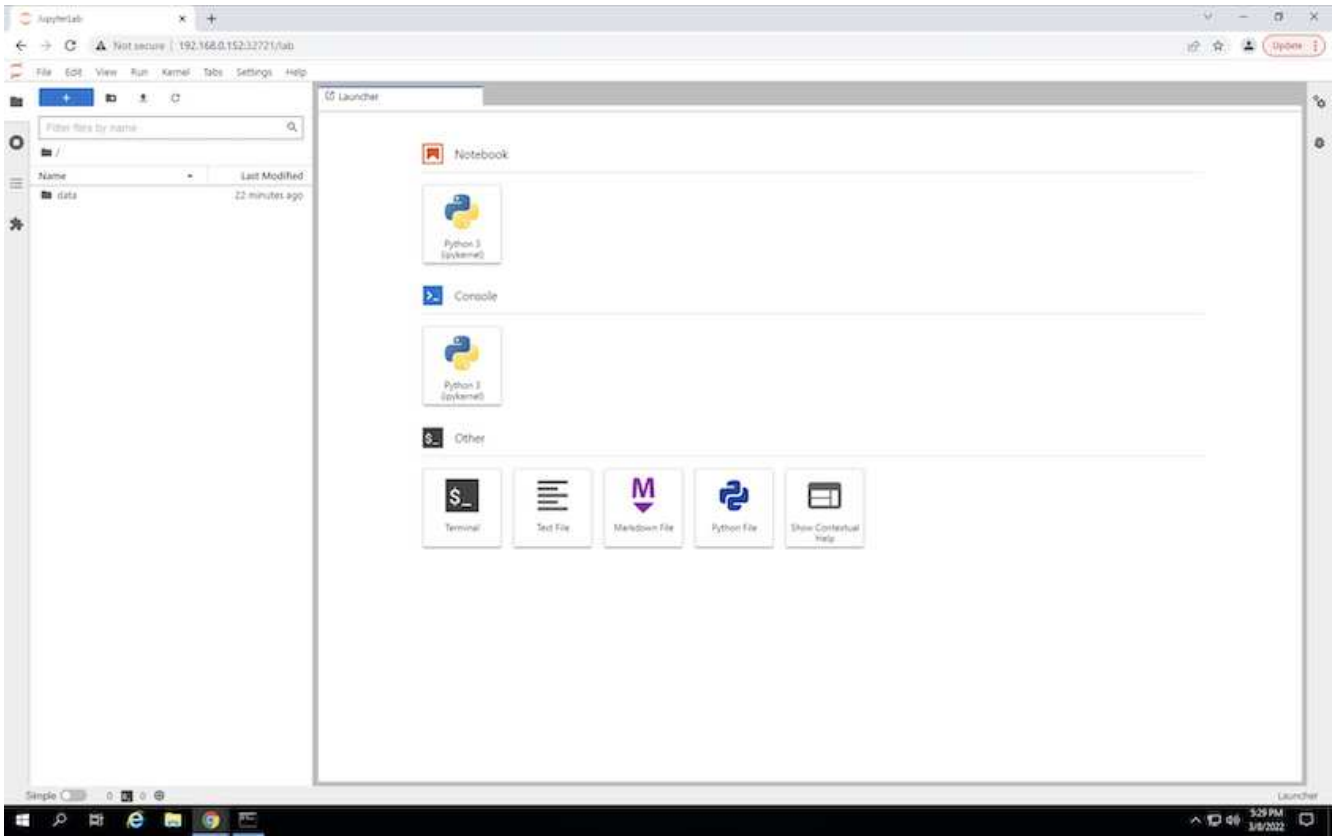
```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=inference-data --size=50Gi
Creating PersistentVolumeClaim (PVC) 'inference-data' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'inference-data' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'inference-data' in namespace 'inference'.
```

3. Utilizza il NetApp DataOps Toolkit per creare un nuovo spazio di lavoro JupyterLab. Montare il volume persistente creato nel passaggio precedente utilizzando `--mount- pvc` opzione. Allocare le GPU NVIDIA nell'area di lavoro secondo necessità utilizzando `-- nvidia-gpu` opzione.

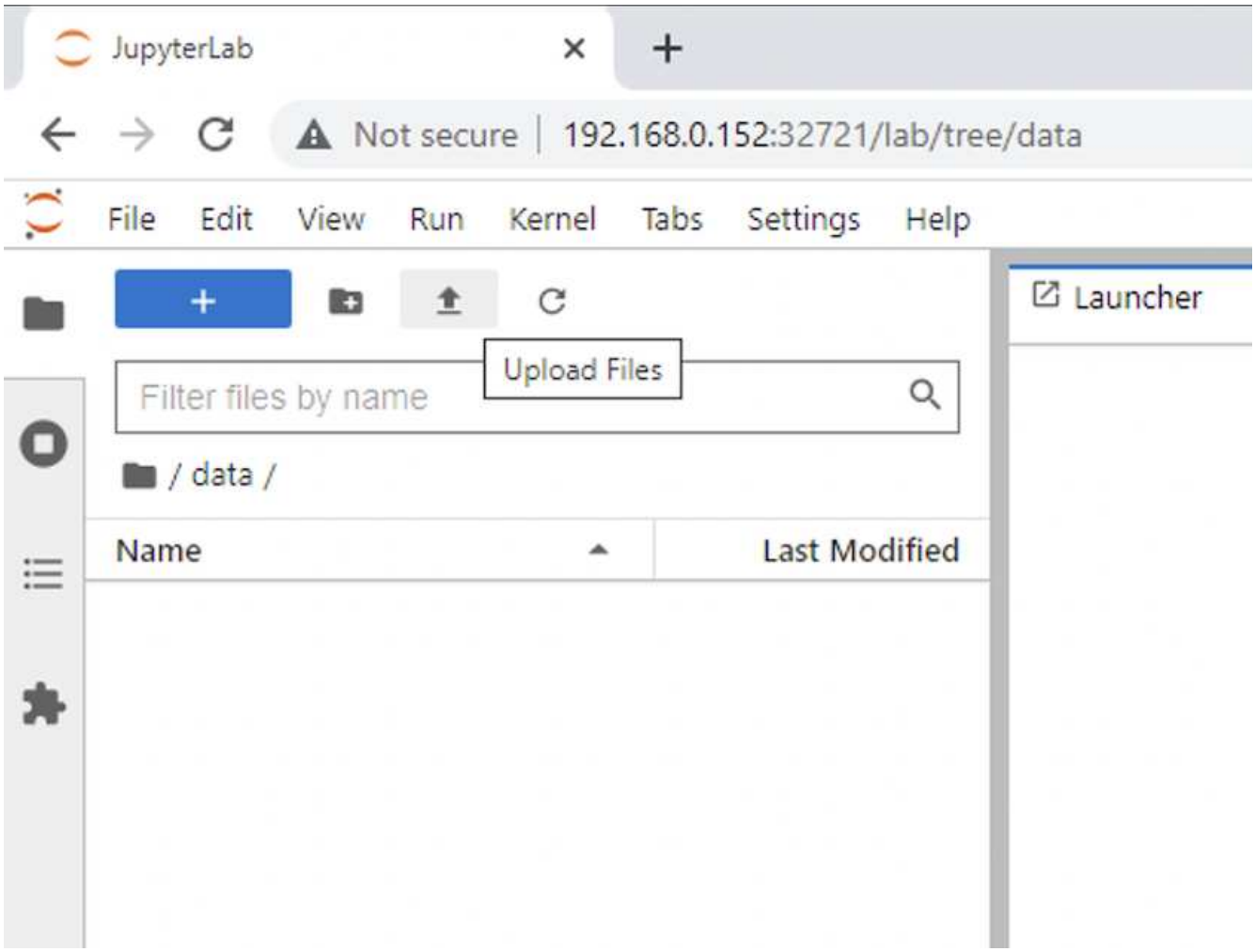
   Nell'esempio seguente, il volume persistente `inference-data` È montato sul container dello spazio di lavoro JupyterLab all'indirizzo `/home/jovyan/data`. Quando si utilizzano le immagini container ufficiali di Project Jupyter, `/home/jovyan` Viene presentato come la directory di primo livello all'interno dell'interfaccia web di JupyterLab.

```
$ netapp_dataops_k8s_cli.py create jupyterlab --namespace=inference
--workspace-name=live-inference --size=50Gi --nvidia-gpu=2 --mount
-pvc=inference-data:/home/jovyan/data
Set workspace password (this password will be required in order to
access the workspace):
Re-enter password:
Creating persistent volume for workspace...
Creating PersistentVolumeClaim (PVC) 'ntap-dsutil-jupyterlab-live-
inference' in namespace 'inference'.
PersistentVolumeClaim (PVC) 'ntap-dsutil-jupyterlab-live-inference'
created. Waiting for Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'ntap-dsutil-jupyterlab-live-inference' in namespace 'inference'.
Creating Service 'ntap-dsutil-jupyterlab-live-inference' in namespace
'inference'.
Service successfully created.
Attaching Additional PVC: 'inference-data' at mount_path:
'/home/jovyan/data'.
Creating Deployment 'ntap-dsutil-jupyterlab-live-inference' in namespace
'inference'.
Deployment 'ntap-dsutil-jupyterlab-live-inference' created.
Waiting for Deployment 'ntap-dsutil-jupyterlab-live-inference' to reach
Ready state.
Deployment successfully created.
Workspace successfully created.
To access workspace, navigate to http://192.168.0.152:32721
```

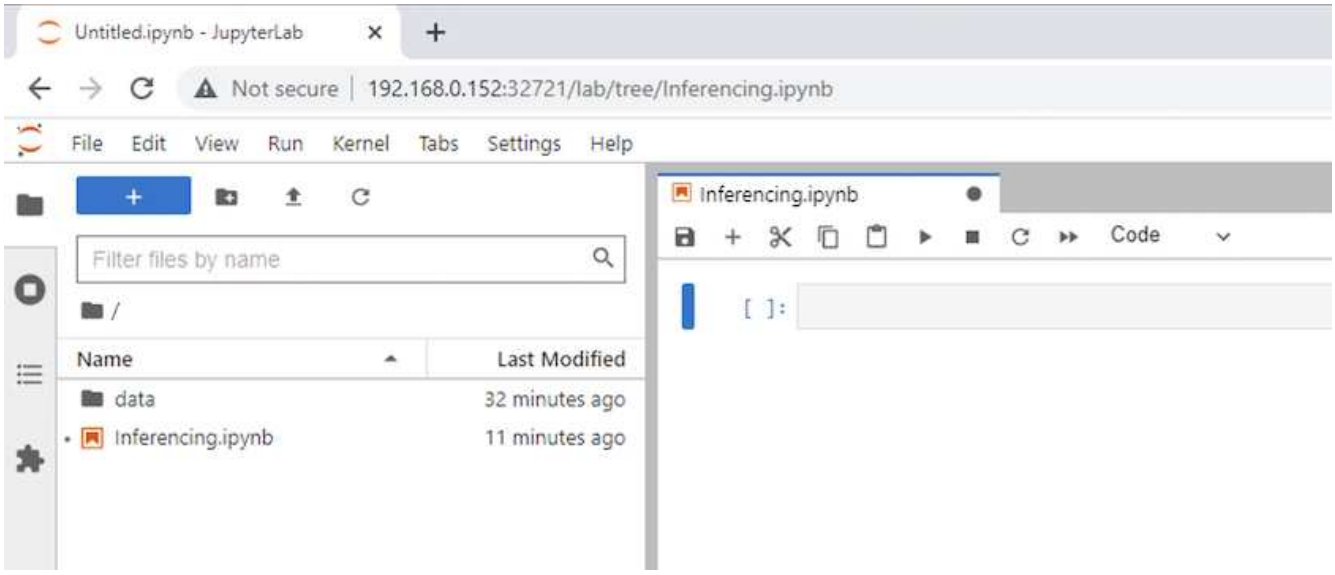
4. Accedere all'area di lavoro di JupyterLab utilizzando l'URL specificato nell'output di `create jupyterlab` comando. La directory dei dati rappresenta il volume persistente montato nell'area di lavoro.



5. Aprire `data` directory e caricare i file su cui eseguire la deduzione. Quando i file vengono caricati nella directory dei dati, vengono memorizzati automaticamente sul volume persistente montato nell'area di lavoro. Per caricare i file, fare clic sull'icona Upload Files (carica file), come mostrato nell'immagine seguente.



6. Tornare alla directory di livello superiore e creare un nuovo notebook.



7. Aggiungere il codice di deduzione al notebook. L'esempio seguente mostra il codice di deduzione per un caso d'uso di rilevamento dell'immagine.

Launcher | image-demo-pytorch.ipynb | Code | Python 3 (ipykernel)

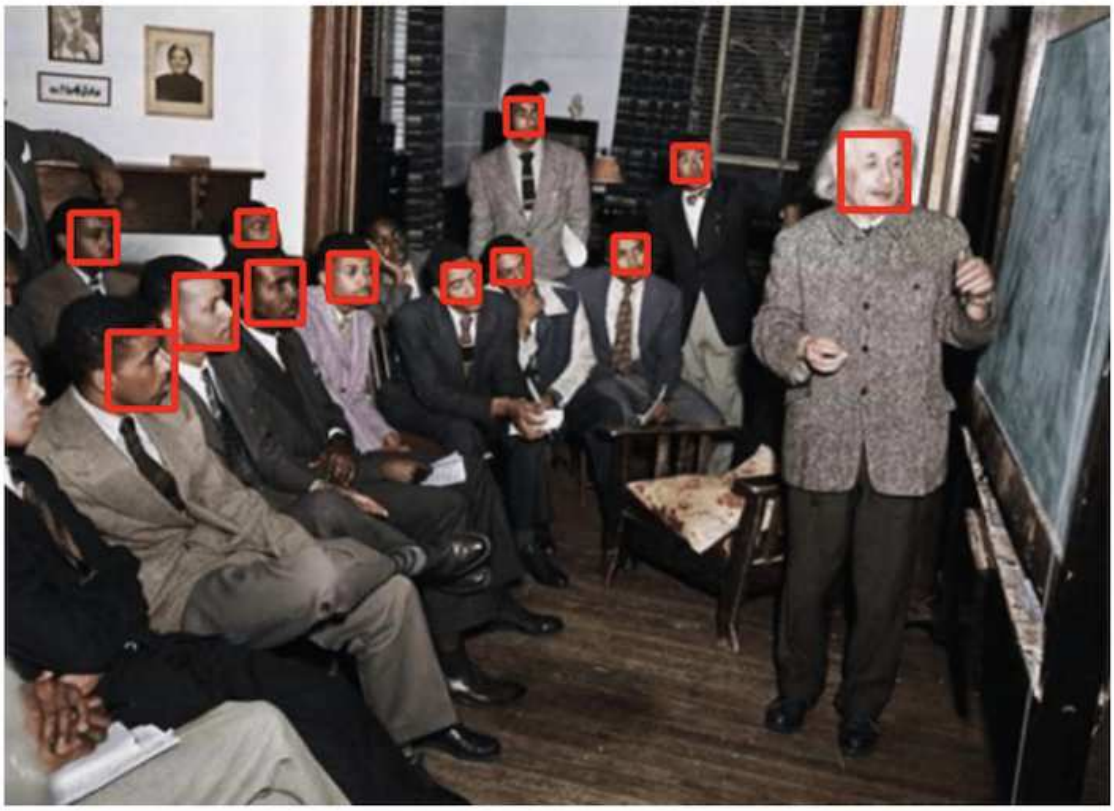
STEP 3-1: Clean (Without obfuscation) detection

[9]:
```
# get current frame
frame = input_image

# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)

# run forward pass
clean_activation = clean_model.forward_head(preprocessed_input)  # runs the first few layers
loc, pred = clean_model.forward_tail(clean_activation)  # runs rest of the layers

# postprocess output
clean_pred = (loc.detach().cpu().numpy(), pred.detach().cpu().numpy())
clean_outputs = postprocess_outputs(
    clean_pred, [[input_image_width, input_image_height]], priors, THRESHOLD
)

# draw rectangles
clean_frame = copy.deepcopy(frame)  # needs to be deep copy
for (x1, y1, x2, y2, s) in clean_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(clean_frame, (x1, y1), (x2, y2), (0, 0, 255), 4)
```

Launcher | image-demo-pytorch.ipynb | Code | Python 3 (ipykernel)

Visualize Clean (Without obfuscation) detection

[10]:
```
show_cv2_image(clean_frame, scale=2)
```



8. Aggiungi l’offuscamento di Protopia al tuo codice di deduzione. Protopia collabora direttamente con i clienti per fornire documentazione specifica per il caso d’utilizzo e non rientra nell’ambito di questo report tecnico. Nell’esempio seguente viene illustrato il codice di deduzione per un caso di utilizzo del rilevamento dell’immagine con l’aggiunta dell’offuscamento di Protopia.

Launcher | image-demo-pytorch.ipynb

Code | Python 3 (ipykernel)

STEP 3-2: Protopia AI (With obfuscation) detection

[11]:
```
# get current frame
frame = input_image

# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)

# run forward pass
not_noisy_activation = noisy_model.forward_head(preprocessed_input)  # runs the first few layers
##################################################################
#          SINGLE ADITIONAL LINE FOR PRIVATE INFERENCE           #
##################################################################
noisy_activation = noisy_model.forward_noise(not_noisy_activation)
##################################################################
loc, pred = noisy_model.forward_tail(noisy_activation)  # runs rest of the layers

# postprocess output
noisy_pred = (loc.detach().cpu().numpy(), pred.detach().cpu().numpy())
noisy_outputs = postprocess_outputs(
    noisy_pred, [[input_image_width, input_image_height]], priors, THRESHOLD * 0.5
)

# get reconstruction of the noisy activation
noisy_reconstruction = decoder_function(noisy_activation)
noisy_reconstruction = noisy_reconstruction.detach().cpu().numpy()[0]
noisy_reconstruction = unpreprocess_output(
    noisy_reconstruction, (input_image_width, input_image_height), True
).astype(np.uint8)

# draw rectangles
for (x1, y1, x2, y2, s) in noisy_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(noisy_reconstruction, (x1, y1), (x2, y2), (0, 0, 255), 4)
```

Launcher | image-demo-pytorch.ipynb

Code | Python 3 (ipykernel)

Visualize Protopia AI (With obfuscation) detection

[12]:
```
show_cv2_image(noisy_reconstruction, scale=2)
```

Scenario 2 – deduzione in batch su Kubernetes

1. Creare uno spazio dei nomi Kubernetes per i carichi di lavoro di inferenza ai/ML.

```
$ kubectl create namespace inference
namespace/inference created
```

2. Utilizza il NetApp DataOps Toolkit per eseguire il provisioning di un volume persistente per l'archiviazione dei dati su cui eseguire l'inferenza.

```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=inference-data --size=50Gi
Creating PersistentVolumeClaim (PVC) 'inference-data' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'inference-data' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'inference-data' in namespace 'inference'.
```

3. Popolare il nuovo volume persistente con i dati su cui eseguire l'deduzione.

   Esistono diversi metodi per caricare i dati su un PVC. Se i tuoi dati sono attualmente memorizzati in una piattaforma di storage a oggetti compatibile con S3, come NetApp StorageGRID o Amazon S3, puoi utilizzare "NetApp DataOps Toolkit S3 Data Mover". Un altro metodo semplice consiste nel creare un'area di lavoro JupyterLab e quindi caricare i file attraverso l'interfaccia web di JupyterLab, come descritto nei passaggi da 3 a 5 della sezione "Scenario 1 – deduzione on-demand in JupyterLab."

4. Creare un lavoro Kubernetes per l'attività di deduzione in batch. Nell'esempio seguente viene illustrato un processo di deduzione in batch per un caso d'uso di rilevamento dell'immagine. Questo lavoro esegue la deduzione su ogni immagine in un set di immagini e scrive le metriche di precisione di deduzione su stdout.

```
$ vi inference-job-raw.yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: netapp-inference-raw
  namespace: inference
spec:
  backoffLimit: 5
  template:
    spec:
      volumes:
      - name: data
        persistentVolumeClaim:
          claimName: inference-data
      - name: dshm
        emptyDir:
          medium: Memory
      containers:
      - name: inference
        image:: netapp-protopia-inference:latest
        imagePullPolicy: IfNotPresent
        command: ["python3", "run-accuracy-measurement.py", "--dataset",
"/data/netapp-face-detection/FDDB"]
        resources:
          limits:
            nvidia.com/gpu: 2
        volumeMounts:
        - mountPath: /data
          name: data
        - mountPath: /dev/shm
          name: dshm
      restartPolicy: Never
$ kubectl create -f inference-job-raw.yaml
job.batch/netapp-inference-raw created
```

5. Verificare che il lavoro di deduzione sia stato completato correttamente.

```
$ kubectl -n inference logs netapp-inference-raw-255sp
100%|██████████| 89/89 [00:52<00:00,  1.68it/s]
Reading Predictions : 100%|██████████| 10/10 [00:01<00:00,  6.23it/s]
Predicting ... : 100%|██████████| 10/10 [00:16<00:00,  1.64s/it]
==================== Results ====================
FDDB-fold-1 Val AP: 0.9491256561145955
FDDB-fold-2 Val AP: 0.9205024466101926
FDDB-fold-3 Val AP: 0.9253013871078468
FDDB-fold-4 Val AP: 0.9399781485863011
FDDB-fold-5 Val AP: 0.9504280149478732
FDDB-fold-6 Val AP: 0.9416473519339292
FDDB-fold-7 Val AP: 0.9241631566241117
FDDB-fold-8 Val AP: 0.9072663297546659
FDDB-fold-9 Val AP: 0.9339648715035469
FDDB-fold-10 Val AP: 0.9447707905560152
FDDB Dataset Average AP: 0.9337148153739079
==================================================
mAP: 0.9337148153739079
```

6. Aggiungi l'offuscamento di Protopia al tuo lavoro di deduzione. È possibile trovare istruzioni specifiche per l'aggiunta di offuscamento Protopia direttamente da Protopia, che non rientra nell'ambito di questo report tecnico. Nell'esempio seguente viene illustrato un processo di deduzione in batch per un caso di utilizzo del rilevamento dei volti con offuscamento di Protopia aggiunto utilizzando un valore ALFA di 0.8. Questo lavoro applica l'offuscamento di Protopia prima di eseguire la deduzione per ogni immagine in un set di immagini e quindi scrive le metriche di precisione dell'inferenza su stdout.

   Abbiamo ripetuto questo passaggio per i valori ALFA 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9 e 0.95. I risultati sono riportati in ""Confronto della precisione delle conferenze"."

```
$ vi inference-job-protopia-0.8.yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: netapp-inference-protopia-0.8
  namespace: inference
spec:
  backoffLimit: 5
  template:
    spec:
      volumes:
      - name: data
        persistentVolumeClaim:
          claimName: inference-data
      - name: dshm
        emptyDir:
          medium: Memory
      containers:
      - name: inference
        image:: netapp-protopia-inference:latest
        imagePullPolicy: IfNotPresent
        env:
        - name: ALPHA
          value: "0.8"
        command: ["python3", "run-accuracy-measurement.py", "--dataset",
"/data/netapp-face-detection/FDDB", "--alpha", "$(ALPHA)", "--noisy"]
        resources:
          limits:
            nvidia.com/gpu: 2
        volumeMounts:
        - mountPath: /data
          name: data
        - mountPath: /dev/shm
          name: dshm
      restartPolicy: Never
$ kubectl create -f inference-job-protopia-0.8.yaml
job.batch/netapp-inference-protopia-0.8 created
```

7. Verificare che il lavoro di deduzione sia stato completato correttamente.

```
$ kubectl -n inference logs netapp-inference-protopia-0.8-b4dkz
100%|██████████| 89/89 [01:05<00:00,  1.37it/s]
Reading Predictions : 100%|██████████| 10/10 [00:02<00:00,  3.67it/s]
Predicting ... : 100%|██████████| 10/10 [00:22<00:00,  2.24s/it]
==================== Results ====================
FDDB-fold-1 Val AP: 0.8953066115834589
FDDB-fold-2 Val AP: 0.8819580264029936
FDDB-fold-3 Val AP: 0.8781107458462862
FDDB-fold-4 Val AP: 0.9085731346308461
FDDB-fold-5 Val AP: 0.9166445508275378
FDDB-fold-6 Val AP: 0.9101178994188819
FDDB-fold-7 Val AP: 0.8383443678423771
FDDB-fold-8 Val AP: 0.8476311547659464
FDDB-fold-9 Val AP: 0.8739624502111121
FDDB-fold-10 Val AP: 0.8905468076424851
FDDB Dataset Average AP: 0.8841195749171925
==================================================
mAP: 0.8841195749171925
```

Scenario 3 – NVIDIA Triton Inference Server

1. Creare uno spazio dei nomi Kubernetes per i carichi di lavoro di inferenza ai/ML.

```
$ kubectl create namespace inference
namespace/inference created
```

2. Utilizza NetApp DataOps Toolkit per eseguire il provisioning di un volume persistente da utilizzare come repository di modelli per NVIDIA Triton Inference Server.

```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=triton-model-repo --size=100Gi
Creating PersistentVolumeClaim (PVC) 'triton-model-repo' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'triton-model-repo' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully created and bound to PersistentVolumeClaim (PVC)
'triton-model-repo' in namespace 'inference'.
```

3. Memorizzare il modello sul nuovo volume persistente in un "formato" Riconosciuto da NVIDIA Triton Inference Server.

   Esistono diversi metodi per caricare i dati su un PVC. Un metodo semplice consiste nel creare un'area di lavoro JupyterLab e quindi caricare i file attraverso l'interfaccia web di JupyterLab, come descritto nei passaggi da 3 a 5 in "Scenario 1 – deduzione on-demand in JupyterLab. "

4. Utilizza NetApp DataOps Toolkit per implementare una nuova istanza di NVIDIA Triton Inference Server.

```
$ netapp_dataops_k8s_cli.py create triton-server --namespace=inference
--server-name=netapp-inference --model-repo-pvc-name=triton-model-repo
Creating Service 'ntap-dsutil-triton-netapp-inference' in namespace
'inference'.
Service successfully created.
Creating Deployment 'ntap-dsutil-triton-netapp-inference' in namespace
'inference'.
Deployment 'ntap-dsutil-triton-netapp-inference' created.
Waiting for Deployment 'ntap-dsutil-triton-netapp-inference' to reach
Ready state.
Deployment successfully created.
Server successfully created.
Server endpoints:
http: 192.168.0.152: 31208
grpc: 192.168.0.152: 32736
metrics: 192.168.0.152: 30009/metrics
```

5. Utilizzare un SDK del client Triton per eseguire un'attività di deduzione. Il seguente estratto di codice Python utilizza l'SDK del client Python di Triton per eseguire un'attività di deduzione per un caso di utilizzo del rilevamento dei volti. Questo esempio chiama l'API Triton e passa un'immagine per la deduzione. Il server di inferenza Triton riceve quindi la richiesta, richiama il modello e restituisce l'output di deduzione come parte dei risultati API.

```
# get current frame
frame = input_image
# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)
# run forward pass
clean_activation = clean_model_head(preprocessed_input)  # runs the
first few layers
#######################################################################
##############
#          pass clean image to Triton Inference Server API for
inferencing           #
#######################################################################
##############
triton_client =
httpclient.InferenceServerClient(url="192.168.0.152:31208",
verbose=False)
model_name = "face_detection_base"
inputs = []
outputs = []
```

```
inputs.append(httpclient.InferInput("INPUT__0", [1, 128, 32, 32],
"FP32"))
inputs[0].set_data_from_numpy(clean_activation.detach().cpu().numpy(),
binary_data=False)
outputs.append(httpclient.InferRequestedOutput("OUTPUT__0",
binary_data=False))
outputs.append(httpclient.InferRequestedOutput("OUTPUT__1",
binary_data=False))
results = triton_client.infer(
    model_name,
    inputs,
    outputs=outputs,
    #query_params=query_params,
    headers=None,
    request_compression_algorithm=None,
    response_compression_algorithm=None)
#print(results.get_response())
statistics =
triton_client.get_inference_statistics(model_name=model_name,
headers=None)
print(statistics)
if len(statistics["model_stats"]) != 1:
    print("FAILED: Inference Statistics")
    sys.exit(1)

loc_numpy = results.as_numpy("OUTPUT__0")
pred_numpy = results.as_numpy("OUTPUT__1")
#######################################################################
##############
# postprocess output
clean_pred = (loc_numpy, pred_numpy)
clean_outputs = postprocess_outputs(
    clean_pred, [[input_image_width, input_image_height]], priors,
THRESHOLD
)
# draw rectangles
clean_frame = copy.deepcopy(frame)  # needs to be deep copy
for (x1, y1, x2, y2, s) in clean_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(clean_frame, (x1, y1), (x2, y2), (0, 0, 255), 4)
```

6. Aggiungi l’offuscamento di Protopia al tuo codice di deduzione. È possibile trovare istruzioni specifiche per il caso d’utilizzo per aggiungere l’offuscamento Protopia direttamente da Protopia; tuttavia, questo processo non rientra nell’ambito di questo report tecnico. Nell’esempio seguente viene illustrato lo stesso codice Python mostrato nel precedente passaggio 5, ma con l’aggiunta dell’offuscamento di Protopia.

Si noti che l'offuscamento Protopia viene applicato all'immagine prima che venga passata all'API Triton. Pertanto, l'immagine non offuscata non lascia mai la macchina locale. Solo l'immagine offuscata viene passata attraverso la rete. Questo flusso di lavoro è applicabile ai casi di utilizzo in cui i dati vengono raccolti all'interno di una zona attendibile, ma devono essere trasferiti all'esterno di tale zona attendibile per l'deduzione. Senza l'offuscamento di Protopia, non è possibile implementare questo tipo di workflow senza che i dati sensibili si allontanino dalla zona di fiducia.

```
# get current frame
frame = input_image
# preprocess input
preprocessed_input = preprocess_input(frame)
preprocessed_input = torch.Tensor(preprocessed_input).to(device)
# run forward pass
not_noisy_activation = noisy_model_head(preprocessed_input)  # runs the
first few layers
#################################################################
#          obfuscate image locally prior to inferencing          #
#          SINGLE ADITIONAL LINE FOR PRIVATE INFERENCE           #
#################################################################
noisy_activation = noisy_model_noise(not_noisy_activation)
#################################################################
#######################################################################
####################
#          pass obfuscated image to Triton Inference Server API for
inferencing           #
#######################################################################
####################
triton_client =
httpclient.InferenceServerClient(url="192.168.0.152:31208",
verbose=False)
model_name = "face_detection_noisy"
inputs = []
outputs = []
inputs.append(httpclient.InferInput("INPUT__0", [1, 128, 32, 32],
"FP32"))
inputs[0].set_data_from_numpy(noisy_activation.detach().cpu().numpy(),
binary_data=False)
outputs.append(httpclient.InferRequestedOutput("OUTPUT__0",
binary_data=False))
outputs.append(httpclient.InferRequestedOutput("OUTPUT__1",
binary_data=False))
results = triton_client.infer(
    model_name,
    inputs,
    outputs=outputs,
    #query_params=query_params,
    headers=None,
```

```
    request_compression_algorithm=None,
    response_compression_algorithm=None)
#print(results.get_response())
statistics =
triton_client.get_inference_statistics(model_name=model_name,
headers=None)
print(statistics)
if len(statistics["model_stats"]) != 1:
    print("FAILED: Inference Statistics")
    sys.exit(1)

loc_numpy = results.as_numpy("OUTPUT__0")
pred_numpy = results.as_numpy("OUTPUT__1")
#######################################################################
###################

# postprocess output
noisy_pred = (loc_numpy, pred_numpy)
noisy_outputs = postprocess_outputs(
    noisy_pred, [[input_image_width, input_image_height]], priors,
THRESHOLD * 0.5
)
# get reconstruction of the noisy activation
noisy_reconstruction = decoder_function(noisy_activation)
noisy_reconstruction = noisy_reconstruction.detach().cpu().numpy()[0]
noisy_reconstruction = unpreprocess_output(
    noisy_reconstruction, (input_image_width, input_image_height), True
).astype(np.uint8)
# draw rectangles
for (x1, y1, x2, y2, s) in noisy_outputs[0]:
    x1, y1 = int(x1), int(y1)
    x2, y2 = int(x2), int(y2)
    cv2.rectangle(noisy_reconstruction, (x1, y1), (x2, y2), (0, 0, 255),
4)
```

Confronto della precisione delle deduzione

Per questa convalida, abbiamo eseguito la deduzione per un caso d'utilizzo di rilevamento dell'immagine utilizzando un set di immagini raw. Quindi, abbiamo eseguito la stessa attività di deduzione sullo stesso set di immagini con l'aggiunta dell'offuscamento di Protopia prima dell'inferenza. Abbiamo ripetuto l'attività utilizzando diversi valori DI ALPHA per la componente di offuscamento Protopia. Nel contesto dell'offuscamento di Protopia, il valore ALFA rappresenta la quantità di offuscamento applicata, con un valore ALFA più alto che rappresenta un livello più elevato di offuscamento. Abbiamo quindi confrontato la precisione delle deduzione in queste

diverse esecuzioni.

Le due tabelle seguenti forniscono dettagli sul nostro caso di utilizzo e delineano i risultati.

Protopia collabora direttamente con i clienti per determinare il valore ALFA appropriato per un caso di utilizzo specifico.

Componente	Dettagli
Modello	FaceBoxes (PyTorch) -
Dataset	Dataset FDDB

Offuscamento di Protopia	ALFA	Precisione
No	N/A.	0.9337148153739079
Sì	0.05	0.9028766627325002
Sì	0.1	0.9024301009661478
Sì	0.2	0.9081836283186224
Sì	0.4	0.9073066107482036
Sì	0.6	0.8847816568680239
Sì	0.8	0.8841195749171925
Sì	0.9	0.8455427675252052
Sì	0.95	0.8455427675252052

Velocità di offuscamento

Per questa convalida, abbiamo applicato l'offuscamento di Protopia a un'immagine da 1920 x 1080 pixel cinque volte e misurato il tempo necessario per completare ogni volta la fase di offuscamento.

Abbiamo utilizzato PyTorch in esecuzione su una singola GPU NVIDIA V100 per applicare l'offuscamento e abbiamo cancellato la cache della GPU tra un'esecuzione e l'altra. La fase di offuscamento ha richiesto rispettivamente 5,47 ms, 5,27 ms, 4,54 ms, 5,24 ms e 4,84 ms per completare le cinque corse. La velocità media è stata di 5,072 ms.

Conclusione

I dati esistono in tre stati: A riposo, in transito e nel calcolo. Una parte importante di qualsiasi servizio di inferenza ai dovrebbe essere la protezione dei dati dalle minacce durante l'intero processo. La protezione dei dati durante le deduzione è fondamentale perché il processo può esporre informazioni private sui clienti esterni e sull'azienda che fornisce il servizio di deduzione. Protopia ai è una soluzione software-only non invadente per deduzione ai riservata nel mercato odierno. Con Protopia, l'ai viene alimentata solo le informazioni trasformate nei record di dati che sono essenziali per eseguire l'attività di ai/ML a portata di mano e niente di più. Questa trasformazione stocastica non è una forma di mascheramento e si basa sul cambiamento matematico della rappresentazione

dei dati utilizzando il rumore curato.

I sistemi storage NetApp con funzionalità ONTAP offrono le stesse performance o migliori dello storage SSD locale e, in combinazione con il toolkit NetApp DataOps, offrono i seguenti vantaggi a data scientist, data engineer, sviluppatori ai/ML e decision maker IT aziendali o aziendali:

- Condivisione semplice dei dati tra sistemi ai, analytics e altri sistemi aziendali critici. Questa condivisione dei dati riduce l'overhead dell'infrastruttura, migliora le performance e ottimizza la gestione dei dati in tutta l'azienda.
- Calcolo e storage scalabili in maniera indipendente per ridurre al minimo i costi e migliorare l'utilizzo delle risorse.
- Workflow di sviluppo e implementazione ottimizzati grazie a copie Snapshot e cloni integrati per spazi di lavoro degli utenti istantanei ed efficienti in termini di spazio, controllo integrato delle versioni e implementazione automatizzata.
- Protezione dei dati di livello Enterprise e governance dei dati per disaster recovery, business continuity e requisiti normativi.
- Invocazione semplificata delle operazioni di gestione dei dati; copie Snapshot delle aree di lavoro dei data scientist per il backup e la tracciabilità dal NetApp DataOps Toolkit nei notebook Jupyter.

La soluzione NetApp e Protopia offre un'architettura scalabile e flessibile, ideale per le implementazioni di inferenza ai di livello Enterprise. Consente la protezione dei dati e fornisce privacy per le informazioni sensibili, laddove i requisiti di inferenza ai confidenziali possono essere soddisfatti con pratiche ai responsabili sia nelle implementazioni on-premise che nel cloud ibrido.

Dove trovare ulteriori informazioni e riconoscimenti

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

- Storage persistente NetApp per container - NetApp Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- NetApp DataOps Toolkit

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

- Storage persistente NetApp per container: NetApp Astra Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- Protopia ai: Inferenza riservata

  "https://protopia.ai/blog/protopia-ai-takes-on-the-missing-link-in-ai-privacy-confidential-inference/"

- Copia e sincronizzazione di NetApp BlueXP

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"

- Server di inferenza NVIDIA Triton

  "https://developer.nvidia.com/nvidia-triton-inference-server"

- Documentazione di NVIDIA Triton Inference Server

  "https://docs.nvidia.com/deeplearning/triton-inference-server/index.html"

- FaceBoxes in PyTorch

  "https://github.com/zisianw/FaceBoxes.PyTorch"

Ringraziamenti

- Mark Cates, Principal Product Manager, NetApp
- Sufian Ahmad, Technical Marketing Engineer, NetApp
- Hadi Esmaeilzadeh, Chief Technology Officer e Professor, Protopia ai

Analisi del sentimento con NetApp ai

TR-4910: Analisi del sentimento da Customer Communications con NetApp ai

Rick Huang, Sathish Thyagarajan e David Arnette, NetApp Diego Sosa-Coba, SFL Scientific

Questo report tecnico fornisce ai clienti indicazioni di progettazione per eseguire l'analisi del sentimento in un centro di supporto globale di livello Enterprise utilizzando le tecnologie di gestione dei dati NetApp con un framework software NVIDIA che utilizza l'apprendimento del trasferimento e l'intelligenza artificiale conversazionale. Questa soluzione è applicabile a qualsiasi settore che desideri ottenere informazioni sui clienti da file di testo o vocali registrati che rappresentano registri di chat, e-mail e altre comunicazioni di testo o audio. Abbiamo implementato una pipeline end-to-end per dimostrare il riconoscimento vocale automatico, l'analisi del sentimento in tempo reale e le funzionalità di riqualificazione del modello di elaborazione del linguaggio naturale di apprendimento approfondito su un cluster di calcolo accelerato dalla GPU con storage all-flash NetApp connesso al cloud. È possibile formare e ottimizzare enormi modelli linguistici all'avanguardia per eseguire rapidamente l'inferenza con il centro di supporto globale, al fine di creare un'esperienza del cliente eccezionale e valutazioni obiettive e a lungo termine delle performance dei dipendenti.

L'analisi del sentimento è un campo di studio all'interno di Natural Language Processing (NLP) attraverso il quale i sentimenti positivi, negativi o neutri vengono estratti dal testo. I sistemi di intelligenza artificiale convergente sono saliti a un livello di integrazione quasi globale man mano che sempre più persone vengono a interagire con loro. L'analisi del sentimento ha una varietà di casi di utilizzo, dalla determinazione delle performance dei dipendenti del centro di supporto nelle conversazioni con i chiamanti e la fornitura di risposte dei chatbot automatizzate appropriate alla previsione del prezzo delle azioni di un'azienda in base alle interazioni tra i rappresentanti dell'azienda e il pubblico alle chiamate trimestrali sui guadagni. Inoltre, l'analisi del sentimento può essere utilizzata per determinare la posizione del cliente sui prodotti, servizi o supporto forniti dal marchio.

Questa soluzione end-to-end utilizza modelli NLP per eseguire un'analisi del sentimento di alto livello che abilita i framework analitici del centro di supporto. Le registrazioni audio vengono elaborate in testo scritto e il sentimento viene estratto da ciascuna frase della conversazione. I risultati, aggregati in una dashboard, possono essere creati per analizzare i sentimenti delle conversazioni, sia storicamente che in tempo reale. Questa soluzione può essere generalizzata ad altre soluzioni con modalità dati e esigenze di output simili. Con i dati appropriati, è possibile eseguire altri casi di utilizzo. Ad esempio, è possibile analizzare le richieste di guadagno dell'azienda per verificare il sentimento utilizzando la stessa pipeline end-to-end. Altre forme di analisi NLP, come la modellazione topica e il riconoscimento di entità nominate (NER), sono possibili anche grazie alla natura flessibile della pipeline.

Queste implementazioni di ai sono state rese possibili da NVIDIA RIVA, NVIDIA TAO Toolkit e NetApp DataOps Toolkit che hanno collaborato. I tool NVIDIA vengono utilizzati per implementare rapidamente soluzioni ai dalle performance elevate utilizzando modelli e pipeline precostruiti. Il NetApp DataOps Toolkit semplifica varie attività di gestione dei dati per accelerare lo sviluppo.

Valore per il cliente

Le aziende vedono il valore di uno strumento di valutazione dei dipendenti e reazione dei clienti per conversazioni testuali, audio e video per l'analisi del sentimento. I manager traggono vantaggio dalle informazioni presentate nella dashboard, consentendo una valutazione dei dipendenti e della soddisfazione dei clienti in base a entrambe le parti della conversazione.

Inoltre, il NetApp DataOps Toolkit gestisce la versione e l'allocazione dei dati all'interno dell'infrastruttura del cliente. Questo porta ad aggiornamenti frequenti delle analisi presentate all'interno della dashboard senza creare costi di storage dei dati ingombranti.

Casi di utilizzo

A causa del numero di chiamate che questi centri di supporto elaborano, la valutazione delle performance delle chiamate potrebbe richiedere molto tempo se eseguita manualmente. I metodi tradizionali, come il conteggio delle parole e altri metodi, possono ottenere una certa automazione, ma questi metodi non acquisiscono aspetti più sfumati e contesto semantico del linguaggio dinamico. È possibile utilizzare tecniche di modellazione ai per eseguire alcune di queste analisi più sfumate in modo automatizzato. Inoltre, con gli attuali tool di modellazione pre-addestrati e all'avanguardia pubblicati da NVIDIA, AWS, Google e altri, una pipeline end-to-end con modelli complessi può ora essere messa in piedi e personalizzata con relativa facilità.

Una pipeline end-to-end per l'analisi del sentimento del centro di supporto consente di acquisire file audio in tempo reale mentre i dipendenti conversano con i chiamanti. Quindi, questi file audio vengono elaborati per l'utilizzo nel componente voce-testo che li converte in un formato di testo. Ogni frase della conversazione riceve un'etichetta indicante il sentimento (positivo, negativo o neutro).

L'analisi del sentimento può fornire un aspetto essenziale delle conversazioni per la valutazione delle performance delle chiamate. Questi sentimenti aggiungono un ulteriore livello di profondità alle interazioni tra dipendenti e chiamanti. Il dashboard dedicato al sentimento assistito dall'ai offre ai manager un monitoraggio in tempo reale del sentimento all'interno di una conversazione, oltre a un'analisi retrospettiva delle chiamate passate del dipendente.

Esistono tool precostruiti che possono essere combinati in modi potenti per creare rapidamente una pipeline di ai end-to-end per risolvere questo problema. In questo caso, la libreria NVIDIA RIVA può essere utilizzata per eseguire le due attività in-series: Trascrizione audio e analisi del sentimento. Il primo è un algoritmo di elaborazione del segnale di apprendimento supervisionato e il secondo è un algoritmo di classificazione NLP

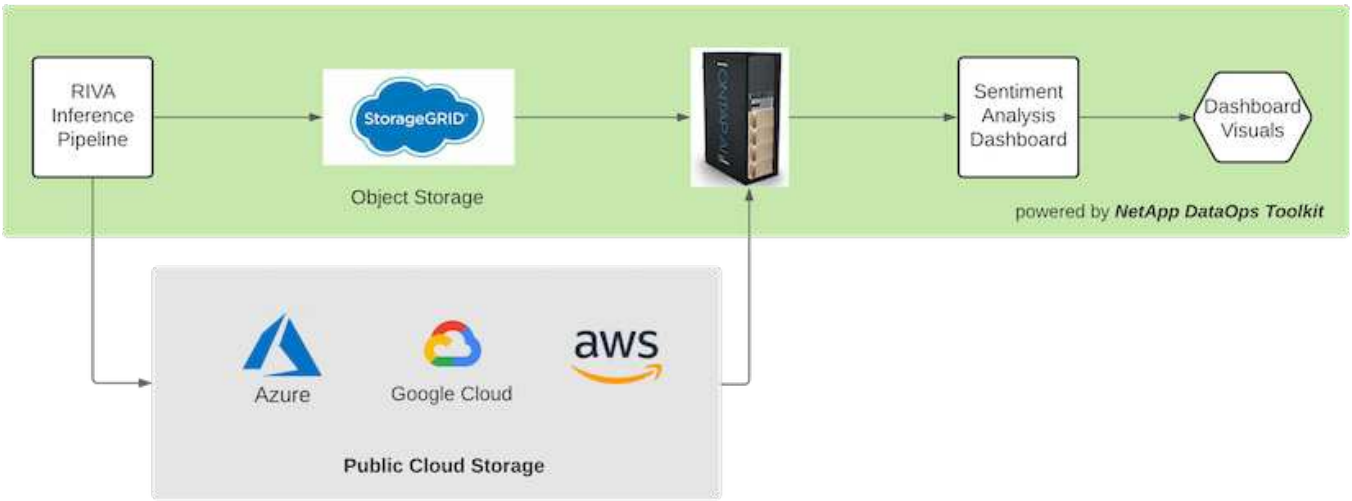
di apprendimento supervisionato. Questi algoritmi pronti all'uso possono essere ottimizzati per qualsiasi caso di utilizzo pertinente con dati rilevanti per l'azienda utilizzando NVIDIA TAO Toolkit. Questo porta a soluzioni più accurate e potenti che vengono costruite solo per una frazione dei costi e delle risorse. I clienti possono incorporare "NVIDIA Maxine" Framework per applicazioni di videoconferenza accelerate dalla GPU nella progettazione del centro di supporto.

I seguenti casi di utilizzo sono alla base di questa soluzione. Entrambi i casi di utilizzo utilizzano il toolkit TAO per la messa a punto del modello e RIVA per l'implementazione del modello.

- Voce-testo
- Analisi del sentimento

Per analizzare le interazioni del centro di supporto tra dipendenti e clienti, ogni conversazione con il cliente sotto forma di chiamate audio può essere eseguita attraverso la pipeline per estrarre sentimenti a livello di frase. Tali sentimenti possono quindi essere verificati da un essere umano per giustificare i sentimenti o modificarli in base alle necessità. I dati etichettati vengono quindi trasferiti nella fase di messa a punto per migliorare le previsioni del sentimento. Se esistono già dati di sentimento etichettati, è possibile accelerare la messa a punto del modello. In entrambi i casi, la pipeline è generalizzabile con altre soluzioni che richiedono l'acquisizione di audio e la classificazione delle frasi.

Gli output di ai sentiment vengono caricati su un database cloud esterno o su un sistema storage gestito dall'azienda. Gli output del sentimento vengono trasferiti da questo database più grande allo storage locale per l'utilizzo all'interno della dashboard che visualizza l'analisi del sentimento per i manager. La funzionalità principale del dashboard consiste nell'interfacciarsi con il dipendente del servizio clienti in tempo reale. I manager possono valutare e fornire un feedback sui dipendenti durante le loro chiamate con aggiornamenti in tempo reale del sentimento di ciascuna frase, nonché una revisione storica delle performance del dipendente o delle reazioni dei clienti.



Il "NetApp DataOps Toolkit" Può continuare a gestire i sistemi di storage dei dati anche dopo che la pipeline di inferenza RIVA ha generato etichette di sentimento. Questi risultati ai possono essere caricati su un sistema storage gestito dal NetApp DataOps Toolkit. I sistemi di storage dei dati devono essere in grado di gestire centinaia di inserti e selezionare ogni minuto. Il sistema di storage dei dispositivi locali esegue query in tempo reale sull'ampio storage dei dati per l'estrazione. È inoltre possibile eseguire query sull'istanza di storage dei dati più grande per ottenere dati storici per migliorare ulteriormente l'esperienza del dashboard. Il NetApp DataOps Toolkit facilita entrambi questi utilizzi clonando rapidamente i dati e distribuirli in tutte le dashboard che li utilizzano.

Pubblico di destinazione

Il pubblico di riferimento per la soluzione comprende i seguenti gruppi:

- Responsabili dei dipendenti
- Data engineer/data scientist
- Amministratori IT (on-premise, cloud o ibridi)

Tenere traccia dei sentimenti durante le conversazioni è uno strumento prezioso per valutare le performance dei dipendenti. Utilizzando la dashboard di ai, i manager possono vedere come dipendenti e chiamanti cambiano le proprie sensazioni in tempo reale, consentendo valutazioni live e sessioni di guida. Inoltre, le aziende possono ottenere preziose informazioni sui clienti dai clienti impegnati in conversazioni vocali, chat di testo e videoconferenze. Tali analisi dei clienti utilizzano le funzionalità di elaborazione multimodale su larga scala con modelli e flussi di lavoro ai moderni e all'avanguardia.

Dal punto di vista dei dati, un gran numero di file audio viene elaborato quotidianamente dal centro di supporto. Il NetApp DataOps Toolkit facilita questa attività di gestione dei dati per la messa a punto periodica di modelli e dashboard di analisi del sentimento.

Gli amministratori IT traggono vantaggio anche dal NetApp DataOps Toolkit, che consente loro di spostare rapidamente i dati tra ambienti di implementazione e produzione. Anche gli ambienti e i server NVIDIA devono essere gestiti e distribuiti per consentire l'inferenza in tempo reale.

Architettura

L'architettura di questa soluzione di Support Center si basa sugli strumenti predefiniti di NVIDIA e sul NetApp DataOps Toolkit. I tool NVIDIA vengono utilizzati per implementare

rapidamente soluzioni ai ad alte performance utilizzando modelli e pipeline precostruiti. Il NetApp DataOps Toolkit semplifica varie attività di gestione dei dati per accelerare lo sviluppo.

Tecnologia della soluzione

"NVIDIA RIVA" È un SDK con accelerazione GPU per la creazione di applicazioni ai converazionali multimodali che offrono performance in tempo reale sulle GPU. Il toolkit NVIDIA Train, Adapt, and Optimize (TAO) offre un modo più rapido e semplice per accelerare la formazione e creare rapidamente modelli di ai altamente accurati e performanti, specifici per il dominio.

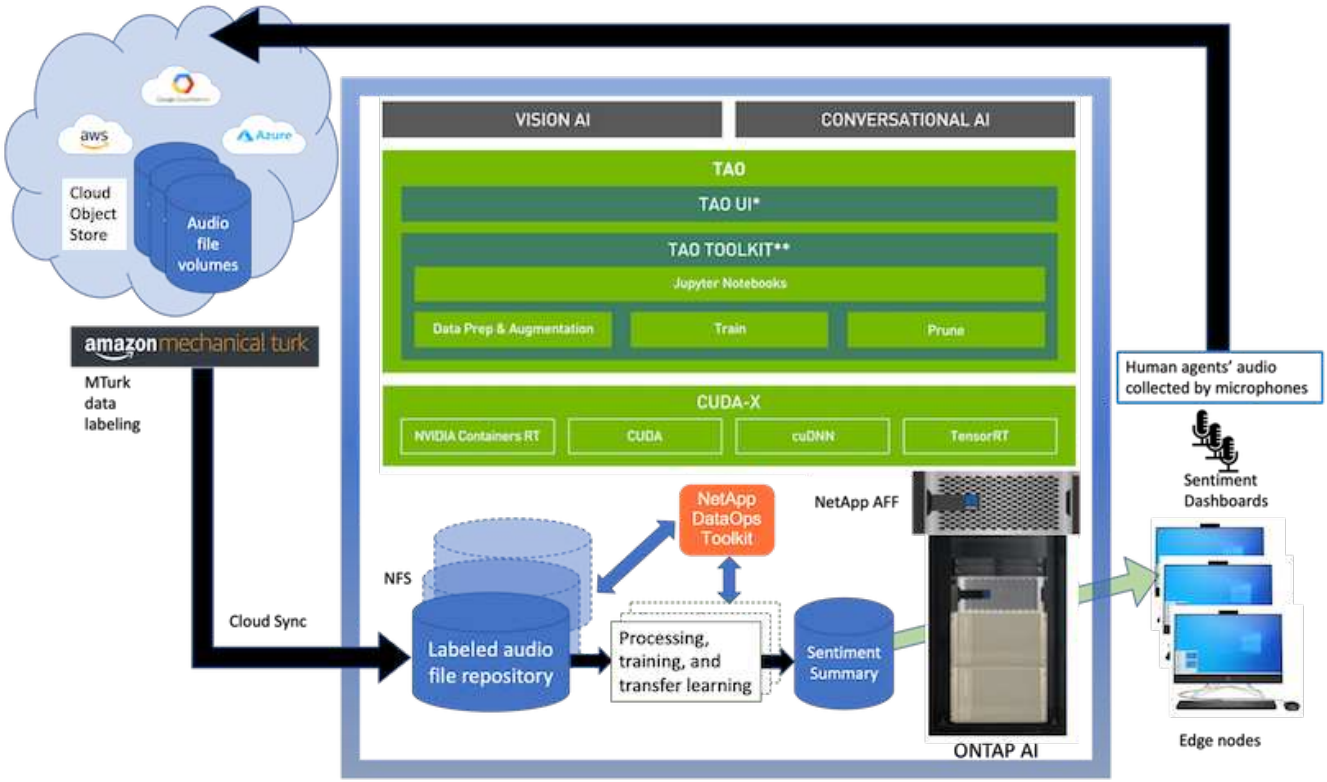
Il NetApp DataOps Toolkit è una libreria Python che semplifica l'esecuzione di varie attività di gestione dei dati da parte di sviluppatori, data scientist, ingegneri DevOps e data engineer. Ciò include il provisioning quasi istantaneo di un nuovo volume di dati o di uno spazio di lavoro JupyterLab, la clonazione quasi istantanea di un volume di dati o di uno spazio di lavoro JupyterLab e lo snap-shoting quasi istantaneo di un volume di dati o di uno spazio di lavoro JupyterLab per la tracciabilità e la baselining.

Diagramma architetturale

Il seguente diagramma illustra l'architettura della soluzione. Esistono tre categorie di ambiente principali: Cloud, core e edge. Ciascuna delle categorie può essere distribuita geograficamente. Ad esempio, il cloud contiene archivi di oggetti con file audio in bucket in regioni diverse, mentre il core potrebbe contenere data center collegati tramite una rete ad alta velocità o Copia e sincronizzazione di NetApp BlueXP. I nodi edge denotano le piattaforme di lavoro quotidiane dei singoli agenti umani, in cui sono disponibili strumenti di dashboard interattivi e microfoni per visualizzare il sentimento e raccogliere dati audio dalle conversazioni con i clienti.

Nei data center con accelerazione GPU, le aziende possono utilizzare NVIDIA "RIVA" Framework per la creazione di applicazioni ai conversazionali, alle quali il "Toolkit Tao" Si connette per il finetuning e la riqualificazione dei modelli utilizzando tecniche di trasferimento L-learning. Queste applicazioni di calcolo e i flussi di lavoro sono basati su "NetApp DataOps Toolkit", Che offre le migliori funzionalità di gestione dei dati offerte da ONTAP. Il toolkit consente ai team di dati aziendali di prototipare rapidamente i propri modelli con dati strutturati e non strutturati associati tramite snapshot e cloni per tracciabilità, controllo delle versioni, test A/B, fornendo così sicurezza, governance, e conformità alle normative. Vedere la sezione "Progettazione dello storage" per ulteriori dettagli.

Questa soluzione dimostra l'elaborazione dei file audio, il training sul modello NLP, l'apprendimento del trasferimento e le fasi dettagliate della gestione dei dati. La pipeline end-to-end risultante genera un riepilogo dei sentimenti che viene visualizzato in tempo reale sui dashboard degli agenti di supporto umano.



Requisiti hardware

La seguente tabella elenca i componenti hardware necessari per implementare la soluzione. I componenti hardware utilizzati in una particolare implementazione della soluzione possono variare in base ai requisiti del cliente.

Test di latenza della risposta	Tempo (millisecondi)
Elaborazione dei dati	10
Deduzione	10

Questi test dei tempi di risposta sono stati eseguiti su oltre 50,000 file audio in 560 conversazioni. Ogni file audio era di ~100 KB come MP3 e ~1 MB quando convertito in WAV. La fase di elaborazione dei dati converte gli MP3 in file WAV. I passaggi di inferenza convertono i file audio in testo ed estraggono un sentimento dal testo. Questi passaggi sono tutti indipendenti l'uno dall'altro e possono essere parallelizzati per accelerare il processo.

Tenendo conto della latenza del trasferimento dei dati tra gli archivi, i manager dovrebbero essere in grado di visualizzare gli aggiornamenti dell'analisi del sentimento in tempo reale entro un secondo dalla fine della frase.

Hardware NVIDIA RIVA

Hardware	Requisiti
SISTEMA OPERATIVO	Linux x86_64
Memoria GPU (ASR)	Modelli di streaming: ~5600 MB modelli senza streaming: ~3100 MB
Memoria GPU (NLP)	~500 MB per modello BERT

Hardware NVIDIA TAO Toolkit

Hardware	Requisiti
RAM di sistema	32 GB
RAM GPU	32 GB
CPU	8 core
GPU	NVIDIA (A100, V100 e RTX 30x0)
SSD	100 GB

Sistema storage flash

NetApp ONTAP 9

ONTAP 9.9, 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 9.9 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.

Copia e sincronizzazione di NetApp BlueXP

"Copia e sincronizzazione di BlueXP" È un servizio NetApp per una sincronizzazione dei dati rapida e sicura che consente di trasferire file tra condivisioni di file NFS o SMB on-premise a una delle seguenti destinazioni:

- NetApp StorageGRID
- NetApp ONTAP S3
- NetApp Cloud Volumes Service
- Azure NetApp Files
- Amazon Simple Storage Service (Amazon S3)
- Amazon Elastic file System (Amazon EFS)
- Azure Blob
- Storage Google Cloud
- Storage a oggetti IBM Cloud

BlueXP Copy e Sync sposta i file dove ne hai bisogno in modo rapido e sicuro. Una volta trasferiti, i dati sono completamente disponibili per l'utilizzo sia sull'origine che sulla destinazione. BlueXP Copy e Sync sincronizza costantemente i dati in base alla pianificazione predefinita, spostando solo i delta, in modo da ridurre al minimo il tempo e il denaro necessari per la replica. BlueXP Copy and Sync è un tool software as a service (SaaS) semplice da configurare e utilizzare. I trasferimenti dei dati attivati da BlueXP Copy e Sync sono effettuati dai broker di dati. Puoi implementare i broker di dati BlueXP Copy e Sync in AWS, Azure, Google Cloud Platform o on-premise.

NetApp StorageGRID

La suite di storage a oggetti software-defined di StorageGRID supporta un'ampia gamma di casi di utilizzo in ambienti multi-cloud pubblici, privati e ibridi. Grazie alle innovazioni leader del settore, NetApp StorageGRID

memorizza, protegge, protegge e preserva i dati non strutturati per un utilizzo multiuso, inclusa la gestione automatica del ciclo di vita per lunghi periodi di tempo. Per ulteriori informazioni, consultare "NetApp StorageGRID" sito.

Requisiti software

La seguente tabella elenca i componenti software necessari per implementare questa soluzione. I componenti software utilizzati in una particolare implementazione della soluzione possono variare in base ai requisiti del cliente.

Computer host	Requisiti
RIVA (in precedenza JARVIS)	1.4.0
TAO Toolkit (in precedenza Transfer Learning Toolkit)	3.0
ONTAP	9.9.1
SISTEMA OPERATIVO DGX	5.1
DOTK	2.0.0

Software NVIDIA RIVA

Software	Requisiti
Docker	>19.02 (con nvidia-docker installato)>=19.03 se non si utilizza DGX
Driver NVIDIA	465.19.01+ 418.40+, 440.33+, 450.51+, 460.27+ per GPU Data Center
Sistema operativo container	Ubuntu 20.04
CUDA	11.3.0
CuBLAS	11.5.1.101
CuDNN	8.2.0.41
NCCL	2.9.6
TensorRT	7.2.3.4
Server di inferenza Triton	2.9.0

Software NVIDIA TAO Toolkit

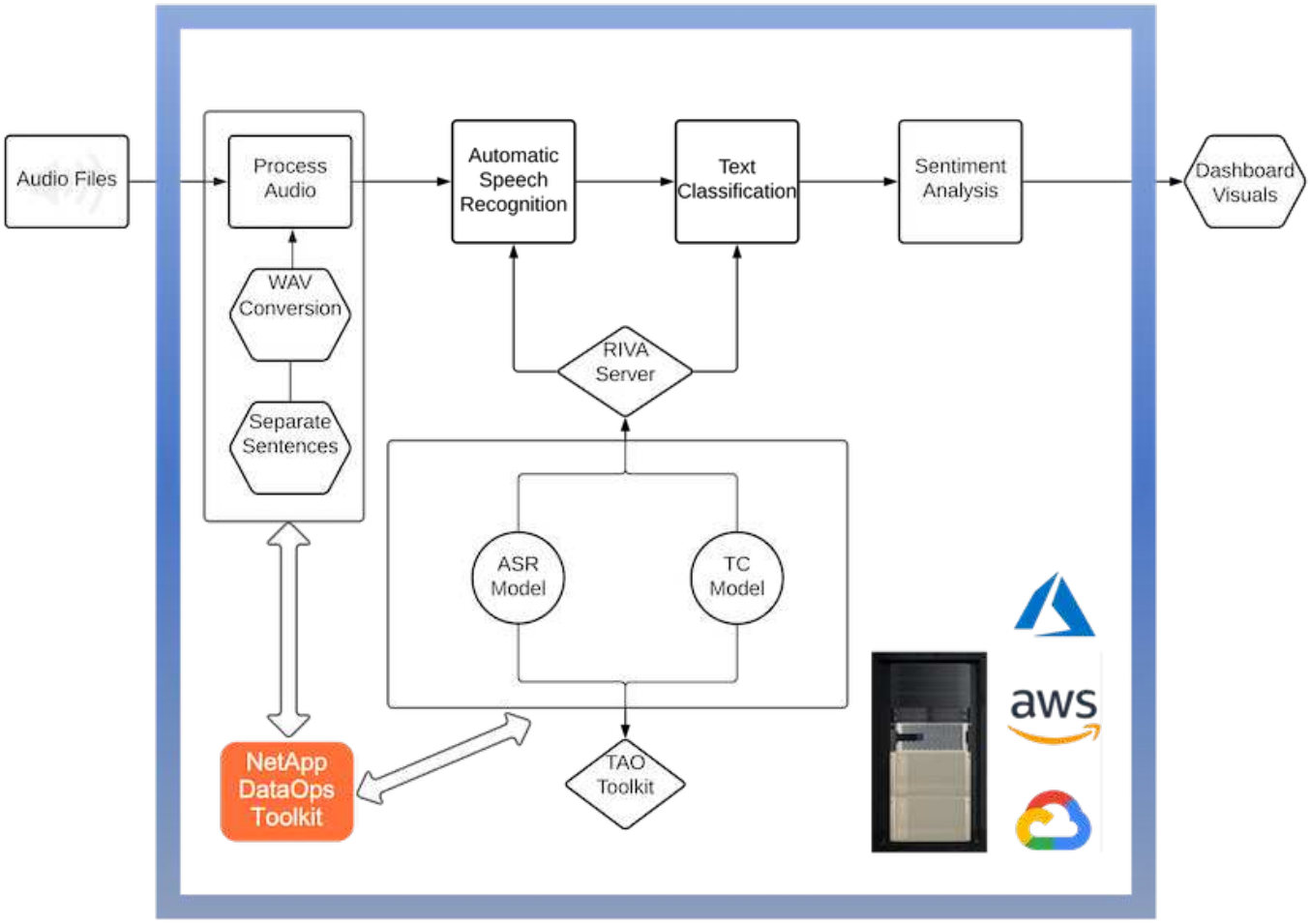
Software	Requisiti
Ubuntu 18.04 LTS	18.04
python	>=3.6.9
docker-ce	>19.03.5
API docker	1.40
nvidia-container-toolkit	>1.3.0-1
nvidia-container-runtime	3.4.0-1

Software	Requisiti
nvidia-docker2	2.5.0-1
driver nvidia	>455
python-pip	>21.06
nvidia-pyindex	Ultima versione

Utilizza i dettagli del caso

Questa soluzione si applica ai seguenti casi di utilizzo:

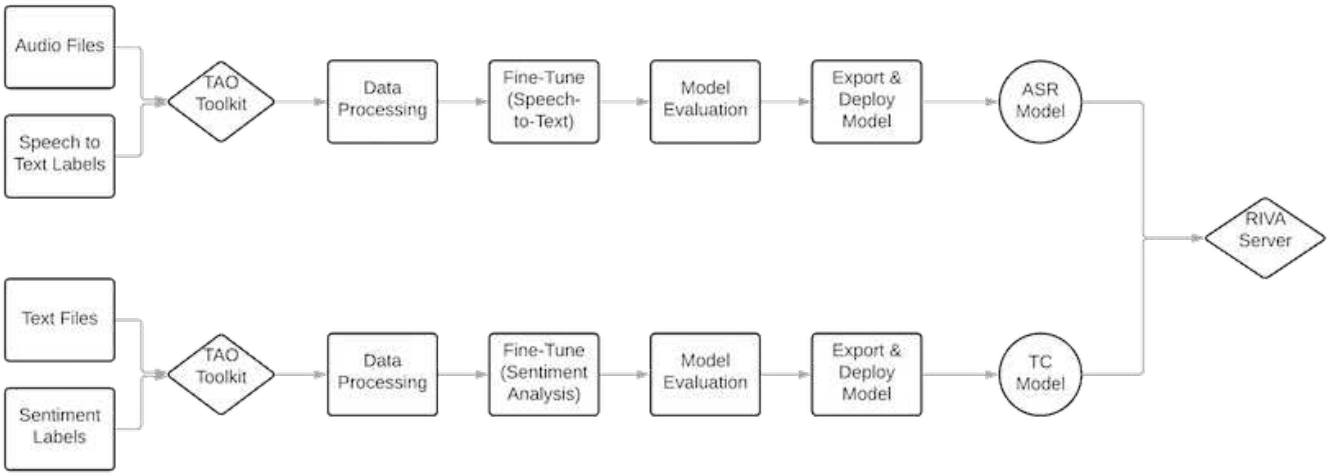
- Voce-testo
- Analisi del sentimento



Il caso d'utilizzo del parlato-to-text inizia con l'acquisizione di file audio per i centri di supporto. Questo audio viene quindi elaborato per adattarsi alla struttura richiesta DA RIVA. Se i file audio non sono già stati suddivisi nelle unità di analisi, è necessario eseguire questa operazione prima di passare l'audio a RIVA. Una volta elaborato, il file audio viene trasmesso al server RIVA come chiamata API. Il server utilizza uno dei numerosi modelli che ospita e restituisce una risposta. Questa voce-testo (parte del riconoscimento vocale automatico) restituisce una rappresentazione testuale dell'audio. Da qui, la pipeline passa alla parte di analisi del sentimento.

Per l'analisi del sentimento, l'output di testo del riconoscimento vocale automatico funge da input per la

classificazione del testo. Text Classification è il componente NVIDIA per la classificazione del testo in un numero qualsiasi di categorie. Le categorie di sentimento variano da positivo a negativo per le conversazioni del centro di supporto. Le performance dei modelli possono essere valutate utilizzando un set di holdout per determinare il successo della fase di fine tuning.



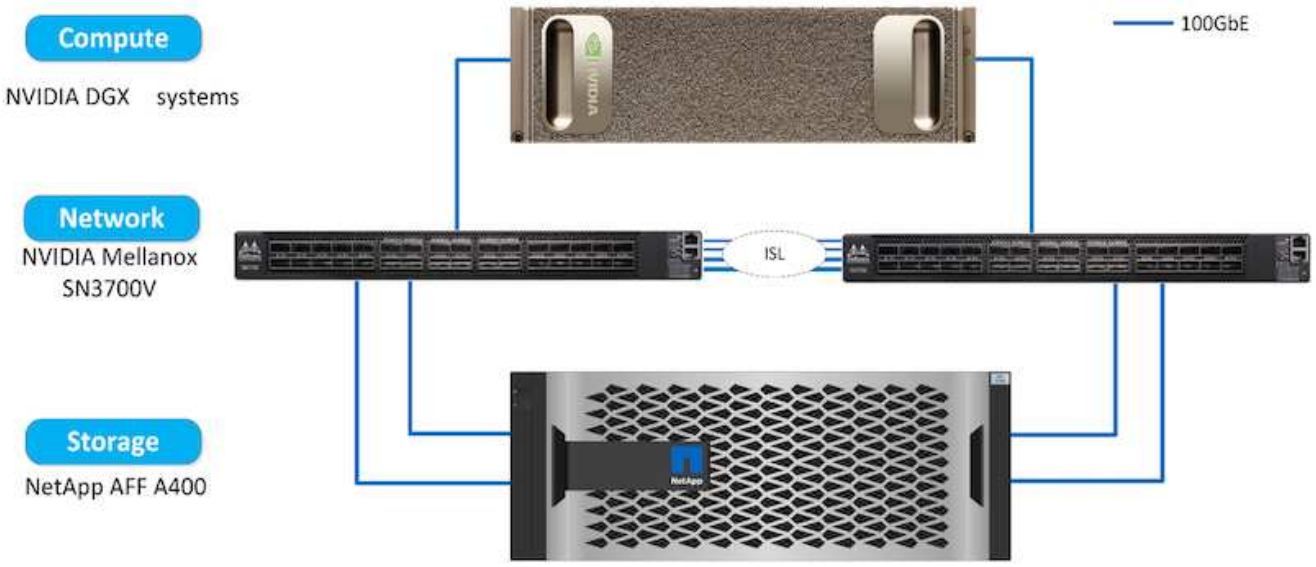
Una pipeline simile viene utilizzata sia per l'analisi del parlato-to-text che per l'analisi del sentimento all'interno del toolkit TAO. La differenza principale è l'utilizzo di etichette necessarie per la messa a punto dei modelli. La pipeline TAO Toolkit inizia con l'elaborazione dei file di dati. Poi i modelli preformati (provenienti da "Catalogo NGC NVIDIA") vengono perfezionati utilizzando i dati del centro di supporto. I modelli perfezionati vengono valutati in base alle metriche di performance corrispondenti e, se sono più performanti dei modelli preformati, vengono implementati sul server RIVA.

Considerazioni di progettazione

In questa sezione vengono descritte le considerazioni di progettazione relative ai diversi componenti di questa soluzione.

Progettazione di rete e calcolo

A seconda delle restrizioni sulla sicurezza dei dati, tutti i dati devono rimanere all'interno dell'infrastruttura del cliente o in un ambiente sicuro.



Progettazione dello storage

Il NetApp DataOps Toolkit funge da servizio principale per la gestione dei sistemi storage. DataOps Toolkit è una libreria Python che consente a sviluppatori, data scientist, ingegneri DevOps e data engineer di eseguire diverse attività di gestione dei dati, come il provisioning quasi istantaneo di un nuovo volume di dati o di un'area di lavoro JupyterLab, la clonazione quasi istantanea di un volume di dati o di un'area di lavoro JupyterLab, E lo snap-shoting quasi istantaneo di un volume di dati o di uno spazio di lavoro JupyterLab per la tracciabilità o il baselining. Questa libreria Python può funzionare come un'utility a riga di comando o una libreria di funzioni che possono essere importate in qualsiasi programma Python o Jupyter notebook.

Best practice RIVA

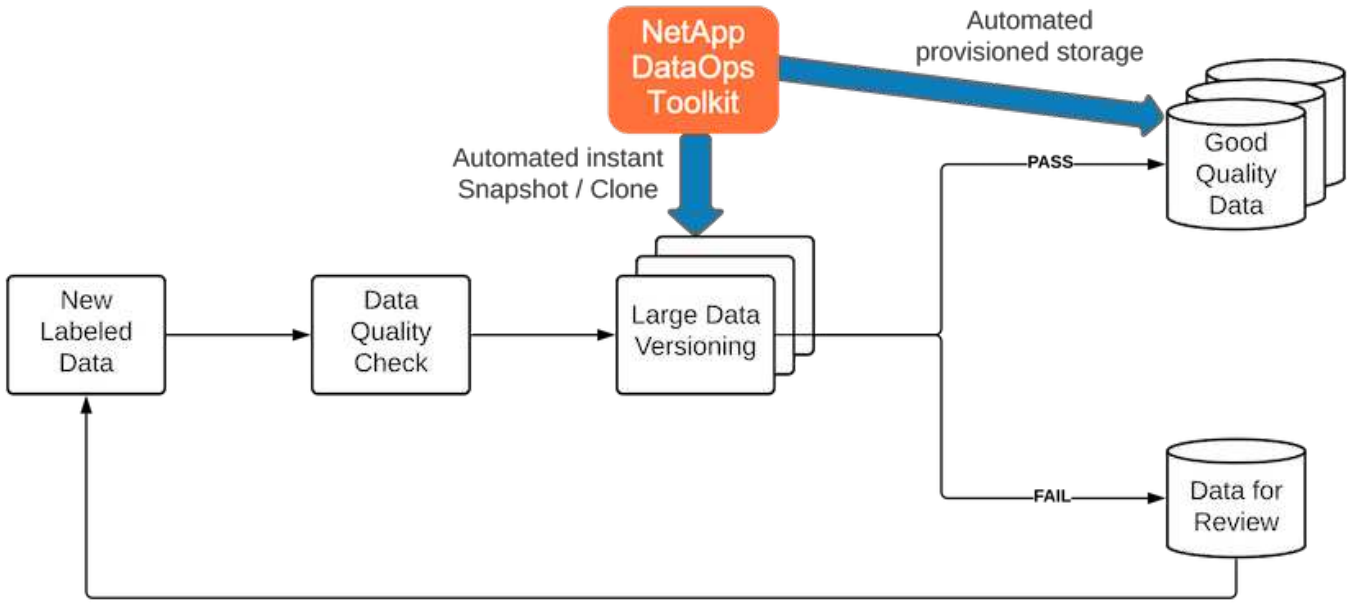
NVIDIA offre diverse funzionalità generali "best practice per i dati" Per utilizzare RIVA:

- **Se possibile, utilizzare formati audio senza perdita di dati.** l'utilizzo di codec con perdita di dati come MP3 può ridurre la qualità.
- **Aumentare i dati di training.** l'aggiunta di rumore di fondo ai dati di training audio può inizialmente ridurre la precisione e aumentare la robustezza.
- **Limitare la dimensione del vocabolario se si utilizza il testo scartato.** molte fonti online contengono messaggi o voci accessorie e parole non comuni. La rimozione di questi elementi può migliorare il modello linguistico.
- **Se possibile, utilizzare una frequenza di campionamento minima di 16 kHz.** tuttavia, provare a non ricampionare, perché in questo modo si riduce la qualità audio.

Oltre a queste Best practice, i clienti devono dare la priorità alla raccolta di un set di dati campione rappresentativo con etichette accurate per ogni fase della pipeline. In altre parole, il set di dati di esempio dovrebbe riflettere in modo proporzionale le caratteristiche specificate esemplificate in un set di dati di destinazione. Allo stesso modo, gli annotatori dei set di dati hanno la responsabilità di bilanciare la precisione e la velocità dell'etichettatura in modo da massimizzare la qualità e la quantità dei dati. Ad esempio, questa soluzione di Support Center richiede file audio, testo etichettato ed etichette di sentimento. La natura sequenziale di questa soluzione significa che gli errori dall'inizio della pipeline vengono propagati fino alla fine Se i file audio sono di scarsa qualità, anche le trascrizioni di testo e le etichette di sentimento saranno.

Questa propagazione degli errori si applica allo stesso modo ai modelli addestrati su questi dati. Se le

previsioni del sentimento sono accurate al 100% ma il modello da voce a testo non funziona correttamente, la pipeline finale è limitata dalle trascrizioni audio-testo iniziali. È essenziale che gli sviluppatori considerino le performance di ciascun modello singolarmente e come un componente di una pipeline più ampia. In questo caso specifico, l'obiettivo finale è sviluppare una pipeline in grado di prevedere con precisione il sentimento. Pertanto, la metrica generale su cui valutare la pipeline è la precisione dei sentimenti, che la trascrizione vocale-testuale influisce direttamente.



Il NetApp DataOps Toolkit integra la pipeline per il controllo della qualità dei dati attraverso l'utilizzo della sua tecnologia di cloning dei dati quasi istantanea. Ogni file etichettato deve essere valutato e confrontato con i file etichettati esistenti. La distribuzione di questi controlli di qualità nei vari sistemi di storage dei dati garantisce che questi controlli vengano eseguiti in modo rapido ed efficiente.

Implementazione dell'analisi del sentimento del centro di supporto

L'implementazione della soluzione comporta i seguenti componenti:

1. NetApp DataOps Toolkit
2. Configurazione NGC
3. Server NVIDIA RIVA
4. Toolkit NVIDIA TAO
5. Esportare i modelli TAO in RIVA

Per eseguire l'implementazione, attenersi alla seguente procedura:

NetApp DataOps Toolkit: Analisi del sentimento del centro di supporto

Per utilizzare "NetApp DataOps Toolkit", completare la seguente procedura:

1. PIP installare il toolkit.

```
python3 -m pip install netapp-dataops-traditional
```

2. Configurare la gestione dei dati

```
netapp_dataops_cli.py config
```

Configurazione NGC: Analisi del sentimento del centro di supporto

Per configurare "NVIDIA NGC", completare la seguente procedura:

1. Scarica NGC.

```
wget -O ngccli_linux.zip
https://ngc.nvidia.com/downloads/ngccli_linux.zip && unzip -o
ngccli_linux.zip && chmod u+x ngc
```

2. Aggiungere la directory corrente al percorso.

```
echo "export PATH=\"\$PATH:$(pwd)\"" >> ~/.bash_profile && source
~/.bash_profile
```

3. È necessario configurare l'interfaccia CLI NGC per l'utilizzo in modo da poter eseguire i comandi. Immettere il seguente comando, inclusa la chiave API quando richiesto.

```
ngc config set
```

Per i sistemi operativi che non sono basati su Linux, visitare il sito "qui".

Server NVIDIA RIVA: Analisi del sentimento del centro di supporto

Per configurare "NVIDIA RIVA", completare la seguente procedura:

1. Scarica I file RIVA da NGC.

```
ngc registry resource download-version
nvidia/riva/riva_quickstart:1.4.0-beta
```

2. Inizializzare la configurazione DI RIVA (`riva_init.sh`).
3. Avviare IL server RIVA (`riva_start.sh`).
4. Avviare IL client RIVA (`riva_start_client.sh`).

5. All'interno del client RIVA, installare la libreria di elaborazione audio ("FFMPEG")

```
apt-get install ffmpeg
```

6. Avviare "Jupyter" server.
7. Eseguire il notebook RIVA Inference Pipeline.

NVIDIA TAO Toolkit: Analisi del sentimento del centro di supporto

Per configurare NVIDIA TAO Toolkit, attenersi alla seguente procedura:

1. Preparare e attivare un "ambiente virtuale" Per TAO Toolkit.
2. Installare "pacchetti richiesti".
3. Estrarre manualmente l'immagine utilizzata durante l'addestramento e la messa a punto.

```
docker pull nvcr.io/nvidia/tao/tao-toolkit-pyt:v3.21.08-py3
```

4. Avviare "Jupyter" server.
5. Eseguire il notebook TAO fine-Tuning.

Esportare i modelli TAO in RIVA: Analisi del sentimento del centro di supporto

Da utilizzare "Modelli TAO Toolkit di RIVA", completare la seguente procedura:

1. Salva i modelli nel notebook TAO fine-Tuning.
2. Copiare i modelli addestrati TAO nella directory dei modelli RIVA.
3. Avviare IL server RIVA (`riva_start.sh`).

Blocchi stradali per l'implementazione

Ecco alcuni elementi da tenere a mente durante lo sviluppo della soluzione:

- Il NetApp DataOps Toolkit viene installato per primo per garantire il funzionamento ottimale del sistema di storage dei dati.
- NVIDIA NGC deve essere installata prima di ogni altra cosa perché autentica il download di immagini e modelli.
- RIVA deve essere installato prima del toolkit TAO. L'installazione DI RIVA configura il daemon del docker per estrarre le immagini in base alle necessità.
- DGX e Docker devono disporre dell'accesso a Internet per scaricare i modelli.

Risultati della convalida

Come indicato nella sezione precedente, gli errori vengono propagati in tutta la pipeline ogni volta che vi sono due o più modelli di apprendimento automatico in esecuzione in sequenza. Per questa soluzione, il sentimento della frase è il fattore più importante nella misurazione del livello di rischio azionario dell'azienda. Il modello da voce a testo,

sebbene essenziale per la pipeline, funge da unità di pre-elaborazione prima che i sentimenti possano essere previsti. Ciò che conta realmente è la differenza di sentimento tra le frasi di verità e le frasi previste. Questo serve come proxy per il tasso di errore di parola (WER). La precisione del parlato-to-text è importante, ma il WER non viene utilizzato direttamente nella metrica finale della pipeline.

```
PIPELINE_SENTIMENT_METRIC = MEAN(DIFF(GT_sentiment, ASR_sentiment))
```

Queste metriche di sentimento possono essere calcolate per il punteggio F1, il richiamo e la precisione di ciascuna frase. I risultati possono quindi essere aggregati e visualizzati all'interno di una matrice di confusione, insieme agli intervalli di confidenza per ciascuna metrica.

Il vantaggio dell'utilizzo del transfer learning è un aumento delle performance del modello per una frazione dei requisiti dei dati, dei tempi di formazione e dei costi. I modelli perfezionati devono anche essere confrontati con le versioni di riferimento per garantire che l'apprendimento del trasferimento migliori le performance invece di comprometterle. In altre parole, il modello ottimizzato dovrebbe funzionare meglio sui dati del centro di supporto rispetto al modello preaddestrato.

Valutazione della pipeline

Caso di test	Dettagli
Numero del test	Metrica del sentimento della pipeline
Prerequisiti del test	Modelli ottimizzati per modelli di analisi del parlato-to-text e del sentimento
Risultato previsto	La metrica del sentimento del modello ottimizzato offre prestazioni migliori rispetto al modello originale preaddestrato.

Metrica del sentimento della pipeline

1. Calcola la metrica del sentimento per il modello di riferimento.
2. Calcola la metrica del sentimento per il modello ottimizzato.
3. Calcola la differenza tra queste metriche.
4. Calcolare la media delle differenze tra tutte le frasi.

Video e demo

Esistono due notebook che contengono la pipeline di analisi del sentimento: ""Support-Center-Model-Transfer-Learning-and-fine-Tuning.ipynb"" e. ""Support-Center-Sentiment-Analysis-pipeline.ipynb"". Insieme, questi notebook dimostrano come sviluppare una pipeline per acquisire i dati del centro di supporto ed estrarre sentimenti da ogni frase utilizzando modelli di deep learning all'avanguardia e ottimizzati sui dati dell'utente.

Support Center - Sentiment Analysis Pipeline.ipynb

Questo notebook contiene la pipeline Inference RIVA per l'acquisizione di audio, la conversione in testo e l'estrazione di sentimenti da utilizzare in una dashboard esterna. I set di dati vengono scaricati ed elaborati

automaticamente, se non è già stato fatto. La prima sezione del notebook è Speech-to-Text, che gestisce la conversione dei file audio in testo. Segue la sezione analisi del sentimento che estrae i sentimenti per ciascuna frase di testo e visualizza i risultati in un formato simile alla dashboard proposta.

> Questo notebook deve essere eseguito prima del training e della messa a punto del modello, in quanto il set di dati MP3 deve essere scaricato e convertito nel formato corretto.

Call Center - Sentiment Analysis Pipeline

This notebook demonstrates how to build a pipeline for sentiment analysis of call center conversations. The goal of this pipeline is to develop sentiment analysis for use within an external dashboard.

This tutorial will guide you through the use of NVIDIA's RIVA for automatic speech recognition and text classification. This tutorial uses NetApp cloud storage for data storage and a pre-trained RIVA model.

Channels

These are the channels on which RIVA is hosting models.

- speech: `51051`
- voice: `61051`

These channels **must** be aligned with `riva_speech_api_port` and `riva_vision_api_port` within `config.sh`

In [4]:
```
speech_channel = "localhost:51051"
voice_channel = "localhost:61051"
```

Speech-To-Text

Automatic Speech Recognition (ASR) takes as input an audio stream or audio buffer and returns one or more text transcripts, along with additional optional metadata. ASR represents a full speech recognition pipeline that is GPU accelerated with optimized performance and accuracy. ASR supports synchronous and streaming recognition modes.

For more information on NVIDIA RIVA's Automatic Speech Recognition, visit here.

Constants

Use these constants to affect different aspects of this pipeline:

- `DATA_DIR` : base folder where data is stored
- `DATASET_NAME` : name of the call center dataset
- `COMPANY_DATE` : folder name identifying the particular call center conversation

Support Center - Model Training e fine-Tuning.ipynb

L'ambiente virtuale TAO Toolkit deve essere configurato prima di eseguire il notebook (per istruzioni sull'installazione, consultare la sezione relativa al toolkit TAO nella Panoramica dei comandi).

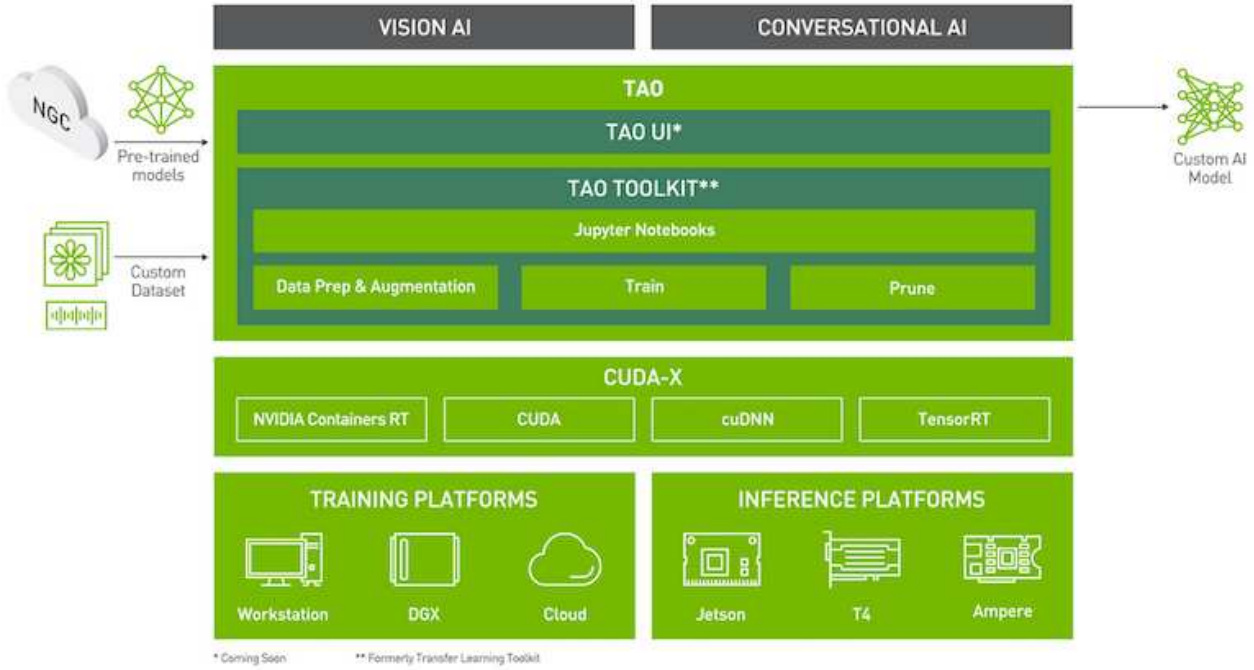
Questo notebook si affida al toolkit TAO per mettere a punto modelli di apprendimento approfondito sui dati dei clienti. Come per il notebook precedente, questo è separato in due sezioni per i componenti Speech-to-Text e analisi del sentimento. Ogni sezione passa attraverso l'elaborazione dei dati, la formazione sui modelli e la messa a punto, la valutazione dei risultati e l'esportazione dei modelli. Infine, è disponibile una sezione finale per l'implementazione di entrambi i modelli ottimizzati per L'utilizzo in RIVA.

Call Center - Model Transfer Learning and Fine-Tuning

TAO Toolkit is a python based AI toolkit for taking purpose-built pre-trained AI models and customizing them with your own data. Transfer learning extracts learned features from an existing neural network to a new one. Transfer learning is often used when creating a large training dataset is not feasible in order to enhance the base performance of state-of-the-art models.

For this call center solution, the speech-to-text and sentiment analysis models are fine-tuned on call center data to augment the model performance on business specific terminology.

For more information on the TAO Toolkit, please visit here.



Installing necessary dependencies

For ease of use, please install TAO Toolkit inside a python virtual environment. We recommend performing this step first and then launching the notebook from the virtual environment. Please refer to the README for these instructions.

Conclusione

Poiché l'esperienza dei clienti è diventata sempre più considerata come un terreno di battaglia competitivo fondamentale, un centro di supporto globale con intelligenza artificiale diventa un componente critico che le aziende di quasi tutti i settori non possono permettersi di trascurare. La soluzione proposta in questo report tecnico è stata dimostrata per supportare l'offerta di esperienze clienti eccezionali di questo tipo e la sfida è ora quella di garantire che le aziende stiano adottando misure per modernizzare l'infrastruttura e i flussi di lavoro dell'ai.

Le migliori implementazioni dell'ai nel servizio clienti non devono sostituire gli agenti umani. Piuttosto, l'ai può consentire loro di creare esperienze eccezionali con i clienti attraverso analisi del sentimento in tempo reale, escalation delle controversie e calcolo affettivo multimodale per rilevare indizi verbali, non verbali e facciali con

i quali modelli ai completi possono fornire consigli su larga scala e integrare ciò che un singolo agente umano potrebbe non essere presente. L'ai può anche fornire una migliore corrispondenza tra un particolare cliente e gli agenti attualmente disponibili. Utilizzando l'ai, le aziende possono estrarre il prezioso sentimento dei clienti riguardo alle loro opinioni e impressioni sui prodotti, i servizi e l'immagine del marchio del provider.

La soluzione può anche essere utilizzata per costruire dati Time-Series per gli agenti di supporto che fungono da metriche di valutazione obiettiva delle performance. I sondaggi convenzionali sulla soddisfazione dei clienti spesso non hanno risposte sufficienti. Raccogliendo il sentimento a lungo termine dei dipendenti e dei clienti, i datori di lavoro possono prendere decisioni informate in merito alle performance degli agenti di supporto.

La combinazione di NetApp, SFL Scientific, framework di orchestrazione open-source e NVIDIA riunisce le più recenti tecnologie come servizi gestiti con grande flessibilità per accelerare l'adozione della tecnologia e migliorare il time-to-market per le nuove applicazioni ai/ML. Questi servizi avanzati vengono forniti on-premise e possono essere facilmente trasferiti per ambienti cloud-native e architetture di implementazione ibride.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i seguenti documenti e/o siti Web:

- Demo interattive 3D

  "www.netapp.com/ai"

- Connettiti direttamente con uno specialista ai di NetApp

  "https://www.netapp.com/artificial-intelligence/"

- NVDIA base Command Platform with NetApp Solution Brief

  https://www.netapp.com/pdf.html?item=/media/32792-DS-4145-NVIDIA-Base-Command-Platform-with-NetApp.pdf

- Infografica sulle buone ragioni di NetApp per ai 10

  "https://www.netapp.com/us/media/netapp-ai-10-good-reasons.pdf"

- Ai in Healthcare: White paper sull'apprendimento approfondito per identificare le lesioni COVID-19 nelle scansioni TC polmonari

  https://www.netapp.com/pdf.html?item=/media/31240-WP-7342.pdf

- Ai in Healthcare: White paper sul monitoraggio dell'utilizzo della maschera facciale in ambito sanitario

  https://www.netapp.com/pdf.html?item=/media/37490-NA-611-Monitoring-face-mask-usage-in-healthcare-settings.pdf

- Ai in Healthcare: Report tecnico di imaging diagnostico

  https://www.netapp.com/pdf.html?item=/media/7395-tr4811.pdf

- Ai per il retail: Ai di NetApp Conversational con NVIDIA RIVA

  "In sintesi"

- Analisi della soluzione ai di NetApp ONTAP

  https://www.netapp.com/pdf.html?item=/media/6736-sb-3939.pdf

- Descrizione della soluzione NetApp DataOps Toolkit

  https://www.netapp.com/pdf.html?item=/media/21480-SB-4111-1220-NA-Data-Science-Toolkit.pdf

- Analisi della soluzione NetApp ai Control Plane

  https://www.netapp.com/pdf.html?item=/media/6737-sb-4055.pdf

- EBook trasformare il settore con Data Drive ai

  "https://www.netapp.com/us/media/na-337.pdf"

- Analisi della soluzione ai NetApp EF-Series

  https://www.netapp.com/pdf.html?item=/media/26708-SB-4136-NetApp-AI-E-Series.pdf

- Analisi della soluzione NetApp ai e Lenovo ThinkSystem for ai Inferencing

  https://www.netapp.com/pdf.html?item=/media/25316-SB-4129.pdf

- Analisi della soluzione NetApp ai e Lenovo ThinkSystem per ai e ML Enterprise

  https://www.netapp.com/pdf.html?item=/media/25317-SB-4128.pdf

- NetApp e NVIDIA – ridefinire le possibilità con i video ai

  https://www.youtube.com/watch?v=38xw65SteUc

Formazione distribuita in Azure - previsione dei tassi click-through

TR-4904: Formazione distribuita in Azure - previsione dei tassi click-through

Rick Huang, Verron Martina, Muneer Ahmad, NetApp

Il lavoro di un data scientist dovrebbe essere incentrato sulla formazione e sulla messa a punto di modelli di apprendimento automatico (ML) e intelligenza artificiale (ai). Tuttavia, secondo una ricerca condotta da Google, i data scientist dedicano circa il 80% del loro tempo a capire come far funzionare i propri modelli con le applicazioni aziendali e a eseguirlo su larga scala.

Per gestire i progetti ai/ML end-to-end, è necessaria una maggiore comprensione dei componenti aziendali. Sebbene DevOps abbia assunto il controllo della definizione, dell'integrazione e dell'implementazione, questi tipi di componenti, le operazioni ML hanno come obiettivo un flusso simile che include i progetti ai/ML. Per avere un'idea di ciò che una pipeline ai/ML end-to-end tocca nell'azienda, consulta il seguente elenco di componenti richiesti:

- Storage

- Networking
- Database
- File system
- Container
- Pipeline ci/CD (Continuous Integration and Continuous Deployment)
- Ambiente di sviluppo integrato (IDE)
- Sicurezza
- Policy di accesso ai dati
- Hardware
- Cloud
- Virtualizzazione
- Set di strumenti e librerie per le scienze dei dati

Pubblico di riferimento

Il mondo della scienza dei dati tocca diverse discipline nell'IT e nel business:

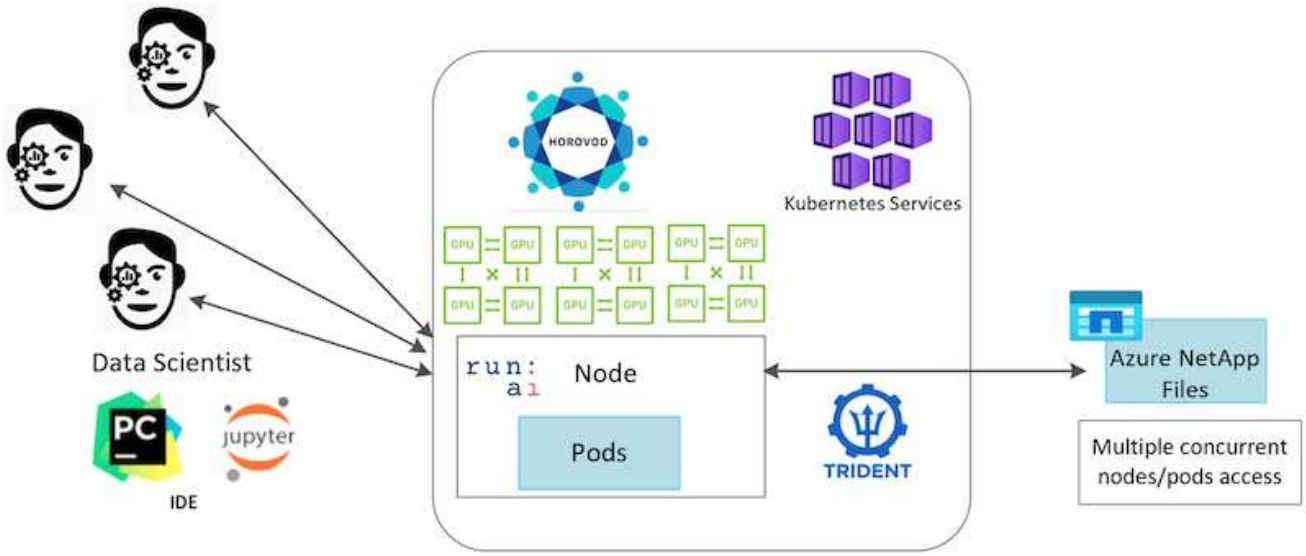
- Il data scientist ha bisogno della flessibilità necessaria per utilizzare i propri strumenti e le librerie preferite.
- Il data engineer deve sapere come i dati scorrono e dove risiedono.
- Un tecnico DevOps ha bisogno dei tool per integrare le nuove applicazioni ai/ML nelle pipeline ci/CD.
- Gli amministratori e gli architetti del cloud devono essere in grado di configurare e gestire le risorse di Azure.
- Gli utenti aziendali desiderano avere accesso alle applicazioni ai/ML.

In questo report tecnico, descriviamo in che modo Azure NetApp Files, RAPIDS ai, DAK e Azure aiutano ciascuno di questi ruoli a portare valore al business.

Panoramica della soluzione

Questa soluzione segue il ciclo di vita di un'applicazione ai/ML. Iniziamo con il lavoro dei data scientist per definire le diverse fasi necessarie per preparare i dati e formare i modelli. Sfruttando RAPIDS su Dask, eseguiamo training distribuiti nel cluster Azure Kubernetes Service (AKS) per ridurre drasticamente i tempi di training rispetto all'approccio convenzionale di Python scikit-Learn. Per completare il ciclo completo, integriamo la pipeline con Azure NetApp Files.

Azure NetApp Files offre diversi livelli di performance. I clienti possono iniziare con un Tier Standard, scalare e scalare fino a un Tier dalle performance elevate senza interruzioni, senza spostare alcun dato. Questa funzionalità consente agli scienziati dei dati di formare modelli su larga scala senza problemi di performance, evitando i silos di dati nel cluster, come mostrato nella figura seguente.



Panoramica della tecnologia

Questa pagina fornisce una panoramica della tecnologia utilizzata in questa soluzione.

Microsoft e NetApp

Da maggio 2019, Microsoft ha fornito un servizio di portale nativo e di prima parte Azure per file service NFS e SMB aziendali basati sulla tecnologia NetApp ONTAP. Questo sviluppo è guidato da una partnership strategica tra Microsoft e NetApp e estende ulteriormente la portata dei servizi dati ONTAP di livello mondiale ad Azure.

Azure NetApp Files

Il servizio Azure NetApp Files è un servizio di storage di file di livello Enterprise, dalle performance elevate e misurato. Azure NetApp Files supporta qualsiasi tipo di carico di lavoro ed è altamente disponibile per impostazione predefinita. È possibile selezionare i livelli di servizio e di performance e impostare le copie Snapshot tramite il servizio. Azure NetApp Files è un servizio Azure first-party per la migrazione e l'esecuzione dei carichi di lavoro dei file aziendali più esigenti nel cloud, inclusi database, SAP e applicazioni di calcolo ad alte performance senza modifiche del codice.

Questa architettura di riferimento offre alle organizzazioni IT i seguenti vantaggi:

- Elimina le complessità di progettazione
- Consente una scalabilità indipendente di calcolo e storage
- Consente ai clienti di partire da piccoli e scalare perfettamente
- Offre una gamma di Tier di storage per diversi punti di costo e performance

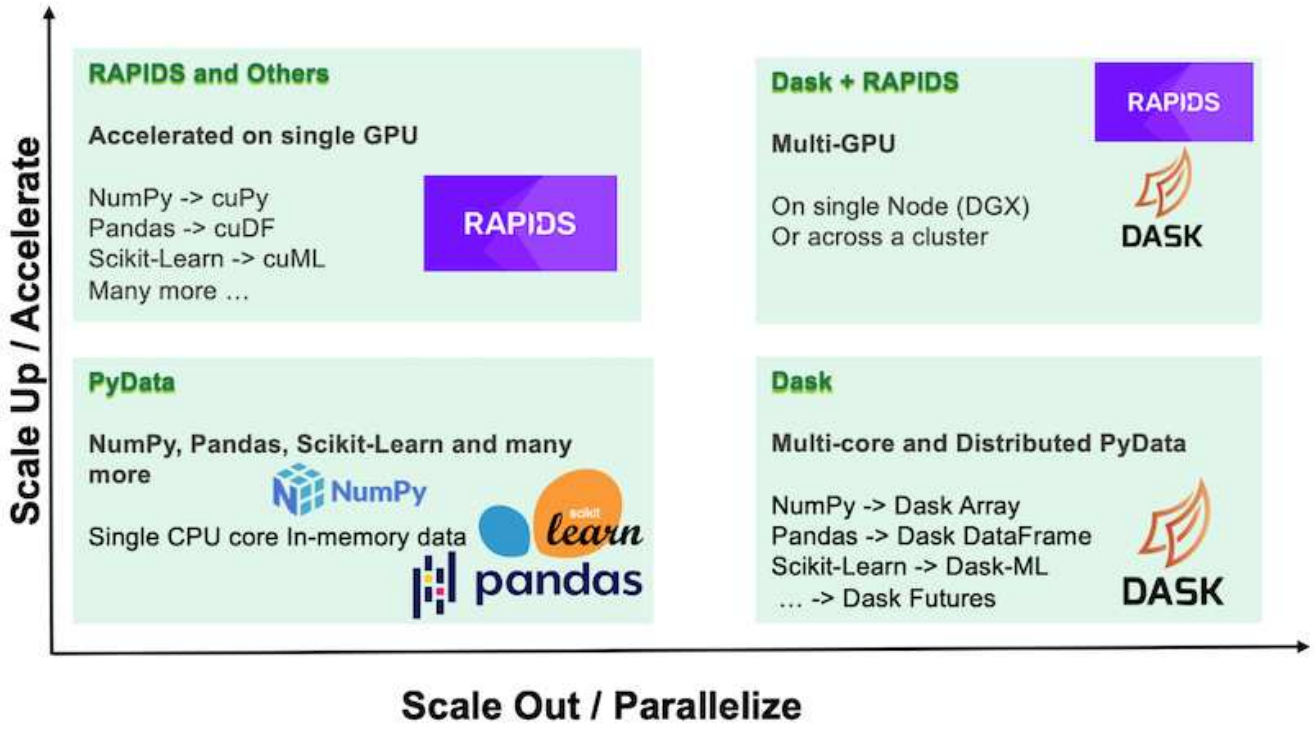
Panoramica di DAK e NVIDIA RAPIDS

Dask è un tool di calcolo parallelo open-source che scala le librerie Python su più macchine e fornisce un'elaborazione più rapida di grandi quantità di dati. Fornisce un'API simile alle librerie Python convenzionali a thread singolo, come Pandas, Numpy e scikit-Learn. Di conseguenza, gli utenti nativi di Python non sono costretti a modificare molto nel codice esistente per utilizzare le risorse all'interno del cluster.

NVIDIA RAPIDS è una suite di librerie open-source che consente di eseguire FLUSSI DI lavoro END-to-end DI

ANALISI DEI dati E ML interamente su GPU. Insieme a Dask, consente di scalare facilmente da workstation GPU (scale-up) a cluster multi-GPU a più nodi (scale-out).

Per l'implementazione di DAK su un cluster, è possibile utilizzare Kubernetes per l'orchestrazione delle risorse. È inoltre possibile scalare in verticale o in orizzontale i nodi di lavoro in base ai requisiti di processo, che a loro volta possono aiutare a ottimizzare il consumo delle risorse del cluster, come illustrato nella figura seguente.



Requisiti software

La seguente tabella elenca i requisiti software necessari per questa soluzione.

Software	Versione
Servizio Azure Kubernetes	1.18.14
IMMAGINE del container RAPIDS e Dask	Repository: "Rapidsai/rapidsai" Tag: 0.17-cuda11.0-runtime-ubuntu18.04
Trident di NetApp	20.01.1
Timone	3.0.0

Requisiti relativi alle risorse cloud

Questa pagina descrive la configurazione delle risorse cloud per Azure NetApp Files.

Configurare Azure NetApp Files

Configurare Azure NetApp Files come descritto in "QuickStart: Configurazione di Azure NetApp Files e creazione di un volume NFS".

È possibile passare alla sezione "Crea volume NFS per Azure NetApp Files" perché si stanno creando volumi

tramite Trident. Prima di continuare, attenersi alla seguente procedura:

1. Registratevi per Azure NetApp Files e per il provider di risorse NetApp (tramite la shell Azure) ("collegamento").
2. Creare un account in Azure NetApp Files ("collegamento").
3. Configurare un pool di capacità (un minimo di 4 TB Standard o Premium, a seconda delle esigenze) ("collegamento").la seguente tabella elenca i requisiti di configurazione di rete per l'impostazione nel cloud. Il cluster e Azure NetApp Files devono trovarsi nella stessa rete virtuale Azure o in una rete VNET peered.

Risorse	Tipo/versione
Servizio Azure Kubernetes	1.18.14
Nodo dell'agente	3x Standard_DS2_v2
Nodo GPU	3x Standard_NC6s_v3
Azure NetApp Files	Pool di capacità standard
Capacità in TB	4

Riepilogo del caso d'uso con la previsione del tasso di click-through

Questo caso d'utilizzo si basa sui dati pubblicamente disponibili "Fare clic su Log di terabyte" dataset da "Criteo ai Lab". Con i recenti progressi nelle piattaforme E nelle applicazioni ML, è ora molto importante concentrarsi sull'apprendimento su larga scala. Il tasso di click-through (CTR) è definito come il numero medio di click-through per cento impressioni di annunci online (espresso in percentuale). È ampiamente adottato come parametro chiave in diversi mercati verticali e casi di utilizzo del settore, tra cui digital marketing, retail, e-commerce e service provider. Di seguito sono riportati alcuni esempi di utilizzo di CTR come metrica importante per il potenziale traffico dei clienti:

- **Digital marketing:** in "Google Analytics", CTR può essere usato per misurare come bene un advertiser o mercantile parole chiavi, annunci, ed elenchi liberi stanno eseguendo. Un CTR elevato è una buona indicazione che gli utenti trovano i tuoi annunci e gli elenchi utili e pertinenti. Il CTR contribuisce anche al CTR previsto dalla parola chiave, che è un componente di "Classifica ad".
- **E-commerce:** oltre a sfruttare "Google Analytics", ci sono almeno alcune statistiche dei visitatori in un backend di e-commerce. Anche se queste statistiche potrebbero non sembrare utili a prima vista, in genere sono facili da leggere e potrebbero essere più accurate di altre informazioni. I set di dati di prima parte composti da tali statistiche sono proprietari e sono quindi i più rilevanti per i venditori, gli acquirenti e le piattaforme di e-commerce. Questi set di dati possono essere utilizzati per impostare benchmark, confrontando i risultati con l'anno scorso e ieri, creando una serie temporale per ulteriori analisi.
- **Retail:** i retailer Brick-and-mortar possono correlare il numero di visitatori e il numero di clienti al CTR. Il numero di clienti può essere visto dalla loro storia del punto vendita. Il CTR proveniente dai siti web dei rivenditori o dal traffico pubblicitario potrebbe comportare le vendite di cui sopra. I programmi fedeltà sono un altro caso d'utilizzo, perché i clienti reindirizzati dagli annunci online o da altri siti Web potrebbero unirsi per guadagnare premi. I retailer possono acquisire i clienti attraverso programmi fedeltà e registrare i comportamenti dalle cronologie di vendita per creare un sistema di raccomandazione che non solo preveda i comportamenti di acquisto dei consumatori in diverse categorie, ma anche personalizza i coupon e riduce il tasso di abbandono.
- **Fornitori di servizi:** le aziende di telecomunicazioni e i provider di servizi Internet dispongono di numerosi dati telemetrici di prima parte per utenti che utilizzano in maniera approfondita ai, ML e analytics. Ad

esempio, una telecomunicazione può sfruttare ogni giorno i log di cronologia dei domini di primo livello per la navigazione sul Web dei propri abbonati mobili per mettere a punto i modelli esistenti e produrre una segmentazione aggiornata del pubblico, prevedere il comportamento dei clienti e collaborare con gli inserzionisti per inserire annunci in tempo reale per una migliore esperienza online. In questo workflow di marketing basato sui dati, il CTR è una metrica importante per riflettere le conversioni.

Nel contesto del digital marketing, "Registri Click di Criteo terabyte" Sono ora il set di dati di riferimento per la valutazione della scalabilità delle piattaforme E degli algoritmi ML. Prevedendo il tasso di click-through, un inserzionista può selezionare i visitatori che hanno più probabilità di rispondere agli annunci, analizzare la cronologia di navigazione e mostrare gli annunci più rilevanti in base agli interessi dell'utente.

La soluzione fornita in questo report tecnico evidenzia i seguenti vantaggi:

- I vantaggi di Azure NetApp Files nella formazione distribuita o su larga scala
- RAPIDA elaborazione dei dati abilitata per CUDA (cuDF, cuPy e così via) e algoritmi ML (cuML)
- Il framework di calcolo parallelo di Dask per la formazione distribuita

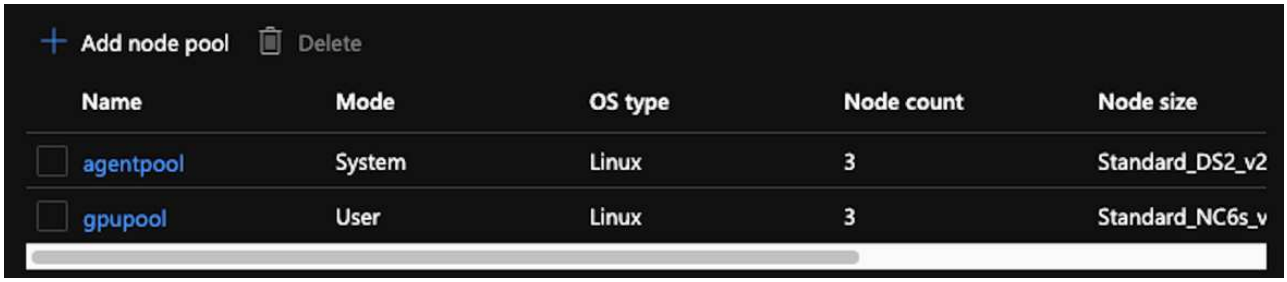
Un workflow end-to-end basato su RAPIDS ai e Azure NetApp Files dimostra il drastico miglioramento dei tempi di training dei modelli di foresta casuali di due ordini di grandezza. Questo miglioramento è significativo rispetto all'approccio Pandas convenzionale quando si tratta di log click reali con 45 GB di dati tabulari strutturati (in media) ogni giorno. Ciò equivale a un DataFrame contenente circa venti miliardi di righe. Illustreremo la configurazione dell'ambiente cluster, l'installazione del framework e della libreria, il caricamento e l'elaborazione dei dati, la formazione convenzionale e distribuita, la visualizzazione e il monitoraggio e il confronto dei risultati critici di runtime end-to-end in questo report tecnico.

Setup (Configurazione)

Installare e configurare il cluster AKS

Per installare e configurare il cluster AKS, consultare la pagina Web "Creare un cluster AKS" quindi completare i seguenti passaggi:

1. Quando si seleziona il tipo di nodo (nodi di sistema [CPU] o di lavoro [GPU]), selezionare quanto segue:
   a. I nodi di sistema primari devono essere DS2v2 standard (`agentpool` tre nodi predefiniti).
   b. Quindi, aggiungere il pool Standard_NC6s_v3 del nodo di lavoro (minimo tre nodi) per il gruppo di utenti (per i nodi GPU) denominato `gpupool`.



2. L'implementazione richiede da 5 a 10 minuti. Al termine, fare clic su Connect to Cluster (Connetti al cluster).
3. Per connettersi al cluster AKS appena creato, installare quanto segue dall'ambiente locale (laptop/pc):
   a. Lo strumento della riga di comando Kubernetes che utilizza "Istruzioni fornite per il sistema operativo in uso"

b. La CLI di Azure come descritto nel documento, "Installare Azure CLI"

4. Per accedere al cluster AKS dal terminale, immettere `az login` e immettere le credenziali.
5. Eseguire i due comandi seguenti:

```
az account set --subscription xxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxxx
aks get-credentials --resource-group resourcegroup --name aksclustername
```

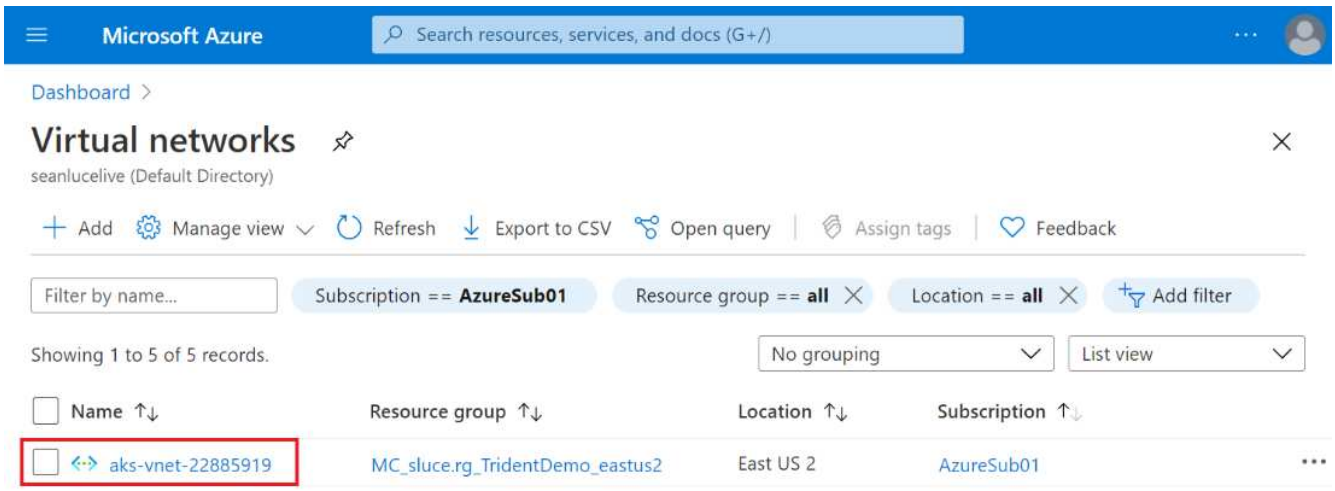
6. Invio `Azure CLI: kubectl get nodes`.
7. Se tutti e sei i nodi sono attivi e in esecuzione, come illustrato nell'esempio seguente, il cluster AKS è pronto e connesso all'ambiente locale

```
verronmartina@verron-mac-0 ~ % kubectl get nodes
NAME                                STATUS   ROLES   AGE   VERSION
aks-agentpool-34613062-vmss000000   Ready    agent   22m   v1.18.14
aks-agentpool-34613062-vmss000001   Ready    agent   22m   v1.18.14
aks-agentpool-34613062-vmss000002   Ready    agent   22m   v1.18.14
aks-gpupool-34613062-vmss000000     Ready    agent   20m   v1.18.14
aks-gpupool-34613062-vmss000001     Ready    agent   20m   v1.18.14
aks-gpupool-34613062-vmss000002     Ready    agent   20m   v1.18.14
verronmartina@verron-mac-0 ~ %
```

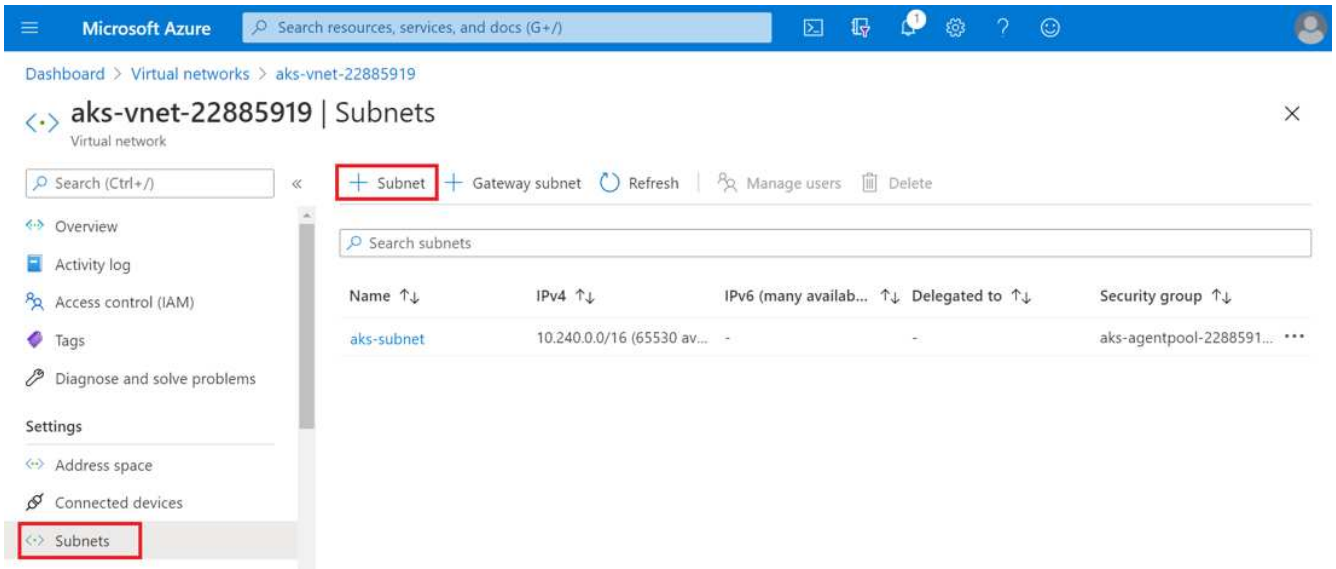
Creare una subnet delegata per Azure NetApp Files

Per creare una subnet delegata per Azure NetApp Files, attenersi alla seguente procedura:

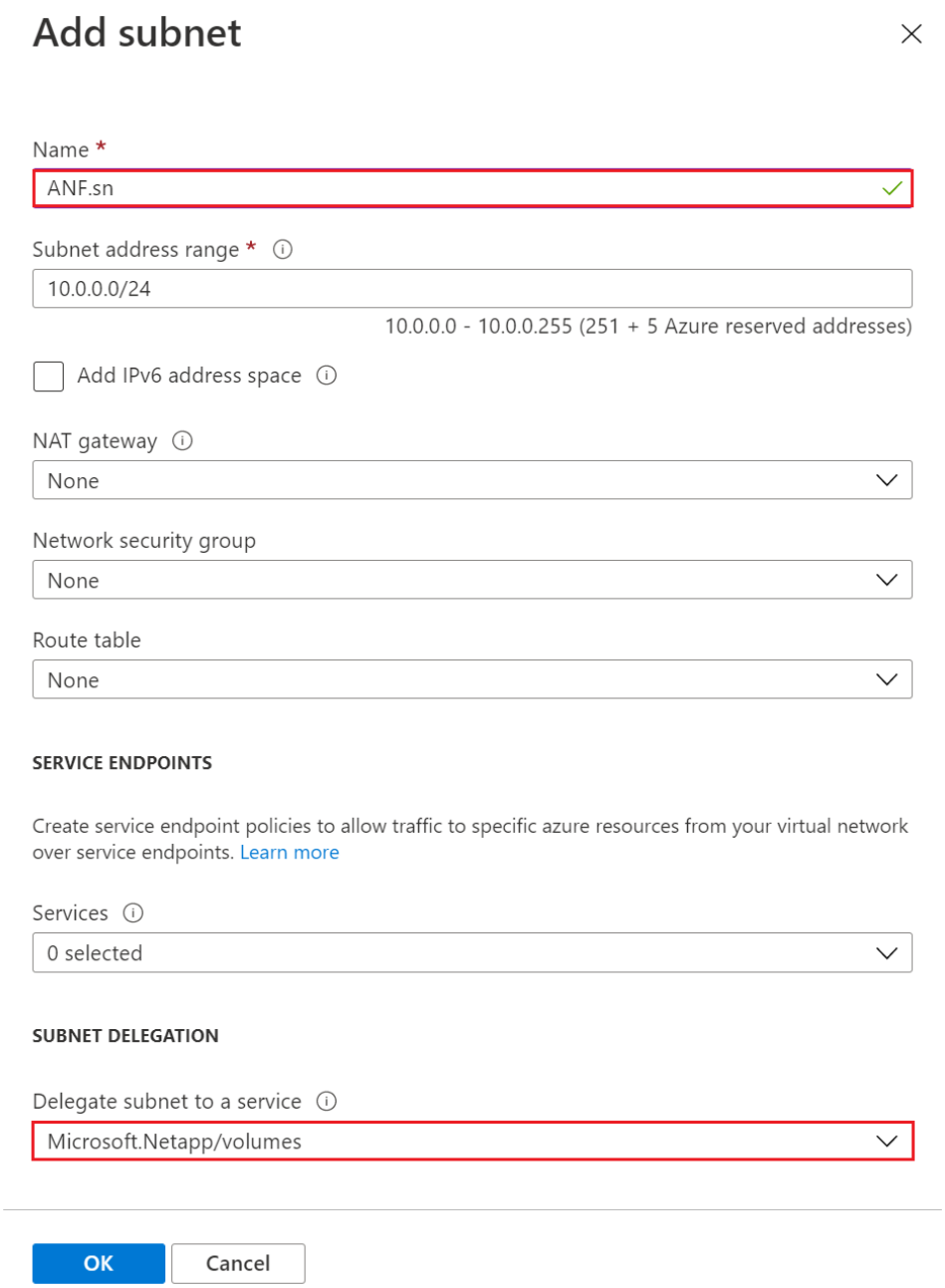
1. Accedere alle reti virtuali all'interno del portale Azure. Trova la tua rete virtuale appena creata. Dovrebbe avere un prefisso come `aks-vnet`.
2. Fare clic sul nome di VNET.



3. Fare clic su subnet e fare clic su +Subnet nella barra degli strumenti superiore.



4. Specificare un nome per la subnet, ad esempio `ANF.sn` Quindi, sotto delega subnet, selezionare `Microsoft.Netapp/volumes`. Non cambiare altro. Fare clic su OK.



I volumi Azure NetApp Files vengono allocati nel cluster di applicazioni e vengono utilizzati come dichiarazioni di volumi persistenti (PVC) in Kubernetes. A sua volta, questo processo offre la flessibilità di mapparli a diversi servizi, come i notebook Jupyter, le funzioni senza server e così via.

Gli utenti dei servizi possono consumare lo storage dalla piattaforma in molti modi. Mentre questo report tecnico illustra gli NFS, i principali vantaggi di Azure NetApp Files sono:

- Fornire agli utenti la possibilità di utilizzare le copie Snapshot.
- Consente agli utenti di memorizzare grandi quantità di dati su volumi Azure NetApp Files.
- Utilizzo dei vantaggi in termini di performance dei volumi Azure NetApp Files quando si eseguono i modelli su grandi set di file.

Peer AKS VNET e Azure NetApp Files VNET

Per eseguire il peer di AKS VNET a Azure NetApp Files VNET, attenersi alla seguente procedura:

1. Immettere Virtual Networks nel campo di ricerca.
2. Selezionare `vnet aks-vnet-name.` Fare clic su di esso e immettere Peerings nel campo di ricerca.
3. Fare clic su +Add (Aggiungi).
4. Immettere i seguenti descrittori:
   a. Il nome del collegamento di peering è `aks-vnet-name_to_anf`.
   b. SubscriptionID e Azure NetApp Files VNET come partner di peering VNET.
   c. Lasciare tutte le sezioni senza asterisco con i valori predefiniti.
5. Fare clic su Aggiungi.

Per ulteriori informazioni, vedere "Creare, modificare o eliminare un peering di rete virtuale".

Installare Trident

Per installare Trident utilizzando Helm, attenersi alla seguente procedura:

1. Installare Helm (per istruzioni sull'installazione, visitare il "origine").
2. Scaricare ed estrarre il programma di installazione di Trident 20.01.1.

```
$wget
$tar -xf trident-installer-21.01.1.tar.gz
```

3. Modificare la directory in `trident-installer`.

```
$cd trident-installer
```

4. Copia `tridentctl` a una directory del sistema `$PATH`.

```
$sudo cp ./tridentctl /usr/local/bin
```

5. Installare Trident sul cluster Kubernetes (K8s) con Helm ("origine"):
   a. Modificare la directory in `helm` directory.

```
$cd helm
```

   b. Installare Trident.

```
$helm install trident trident-operator-21.01.1.tgz --namespace
trident --create-namespace
```

c. Controllare lo stato dei pod Trident.

```
$kubectl -n trident get pods
```

Se tutti i pod sono in funzione, Trident viene installato ed è possibile procedere.

6. Impostare il backend Azure NetApp Files e la classe storage per AKS.

a. Creare un principio di servizio Azure.

Il service principal è il modo in cui Trident comunica con Azure per gestire le risorse Azure NetApp Files.

```
$az ad sp create-for-rbac --name ""
```

L'output dovrebbe essere simile al seguente esempio:

```
{
"appId": "xxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx",
"displayName": "netapptrident",
"name": "",
"password": "xxxxxxxxxxxxxxx.xxxxxxxxxxxxxx",
"tenant": "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxx"
}
```

7. Creare un file json backend Trident, nome di esempio `anf-backend.json`.
8. Utilizzando l'editor di testo preferito, completare i seguenti campi all'interno di `anf-backend.json` file:

```
{
    "version": 1,
    "storageDriverName": "azure-netapp-files",
    "subscriptionID": "fakec765-4774-fake-ae98-a721add4fake",
    "tenantID": "fakef836-edc1-fake-bff9-b2d865eefake",
    "clientID": "fake0f63-bf8e-fake-8076-8de91e57fake",
    "clientSecret": "SECRET",
    "location": "westeurope",
    "serviceLevel": "Standard",
    "virtualNetwork": "anf-vnet",
    "subnet": "default",
    "nfsMountOptions": "vers=3,proto=tcp",
    "limitVolumeSize": "500Gi",
    "defaults": {
    "exportRule": "0.0.0.0/0",
    "size": "200Gi"
}
```

9. Sostituire i seguenti campi:
   - `subscriptionID`. Il tuo ID di abbonamento Azure.
   - `tenantID`. Il tuo ID tenant Azure dall'output di `az ad sp` nella fase precedente.
   - `clientID`. Il tuo appID dall'output di `az ad sp` nella fase precedente.
   - `clientSecret`. La password dall'output di `az ad sp` nella fase precedente.
10. Chiedere a Trident di creare il backend Azure NetApp Files in `trident` namespace con `anf-backend.json` come file di configurazione:

```
$tridentctl create backend -f anf-backend.json -n trident
```

```
+------------------------+--------------------+--------------------------------------+--------+---------+
|          NAME          |   STORAGE DRIVER   |                 UUID                 | STATE  | VOLUMES |
+------------------------+--------------------+--------------------------------------+--------+---------+
| azurenetappfiles_86181 | azure-netapp-files | 2ca85462-59ac-4946-be05-c03f5575a2ad | online |       0 |
+------------------------+--------------------+--------------------------------------+--------+---------+
```

11. Creare una classe di storage. Kubernetes consente agli utenti di eseguire il provisioning dei volumi utilizzando PVC che specificano una classe di storage in base al nome. Chiedere a K8s di creare una classe di storage `azurenetappfiles` Che fa riferimento al backend Trident creato nel passaggio precedente.
12. Creare un YAML (`anf-storage-class.yaml`) per la copia e la classe di storage.

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
name: azurenetappfiles
provisioner: netapp.io/trident
parameters:
backendType: "azure-netapp-files"
$kubectl create -f anf-storage-class.yaml
```

13. Verificare che la classe di storage sia stata creata.

```
kubectl get sc azurenetappfiles
```

```
NAME               PROVISIONER            RECLAIMPOLICY   VOLUMEBINDINGMODE   ALLOWVOLUMEEXPANSION   AGE
azurenetappfiles   csi.trident.netapp.io  Delete          Immediate           false                  98s
```

Impostare Dask con L'implementazione DI RAPIDS su AKS utilizzando Helm

Per configurare la distribuzione di Dask con RAPIDS su AKS utilizzando Helm, attenersi alla seguente procedura:

1. Creare uno spazio dei nomi per l'installazione di Dask con RAPIDS.

```
kubectl create namespace rapids-dask
```

2. Creare un PVC per memorizzare il set di dati del tasso di click-through:
   a. Salvare il seguente contenuto YAML in un file per creare un PVC.

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: pvc-criteo-data
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 1000Gi
  storageClassName: azurenetappfiles
```

   b. Applicare il file YAML al cluster Kubernetes.

```
kubectl -n rapids-dask apply -f <your yaml file>
```

3. Clonare il `rapidsai git` repository ("https://github.com/rapidsai/helm-chart").

```
git clone https://github.com/rapidsai/helm-chart helm-chart
```

4. Modificare `values.yaml` E includere il PVC creato in precedenza per i lavoratori e l'area di lavoro Jupyter.
   a. Accedere alla `rapidsai` directory del repository.

```
cd helm-chart/rapidsai
```

   b. Aggiornare `values.yaml` Archiviare e montare il volume utilizzando PVC.

```
dask:
  …
  worker:
    name: worker
    …
    mounts:
      volumes:
        - name: data
          persistentVolumeClaim:
            claimName: pvc-criteo-data
      volumeMounts:
        - name: data
          mountPath: /data
    …
  jupyter:
    name: jupyter
    …
    mounts:
      volumes:
        - name: data
          persistentVolumeClaim:
            claimName: pvc-criteo-data
      volumeMounts:
        - name: data
          mountPath: /data
    …
```

5. Accedere alla home directory del repository e implementare Dask con tre nodi di lavoro su AKS utilizzando

Helm.

```
cd ..
helm dep update rapidsai
helm install rapids-dask --namespace rapids-dask rapidsai
```

Livelli di performance Azure NetApp Files

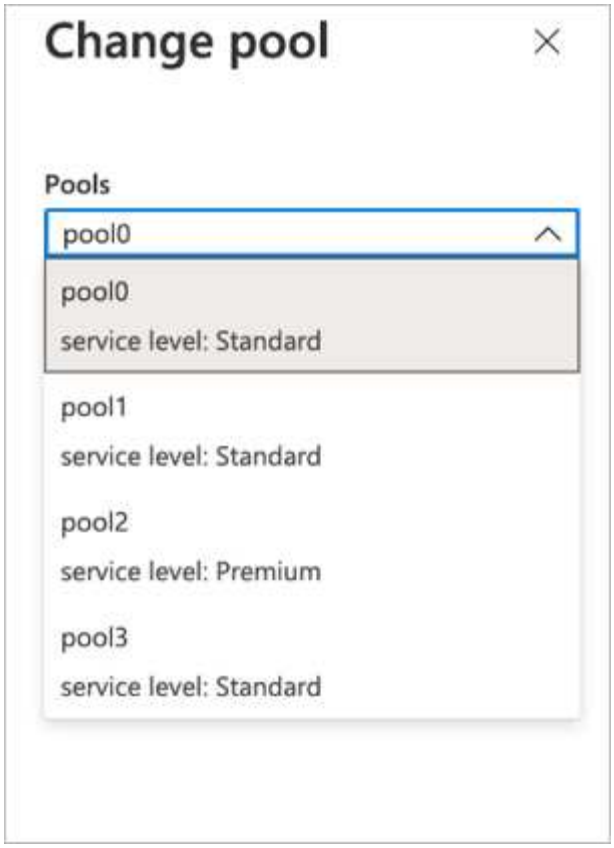
È possibile modificare il livello di servizio di un volume esistente spostando il volume in un altro pool di capacità che utilizza il livello di servizio desiderato per il volume. Questa soluzione consente ai clienti di iniziare con un piccolo set di dati e un piccolo numero di GPU nel Tier standard e scalare in orizzontale o in verticale fino al Tier Premium con l'aumentare della quantità di dati e GPU. Il livello Premium offre un throughput per terabyte quattro volte superiore rispetto al livello Standard e la scalabilità verticale viene eseguita senza dover spostare alcun dato per modificare il livello di servizio di un volume.

Modificare dinamicamente il livello di servizio di un volume

Per modificare dinamicamente il livello di servizio di un volume, attenersi alla seguente procedura:

1. Nella pagina Volumes (volumi), fare clic con il pulsante destro del mouse sul volume di cui si desidera modificare il livello di servizio. Selezionare Cambia pool.

2. Nella finestra Change Pool, selezionare il pool di capacità in cui si desidera spostare il volume.



3. Fare clic su OK.

Automatizza la modifica del Tier delle performance

Sono disponibili le seguenti opzioni per automatizzare le modifiche del Tier di performance:

- La modifica dinamica del livello di servizio è ancora in Public Preview (Anteprima pubblica) e non è attivata per impostazione predefinita. Per attivare questa funzione nell'abbonamento Azure, consultare la presente documentazione "Modificare dinamicamente il livello di servizio di un volume".
- I comandi di modifica del pool di volumi Azure CLI sono forniti in "documentazione per la modifica del pool di volumi" e nel seguente esempio:

```
az netappfiles volume pool-change -g mygroup --account-name myaccname
--pool-name mypoolname --name myvolname --new-pool-resource-id
mynewresourceid
```

- PowerShell "Set-AzNetAppFilesVolumePool cmdlet" Modifica il pool di un volume Azure NetApp Files e viene mostrato nell'esempio seguente:

```
Set-AzNetAppFilesVolumePool
-ResourceGroupName "MyRG"
-AccountName "MyAnfAccount"
-PoolName "MyAnfPool"
-Name "MyAnfVolume"
-NewPoolResourceId 7d6e4069-6c78-6c61-7bf6-c60968e45fbf
```

Fare clic per valutare l'elaborazione dei dati di previsione e modellare la formazione

Librerie per l'elaborazione dei dati e la formazione sui modelli

La tabella seguente elenca le librerie e i framework utilizzati per creare questa attività. Tutti questi componenti sono stati completamente integrati con i controlli di sicurezza e accesso basati sui ruoli di Azure.

Librerie/framework	Descrizione
CuML di Dask	Per CONSENTIRE A ML di lavorare su GPU, il "Libreria cuML" Fornisce l'accesso al pacchetto RAPIDS cuML con DAK. RAPIDS cuML implementa i più diffusi algoritmi ML, tra cui clustering, riduzione delle dimensioni e approcci di regressione, con implementazioni basate su GPU ad alte performance, che offrono velocità fino a 100 volte superiori rispetto agli approcci basati su CPU.
Dask cuDF	CuDF include diverse altre funzioni che supportano l'estrazione, la trasformazione, il carico (ETL) con accelerazione GPU, come il sottosetting dei dati, le trasformazioni, la codifica one-hot e molto altro ancora. Il team RAPIDS gestisce un "libreria dask-cudf" Sono inclusi i metodi di supporto per l'utilizzo di Dask e cuDF.
Scikit Impara	Scikit-Learn offre decine di algoritmi e modelli di apprendimento automatico integrati, chiamati stimatori. Ciascuno "stimatore" può essere adattato ad alcuni dati utilizzando its "adatta" metodo.

Abbiamo utilizzato due notebook per costruire LE pipeline ML per il confronto; uno è l'approccio convenzionale Pandas scikit-Learn e l'altro è la formazione distribuita con RAPIDS e Dask. Ciascun notebook può essere testato singolarmente per verificarne le prestazioni in termini di tempo e scalabilità. Copriamo ogni notebook singolarmente per dimostrare i vantaggi della formazione distribuita utilizzando RAPIDS e Dask.

Load Criteo fare clic su Logs giorno 15 in Pandas e formare un modello di foresta casuale scikit-learn

In questa sezione viene descritto come abbiamo utilizzato Pandas e Dask DataFrame per caricare i dati dei registri Click dall'insieme di dati Criteo Terabyte. Il caso d'utilizzo è importante nella pubblicità digitale per gli scambi di annunci per creare i profili degli utenti

prevedendo se gli annunci verranno cliccati o se lo scambio non utilizza un modello accurato in una pipeline automatica.

Abbiamo caricato i dati del giorno 15 dal set di dati Click Logs, per un totale di 45 GB. Eseguire la seguente cella nel notebook Jupyter `CTR-PandasRF-collated.ipynb` Crea un Pandas DataFrame che contiene i primi 50 milioni di righe e genera un modello di foresta casuale scikit-Learn.

```
%%time
import pandas as pd
import numpy as np
header = ['col'+str(i) for i in range (1,41)] #note that according to
criteo, the first column in the dataset is Click Through (CT). Consist of
40 columns
first_row_taken = 50_000_000 # use this in pd.read_csv() if your compute
resource is limited.
# total number of rows in day15 is 20B
# take 50M rows
"""
Read data & display the following metrics:
1. Total number of rows per day
2. df loading time in the cluster
3. Train a random forest model
"""
df = pd.read_csv(file, nrows=first_row_taken, delimiter='\t',
names=header)
# take numerical columns
df_sliced = df.iloc[:, 0:14]
# split data into training and Y
Y = df_sliced.pop('col1') # first column is binary (click or not)
# change df_sliced data types & fillna
df_sliced = df_sliced.astype(np.float32).fillna(0)
from sklearn.ensemble import RandomForestClassifier
# Random Forest building parameters
# n_streams = 8 # optimization
max_depth = 10
n_bins = 16
n_trees = 10
rf_model = RandomForestClassifier(max_depth=max_depth,
n_estimators=n_trees)
rf_model.fit(df_sliced, Y)
```

Per eseguire la previsione utilizzando un modello di foresta casuale con formazione, eseguire il paragrafo seguente in questo notebook. Abbiamo preso gli ultimi un milione di righe dal giorno 15 come set di test per evitare qualsiasi duplicazione. La cella calcola anche la precisione della previsione, definita come la percentuale di occorrenze che il modello prevede accuratamente se un utente fa clic su un annuncio o meno. Per esaminare eventuali componenti non familiari presenti in questo notebook, consultare la sezione "documentazione ufficiale scikit-learn".

```
# testing data, last 1M rows in day15
test_file = '/data/day_15_test'
with open(test_file) as g:
    print(g.readline())

# dataFrame processing for test data
test_df = pd.read_csv(test_file, delimiter='\t', names=header)
test_df_sliced = test_df.iloc[:, 0:14]
test_Y = test_df_sliced.pop('col1')
test_df_sliced = test_df_sliced.astype(np.float32).fillna(0)
# prediction & calculating error
pred_df = rf_model.predict(test_df_sliced)
from sklearn import metrics
# Model Accuracy
print("Accuracy:",metrics.accuracy_score(test_Y, pred_df))
```

Caricare il giorno 15 in Dask e formare un modello di foresta casuale di Dask cuML

In modo simile alla sezione precedente, caricare Criteo Click Logs Day 15 in Pandas e formare un modello di foresta casuale scikit-learn. In questo esempio, è stato eseguito il caricamento di DataFrame con Dask cuDF e il training di un modello di foresta casuale in Dask cuML. Nella sezione abbiamo confrontato le differenze di tempo e di scala per la formazione ""Confronto tra i tempi di formazione"."

criteo_dask_RF.ipynb

Questo notebook importa `numpy`, `cuml`e il necessario `dask` librerie, come mostrato nell'esempio seguente:

```
import cuml
from dask.distributed import Client, progress, wait
import dask_cudf
import numpy as np
import cudf
from cuml.dask.ensemble import RandomForestClassifier as cumlDaskRF
from cuml.dask.common import utils as dask_utils
```

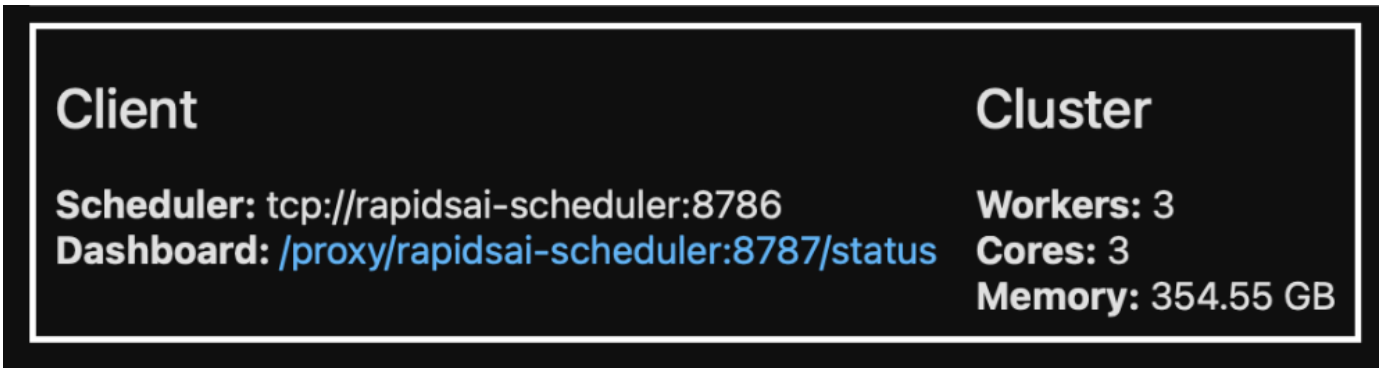
Avviare Dask Client().

```
client = Client()
```

Se il cluster è configurato correttamente, è possibile visualizzare lo stato dei nodi di lavoro.

```
client
workers = client.has_what().keys()
n_workers = len(workers)
n_streams = 8 # Performance optimization
```

Nel nostro cluster AKS viene visualizzato il seguente stato:



Si noti che Dask utilizza il paradigma di esecuzione pigro: Invece di eseguire il codice di elaborazione istantaneamente, Dask crea invece un DAG (Directed Acyclic Graph) di esecuzione. IL DAG contiene una serie di attività e le relative interazioni che ciascun lavoratore deve eseguire. Questo layout significa che i task non vengono eseguiti finché l'utente non dice a Task di eseguirli in un modo o nell'altro. Con Dask hai tre opzioni principali:

- **Call compute() su un DataFrame.** questa chiamata elabora tutte le partizioni e restituisce i risultati allo scheduler per l'aggregazione finale e la conversione in cuDF DataFrame. Questa opzione deve essere utilizzata con parsimonia e solo in caso di risultati fortemente ridotti, a meno che il nodo dello scheduler non esaurisca la memoria.
- **Call Persistent() su un DataFrame.** questa chiamata esegue il grafico, ma, invece di restituire i risultati al nodo scheduler, li mantiene in memoria nel cluster in modo che l'utente possa riutilizzare questi risultati intermedi lungo la pipeline senza dover eseguire nuovamente la stessa elaborazione.
- **Call head() su un DataFrame.** proprio come con cuDF, questa chiamata restituisce 10 record al nodo Scheduler. Questa opzione consente di verificare rapidamente se il DataFrame contiene il formato di output desiderato o se i record stessi hanno senso, a seconda dell'elaborazione e del calcolo.

Pertanto, a meno che l'utente non chiami una di queste azioni, i lavoratori sono inattivi in attesa che lo scheduler avvii l'elaborazione. Questo paradigma di esecuzione pigro è comune nei moderni framework di calcolo distribuiti e paralleli come Apache Spark.

Il paragrafo seguente forma un modello di foresta casuale utilizzando Dask cuML per il calcolo distribuito con accelerazione GPU e calcola la precisione di previsione del modello.

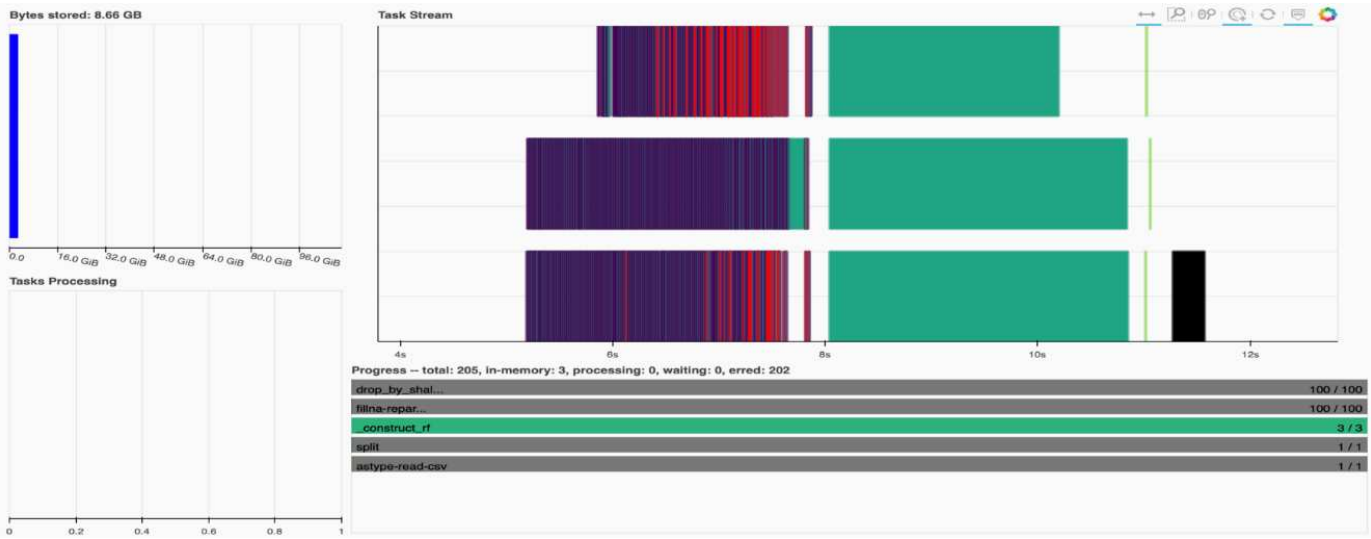
```
Adsf
# Random Forest building parameters
n_streams = 8 # optimization
max_depth = 10
n_bins = 16
n_trees = 10
cuml_model = cumlDaskRF(max_depth=max_depth, n_estimators=n_trees,
n_bins=n_bins, n_streams=n_streams, verbose=True, client=client)
cuml_model.fit(gdf_sliced_small, Y)
# Model prediction
pred_df = cuml_model.predict(gdf_test)
# calculate accuracy
cu_score = cuml.metrics.accuracy_score( test_y, pred_df )
```

Monitorate la Task utilizzando la dashboard nativa dei Task Streams

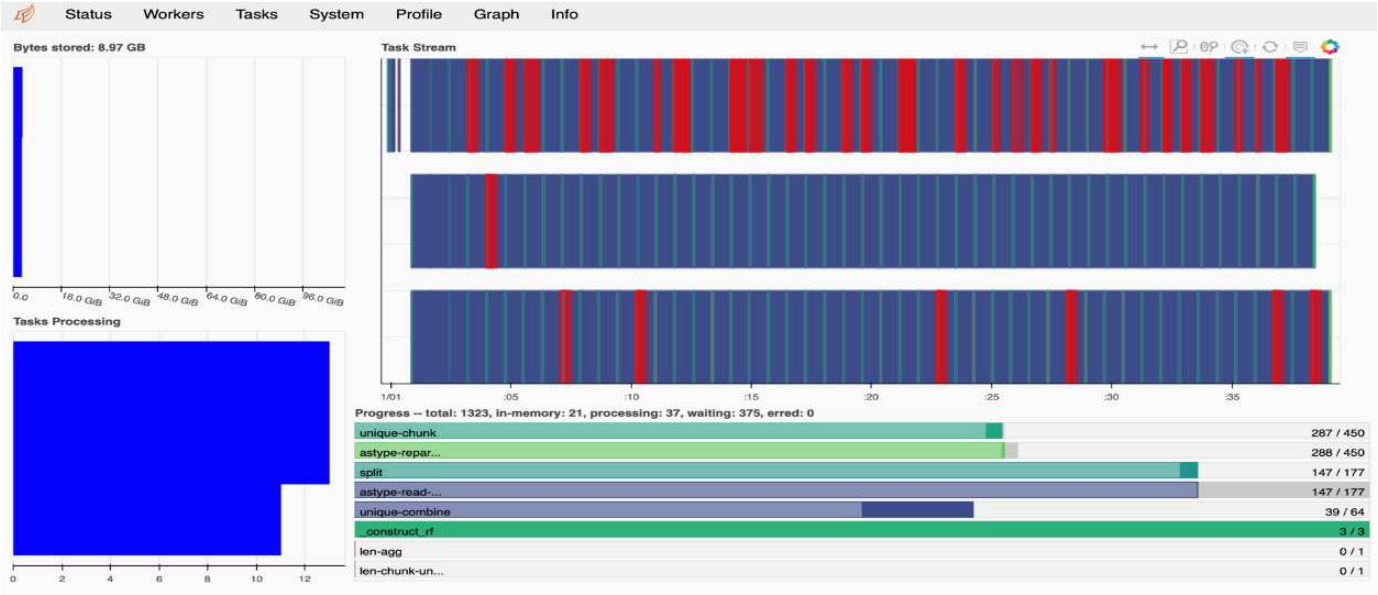
Il "Scheduler distribuito di Dask" fornisce feedback live in due forme:

- Una dashboard interattiva contenente numerosi grafici e tabelle con informazioni in tempo reale
- Barra di avanzamento adatta per l'utilizzo interattivo in console o notebook

Nel nostro caso, la figura seguente mostra come è possibile monitorare l'avanzamento del task, inclusi i byte memorizzati, il Task Stream con una dettagliata suddivisione del numero di flussi e l'avanzamento in base ai nomi delle attività con le funzioni associate eseguite. Nel nostro caso, poiché abbiamo tre nodi di lavoro, ci sono tre blocchi principali di flusso e i codici colore indicano attività diverse all'interno di ogni flusso.



È possibile analizzare le singole attività ed esaminare il tempo di esecuzione in millisecondi o identificare eventuali ostacoli o ostacoli. Ad esempio, la figura seguente mostra i flussi di attività per la fase di adattamento del modello di foresta casuale. Le funzioni eseguite sono notevolmente più numerose, tra cui il chunk unico per l'elaborazione di DataFrame, _Construct_rf per l'adattamento della foresta casuale e così via. La maggior parte del tempo è stato dedicato alle operazioni DataFrame a causa delle grandi dimensioni (45 GB) di un giorno di dati provenienti dai Click Logs di Criteo.



Confronto dei tempi di training

In questa sezione viene confrontato il tempo di training del modello utilizzando i Panda convenzionali rispetto a quello di Dask. Per Pandas, abbiamo caricato una quantità inferiore di dati a causa della natura del tempo di elaborazione più lento per evitare l'overflow della memoria. Pertanto, abbiamo interpolato i risultati per offrire un confronto equo.

La tabella seguente mostra il confronto dei tempi di training raw quando i dati utilizzati per il modello di foresta casuale Pandas sono significativamente inferiori (50 milioni di righe su 20 miliardi al giorno 15 del set di dati). Questo esempio utilizza solo meno del 0.25% di tutti i dati disponibili. Mentre per Dask-cuML abbiamo addestrato il modello di foresta casuale su tutti i 20 miliardi di righe disponibili. I due approcci hanno consentito di ottenere tempi di formazione comparabili.

Approccio	Tempo di training
Scikit-Learn: Utilizzando solo 50M righe nel giorno 15 come dati di training	47 minuti e 21 secondi
RAPIDS-Dask: Utilizzo di tutte le 20B righe del giorno 15 come dati di training	1 ora, 12 minuti e 11 secondi

Se si interpolano i risultati dei tempi di training in modo lineare, come mostrato nella tabella seguente, si ha un vantaggio significativo nell'utilizzo della formazione distribuita con Dask. L'approccio convenzionale Pandas scikit-Learn richiede 13 giorni per elaborare e formare 45 GB di dati per un singolo giorno di log click, mentre L'approccio RAPIDS-Dask elabora la stessa quantità di dati 262.39 volte più velocemente.

Approccio	Tempo di training
Scikit-Learn: Utilizzando tutte le 20B righe del giorno 15 come dati di training	13 giorni, 3 ore, 40 minuti e 11 secondi
RAPIDS-Dask: Utilizzo di tutte le 20B righe del giorno 15 come dati di training	1 ora, 12 minuti e 11 secondi

Nella tabella precedente, è possibile osservare che, utilizzando RAPIDS con Dask per distribuire

l'elaborazione dei dati e modellare la formazione su più istanze GPU, il tempo di esecuzione è significativamente più breve rispetto all'elaborazione convenzionale di Pandas DataFrame con il training del modello scikit-Learn. Questo framework consente la scalabilità verticale e orizzontale nel cloud e on-premise in un cluster multi-GPU a più nodi.

Monitoraggio di Dask e RAPIDE con Prometheus e Grafana

Una volta implementato tutto, esegui le inferenze sui nuovi dati. I modelli prevedono se un utente fa clic su un annuncio in base alle attività di navigazione. I risultati della previsione sono memorizzati in un cuDF di Dask. Puoi monitorare i risultati con Prometheus e visualizzarli nelle dashboard Grafana.

Per ulteriori informazioni, consulta questa sezione "RAPIDS ai Medium post".

Versione di set di dati e modelli con NetApp DataOps Toolkit

Il NetApp DataOps Toolkit per Kubernetes astratta le risorse di storage e i carichi di lavoro Kubernetes fino al livello di spazio di lavoro per la scienza dei dati. Queste funzionalità sono integrate in un'interfaccia semplice e facile da usare, progettata per data scientist e data engineer. Utilizzando la forma familiare di un programma Python, il Toolkit consente a data scientist e ingegneri di eseguire il provisioning e la distruzione delle aree di lavoro di JupyterLab in pochi secondi. Queste aree di lavoro possono contenere terabyte, o persino petabyte, di capacità di storage, consentendo agli scienziati dei dati di memorizzare tutti i set di dati di training direttamente nelle aree di lavoro dei progetti. Sono finiti i tempi della gestione separata degli spazi di lavoro e dei volumi di dati.

Per ulteriori informazioni, visitare il Toolkit "Repository di GitHub".

Notebook Jupyter come riferimento

Al report tecnico sono associati due notebook Jupyter:

- **"CTR-PandasRF-collated.ipynb."** Questo notebook carica il giorno 15 dal set di dati Click Logs di Criteo Terabyte, elabora e formatta i dati in un Pandas DataFrame, forma un modello di foresta casuale Scikit-learn, esegue la previsione e calcola la precisione.
- **"criteo_dask_RF.ipynb."** Questo notebook carica il giorno 15 dal set di dati Click Logs di Criteo Terabyte, elabora e formatta i dati in un cuDF Dask, forma un modello di foresta casuale cuML Dask, esegue la previsione e calcola la precisione. Sfruttando nodi di lavoro multipli con GPU, questo approccio di elaborazione e formazione dei dati distribuiti e dei modelli è altamente efficiente. Maggiore è il numero di dati elaborati, maggiore è il risparmio di tempo rispetto a un approccio ML convenzionale. È possibile implementare questo notebook nel cloud, on-premise o in un ambiente ibrido in cui il cluster Kubernetes contiene calcolo e storage in posizioni diverse, purché la configurazione di rete consenta il libero spostamento dei dati e la distribuzione dei modelli.

Conclusione

Azure NetApp Files, RAPIDS e Dak accelerano e semplificano l'implementazione dell'elaborazione E della formazione ML su larga scala integrandosi con strumenti di orchestrazione come Docker e Kubernetes. Unificando la pipeline di dati end-to-end,

questa soluzione riduce la latenza e la complessità inerenti a molti carichi di lavoro di calcolo avanzati, colmando efficacemente il divario tra sviluppo e operazioni. I data scientist possono eseguire query su set di dati di grandi dimensioni e condividere in modo sicuro dati e modelli algoritmici con altri utenti durante la fase di training.

Quando si creano pipeline ai/ML personalizzate, configurare l'integrazione, la gestione, la sicurezza e l'accessibilità dei componenti in un'architettura è un'attività complessa. Fornire agli sviluppatori l'accesso e il controllo del proprio ambiente presenta un'altra serie di sfide.

Creando un modello di training distribuito end-to-end e una pipeline di dati nel cloud, abbiamo dimostrato un miglioramento di due ordini di grandezza nel tempo totale di completamento del workflow rispetto a un approccio open-source convenzionale che non ha sfruttato i framework di elaborazione e di elaborazione dei dati accelerati dalla GPU.

La combinazione di NetApp, Microsoft, framework di orchestrazione open-source e NVIDIA riunisce le più recenti tecnologie come servizi gestiti con grande flessibilità per accelerare l'adozione della tecnologia e migliorare il time-to-market per le nuove applicazioni ai/ML. Questi servizi avanzati vengono forniti in un ambiente cloud nativo che può essere facilmente trasferito per architetture di implementazione on-premise e ibride.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare le seguenti risorse:

- Azure NetApp Files:
  - Pagina dell'architettura delle soluzioni per Azure NetApp Files

    "https://docs.microsoft.com/azure/azure-netapp-files/azure-netapp-files-solution-architectures"
- Storage persistente Trident per container:
  - Azure NetApp Files e Trident

    "https://netapptrident.readthedocs.io/en/stablev20.07/kubernetes/operations/tasks/backends/anf.html"
- Dask e RAPIDE:
  - Dek

    "https://docs.dask.org/en/latest/"
  - Installare Dask

    "https://docs.dask.org/en/latest/install.html"
  - API di Dask

    "https://docs.dask.org/en/latest/api.html"
  - Apprendimento automatico di Dask

    "https://examples.dask.org/machine-learning.html"

- DAK Distributed Diagnostics

  "https://docs.dask.org/en/latest/diagnostics-distributed.html"

- Framework E tool ML:
  - TensorFlow: Un framework di apprendimento automatico open-source per tutti

    "https://www.tensorflow.org/"

  - Docker

    "https://docs.docker.com"

  - Kubernetes

    "https://kubernetes.io/docs/home/"

  - Kubeflow

    "http://www.kubeflow.org/"

  - Jupyter notebook Server

    "http://www.jupyter.org/"

TR-4896: Formazione distribuita in Azure: Rilevamento della corsia - progettazione della soluzione

Muneer Ahmad e Verron Martina, NetApp Ronen Dar, RUN:ai

Da maggio 2019, Microsoft offre un servizio di portale nativo di Azure per i file service NFS e SMB aziendali basati sulla tecnologia NetApp ONTAP. Questo sviluppo è guidato da una partnership strategica tra Microsoft e NetApp e estende ulteriormente la portata dei servizi dati ONTAP di livello mondiale ad Azure.

NetApp, un provider leader di servizi dati cloud, ha collaborato con RUN: Ai, un'azienda che virtualizza l'infrastruttura ai, per consentire una sperimentazione ai più rapida con un utilizzo completo della GPU. La partnership consente ai team di accelerare l'ai eseguendo numerosi esperimenti in parallelo, con accesso rapido ai dati e sfruttando risorse di calcolo illimitate. RUN: L'ai consente l'utilizzo completo della GPU automatizzando l'allocazione delle risorse, mentre l'architettura comprovata di Azure NetApp Files consente di eseguire ogni esperimento alla massima velocità eliminando le ostruzioni della pipeline dei dati.

NetApp e RUN: L'ai ha Unito le forze per offrire ai clienti una piattaforma a prova di futuro per il loro viaggio nell'ai in Azure. Dagli analytics al calcolo ad alte performance (HPC) alle decisioni autonome (in cui i clienti possono ottimizzare i propri investimenti IT pagando solo ciò di cui hanno bisogno, quando ne hanno bisogno), l'alleanza tra NetApp e RUN: L'ai offre una singola esperienza unificata in Azure Cloud.

Panoramica della soluzione

In questa architettura, l'attenzione si concentra sulla parte più intensiva dal punto di vista computazionale del processo di training distribuito ai o di machine learning (ML) del

rilevamento di corsia. Il rilevamento della corsia è una delle attività più importanti nella guida autonoma, che aiuta a guidare i veicoli attraverso la localizzazione delle linee di demarcazione della corsia. Componenti statici come le linee di demarcazione della corsia guidano il veicolo a guidare in autostrada in modo interattivo e sicuro.

Gli approcci convoluzionali basati sulla rete neurale (CNN) hanno portato la comprensione e la segmentazione della scena a un nuovo livello. Anche se non funziona bene per oggetti con strutture e zone lunghe che potrebbero essere occluse (ad esempio, poli, ombre sulla corsia e così via). La rete neurale convoluzionale spaziale (SCNN) generalizza la CNN a un livello spaziale ricco. Consente la propagazione delle informazioni tra neuroni nello stesso livello, il che lo rende più adatto per oggetti strutturati come corsie, pali o camion con occlusioni. Questa compatibilità è dovuta al fatto che le informazioni spaziali possono essere rafforzate e preservano uniformità e continuità.

Migliaia di immagini di scena devono essere iniettate nel sistema per consentire al modello di apprendere e distinguere i vari componenti del set di dati. Queste immagini includono condizioni meteo, diurne o notturne, strade a più corsie e altre condizioni di traffico.

Per la formazione, è necessario disporre di una buona qualità e quantità di dati. Una singola GPU o più GPU possono richiedere da giorni a settimane per completare il training. La formazione distribuita sui dati può accelerare il processo utilizzando GPU multiple e multinode. Horovod è un framework di questo tipo che garantisce la formazione distribuita, ma la lettura dei dati tra cluster di GPU potrebbe costituire un ostacolo. Azure NetApp Files offre un throughput ultraveloce e elevato e una latenza ridotta e sostenuta per fornire funzionalità scale-out/scale-up in modo che le GPU vengano sfruttate al meglio della loro capacità di calcolo. I nostri esperimenti hanno verificato che tutte le GPU nel cluster vengono utilizzate in media più del 96% per l'addestramento del rilevamento di corsia mediante SCNN.

Pubblico di riferimento

La scienza dei dati incorpora diverse discipline nell'IT e nel business, pertanto più persone fanno parte del nostro pubblico di riferimento:

- Gli scienziati dei dati hanno bisogno della flessibilità necessaria per utilizzare gli strumenti e le librerie di loro scelta.
- I data engineer devono sapere come i dati scorrono e dove risiedono.
- Esperti di casi d'utilizzo per la guida autonoma.
- Amministratori e architetti del cloud per configurare e gestire le risorse cloud (Azure).
- Un tecnico DevOps ha bisogno dei tool per integrare le nuove applicazioni ai/ML nelle pipeline di integrazione continua e implementazione continua (ci/CD).
- Gli utenti aziendali desiderano avere accesso alle applicazioni ai/ML.

In questo documento, descriviamo in che modo Azure NetApp Files, RUN: Ai e Microsoft Azure aiutano ciascuno di questi ruoli a portare valore al business.

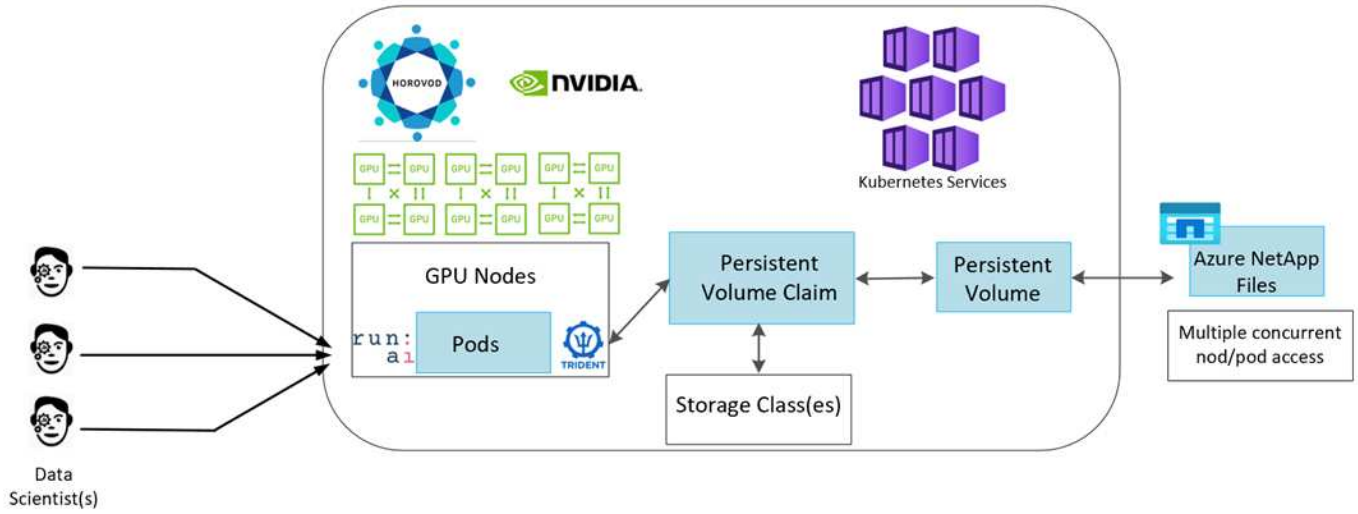
Tecnologia della soluzione

In questa sezione vengono illustrati i requisiti tecnologici per il caso di utilizzo del rilevamento di corsia implementando una soluzione di training distribuita su larga scala che viene eseguita completamente nel cloud Azure. La figura seguente fornisce una panoramica dell'architettura della soluzione.

Gli elementi utilizzati in questa soluzione sono:

- Servizio Azure Kubernetes (AKS)
- Azure Compute SKU con GPU NVIDIA
- Azure NetApp Files
- ESECUZIONE: AI
- Trident di NetApp

I collegamenti a tutti gli elementi menzionati sono elencati nella "Ulteriori informazioni" sezione.



Requisiti di risorse e servizi cloud

La seguente tabella elenca i componenti hardware necessari per implementare la soluzione. I componenti cloud utilizzati in qualsiasi implementazione della soluzione possono variare in base ai requisiti del cliente.

Cloud	Quantità
AKS	Almeno tre nodi di sistema e tre nodi di lavoro GPU
Nodi di sistema delle SKU delle macchine virtuali (VM)	Tre Standard_DS2_v2
Nodi di lavoro GPU SKU VM	Tre standard_NC6s_v3
Azure NetApp Files	Tier standard da 4 TB

Requisiti software

La seguente tabella elenca i componenti software necessari per implementare la soluzione. I componenti software utilizzati in qualsiasi implementazione della soluzione possono variare in base ai requisiti del cliente.

Software	Versione o altre informazioni
AKS - versione di Kubernetes	1.18.14
ESEGUI:AI CLI	v2.2.25
RUN:ai Orchestration Kubernetes Operator version	1.0.109
Horovod	0.21.2

Software	Versione o altre informazioni
Trident di NetApp	20.01.1
Timone	3.0.0

Lane Detection – formazione distribuita con RUN:ai

Questa sezione fornisce dettagli sulla configurazione della piattaforma per l'esecuzione del training distribuito di rilevamento della corsia su larga scala utilizzando L'ORCHESTRATOR DI intelligenza artificiale. Discutiamo dell'installazione di tutti gli elementi della soluzione e dell'esecuzione del lavoro di training distribuito sulla piattaforma suddetta. IL controllo della versione ML viene completato utilizzando NetApp SnapshotTM collegato a ESPERIMENTI DI RUN: Ai per ottenere la riproducibilità dei dati e dei modelli. IL controllo delle versioni DI ML svolge un ruolo cruciale nel monitoraggio dei modelli, nella condivisione del lavoro tra i membri del team, nella riproducibilità dei risultati, nel passaggio in produzione delle nuove versioni dei modelli e nella provenienza dei dati. NetApp ML version control (Snapshot) è in grado di acquisire versioni point-in-time dei dati, dei modelli addestrati e dei registri associati a ciascun esperimento. Grazie al supporto API completo, è facile da integrare con LA piattaforma DI ESECUZIONE: Ai; devi solo attivare un evento in base allo stato del training. Inoltre, è necessario acquisire lo stato dell'intero esperimento senza modificare nulla nel codice o nei container eseguiti su Kubernetes (K8s).

Infine, questo report tecnico si conclude con la valutazione delle performance su più nodi abilitati alla GPU in AKS.

Training distribuito per il caso di utilizzo del rilevamento di corsia utilizzando il set di dati TuSimple

In questo report tecnico, viene eseguito un training distribuito sul set di dati TuSimple per il rilevamento della corsia. Horovod viene utilizzato nel codice di training per condurre training distribuiti su più nodi GPU contemporaneamente nel cluster Kubernetes tramite AKS. Il codice viene confezionato come immagini container per il download e l'elaborazione dei dati TuSimple. I dati elaborati vengono memorizzati su volumi persistenti allocati dal plug-in di NetApp Trident. Per il training, viene creata un'altra immagine container che utilizza i dati memorizzati nei volumi persistenti creati durante il download dei dati.

Per inviare i dati e il lavoro di training, utilizza RUN: Ai per orchestrare l'allocazione e la gestione delle risorse. ESEGUI: L'ai consente di eseguire operazioni MPI (message Passing Interface) necessarie per Horovod. Questo layout consente a più nodi GPU di comunicare tra loro per aggiornare i pesi di training dopo ogni mini batch di training. Consente inoltre di monitorare la formazione attraverso l'interfaccia utente e la CLI, semplificando il monitoraggio dei progressi degli esperimenti.

NetApp Snapshot è integrato nel codice di training e acquisisce lo stato dei dati e il modello formativo per ogni esperimento. Questa funzionalità consente di tenere traccia della versione dei dati e del codice utilizzati e del modello di formazione associato generato.

Installazione e configurazione di AKS

Per la configurazione e l'installazione del cluster AKS, visitare il sito Web all'indirizzo "Creare un cluster AKS". Quindi, attenersi alla seguente serie di passaggi:

1. Quando si seleziona il tipo di nodi (che si tratti di nodi di sistema (CPU) o di lavoro (GPU)), selezionare quanto segue:

   a. Aggiungere il nodo di sistema primario denominato `agentpool` su `Standard_DS2_v2` dimensione. Utilizzare i tre nodi predefiniti.

   b. Aggiungi nodo di lavoro `gpupool` con `the Standard_NC6s_v3` dimensioni del pool. Utilizzare almeno tre nodi per i nodi GPU.

L'implementazione richiede 10 minuti.

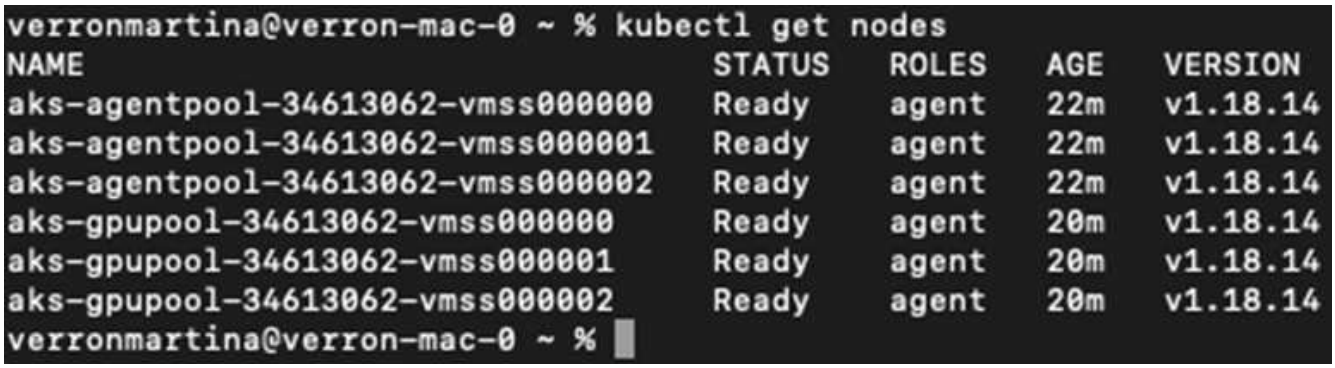
2. Al termine dell'implementazione, fare clic su Connect to Cluster (Connetti al cluster). Per connettersi al cluster AKS appena creato, installare il tool della riga di comando Kubernetes dall'ambiente locale (laptop/PC). Visitare il sito "Strumenti di installazione" Per installarlo in base al sistema operativo in uso.
3. "Installare Azure CLI nell'ambiente locale".
4. Per accedere al cluster AKS dal terminale, immettere `az login` e inserire le credenziali.
5. Eseguire i due comandi seguenti:

```
az account set --subscription xxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxxx
aks get-credentials --resource-group resourcegroup --name aksclustername
```

6. Immettere questo comando nella riga di comando Azure:

```
kubectl get nodes
```

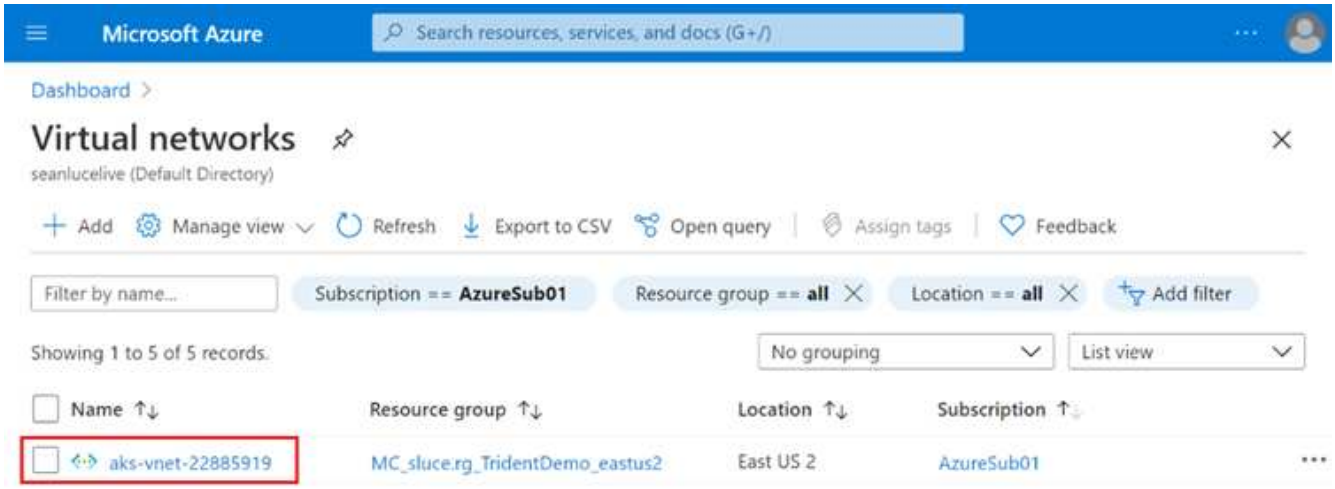
Se tutti e sei i nodi sono attivi e in esecuzione, il cluster AKS è pronto e connesso all'ambiente locale.



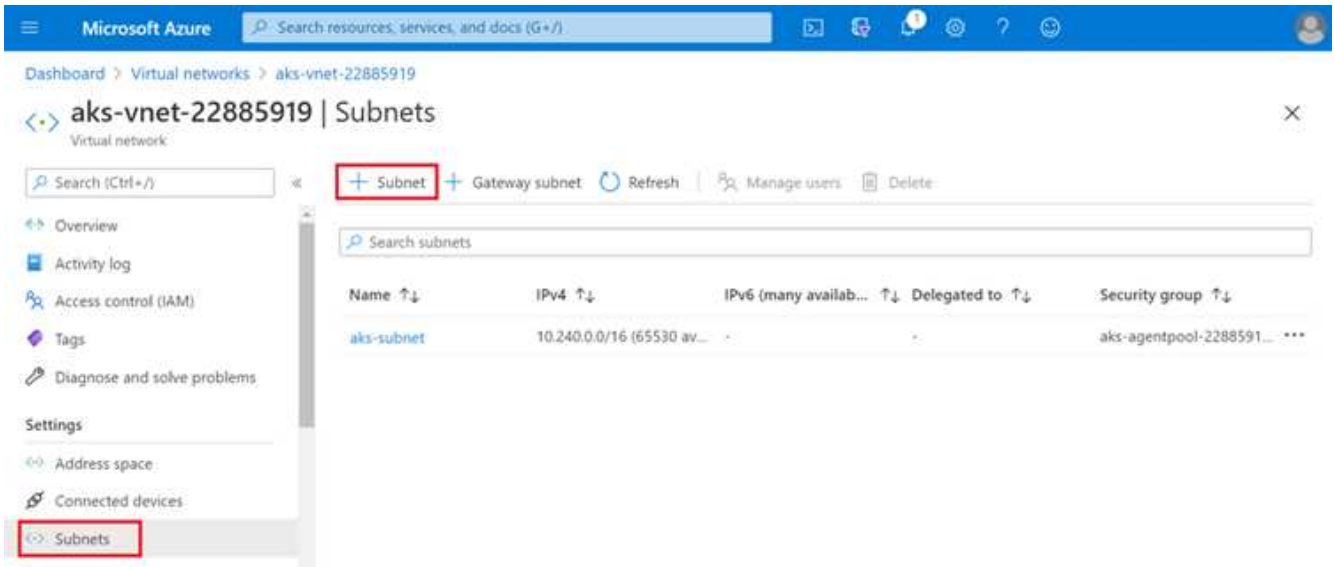
Creare una subnet delegata per Azure NetApp Files

Per creare una subnet delegata per Azure NetApp Files, seguire questa serie di passaggi:

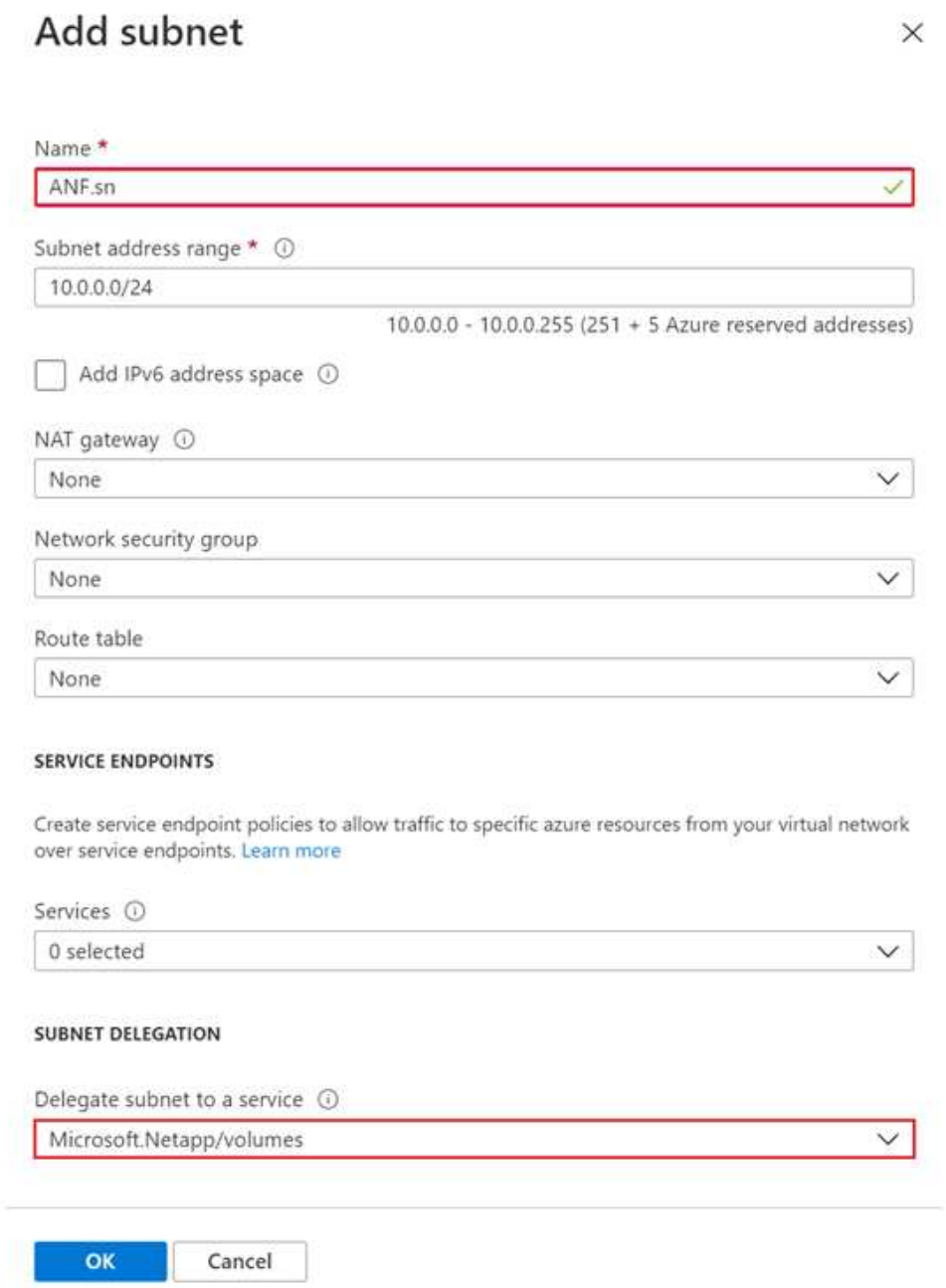
1. Accedere alle reti virtuali all'interno del portale Azure. Trova la tua rete virtuale appena creata. Dovrebbe avere un prefisso come aks-vnet, come mostrato qui. Fare clic sul nome della rete virtuale.



2. Fare clic su subnet e selezionare +Subnet nella barra degli strumenti superiore.



3. Specificare un nome per la subnet, ad esempio `ANF.sn` E sotto l'intestazione Subnet Delegation (delega subnet), selezionare Microsoft.NetApp/volumes. Non cambiare altro. Fare clic su OK.



I volumi Azure NetApp Files vengono allocati nel cluster di applicazioni e vengono utilizzati come dichiarazioni di volumi persistenti (PVC) in Kubernetes. A sua volta, questa allocazione ci offre la flessibilità di mappare i volumi a diversi servizi, sia che si trattino di notebook Jupyter, funzioni senza server e così via

Gli utenti dei servizi possono consumare lo storage dalla piattaforma in molti modi. I principali vantaggi di Azure NetApp Files sono:

- Offre agli utenti la possibilità di utilizzare le snapshot.
- Consente agli utenti di memorizzare grandi quantità di dati su volumi Azure NetApp Files.
- Ottenere i vantaggi in termini di performance dei volumi Azure NetApp Files quando si eseguono i modelli su grandi set di file.

Configurazione di Azure NetApp Files

Per completare la configurazione di Azure NetApp Files, è necessario configurarla come descritto in "QuickStart: Configurazione di Azure NetApp Files e creazione di un volume NFS".

Tuttavia, è possibile omettere la procedura per creare un volume NFS per Azure NetApp Files, poiché si creeranno volumi tramite Trident. Prima di continuare, assicurarsi di disporre di:

1. "Registrato per Azure NetApp Files e per il provider di risorse NetApp (tramite la shell cloud di Azure)".
2. "Creato un account in Azure NetApp Files".
3. "Impostare un pool di capacità" (Minimo 4 TiB Standard o Premium a seconda delle esigenze).

Peering della rete virtuale AKS e della rete virtuale Azure NetApp Files

Quindi, eseguire il peer della rete virtuale AKS con Azure NetApp Files VNET seguendo questa procedura:

1. Nella casella di ricerca nella parte superiore del portale Azure, digitare virtual networks (reti virtuali).
2. Fare clic su VNET aks- vnet-name, quindi immettere Peerings nel campo di ricerca.
3. Fare clic su +Add (Aggiungi) e inserire le informazioni fornite nella tabella seguente:

Campo	Valore o descrizione #
Nome del collegamento peering	aks-vnet-name_to_an
SubscriptionID	Iscrizione a Azure NetApp Files VNET a cui stai eseguendo il peering
Partner di peering VNET	Azure NetApp Files VNET

> Lasciare tutte le sezioni non contrassegnate come predefinite

4. Fare clic su ADD (AGGIUNGI) o su OK per aggiungere il peering alla rete virtuale.

Per ulteriori informazioni, visitare il sito "Creare, modificare o eliminare un peering di rete virtuale".

Trident

Trident è un progetto open-source che NetApp gestisce per lo storage persistente dei container delle applicazioni. Trident è stato implementato come un provisioning controller esterno che viene eseguito come pod stesso, monitorando i volumi e automatizzando completamente il processo di provisioning.

NetApp Trident consente un'integrazione perfetta con K8s creando e allegando volumi persistenti per l'archiviazione di set di dati di training e modelli di training. Questa funzionalità semplifica l'utilizzo di K8 da parte di data scientist e data engineer senza il fastidio di memorizzare e gestire manualmente i set di dati. Trident elimina inoltre la necessità per i data scientist di imparare a gestire nuove piattaforme dati, poiché integra le attività correlate alla gestione dei dati attraverso l'integrazione API logica.

Installare Trident

Per installare il software Trident, attenersi alla seguente procedura:

1. "Installare prima il timone".
2. Scaricare ed estrarre il programma di installazione di Trident 21.01.1.

```
wget
https://github.com/NetApp/trident/releases/download/v21.01.1/trident-
installer-21.01.1.tar.gz
tar -xf trident-installer-21.01.1.tar.gz
```

3. Modificare la directory in `trident-installer`.

```
cd trident-installer
```

4. Copia `tridentctl` a una directory del sistema `$PATH`.

```
cp ./tridentctl /usr/local/bin
```

5. Installare Trident sul cluster K8s con Helm:
   a. Cambiare la directory in Helm directory.

   ```
   cd helm
   ```

   b. Installare Trident.

   ```
   helm install trident trident-operator-21.01.1.tgz --namespace trident
   --create-namespace
   ```

   c. Verificare lo stato dei pod Trident nel modo consueto di K8s:

   ```
   kubectl -n trident get pods
   ```

   d. Se tutti i pod sono in funzione, Trident è installato e si è bene andare avanti.

Configurare il back-end Azure NetApp Files e la classe di storage

Per configurare il back-end Azure NetApp Files e la classe di storage, attenersi alla seguente procedura:

1. Tornare alla home directory.

```
cd ~
```

2. Clonare il "repository di progetto" `lane-detection-SCNN-horovod`.
3. Accedere alla `trident-config` directory.

```
cd ./lane-detection-SCNN-horovod/trident-config
```

4. Creare un principio di servizio Azure (il principio di servizio è il modo in cui Trident comunica con Azure per accedere alle risorse Azure NetApp Files).

```
az ad sp create-for-rbac --name
```

L'output dovrebbe essere simile al seguente esempio:

```
{
  "appId": "xxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx",
   "displayName": "netapptrident",
    "name": "http://netapptrident",
    "password": "xxxxxxxxxxxxxxx.xxxxxxxxxxxxxx",
    "tenant": "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxx"
 }
```

5. Creare il Trident `backend json` file.
6. Utilizzando l'editor di testo preferito, completare i seguenti campi della tabella riportata di seguito all'interno di `anf-backend.json` file.

Campo	Valore
SubscriptionID	Il tuo ID di abbonamento Azure
ID tenant	Il tuo ID tenant Azure (dall'output di az ad sp nel passaggio precedente)
ID cliente	Il tuo appID (dall'output di az ad sp nel passaggio precedente)
ClientSecret	La tua password (dall'output di az ad sp nel passaggio precedente)

Il file dovrebbe essere simile al seguente esempio:

```
{
    "version": 1,
    "storageDriverName": "azure-netapp-files",
    "subscriptionID": "fakec765-4774-fake-ae98-a721add4fake",
    "tenantID": "fakef836-edc1-fake-bff9-b2d865eefake",
    "clientID": "fake0f63-bf8e-fake-8076-8de91e57fake",
    "clientSecret": "SECRET",
    "location": "westeurope",
    "serviceLevel": "Standard",
    "virtualNetwork": "anf-vnet",
    "subnet": "default",
    "nfsMountOptions": "vers=3,proto=tcp",
    "limitVolumeSize": "500Gi",
    "defaults": {
    "exportRule": "0.0.0.0/0",
    "size": "200Gi"
}
```

7. Chiedere a Trident di creare il back-end Azure NetApp Files in `trident` namespace, utilizzando `anf-backend.json` come il file di configurazione come segue:

```
tridentctl create backend -f anf-backend.json -n trident
```

8. Creare la classe di storage:

   a. Gli utenti K8 eseguono il provisioning dei volumi utilizzando PVC che specificano una classe di storage in base al nome. Chiedere a K8s di creare una classe di storage `azurenetappfiles` Questo farà riferimento al back-end Azure NetApp Files creato nel passaggio precedente utilizzando quanto segue:

   ```
   kubectl create -f anf-storage-class.yaml
   ```

   b. Verificare che la classe di storage venga creata utilizzando il seguente comando:

   ```
   kubectl get sc azurenetappfiles
   ```

   L'output dovrebbe essere simile al seguente esempio:

```
NAME               PROVISIONER             RECLAIMPOLICY   VOLUMEBINDINGMODE   ALLOWVOLUMEEXPANSION   AGE
azurenetappfiles   csi.trident.netapp.io   Delete          Immediate           false                  98s
```

Implementare e configurare i componenti di snapshot dei volumi su AKS

Se il cluster non viene fornito con i componenti di snapshot del volume corretti, è possibile installare manualmente questi componenti eseguendo i seguenti passaggi:

AKS 1.18.14 non dispone di Snapshot Controller preinstallato.

1. Installare i CRD Snapshot Beta utilizzando i seguenti comandi:

```
kubectl create -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-
3.0/client/config/crd/snapshot.storage.k8s.io_volumesnapshotclasses.yaml
kubectl create -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-
3.0/client/config/crd/snapshot.storage.k8s.io_volumesnapshotcontents.yam
l
kubectl create -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-
3.0/client/config/crd/snapshot.storage.k8s.io_volumesnapshots.yaml
```

2. Installare Snapshot Controller utilizzando i seguenti documenti di GitHub:

```
kubectl apply -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-3.0/deploy/kubernetes/snapshot-
controller/rbac-snapshot-controller.yaml
kubectl apply -f https://raw.githubusercontent.com/kubernetes-
csi/external-snapshotter/release-3.0/deploy/kubernetes/snapshot-
controller/setup-snapshot-controller.yaml
```

3. Impostare K8s `volumesnapshotclass`Prima di creare uno snapshot di volume "classe di snapshot del volume" deve essere configurato. Creare una classe di snapshot di volume per Azure NetApp Files e utilizzarla per ottenere IL controllo delle versioni ML utilizzando la tecnologia NetApp Snapshot. Creare `volumesnapshotclass netapp-csi-snapclass` e impostarlo sul valore predefinito `volumesnapshotclass `come tale:

```
kubectl create -f netapp-volume-snapshot-class.yaml
```

L'output dovrebbe essere simile al seguente esempio:

```
volumesnapshotclass.snapshot.storage.k8s.io/netapp-csi-snapclass created
```

4. Verificare che la classe di copia Snapshot del volume sia stata creata utilizzando il seguente comando:

```
kubectl get volumesnapshotclass
```

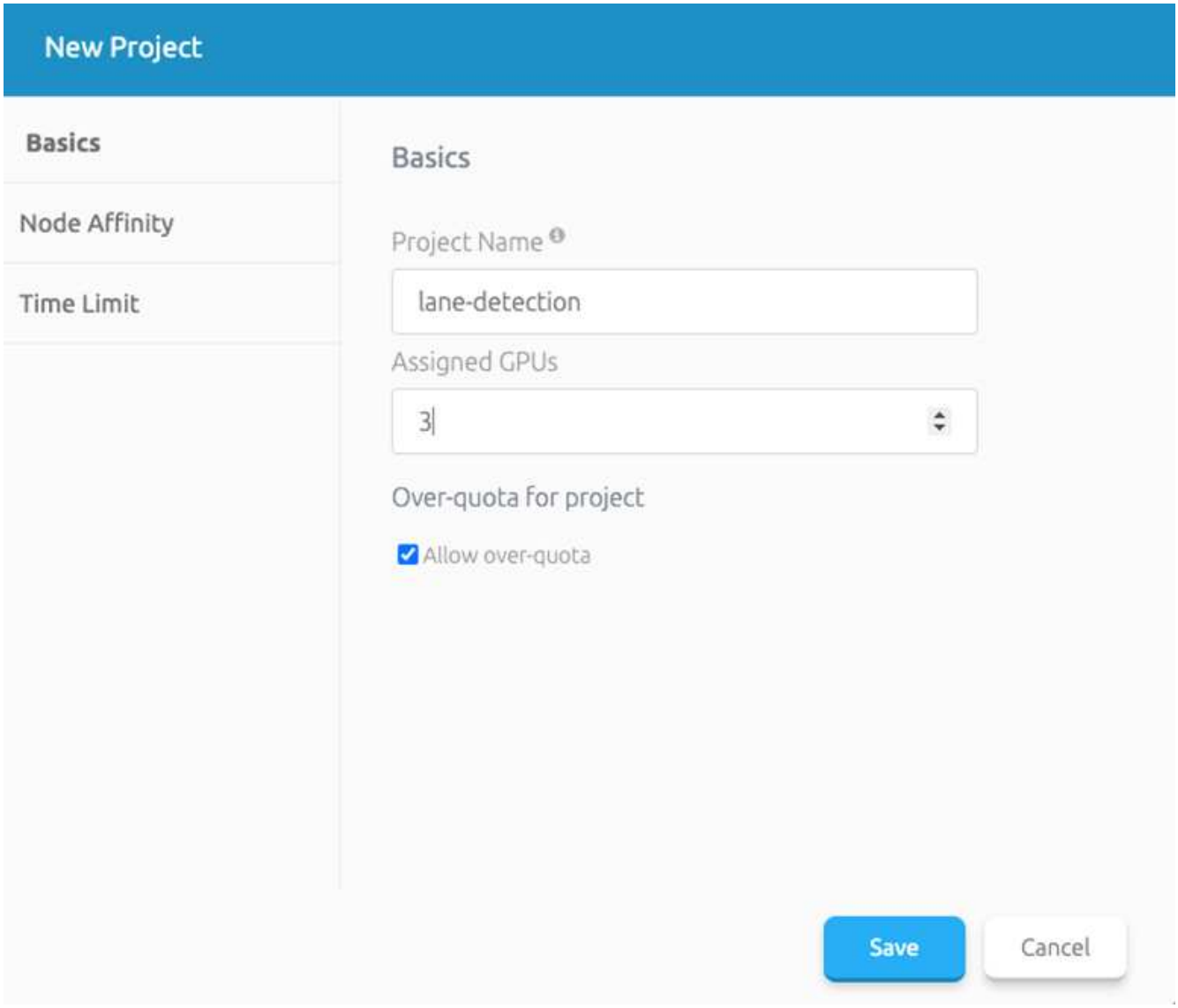
L'output dovrebbe essere simile al seguente esempio:

```
NAME                    DRIVER                  DELETIONPOLICY   AGE
netapp-csi-snapclass    csi.trident.netapp.io   Delete           63s
```

ESEGUI:installazione ai

Per installare RUN:ai, attenersi alla seguente procedura:

1. "Installare IL cluster RUN:ai su AKS".
2. Accedere a app.runai.ai, fare clic su Create New Project (Crea nuovo progetto) e assegnargli il nome di rilevamento della corsia. Verrà creato uno spazio dei nomi su un cluster K8s a partire da `runai-` seguito dal nome del progetto. In questo caso, lo spazio dei nomi creato sarà runai-lane-detection.



3. "INSTALLARE RUN:AI CLI".
4. Sul terminale, impostare il rilevamento di corsia come UN progetto di default RUN: Ai utilizzando il seguente comando:

```
`runai config project lane-detection`
```

L'output dovrebbe essere simile al seguente esempio:

```
Project lane-detection has been set as default project
```

5. Creare ClusterRole e ClusterRoleBinding per lo spazio dei nomi del progetto (ad esempio, `lane-detection)` quindi, l'account di servizio predefinito appartenente a. `runai-lane-detection` lo spazio dei nomi dispone dell'autorizzazione per eseguire le operazioni `volumesnapshot` operazioni durante l'esecuzione del processo:

   a. Elencare gli spazi dei nomi per controllarli `runai-lane-detection` esiste utilizzando questo comando:

   ```
   kubectl get namespaces
   ```

   L'output dovrebbe apparire come nell'esempio seguente:

```
NAME                   STATUS   AGE
default                Active   130m
kube-node-lease        Active   130m
kube-public            Active   130m
kube-system            Active   130m
runai                  Active   4m44s
runai-lane-detection   Active   13s
trident                Active   102m
```

6. Creare ClusterRole `netappsnapshot` E ClusterRoleBinding `netappsnapshot` utilizzando i seguenti comandi:

```
`kubectl create -f runai-project-snap-role.yaml`
`kubectl create -f runai-project-snap-role-binding.yaml`
```

Scaricare ed elaborare il set di dati TuSimple come lavoro RUN:ai

Il processo per scaricare ed elaborare il set di dati TuSimple come UN processo DI ESECUZIONE: Ai è facoltativo. La procedura prevede i seguenti passaggi:

1. Creare e inviare l'immagine del docker o omettere questo passaggio se si desidera utilizzare un'immagine del docker esistente (ad esempio, `muneer7589/download-tusimple:1.0)`

   a. Passare alla home directory:

   ```
   cd ~
   ```

   b. Accedere alla directory dei dati del progetto `lane-detection-SCNN-horovod`:

```
cd ./lane-detection-SCNN-horovod/data
```

c. Modificare `build_image.sh` shell script e modifica il repository di docker in base al tuo. Ad esempio, sostituire `muneer7589` con il nome del repository di docker. È anche possibile modificare il nome e IL TAG dell'immagine del docker (ad esempio `download-tusimple` e. `1.0`):

```
#!/bin/bash
#
# A simple script to build the Docker image.
#
# $ build_image.sh
set -ex

IMAGE=muneer7589/download-tusimple
TAG=1.0

# Build image
echo "Building image: "$IMAGE
docker build . -f Dockerfile \
 --tag "${IMAGE}:${TAG}"
echo "Finished building image: "$IMAGE

# Push image
echo "Pushing image: "$IMAGE
docker push "${IMAGE}:${TAG}"
echo "Finished pushing image: "$IMAGE
```

d. Eseguire lo script per creare l'immagine del docker e inserirla nel repository del docker utilizzando i seguenti comandi:

```
chmod +x build_image.sh
./build_image.sh
```

2. Inviare il lavoro DI ESECUZIONE: Ai per scaricare, estrarre, pre-elaborare e memorizzare il set di dati di rilevamento della corsia TuSimple in un `pvc`, Creata dinamicamente da NetApp Trident:

a. Utilizzare i seguenti comandi per inviare LA SERIOGRAFIA: AI job:

```
runai submit
--name download-tusimple-data
--pvc azurenetappfiles:100Gi:/mnt
--image muneer7589/download-tusimple:1.0
```

b. Inserire le informazioni dalla tabella seguente per inviare il job RUN:ai:

Campo	Valore o descrizione
-name	Nome del lavoro
pvc	PVC del formato [StorageClassName]:Size:ContainerMountPath nell'invio del job di cui sopra, si sta creando un PVC basato su richiesta utilizzando Trident con azurenetappfile di classe storage. La capacità del volume persistente qui è di 100 Gi ed è montata in path /mnt.
-immagine	Immagine Docker da utilizzare durante la creazione del contenitore per questo lavoro

L'output dovrebbe essere simile al seguente esempio:

```
The job 'download-tusimple-data' has been submitted successfully
You can run `runai describe job download-tusimple-data -p lane-detection` to check the job status
```

c. Elencare i job RUN:ai inviati.

```
runai list jobs
```

```
Showing jobs for project lane-detection
NAME                  STATUS            AGE  NODE                          IMAGE                          TYPE  PROJECT        USER          GPUs Allocated (Requested)
PODs Running (Pending)  SERVICE URL(S)
download-tusimple-data  ContainerCreating  1m   aks-agentpool-34613062-vmss00000a  muneer7589/download-tusimple:1.0  Train  lane-detection  verronmartina  0 (0)
1 (0)
```

d. Controllare i log dei lavori inoltrati.

```
runai logs download-tusimple-data -t 10
```

```
751150K .......... .......... .......... .......... ..........  6% 16.2M 20m37s
751200K .......... .......... .......... .......... ..........  6% 11.1M 20m37s
751250K .......... .......... .......... .......... ..........  6% 12.5M 20m36s
751300K .......... .......... .......... .......... ..........  6% 11.3M 20m36s
751350K .......... .......... .......... .......... ..........  6% 15.2M 20m36s
751400K .......... .......... .......... .......... ..........  6% 10.5M 20m36s
751450K .......... .......... .......... .......... ..........  6% 15.2M 20m36s
751500K .......... .......... .......... .......... ..........  6% 14.1M 20m36s
751550K .......... .......... .......... .......... ..........  6% 24.3M 20m36s
751600K .......... .......... .......... .......... ..........  6% 26.3M 20m36s
```

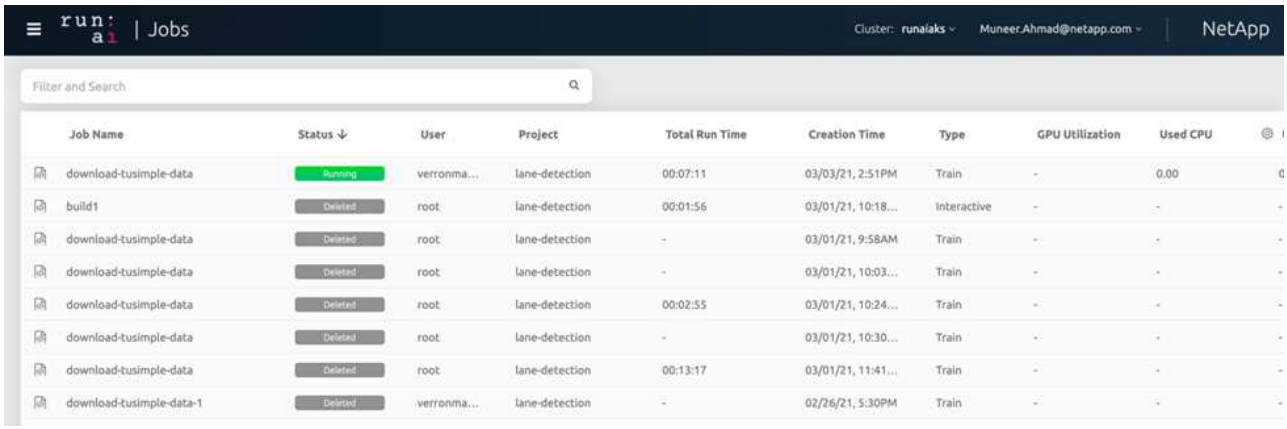
e. Elencare `pvc` creato. Utilizzare questo `pvc` comando per la formazione nella fase successiva.

```
kubectl get pvc | grep download-tusimple-data
```

L'output dovrebbe essere simile al seguente esempio:

```
pvc-download-tusimple-data-0   Bound    pvc-bb03b74d-2c17-40c4-a445-79f3de8d16d5   100Gi      RWO            azurenetappfiles   4m47s
```

a. Controllare il lavoro IN ESECUZIONE: Ai UI (o. `app.run.ai`).



Eseguire un training di rilevamento di corsia distribuito utilizzando Horovod

L'esecuzione di un training di rilevamento di corsia distribuito con Horovod è un processo facoltativo. Tuttavia, di seguito sono riportati i passaggi:

1. Creare e inviare l'immagine del docker o saltare questo passaggio se si desidera utilizzare l'immagine del docker esistente (ad esempio, `muneer7589/dist-lane-detection:3.1)`:

   a. Passare alla home directory.

   ```
   cd ~
   ```

   b. Accedere alla directory del progetto `lane-detection-SCNN-horovod.`

   ```
   cd ./lane-detection-SCNN-horovod
   ```

   c. Modificare il `build_image.sh` shell script e modifica il repository di docker in base al tuo (ad esempio, sostituire `muneer7589` con il nome del repository del docker). È anche possibile modificare il nome e IL TAG dell'immagine del docker (`dist-lane-detection` e. `3.1, for example).`

```
#!/bin/bash
#
# A simple script to build the distributed Docker image.
#
# $ build_image.sh
set -ex

IMAGE=muneer7589/dist-lane-detection
TAG=3.0

# Build image
echo "Building image: "$IMAGE
docker build . -f Dockerfile \
 --tag "${IMAGE}:${TAG}"
echo "Finished building image: "$IMAGE

# Push image
echo "Pushing image: "$IMAGE
docker push "${IMAGE}:${TAG}"
echo "Finished pushing image: "$IMAGE
```

d. Eseguire lo script per creare l'immagine del docker e passare al repository del docker.

```
chmod +x build_image.sh
./build_image.sh
```

2. Inviare la CORSA: Lavoro ai per l'esecuzione del training distribuito (MPI):

a. Utilizzo di submit of RUN: L'ai per la creazione automatica del PVC nella fase precedente (per il download dei dati) consente solo l'accesso RWO, che non consente a più pod o nodi di accedere allo stesso PVC per la formazione distribuita. Aggiornare la modalità di accesso a ReadWriteMany e utilizzare la patch Kubernetes per eseguire questa operazione.

b. Innanzitutto, ottenere il nome del volume del PVC eseguendo il seguente comando:

```
kubectl get pvc | grep download-tusimple-data
```

```
root@ai-w-gpu-2:/mnt/ai_data/anf_runai/lane-detection-SCNN-horovod# kubectl get pvc | grep download-tusimple-data
pvc-download-tusimple-data-0   Bound    pvc-bb03b74d-2c17-40c4-a445-79f3de8d16d5   100Gi      RWX            azurenetappfiles   2d4h
```

c. Applicare la patch al volume e aggiornare la modalità di accesso a ReadWriteMany (sostituire il nome del volume con il proprio nel seguente comando):

```
kubectl patch pv pvc-bb03b74d-2c17-40c4-a445-79f3de8d16d5 -p
'{"spec":{"accessModes":["ReadWriteMany"]}}'
```

d. Inviare la CORSA: Lavoro ai MPI per l'esecuzione del lavoro di training distribuito` utilizzando le informazioni della tabella seguente:

```
runai submit-mpi
--name dist-lane-detection-training
--large-shm
--processes=3
--gpu 1
--pvc pvc-download-tusimple-data-0:/mnt
--image muneer7589/dist-lane-detection:3.1
-e USE_WORKERS="true"
-e NUM_WORKERS=4
-e BATCH_SIZE=33
-e USE_VAL="false"
-e VAL_BATCH_SIZE=99
-e ENABLE_SNAPSHOT="true"
-e PVC_NAME="pvc-download-tusimple-data-0"
```

Campo	Valore o descrizione
nome	Nome del lavoro di formazione distribuito
grande shm	Montare un grande dispositivo /dev/shm si tratta di un file system condiviso montato sulla RAM e fornisce una memoria condivisa abbastanza grande per consentire a più lavoratori della CPU di elaborare e caricare batch nella RAM della CPU.
processi	Numero di processi di formazione distribuiti
gpu	Numero di GPU/processi da allocare per il processo in questo processo, esistono tre processi di lavoro GPU (--processi=3), ciascuno allocato con una singola GPU (--gpu 1)
pvc	Utilizza il volume persistente esistente (pvc-download-tusemplici-data-0) creato dal job precedente (download-tusemplici-data) e viene montato nel percorso /mnt
immagine	Immagine Docker da utilizzare durante la creazione del contenitore per questo lavoro
Definire le variabili di ambiente da impostare nel container	
LAVORATORI_DI_UTILIZZO	Impostando l'argomento su true si attiva il caricamento dei dati multi-processo
NUM_WORKERS	Numero di processi di lavoro del data loader
BATCH_SIZE	Dimensione del batch di training
VALORE_UTILIZZO	L'impostazione dell'argomento su true consente la convalida
VAL_BATCH_SIZE	Dimensione del batch di convalida

Campo	Valore o descrizione
ENABLE_SNAPSHOT	Impostando l'argomento su true, è possibile acquisire dati e snapshot dei modelli con formazione per scopi di versioning ML
NOME_PVC	Nome del pvc di cui eseguire un'istantanea. Nell'invio del job di cui sopra, si sta prendendo un'istantanea di pvc-download-tusSimple-data-0, che consiste di dataset e modelli addestrati

L'output dovrebbe essere simile al seguente esempio:

```
The job 'dist-lane-detection-training' has been submitted successfully
You can run `runai describe job dist-lane-detection-training -p lane-detection` to check the job status
```

e. Elencare il lavoro inoltrato.

```
runai list jobs
```

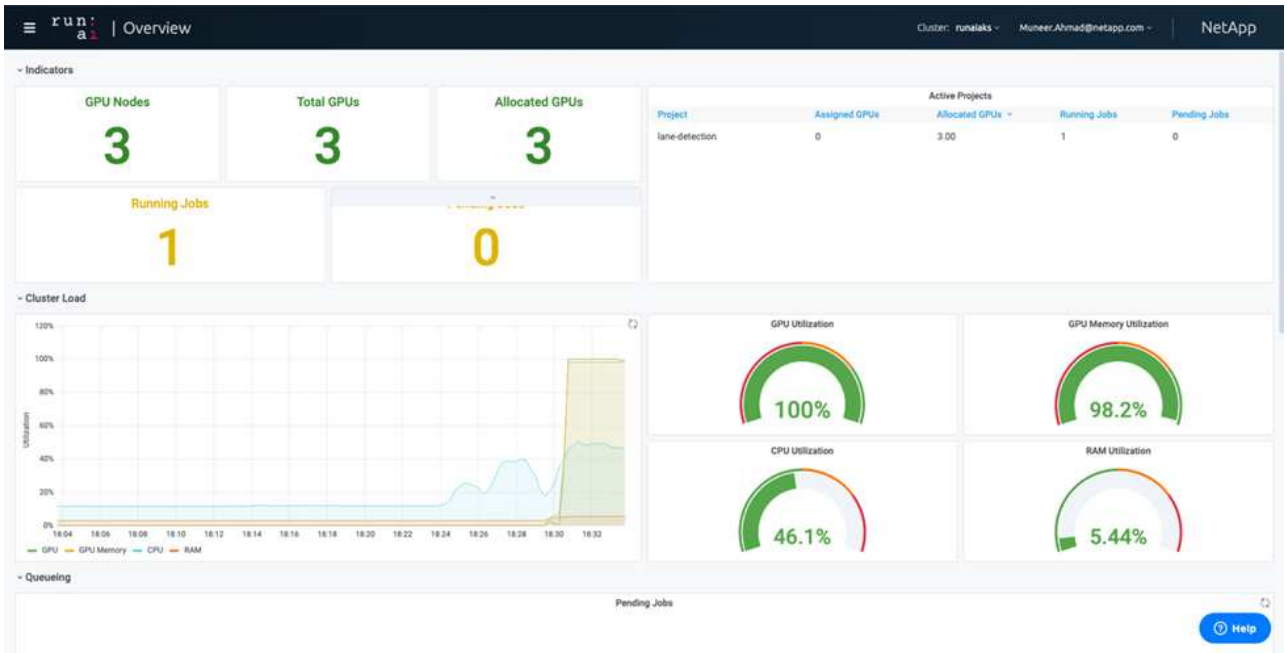
```
NAME                          STATUS     AGE  NODE        IMAGE                              TYPE   PROJECT         USER           GPUs Allocated (Requested)  PODs
 SERVICE URL(S)
download-tusimple-data        Succeeded  1d                muneer7589/download-tusimple:1.0   Train  lane-detection  verronmartina  - (0)                       0 (0)

dist-lane-detection-training  Init:0/1   2m   <multiple>  muneer7589/dist-lane-detection:3.1 Train  lane-detection  root           3 (3)                       4 (0)
```

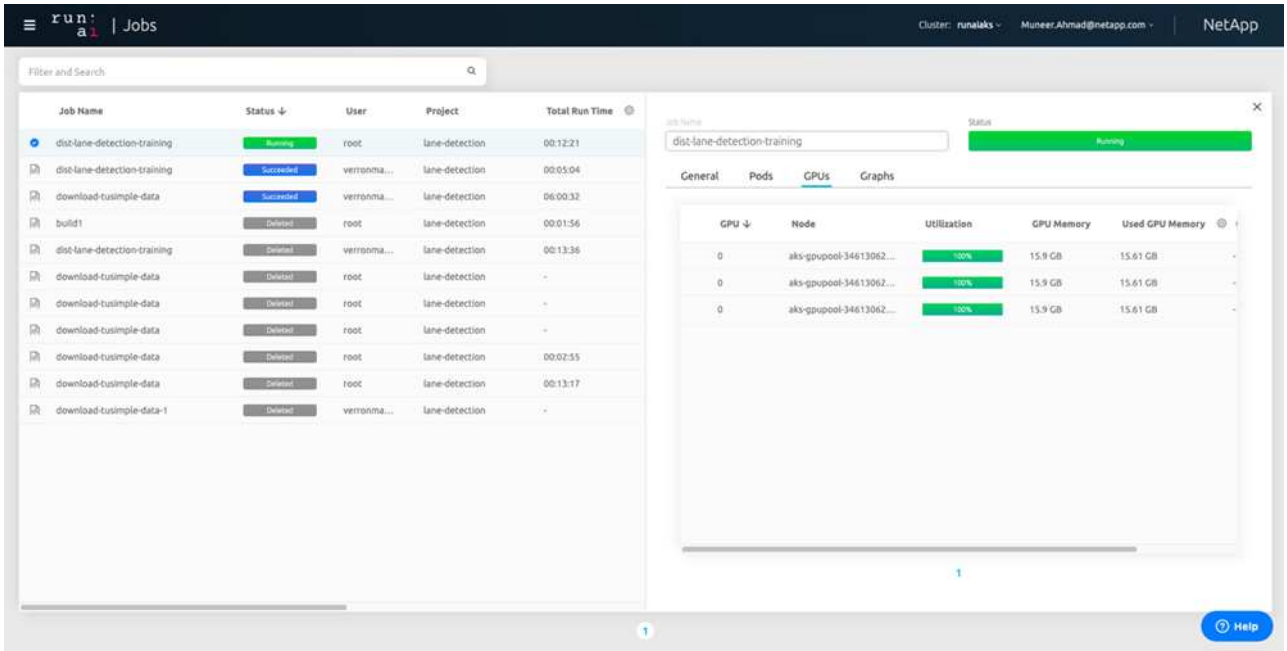
f. Log dei lavori inoltrati:

```
runai logs dist-lane-detection-training
```

```
root@ai-w-gpu-2:~/runai# runai logs dist-lane-detection-training
Running with 3 workers
2021-03-04 17:29:23.158449: I tensorflow/stream_executor/platform/default/dso_loader.cc:48] Successfully opened dynamic library libcudart.so.10.1
+ POD_NAME=dist-lane-detection-training-worker-0
+ [ d = - ]
+ shift
+ /opt/kube/kubectl cp /opt/kube/hosts dist-lane-detection-training-worker-0:/etc/hosts_of_nodes
+ POD_NAME=dist-lane-detection-training-worker-2
+ [ d = - ]
+ shift
+ /opt/kube/kubectl cp /opt/kube/hosts dist-lane-detection-training-worker-2:/etc/hosts_of_nodes
+ POD_NAME=dist-lane-detection-training-worker-1
```

g. Controllare il lavoro di training in CORSO: Ai GUI (o app.runai.ai): RUN: Ai Dashboard, come mostrato nelle figure seguenti. La prima figura descrive in dettaglio tre GPU allocate per il lavoro di training distribuito su tre nodi su AKS e la seconda ESECUZIONE:job ai:





h. Al termine del training, controlla la copia Snapshot di NetApp creata e collegata al lavoro RUN: Ai.

```
runai logs dist-lane-detection-training --tail 1
```

```
[1,0]<stdout>:Snapshot snap-pvc-download-tusimple-data-0-dist-lane-detection-training-launcher-2021-03-05-16-23-42 created in namespace runai-lane-detection
```

```
kubectl get volumesnapshots | grep download-tusimple-data-0
```

Ripristinare i dati dalla copia Snapshot di NetApp

Per ripristinare i dati dalla copia Snapshot di NetApp, attenersi alla seguente procedura:

1. Passare alla home directory.

```
cd ~
```

2. Accedere alla directory del progetto `lane-detection-SCNN-horovod`.

```
cd ./lane-detection-SCNN-horovod
```

3. Modificare `restore-snaphot-pvc.yaml` e aggiornare `dataSource name` Nella copia Snapshot da cui si desidera ripristinare i dati. È anche possibile modificare il nome PVC in cui verranno ripristinati i dati, in questo esempio ITS `restored-tusimple`.

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: restored-tusimple
spec:
  storageClassName: azurenetappfiles
  dataSource:
    name: snap-pvc-download-tusimple-data-0-dist-lane-detection-training-launcher-2021-03-05-16-23-42
    kind: VolumeSnapshot
    apiGroup: snapshot.storage.k8s.io
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 100Gi
```

4. Creare un nuovo PVC utilizzando `restore-snapshot-pvc.yaml`.

```
kubectl create -f restore-snapshot-pvc.yaml
```

L'output dovrebbe essere simile al seguente esempio:

```
persistentvolumeclaim/restored-tusimple created
```

5. Se si desidera utilizzare i dati appena ripristinati per la formazione, l'invio del lavoro rimane lo stesso di prima; sostituire solo `PVC_NAME` con il ripristinato `PVC_NAME` quando si invia il lavoro di formazione, come indicato nei seguenti comandi:

```
runai submit-mpi
--name dist-lane-detection-training
--large-shm
--processes=3
--gpu 1
--pvc restored-tusimple:/mnt
--image muneer7589/dist-lane-detection:3.1
-e USE_WORKERS="true"
-e NUM_WORKERS=4
-e BATCH_SIZE=33
-e USE_VAL="false"
-e VAL_BATCH_SIZE=99
-e ENABLE_SNAPSHOT="true"
-e PVC_NAME="restored-tusimple"
```

Valutazione delle performance

Per mostrare la scalabilità lineare della soluzione, sono stati eseguiti test delle performance per due scenari: Una GPU e tre GPU. L'allocazione della GPU, l'utilizzo della GPU e della memoria, diverse metriche a nodo singolo e a tre nodi sono state acquisite durante il training sul set di dati di rilevamento della corsia TuSimple. I dati vengono aumentati di cinque volte solo per analizzare l'utilizzo delle risorse durante i processi di training.

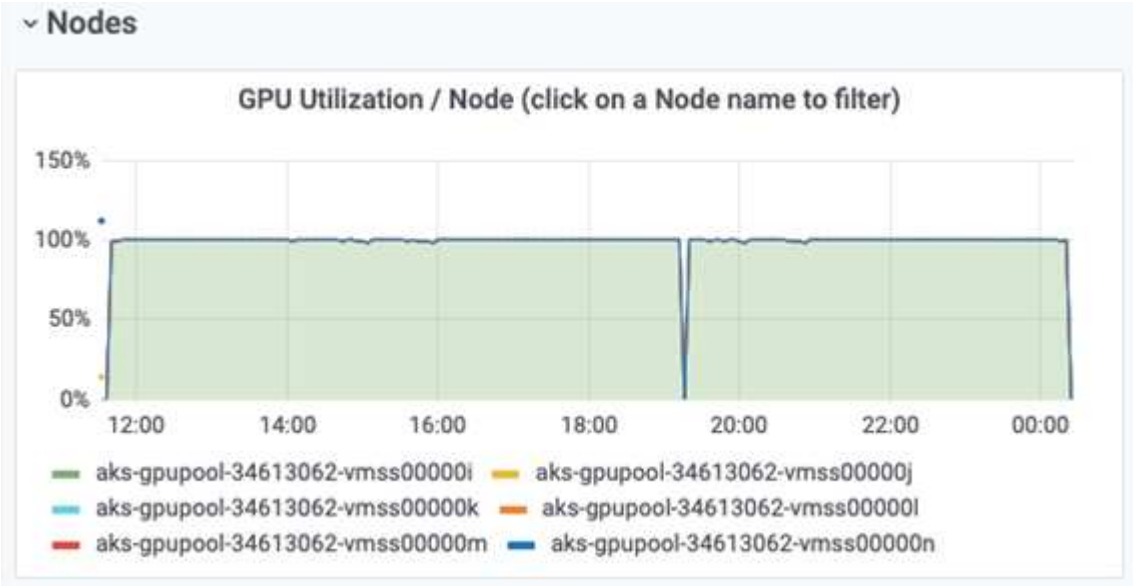
La soluzione consente ai clienti di iniziare con un piccolo set di dati e poche GPU. Quando la quantità di dati e la domanda di GPU aumentano, i clienti possono scalare dinamicamente i terabyte nel Tier Standard e scalare rapidamente fino al Tier Premium per ottenere un throughput quattro volte superiore per terabyte senza spostare alcun dato. Questo processo viene spiegato ulteriormente nella sezione, "Livelli di servizio Azure NetApp Files".

Il tempo di elaborazione su una GPU era di 12 ore e 45 minuti. Il tempo di elaborazione su tre GPU su tre nodi era di circa 4 ore e 30 minuti.

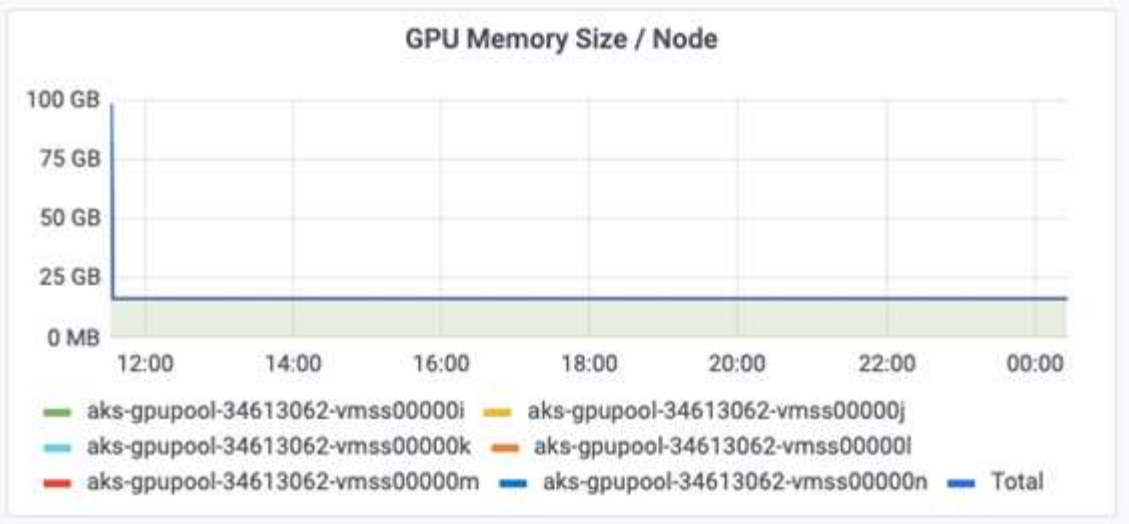
Le figure mostrate nel resto di questo documento illustrano esempi di performance e scalabilità in base alle singole esigenze aziendali.

La figura seguente illustra l'allocazione e l'utilizzo della memoria di 1 GPU.

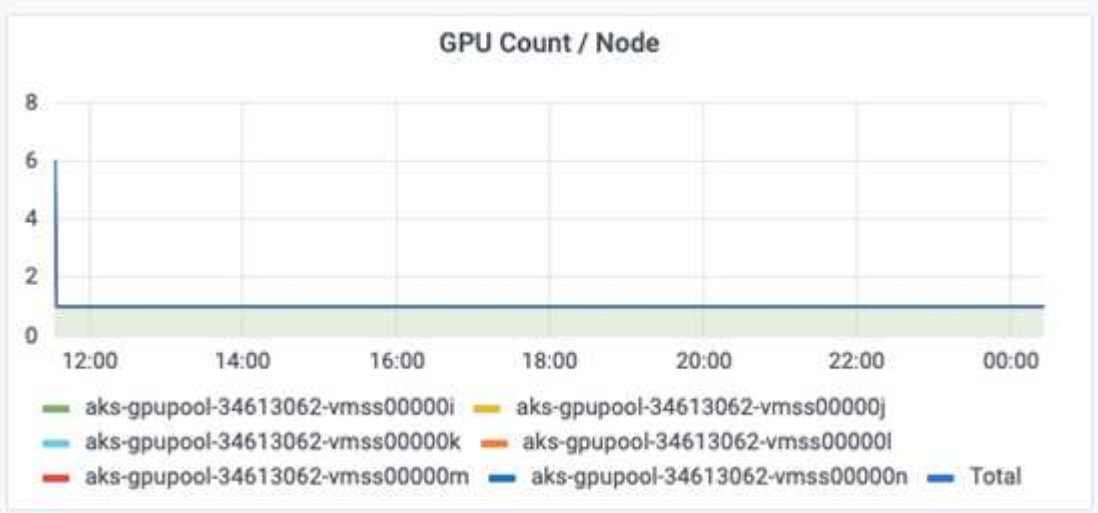
La figura seguente illustra l'utilizzo della GPU a nodo singolo.



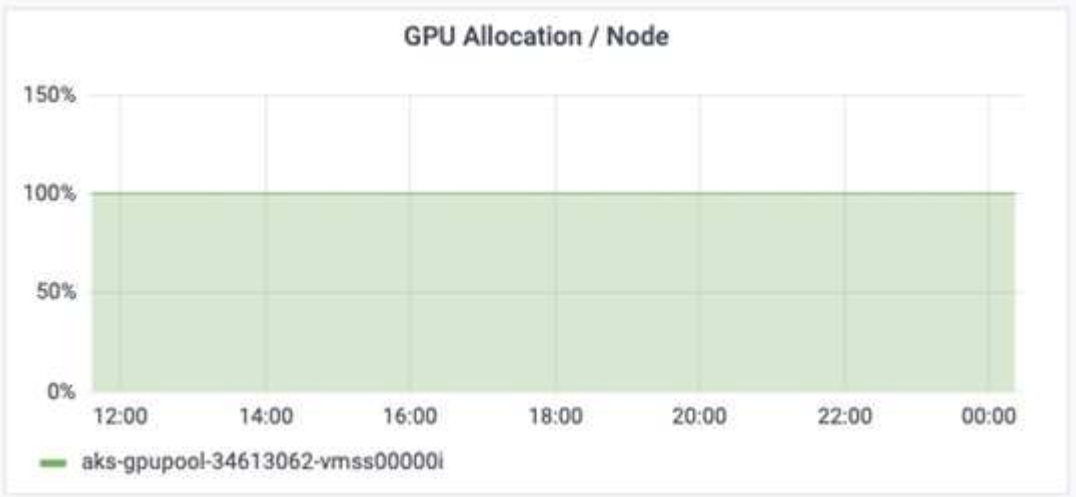
La figura seguente illustra le dimensioni della memoria a nodo singolo (16 GB).



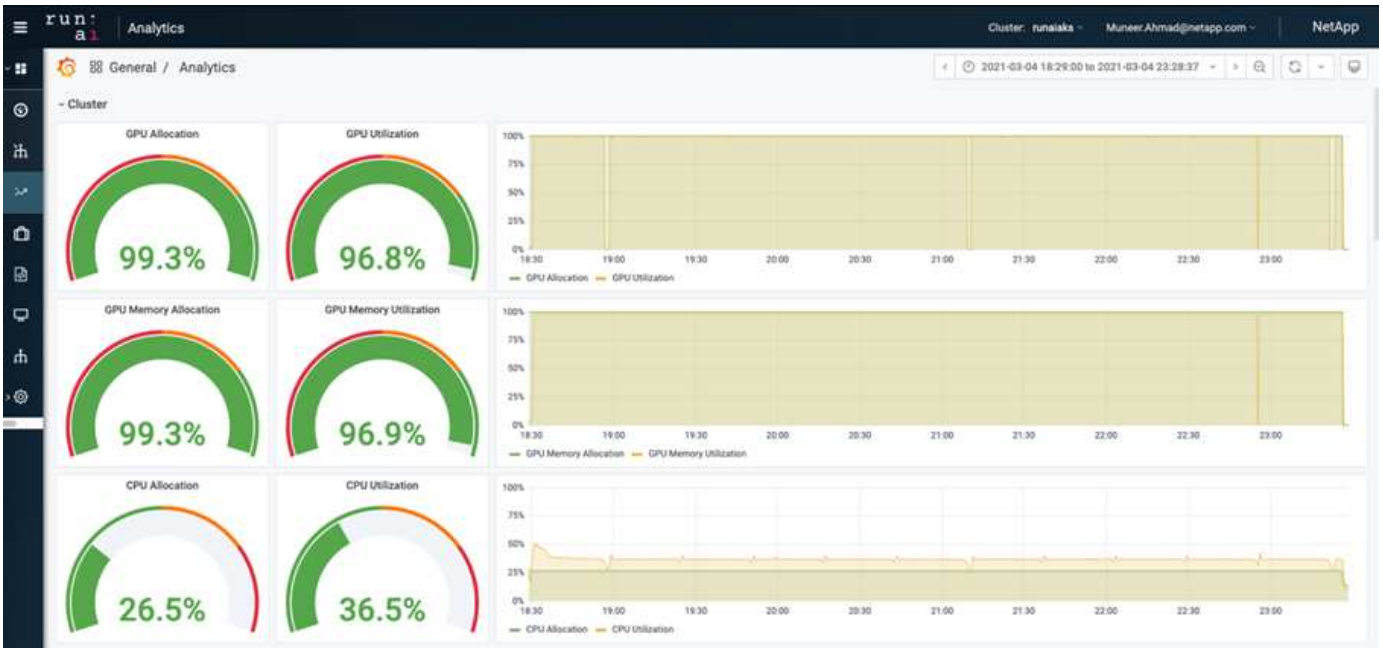
La figura seguente illustra il numero di GPU a nodo singolo (1).



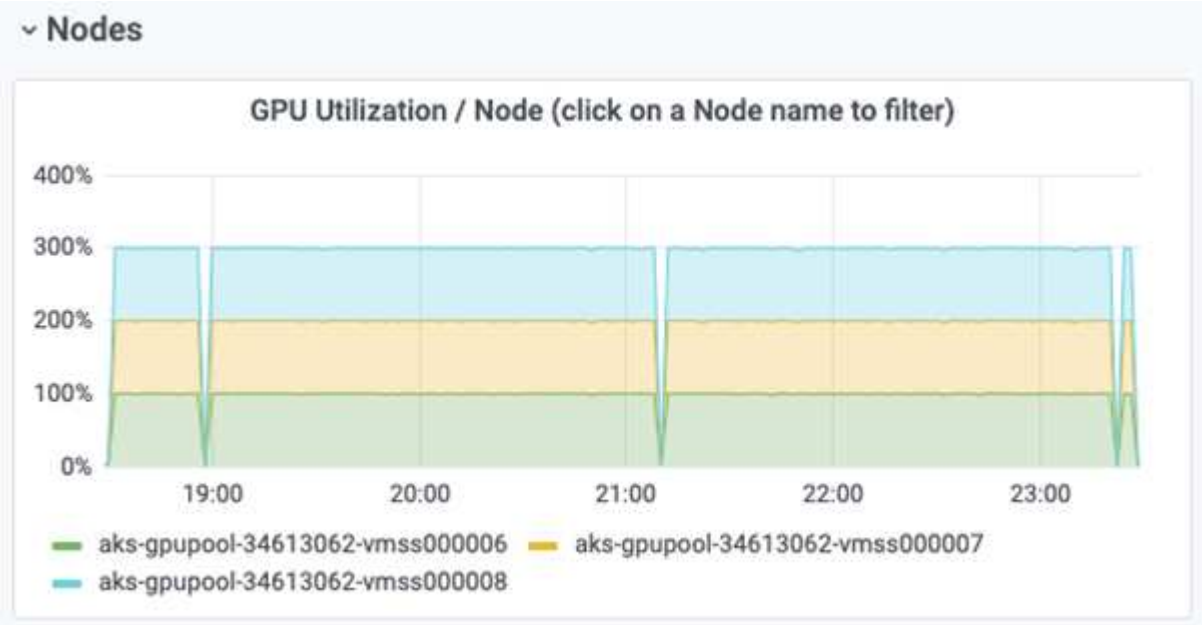
La figura seguente illustra l'allocazione della GPU a nodo singolo (%).



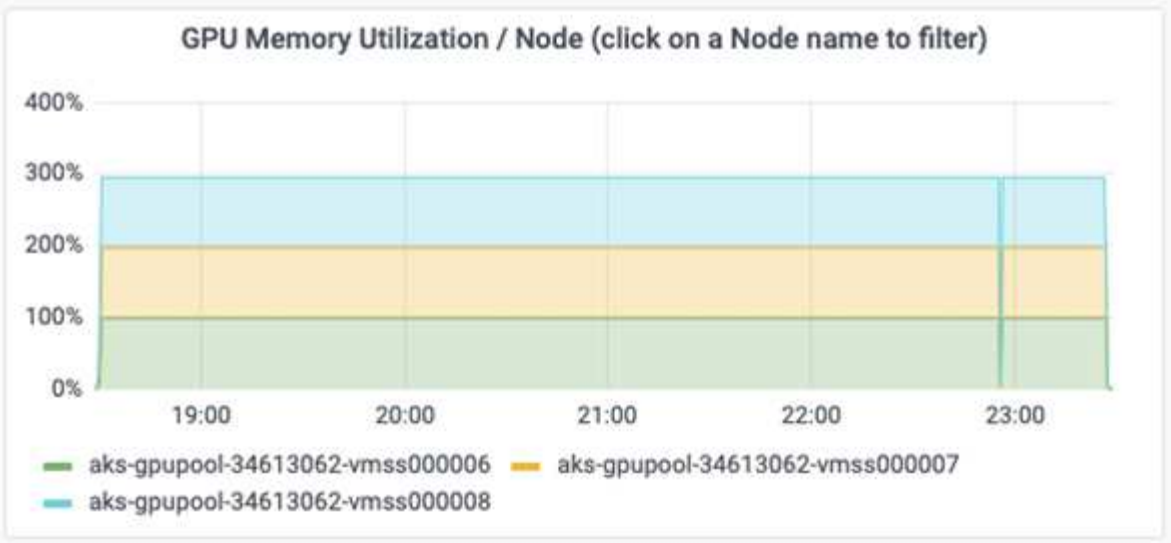
La figura seguente illustra tre GPU su tre nodi: Allocazione e memoria delle GPU.



La figura seguente illustra tre GPU in tre nodi utilizzati (%).



La figura seguente illustra tre GPU in tre nodi di utilizzo della memoria (%).



Livelli di servizio Azure NetApp Files

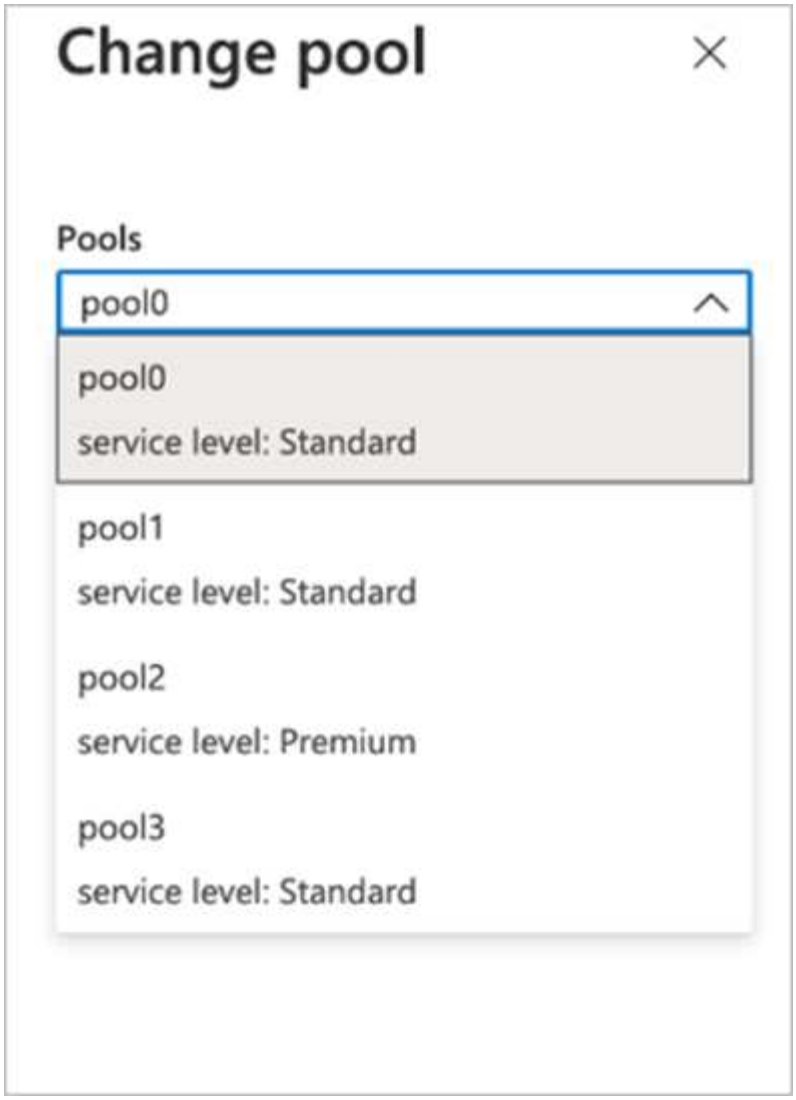
È possibile modificare il livello di servizio di un volume esistente spostando il volume in un altro pool di capacità che utilizza "livello di servizio" si desidera per il volume. Questa modifica del livello di servizio esistente per il volume non richiede la migrazione dei dati. Inoltre, non influisce sull'accesso al volume.

Modificare dinamicamente il livello di servizio di un volume

Per modificare il livello di servizio di un volume, attenersi alla seguente procedura:

1. Nella pagina Volumes (volumi), fare clic con il pulsante destro del mouse sul volume di cui si desidera modificare il livello di servizio. Selezionare Cambia pool.

2. Nella finestra Change Pool, selezionare il pool di capacità in cui si desidera spostare il volume. Quindi, fare clic su OK.



Automatizzare la modifica del livello di servizio

La modifica dinamica del livello di servizio è ancora in Public Preview, ma non è attivata per impostazione predefinita. Per attivare questa funzione nell'abbonamento Azure, seguire la procedura descritta nel documento " "Modificare dinamicamente il livello di servizio di un volume"."

- Per Azure è inoltre possibile utilizzare i seguenti comandi: CLI. Per ulteriori informazioni su come modificare le dimensioni del pool di Azure NetApp Files, visitare il sito "Volume netappfiles az: Gestione delle risorse dei volumi ANF (Azure NetApp Files)".

```
az netappfiles volume pool-change -g mygroup
--account-name myaccname
-pool-name mypoolname
--name myvolname
--new-pool-resource-id mynewresourceid
```

- Il `set- aznetappfilesvolumepool` Il cmdlet illustrato può modificare il pool di un volume Azure NetApp Files. Per ulteriori informazioni sulla modifica delle dimensioni del pool di volumi e di Azure PowerShell, visitare il sito Web "Modifica del pool per un volume Azure NetApp Files".

```
Set-AzNetAppFilesVolumePool
-ResourceGroupName "MyRG"
-AccountName "MyAnfAccount"
-PoolName "MyAnfPool"
-Name "MyAnfVolume"
-NewPoolResourceId 7d6e4069-6c78-6c61-7bf6-c60968e45fbf
```

Conclusione

NetApp e RUN: L'ai ha collaborato alla creazione di questo report tecnico per dimostrare le funzionalità uniche di Azure NetApp Files insieme alla piattaforma RUN: Ai per semplificare l'orchestrazione dei carichi di lavoro ai. Questo report tecnico fornisce un'architettura di riferimento per semplificare il processo di pipeline di dati e orchestrazione dei carichi di lavoro per il training di rilevamento della corsia distribuita.

In conclusione, per quanto riguarda la formazione distribuita su larga scala (soprattutto in un ambiente di cloud pubblico), il componente di orchestrazione delle risorse e storage è una parte critica della soluzione. Assicurarsi che la gestione dei dati non ostacoli mai l'elaborazione di più GPU, per cui si ottiene un utilizzo ottimale dei cicli GPU. Pertanto, rendendo il sistema il più conveniente possibile per scopi di formazione distribuita su larga scala.

Il data fabric fornito da NetApp supera la sfida consentendo a data scientist e data engineer di connettersi tra loro on-premise e nel cloud per avere dati sincroni, senza eseguire alcun intervento manuale. In altre parole, il data fabric rende più uniforme il processo di gestione del workflow ai distribuito in più sedi. Inoltre, facilita la disponibilità dei dati on-demand avvicinando i dati al calcolo ed eseguendo analisi, training e validazione, dove e quando necessario. Questa funzionalità non solo consente l'integrazione dei dati, ma anche la protezione e la sicurezza dell'intera pipeline di dati.

Ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i seguenti documenti e/o siti Web:

- Dataset: TuSimple

  "https://github.com/TuSimple/tusimple-benchmark/tree/master/doc/lane_detection"

- Deep Learning Network Architecture: Rete neurale spaziale convoluzionale

  "https://arxiv.org/abs/1712.06080"

- Framework distribuito per il deep learning: Horovod

  "https://horovod.ai/"

- ESEGUI: Soluzione di orchestrazione dei container ai: ESEGUI: Introduzione al prodotto ai

  "https://docs.run.ai/home/components/"

- ESEGUI: Documentazione di installazione ai

  "https://docs.run.ai/Administrator/Cluster-Setup/cluster-install/#step-3-install-runai"
  "https://docs.run.ai/Administrator/Researcher-Setup/cli-install/#runai-cli-installation"

- Invio di job in ESECUZIONE: Ai CLI

  "https://docs.run.ai/Researcher/cli-reference/runai-submit/"

  "https://docs.run.ai/Researcher/cli-reference/runai-submit-mpi/"

- Risorse cloud di Azure: Azure NetApp Files

  "https://docs.microsoft.com/azure/azure-netapp-files/"

- Servizio Azure Kubernetes

  "https://azure.microsoft.com/services/kubernetes-service/-features"

- SKU di Azure VM

  "https://azure.microsoft.com/services/virtual-machines/"

- Macchine virtuali Azure con SKU GPU

  "https://docs.microsoft.com/azure/virtual-machines/sizes-gpu"

- Trident di NetApp

  "https://github.com/NetApp/trident/releases"

- Data fabric basato su NetApp

  "https://www.netapp.com/data-fabric/what-is-data-fabric/"

- Documentazione sui prodotti NetApp

  "https://www.netapp.com/support-and-training/documentation/"

TR-4841: Sistema operativo ai per il cloud ibrido con caching dei dati

Rick Huang, David Arnette, NetApp Yochay Ettun, cnvrg.io

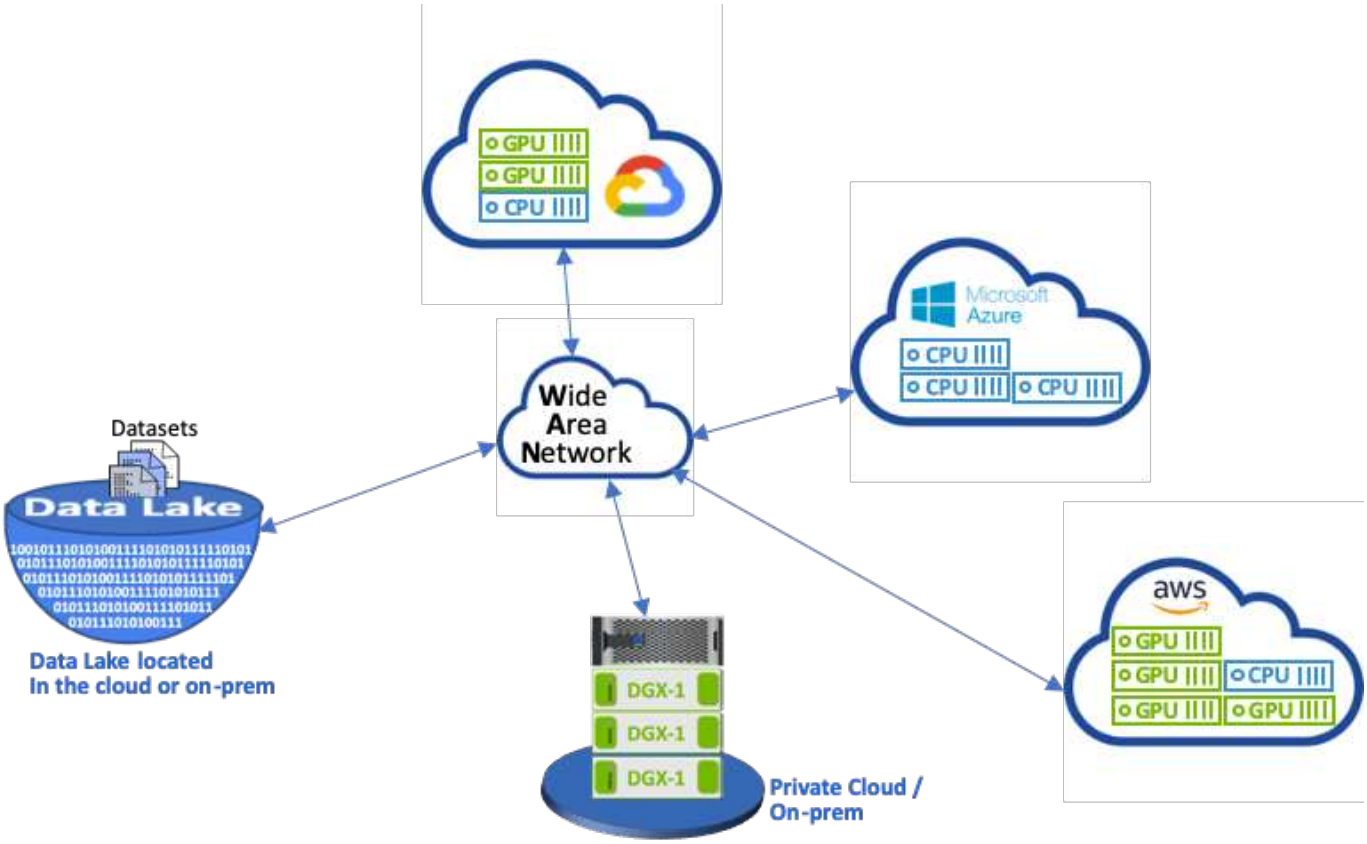
La crescita esplosiva dei dati e la crescita esponenziale di ML e ai sono convergenti per creare un'economia di zettabyte con sfide di sviluppo e implementazione uniche.

Sebbene sia noto che i modelli ML sono affamati di dati e richiedono uno storage dei dati ad alte performance prossimale alle risorse di calcolo, in pratica non è così semplice implementare questo modello, soprattutto con il cloud ibrido e le istanze di calcolo elastiche. In genere, enormi quantità di dati vengono memorizzate in data Lake a basso costo, dove le risorse di calcolo ai dalle performance elevate, come le GPU, non possono accedervi in modo efficiente. Questo problema è aggravato in un'infrastruttura di cloud ibrido in cui alcuni carichi di lavoro operano nel cloud e alcuni si trovano on-premise o in un ambiente HPC completamente diverso.

In questo documento, presentiamo una nuova soluzione che consente ai professionisti IT e ai data engineer di creare una piattaforma di cloud ai realmente ibrido con un data hub consapevole della topologia che consente ai data scientist di creare istantaneamente e automaticamente una cache dei propri set di dati in prossimità delle proprie risorse di calcolo. ovunque si trovino. Di conseguenza, non solo è possibile ottenere un training con modelli ad alte performance, ma si creano anche ulteriori benefici, tra cui la collaborazione di diversi professionisti dell'ai, che hanno accesso immediato a cache, versioni e linee di dati all'interno di un hub di versione del set di dati.

Panoramica del caso d'utilizzo e Problem Statement

I set di dati e le versioni dei set di dati si trovano in genere in un data Lake, come lo storage basato su oggetti NetApp StorageGRID, che offre costi ridotti e altri vantaggi operativi. Gli scienziati dei dati estraggono questi set di dati e li progettano in più fasi per prepararli alla formazione con un modello specifico, spesso creando più versioni lungo il percorso. Come fase successiva, il data scientist deve scegliere risorse di calcolo ottimizzate (GPU, istanze di CPU high-end, un cluster on-premise e così via) per eseguire il modello. La figura seguente mostra la mancanza di prossimità del dataset in un ambiente di calcolo ML.



Tuttavia, è necessario eseguire più esperimenti di training in parallelo in diversi ambienti di calcolo, ciascuno dei quali richiede il download del dataset dal data Lake, un processo costoso e lungo. La prossimità del set di dati all'ambiente di calcolo (in particolare per un cloud ibrido) non è garantita. Inoltre, gli altri membri del team che eseguono i propri esperimenti con lo stesso set di dati devono eseguire lo stesso arduo processo. Al di là dell'evidente rallentamento dell'accesso ai dati, le sfide includono il monitoraggio delle versioni dei set di dati, la condivisione dei set di dati, la collaborazione e la riproducibilità.

Requisiti del cliente

I requisiti dei clienti possono variare per ottenere esecuzioni ML dalle performance elevate utilizzando le risorse in modo efficiente; ad esempio, i clienti potrebbero richiedere quanto segue:

- Accesso rapido ai set di dati da ogni istanza di calcolo che esegue il modello di training senza incorrere in costose complessità di download e accesso ai dati
- L'utilizzo di qualsiasi istanza di calcolo (GPU o CPU) nel cloud o on-premise senza preoccuparsi della posizione dei set di dati
- Maggiore efficienza e produttività grazie all'esecuzione di più esperimenti di training in parallelo con diverse risorse di calcolo sullo stesso set di dati senza ritardi e latenza dei dati non necessari
- Costi delle istanze di calcolo ridotti al minimo
- Riproducibilità migliorata grazie a tool per la conservazione dei record dei set di dati, della loro discendenza, delle versioni e di altri dettagli sui metadati
- Condivisione e collaborazione migliorate per consentire a qualsiasi membro autorizzato del team di accedere ai set di dati ed eseguire esperimenti

Per implementare il caching dei set di dati con il software per la gestione dei dati NetApp ONTAP, i clienti devono eseguire le seguenti attività:

- Configurare e impostare lo storage NFS più vicino alle risorse di calcolo.
- Determinare il set di dati e la versione da memorizzare nella cache.
- Monitorare la memoria totale impegnata nei set di dati memorizzati nella cache e la quantità di storage NFS disponibile per ulteriori commit di cache (ad esempio, gestione della cache).
- Esaurire i set di dati nella cache se non sono stati utilizzati in un determinato periodo di tempo. L'impostazione predefinita è un giorno; sono disponibili altre opzioni di configurazione.

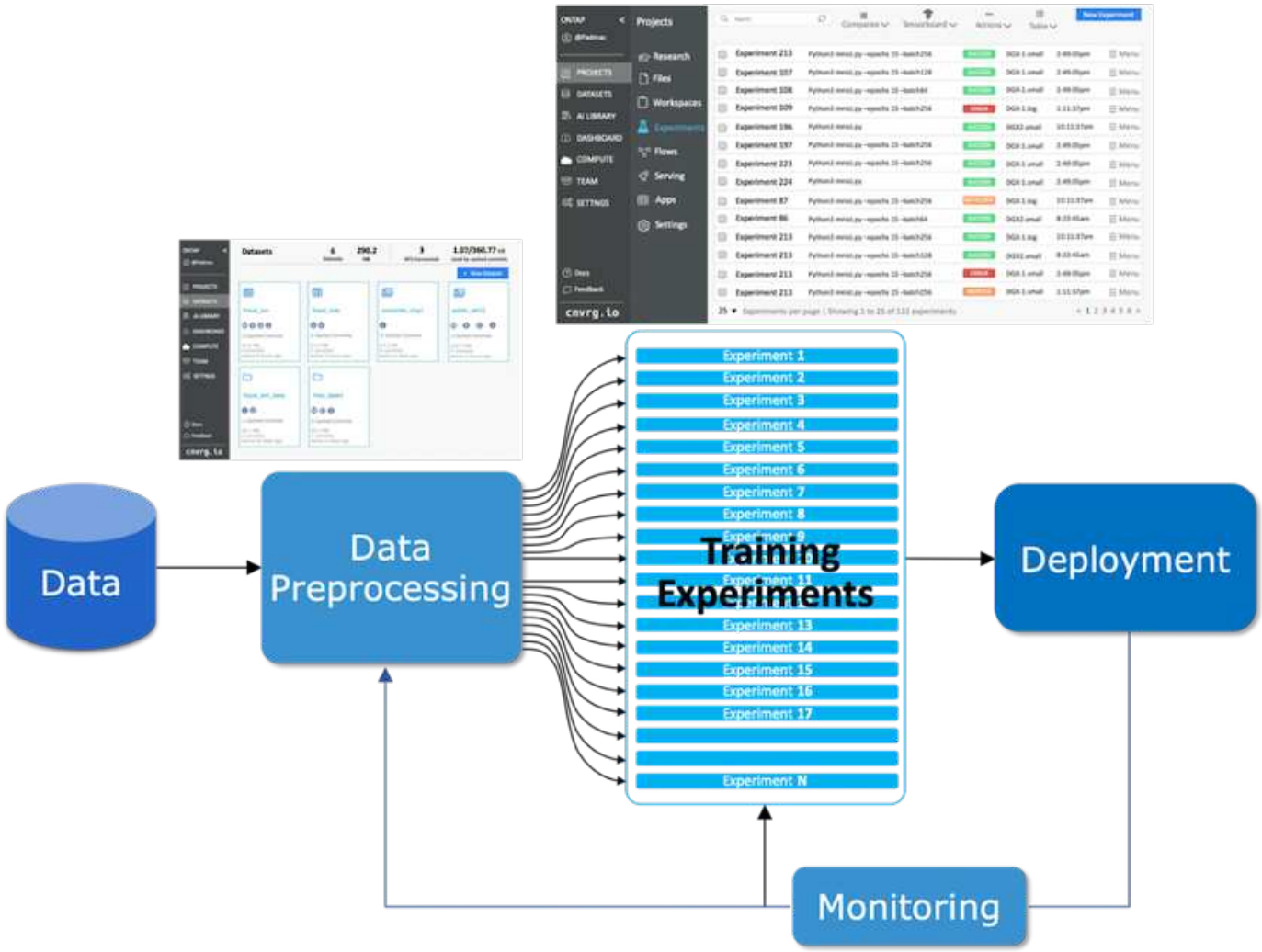
Panoramica della soluzione

In questa sezione viene descritta una pipeline convenzionale per la scienza dei dati e i relativi inconvenienti. Presenta inoltre l'architettura della soluzione di caching dei set di dati proposta.

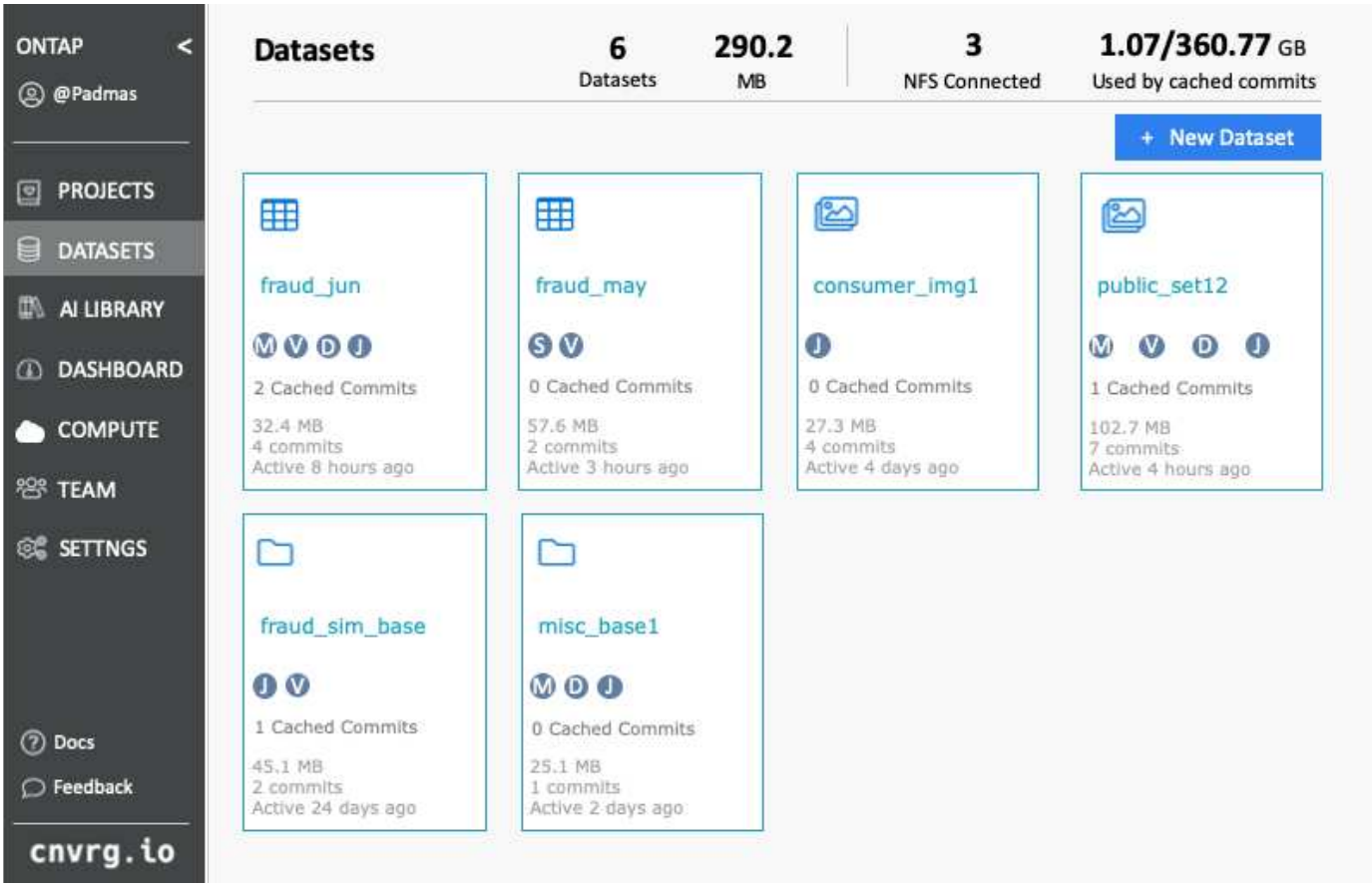
Pipeline e svantaggi convenzionali di Data Science

Una sequenza tipica di sviluppo e implementazione del modello ML prevede passaggi iterativi che includono:

- Acquisizione dei dati
- Pre-elaborazione dei dati (creazione di più versioni dei set di dati)
- Esecuzione di esperimenti multipli che coinvolgono l'ottimizzazione degli hyperparameter, modelli diversi e così via
- Implementazione
- Monitoringcnvrg.io ha sviluppato una piattaforma completa per automatizzare tutte le attività, dalla ricerca all'implementazione. Un piccolo esempio di schermate della dashboard relative alla pipeline è illustrato nella figura seguente.



È molto comune avere più set di dati in gioco da repository pubblici e dati privati. Inoltre, è probabile che ogni set di dati disponga di più versioni risultanti dalla pulizia dei set di dati o dall'ingegneria delle funzionalità. Una dashboard che fornisce un hub di set di dati e una versione hub è necessaria per garantire che i tool di collaborazione e coerenza siano disponibili per il team, come illustrato nella figura seguente.



La fase successiva della pipeline è la formazione, che richiede più istanze parallele di modelli di training, ciascuna associata a un dataset e a una determinata istanza di calcolo. L'associazione di un dataset a un certo esperimento con una determinata istanza di calcolo è una sfida perché è possibile che alcuni esperimenti vengano eseguiti da istanze GPU da Amazon Web Services (AWS), mentre altri esperimenti vengono eseguiti da istanze DGX-1 o DGX-2 on-premise. Altri esperimenti potrebbero essere eseguiti nei server CPU in GCP, mentre la posizione del set di dati non si trova in prossimità delle risorse di calcolo che eseguono il training. Una vicinanza ragionevole avrebbe una connettività completa a 10 GbE o più a bassa latenza dallo storage del dataset all'istanza di calcolo.

È pratica comune per i data scientist scaricare il set di dati nell'istanza di calcolo che esegue il training ed esegue l'esperimento. Tuttavia, questo approccio può comportare diversi problemi:

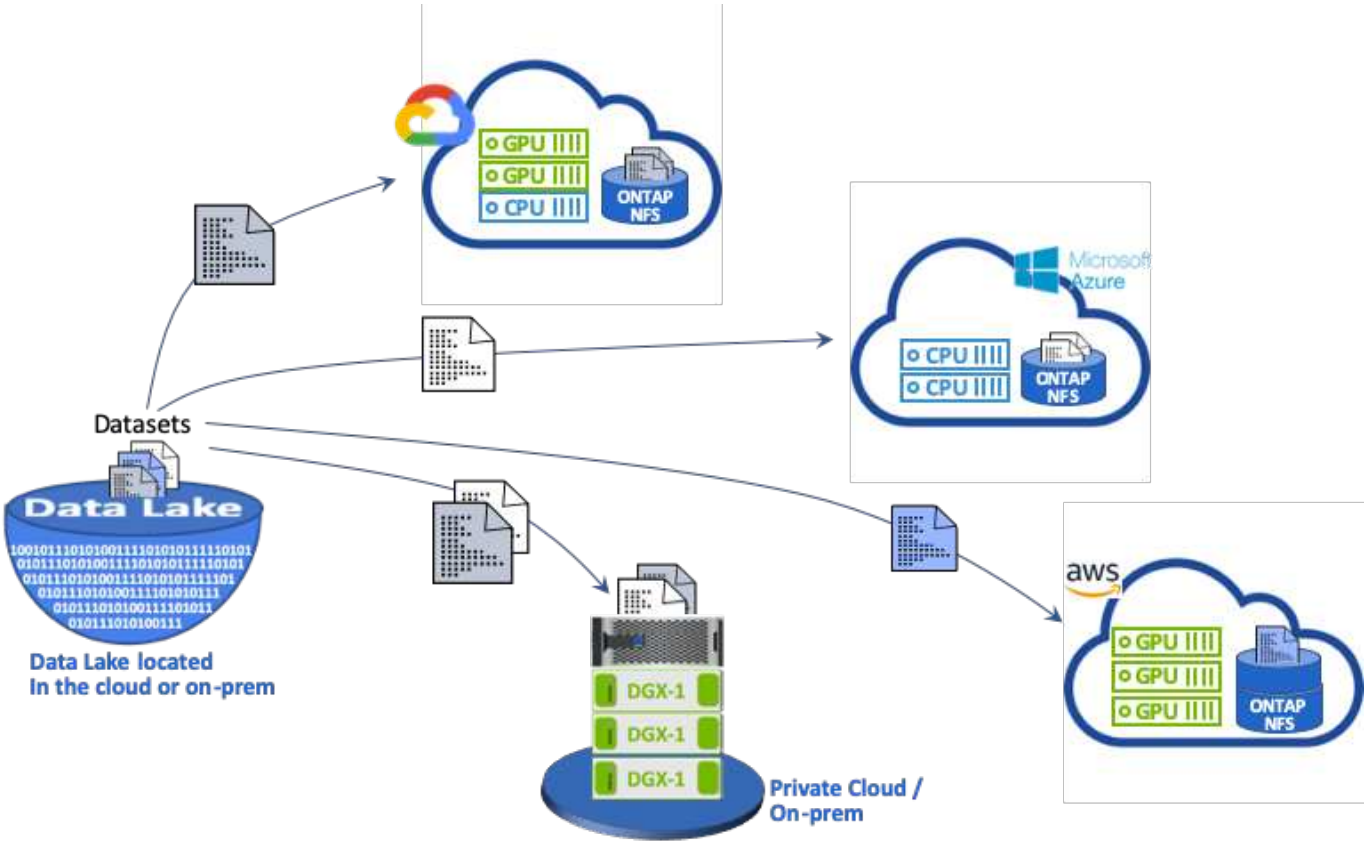
- Quando il data scientist scarica il dataset in un'istanza di calcolo, non vi sono garanzie che lo storage di calcolo integrato sia dalle performance elevate (un esempio di sistema dalle performance elevate sarebbe la soluzione NVMe ONTAP AFF A800).
- Quando il set di dati scaricato risiede in un nodo di calcolo, lo storage può diventare un collo di bottiglia quando i modelli distribuiti vengono eseguiti su più nodi (a differenza dello storage distribuito dalle performance elevate di NetApp ONTAP).
- La successiva iterazione dell'esperimento di training potrebbe essere eseguita in un'istanza di calcolo diversa a causa di conflitti di coda o priorità, creando nuovamente una distanza di rete significativa dal dataset alla posizione di calcolo.
- Gli altri membri del team che eseguono esperimenti di training sullo stesso cluster di calcolo non possono condividere questo set di dati; ciascuno esegue il (costoso) download del set di dati da una posizione arbitraria.
- Se sono necessari altri set di dati o versioni dello stesso set di dati per i successivi lavori di formazione, i data scientist devono eseguire nuovamente il (costoso) download del set di dati nell'istanza di calcolo che

esegue training.NetApp e cnvrg.io hanno creato una nuova soluzione di caching del set di dati che elimina questi ostacoli. La soluzione crea un'esecuzione accelerata della pipeline ML memorizzando nella cache i set di dati hot sul sistema storage ad alte performance ONTAP. Con ONTAP NFS, i set di dati vengono memorizzati nella cache una sola volta (e una sola volta) in un data fabric basato su NetApp (ad esempio AFF A800), che viene posizionato insieme al calcolo. Poiché lo storage NetApp ONTAP NFS ad alta velocità può servire più nodi di calcolo ML, le performance dei modelli di training sono ottimizzate, offrendo risparmi sui costi, produttività ed efficienza operativa all'organizzazione.

Architettura della soluzione

Questa soluzione di NetApp e cnvrg.io fornisce il caching dei set di dati, come mostrato nella figura seguente. Il caching dei set di dati consente agli scienziati dei dati di scegliere una versione di set di dati o set di dati desiderata e di spostarla nella cache NFS di ONTAP, che si trova in prossimità del cluster di calcolo ML. Il data scientist può ora eseguire più esperimenti senza incorrere in ritardi o download. Inoltre, tutti i tecnici che collaborano possono utilizzare lo stesso set di dati con il cluster di calcolo collegato (con la libertà di scegliere qualsiasi nodo) senza ulteriori download dal data Lake. Ai data scientist viene offerta una dashboard che tiene traccia e monitora tutti i set di dati e le versioni e fornisce una vista dei set di dati memorizzati nella cache.

La piattaforma cnvrg.io rileva automaticamente i set di dati vecchi che non sono stati utilizzati per un certo periodo di tempo e li eludono dalla cache, mantenendo spazio libero nella cache NFS per i set di dati più utilizzati. È importante notare che il caching dei set di dati con ONTAP funziona nel cloud e on-premise, fornendo così la massima flessibilità.



Concetti e componenti

In questa sezione vengono illustrati i concetti e i componenti associati al caching dei dati in un workflow ML.

Apprendimento automatico

ML sta diventando rapidamente essenziale per molte aziende e organizzazioni in tutto il mondo. Pertanto, I team IT e DevOps devono ora affrontare la sfida della standardizzazione dei carichi DI lavoro ML e del provisioning di cloud, risorse di calcolo on-premise e ibride che supportano i flussi di lavoro dinamici e intensivi richiesti dai processi E dalle pipeline ML.

Machine Learning e Kubernetes basati su container

I container sono istanze isolate dello spazio utente eseguite su un kernel del sistema operativo host condiviso. L'adozione dei container è in rapida crescita. I container offrono molti degli stessi vantaggi offerti dalle macchine virtuali (VM) per il sandboxing delle applicazioni. Tuttavia, poiché l'hypervisor e i livelli del sistema operativo guest su cui si basano le macchine virtuali sono stati eliminati, i container sono molto più leggeri.

I container consentono inoltre un efficiente packaging delle dipendenze delle applicazioni, dei tempi di esecuzione e così via direttamente con un'applicazione. Il formato di packaging dei container più comunemente utilizzato è Docker Container. Un'applicazione che è stata containerizzata nel formato Docker container può essere eseguita su qualsiasi computer in grado di eseguire i container Docker. Ciò è vero anche se le dipendenze dell'applicazione non sono presenti sul computer, perché tutte le dipendenze sono contenute nel container stesso. Per ulteriori informazioni, visitare il "Sito web di Docker".

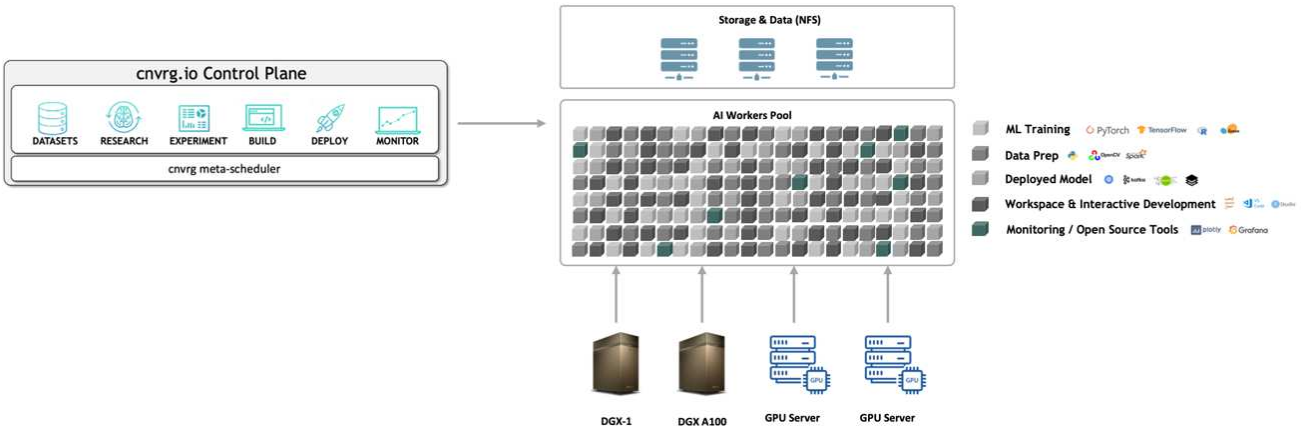
Kubernetes, il popolare container orchestrator, consente agli scienziati dei dati di lanciare processi e pipeline flessibili e basati su container. Consente inoltre ai team dell'infrastruttura di gestire e monitorare i carichi DI lavoro ML in un singolo ambiente gestito e nativo del cloud. Per ulteriori informazioni, visitare il "Sito web di Kubernetes".

cnvrg.io

Cnvrg.io è un sistema operativo ai che trasforma il modo in cui le aziende gestiscono, scalano e accelerano l'ai e lo sviluppo di data science dalla ricerca alla produzione. La piattaforma code-first è costruita dai data scientist per i data scientist e offre flessibilità per l'esecuzione on-premise o nel cloud. Grazie alla gestione dei modelli, agli MLOps e alle soluzioni DI ML continuo, cnvrg.io porta la tecnologia top di gamma ai team di data science in modo che possano dedicare meno tempo a DevOps e concentrarsi sugli algoritmi più magici. Da quando si utilizza cnvrg.io, i team di diversi settori hanno ottenuto più modelli in produzione, con un conseguente aumento del valore di business.

Meta-Scheduler cnvrg.io

cnvrg. io ha un'architettura unica che consente A IT e ingegneri di collegare diverse risorse di calcolo allo stesso piano di controllo e di fare in modo che cnvrg.io gestisca i lavori ML in tutte le risorse. Ciò significa che può collegare più cluster Kubernetes on-premise, server VM e account cloud ed eseguire carichi DI lavoro ML su tutte le risorse, come mostrato nella figura seguente.



Caching dei dati cnvrg.io

cnvrg.io consente agli scienziati dei dati di definire le versioni dei set di dati hot e cold con la sua tecnologia di caching dei dati. Per impostazione predefinita, i set di dati vengono memorizzati in un database di storage a oggetti centralizzato. Quindi, i data scientist possono memorizzare nella cache una versione specifica dei dati sulla risorsa di calcolo selezionata per risparmiare tempo durante il download e quindi aumentare LO sviluppo E la produttività DI ML. I set di dati memorizzati nella cache e non utilizzati per alcuni giorni vengono cancellati automaticamente dal NFS selezionato. Il caching e la cancellazione della cache possono essere eseguiti con un singolo clic; non sono richiesti né processi di codifica, NÉ OPERAZIONI IT o DevOps.

Flussi cnvrg.io e PIPELINE ML

Cnvrg.io Flows è uno strumento per la creazione di pipeline ML di produzione. Ogni componente di un flusso è uno script/codice in esecuzione su un calcolo selezionato con un'immagine di base del docker. Questo design consente a data scientist e ingegneri di creare una singola pipeline che può essere eseguita sia on-premise che nel cloud. cnvrg.io garantisce lo spostamento di dati, parametri e artefatti tra i diversi componenti. Inoltre, ogni flusso viene monitorato e monitorato per ottenere una data science riproducibile al 100%.

CORE cnvrg.io

Cnvrg.io CORE è una piattaforma gratuita per la community di data science per aiutare i data scientist a concentrarsi maggiormente sulla data science e meno su DevOps. L'infrastruttura flessibile DI CORE offre ai data scientist il controllo di utilizzare qualsiasi linguaggio, framework ai o ambiente di calcolo, sia on-premise che nel cloud, in modo che possano fare ciò che fanno meglio, costruire algoritmi. Il CORE cnvrg.io può essere facilmente installato con un singolo comando su qualsiasi cluster Kubernetes.

NetApp ONTAP ai

ONTAP ai è un'architettura di riferimento per data center per i carichi di lavoro ML e deep learning (DL) che utilizza i sistemi storage NetApp AFF e i sistemi NVIDIA DGX con GPU Tesla V100. ONTAP ai si basa sul protocollo file NFS standard di settore su Ethernet da 100 GB, offrendo ai clienti un'infrastruttura ML/DL dalle performance elevate che utilizza tecnologie standard per data center per ridurre l'overhead di implementazione e amministrazione. L'utilizzo di protocolli e reti standardizzati consente a ONTAP ai di integrarsi in ambienti di cloud ibrido mantenendo al contempo coerenza e semplicità operativa. Come soluzione di infrastruttura pre-validata, ONTAP ai riduce i tempi e i rischi di implementazione e riduce significativamente l'overhead amministrativo, consentendo ai clienti di ottenere un time-to-value più rapido.

NVIDIA DeepOps

DeepOps è un progetto open source di NVIDIA che, utilizzando Ansible, automatizza l'implementazione dei cluster di server GPU in base alle Best practice. DeepOps è modulare e può essere utilizzato per varie attività di implementazione. Per questo documento e per l'esercizio di convalida descritto, DeepOps viene utilizzato per implementare un cluster Kubernetes costituito da nodi di lavoro del server GPU. Per ulteriori informazioni, visitare il "Sito Web di DeepOps".

Trident di NetApp

Trident è un orchestratore di storage open source sviluppato e gestito da NetApp che semplifica notevolmente la creazione, la gestione e il consumo dello storage persistente per i carichi di lavoro Kubernetes. Trident è un'applicazione nativa di Kubernetes che viene eseguita direttamente all'interno di un cluster Kubernetes. Con Trident, gli utenti di Kubernetes (sviluppatori, data scientist, amministratori di Kubernetes e così via) possono creare, gestire e interagire con volumi di storage persistenti nel formato standard di Kubernetes che già conoscono. Allo stesso tempo, possono sfruttare le funzionalità avanzate di gestione dei dati di NetApp e un data fabric basato sulla tecnologia NetApp. Trident astratta le complessità dello storage persistente e lo rende semplice da utilizzare. Per ulteriori informazioni, visitare il "Sito web di Trident".

NetApp StorageGRID

NetApp StorageGRID è una piattaforma di storage a oggetti software-defined progettata per soddisfare queste esigenze fornendo uno storage semplice e simile al cloud a cui gli utenti possono accedere utilizzando il protocollo S3. StorageGRID è un sistema scale-out progettato per supportare più nodi nei siti connessi a Internet, indipendentemente dalla distanza. Con il motore intelligente delle policy di StorageGRID, gli utenti possono scegliere oggetti di erasure coding tra i siti per georesilienza o replica di oggetti tra siti remoti per ridurre al minimo la latenza di accesso WAN. StorageGRID offre un eccellente data Lake di storage a oggetti primario per il cloud privato in questa soluzione.

NetApp Cloud Volumes ONTAP

Il software per la gestione dei dati NetApp Cloud Volumes ONTAP offre controllo, protezione ed efficienza ai dati degli utenti con la flessibilità dei provider di cloud pubblico, tra cui AWS, Google Cloud Platform e Microsoft Azure. Cloud Volumes ONTAP è un software per la gestione dei dati nativo del cloud basato sul software di storage NetApp ONTAP, che offre agli utenti una piattaforma di storage universale di livello superiore in grado di soddisfare le loro esigenze di dati nel cloud. La disponibilità dello stesso software di storage nel cloud e on-premise offre agli utenti il valore di un data fabric senza dover formare il personale IT in nuovi metodi per la gestione dei dati.

Per i clienti interessati ai modelli di implementazione del cloud ibrido, Cloud Volumes ONTAP è in grado di fornire le stesse funzionalità e performance leader di settore nella maggior parte dei cloud pubblici per offrire un'esperienza utente coerente e perfetta in qualsiasi ambiente.

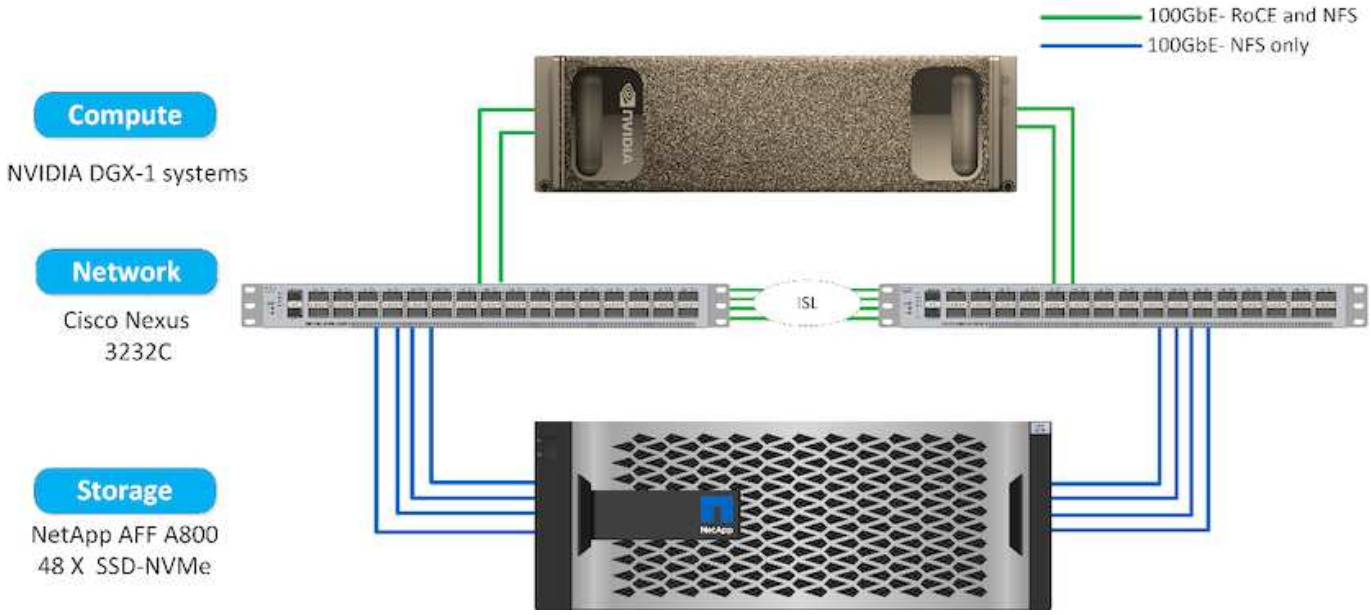
Requisiti hardware e software

In questa sezione vengono illustrati i requisiti tecnologici per la soluzione ai di ONTAP.

Requisiti hardware

Sebbene i requisiti hardware dipendano da carichi di lavoro specifici dei clienti, ONTAP ai può essere implementato su qualsiasi scala per data engineering, formazione sui modelli e deduzione di produzione da una singola GPU fino a configurazioni su scala rack per operazioni ML/DL su larga scala. Per ulteriori informazioni su ONTAP ai, vedere "Sito web ONTAP ai".

Questa soluzione è stata validata utilizzando un sistema DGX-1 per il calcolo, un sistema storage NetApp AFF A800 e Cisco Nexus 3232C per la connettività di rete. Il sistema AFF A800 utilizzato per questa convalida può supportare fino a 10 sistemi DGX-1 per la maggior parte dei carichi di lavoro ML/DL. La figura seguente mostra la topologia ONTAP ai utilizzata per il training sui modelli in questa convalida.



Per estendere questa soluzione a un cloud pubblico, Cloud Volumes ONTAP può essere implementato insieme alle risorse di calcolo della GPU del cloud e integrato in un data fabric del cloud ibrido che consente ai clienti di utilizzare le risorse appropriate per un determinato carico di lavoro.

Requisiti software

La seguente tabella mostra le versioni software specifiche utilizzate per la convalida della soluzione.

Componente	Versione
Ubuntu	18.04.4 LTS
SISTEMA OPERATIVO NVIDIA DGX	4.4.0
NVIDIA DeepOps	20.02.1
Kubernetes	1.15
Timone	3.1.0
cnvrg.io	3.0.0
NetApp ONTAP	9.6P4

Per la convalida di questa soluzione, Kubernetes è stato implementato come cluster a nodo singolo nel sistema DGX-1. Per le implementazioni su larga scala, è necessario implementare nodi master Kubernetes indipendenti per fornire un'elevata disponibilità dei servizi di gestione e riservare preziose risorse DGX per i carichi di lavoro ML e DL.

Dettagli sulla distribuzione e sulla convalida della soluzione

Le sezioni seguenti illustrano i dettagli dell'implementazione e della convalida della soluzione.

Implementazione di ONTAP ai

L'implementazione di ONTAP ai richiede l'installazione e la configurazione dell'hardware di rete, calcolo e storage. Le istruzioni specifiche per l'implementazione dell'infrastruttura ai di ONTAP esulano dall'ambito di questo documento. Per informazioni dettagliate sull'implementazione, vedere "NVA-1121-DEPLOY: NetApp ONTAP ai, basato su NVIDIA".

Per la convalida di questa soluzione, è stato creato un singolo volume e montato sul sistema DGX-1. Tale punto di montaggio è stato quindi montato sui container per rendere i dati accessibili per la formazione. Per implementazioni su larga scala, NetApp Trident automatizza la creazione e il montaggio dei volumi per eliminare i costi amministrativi e consentire la gestione delle risorse da parte dell'utente finale.

Implementazione di Kubernetes

Per implementare e configurare il cluster Kubernetes con NVIDIA DeepOps, eseguire le seguenti operazioni da un host di distribuzione jump:

1. Scaricare NVIDIA DeepOps seguendo le istruzioni sul "Pagina introduttiva" Sul sito NVIDIA DeepOps GitHub.
2. Implementare Kubernetes nel cluster seguendo le istruzioni sul "Guida all'implementazione di Kubernetes" Sul sito NVIDIA DeepOps GitHub.

> Affinché l'implementazione di DeepOps Kubernetes funzioni, lo stesso utente deve esistere su tutti i nodi master e worker di Kubernetes.

Se l'implementazione non riesce, modificare il valore di `kubectl_localhost` a false in `deepops/config/group_vars/k8s-cluster.yml` e ripetere il punto 2. Il `Copy kubectl binary to ansible host` attività, che viene eseguita solo quando il valore di `kubectl_localhost` È vero, si basa sul modulo fetch Ansible, che presenta problemi noti di utilizzo della memoria. A volte, questi problemi di utilizzo della memoria possono causare un errore nell'attività. Se l'operazione non riesce a causa di un problema di memoria, il resto dell'operazione di implementazione non viene completata correttamente.

Se l'implementazione viene completata correttamente dopo aver modificato il valore di `kubectl_localhost` a. `false`, quindi è necessario copiare manualmente `kubectl binary` Da un nodo master Kubernetes all'host di salto per l'implementazione. È possibile trovare la posizione di `kubectl binary` su un nodo master specifico eseguendo `which kubectl` comando direttamente su quel nodo.

Implementazione di cnvrg.io

In questa sezione vengono forniti i dettagli per la distribuzione del CORE cnvrg utilizzando i grafici Helm.

Implementare il CORE cnvrg utilizzando Helm

Helm è il modo più semplice per implementare rapidamente cnvrg utilizzando qualsiasi cluster, on-premise, Minikube o qualsiasi cluster cloud (come AKS, EKS e GKE). Questa sezione descrive come cnvrg è stato installato su un'istanza on-premise (DGX-1) con Kubernetes installato.

Prerequisiti

Prima di completare l'installazione, è necessario installare e preparare le seguenti dipendenze sul computer locale:

- Kubectl
- Timone 3.x
- Kubernetes cluster 1.15+

Implementazione con Helm

1. Per scaricare i grafici di comando più aggiornati, eseguire il seguente comando:

```
helm repo add cnvrg https://helm.cnvrg.io
helm repo update
```

2. Prima di implementare cnvrg, è necessario disporre dell'indirizzo IP esterno del cluster e del nome del nodo su cui verrà implementato cnvrg. Per implementare cnvrg in un cluster Kubernetes on-premise, eseguire il seguente comando:

```
helm install cnvrg cnvrg/cnvrg --timeout 1500s  --wait \ --set
global.external_ip=<ip_of_cluster> \ --set global.node=<name_of_node>
```

3. Eseguire `helm install` comando. Tutti i servizi e i sistemi vengono installati automaticamente sul cluster. Il processo può richiedere fino a 15 minuti.
4. Il `helm install` il comando può richiedere fino a 10 minuti. Una volta completata l'implementazione, accedere all'URL del cnvrg appena distribuito o aggiungere il nuovo cluster come risorsa all'interno dell'organizzazione. Il `helm` Il comando indica l'URL corretto.

```
Thank you for installing cnvrg.io!
Your installation of cnvrg.io is now available, and can be reached via:
Talk to our team via email at
```

5. Quando lo stato di tutti i container è in esecuzione o completo, cnvrg è stato implementato correttamente. Dovrebbe essere simile al seguente output di esempio:

```
NAME                                READY   STATUS      RESTARTS   AGE
cnvrg-app-69fbb9df98-6xrgf                  1/1     Running     0          2m
cnvrg-sidekiq-b9d54d889-5x4fc               1/1     Running     0          2m
controller-65895b47d4-s96v6                 1/1     Running     0          2m
init-app-vs-config-wv9c4                    0/1     Completed   0          9m
init-gateway-vs-config-2zbpp                0/1     Completed   0          9m
init-minio-vs-config-cd2rg                  0/1     Completed   0          9m
minio-0                                     1/1     Running     0          2m
postgres-0                                  1/1     Running     0          2m
redis-695c49c986-kcbt9                      1/1     Running     0          2m
seeder-wh655                                0/1     Completed   0          2m
speaker-5sghr                               1/1     Running     0          2m
```

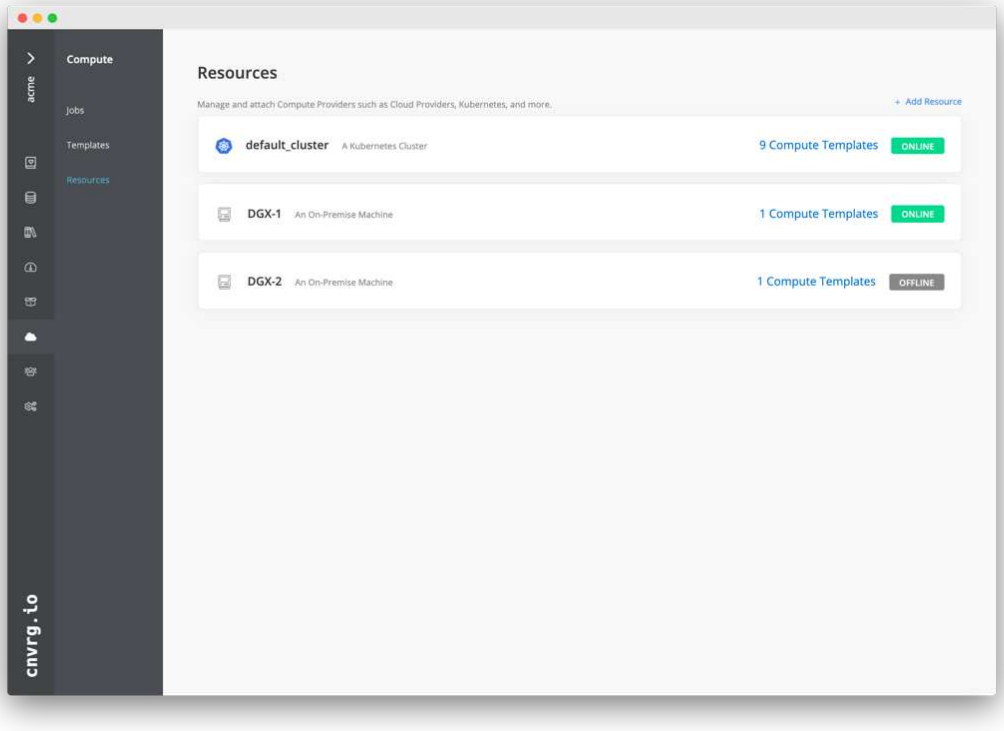
Formazione sul modello di visione artificiale con ResNet50 e il set di dati radiologici Chest

Il sistema operativo ai cnvrg.io è stato implementato su una configurazione Kubernetes su un'architettura NetApp ONTAP ai basata sul sistema NVIDIA DGX. Per la convalida, abbiamo utilizzato il set di dati radiologici NIH Chest costituito da immagini anonimizzate dei raggi X del torace. Le immagini erano in formato PNG. I dati sono stati forniti dal NIH Clinical Center e sono disponibili tramite "Sito di download NIH". Abbiamo utilizzato un campione di 250 GB dei dati con 627, 615 immagini in 15 classi.

Il set di dati è stato caricato sulla piattaforma cnvrg ed è stato memorizzato nella cache di un'esportazione NFS dal sistema di storage NetApp AFF A800.

Impostare le risorse di calcolo

L'architettura cnvrg e la funzionalità di meta-scheduling consentono a tecnici e professionisti IT di collegare diverse risorse di calcolo a una singola piattaforma. Nella nostra configurazione, abbiamo utilizzato lo stesso cluster cnvrg implementato per l'esecuzione dei carichi di lavoro di deep-learning. Se è necessario collegare altri cluster, utilizzare la GUI, come mostrato nella seguente schermata.



Caricare i dati

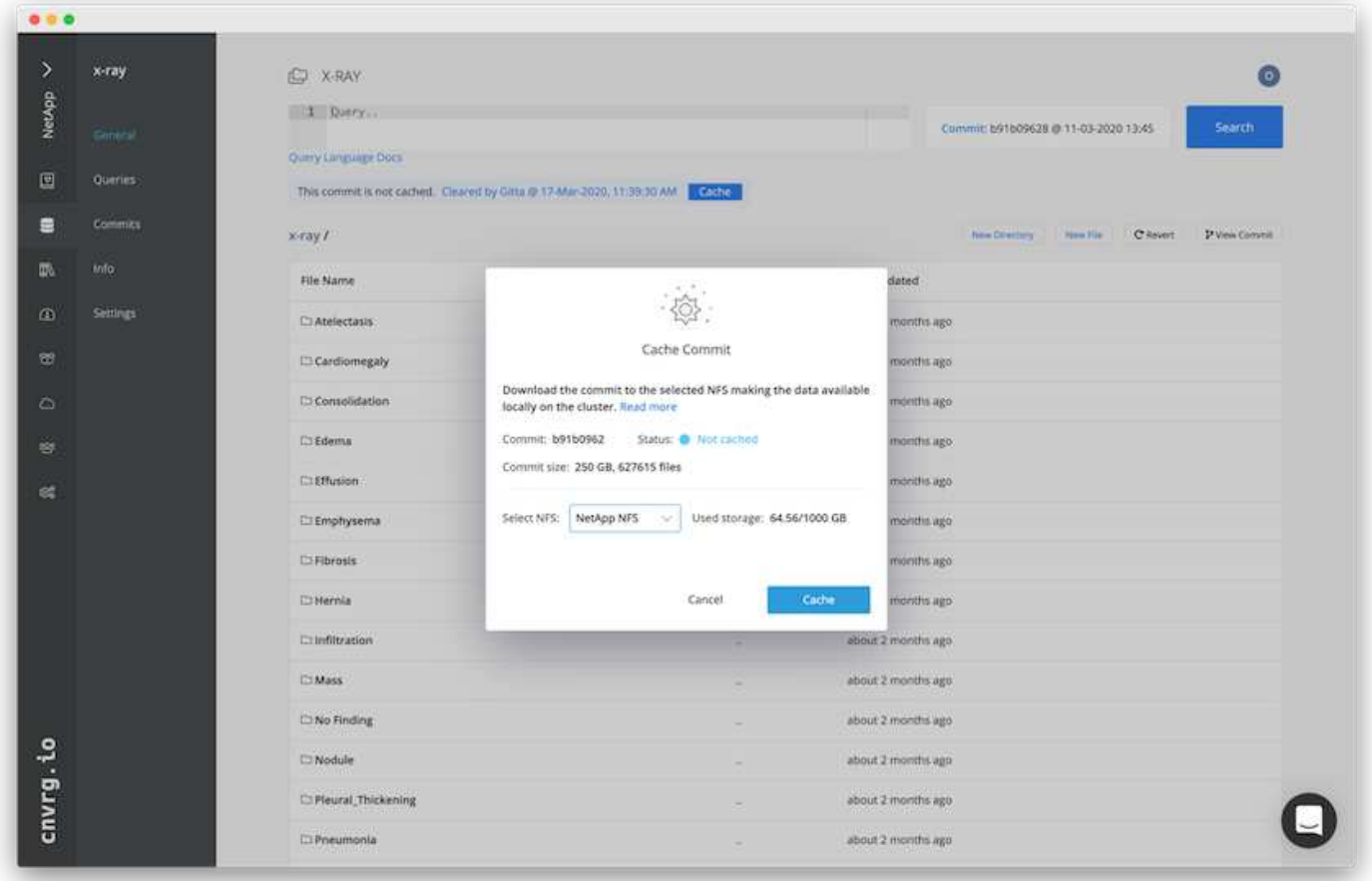
Per caricare i dati sulla piattaforma cnvrg, è possibile utilizzare la GUI o la CLI cnvrg. Per i set di dati di grandi dimensioni, NetApp consiglia di utilizzare CLI perché si tratta di uno strumento potente, scalabile e affidabile in grado di gestire un gran numero di file.

Per caricare i dati, attenersi alla seguente procedura:

1. Scaricare il "CLI cnvrg".
2. accedere alla directory dei raggi x.
3. Inizializzare il set di dati nella piattaforma con `cnvrg data init` comando.
4. Caricare tutti i contenuti della directory nel data Lake centrale con `cnvrg data sync` Command.una volta caricati i dati nell'archivio centrale di oggetti (StorageGRID, S3 o altri), è possibile navigare con la GUI. La figura seguente mostra un file PNG di immagine della fibrosi a raggi X del torace caricato. Inoltre, cnvrg consente di eseguire la versione dei dati in modo che qualsiasi modello creato possa essere riprodotto fino alla versione dei dati.

Dati di cach

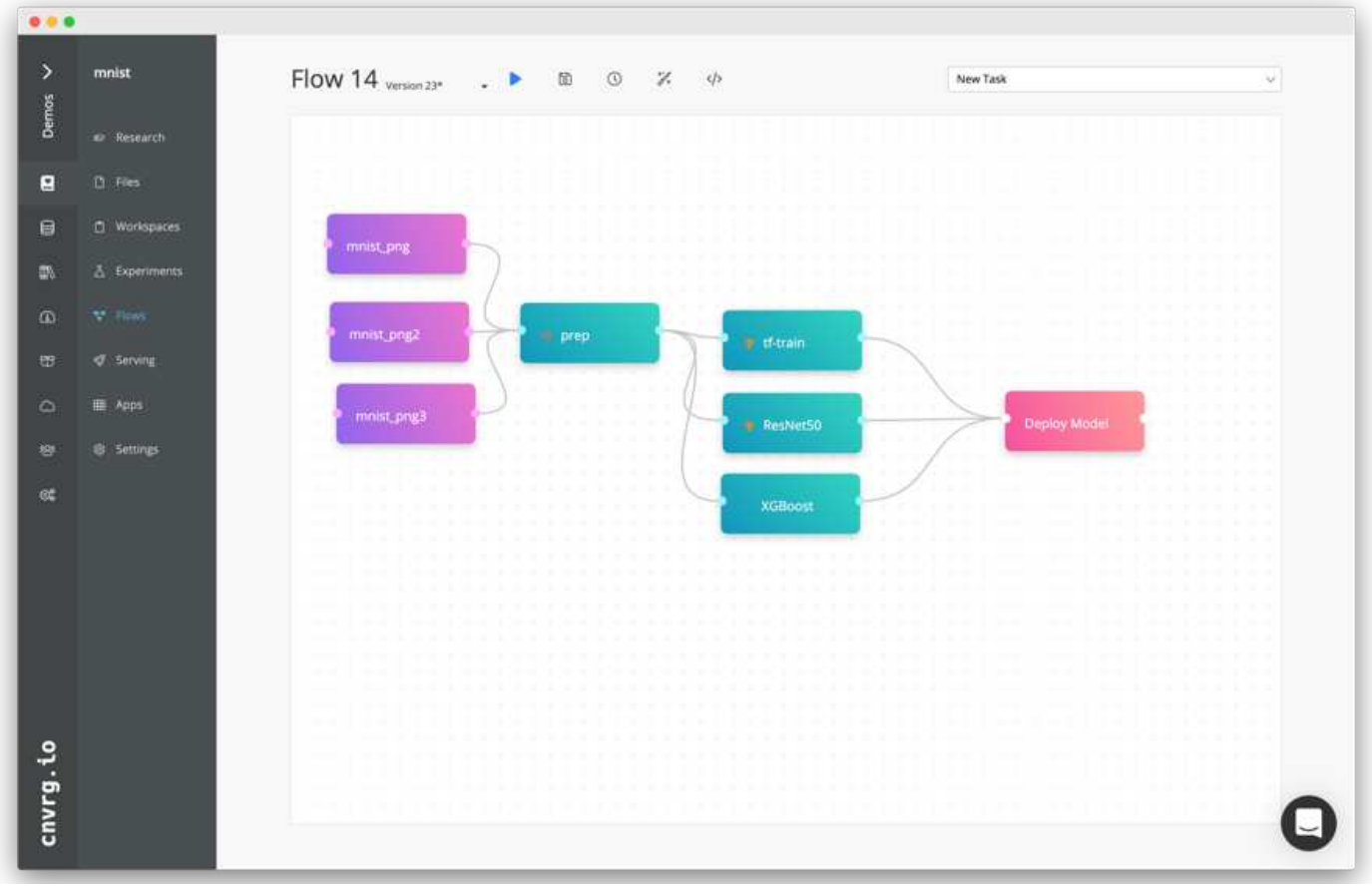
Per accelerare il training ed evitare il download di oltre 600.000 file per ciascun modello di training ed esperimento, abbiamo utilizzato la funzionalità di caching dei dati dopo che i dati sono stati inizialmente caricati nell'archivio centrale di oggetti data-Lake.



Dopo che gli utenti hanno fatto clic su cache, cnvrg scarica i dati nel relativo commit specifico dall'archivio remoto di oggetti e li memorizza nella cache del volume NFS di ONTAP. Al termine, i dati saranno disponibili per il training istantaneo. Inoltre, se i dati non vengono utilizzati per alcuni giorni (ad esempio, per la formazione o l'esplorazione del modello), cnvrg cancella automaticamente la cache.

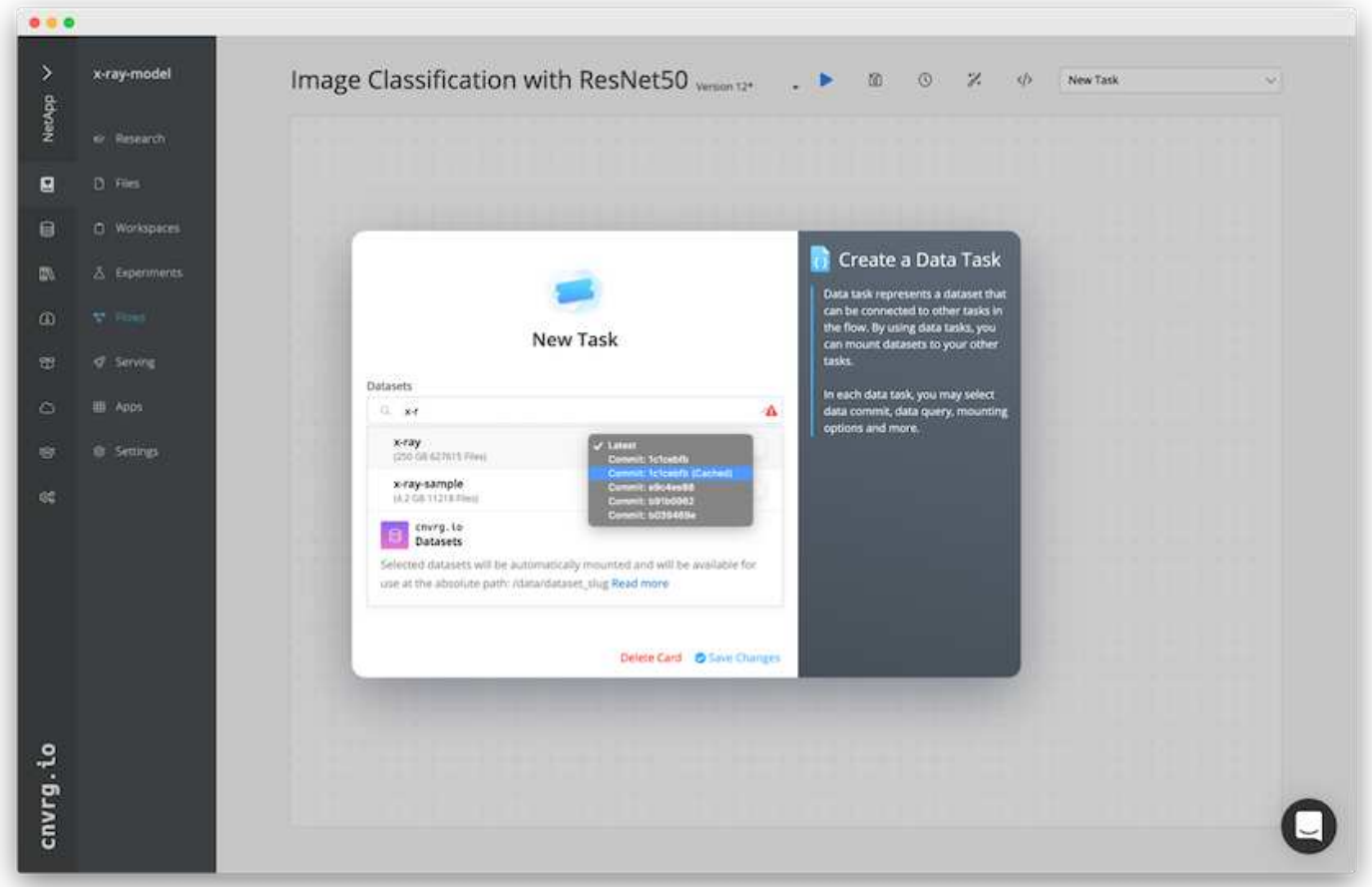
Crea una pipeline ML con i dati memorizzati nella cache

Cnvrg Flows consente di creare facilmente pipeline ML di produzione. I flussi sono flessibili, possono funzionare per qualsiasi tipo di caso d'utilizzo DI ML e possono essere creati attraverso la GUI o il codice. Ogni componente di un flusso può essere eseguito su una diversa risorsa di calcolo con un'immagine Docker diversa, il che rende possibile la creazione di cloud ibrido e pipeline ML ottimizzate.



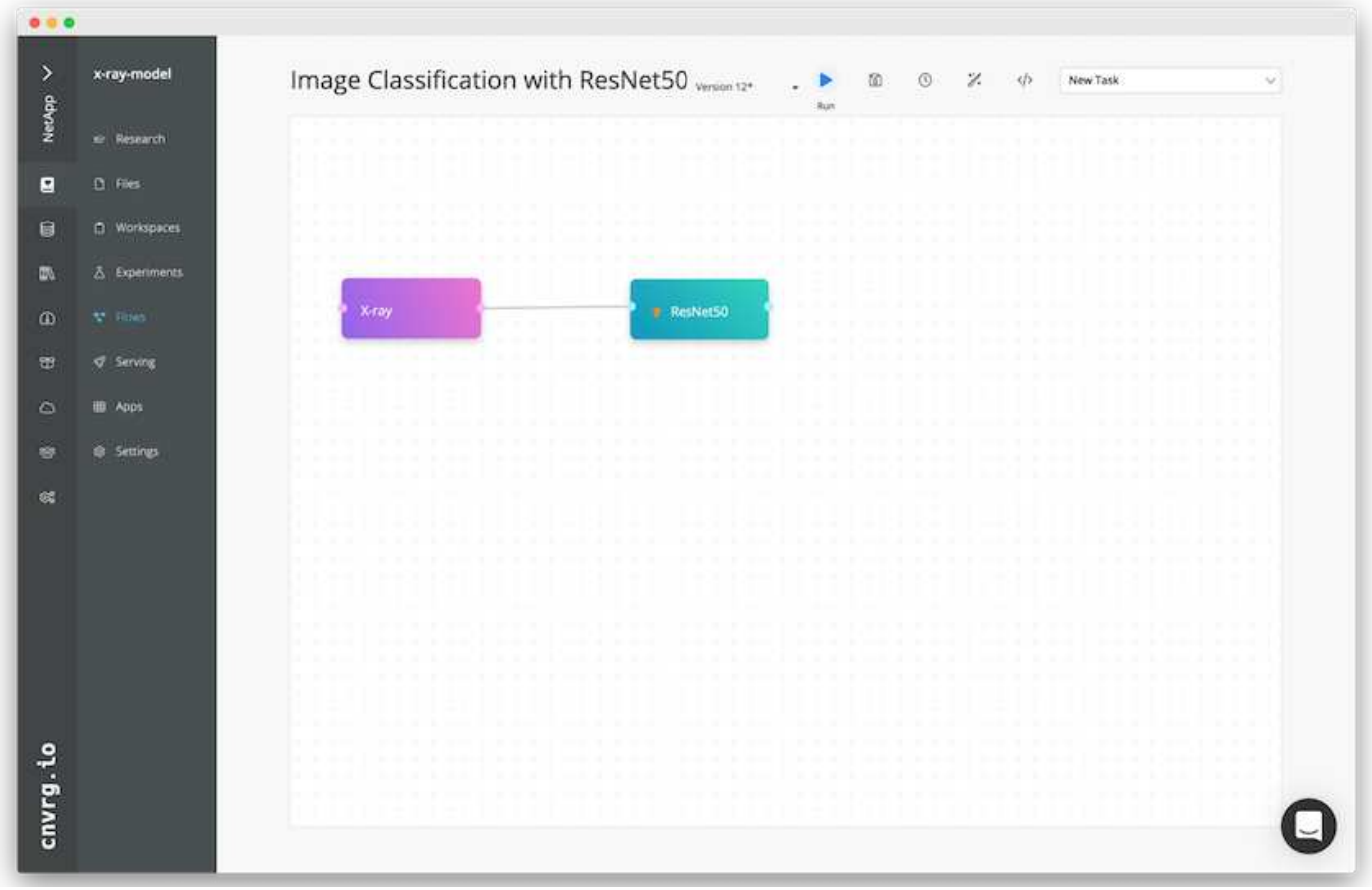
Creazione del flusso di raggi X del torace: Impostazione dei dati

Abbiamo aggiunto il nostro set di dati a un flusso appena creato. Quando si aggiunge il dataset, è possibile selezionare la versione specifica (commit) e indicare se si desidera la versione memorizzata nella cache. In questo esempio, è stato selezionato il commit memorizzato nella cache.



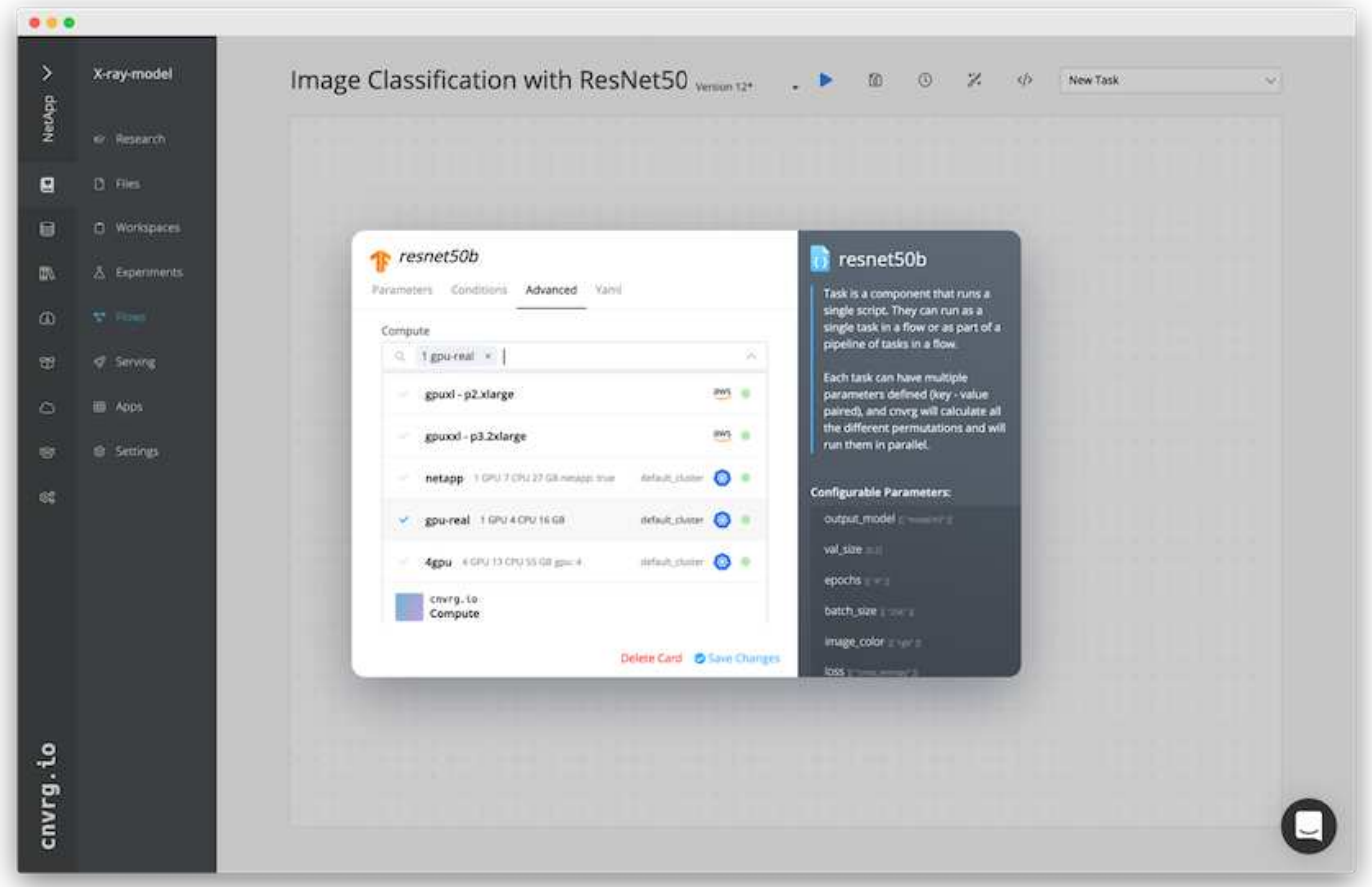
Creazione del flusso di raggi X del torace: Impostazione del modello di training: ResNet50

Nella pipeline, è possibile aggiungere qualsiasi tipo di codice personalizzato desiderato. In cnvrg è disponibile anche la libreria ai, una raccolta di componenti ML riutilizzabili. Nella libreria ai sono presenti algoritmi, script, origini dati e altre soluzioni che possono essere utilizzate in qualsiasi ML o flusso di deep learning. In questo esempio, è stato selezionato il modulo ResNet50 preinstallato. Abbiamo utilizzato parametri predefiniti come batch_size:128, epochs:10 e molto altro ancora. Questi parametri possono essere visualizzati nei documenti della ai Library. La seguente schermata mostra il nuovo flusso con il set di dati radiologici collegato a ResNet50.



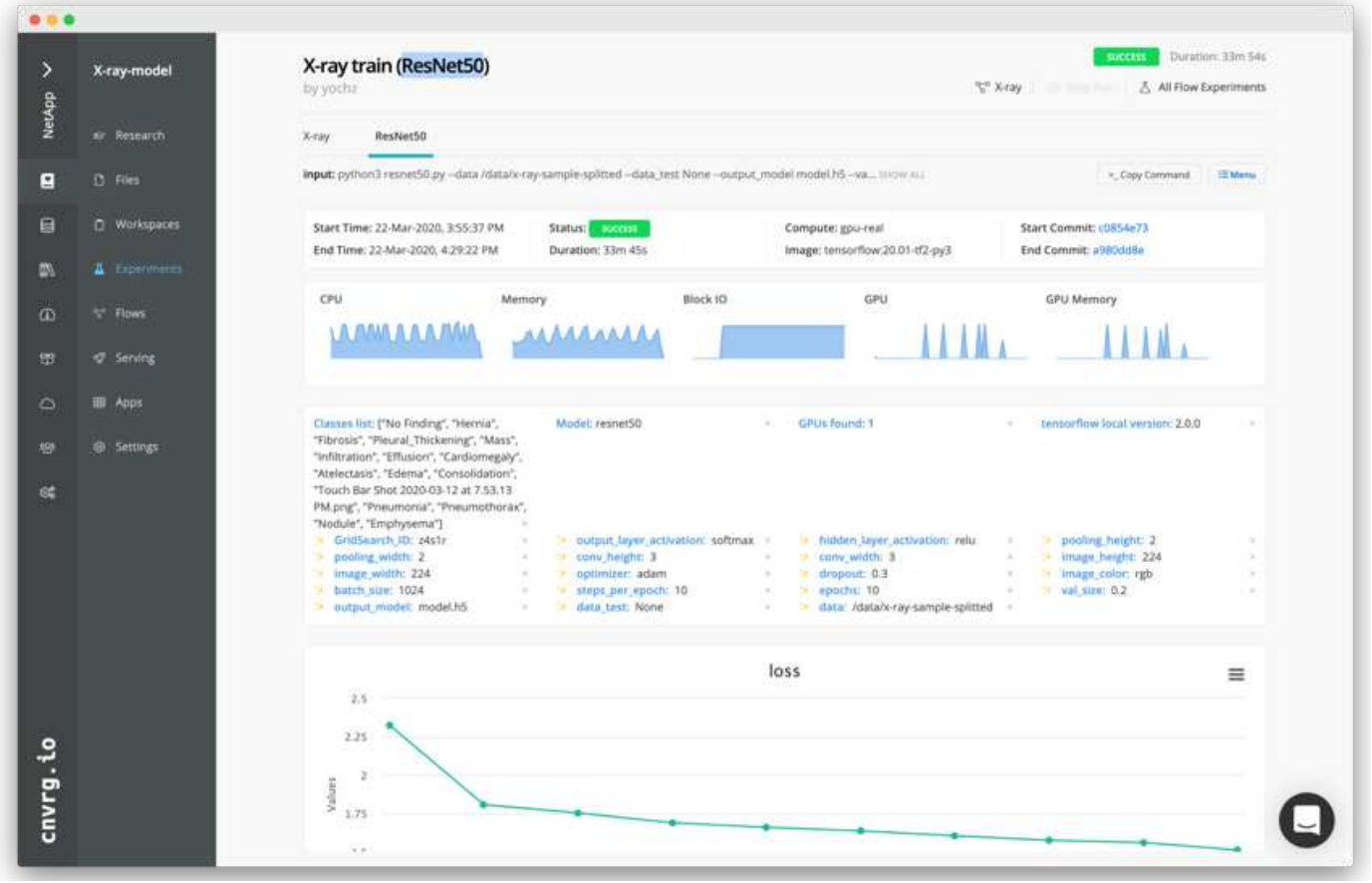
Definire la risorsa di calcolo per ResNet50

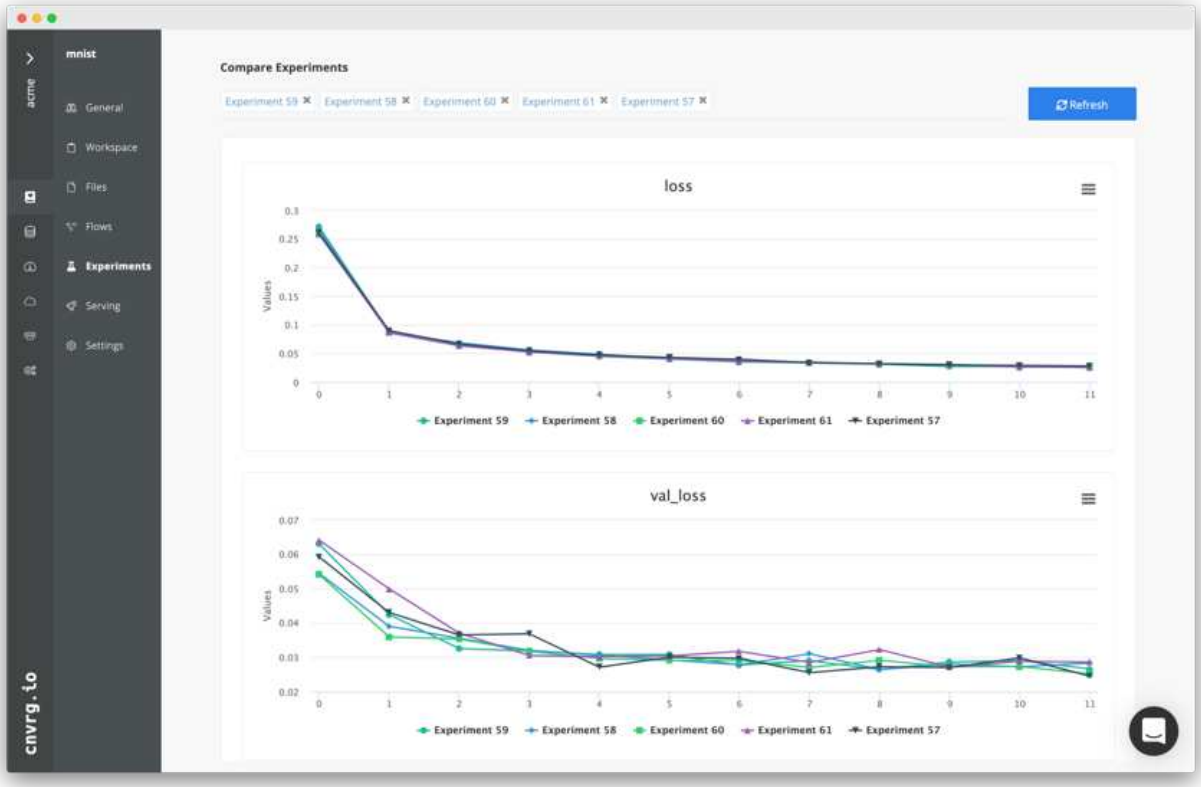
Ogni algoritmo o componente nei flussi cnvrg può essere eseguito su un'istanza di calcolo diversa, con un'immagine Docker diversa. Nella nostra configurazione, volevamo eseguire l'algoritmo di training sui sistemi NVIDIA DGX con l'architettura NetApp ONTAP ai. Nella figura seguente, è stato selezionato `gpu-real`, che è un modello di calcolo e una specifica per il nostro cluster on-premise. Abbiamo anche creato una coda di modelli e selezionato più modelli. In questo modo, se il `gpu-real` non è possibile allocare le risorse (se, ad esempio, altri data scientist le stanno utilizzando), quindi è possibile attivare la diffusione automatica del cloud aggiungendo un modello di cloud provider. La seguente schermata mostra l'utilizzo di gpu-real come nodo di calcolo per ResNet50.



Monitoraggio e monitoraggio dei risultati

Una volta eseguito un flusso, cnvrg attiva il motore di monitoraggio e tracciamento. Ogni esecuzione di un flusso viene documentata e aggiornata automaticamente in tempo reale. Hyperparameters, metriche, utilizzo delle risorse (utilizzo della GPU e altro ancora), versione del codice, artefatti, log, E così via sono disponibili automaticamente nella sezione Experiments (esperimenti), come mostrato nelle due schermate seguenti.


X-ray-model
X-ray train (ResNet50)
by yochz
SUCCESS
Duration: 33m 54s
X-ray
All Flow Experiments
Research
Files
Workspaces
Experiments
Flows
Serving
Apps
Settings
X-ray
ResNet50
input: python3 resnet50.py --data /data/x-ray-sample-splitted --data_test None --output_model model.h5 --va...
Copy Command
Menu
Start Time: 22-Mar-2020, 3:55:37 PM
End Time: 22-Mar-2020, 4:29:22 PM
Status: SUCCESS
Duration: 33m 45s
Compute: gpu-real
Image: tensorflow:20.01-tf2-py3
Start Commit: c0854e73
End Commit: a980dd8e
CPU
Memory
Block IO
GPU
GPU Memory
Classes list: ["No Finding", "Hernia", "Fibrosis", "Pleural_Thickening", "Mass", "Infiltration", "Effusion", "Cardiomegaly", "Atelectasis", "Edema", "Consolidation", "Touch Bar Shot 2020-03-12 at 7.53.13 PM.png", "Pneumonia", "Pneumothorax", "Nodule", "Emphysema"]
Model: resnet50
GPUs found: 1
tensorflow local version: 2.0.0
GridSearch_ID: z4s1r
pooling_width: 2
image_width: 224
batch_size: 1024
output_model: model.h5
output_layer_activation: softmax
conv_height: 3
optimizer: adam
steps_per_epoch: 10
data_test: None
hidden_layer_activation: relu
conv_width: 3
dropout: 0.3
epochs: 10
data: /data/x-ray-sample-splitted
pooling_height: 2
image_height: 224
image_color: rgb
val_size: 0.2
loss
Values
cnvrg.io


acme
mnist
General
Workspace
Files
Flows
Experiments
Serving
Settings
Compare Experiments
Experiment 59
Experiment 58
Experiment 60
Experiment 61
Experiment 57
Refresh
loss
Values
Experiment 59
Experiment 58
Experiment 60
Experiment 61
Experiment 57
val_loss
Values
Experiment 59
Experiment 58
Experiment 60
Experiment 61
Experiment 57
cnvrg.io

Conclusione

NetApp e cnvrg.io hanno collaborato per offrire ai clienti una soluzione completa per la gestione dei dati per lo sviluppo di software ML e DL. ONTAP ai offre calcolo e storage dalle performance elevate per qualsiasi scala operativa, mentre il software cnvrg.io ottimizza i flussi di lavoro di data science e migliora l'utilizzo delle risorse.

Ringraziamenti

- Mike Oglesby, Technical Marketing Engineer, NetApp
- Santosh Rao, Senior Technical Director di NetApp

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare le seguenti risorse:

- Cnvrg.io ("https://cnvrg.io"):
  - CORE Cnvrg (piattaforma ML gratuita)

    https://cnvrg.io/platform/core
  - Documenti Cnvrg

    "https://app.cnvrg.io/docs"
- Server NVIDIA DGX-1:
  - Server NVIDIA DGX-1

    https://www.nvidia.com/en-us/data-center/dgx-1/
  - NVIDIA Tesla V100 Tensor Core GPU

    https://www.nvidia.com/en-us/data-center/tesla-v100/
  - NVIDIA GPU CLOUD (NGC)

    https://www.nvidia.com/en-us/gpu-cloud/
- Sistemi NetApp AFF:
  - Scheda informativa su AFF

    https://www.netapp.com/us/media/d-3582.pdf
  - NetApp FlashAdvantage per AFF

    https://www.netapp.com/us/media/ds-3733.pdf
  - Documentazione di ONTAP 9.x.

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

- Report tecnico di NetApp FlexGroup

  https://www.netapp.com/us/media/tr-4557.pdf

- Storage persistente NetApp per container:
  - Trident di NetApp

    https://netapp.io/persistent-storage-provisioner-for-kubernetes/

- Matrice di interoperabilità NetApp:
  - Tool di matrice di interoperabilità NetApp

    https://mysupport.netapp.com/matrix/#welcome

- Networking ai ONTAP:
  - Switch Cisco Nexus 3232C

    https://www.cisco.com/c/en/us/products/switches/nexus-3232c-switch/index.html

  - Switch Mellanox Spectrum serie 2000

    http://www.mellanox.com/page/products_dyn?product_family=251&mtag=sn2000

- Framework E tool ML:
  - DALI

    https://github.com/NVIDIA/DALI

  - TensorFlow: Un framework di apprendimento automatico open-source per tutti

    https://www.tensorflow.org/

  - Horovod: Framework di deep learning distribuito open-source di Uber per TensorFlow

    https://eng.uber.com/horovod/

  - Abilitazione delle GPU nell'ecosistema di runtime container

    https://devblogs.nvidia.com/gpu-containers-runtime/

  - Docker

    https://docs.docker.com

  - Kubernetes

    https://kubernetes.io/docs/home/

  - NVIDIA DeepOps

    https://github.com/NVIDIA/deepops

  - Kubeflow

http://www.kubeflow.org/

- Jupyter notebook Server

  http://www.jupyter.org/

- Set di dati e benchmark:
  - Set di dati per radiografia toracica NIH

    https://nihcc.app.box.com/v/ChestXray-NIHCC

  - Xiaosong Wang, Yifan Peng, le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald Summers, ChestX-ray8: Database dei raggi X del torace su scala ospedaliera e benchmark sulla classificazione e localizzazione con supervisione debole delle malattie toraciche comuni, IEEE CVPR, pp 3462-3471, 2017TR-4841-0620

TR-4732: Dai dati di analisi dei big data all'intelligenza artificiale

Karthikeyan Nagalingam, NetApp

Questo documento descrive come spostare i dati di analisi dei big data e i dati HPC nell'ai. L'ai elabora i dati NFS attraverso le esportazioni NFS, mentre i clienti spesso dispongono dei propri dati ai in una piattaforma di analisi dei big data, come lo storage HDFS, Blob o S3, oltre a piattaforme HPC come GPFS. Questo documento fornisce linee guida per lo spostamento dei dati di analisi dei big data e dei dati HPC nell'ai utilizzando NetApp XCP e NIPAM. Discutiamo inoltre dei vantaggi per il business derivanti dal passaggio dei dati da big data e HPC all'ai.

Concetti e componenti

Storage per l'analisi dei big data

L'analisi dei big data è il principale provider di storage per HDFS. Un cliente utilizza spesso un file system compatibile con Hadoop (HCFS) come Windows Azure Blob Storage, MapR file System (MapR-FS) e lo storage a oggetti S3.

File system parallelo generale

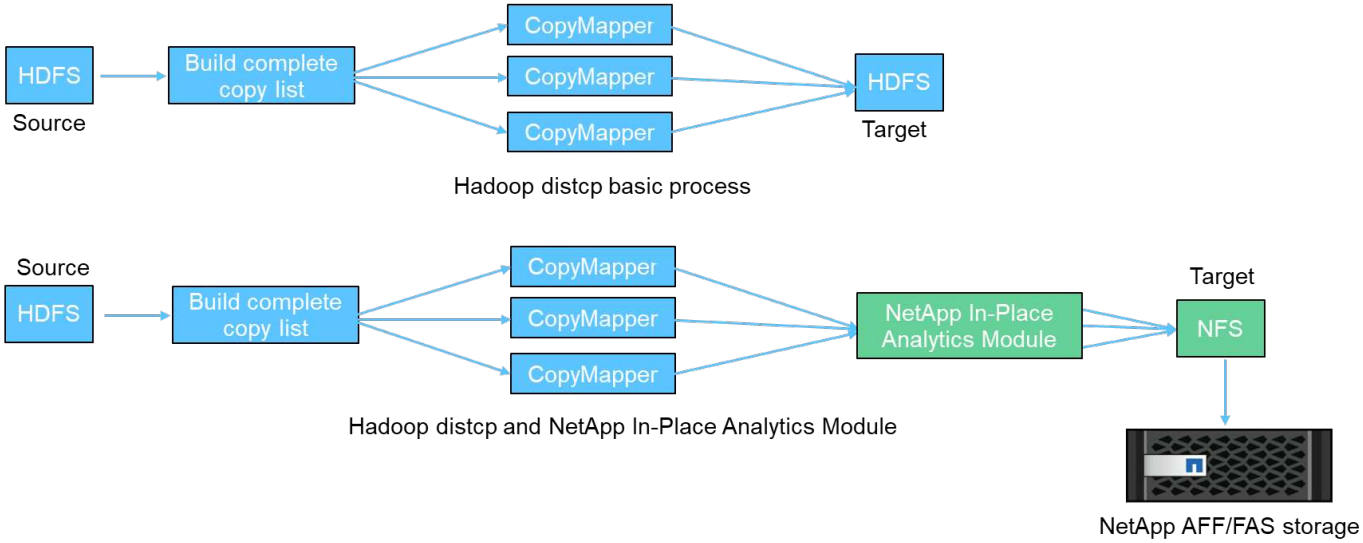
Il GPFS di IBM è un file system aziendale che offre un'alternativa a HDFS. LE GPF offrono alle applicazioni la flessibilità necessaria per decidere le dimensioni dei blocchi e il layout di replica, garantendo buone performance ed efficienza.

Modulo NetApp in-place Analytics

Il NetApp in-place Analytics Module (NIPAM) funge da driver per i cluster Hadoop per accedere ai dati NFS. Ha quattro componenti: Un pool di connessioni, un NFS InputStream, una cache di handle di file e un NFS OutputStream. Per ulteriori informazioni, vedere "TR-4382: Modulo NetApp in-place Analytics."

Copia distribuita Hadoop

La copia distribuita di Hadoop (DistCp) è uno strumento di copia distribuita utilizzato per attività di coping tra cluster e intra-cluster di grandi dimensioni. Questo strumento utilizza MapReduce per la distribuzione dei dati, la gestione degli errori e il reporting. Espande l'elenco di file e directory e li inserisce per mappare le attività per copiare i dati dall'elenco di origine. L'immagine seguente mostra l'operazione DistCp in HDFS e non in HDFS.



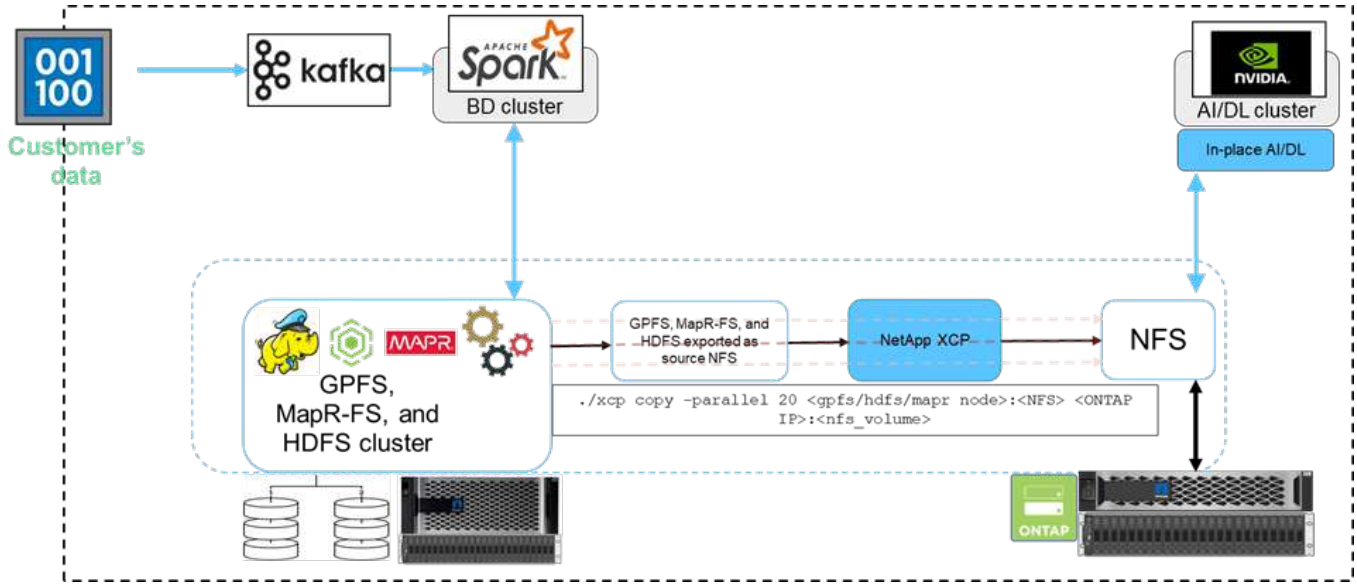
Hadoop DistCp sposta i dati tra i due sistemi HDFS senza utilizzare un driver aggiuntivo. NetApp fornisce il driver per i sistemi non HDFS. Per una destinazione NFS, NIPAM fornisce il driver per copiare i dati utilizzati da Hadoop DistCp per comunicare con le destinazioni NFS durante la copia dei dati.

NetApp Cloud Volumes Service

NetApp Cloud Volumes Service è un file service nativo del cloud con performance estreme. Questo servizio aiuta i clienti ad accelerare il time-to-market, aumentando e diminuendo rapidamente le risorse e utilizzando le funzionalità NetApp per migliorare la produttività e ridurre i tempi di inattività del personale. Cloud Volumes Service è la giusta alternativa per il disaster recovery e il backup nel cloud, in quanto riduce l'impatto complessivo del data center e consuma meno storage di cloud pubblico nativo.

XCP di NetApp

NetApp XCP è un software client che consente una migrazione dei dati rapida e affidabile da qualsiasi a NetApp e da NetApp a NetApp. Questo tool è progettato per copiare una grande quantità di dati NAS non strutturati da qualsiasi sistema NAS a un controller di storage NetApp. XCP Migration Tool utilizza un motore di streaming i/o multicore e multicanale in grado di elaborare molte richieste in parallelo, ad esempio migrazione dei dati, elenchi di file o directory e report di spazio. Questo è il tool di migrazione dei dati NetApp predefinito. È possibile utilizzare XCP per copiare i dati da un cluster Hadoop e HPC allo storage NetApp NFS. Il diagramma seguente mostra il trasferimento dei dati da un cluster Hadoop e HPC a un volume NetApp NFS utilizzando XCP.



Copia e sincronizzazione di NetApp BlueXP

NetApp BlueXP Copy and Sync è un software-as-a-service di replica dei dati ibridi che trasferisce e sincronizza i dati NFS, S3 e CIFS in modo perfetto e sicuro tra storage on-premise e cloud storage. Questo software viene utilizzato per la migrazione dei dati, l'archiviazione, la collaborazione, l'analisi e altro ancora. Una volta trasferiti i dati, BlueXP Copy e Sync sincronizza costantemente i dati tra origine e destinazione. In futuro, trasferisce il delta. Inoltre, protegge i dati all'interno della tua rete, nel cloud o on-premise. Questo software si basa su un modello pay-as-you-go, che fornisce una soluzione conveniente e offre funzionalità di monitoraggio e reporting per il trasferimento dei dati.

Ai Inferencing at the Edge - NetApp con Lenovo ThinkSystem - progettazione di soluzioni

TR-4886: Ai Inferencing at the Edge - NetApp con Lenovo ThinkSystem - progettazione di soluzioni

Sathish Thyagarajan, NetApp Miroslav Hodak, Lenovo

Questo documento descrive un'architettura di calcolo e storage per implementare deduzione di intelligenza artificiale (ai) basata su GPU su storage controller NetApp e server Lenovo ThinkSystem in un ambiente edge che soddisfa gli scenari applicativi emergenti.

Riepilogo

Diversi scenari applicativi emergenti, come i sistemi avanzati di assistenza alla guida (ADAS), Industry 4.0, smart cities e Internet of Things (IoT), richiedono l'elaborazione di flussi di dati continui con una latenza quasi nulla. Questo documento descrive un'architettura di calcolo e storage per implementare deduzione di intelligenza artificiale (ai) basata su GPU su storage controller NetApp e server Lenovo ThinkSystem in un ambiente edge che soddisfi questi requisiti. Questo documento fornisce inoltre dati sulle performance per il benchmark MLPerf Inference standard di settore, valutando varie attività di inferenza su edge server dotati di GPU NVIDIA T4. Analizziamo le performance degli scenari di inferenza offline, single stream e multistream e mostriamo che l'architettura con un sistema di storage condiviso in rete a costi contenuti è altamente performante e fornisce un punto centrale per la gestione di dati e modelli per più edge server.

Introduzione

Le aziende stanno generando sempre più enormi volumi di dati all'edge della rete. Per ottenere il massimo valore dai sensori intelligenti e dai dati IoT, le organizzazioni sono alla ricerca di una soluzione di streaming di eventi in tempo reale che consenta l'edge computing. I lavori più impegnativi dal punto di vista computazionale vengono quindi eseguiti sempre più all'edge, al di fuori dei data center. L'inferenza dell'ai è uno dei fattori trainanti di questa tendenza. Gli edge server forniscono una potenza di calcolo sufficiente per questi carichi di lavoro, soprattutto quando si utilizzano acceleratori, ma lo storage limitato è spesso un problema, soprattutto negli ambienti multisserver. In questo documento mostreremo come puoi implementare un sistema storage condiviso nell'ambiente edge e in che modo esso benefici per i carichi di lavoro di inferenza ai senza imporre penalizzazioni in termini di performance.

Questo documento descrive un'architettura di riferimento per l'inferenza ai ai edge. Combina più edge server Lenovo ThinkSystem con un sistema storage NetApp per creare una soluzione semplice da implementare e gestire. Si tratta di una guida di riferimento per implementazioni pratiche in varie situazioni, come ad esempio il reparto di produzione con telecamere multiple e sensori industriali, sistemi POS (Point of sale) nelle transazioni al dettaglio o sistemi FSD (Full Self-Driving) che identificano anomalie visive nei veicoli autonomi.

Il presente documento illustra il test e la convalida di una configurazione di calcolo e storage costituita da un server edge Lenovo ThinkSystem SE350 e da un sistema storage entry-level NetApp AFF ed EF-Series. Le architetture di riferimento offrono una soluzione efficiente e conveniente per le implementazioni ai, fornendo al contempo servizi dati completi, protezione integrata dei dati, scalabilità perfetta e storage dei dati connesso al cloud con NetApp ONTAP e il software di gestione dei dati NetApp SANtricity.

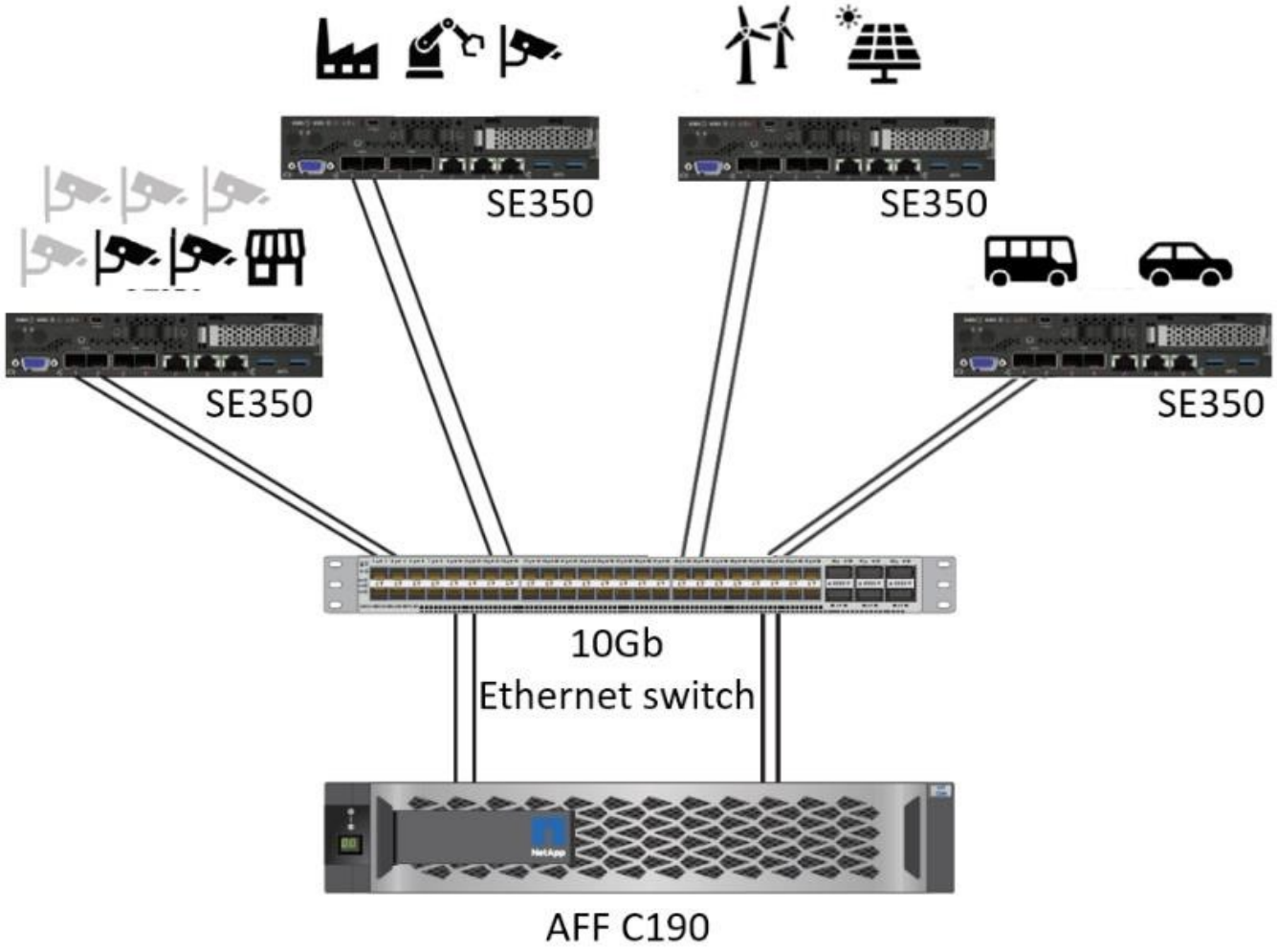
Pubblico di riferimento

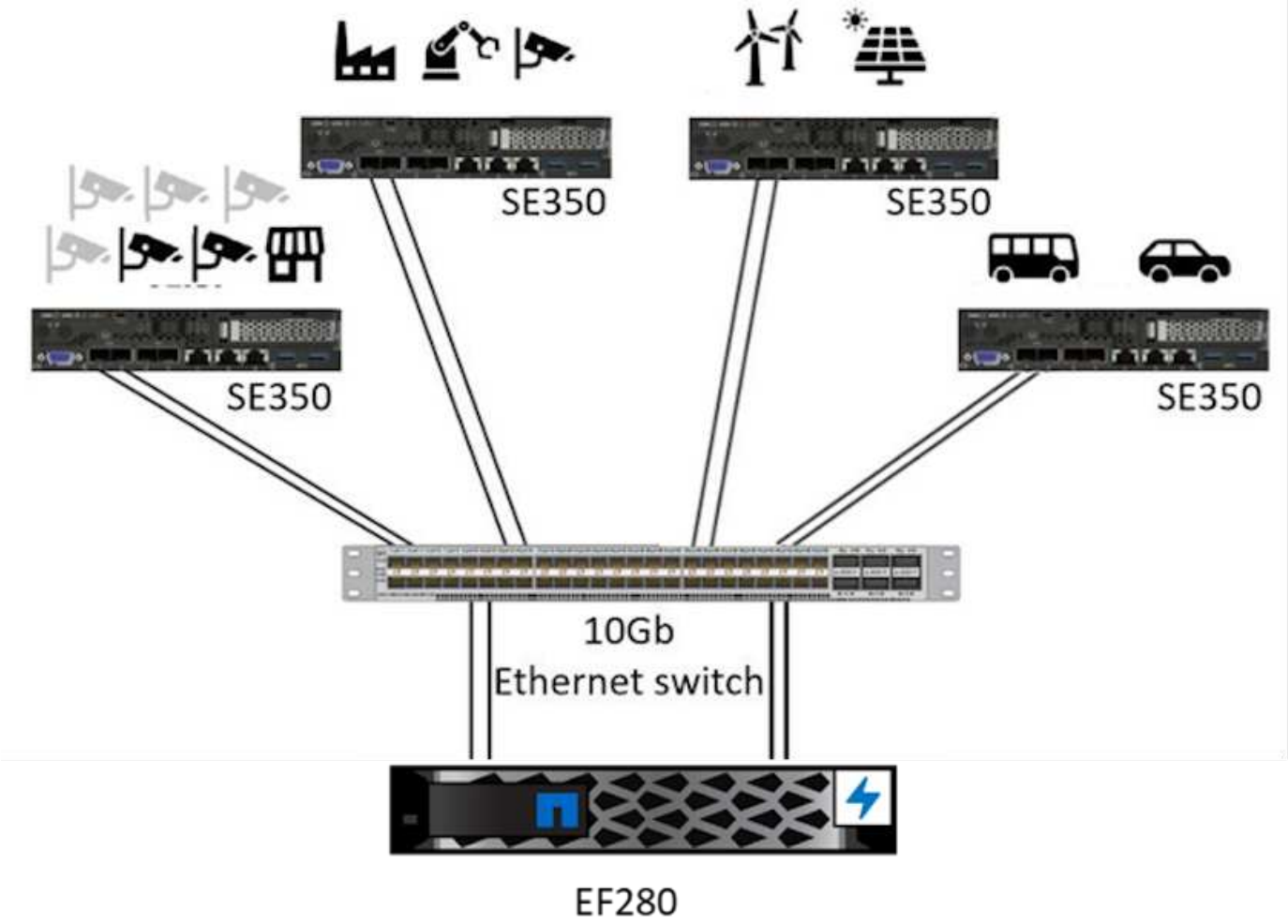
Il presente documento è destinato ai seguenti destinatari:

- Business leader e Enterprise architect che vogliono produrre l'ai alla periferia della rete.
- Data scientist, data engineer, ricercatori ai/machine learning (ML) e sviluppatori di sistemi ai.
- Architetti aziendali che progettano soluzioni per lo sviluppo di modelli e applicazioni ai/ML.
- Data scientist e ingegneri ai alla ricerca di modi efficienti per implementare modelli di deep learning (DL) e ML.
- Edge Device Manager e amministratori degli edge server responsabili dell'implementazione e della gestione dei modelli di inferenza edge.

Architettura della soluzione

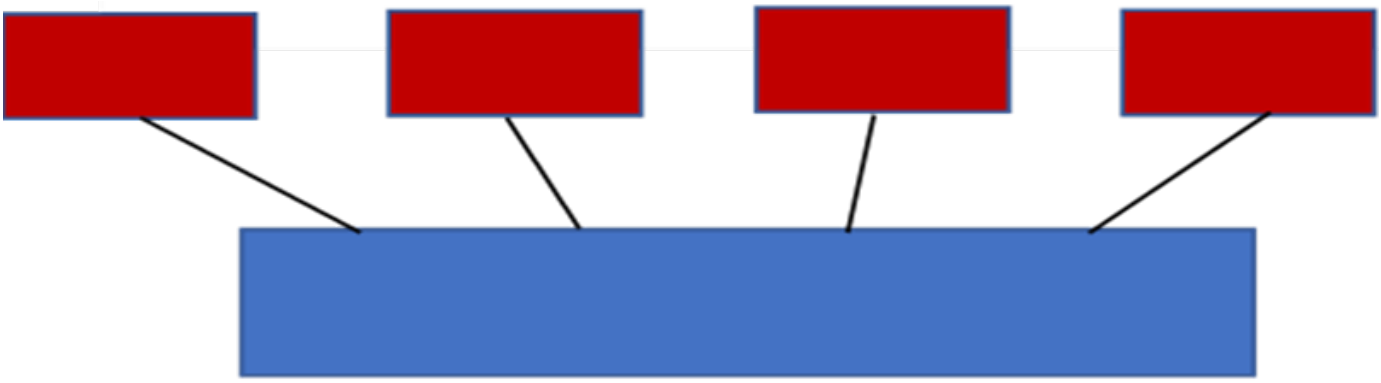
Questo server Lenovo ThinkSystem e la soluzione di storage NetApp ONTAP o NetApp SANtricity sono progettati per gestire l'inferenza ai su set di dati di grandi dimensioni utilizzando la potenza di elaborazione delle GPU insieme alle CPU tradizionali. Questa convalida dimostra performance elevate e una gestione ottimale dei dati con un'architettura che utilizza uno o più edge server Lenovo SR350 interconnessi con un singolo sistema di storage NetApp AFF, come mostrato nelle due figure seguenti.


SE350
SE350
SE350
SE350
10Gb
Ethernet switch
AFF C190



La panoramica dell'architettura logica nella figura seguente mostra i ruoli degli elementi di calcolo e storage in questa architettura. In particolare, viene mostrato quanto segue:

- I dispositivi di calcolo edge che eseguono l'inferenza sui dati ricevuti da telecamere, sensori e così via.
- Un elemento storage condiviso che serve più scopi:
  - Fornisce una posizione centrale per i modelli di inferenza e altri dati necessari per eseguire l'inferenza. I server di calcolo accedono direttamente allo storage e utilizzano modelli di inferenza in tutta la rete senza la necessità di copiarli localmente.
  - I modelli aggiornati vengono inviati qui.
  - Archivia i dati di input ricevuti dagli edge server per un'analisi successiva. Ad esempio, se i dispositivi edge sono collegati alle telecamere, l'elemento di storage mantiene i video acquisiti dalle telecamere.



rosso	blu
Sistema di calcolo Lenovo	Sistema storage NetApp AFF
Dispositivi edge che eseguono deduzione sugli input provenienti da telecamere, sensori e così via.	Storage condiviso con modelli di inferenza e dati provenienti da dispositivi edge per analisi successive.

Questa soluzione NetApp e Lenovo offre i seguenti vantaggi principali:

- La GPU ha accelerato il computing all'edge della rete.
- Implementazione di più edge server supportata e gestita da uno storage condiviso.
- Protezione dei dati efficace per soddisfare gli obiettivi RPO (Recovery Point Objective) e RTO (Recovery Time Objective) ridotti senza perdita di dati.
- Gestione dei dati ottimizzata con copie Snapshot e cloni NetApp per ottimizzare i flussi di lavoro di sviluppo.

Come utilizzare questa architettura

Questo documento convalida la progettazione e le performance dell'architettura proposta. Tuttavia, non abbiamo testato alcuni componenti a livello di software, come la gestione di container, workload o modelli e la sincronizzazione dei dati con il cloud o il data center on-premise, perché sono specifici di uno scenario di implementazione. In questo caso, esistono diverse scelte.

A livello di gestione dei container, Kubernetes Container Management è una buona scelta ed è ben supportato sia in una versione completamente upstream (Canonical) che in una versione modificata adatta per le implementazioni Enterprise (Red Hat). Il "Piano di controllo ai di NetApp" Che utilizza NetApp Trident e il nuovo aggiunto "NetApp DataOps Toolkit" Offre tracciabilità integrata, funzioni di gestione dei dati, interfacce e strumenti per data scientist e data engineer da integrare con lo storage NetApp. Kubeflow, IL toolkit ML per Kubernetes, offre funzionalità ai aggiuntive, oltre al supporto per il controllo delle versioni dei modelli e KFServing su diverse piattaforme, come TensorFlow Serving o NVIDIA Triton Inference Server. Un'altra opzione è la piattaforma NVIDIA EGX, che offre la gestione dei carichi di lavoro e l'accesso a un catalogo di container di inferenza ai abilitati per GPU. Tuttavia, queste opzioni potrebbero richiedere sforzi ed esperienza significativi per metterle in produzione e potrebbero richiedere l'assistenza di un vendor di software indipendente (ISV) o di un consulente di terze parti.

Aree di soluzione

Il vantaggio principale dell'inferenza ai e dell'edge computing è la capacità dei dispositivi di calcolare, elaborare e analizzare i dati con un elevato livello di qualità senza latenza. Esistono troppi esempi di casi d'utilizzo di edge computing da descrivere in questo documento, ma di seguito sono riportati alcuni esempi importanti:

Automobili: Veicoli autonomi

La classica illustrazione dell'edge computing si trova nei sistemi avanzati di assistenza alla guida (ADAS) nei veicoli autonomi (AV). L'ai nelle auto senza conducente deve elaborare rapidamente una grande quantità di dati provenienti da telecamere e sensori per essere un pilota sicuro e di successo. Un'interpretazione troppo lunga tra un oggetto e un essere umano può significare vita o morte, pertanto è fondamentale essere in grado di elaborare i dati il più vicino possibile al veicolo. In questo caso, uno o più server di calcolo edge gestiscono l'input da telecamere, RADAR, LDAR e altri sensori, mentre lo storage condiviso contiene modelli di inferenza e memorizza i dati di input provenienti dai sensori.

Settore sanitario: Monitoraggio dei pazienti

Uno dei maggiori impatti dell'intelligenza artificiale e dell'edge computing è la sua capacità di migliorare il monitoraggio continuo dei pazienti per le malattie croniche sia nelle strutture di assistenza domiciliare che nelle unità di terapia intensiva (ICU). I dati provenienti da dispositivi periferici che monitorano i livelli di insulina, la respirazione, l'attività neurologica, il ritmo cardiaco e le funzioni gastrointestinali richiedono un'analisi istantanea dei dati che devono essere utilizzati immediatamente, in quanto il tempo necessario per agire è limitato per salvare la vita di qualcuno.

Retail: Pagamento senza cassa

L'edge computing può potenziare ai e ML per aiutare i retailer a ridurre i tempi di checkout e aumentare il traffico. I sistemi senza cassiere supportano diversi componenti, ad esempio:

- Autenticazione e accesso. Collegare l'acquirente fisico a un account validato e consentire l'accesso allo spazio di vendita al dettaglio.
- Monitoraggio dell'inventario. Utilizzo di sensori, tag RFID e sistemi di visione computerizzata per confermare la selezione o la deselezione degli articoli da parte degli acquirenti.

  In questo caso, ciascuno degli edge server gestisce ciascun contatore di cassa e il sistema di storage condiviso funge da punto di sincronizzazione centrale.

Servizi finanziari: Sicurezza umana nei chioschi e prevenzione delle frodi

Le organizzazioni bancarie utilizzano l'ai e l'edge computing per innovare e creare esperienze bancarie personalizzate. I chioschi interattivi che utilizzano l'analisi dei dati in tempo reale e l'inferenza ai consentono ora agli ATM non solo di aiutare i clienti a prelevare denaro, ma anche di monitorare in modo proattivo i chioschi attraverso le immagini acquisite dalle telecamere per identificare i rischi per la sicurezza umana o i comportamenti fraudolenti. In questo scenario, i server di calcolo edge e i sistemi storage condivisi sono collegati a chioschi e telecamere interattivi per aiutare le banche a raccogliere ed elaborare i dati con modelli di inferenza ai.

Produzione: Settore 4.0

È iniziata la quarta rivoluzione industriale (Industry 4.0), insieme a trend emergenti come Smart Factory e stampa 3D. Per prepararsi a un futuro basato sui dati, la comunicazione machine-to-machine (M2M) e l'IoT su larga scala sono integrati per una maggiore automazione senza la necessità di un intervento umano. La produzione è già altamente automatizzata e l'aggiunta di funzionalità di ai è una naturale continuazione della tendenza a lungo termine. L'ai consente di automatizzare le operazioni che possono essere automatizzate con l'aiuto di computer Vision e altre funzionalità di ai. È possibile automatizzare il controllo di qualità o le attività che si basano sulla visione umana o sul processo decisionale per eseguire analisi più rapide dei materiali sulle linee di assemblaggio nei piani della fabbrica, in modo da aiutare gli impianti di produzione a soddisfare gli standard ISO richiesti per la gestione della qualità e della sicurezza. In questo caso, ogni edge server di calcolo è connesso a un array di sensori che monitorano il processo di produzione e i modelli di inferenza aggiornati vengono inviati allo storage condiviso, in base alle necessità.

Telecomunicazioni: Rilevamento della ruggine, ispezione della torre e ottimizzazione della rete

Il settore delle telecomunicazioni utilizza tecniche di visione computerizzata e ai per elaborare immagini che rilevano automaticamente la ruggine e identificano le torri cellulari che contengono corrosione e, di conseguenza, richiedono un'ulteriore ispezione. Negli ultimi anni è aumentato l'utilizzo di immagini drone e modelli ai per identificare regioni distinte di una torre per analizzare ruggine, crepe superficiali e corrosione. La domanda continua a crescere per le tecnologie ai che consentono di ispezionare in modo efficiente l'infrastruttura di telecomunicazione e i ripetitori cellulari, valutarne regolarmente il degrado e ripararli

tempestivamente quando necessario.

Inoltre, un altro caso d'utilizzo emergente nel settore delle telecomunicazioni è l'utilizzo di algoritmi ai e ML per prevedere i modelli di traffico dati, rilevare i dispositivi compatibili con il 5G e automatizzare e aumentare la gestione dell'energia MIMO (Multiple-Input and Multiple-Output). L'hardware MIMO viene utilizzato nelle radio tower per aumentare la capacità di rete, ma ciò comporta costi energetici aggiuntivi. I modelli ML per la "modalità di sospensione MIMO" implementati nei siti cellulari possono prevedere l'utilizzo efficiente delle radio e contribuire a ridurre i costi di consumo energetico per gli operatori di reti mobili (MNOS). Le soluzioni di inferenza ai e edge computing aiutano gli MNOS a ridurre la quantità di dati trasmessi avanti e indietro ai data center, ridurre il TCO, ottimizzare le operazioni di rete e migliorare le performance complessive per gli utenti finali.

Panoramica della tecnologia

Questa sezione descrive le basi tecnologiche di questa soluzione ai.

Sistemi NetApp AFF

I sistemi storage NetApp AFF all'avanguardia consentono implementazioni di inferenza ai ai ai edge per soddisfare i requisiti di storage Enterprise con performance leader di settore, flessibilità superiore, integrazione nel cloud e gestione dei dati Best-in-class. Progettati appositamente per la tecnologia flash, i sistemi NetApp AFF aiutano ad accelerare, gestire e proteggere i dati business-critical.

- I sistemi storage entry-level NetApp AFF sono basati su hardware FAS2750 e supporti flash SSD
- Due controller in configurazione ha

I sistemi storage AFF C190 entry-level di NetApp supportano le seguenti funzionalità:

- Un numero massimo di dischi SSD 24x 960 GB
- Due possibili configurazioni:
  - Ethernet (10 GbE): 4 porte 10GBASE-T (RJ-45)
  - Unified (16 GB FC o 10 GbE): 4 porte UTA2 (Unified Target Adapter)
- Capacità effettiva massima di 50,5 TB

Per i carichi di lavoro NAS, un singolo sistema AFF C190 entry-level supporta un throughput di 4,4 Gbps per letture sequenziali e 230 K IOPS per piccole letture casuali a latenze di 1 ms o inferiori.

NetApp AFF A220

NetApp offre anche altri sistemi storage entry-level che offrono performance e scalabilità superiori per implementazioni su larga scala. Per i carichi di lavoro NAS, un singolo sistema AFF A220 entry-level supporta:

- Throughput di 6,2 Gbps per letture sequenziali
- 375.000 IOPS per piccole letture casuali con latenze di 1 ms o meno
- Numero massimo di dischi SSD 144 x 960 GB, 3,8 TB o 7,6 TB
- AFF A220 è in grado di scalare fino a un massimo di 1 PB di capacità effettiva

NetApp AFF A250

- La capacità effettiva massima è di 35 PB con una scalabilità massima di 2-24 nodi (12 coppie ha)
- Offre un aumento delle performance di ≥ 45% rispetto a AFF A220
- 440.000 letture casuali IOPS @1 ms.
- Basato sull'ultima release di NetApp ONTAP: ONTAP 9.8
- Sfrutta due porte Ethernet da 25 GB per l'interconnessione di ha e cluster

Sistemi NetApp e-Series EF

EF-Series è una famiglia di storage array SAN all-flash entry-level e mid-range in grado di accelerare l'accesso ai dati e di trarne valore più rapidamente con il software NetApp SANtricity. Questi sistemi offrono storage flash SAS e NVMe e offrono IOPS da convenienti a estremi, tempi di risposta inferiori a 100 microsecondi e larghezza di banda fino a 44 Gbps, il che li rende ideali per carichi di lavoro misti e applicazioni esigenti come l'inferenza ai e l'High Performance Computing (HPC).

La figura seguente mostra il sistema storage NetApp EF280.

NetApp EF280

- Supporto FC 32 GB/16 GB, iSCSI 25 GB/10 GB e SAS 12 GB

- La capacità effettiva massima è di 96 dischi per un totale di 1,5 PB
- Throughput di 10 Gbps (letture sequenziali)
- 300.000 IOPS (letture casuali)
- NetApp EF280 è l'array all-flash (AFA) più economico del portfolio NetApp

NetApp EF300

- 24 unità SSD NVMe per una capacità totale di 367 TB
- Opzioni di espansione per un totale di 240x HDD NL-SAS, 96x SSD SAS o una combinazione
- 100 GB di NVMe/IB, NVMe/RoCE, iSER/IB e SRP/IB
- 32 GB NVME/FC, FCP
- 25 GB iSCSI
- 20 Gbps (letture sequenziali)
- 670.000 IOPS (letture casuali)

Per ulteriori informazioni, consultare "Scheda informativa sugli array all-flash NetApp EF-Series EF600, F300, EF570 e EF280".

NetApp ONTAP 9

ONTAP 9.8.1, l'ultima generazione di software per la gestione dello storage NetApp, consente alle aziende di modernizzare l'infrastruttura e passare a un data center cloud-ready. 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 9.8.1 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, in modo da utilizzare le risorse appropriate per applicazioni e set di dati. ONTAP include le seguenti funzionalità per ottimizzare e semplificare le operazioni e ridurre il costo totale delle operazioni:

- **Compattazione 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 adattiva.** 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.** questa funzione 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".

Accelera e proteggi i dati

ONTAP 9 offre livelli superiori di performance e protezione dei dati ed estende queste funzionalità nei seguenti modi:

- **Prestazioni 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.
- **Crittografia dei volumi NetApp (NVE).** ONTAP offre crittografia nativa a livello di volume con supporto per la gestione delle chiavi integrata ed esterna.
- **Multitenancy 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 9 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 di 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 applicazioni emergenti.** ONTAP offre servizi dati di livello Enterprise per piattaforme e applicazioni di prossima generazione, come veicoli autonomi, città intelligenti e Industry 4.0, utilizzando la stessa infrastruttura che supporta le applicazioni aziendali esistenti.

NetApp SANtricity

NetApp SANtricity è progettato per offrire performance, affidabilità e semplicità leader di settore agli array all-flash ibridi e EF-Series. Ottieni il massimo delle performance e dell'utilizzo degli array all-flash ibridi e EF-Series per applicazioni con carichi di lavoro elevati, tra cui analisi dei dati, videosorveglianza e backup e recovery. Con SANtricity, è possibile completare la modifica della configurazione, la manutenzione, l'espansione della capacità e altre attività mentre lo storage rimane online. SANtricity offre inoltre una protezione dei dati superiore, un monitoraggio proattivo e una sicurezza certificata, il tutto accessibile tramite l'interfaccia di System Manager, semplice da utilizzare e integrata. Per ulteriori informazioni, consultare "Scheda informativa sul software NetApp e-Series SANtricity".

Prestazioni ottimizzate

Il software SANtricity ottimizzato per le performance offre dati, con IOPS elevati, throughput elevato e bassa latenza, a tutte le applicazioni di analisi dei dati, videosorveglianza e backup. Accelera le performance per applicazioni a bassa latenza, IOPS elevati e applicazioni a elevata larghezza di banda e throughput elevato.

Massimizzare l'uptime

Completa tutte le tue attività di gestione mentre lo storage rimane online. Modificare le configurazioni, eseguire la manutenzione o espandere la capacità senza interrompere l'i/O. Ottieni un'affidabilità Best-in-class con funzionalità automatizzate, configurazione online, tecnologia all'avanguardia Dynamic Disk Pools (DPP) e molto altro ancora.

Resto facile

Il software SANtricity offre una protezione dei dati superiore, un monitoraggio proattivo e una sicurezza certificata, il tutto tramite l'interfaccia di System Manager, semplice da utilizzare e integrata. Semplifica le attività di gestione dello storage. Ottieni la flessibilità necessaria per il tuning avanzato di tutti i sistemi storage

e-Series. Gestisci il tuo sistema NetApp e-Series, sempre e ovunque. La nostra interfaccia on-box basata sul web ottimizza il tuo workflow di gestione.

Trident di NetApp

"Trident" NetApp è uno storage dinamico open-source orchestrator per Docker e Kubernetes che semplifica la creazione, la gestione e il consumo dello storage persistente. Trident, un'applicazione nativa di Kubernetes, viene eseguita direttamente all'interno di un cluster Kubernetes. Trident consente ai clienti di implementare senza problemi le immagini dei container DL sullo storage NetApp e offre un'esperienza di livello Enterprise per le implementazioni dei container ai. Gli utenti di Kubernetes (come sviluppatori ML e data scientist) possono creare, gestire e automatizzare orchestrazione e cloning per sfruttare le funzionalità avanzate di gestione dei dati di NetApp basate sulla tecnologia NetApp.

Copia e sincronizzazione di NetApp BlueXP

"Copia e sincronizzazione di BlueXP" È un servizio NetApp per una sincronizzazione dei dati rapida e sicura. Sia che tu debba trasferire file tra condivisioni di file SMB o NFS on-premise, NetApp StorageGRID, NetApp ONTAP S3, NetApp Cloud Volumes Service, Azure NetApp Files, Amazon Simple Storage Service (Amazon S3), Amazon Elastic file System (Amazon EFS), BLOB di Azure, Google Cloud Storage, o IBM Cloud Object Storage, BlueXP Copy and Sync sposta i file dove ne hai bisogno in modo rapido e sicuro. Una volta trasferiti, i dati sono completamente disponibili per l'utilizzo sia sull'origine che sulla destinazione. BlueXP Copy e Sync sincronizza costantemente i dati in base alla pianificazione predefinita, spostando solo i delta, in modo da poter ridurre al minimo tempo e denaro necessari per la replica. BlueXP Copy and Sync è un tool software as a service (SaaS) estremamente semplice da configurare e utilizzare. I trasferimenti dei dati attivati da BlueXP Copy e Sync sono effettuati dai broker di dati. Puoi implementare i broker di dati BlueXP Copy e Sync in AWS, Azure, Google Cloud Platform o on-premise.

Server Lenovo ThinkSystem

I server Lenovo ThinkSystem sono dotati di hardware, software e servizi innovativi che risolvono le sfide attuali dei clienti e offrono un approccio di progettazione modulare e evolutivo, adatto allo scopo, per affrontare le sfide del futuro. Questi server si basano su tecnologie Best-in-class e standard di settore, unite a innovazioni Lenovo differenziate per offrire la massima flessibilità possibile nei server x86.

I vantaggi principali dell'implementazione dei server Lenovo ThinkSystem includono:

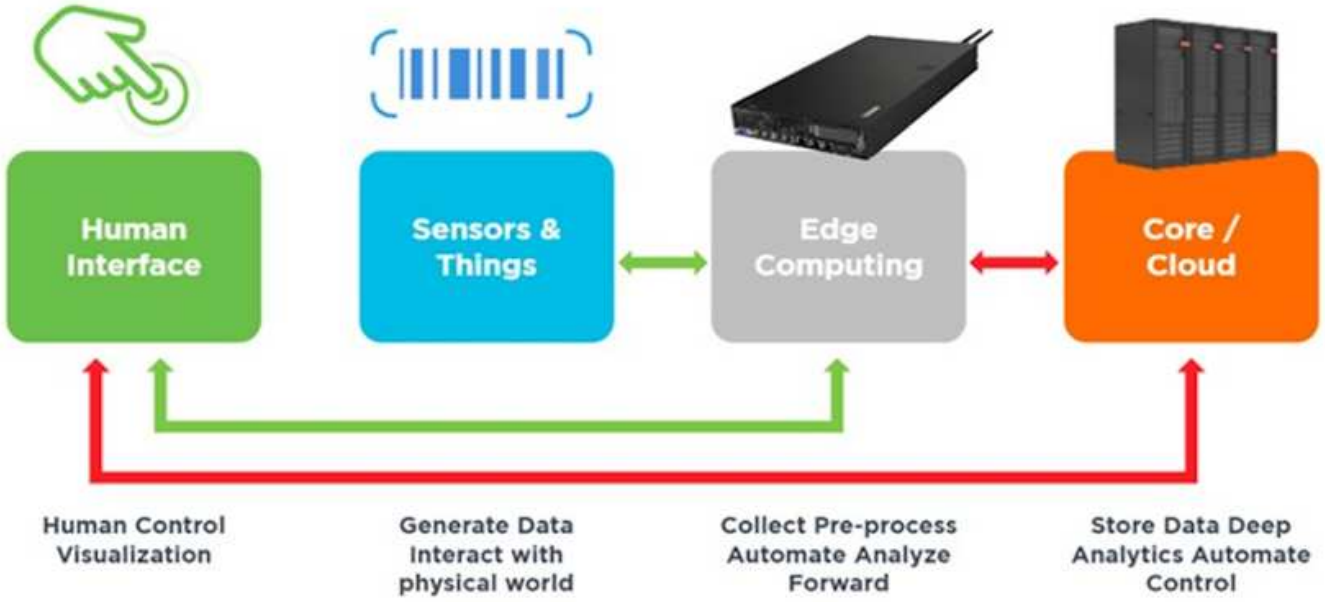
- Design altamente scalabili e modulari per crescere insieme al tuo business
- Resilienza leader del settore per risparmiare ore di costosi downtime non pianificati
- Tecnologie flash veloci per latenze inferiori, tempi di risposta più rapidi e gestione dei dati più intelligente in tempo reale

Nell'area dell'ai, Lenovo sta adottando un approccio pratico per aiutare le aziende a comprendere e adottare i vantaggi di ML e ai per i propri carichi di lavoro. I clienti Lenovo possono esplorare e valutare le offerte Lenovo ai nei Lenovo ai Innovation Center per comprendere appieno il valore del loro caso di utilizzo specifico. Per migliorare il time-to-value, questo approccio incentrato sul cliente offre ai clienti una prova di concetto per le piattaforme di sviluppo di soluzioni pronte all'uso e ottimizzate per l'ai.

Lenovo ThinkSystem SE350 Edge Server

L'edge computing consente di analizzare i dati provenienti dai dispositivi IoT all'edge della rete prima di inviarli al data center o al cloud. Lenovo ThinkSystem SE350, come illustrato nella figura seguente, è progettato per soddisfare i requisiti esclusivi di implementazione alla periferia della rete, con particolare attenzione a flessibilità, connettività, sicurezza e gestibilità remota in un fattore di forma compatto e rinforzato dal punto di vista ambientale.

Dotato del processore Intel Xeon D con la flessibilità di supportare l'accelerazione per i carichi di lavoro ai edge, il SE350 è costruito ad hoc per affrontare la sfida delle implementazioni dei server in una varietà di ambienti esterni al data center.



MLPerf

MLPerf è la suite di benchmark leader del settore per la valutazione delle performance ai. Copre molte aree dell'ai applicata, tra cui classificazione delle immagini, rilevamento degli oggetti, imaging medico e NLP (Natural Language Processing). In questa convalida, abbiamo utilizzato i carichi di lavoro Inference v0.7, che è l'ultima iterazione dell'inferenza MLPerf al completamento di questa convalida. Il "MLPerf Inference v0.7" la suite include quattro nuovi benchmark per data center e sistemi edge:

- **BERT.** rappresentazione del codificatore bidirezionale da Transformers (BERT) ottimizzata per la risposta alle domande utilizzando il set di dati della squadra.
- **DLRM.** Deep Learning Recommendation Model (DLRM) è un modello di personalizzazione e raccomandazione che viene addestrato per ottimizzare i tassi di click-through (CTR).

- **3D U-Net.** l'architettura 3D U-Net viene addestrata sul set di dati Brain Tumor Segmentation (Brats).
- **RNN-T.** il trasduttore di rete neurale ricorrente (RNN-T) è un modello di riconoscimento vocale automatico (ASR) che viene addestrato su un sottoinsieme di LibriSpeech. I risultati e il codice dell'inferenza MLPerf sono pubblicamente disponibili e rilasciati sotto licenza Apache. MLPerf Inference dispone di una divisione Edge, che supporta i seguenti scenari:
- **Single stream.** questo scenario imita i sistemi in cui la reattività è un fattore critico, come le query ai offline eseguite sugli smartphone. Le singole query vengono inviate al sistema e i tempi di risposta vengono registrati. come risultato viene riportata la latenza del 90° percentile di tutte le risposte.
- **Multistream.** questo benchmark è per i sistemi che elaborano input da più sensori. Durante il test, le query vengono inviate a un intervallo di tempo fisso. Viene imposto un vincolo QoS (latenza massima consentita). Il test indica il numero di flussi che il sistema può elaborare rispettando il limite di QoS.
- **Offline.** questo è lo scenario più semplice che copre le applicazioni di elaborazione in batch e la metrica è il throughput in campioni al secondo. Tutti i dati sono disponibili per il sistema e il benchmark misura il tempo necessario per elaborare tutti i campioni.

Lenovo ha pubblicato i punteggi di inferenza MLPerf per SE350 con T4, il server utilizzato in questo documento. Vedere i risultati all'indirizzo "https://mlperf.org/inference-results-0-7/" Nella sezione "Edge, CLOSED Division" della voce 0.7-145.

Piano di test

Questo documento segue l'inferenza MLPerf v0.7 "codice", MLPerf Inference v1.1 "codice", e. "regole". Abbiamo eseguito benchmark MLPerf progettati per l'inferenza ai margini, come definito nella tabella seguente.

Area	Attività	Modello	Dataset	Dimensione QSL	Qualità	Vincolo di latenza multi-stream
Visione	Classificazione delle immagini	Resnet50v1.5	ImageNet (224 x 224)	1024	99% del 32° PQ	50 ms.
Visione	Rilevamento di oggetti (grande)	SSD-ResNet34	COCO (1200 x 1200)	64	99% del 32° PQ	66 ms.
Visione	Rilevamento di oggetti (piccolo)	SSD - MobileNetsv1	COCO (300 x 300)	256	99% del 32° PQ	50 ms.
Visione	Segmentazione delle immagini mediche	UNET 3D	Brat 2019 (224 x 224 x 160)	16	99% e 99.9% del 32° PQ	n/a.
Discorso	Voce-testo	RNNT	Sviluppo di Librispeech-clean	2513	99% del 32° PQ	n/a.
Lingua	Elaborazione della lingua	BERT	Squadra v1.1	10833	99% del 32° PQ	n/a.

Nella tabella seguente sono illustrati gli scenari di benchmark Edge.

Area	Attività	Scenari
Visione	Classificazione delle immagini	Single stream, offline, multistream
Visione	Rilevamento di oggetti (grande)	Single stream, offline, multistream
Visione	Rilevamento di oggetti (piccolo)	Single stream, offline, multistream
Visione	Segmentazione delle immagini mediche	Single stream, offline
Discorso	Voce-testo	Single stream, offline
Lingua	Elaborazione della lingua	Single stream, offline

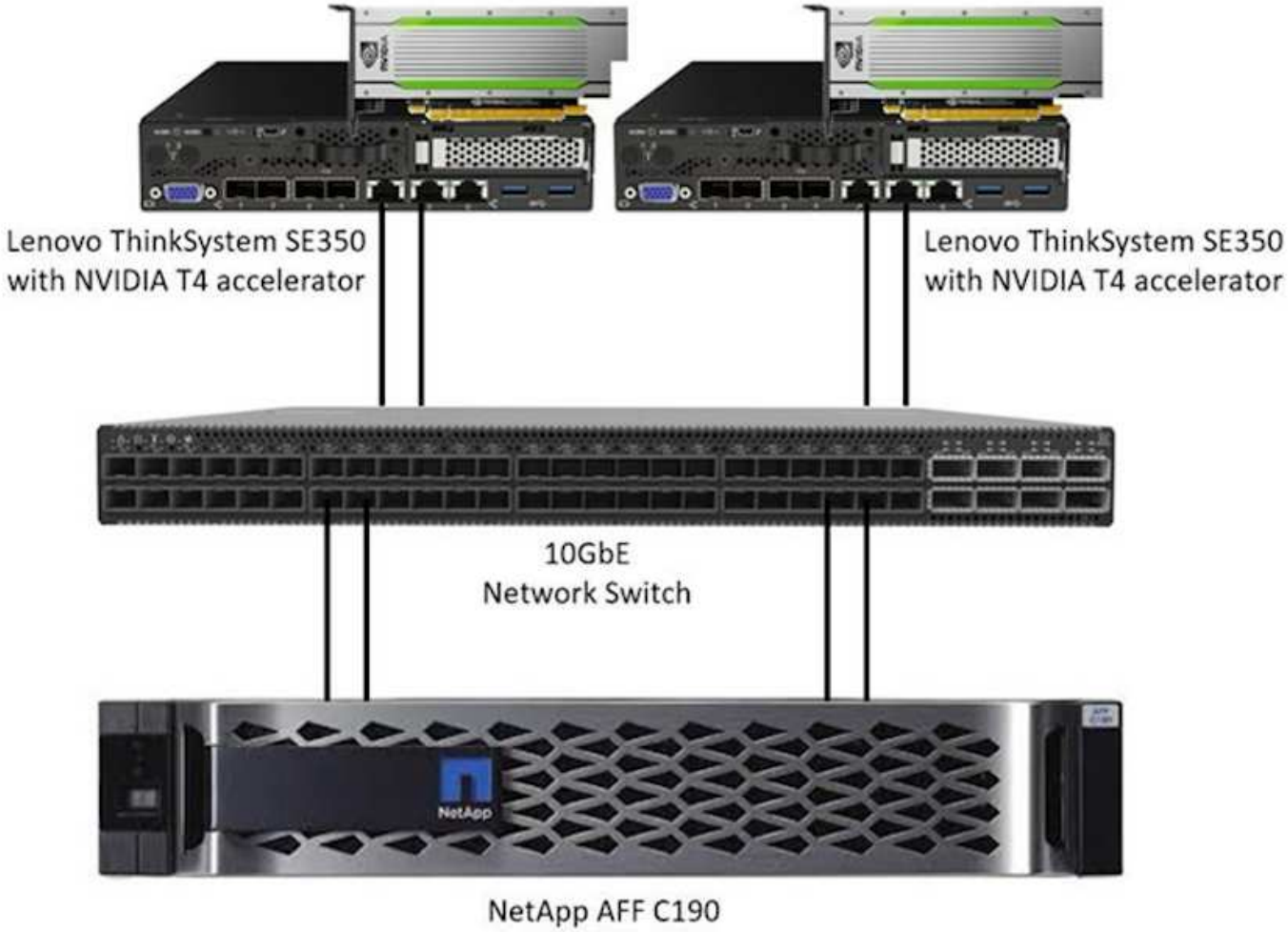
Abbiamo eseguito questi benchmark utilizzando l'architettura di storage di rete sviluppata in questa convalida e confrontato i risultati con quelli delle esecuzioni locali sugli edge server precedentemente inviati a MLPerf. Il confronto serve a determinare l'impatto dello storage condiviso sulle performance di inferenza.

Eseguire il test della configurazione

La seguente figura mostra la configurazione del test. Abbiamo utilizzato il sistema storage NetApp AFF C190 e due server Lenovo ThinkSystem SE350 (ciascuno con un acceleratore NVIDIA T4). Questi componenti sono collegati tramite uno switch di rete 10 GbE. Lo storage di rete contiene set di dati di convalida/test e modelli preformati. I server offrono funzionalità di calcolo e l'accesso allo storage avviene attraverso il protocollo NFS.

Questa sezione descrive le configurazioni testate, l'infrastruttura di rete, il server SE350 e i dettagli relativi al provisioning dello storage. La tabella seguente elenca i componenti di base per l'architettura della soluzione.

Componenti della soluzione	Dettagli
Server Lenovo ThinkSystem	• 2 server SE350 ciascuno con una scheda NVIDIA T4 GPU
	• Ogni server contiene una CPU Intel Xeon D-2123IT con quattro core fisici a 2,20 GHz e 128 GB di RAM
Sistema storage entry-level NetApp AFF (coppia ha)	• Software NetApp ONTAP 9 • 24x 960 GB SSD • Protocollo NFS • Un gruppo di interfacce per controller, con quattro indirizzi IP logici per i punti di montaggio



La seguente tabella elenca la configurazione dello storage: AFF C190 con 2 RU, 24 slot per unità.

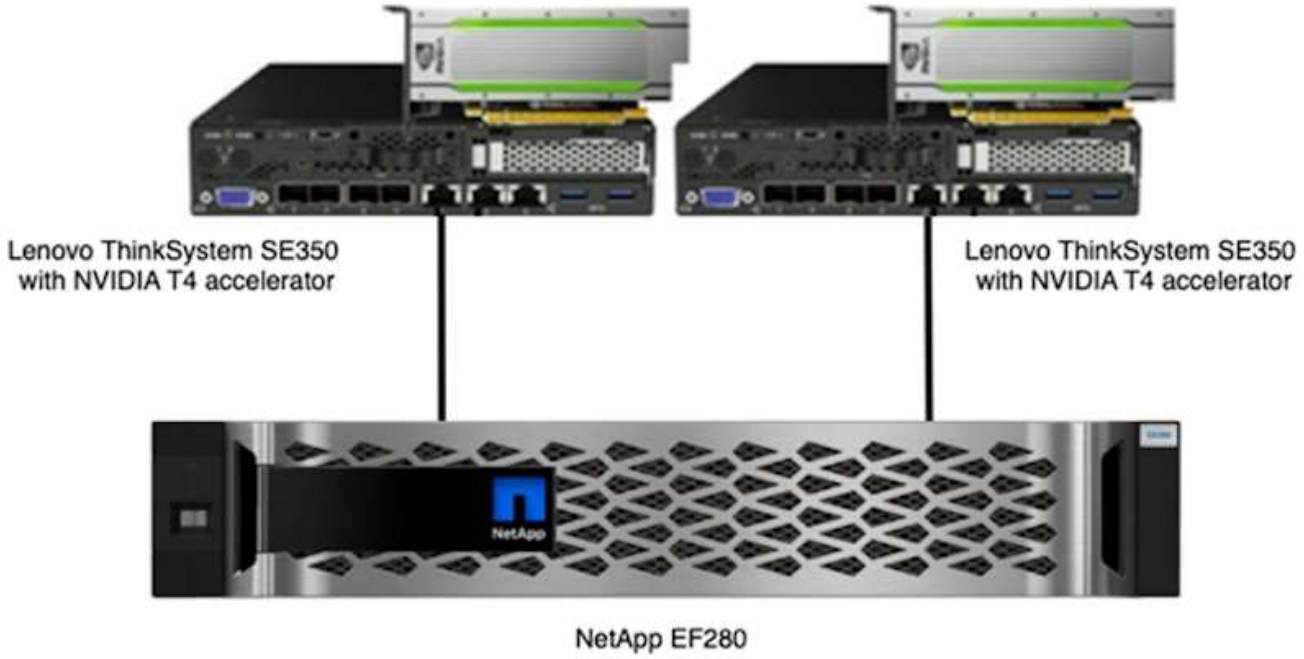
Controller	Aggregato	Volume FlexGroup	Aggregatesize	Volumesize (Volumesize)	Punto di montaggio del sistema operativo
Controller 1	Aggr1	/Netappelenovo_ai_fg	8,42TiB	15 TB	/netapp_lenovo_fg
Controller 2	Aggr2		8,42TiB		

La cartella /netappLenovo_ai_fg contiene i set di dati utilizzati per la convalida del modello.

La figura seguente mostra la configurazione del test. Abbiamo utilizzato il sistema storage NetApp EF280 e due server Lenovo ThinkSystem SE350 (ciascuno con un acceleratore NVIDIA T4). Questi componenti sono collegati tramite uno switch di rete 10 GbE. Lo storage di rete contiene set di dati di convalida/test e modelli preformati. I server offrono funzionalità di calcolo e l'accesso allo storage avviene attraverso il protocollo NFS.

La seguente tabella elenca la configurazione dello storage per EF280.

Controller	Gruppo di volumi	Volume	Volumesize (Volumesize)	DDDP	Metodo di connessione
Controller 1	DDP1	Volume 1	8,42TiB	16 TB	Da SE350-1 a LUN iSCSI 0
Controller 2		Volume 2	8,42TiB		Da SE350-2 a LUN iSCSI 1



Procedura di test

Questa sezione descrive le procedure di test utilizzate per validare questa soluzione.

Configurazione del sistema operativo e dell'inferenza ai

Per AFF C190, abbiamo utilizzato Ubuntu 18.04 con driver NVIDIA e docker con supporto per GPU NVIDIA e abbiamo utilizzato MLPerf "codice" Disponibile come parte dell'invio di Lenovo a MLPerf Inference v0.7.

Per EF280, abbiamo utilizzato Ubuntu 20.04 con driver NVIDIA e docker con supporto per GPU NVIDIA e MLPerf "codice" Disponibile come parte dell'invio di Lenovo a MLPerf Inference v1.1.

Per impostare l'inferenza ai, segui questi passaggi:

1. Scarica i set di dati che richiedono la registrazione, il set di convalida ImageNet 2012, il set di dati Criteo Terabyte e il set di training Brats 2019, quindi decomprimere i file.
2. Creare una directory di lavoro con almeno 1 TB e definire la variabile ambientale `MLPERF_SCRATCH_PATH` facendo riferimento alla directory.

   È necessario condividere questa directory sullo storage condiviso per il caso di utilizzo dello storage di rete o sul disco locale durante il test con dati locali.

3. Esegui il make `prebuild` che crea e avvia il contenitore del docker per le attività di inferenza richieste.

> I seguenti comandi vengono eseguiti tutti dall'interno del contenitore di docker in esecuzione:

- Scarica i modelli ai preformati per le attività di inferenza MLPerf: `make download_model`
- Scarica altri set di dati scaricabili gratuitamente: `make download_data`
- Pre-elaborare i dati: Make `preprocess_data`
- Esecuzione: `make build`.
- Creazione di motori di inferenza ottimizzati per la GPU nei server di calcolo: `make generate_engines`
- Per eseguire i carichi di lavoro di inferenza, eseguire quanto segue (un comando):

```
make run_harness RUN_ARGS="--benchmarks=<BENCHMARKS>
--scenarios=<SCENARIOS>"
```

L'inferenza ai è in esecuzione

Sono stati eseguiti tre tipi di esecuzione:

- Inferenza ai su server singolo utilizzando lo storage locale
- Inferenza ai su server singolo utilizzando lo storage di rete
- Inferenza ai multi-server utilizzando lo storage di rete

Risultati del test

Sono stati eseguiti numerosi test per valutare le performance dell'architettura proposta.

Esistono sei diversi carichi di lavoro (classificazione delle immagini, rilevamento degli oggetti [piccoli], rilevamento degli oggetti [grandi], imaging medico, comunicazione vocale, E Natural Language Processing [NLP]), che è possibile eseguire in tre diversi scenari: Offline, single stream e multistream.

> L'ultimo scenario viene implementato solo per la classificazione delle immagini e il rilevamento degli oggetti.

Ciò offre 15 possibili carichi di lavoro, tutti testati in tre diverse configurazioni:

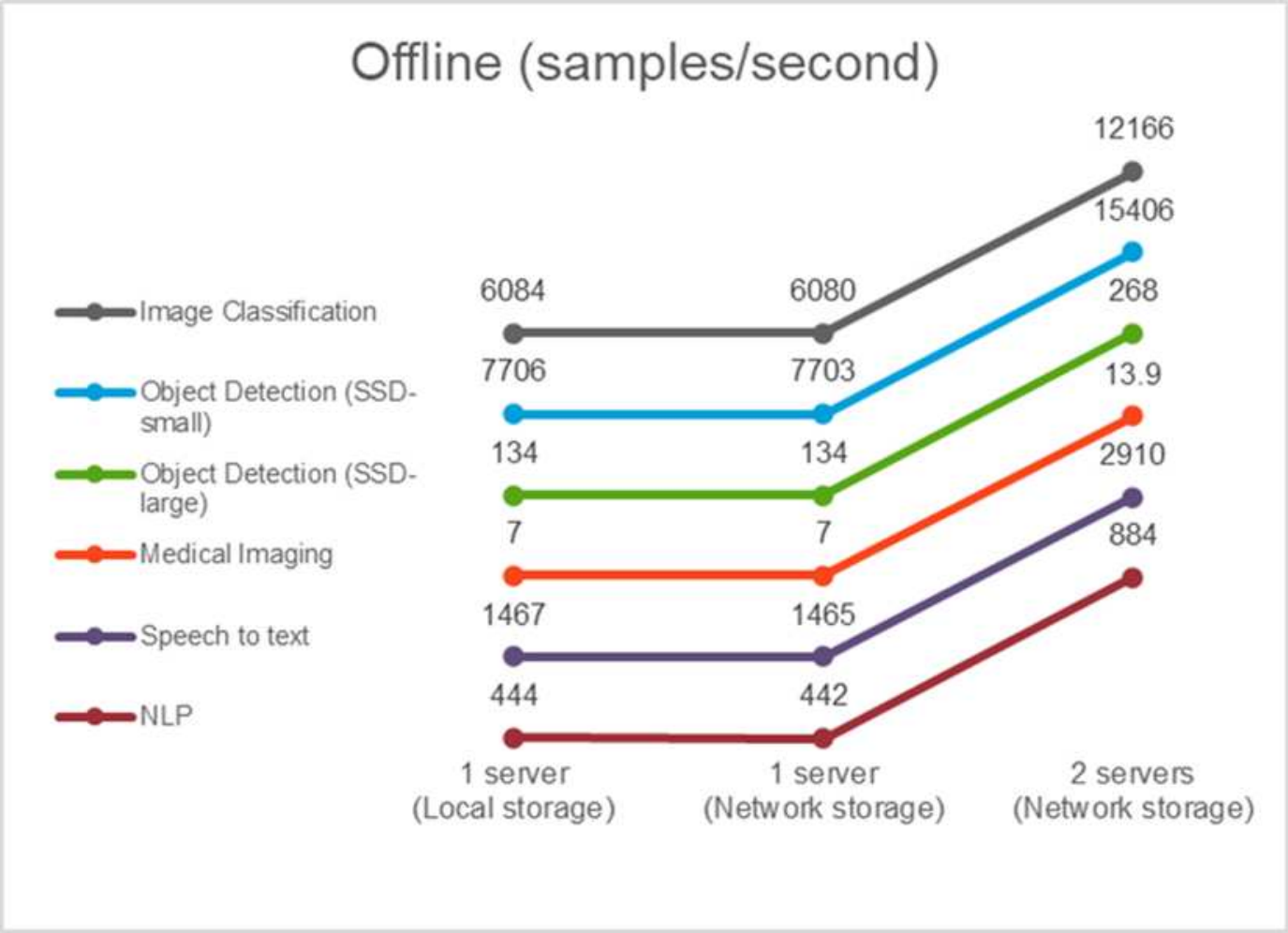
- Server singolo/storage locale
- Storage di rete/server singolo
- Storage di rete/multi-server

I risultati sono descritti nelle sezioni seguenti.

Inferenza ai nello scenario offline per AFF

In questo scenario, tutti i dati erano disponibili per il server e il tempo impiegato per elaborare tutti i campioni è

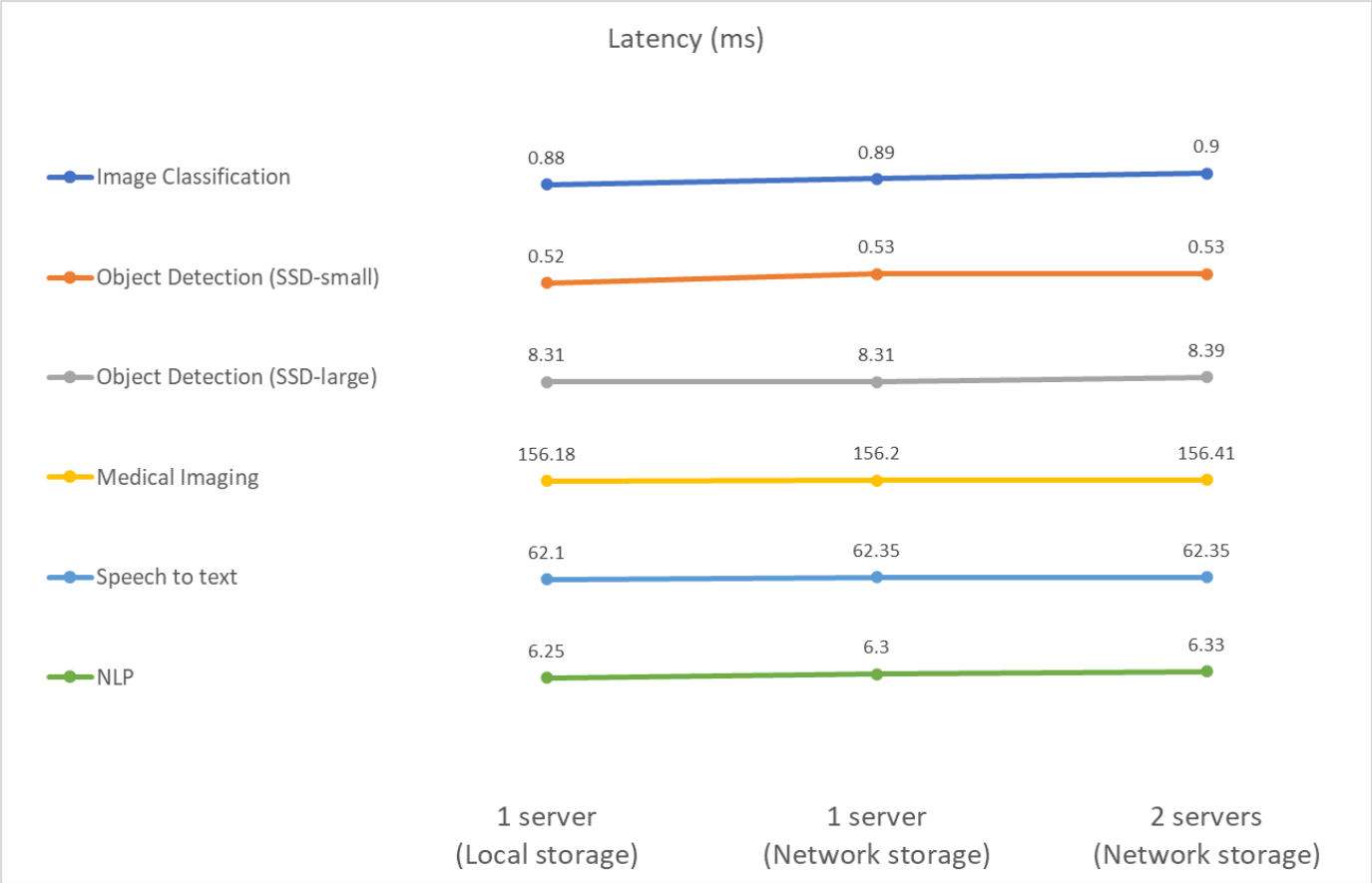
stato misurato. I risultati dei test riportano le larghezze di banda in campioni al secondo. Quando sono stati utilizzati più server di calcolo, si riporta la larghezza di banda totale sommata su tutti i server. I risultati per tutti e tre i casi di utilizzo sono mostrati nella figura seguente. Per il caso di due server, segnaliamo la larghezza di banda combinata di entrambi i server.



I risultati mostrano che lo storage di rete non influisce negativamente sulle performance: La modifica è minima e per alcune attività non viene rilevata alcuna. Quando si aggiunge il secondo server, la larghezza di banda totale raddoppia esattamente o, nel peggiore dei casi, la modifica è inferiore all'1%.

Inferenza ai in uno scenario a flusso singolo per AFF

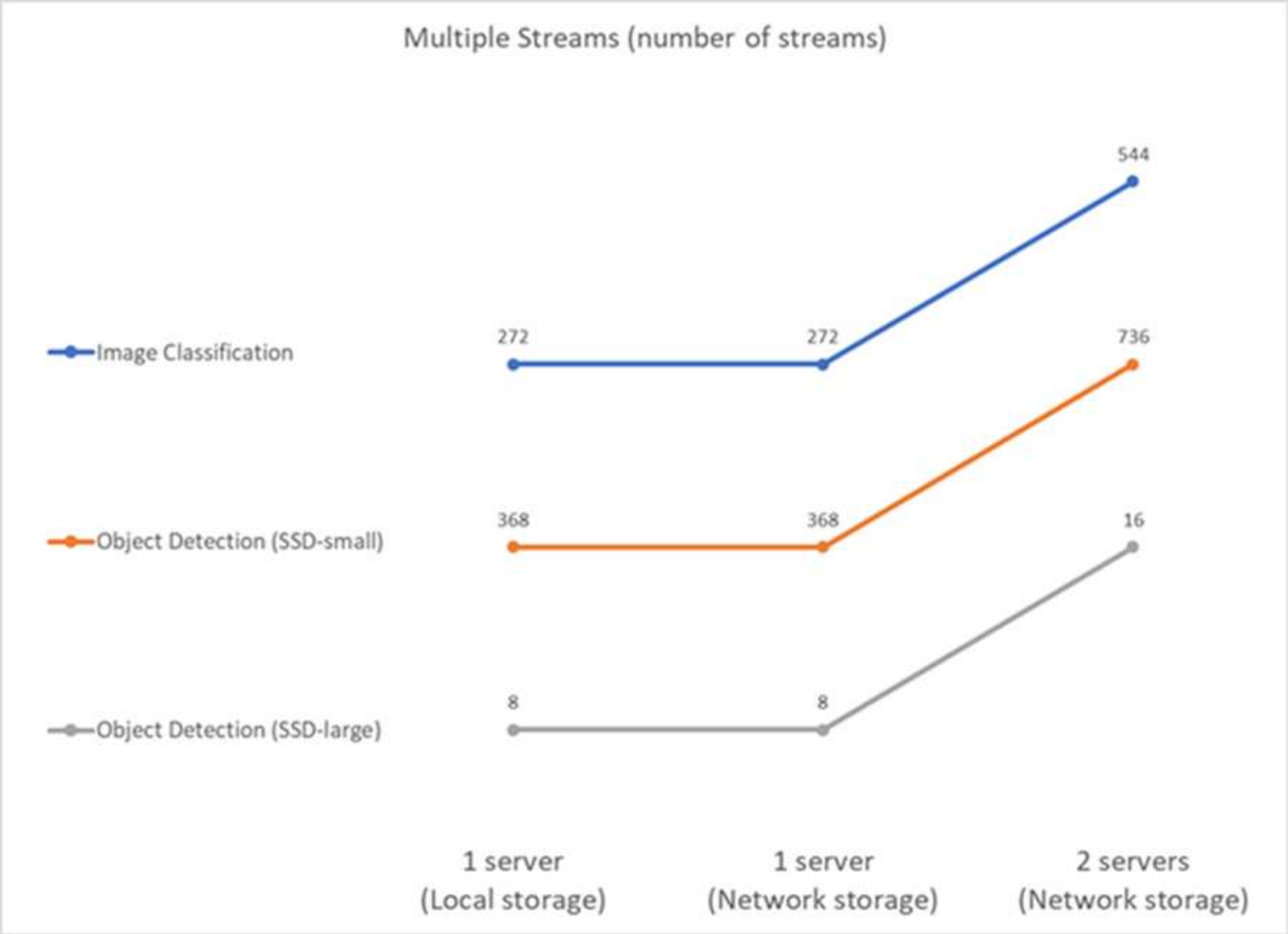
Questo benchmark misura la latenza. Per il caso di più server di calcolo, segnaliamo la latenza media. I risultati per la suite di attività sono riportati nella figura seguente. Per il caso di due server, segnaliamo la latenza media di entrambi i server.



I risultati, ancora una volta, mostrano che lo storage di rete è sufficiente per gestire le attività. La differenza tra storage locale e di rete nel caso di un server è minima o nulla. Allo stesso modo, quando due server utilizzano lo stesso storage, la latenza su entrambi i server rimane la stessa o cambia di molto.

Inferenza ai nello scenario multistream per AFF

In questo caso, il risultato è il numero di flussi che il sistema è in grado di gestire, soddisfacendo al tempo stesso il limite di QoS. Pertanto, il risultato è sempre un numero intero. Per più di un server, viene riportato il numero totale di flussi sommati su tutti i server. Non tutti i carichi di lavoro supportano questo scenario, ma abbiamo eseguito quelli che lo fanno. I risultati dei nostri test sono riassunti nella figura seguente. Per il caso di due server, segnaliamo il numero combinato di flussi da entrambi i server.



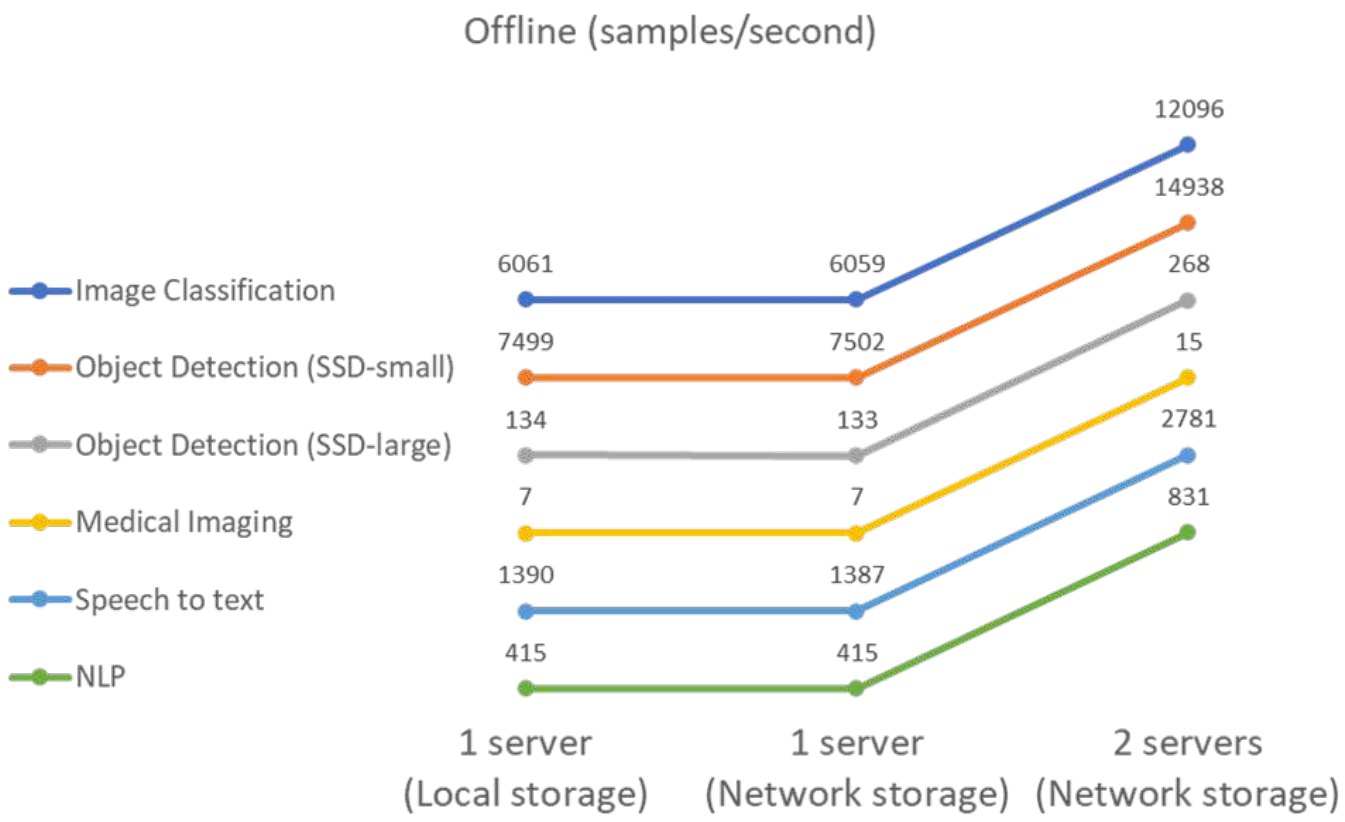
I risultati mostrano le performance perfette dell'installazione: Lo storage locale e di rete offrono gli stessi risultati e l'aggiunta del secondo server raddoppia il numero di flussi gestibili dall'installazione proposta.

Risultati del test per EF

Sono stati eseguiti numerosi test per valutare le performance dell'architettura proposta. Esistono sei diversi carichi di lavoro (classificazione delle immagini, rilevamento degli oggetti [piccoli], rilevamento degli oggetti [grandi], imaging medico, comunicazione vocale, E Natural Language Processing [NLP]), eseguiti in due scenari diversi: Offline e single stream. I risultati sono descritti nelle sezioni seguenti.

Inferenza ai nello scenario offline per EF

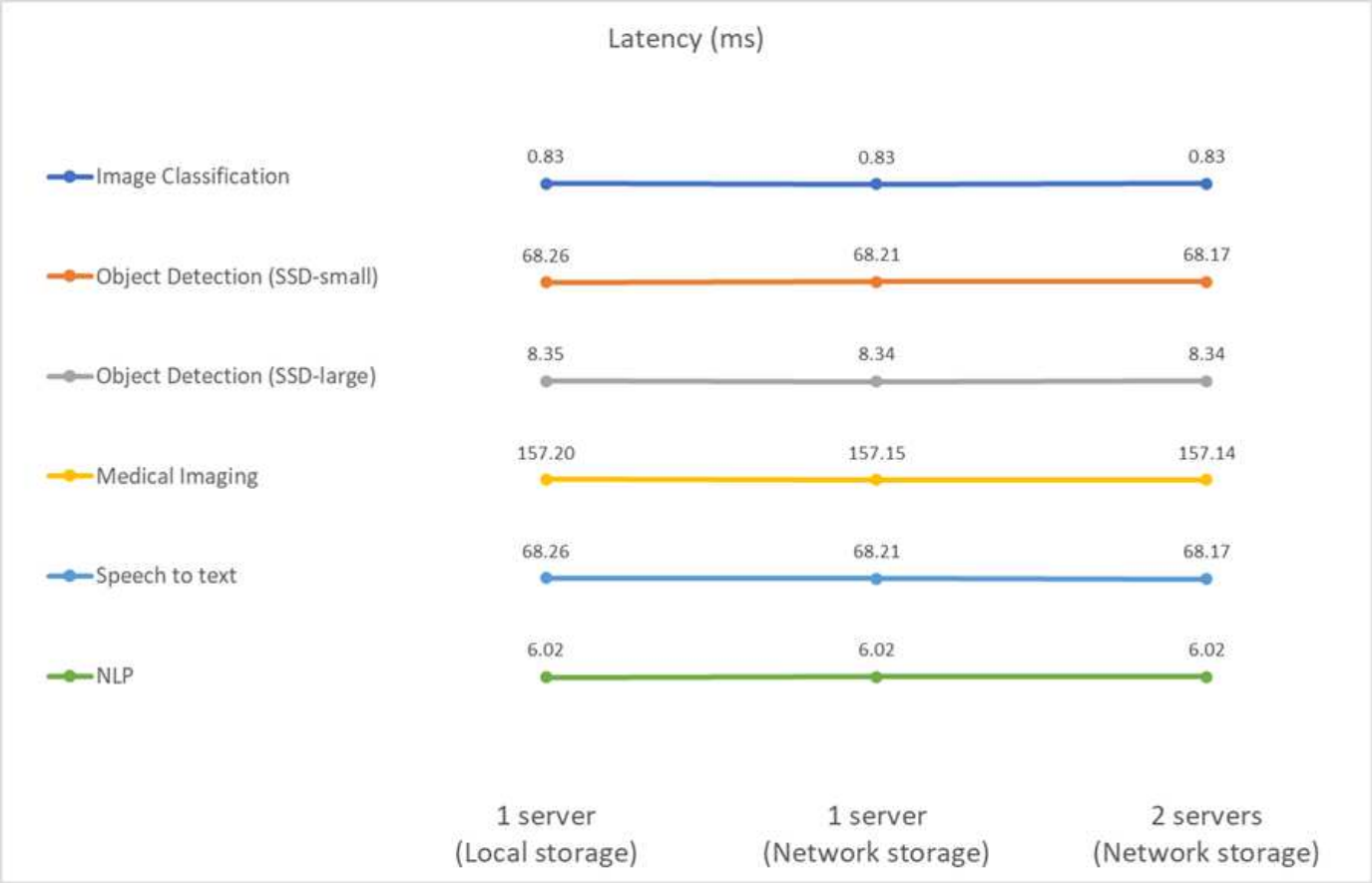
In questo scenario, tutti i dati erano disponibili per il server e il tempo impiegato per elaborare tutti i campioni è stato misurato. I risultati dei test riportano le larghezze di banda in campioni al secondo. Per le esecuzioni a nodo singolo, segnaliamo la media di entrambi i server, mentre per due esecuzioni a server segnaliamo la larghezza di banda totale sommata su tutti i server. I risultati dei casi di utilizzo sono mostrati nella figura seguente.



I risultati mostrano che lo storage di rete non influisce negativamente sulle performance: La modifica è minima e per alcune attività non viene rilevata alcuna. Quando si aggiunge il secondo server, la larghezza di banda totale raddoppia esattamente o, nel peggiore dei casi, la modifica è inferiore all'1%.

Inferenza ai in uno scenario a flusso singolo per EF

Questo benchmark misura la latenza. In tutti i casi, segnaliamo la latenza media su tutti i server coinvolti nelle esecuzioni. Vengono forniti i risultati per la suite di attività.



I risultati mostrano ancora una volta che lo storage di rete è sufficiente per gestire le attività. La differenza tra lo storage locale e di rete nel caso di un server è minima o nulla. Allo stesso modo, quando due server utilizzano lo stesso storage, la latenza su entrambi i server rimane la stessa o cambia di molto.

Opzioni di dimensionamento dell'architettura

È possibile regolare l'impostazione utilizzata per la convalida in modo che si adatti ad altri casi di utilizzo.

Server di calcolo

Abbiamo utilizzato una CPU Intel Xeon D-2123IT, che è il livello più basso di CPU supportato in SE350, con quattro core fisici e TDP da 60 W. Anche se il server non supporta la sostituzione delle CPU, può essere ordinato con una CPU più potente. La CPU più alta supportata è Intel Xeon D-2183IT con 16 core, 100 W a 2,20 GHz. Ciò aumenta notevolmente la capacità di calcolo della CPU. Anche se la CPU non era un collo di bottiglia per l'esecuzione dei carichi di lavoro di inferenza, aiuta nell'elaborazione dei dati e in altre attività correlate all'inferenza. Attualmente, NVIDIA T4 è l'unica GPU disponibile per i casi di utilizzo edge; pertanto, attualmente, non è possibile aggiornare o eseguire il downgrade della GPU.

Storage condiviso

Per il test e la convalida, il sistema NetApp AFF C190, con una capacità di storage massima di 50,5 TB, un throughput di 4,4 Gbps per letture sequenziali e 230 K IOPS per letture casuali di piccole dimensioni, è stato utilizzato per lo scopo di questo documento ed è stato dimostrato adatto per i carichi di lavoro di inferenza edge.

Tuttavia, se è necessaria una maggiore capacità di storage o velocità di rete più elevate, è consigliabile

utilizzare i sistemi di storage NetApp AFF A220 o NetApp AFF A250. Inoltre, per la convalida della soluzione è stato utilizzato anche il sistema NetApp EF280, con una capacità massima di 1,5 PB e una larghezza di banda di 10 Gbps. È possibile utilizzare una maggiore capacità di storage con una larghezza di banda superiore "NetApp EF300" .

Conclusione

L'automazione basata sull'ai e l'edge computing sono un approccio leader per aiutare le organizzazioni aziendali a ottenere la trasformazione digitale e massimizzare l'efficienza e la sicurezza delle operazioni. Con l'edge computing, i dati vengono elaborati molto più velocemente perché non devono viaggiare da e verso un data center. Di conseguenza, il costo associato all'invio dei dati avanti e indietro ai data center o al cloud è diminuito. Una latenza inferiore e una velocità maggiore possono essere utili quando le aziende devono prendere decisioni quasi in tempo reale utilizzando modelli di inferenza ai implementati all'edge della rete.

I sistemi storage NetApp offrono performance uguali o migliori rispetto allo storage SSD locale e offrono i seguenti vantaggi a data scientist, data engineer, sviluppatori ai/ML e decision maker aziendali o IT:

- Condivisione semplice dei dati tra sistemi ai, analytics e altri sistemi aziendali critici. Questa condivisione dei dati riduce l'overhead dell'infrastruttura, migliora le performance e ottimizza la gestione dei dati in tutta l'azienda.
- Calcolo e storage scalabili in maniera indipendente per ridurre al minimo i costi e migliorare l'utilizzo delle risorse.
- Workflow di sviluppo e implementazione ottimizzati grazie a copie Snapshot e cloni integrati per spazi di lavoro degli utenti istantanei ed efficienti in termini di spazio, controllo integrato delle versioni e implementazione automatizzata.
- Protezione dei dati di livello Enterprise per disaster recovery e business continuity. La soluzione NetApp e Lenovo presentata in questo documento è un'architettura scalabile e flessibile, ideale per le implementazioni di inferenza ai di livello Enterprise all'edge della rete.

Ringraziamenti

- J.J. Falkanger, Sr Manager, soluzioni HPC e ai, Lenovo
- Dave Arnette, Technical Marketing Engineer, NetApp
- Joey Parnell, Tech Lead e-Series ai Solutions, NetApp
- Cody Harryman, QA Engineer, NetApp

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i seguenti documenti e/o siti Web:

- Pagina del prodotto array NetApp AFF A-Series

  "https://www.netapp.com/data-storage/aff-a-series/"

- Software per la gestione dei dati NetApp ONTAP: Libreria di informazioni ONTAP 9

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

- TR-4727: Introduzione a NetApp EF-Series

  https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf

- Scheda informativa sul software NetApp e-Series SANtricity

  https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf

- Storage persistente NetApp per container - NetApp Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- MLPerf
  - "https://mlcommons.org/en/"
  - "http://www.image-net.org/"
  - "https://mlcommons.org/en/news/mlperf-inference-v11/"
- Copia e sincronizzazione di NetApp BlueXP

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"

- Benchmark TensorFlow

  "https://github.com/tensorflow/benchmarks"

- Lenovo ThinkSystem SE350 Edge Server

  "https://lenovopress.com/lp1168"

- Array di storage flash unificato Lenovo ThinkSystem DM5100F

  "https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"

WP-7328: Ai di NetApp Conversational con NVIDIA Jarvis

Rick Huang, Sung-Han Lin, NetApp Davide Onofrio, NVIDIA

La famiglia di sistemi NVIDIA DGX è costituita dai primi sistemi al mondo basati su intelligenza artificiale integrata (ai) costruiti appositamente per l'ai aziendale. I sistemi storage NetApp AFF offrono performance estreme e funzionalità di gestione dei dati del cloud ibrido leader di settore. NetApp e NVIDIA hanno collaborato per creare l'architettura di riferimento ai di NetApp ONTAP, una soluzione chiavi in mano per i carichi di lavoro di ai e machine learning (ML) che offre performance, affidabilità e supporto di livello Enterprise.

Questo white paper fornisce una guida direzionale ai clienti che sviluppano sistemi di ai conversazionali a supporto di diversi casi di utilizzo in diversi mercati verticali del settore. Include informazioni sull'implementazione del sistema utilizzando NVIDIA Jarvis. I test sono stati eseguiti utilizzando una stazione NVIDIA DGX e un sistema storage NetApp AFF A220.

Il pubblico di riferimento per la soluzione comprende i seguenti gruppi:

- Enterprise Architect che progettano soluzioni per lo sviluppo di modelli di ai e software per casi di utilizzo conversa dell'ai, come un assistente virtuale al dettaglio
- Data scientist alla ricerca di modi efficienti per raggiungere gli obiettivi di sviluppo della modellazione linguistica
- Data engineer incaricati di gestire ed elaborare dati di testo come domande dei clienti e trascrizioni di dialoghi
- Dirigenti e responsabili delle decisioni IT e business leader interessati a trasformare l'esperienza di intelligenza artificiale conversazionale e a ottenere il più rapido time-to-market dalle iniziative di intelligenza artificiale

Panoramica della soluzione

Questo documento offre una panoramica del modello ai conversazionale per ONTAP ai e NVIDIA DGX.

NetApp ONTAP ai e BlueXP Copy and Sync

L'architettura NetApp ONTAP ai, basata su sistemi NVIDIA DGX e sistemi storage connessi al cloud, è stata sviluppata e verificata da NetApp e NVIDIA. Questa architettura di riferimento offre alle organizzazioni IT i seguenti vantaggi:

- Elimina le complessità di progettazione
- Consente una scalabilità indipendente di calcolo e storage
- Consente ai clienti di partire da piccoli e scalare perfettamente
- Offre una vasta gamma di opzioni di storage per diverse esigenze di performance e costi NetApp ONTAP ai integra perfettamente i sistemi DGX e i sistemi storage NetApp AFF A220 con networking all'avanguardia. I sistemi NetApp ONTAP ai e DGX semplificano le implementazioni ai eliminando la complessità e le congetture di progettazione. I clienti possono iniziare a crescere in maniera ininterrotta e allo stesso tempo gestire in modo intelligente i dati dall'edge al core, fino al cloud e viceversa.

NetApp BlueXP Copy e Sync ti permette di spostare facilmente i dati tra vari protocolli, tra due NFS share, due CIFS share, oppure un file share e lo storage Amazon S3, Amazon Elastic file System (EFS) o Azure Blob. Il funzionamento Active-Active consente di continuare a lavorare contemporaneamente con l'origine e la destinazione, sincronizzando in modo incrementale le modifiche dei dati quando necessario. Consentendoti di spostare e sincronizzare in modo incrementale i dati tra qualsiasi sistema di origine e destinazione, sia on-premise che basato sul cloud, BlueXP Copy and Sync apre una vasta gamma di nuovi modi in cui puoi utilizzare i dati. La migrazione dei dati tra sistemi on-premise, cloud on-boarding e migrazione del cloud o collaboration e analytics dei dati diventa facilmente realizzabile. La figura seguente mostra le fonti e le destinazioni disponibili.

Nei sistemi ai conversazionali, gli sviluppatori possono sfruttare BlueXP Copy e Sync per archiviare la cronologia delle conversazioni dal cloud ai data center, consentendo il training offline dei modelli di elaborazione del linguaggio naturale (NLP). Attraverso modelli di training per riconoscere più intenti, il sistema di ai convergenti sarà meglio attrezzato per gestire domande più complesse da parte degli utenti finali.

Framework multimodale NVIDIA Jarvis

"NVIDIA Jarvis" È un framework end-to-end per la creazione di servizi di ai conversivi. Include i seguenti servizi ottimizzati per GPU:

- Riconoscimento vocale automatico (ASR)
- Comprensione del linguaggio naturale (NLU)
- Integrazione con servizi di adempimento specifici del dominio
- Text-to-speech (TTS)
- I servizi basati su computer Vision (CV) Jarvis utilizzano modelli di deep learning all'avanguardia per affrontare il complesso e impegnativo compito dell'ai conversazionale in tempo reale. Per consentire un'interazione naturale e in tempo reale con un utente finale, i modelli devono completare il calcolo in meno di 300 millisecondi. Le interazioni naturali sono impegnative e richiedono un'integrazione sensoriale multimodale. Anche le pipeline dei modelli sono complesse e richiedono un coordinamento tra i servizi indicati sopra.

Jarvis è un framework applicativo completamente accelerato per la creazione di servizi ai di conversazione multimodale che utilizzano una pipeline di deep learning end-to-end. Il framework Jarvis include modelli di ai conversazionali preformati, strumenti e servizi end-to-end ottimizzati per le attività vocali, di visione e NLU. Oltre ai servizi di intelligenza artificiale, Jarvis ti consente di unire contemporaneamente vision, audio e altri input dei sensori per offrire funzionalità come conversazioni multi-utente e multi-contesto in applicazioni come assistenti virtuali, diarizzazione multiutente e assistenti di call center.

NVIDIA NEMO

"NVIDIA NEMO" È un toolkit Python open-source per la creazione, la formazione e la messa a punto di modelli di ai conversazionali allo stato dell'arte con accelerazione GPU utilizzando interfacce di programmazione applicativa (API) di facile utilizzo. NEMO esegue calcoli misti di precisione utilizzando core Tensor in GPU NVIDIA e può scalare facilmente fino a più GPU per offrire le migliori performance di training possibili. NEMO viene utilizzato per creare modelli per applicazioni ASR, NLP e TTS in tempo reale, come trascrizioni di videochiamate, assistenti video intelligenti e supporto automatizzato di call center in diversi mercati verticali del settore, tra cui settore sanitario, finanziario, retail e telecomunicazioni.

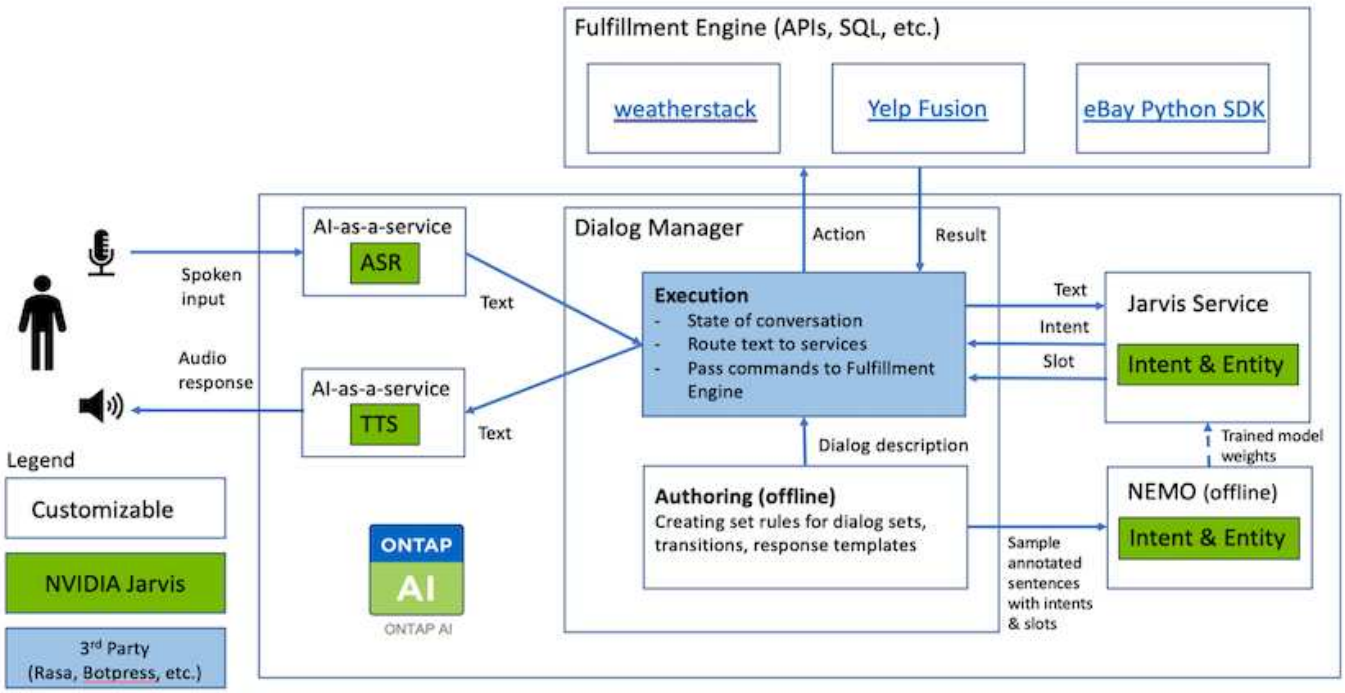
Abbiamo utilizzato NEMO per formare modelli che riconoscono intenti complessi dalle domande degli utenti nella cronologia delle conversazioni archiviate. Questo training estende le funzionalità dell'assistente virtuale al dettaglio oltre il supporto fornito da Jarvis.

Riepilogo dei casi di utilizzo al dettaglio

Utilizzando NVIDIA Jarvis, abbiamo creato un assistente virtuale al dettaglio che accetta l'input vocale o di testo e risponde a domande relative a meteo, punti di interesse e prezzi di inventario. Il sistema ai conversazionale è in grado di ricordare il flusso di conversazione, ad esempio, porre una domanda di follow-up se l'utente non specifica la posizione per il meteo o i punti di interesse. Il sistema riconosce anche entità complesse come "cibo tailandese" o "memoria per laptop". Comprende domande di linguaggio naturale come "pioverà la prossima settimana a Los Angeles?" Una dimostrazione dell'assistente virtuale per la vendita al dettaglio è disponibile in "Personalizza gli stati e i flussi per i casi d'utilizzo retail".

Tecnologia della soluzione

La figura seguente illustra l'architettura di sistema ai conversazionale proposta. È possibile interagire con il sistema sia con il segnale vocale che con l'immissione di testo. Se viene rilevato un input vocale, Jarvis ai-as-service (AIaaS) esegue ASR per produrre testo per Dialog Manager. Dialog Manager memorizza gli stati di conversazione, indirizza il testo ai servizi corrispondenti e passa i comandi al motore di adempimento. Jarvis NLP Service prende il testo, riconosce intenti ed entità e restituisce tali intenti e slot di entità a Dialog Manager, che invia quindi Action al motore di adempimento. Fulfillment Engine è costituito da API di terze parti o database SQL che rispondono alle query degli utenti. Dopo aver ricevuto il risultato da Fulfillment Engine, Dialog Manager indirizza il testo a Jarvis TTS AIaaS per produrre una risposta audio per l'utente finale. Possiamo archiviare la cronologia delle conversazioni, annotare frasi con intenti e slot per il training NEMO in modo che il servizio NLP migliori man mano che un maggior numero di utenti interagisce con il sistema.



Requisiti hardware

Questa soluzione è stata validata utilizzando una stazione DGX e un sistema storage AFF A220. Jarvis richiede una GPU T4 o V100 per eseguire calcoli di rete neurali profondi.

La seguente tabella elenca i componenti hardware necessari per implementare la soluzione come testata.

Hardware	Quantità
GPU T4 O V100	1
Stazione NVIDIA DGX	1

Requisiti software

La seguente tabella elenca i componenti software necessari per implementare la soluzione come testata.

Software	Versione o altre informazioni
Software per la gestione dei dati NetApp ONTAP	9.6
Firmware dello switch Cisco NX-OS	7.0(3)I6(1)
SISTEMA OPERATIVO NVIDIA DGX	4.0.4 - Ubuntu 18.04 LTS
NVIDIA Jarvis Framework	EA v0.2
NVIDIA NEMO	nvcr.io/nvidia/nemo:v0.10
Piattaforma container Docker	18.06.1-ce [e68fc7a]

Panoramica

In questa sezione vengono fornite informazioni dettagliate sull'implementazione di Virtual Retail Assistant.

Implementazione di Jarvis

Puoi iscriverti a. "Programma Jarvis Early Access" Per accedere ai container Jarvis su NVIDIA GPU Cloud (NGC). Dopo aver ricevuto le credenziali da NVIDIA, puoi implementare Jarvis seguendo questa procedura:

1. Accedi a NGC.
2. Imposta la tua organizzazione su NGC: `ea-2-jarvis`.
3. Individuare le risorse Jarvis EA v0.2: I container Jarvis sono in `Private Registry` > `Organization Containers`.
4. Selezionare Jarvis: Selezionare `Model Scripts` e fare clic su `Jarvis Quick Start`
5. Verificare che tutte le risorse funzionino correttamente.
6. Trova la documentazione per creare le tue applicazioni: I PDF sono disponibili in `Model Scripts` > `Jarvis Documentation` > `File Browser`.

Personalizza gli stati e i flussi per i casi d'utilizzo retail

È possibile personalizzare gli stati e i flussi di Dialog Manager in base ai casi di utilizzo specifici. Nel nostro esempio di vendita al dettaglio, abbiamo i seguenti quattro file yaml per indirizzare la conversazione in base a diversi intenti.

Se il seguente elenco di nomi di file e la descrizione di ciascun file:

- `main_flow.yml`: Definisce i flussi e gli stati principali della conversazione e indirizza il flusso agli altri tre file yaml, se necessario.
- `retail_flow.yml`: Contiene stati relativi a domande al dettaglio o punti di interesse. Il sistema fornisce le informazioni del negozio più vicino o il prezzo di un dato articolo.
- `weather_flow.yml`: Contiene gli stati relativi alle domande sul meteo. Se non è possibile determinare la posizione, il sistema pone una domanda di follow-up per chiarire.
- `error_flow.yml`: Gestisce i casi in cui gli intenti dell'utente non rientrano nei tre file yaml precedenti. Dopo aver visualizzato un messaggio di errore, il sistema torna ad accettare le domande dell'utente.le sezioni seguenti contengono le definizioni dettagliate per questi file yaml.

main_flow.yml

```
name: JarvisRetail
intent_transitions:
  jarvis_error: error
  price_check: retail_price_check
  inventory_check: retail_inventory_check
  store_location: retail_store_location
  weather.weather: weather
  weather.temperature: temperature
  weather.sunny: sunny
  weather.cloudy: cloudy
  weather.snow: snow
  weather.rainfall: rain
  weather.snow_yes_no: snowfall
  weather.rainfall_yes_no: rainfall
  weather.temperature_yes_no: tempyesno
  weather.humidity: humidity
  weather.humidity_yes_no: humidity
  navigation.startnavigationpoi: retail # Transitions should be context
and slot based. Redirecting for now.
  navigation.geteta: retail
  navigation.showdirection: retail
  navigation.showmappoi: idk_what_you_talkin_about
  nomatch.none: idk_what_you_talkin_about
states:
  init:
    type: message_text
    properties:
```

```
      text: "Hi, welcome to NARA retail and weather service. How can I
help you?"
  input_intent:
    type: input_context
    properties:
      nlp_type: jarvis
      entities:
        intent: dontcare
# This state is executed if the intent was not understood
  dont_get_the intent:
    type: message_text_random
    properties:
      responses:
        - "Sorry I didn't get that! Please come again."
        - "I beg your pardon! Say that again?"
        - "Are we talking about weather? What would you like to know?"
        - "Sorry I know only about the weather"
        - "You can ask me about the weather, the rainfall, the
temperature, I don't know much more"
      delay: 0
    transitions:
      next_state: input_intent
  idk_what_you_talkin_about:
    type: message_text_random
    properties:
      responses:
        - "Sorry I didn't get that! Please come again."
        - "I beg your pardon! Say that again?"
        - "Are we talking about retail or weather? What would you like to
know?"
        - "Sorry I know only about retail and the weather"
        - "You can ask me about retail information or the weather, the
rainfall, the temperature. I don't know much more."
      delay: 0
    transitions:
      next_state: input_intent
  error:
    type: change_context
    properties:
        update_keys:
           intent: 'error'
    transitions:
        flow: error_flow
  retail_inventory_check:
    type: change_context
    properties:
```

```
        update_keys:
           intent: 'retail_inventory_check'
    transitions:
        flow: retail_flow
  retail_price_check:
    type: change_context
    properties:
        update_keys:
           intent: 'check_item_price'
    transitions:
        flow: retail_flow
  retail_store_location:
    type: change_context
    properties:
        update_keys:
           intent: 'find_the_store'
    transitions:
        flow: retail_flow
  weather:
    type: change_context
    properties:
        update_keys:
           intent: 'weather'
    transitions:
        flow: weather_flow
  temperature:
    type: change_context
    properties:
        update_keys:
           intent: 'temperature'
    transitions:
        flow: weather_flow
  rainfall:
    type: change_context
    properties:
        update_keys:
           intent: 'rainfall'
    transitions:
        flow: weather_flow
  sunny:
    type: change_context
    properties:
        update_keys:
           intent: 'sunny'
    transitions:
        flow: weather_flow
```

```
  cloudy:
    type: change_context
    properties:
        update_keys:
           intent: 'cloudy'
    transitions:
        flow: weather_flow
  snow:
    type: change_context
    properties:
        update_keys:
           intent: 'snow'
    transitions:
        flow: weather_flow
  rain:
    type: change_context
    properties:
        update_keys:
           intent: 'rain'
    transitions:
        flow: weather_flow
  snowfall:
      type: change_context
      properties:
          update_keys:
             intent: 'snowfall'
      transitions:
          flow: weather_flow
  tempyesno:
      type: change_context
      properties:
          update_keys:
             intent: 'tempyesno'
      transitions:
          flow: weather_flow
  humidity:
      type: change_context
      properties:
          update_keys:
             intent: 'humidity'
      transitions:
          flow: weather_flow
  end_state:
    type: reset
    transitions:
      next_state: init
```

retail_flow.yml

```yaml
name: retail_flow
states:
  store_location:
    type: conditional_exists
    properties:
      key: '{{location}}'
    transitions:
      exists: retail_state
      notexists: ask_retail_location
  retail_state:
    type: Retail
    properties:
    transitions:
      next_state: output_retail
  output_retail:
      type: message_text
      properties:
        text: '{{retail_status}}'
      transitions:
        next_state: input_intent
  ask_retail_location:
    type: message_text
    properties:
      text: "For which location? I can find the closest store near you."
    transitions:
      next_state: input_retail_location
  input_retail_location:
    type: input_user
    properties:
      nlp_type: jarvis
      entities:
        slot: location
      require_match: true
    transitions:
      match: retail_state
      notmatch: check_retail_jarvis_error
  output_retail_acknowledge:
    type: message_text_random
    properties:
      responses:
        - 'ok in {{location}}'
        - 'the store in {{location}}'
        - 'I always wanted to shop in {{location}}'
      delay: 0
```

```
    transitions:
      next_state: retail_state
  output_retail_notlocation:
    type: message_text
    properties:
      text: "I did not understand the location. Can you please repeat?"
    transitions:
      next_state: input_intent
  check_rerail_jarvis_error:
    type: conditional_exists
    properties:
      key: '{{jarvis_error}}'
    transitions:
      exists: show_retail_jarvis_api_error
      notexists: output_retail_notlocation
  show_retail_jarvis_api_error:
    type: message_text
    properties:
      text: "I am having troubled understanding right now. Come again on
that?"
    transitions:
      next_state: input_intent
```

weather_flow.yml

```
name: weather_flow
states:
  check_weather_location:
    type: conditional_exists
    properties:
      key: '{{location}}'
    transitions:
      exists: weather_state
      notexists: ask_weather_location
  weather_state:
    type: Weather
    properties:
    transitions:
      next_state: output_weather
  output_weather:
      type: message_text
      properties:
        text: '{{weather_status}}'
      transitions:
        next_state: input_intent
```

```
  ask_weather_location:
    type: message_text
    properties:
      text: "For which location?"
    transitions:
      next_state: input_weather_location
  input_weather_location:
    type: input_user
    properties:
      nlp_type: jarvis
      entities:
        slot: location
      require_match: true
    transitions:
      match: weather_state
      notmatch: check_jarvis_error
  output_weather_acknowledge:
    type: message_text_random
    properties:
      responses:
        - 'ok in {{location}}'
        - 'the weather in {{location}}'
        - 'I always wanted to go in {{location}}'
      delay: 0
    transitions:
      next_state: weather_state
  output_weather_notlocation:
    type: message_text
    properties:
      text: "I did not understand the location, can you please repeat?"
    transitions:
      next_state: input_intent
  check_jarvis_error:
    type: conditional_exists
    properties:
      key: '{{jarvis_error}}'
    transitions:
      exists: show_jarvis_api_error
      notexists: output_weather_notlocation
  show_jarvis_api_error:
    type: message_text
    properties:
      text: "I am having troubled understanding right now. Come again on
that, else check jarvis services?"
    transitions:
      next_state: input_intent
```

error_flow.yml

```
name: error_flow
states:
  error_state:
    type: message_text_random
    properties:
      responses:
        - "Sorry I didn't get that!"
        - "Are we talking about retail or weather? What would you like to
know?"
        - "Sorry I know only about retail information or the weather"
        - "You can ask me about retail information or the weather, the
rainfall, the temperature. I don't know much more"
        - "Let's talk about retail or the weather!"
      delay: 0
    transitions:
      next_state: input_intent
```

Connettersi alle API di terze parti come motore di adempimento

Abbiamo collegato le seguenti API di terze parti come motore di adempimento per rispondere alle domande:

- "API di WeatherStack": restituisce meteo, temperatura, pioggia e neve in una determinata posizione.
- "API Fusion di Yelp": restituisce le informazioni del negozio più vicino in una determinata posizione.
- "SDK di eBay Python": restituisce il prezzo di un dato articolo.

Dimostrazione di NetApp Retail Assistant

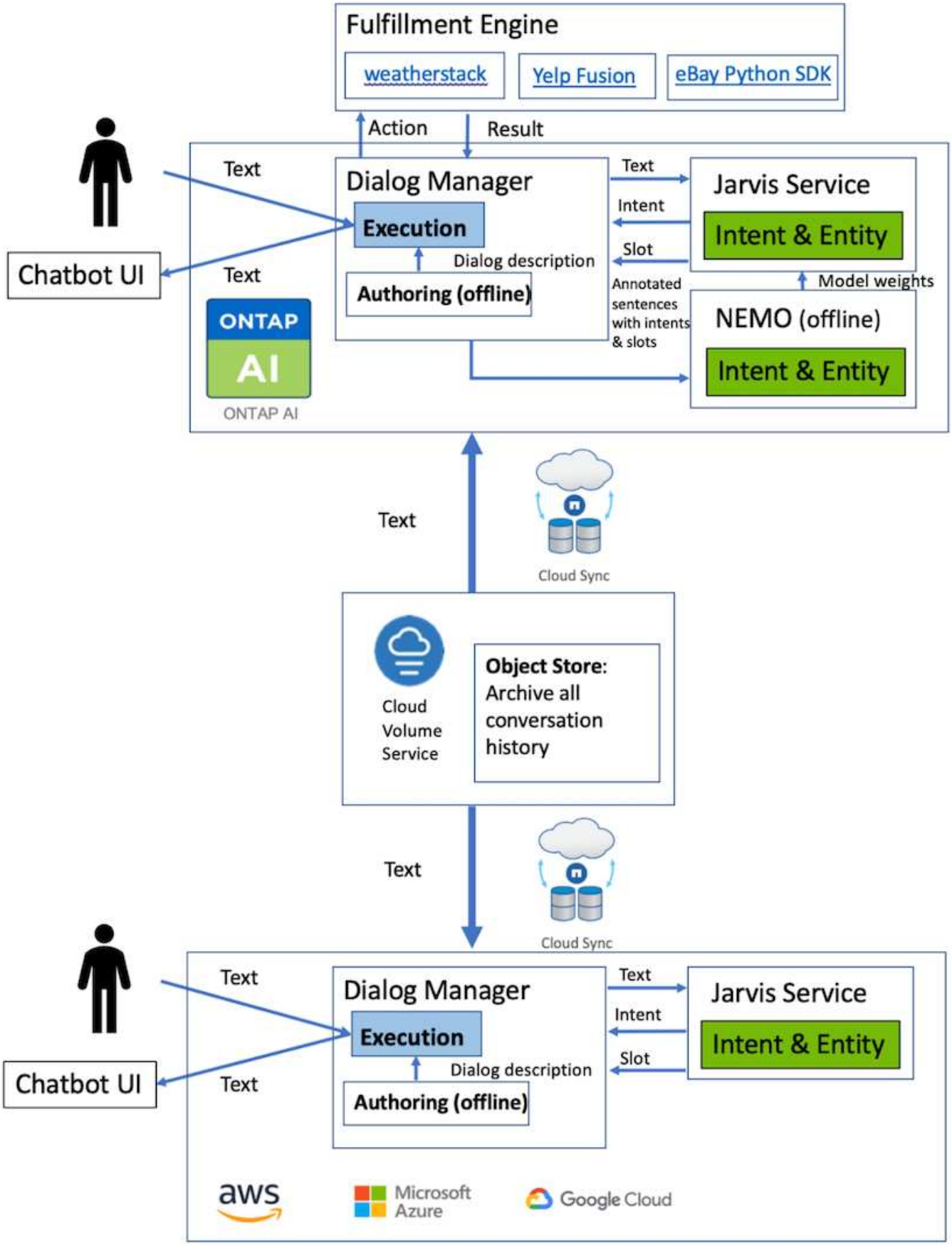
Abbiamo registrato un video dimostrativo di NetApp Retail Assistant (NARA).

Video dimostrativo DI NARA

Video dimostrativo DI NARA

Utilizza la copia e sincronizzazione di NetApp BlueXP per archiviare la cronologia delle conversazioni

Scaricando la cronologia delle conversazioni in un file CSV una volta al giorno, possiamo quindi sfruttare BlueXP Copy e Sync per scaricare i file di log nello storage locale. La figura seguente mostra l'architettura di avere implementato Jarvis on-premise e nei cloud pubblici, mentre utilizza BlueXP Copy e Sync per inviare la cronologia delle conversazioni per il training NEMO. I dettagli del training NEMO sono disponibili nella sezione "Espandi i modelli di intento utilizzando NEMO Training".


Fulfillment Engine
weatherstack
Yelp Fusion
eBay Python SDK
Action
Result
Text
Chatbot UI
Text
ONTAP
AI
ONTAP AI
Dialog Manager
Execution
Dialog description
Authoring (offline)
Text
Intent
Slot
Annotated sentences with intents & slots
Jarvis Service
Intent & Entity
Model weights
NEMO (offline)
Intent & Entity
Text
Cloud Sync
Cloud Volume Service
Object Store: Archive all conversation history
Text
Cloud Sync
Text
Chatbot UI
Text
Dialog Manager
Execution
Dialog description
Authoring (offline)
Text
Intent
Slot
Jarvis Service
Intent & Entity
aws
Microsoft Azure
Google Cloud

Espandi i modelli di intento utilizzando NEMO Training

NVIDIA NEMO è un toolkit creato da NVIDIA per la creazione di applicazioni ai conversazionali. Questo toolkit include raccolte di moduli pre-formati per ASR, NLP e TTS, che consentono a ricercatori e data scientist di comporre facilmente architetture di rete neurali complesse e di concentrarsi maggiormente sulla progettazione delle proprie applicazioni.

Come illustrato nell'esempio precedente, NARA può gestire solo un tipo limitato di domanda. Questo perché il modello di NLP pre-addestrato si allena solo su questi tipi di domande. Se vogliamo consentire A NARA di gestire una gamma più ampia di domande, dobbiamo rielaborare il sistema con i nostri set di dati. In questo caso, dimostreremo come possiamo utilizzare NEMO per estendere il modello NLP in modo da soddisfare i requisiti. Iniziamo convertendo il log raccolto da NARA nel formato NEMO, quindi ci alleniamo con il set di dati per migliorare il modello NLP.

Modello

Il nostro obiettivo è consentire A NARA di ordinare gli elementi in base alle preferenze dell'utente. Ad esempio, potremmo chiedere A NARA di suggerire il ristorante di sushi più classificato o di cercare I jeans CON il prezzo più basso. A tal fine, utilizziamo il modello di rilevamento degli intenti e di riempimento degli slot fornito in NEMO come modello di training. Questo modello consente A NARA di comprendere l'intento della ricerca delle preferenze.

Preparazione dei dati

Per formare il modello, raccogliamo il dataset per questo tipo di domanda e lo convertiamo nel formato NEMO. Qui sono elencati i file utilizzati per la formazione del modello.

dict.intents.csv

Questo file elenca tutti gli intenti che vogliamo che NEMO comprenda. In questo caso, abbiamo due intenti primari e un solo intento utilizzato per classificare le domande che non si inseriscono in nessuno degli intenti primari.

```
price_check
find_the_store
unknown
```

dict.slots.csv

Questo file elenca tutti gli slot che possiamo etichettare sulle nostre domande di training.

```
B-store.type
B-store.name
B-store.status
B-store.hour.start
B-store.hour.end
B-store.hour.day
B-item.type
```

```
B-item.name
B-item.color
B-item.size
B-item.quantity
B-location
B-cost.high
B-cost.average
B-cost.low
B-time.period_of_time
B-rating.high
B-rating.average
B-rating.low
B-interrogative.location
B-interrogative.manner
B-interrogative.time
B-interrogative.personal
B-interrogative
B-verb
B-article
I-store.type
I-store.name
I-store.status
I-store.hour.start
I-store.hour.end
I-store.hour.day
I-item.type
I-item.name
I-item.color
I-item.size
I-item.quantity
I-location
I-cost.high
I-cost.average
I-cost.low
I-time.period_of_time
I-rating.high
I-rating.average
I-rating.low
I-interrogative.location
I-interrogative.manner
I-interrogative.time
I-interrogative.personal
I-interrogative
I-verb
I-article
O
```

train.sv

Questo è il set di dati di training principale. Ogni riga inizia con la domanda che segue l'elenco delle categorie di intento nel file dict.intent.csv. L'etichetta viene enumerata a partire da zero.

train_slot.sv

```
20 46 24 25 6 32 6
52 52 24 6
23 52 14 40 52 25 6 32 6
…
```

Formare il modello

```
docker pull nvcr.io/nvidia/nemo:v0.10
```

Quindi, viene utilizzato il seguente comando per avviare il container. In questo comando, limitiamo il container a utilizzare una singola GPU (ID GPU = 1), poiché si tratta di un esercizio di formazione leggero. Inoltre, mappiamo la nostra area di lavoro locale /Workspace/nemo/ nella cartella all'interno di container /nemo.

```
NV_GPU='1' docker run --runtime=nvidia -it --shm-size=16g \
                        --network=host --ulimit memlock=-1 --ulimit
stack=67108864 \
                        -v /workspace/nemo:/nemo\
                        --rm nvcr.io/nvidia/nemo:v0.10
```

All'interno del container, se si desidera partire dal modello BERT originale pre-addestrato, è possibile utilizzare il seguente comando per avviare la procedura di training. data_dir è l'argomento per impostare il percorso dei dati di training. work_dir consente di configurare la posizione in cui si desidera memorizzare i file del punto di verifica.

```
cd examples/nlp/intent_detection_slot_tagging/
python joint_intent_slot_with_bert.py \
    --data_dir /nemo/training_data\
    --work_dir /nemo/log
```

Se abbiamo nuovi set di dati di training e vogliamo migliorare il modello precedente, possiamo utilizzare il seguente comando per continuare dal punto in cui ci siamo fermati. checkpoint_dir porta il percorso alla cartella checkpoint precedente.

```
cd examples/nlp/intent_detection_slot_tagging/
python joint_intent_slot_infer.py \
    --data_dir /nemo/training_data \
    --checkpoint_dir /nemo/log/2020-05-04_18-34-20/checkpoints/ \
    --eval_file_prefix test
```

Deduzione del modello

Dobbiamo convalidare le performance del modello formatosi dopo un certo numero di epoche. Il seguente comando consente di eseguire il test della query uno per uno. Ad esempio, in questo comando, si desidera verificare se il modello è in grado di identificare correttamente l'intenzione della query `where can I get the best pasta`.

```
cd examples/nlp/intent_detection_slot_tagging/
python joint_intent_slot_infer_b1.py \
--checkpoint_dir /nemo/log/2020-05-29_23-50-58/checkpoints/ \
--query "where can i get the best pasta" \
--data_dir /nemo/training_data/ \
--num_epochs=50
```

Di seguito viene riportato l'output dell'inferenza. Nell'output, possiamo vedere che il nostro modello addestrato può prevedere correttamente l'intenzione find_the_store e restituire le parole chiave a cui siamo interessati. Con queste parole chiave, consentiamo A NARA di cercare ciò che gli utenti desiderano e di effettuare una ricerca più precisa.

```
[NeMo I 2020-05-30 00:06:54 actions:728] Evaluating batch 0 out of 1
[NeMo I 2020-05-30 00:06:55 inference_utils:34] Query: where can i get the
best pasta
[NeMo I 2020-05-30 00:06:55 inference_utils:36] Predicted intent:       1
find_the_store
[NeMo I 2020-05-30 00:06:55 inference_utils:50] where   B-
interrogative.location
[NeMo I 2020-05-30 00:06:55 inference_utils:50] can     O
[NeMo I 2020-05-30 00:06:55 inference_utils:50] i       O
[NeMo I 2020-05-30 00:06:55 inference_utils:50] get     B-verb
[NeMo I 2020-05-30 00:06:55 inference_utils:50] the     B-article
[NeMo I 2020-05-30 00:06:55 inference_utils:50] best    B-rating.high
[NeMo I 2020-05-30 00:06:55 inference_utils:50] pasta   B-item.type
```

Conclusione

Un vero e proprio sistema di ai conversa si impegna in un dialogo umano, comprende il contesto e fornisce risposte intelligenti. Tali modelli di ai sono spesso enormi e altamente complessi. Con le GPU NVIDIA e lo storage NetApp, è possibile formare e ottimizzare

modelli di linguaggio all'avanguardia per eseguire rapidamente l'inferenza. Si tratta di un importante passo avanti verso la fine del compromesso tra un modello di ai veloce e uno grande e complesso. I modelli di comprensione del linguaggio ottimizzati per la GPU possono essere integrati nelle applicazioni di ai per settori come l'assistenza sanitaria, la vendita al dettaglio e i servizi finanziari, alimentando assistenti vocali digitali avanzati in altoparlanti intelligenti e linee di assistenza clienti. Questi sistemi di ai convergenti di alta qualità consentono alle aziende di tutti i mercati verticali di fornire servizi personalizzati precedentemente irraggiungibili quando si impegnano con i clienti.

Jarvis consente l'implementazione di casi di utilizzo come assistenti virtuali, avatar digitali, Fusion del sensore multimodale (CV fuso con ASR/NLP/TTS) o qualsiasi caso di utilizzo autonomo ASR/NLP/TTS/CV, ad esempio la trascrizione. Abbiamo creato un assistente virtuale al dettaglio in grado di rispondere a domande relative a meteo, punti di interesse e prezzi dell'inventario. Abbiamo anche dimostrato come migliorare le funzionalità di comprensione del linguaggio naturale del sistema ai conversazionale archiviando la cronologia delle conversazioni utilizzando BlueXP Copy and Sync e formando i modelli NEMO sui nuovi dati.

Ringraziamenti

Gli autori riconoscono con gratitudine i contributi che sono stati apportati a questo white paper dai nostri stimati colleghi di NVIDIA: Davide Onofrio, Alex Qi, Sicong Ji, Marty Jain e Robert Sohigian. Gli autori desiderano inoltre ringraziare i principali membri del team NetApp: Santosh Rao, David Arnette, Michael Oglesby, Brent Davis, Andy Sayare, Erik Mulder e Mike McNamara.

Il nostro sincero apprezzamento e ringraziamento va a tutti questi individui, che hanno fornito informazioni e competenze che hanno contribuito enormemente alla creazione di questo documento.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare le seguenti risorse:

- NVIDIA DGX Station, V100 GPU, GPU Cloud
  - Stazione NVIDIA DGXhttps://www.nvidia.com/en-us/data-center/dgx-station/["https://www.nvidia.com/en-us/data-center/dgx-station/"^]
  - NVIDIA V100 Tensor Core GPUhttps://www.nvidia.com/en-us/data-center/tesla-v100/["https://www.nvidia.com/en-us/data-center/tesla-v100/"^]
  - NVIDIA NGChttps://www.nvidia.com/en-us/gpu-cloud/["https://www.nvidia.com/en-us/gpu-cloud/"^]
- Framework multimodale NVIDIA Jarvis
  - NVIDIA Jarvishttps://developer.nvidia.com/nvidia-jarvis["https://developer.nvidia.com/nvidia-jarvis"^]
  - Accesso anticipato a NVIDIA Jarvishttps://developer.nvidia.com/nvidia-jarvis-early-access["https://developer.nvidia.com/nvidia-jarvis-early-access"^]
- NVIDIA NEMO
  - NVIDIA NEMOhttps://developer.nvidia.com/nvidia-nemo["https://developer.nvidia.com/nvidia-nemo"^]
  - Guida per sviluppatorihttps://nvidia.github.io/NeMo/["https://nvidia.github.io/NeMo/"^]
- Sistemi NetApp AFF

    - Scheda informativa su NetApp AFF Serie A.https://www.netapp.com/us/media/ds-3582.pdf["https://www.netapp.com/us/media/ds-3582.pdf"^]
    - NetApp Flash Advantage per All Flash FAShttps://www.netapp.com/us/media/ds-3733.pdf["https://www.netapp.com/us/media/ds-3733.pdf"^]
    - Raccolta di informazioni su ONTAP 9http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286["http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286"^]
    - Report tecnico NetApp ONTAP FlexGroup Volumeshttps://www.netapp.com/us/media/tr-4557.pdf["https://www.netapp.com/us/media/tr-4557.pdf"^]
- NetApp ONTAP ai
    - Guida alla progettazione di reti ONTAP ai con DGX-1 e Ciscohttps://www.netapp.com/us/media/nva-1121-design.pdf["https://www.netapp.com/us/media/nva-1121-design.pdf"^]
    - Guida all'implementazione di ONTAP ai con DGX-1 e Cisco Networkinghttps://www.netapp.com/us/media/nva-1121-deploy.pdf["https://www.netapp.com/us/media/nva-1121-deploy.pdf"^]
    - Guida alla progettazione di reti ONTAP ai con DGX-1 e Mellanoxhttp://www.netapp.com/us/media/nva-1138-design.pdf["http://www.netapp.com/us/media/nva-1138-design.pdf"^]
    - Guida alla progettazione di ONTAP ai con DGX-2https://www.netapp.com/us/media/nva-1135-design.pdf["https://www.netapp.com/us/media/nva-1135-design.pdf"^]

TR-4858: Soluzione di orchestrazione NetApp con Run:ai

Rick Huang, David Arnette, Sung-Han Lin, NetApp Yaron Goldberg, Run:ai

I sistemi storage NetApp AFF offrono performance estreme e funzionalità di gestione dei dati del cloud ibrido leader di settore. NetApp e Run:ai hanno collaborato per dimostrare le funzionalità esclusive della soluzione NetApp ONTAP ai per i carichi di lavoro di intelligenza artificiale (ai) e machine learning (ML) che offrono performance, affidabilità e supporto di livello Enterprise. Run:l'orchestrazione ai dei carichi di lavoro ai aggiunge una piattaforma di scheduling e utilizzo delle risorse basata su Kubernetes per aiutare i ricercatori a gestire e ottimizzare l'utilizzo della GPU. Insieme ai sistemi NVIDIA DGX, la soluzione combinata di NetApp, NVIDIA e Run:ai offre uno stack di infrastruttura costruito ad hoc per i carichi di lavoro ai aziendali. Questo report tecnico fornisce una guida direzionale ai clienti che sviluppano sistemi di ai conversazionali a supporto di vari casi di utilizzo e mercati verticali del settore. Include informazioni sull'implementazione di Run:ai e di un sistema storage NetApp AFF A800 e funge da architettura di riferimento per il modo più semplice per ottenere un'implementazione rapida e di successo delle iniziative ai.

Il pubblico di riferimento per la soluzione comprende i seguenti gruppi:

- Architetti aziendali che progettano soluzioni per lo sviluppo di modelli ai e software per casi di utilizzo basati su Kubernetes, come i microservizi containerizzati
- Data scientist alla ricerca di modi efficienti per raggiungere obiettivi di sviluppo dei modelli efficienti in un ambiente cluster con più team e progetti
- Data engineer responsabili della manutenzione e dell'esecuzione dei modelli di produzione

- Decision maker e dirigenti IT e business leader che desiderano creare l'esperienza ottimale di utilizzo delle risorse cluster di Kubernetes e ottenere il più rapido time-to-market dalle iniziative di ai

Panoramica della soluzione

Questa sezione offre una panoramica della soluzione Run:ai per ONTAP ai.

Piano di controllo ai e ai di NetApp ONTAP

L'architettura NetApp ONTAP ai, sviluppata e verificata da NetApp e NVIDIA, è basata su sistemi NVIDIA DGX e sistemi storage connessi al cloud. Questa architettura di riferimento offre alle organizzazioni IT i seguenti vantaggi:

- Elimina le complessità di progettazione
- Consente una scalabilità indipendente di calcolo e storage
- Consente ai clienti di partire da piccoli e scalare perfettamente
- Offre una gamma di opzioni di storage per diverse performance e costi

NetApp ONTAP ai integra perfettamente i sistemi DGX e i sistemi storage NetApp AFF A800 con reti all'avanguardia. I sistemi NetApp ONTAP ai e DGX semplificano le implementazioni ai eliminando la complessità e le congetture di progettazione. I clienti possono iniziare a crescere in maniera ininterrotta e allo stesso tempo gestire in modo intelligente i dati dall'edge al core, fino al cloud e viceversa.

NetApp ai Control Plane è una soluzione per la gestione di dati e esperimenti di ai, ML e deep learning (DL) per data scientist e data engineer. Man mano che le organizzazioni aumentano l'utilizzo dell'ai, devono affrontare molte sfide, tra cui la scalabilità dei workload e la disponibilità dei dati. NetApp ai Control Plane affronta queste sfide attraverso funzionalità, come la clonazione rapida di uno spazio dei nomi dei dati come faresti con un Git repo, e la definizione e l'implementazione di workflow di training ai che incorporano la creazione quasi istantanea di dati e linee di base dei modelli per la tracciabilità e il controllo delle versioni. Con NetApp ai Control Plane, puoi replicare perfettamente i dati tra siti e regioni e fornire rapidamente spazi di lavoro Jupyter notebook con accesso a set di dati di grandi dimensioni.

Run:ai Platform for ai workload Orchestration

Run:ai ha costruito la prima piattaforma di orchestrazione e virtualizzazione al mondo per l'infrastruttura ai. Astrando i carichi di lavoro dall'hardware sottostante, Run:ai crea un pool condiviso di risorse GPU che può essere sottoposto a provisioning dinamico, consentendo un'orchestrazione efficiente dei carichi di lavoro ai e un utilizzo ottimizzato delle GPU. I data scientist possono consumare senza problemi enormi quantità di energia GPU per migliorare e accelerare la ricerca, mentre i team IT mantengono un controllo centralizzato e cross-site e una visibilità in tempo reale su provisioning, accodamento e utilizzo delle risorse. La piattaforma Run:ai si basa su Kubernetes, consentendo una semplice integrazione con i flussi di lavoro IT e di data science esistenti.

La piattaforma Run:ai offre i seguenti vantaggi:

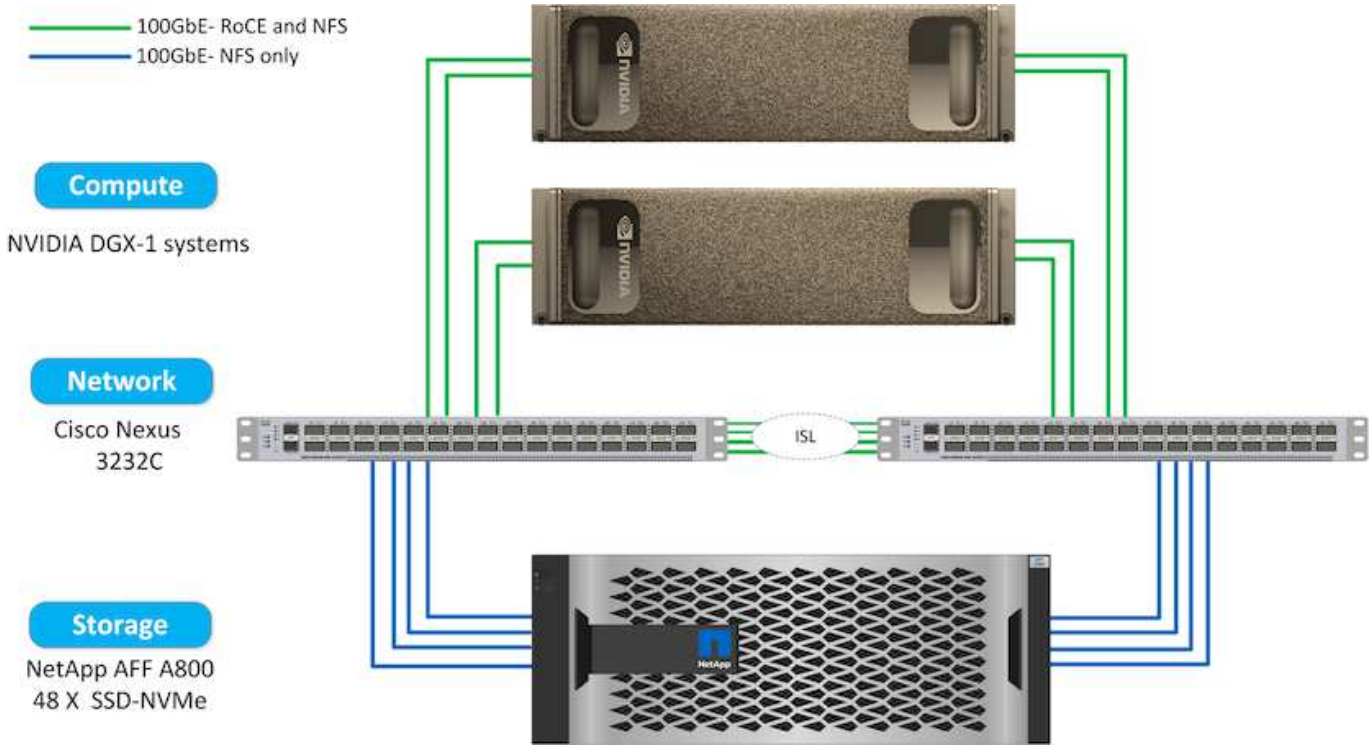
- **Time-to-innovation più veloce.** utilizzando i meccanismi di pool di risorse Run:ai, accodamento e prioritizzazione insieme a un sistema storage NetApp, i ricercatori vengono rimossi dai problemi di gestione dell'infrastruttura e possono concentrarsi esclusivamente sulla scienza dei dati. Esegui: I clienti ai e NetApp aumentano la produttività eseguendo tutti i carichi di lavoro necessari senza colli di bottiglia della pipeline di dati o di calcolo.
- **Maggiore produttività del team.** gli algoritmi Run:ai Fairness garantiscono che tutti gli utenti e i team ottenano la loro giusta quota di risorse. È possibile preimpostare le policy relative ai progetti prioritari e la

piattaforma consente l'allocazione dinamica delle risorse da un utente o team all'altro, aiutando gli utenti a ottenere un accesso tempestivo alle risorse GPU più ambite.

- **Utilizzo migliorato della GPU.** il programma Run:ai Scheduler consente agli utenti di utilizzare facilmente GPU frazionali, GPU interi e nodi multipli di GPU per la formazione distribuita su Kubernetes. In questo modo, i carichi di lavoro ai vengono eseguiti in base alle tue esigenze, non alla capacità. I team di data science sono in grado di eseguire più esperimenti di ai sulla stessa infrastruttura.

Tecnologia della soluzione

Questa soluzione è stata implementata con un sistema NetApp AFF A800, due server DGX-1 e due switch Cisco Nexus 3232C 100GbE. Ciascun server DGX-1 è connesso agli switch Nexus con quattro connessioni da 100 GbE utilizzate per le comunicazioni tra GPU utilizzando RDMA (Remote Direct Memory Access) su RoCE (Converged Ethernet). Anche le comunicazioni IP tradizionali per l'accesso allo storage NFS avvengono su questi collegamenti. Ogni controller di storage è collegato agli switch di rete utilizzando quattro collegamenti da 100 GbE. La figura seguente mostra l'architettura della soluzione ONTAP ai utilizzata in questo report tecnico per tutti gli scenari di test.



Hardware utilizzato in questa soluzione

Questa soluzione è stata validata utilizzando l'architettura di riferimento ONTAP ai, due nodi DGX-1 e un sistema storage AFF A800. Vedere "NVA-1121" per ulteriori informazioni sull'infrastruttura utilizzata in questa convalida.

La seguente tabella elenca i componenti hardware necessari per implementare la soluzione come testata.

Hardware	Quantità
Sistemi DGX-1	2

Hardware	Quantità
AFF A800	1
Switch Nexus 3232C	2

Requisiti software

Questa soluzione è stata convalidata utilizzando un'implementazione Kubernetes di base con l'operatore Run:ai installato. Kubernetes è stato implementato utilizzando "NVIDIA DeepOps" motore di implementazione, che implementa tutti i componenti necessari per un ambiente pronto per la produzione. DeepOps implementato automaticamente "Trident di NetApp" Per l'integrazione persistente dello storage con l'ambiente k8s, sono state create classi di storage predefinite in modo che i container sfruttino lo storage del sistema storage AFF A800. Per ulteriori informazioni su Trident con Kubernetes su ONTAP ai, vedere "TR-4798".

La seguente tabella elenca i componenti software necessari per implementare la soluzione come testata.

Software	Versione o altre informazioni
Software per la gestione dei dati NetApp ONTAP	9.6p4
Firmware dello switch Cisco NX-OS	7.0(3)I6(1)
SISTEMA OPERATIVO NVIDIA DGX	4.0.4 - Ubuntu 18.04 LTS
Versione di Kubernetes	1.17
Versione di Trident	20.04.0
Esegui:ai CLI	v2.1.13
Run:ai Orchestration Kubernetes Operator version	1.0.39
Piattaforma container Docker	18.06.1-ce [e68fc7a]

Ulteriori requisiti software per Run:ai sono disponibili all'indirizzo "Esegui: Prerequisiti del cluster GPU ai".

Utilizzo ottimale di cluster e GPU con Run:ai

Le sezioni seguenti forniscono dettagli sull'installazione Run:ai, sugli scenari di test e sui risultati ottenuti in questa convalida.

Abbiamo validato il funzionamento e le performance di questo sistema utilizzando tool di benchmark standard di settore, inclusi i benchmark TensorFlow. Il set di dati ImageNet è stato utilizzato per formare ResNet-50, un famoso modello DL della rete neurale convoluzionale (CNN) per la classificazione delle immagini. ResNet-50 offre un risultato di training accurato con un tempo di elaborazione più rapido, che ci ha consentito di gestire una domanda sufficiente sullo storage.

Esegui:Installazione ai

Per installare Run:ai, attenersi alla seguente procedura:

1. Installare il cluster Kubernetes utilizzando DeepOps e configurare la classe di storage predefinita di NetApp.
2. Preparare i nodi GPU:

a. Verificare che i driver NVIDIA siano installati sui nodi GPU.

b. Verificare che `nvidia-docker` è installato e configurato come runtime predefinito di docker.

3. Esecuzione dell'installazione:ai:

a. Accedere a. "Esegui: UI Admin ai" per creare il cluster.

b. Scarica il creato `runai-operator-<clustername>.yaml` file.

c. Applicare la configurazione dell'operatore al cluster Kubernetes.

```
kubectl apply -f runai-operator-<clustername>.yaml
```

4. Verificare l'installazione:

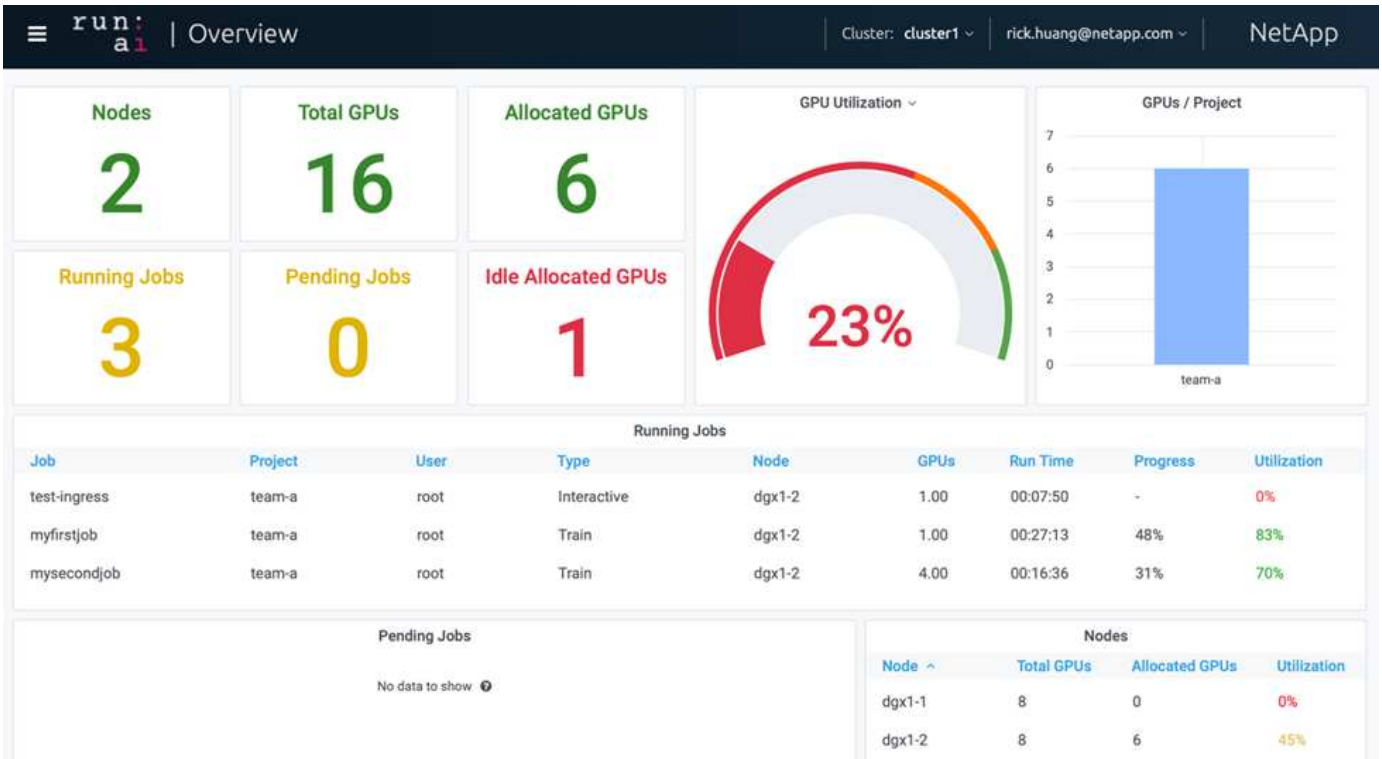
a. Passare a. "https://app.run.ai/".

b. Accedere alla dashboard Panoramica.

c. Verificare che il numero di GPU in alto a destra rifletta il numero previsto di GPU e che i nodi GPU siano tutti nell'elenco dei server.per ulteriori informazioni sull'implementazione di Run:ai, vedere "Installazione di Run:ai su un cluster Kubernetes on-premise" e. "Installazione della CLI Run:ai".

Run:ai Dashboard e viste

Dopo aver installato Run:ai sul cluster Kubernetes e aver configurato correttamente i container, vengono visualizzate le seguenti dashboard e viste "https://app.run.ai" nel browser, come mostrato nella figura seguente.



Nel cluster sono presenti 16 GPU totali fornite da due nodi DGX-1. È possibile visualizzare il numero di nodi, il totale delle GPU disponibili, le GPU allocate assegnate con i carichi di lavoro, il numero totale di processi in esecuzione, i processi in sospeso e le GPU allocate inattive. Sul lato destro, il diagramma a barre mostra le

GPU per progetto, che riepiloga il modo in cui i diversi team utilizzano la risorsa del cluster. Al centro è riportato l'elenco dei job attualmente in esecuzione con i relativi dettagli, inclusi nome del job, progetto, utente, tipo di job, Il nodo su cui ciascun processo è in esecuzione, il numero di GPU allocate per tale processo, il tempo di esecuzione corrente del processo, l'avanzamento del processo in percentuale e l'utilizzo della GPU per tale processo. Si noti che il cluster è sottoutilizzato (utilizzo della GPU al 23%) perché ci sono solo tre job in esecuzione inviati da un singolo team (`team-a`).

Nella sezione seguente, mostreremo come creare più team nella scheda progetti e allocare GPU per ciascun team per massimizzare l'utilizzo del cluster e gestire le risorse quando sono presenti molti utenti per cluster. Gli scenari di test imitano gli ambienti aziendali in cui le risorse di memoria e GPU sono condivise tra carichi di lavoro di training, deduzione e interattivi.

Creazione di progetti per i team Data Science e allocazione delle GPU

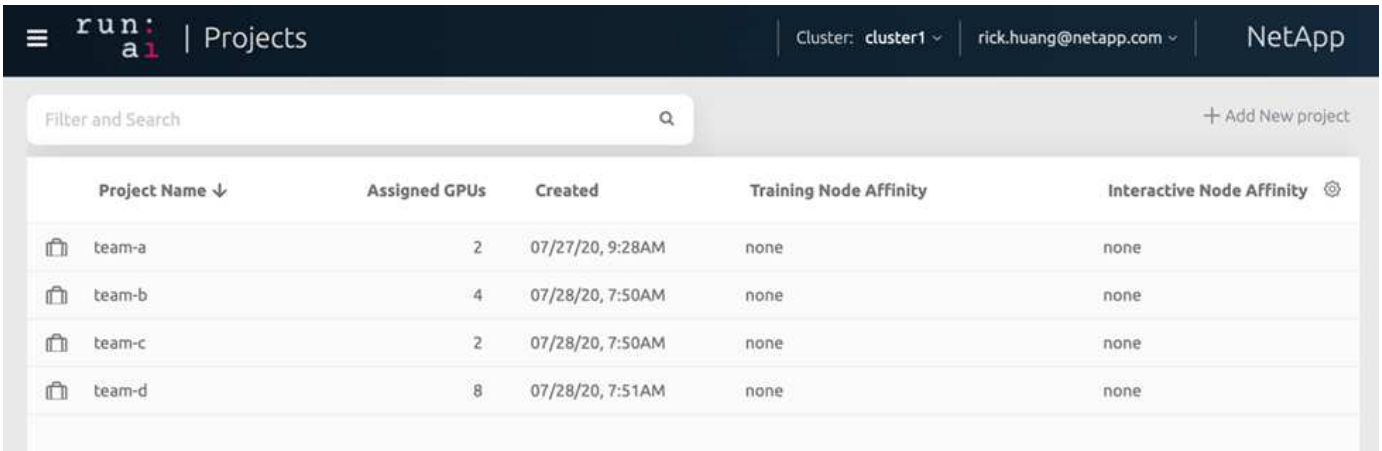
I ricercatori possono inviare i carichi di lavoro attraverso la CLI Run:ai, Kubeflow o processi simili. Per ottimizzare l'allocazione delle risorse e creare priorità, Run:ai introduce il concetto di progetti. I progetti sono entità di quota che associano un nome di progetto all'allocazione e alle preferenze della GPU. Si tratta di un metodo semplice e conveniente per gestire più team di data science.

Un ricercatore che invia un workload deve associare un progetto a una richiesta di workload. Lo scheduler Run:ai confronta la richiesta con le allocazioni correnti e il progetto e determina se il carico di lavoro può essere allocato o se deve rimanere in uno stato in sospeso.

In qualità di amministratore di sistema, è possibile impostare i seguenti parametri nella scheda Run:ai Projects (Esegui: Progetti ai):

- **Model projects.** Imposta un progetto per utente, imposta un progetto per team di utenti e imposta un progetto per un progetto organizzativo reale.
- **Quote di progetto.** ogni progetto è associato a una quota di GPU che può essere allocata per questo progetto contemporaneamente. Si tratta di una quota garantita nel senso che i ricercatori che utilizzano questo progetto possono ottenere questo numero di GPU indipendentemente dallo stato del cluster. Di norma, la somma dell'allocazione del progetto deve essere uguale al numero di GPU nel cluster. Oltre a questo, un utente di questo progetto può ricevere una quota eccessiva. Finché le GPU non vengono utilizzate, un ricercatore che utilizza questo progetto può ottenere più GPU. In vengono illustrati scenari di test con quote superiori e considerazioni di equità "Elevato utilizzo del cluster con allocazione della GPU con quota eccessiva", "Equità nell'allocazione delle risorse di base", e. "Equità nell'overquota".
- Creare un nuovo progetto, aggiornare un progetto esistente ed eliminare un progetto esistente.
- **Limita l'esecuzione dei job su gruppi di nodi specifici**. È possibile assegnare progetti specifici da eseguire solo su nodi specifici. Ciò è utile quando il team di progetto ha bisogno di hardware specializzato, ad esempio con memoria sufficiente. In alternativa, un team di progetto potrebbe essere il proprietario di hardware specifico acquistato con un budget specializzato, oppure quando potrebbe essere necessario indirizzare i carichi di lavoro di build o interattivi per lavorare su hardware più debole e indirizzare i carichi di lavoro di formazione più lunghi o non presidiati su nodi più veloci. Per i comandi per raggruppare i nodi e impostare l'affinità per un progetto specifico, vedere "Run:documentazione ai".
- **Limitare la durata dei lavori interattivi**. I ricercatori spesso si dimenticano di chiudere lavori interattivi. Ciò potrebbe comportare uno spreco di risorse. Alcune organizzazioni preferiscono limitare la durata dei lavori interattivi e chiuderli automaticamente.

La figura seguente mostra la vista progetti con quattro team creati. A ciascun team viene assegnato un numero diverso di GPU per i diversi carichi di lavoro, con il numero totale di GPU pari a quello delle GPU totali disponibili in un cluster costituito da due DGX-1.



Invio di job in Run:ai CLI

Questa sezione fornisce i dettagli sui comandi Run:ai di base che è possibile utilizzare per eseguire qualsiasi lavoro Kubernetes. È suddiviso in tre parti in base al tipo di carico di lavoro. I carichi di lavoro ai/ML/DL possono essere suddivisi in due tipi generici:

- **Sessioni di training non presidiate**. Con questi tipi di carichi di lavoro, il data scientist prepara un carico di lavoro a esecuzione automatica e lo invia per l'esecuzione. Durante l'esecuzione, il cliente può esaminare i risultati. Questo tipo di carico di lavoro viene spesso utilizzato in produzione o quando lo sviluppo del modello si trova in una fase in cui non è richiesto alcun intervento umano.
- **Sessioni di build interattive**. Con questi tipi di carichi di lavoro, il data scientist apre una sessione interattiva con Bash, Jupyter notebook, PyCharm remoto o IDE simili e accede direttamente alle risorse GPU. Abbiamo incluso un terzo scenario per l'esecuzione di workload interattivi con porte connesse per rivelare una porta interna all'utente del container.

Carichi di lavoro di training non presidiati

Dopo aver impostato i progetti e allocato le GPU, è possibile eseguire qualsiasi carico di lavoro Kubernetes utilizzando il seguente comando nella riga di comando:

```
$ runai project set team-a runai submit hyper1 -i gcr.io/run-ai-
demo/quickstart -g 1
```

Questo comando avvia un processo di training non assistito per il team-a con un'allocazione di una singola GPU. Il lavoro si basa su un'immagine del docker di esempio, `gcr.io/run-ai-demo/quickstart`. Abbiamo nominato il lavoro `hyper1`. È quindi possibile monitorare l'avanzamento del lavoro eseguendo il seguente comando:

```
$ runai list
```

La figura seguente mostra il risultato di `runai list` comando. Gli stati tipici che potrebbero essere visualizzati includono:

- `ContainerCreating`. Il container del docker viene scaricato dal repository cloud.
- `Pending`. Il lavoro è in attesa di essere pianificato.

- `Running`. Il processo è in esecuzione.

```
~> runai list
Showing jobs for project team-a
NAME    STATUS   AGE  NODE                                    IMAGE                        TYPE        PROJECT  USER   GPUs
hyper1  Running  11s  gke-dev-yaron1-gpu-4-pool-154f511d-5nk5  gcr.io/run-ai-demo/quickstart  Train       team-a   yaron  1
```

Per ottenere uno stato aggiuntivo sul lavoro, eseguire il seguente comando:

```
$ runai get hyper1
```

Per visualizzare i log del lavoro, eseguire `runai logs <job-name>` comando:

```
$ runai logs hyper1
```

In questo esempio, dovresti visualizzare il registro di una sessione DL in esecuzione, inclusi l'epoca di training corrente, l'ETA, il valore della funzione di perdita, l'accuratezza e il tempo trascorso per ogni fase.

È possibile visualizzare lo stato del cluster nell'interfaccia utente Run:ai all'indirizzo "https://app.run.ai/". In Dashboard > Panoramica, è possibile monitorare l'utilizzo della GPU.

Per arrestare questo carico di lavoro, eseguire il seguente comando:

```
$ runai delte hyper1
```

Questo comando interrompe il carico di lavoro del training. È possibile verificare questa azione eseguendo `runai list` di nuovo. Per ulteriori informazioni, vedere "lancio di workload di training non presidiati".

Workload di build interattivi

Dopo aver impostato i progetti e allocato le GPU, è possibile eseguire un carico di lavoro di build interattivo utilizzando il seguente comando dalla riga di comando:

```
$ runai submit build1 -i python -g 1 --interactive --command sleep --args
infinity
```

Il lavoro si basa su un python immagine del docker di esempio. Abbiamo chiamato la creazione di job 1.

> Il `-- interactive` flag indica che il lavoro non ha inizio o fine È responsabilità del ricercatore chiudere il lavoro. L'amministratore può definire un limite di tempo per i lavori interattivi dopo il quale vengono terminati dal sistema.

Il `--g 1` Flag assegna una singola GPU a questo lavoro. Il comando e l'argomento forniti sono `--command sleep—args infinity`. È necessario fornire un comando, altrimenti il container viene avviato e quindi chiuso immediatamente.

I seguenti comandi funzionano in modo simile ai comandi descritti in Carichi di lavoro di training non presidiati:

- `runai list`: Mostra il nome, lo stato, l'età, il nodo, l'immagine, Progetto, utente e GPU per i lavori.
- `runai get build1`: Visualizza lo stato aggiuntivo nella creazione del job 1.
- `runai delete build1`: Interrompe la creazione interattiva del workload 1.per ottenere una shell bash nel container, utilizzare il seguente comando:

```
$ runai bash build1
```

Questo fornisce una shell diretta nel computer. I data scientist possono quindi sviluppare o perfezionare i propri modelli all'interno del container.

È possibile visualizzare lo stato del cluster nell'interfaccia utente Run:ai all'indirizzo "https://app.run.ai". Per ulteriori informazioni, vedere "avvio e utilizzo di workload di build interattivi".

Carichi di lavoro interattivi con porte connesse

Come estensione dei carichi di lavoro di build interattivi, è possibile rivelare le porte interne all'utente del container quando si avvia un container con la CLI Run:ai. Questo è utile per ambienti cloud, per lavorare con i notebook Jupyter o per connettersi ad altri microservizi. "Ingresso" Consente l'accesso ai servizi Kubernetes dall'esterno del cluster Kubernetes. È possibile configurare l'accesso creando un insieme di regole che definiscono quali connessioni in entrata raggiungono i servizi.

Per una migliore gestione dell'accesso esterno ai servizi in un cluster, si consiglia agli amministratori del cluster di eseguire l'installazione "Ingresso" E configurare LoadBalancer.

Per utilizzare Ingress come tipo di servizio, eseguire il seguente comando per impostare il tipo di metodo e le porte durante l'invio del carico di lavoro:

```
$ runai submit test-ingress -i jupyter/base-notebook -g 1 \
  --interactive --service-type=ingress --port 8888 \
  --args="--NotebookApp.base_url=test-ingress" --command=start-notebook.sh
```

Una volta avviato il container, eseguire `runai list` per visualizzare `SERVICE URL(S)` Con cui accedere al Jupyter notebook. L'URL è composto dall'endpoint di ingresso, dal nome del processo e dalla porta.

Per ulteriori informazioni, vedere "lancio di un workload di build interattivo con porte connesse".

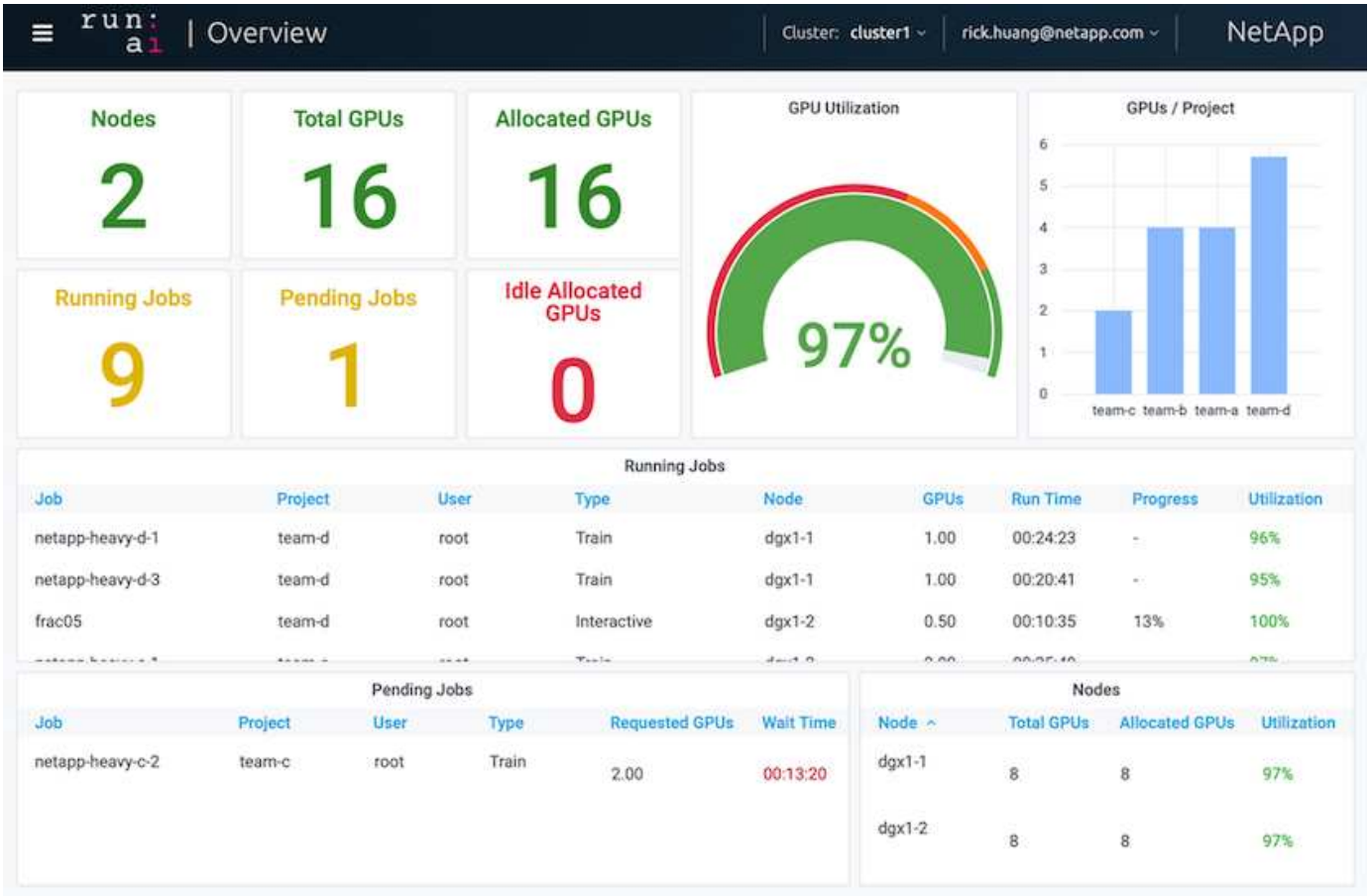
Elevato utilizzo del cluster

In questa sezione, emuliamo uno scenario realistico in cui quattro team di data science inviano ciascuno i propri carichi di lavoro per dimostrare la soluzione di orchestrazione Run:ai che raggiunge un elevato utilizzo del cluster mantenendo al contempo la prioritizzazione e il bilanciamento delle risorse GPU. Iniziamo utilizzando il benchmark ResNet-50 descritto nella sezione "ResNet-50 con ImageNet dataset Benchmark Summary":

```
$ runai submit netapp1 -i netapp/tensorflow-tf1-py3:20.01.0 --local-image
--large-shm  -v /mnt:/mnt -v /tmp:/tmp --command python --args
"/netapp/scripts/run.py" --args "--
dataset_dir=/mnt/mount_0/dataset/imagenet/imagenet_original/" --args "--
num_mounts=2"  --args "--dgx_version=dgx1" --args "--num_devices=1" -g 1
```

Abbiamo eseguito lo stesso benchmark ResNet-50 di in "NVA-1121". Abbiamo utilizzato la bandiera `--local -image` per i container che non risiedono nel repository del dock pubblico. Abbiamo montato le directory `/mnt` e. `/tmp` Sul nodo host DGX-1 a. `/mnt` e. `/tmp` al container, rispettivamente. Il set di dati è disponibile presso NetApp AFFA800 con `dataset_dir` argomento che punta alla directory. Entrambi `--num_devices=1` e. `-g 1` Significa che allociamo una GPU per questo lavoro. Il primo è un argomento per `run.py` script, mentre quest'ultimo è un flag per `runai submit` comando.

La figura seguente mostra una dashboard panoramica del sistema con il 97% di utilizzo della GPU e tutte le sedici GPU disponibili allocate. È possibile visualizzare facilmente il numero di GPU allocate per ciascun team nel grafico a barre GPU/progetto. Il riquadro dei job in esecuzione mostra i nomi dei job in esecuzione, il progetto, l'utente, il tipo, il nodo, GPU consumate, tempo di esecuzione, avanzamento e dettagli di utilizzo. Un elenco dei workload in coda con il relativo tempo di attesa viene visualizzato in lavori in sospeso. Infine, la casella Nodes offre i numeri GPU e l'utilizzo per i singoli nodi DGX-1 nel cluster.



Allocazione frazionale della GPU per carichi di lavoro meno impegnativi o interattivi

Quando ricercatori e sviluppatori stanno lavorando sui propri modelli, sia nelle fasi di sviluppo, tuning di iperparametri o debug, tali carichi di lavoro richiedono di solito meno

risorse di calcolo. È quindi più efficiente eseguire il provisioning di GPU e memoria frazionarie in modo che la stessa GPU possa essere allocata contemporaneamente ad altri carichi di lavoro. La soluzione di orchestrazione di Run:ai offre un sistema di condivisione della GPU frazionale per carichi di lavoro containerizzati su Kubernetes. Il sistema supporta i carichi di lavoro che eseguono programmi CUDA ed è particolarmente adatto per attività ai leggere come inferenza e costruzione di modelli. Il sistema di GPU frazionale offre ai team di progettazione ai e data science la possibilità di eseguire più carichi di lavoro contemporaneamente su una singola GPU. Ciò consente alle aziende di eseguire più carichi di lavoro, ad esempio visione artificiale, riconoscimento vocale ed elaborazione del linguaggio naturale sullo stesso hardware, riducendo così i costi.

Esegui: Il sistema di GPU frazionale di ai crea efficacemente GPU logiche virtualizzate con la propria memoria e spazio di calcolo che i container possono utilizzare e accedere come se fossero processori autonomi. In questo modo, è possibile eseguire diversi carichi di lavoro in container, uno accanto all'altro, sulla stessa GPU senza interferire l'uno con l'altro. La soluzione è trasparente, semplice e portatile e non richiede modifiche ai container stessi.

Un'usecase tipica potrebbe visualizzare da due a otto lavori in esecuzione sulla stessa GPU, il che significa che è possibile eseguire otto volte il lavoro con lo stesso hardware.

Per il lavoro `frac05` appartenente al progetto `team-d` Nella figura seguente, è possibile vedere che il numero di GPU allocate era 0.50. Questo è ulteriormente verificato da `nvidia-smi` Che indica che la memoria GPU disponibile per il container era di 16,255 MB: Metà dei 32 GB per GPU V100 nel nodo DGX-1.

```
root@run-deploy:~# runai bash frac05 -p team-d
root@frac05-0:/workload# nvidia-smi
Tue Jul 28 15:17:03 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.05    Driver Version: 450.51.05    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:07:00.0 Off |                    0 |
| N/A   57C    P0   240W / 300W |  15525MiB / 16255MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A       156      C   python3                         15525MiB |
+-----------------------------------------------------------------------------+
```

Elevato utilizzo del cluster con allocazione della GPU con quota eccessiva

In questa sezione e nelle sezioni "Equità nell'allocazione delle risorse di base", e. "Equità nell'overquota", Abbiamo ideato scenari di test avanzati per dimostrare le funzionalità di orchestrazione Run:ai per la gestione di workload complessi, la pianificazione preventiva automatica e il provisioning di GPU con overquota. Lo abbiamo fatto per ottenere un elevato utilizzo delle risorse del cluster e ottimizzare la produttività del team di data science di livello Enterprise in un ambiente ai ONTAP.

Per queste tre sezioni, impostare i seguenti progetti e quote:

Progetto	Quota
squadra a.	4
team-b	2
team-c	2
team-d	8

Inoltre, per queste tre sezioni vengono utilizzati i seguenti container:

- Notebook Jupyter: `jupyter/base-notebook`
- Run:Avvio rapido ai: `gcr.io/run-ai-demo/quickstart`

Per questo scenario di test sono stati stabiliti i seguenti obiettivi:

- Mostra la semplicità del provisioning delle risorse e il modo in cui le risorse vengono estratte dagli utenti
- Mostrare come gli utenti possono eseguire facilmente il provisioning di frazioni di GPU e numero intero di GPU
- Mostra come il sistema elimina i colli di bottiglia di calcolo consentendo a team o utenti di superare la quota di risorse se nel cluster sono presenti GPU gratuite
- Mostra come vengono eliminati i colli di bottiglia della pipeline di dati utilizzando la soluzione NetApp durante l'esecuzione di processi a elaborazione intensiva, come il container NetApp
- Mostrare come vengono eseguiti diversi tipi di container utilizzando il sistema
  - Notebook Jupyter
  - Container Run:ai
- Mostra un utilizzo elevato quando il cluster è pieno

Per informazioni dettagliate sulla sequenza di comandi effettiva eseguita durante il test, vedere "Dettagli sui test per la Sezione 4.8".

Una volta inviati tutti i 13 carichi di lavoro, è possibile visualizzare un elenco di nomi di container e GPU allocati, come mostrato nella figura seguente. Disponiamo di sette corsi di formazione e sei lavori interattivi, che simulano quattro team di data science, ciascuno con i propri modelli in esecuzione o in fase di sviluppo. Per i lavori interattivi, i singoli sviluppatori utilizzano Jupyter Notebooks per scrivere o eseguire il debug del codice. Pertanto, è adatto per eseguire il provisioning delle frazioni GPU senza utilizzare troppe risorse del cluster.

```
root@run-deploy:~# runai list -A
NAME           STATUS   AGE  NODE    IMAGE                          TYPE        PROJECT  USER  GPUs  CREATED BY CLI  SERVICE URL(S)
b-4-gg         Running  2m   dgx1-2  gcr.io/run-ai-demo/quickstart  Train       team-b   root  2     true
c-5-g          Running  2m   dgx1-2  gcr.io/run-ai-demo/quickstart  Train       team-c   root  1     true
c-4-gg         Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-c   root  2     true
b-3-g          Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-b   root  1     true
c-3-g02        Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Interactive team-c   root  0.2   true
d-1-gggg       Running  2m   dgx1-2  gcr.io/run-ai-demo/quickstart  Train       team-d   root  4     true
c-2-g03        Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Interactive team-c   root  0.3   true
c-1-g05        Running  2m   dgx1-1  gcr.io/run-ai-demo/quickstart  Interactive team-c   root  0.5   true
a-2-gg         Running  3m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-a   root  2     true
b-2-g04        Running  3m   dgx1-2  gcr.io/run-ai-demo/quickstart  Interactive team-b   root  0.4   true
a-1-g          Running  3m   dgx1-1  gcr.io/run-ai-demo/quickstart  Train       team-a   root  1     true
b-1-g06        Running  3m   dgx1-2  gcr.io/run-ai-demo/quickstart  Interactive team-b   root  0.6   true
a-1-1-jupyter  Running  3m   dgx1-1  jupyter/base-notebook          Interactive team-a   root  1     true            http://10.61.218.134/a-1-1-jupyter,
https://10.61.218.134/a-1-1-jupyter
```

I risultati di questo scenario di test mostrano quanto segue:

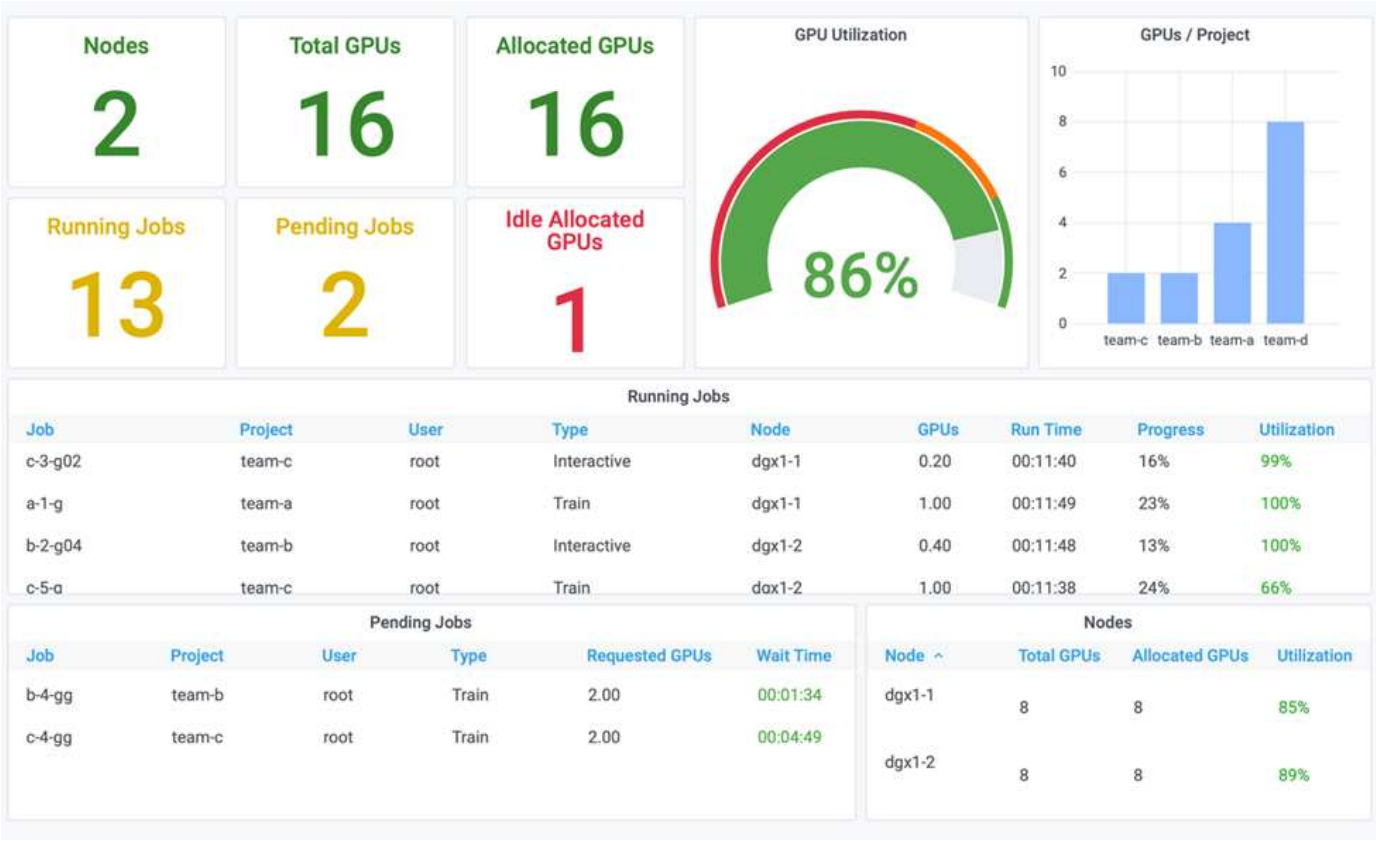
- Il cluster deve essere pieno: Vengono utilizzate 16/16 GPU.
- Elevato utilizzo del cluster.
- Più esperimenti rispetto alle GPU a causa dell'allocazione frazionale.
- `team-d` non utilizza tutta la quota; pertanto, `team-b` e. `team-c` Possibilità di utilizzare GPU aggiuntive per i propri esperimenti, con conseguente riduzione dei tempi di innovazione.

Equità nell'allocazione delle risorse di base

In questa sezione, mostreremo quando `team-d` Richiede più GPU (sono sotto la loro quota), il sistema mette in pausa i carichi di lavoro di `team-b` e. `team-c` e li sposta in uno stato in sospeso in modo equo e condiviso.

Per ulteriori informazioni, tra cui invio dei job, immagini container utilizzate e sequenze di comandi eseguite, vedere la sezione "Dettagli sui test per la Sezione 4.9".

La figura seguente mostra l'utilizzo del cluster risultante, le GPU allocate per team e i processi in sospeso a causa del bilanciamento automatico del carico e della pianificazione preventiva. Possiamo osservare che quando il numero totale di GPU richieste da tutti i carichi di lavoro del team supera il numero totale di GPU disponibili nel cluster, l'algoritmo di equità interna di Run:ai mette in pausa un job per ciascuno `team-b` e. `team-c` perché hanno soddisfatto la quota di progetto. In questo modo si ottiene un elevato utilizzo generale del cluster, mentre i team di data science continuano a lavorare sotto i limiti delle risorse stabiliti da un amministratore.



I risultati di questo scenario di test dimostrano quanto segue:

- **Bilanciamento automatico del carico.** il sistema bilancia automaticamente la quota delle GPU, in modo che ogni team utilizzi ora la propria quota. I carichi di lavoro che sono stati sospesi appartengono ai team che hanno superato la quota.
- **Fair share pause.** il sistema sceglie di arrestare il carico di lavoro di un team che ha superato la quota e quindi di arrestare il carico di lavoro dell'altro team. Run:ai dispone di algoritmi interni per la correttezza.

Equità nell'overquota

In questa sezione, espandiamo lo scenario in cui più team inviano carichi di lavoro e superano la loro quota. In questo modo, dimostreremo come l'algoritmo di equità di Run:ai alloca le risorse del cluster in base al rapporto delle quote preimpostate.

Obiettivi per questo scenario di test:

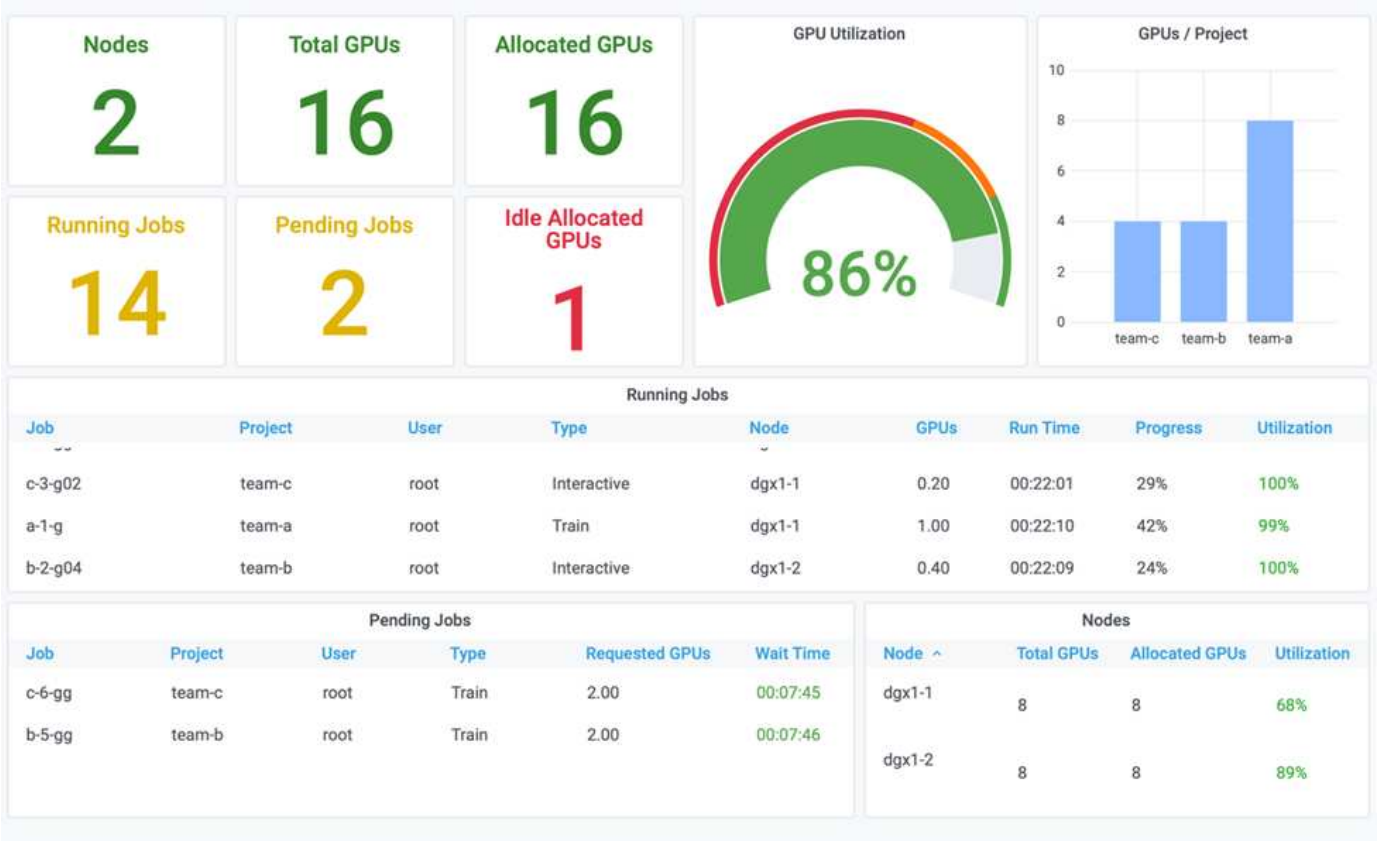
- Mostra il meccanismo di accodamento quando più team richiedono GPU sulla propria quota.
- Mostrare come il sistema distribuisce una quota equa del cluster tra più team che superano la quota in base al rapporto tra le quote, in modo che il team con la quota maggiore ottenga una quota maggiore della capacità di riserva.

Alla fine di "Equità nell'allocazione delle risorse di base", sono presenti due carichi di lavoro in coda: uno per `team-b` e uno per `team-c`. In questa sezione, vengono accodati carichi di lavoro aggiuntivi.

Per ulteriori informazioni, tra cui invio di lavori, immagini container utilizzate e sequenze di comandi eseguite, vedere "Dettagli sui test per la sezione 4.10".

Quando tutti i lavori vengono inoltrati in base alla sezione "Dettagli sui test per la sezione 4.10", il dashboard di

sistema lo mostra `team-a`, `team-b`, e. `team-c` Tutti hanno più GPU rispetto alla quota preimpostata. `team-a` Occupa quattro GPU in più rispetto alla quota soft preimpostata (quattro), mentre `team-b` e. `team-c` Ciascuna di esse occupa due GPU in più rispetto alla propria quota soft (due). Il rapporto delle GPU con overquota allocate è uguale a quello della quota preimpostata. Questo perché il sistema ha utilizzato la quota preimpostata come riferimento di priorità e fornito di conseguenza quando più team richiedono più GPU, superando la quota. Tale bilanciamento automatico del carico offre equità e prioritizzazione quando i team di data science aziendali sono attivamente impegnati nello sviluppo e nella produzione di modelli di ai.



I risultati di questo scenario di test mostrano quanto segue:

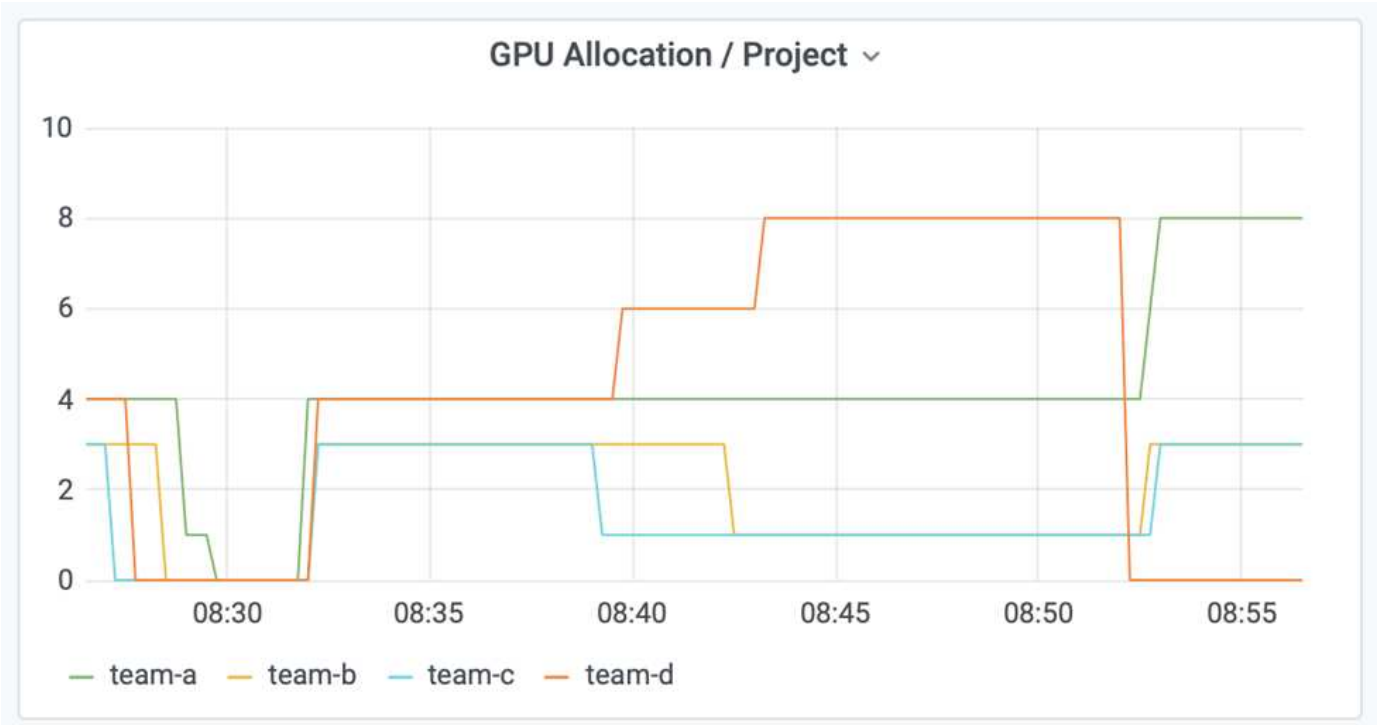
- Il sistema inizia a demettere in coda i carichi di lavoro di altri team.
- L'ordine di dequeuing viene stabilito in base agli algoritmi di equità, in modo che `team-b` e. `team-c` Ottenere la stessa quantità di GPU con quote eccessive (poiché hanno una quota simile), e. `team-a` Ottiene una quantità doppia di GPU poiché la loro quota è due volte superiore alla quota di `team-b` e. `team-c`.
- Tutta l'allocazione viene eseguita automaticamente.

Pertanto, il sistema dovrebbe stabilizzarsi nei seguenti stati:

Progetto	GPU allocate	Commento
squadra a.	8/4	Quattro GPU oltre la quota. Coda vuota.
team-b	4/2	Due GPU oltre la quota. Un carico di lavoro in coda.
team-c	4/2	Due GPU oltre la quota. Un carico di lavoro in coda.

Progetto	GPU allocate	Commento
team-d	0/8	Non utilizzare le GPU, non i carichi di lavoro in coda.

La figura seguente mostra l'allocazione della GPU per progetto nel tempo nella dashboard Run:ai Analytics per le sezioni "Elevato utilizzo del cluster con allocazione della GPU con quota eccessiva", "Equità nell'allocazione delle risorse di base", e. "Equità nell'overquota". Ciascuna riga della figura indica il numero di GPU fornite per un determinato team di data science in qualsiasi momento. Possiamo vedere che il sistema alloca dinamicamente le GPU in base ai carichi di lavoro inviati. Ciò consente ai team di superare la quota quando nel cluster sono disponibili GPU, quindi di prevenire i lavori in base all'equità, prima di raggiungere infine uno stato stabile per tutti e quattro i team.



Salvataggio dei dati in un PersistentVolume con provisioning Trident

NetApp Trident è un progetto open source completamente supportato, progettato per aiutarti a soddisfare le sofisticate esigenze di persistenza delle tue applicazioni containerizzate. È possibile leggere e scrivere i dati su un volume di prestazioni (PV) Kubernetes con provisioning Trident, con il vantaggio aggiunto di tiering dei dati, crittografia, tecnologia Snapshot NetApp, conformità e performance elevate offerte dal software di gestione dei dati NetApp ONTAP.

Riutilizzo dei PVC in uno spazio dei nomi esistente

Per i progetti ai più grandi, potrebbe essere più efficiente per diversi container leggere e scrivere i dati sullo stesso PV Kubernetes. Per riutilizzare un PVC (Persistent Volume Claim) Kubernetes, l'utente deve aver già creato un PVC. Vedere "Documentazione di NetApp Trident" Per informazioni dettagliate sulla creazione di un PVC. Ecco un esempio di riutilizzo di un PVC esistente:

```
$ runai submit pvc-test -p team-a --pvc test:/tmp/pvc1mount -i gcr.io/run-ai-demo/quickstart -g 1
```

Eseguire il seguente comando per visualizzare lo stato del lavoro `pvc-test` per il progetto `team-a`:

```
$ runai get pvc-test -p team-a
```

Dovrebbe essere montato PV /tmp/pvc1mount su `team-a` lavoro `pvc-test`. In questo modo, più container possono leggere dallo stesso volume, il che è utile quando ci sono più modelli concorrenti in fase di sviluppo o in produzione. Gli scienziati dei dati possono creare un insieme di modelli e quindi combinare i risultati delle previsioni con il voto a maggioranza o altre tecniche.

Per accedere alla shell container, utilizzare quanto segue:

```
$ runai bash pvc-test -p team-a
```

È quindi possibile controllare il volume montato e accedere ai dati all'interno del container.

Questa funzionalità di riutilizzo dei PVC funziona con i volumi NetApp FlexVol e NetApp ONTAP FlexGroup, consentendo ai data engineer di utilizzare opzioni di gestione dei dati più flessibili e solide per sfruttare il data fabric basato su NetApp.

Conclusione

NetApp e Run:ai hanno collaborato a questo report tecnico per dimostrare le funzionalità esclusive della soluzione ai di NetApp ONTAP insieme alla piattaforma Run:ai per semplificare l'orchestrazione dei carichi di lavoro ai. I passaggi precedenti forniscono un'architettura di riferimento per ottimizzare il processo di pipeline di dati e orchestrazione dei carichi di lavoro per un apprendimento approfondito. I clienti che desiderano implementare queste soluzioni sono incoraggiati a contattare NetApp e Run:ai per ulteriori informazioni.

Dettagli sui test per la Sezione 4.8

Questa sezione contiene i dettagli del test per la sezione "Elevato utilizzo del cluster con allocazione della GPU con quota eccessiva".

Inoltrare i lavori nel seguente ordine:

Progetto	Immagine	N. di GPU	Totale	Commento
squadra a.	Jupyter	1	1/4	–
squadra a.	NetApp	1	2/4	–
squadra a.	Esegui:ai	2	4/4	Utilizzando tutta la quota

Progetto	Immagine	N. di GPU	Totale	Commento
team-b	Esegui:ai	0.6	0.6/2	GPU frazionale
team-b	Esegui:ai	0.4	1/2	GPU frazionale
team-b	NetApp	1	2/2	–
team-b	NetApp	2	4/2	Due in eccesso di quota
team-c	Esegui:ai	0.5	0.5/2	GPU frazionale
team-c	Esegui:ai	0.3	0.8/2	GPU frazionale
team-c	Esegui:ai	0.2	1/2	GPU frazionale
team-c	NetApp	2	3/2	Uno sopra la quota
team-c	NetApp	1	4/2	Due in eccesso di quota
team-d	NetApp	4	4/8	Utilizzando metà della quota

Struttura dei comandi:

```
$ runai submit <job-name> -p <project-name> -g <#GPUs> -i <image-name>
```

Sequenza di comandi effettiva utilizzata nel test:

```
$ runai submit a-1-1-jupyter -i jupyter/base-notebook -g 1 \
  --interactive --service-type=ingress --port 8888 \
  --args="--NotebookApp.base_url=team-a-test-ingress" --command=start
-notebook.sh -p team-a
$ runai submit a-1-g -i gcr.io/run-ai-demo/quickstart -g 1 -p team-a
$ runai submit a-2-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-a
$ runai submit b-1-g06 -i gcr.io/run-ai-demo/quickstart -g 0.6
--interactive -p team-b
$ runai submit b-2-g04 -i gcr.io/run-ai-demo/quickstart -g 0.4
--interactive -p team-b
$ runai submit b-3-g -i gcr.io/run-ai-demo/quickstart -g 1 -p team-b
$ runai submit b-4-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-b
$ runai submit c-1-g05 -i gcr.io/run-ai-demo/quickstart -g 0.5
--interactive -p team-c
$ runai submit c-2-g03 -i gcr.io/run-ai-demo/quickstart -g 0.3
--interactive -p team-c
$ runai submit c-3-g02 -i gcr.io/run-ai-demo/quickstart -g 0.2
--interactive -p team-c
$ runai submit c-4-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-c
$ runai submit c-5-g -i gcr.io/run-ai-demo/quickstart -g 1 -p team-c
$ runai submit d-1-gggg -i gcr.io/run-ai-demo/quickstart -g 4 -p team-d
```

A questo punto, si dovrebbero avere i seguenti stati:

Progetto	GPU allocate	Carichi di lavoro in coda
squadra a.	4/4 (quota soft/allocazione effettiva)	Nessuno
team-b	4/2	Nessuno
team-c	4/2	Nessuno
team-d	4/8	Nessuno

Vedere la sezione "Elevato utilizzo del cluster con allocazione GPU over-uota" per discussioni sullo scenario di test.

Dettagli sui test per la Sezione 4.9

Questa sezione contiene i dettagli dei test per la sezione "Equità nell'allocazione delle risorse di base".

Inoltrare i lavori nel seguente ordine:

Progetto	N. di GPU	Totale	Commento
team-d	2	6/8	Il carico di lavoro del team-b/c viene messo in pausa e spostato a. `pending`.
team-d	2	8/8	I carichi di lavoro degli altri team (b/c) vengono interrompiti e spostati `pending`.

Vedere la seguente sequenza di comandi eseguiti:

```
$ runai submit d-2-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-d$
runai submit d-3-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-d
```

A questo punto, si dovrebbero avere i seguenti stati:

Progetto	GPU allocate	Carichi di lavoro in coda
squadra a.	4/4	Nessuno
team-b	2/2	Nessuno
team-c	2/2	Nessuno
team-d	8/8	Nessuno

Vedere la sezione "Equità nell'allocazione delle risorse di base" per una discussione sullo scenario di test.

Dettagli sui test per la Sezione 4.10

Questa sezione contiene i dettagli dei test per la sezione "Equità nell'overquota".

Inoltrare i lavori nel seguente ordine per `team-a`, `team-b`, e. `team-c`:

Progetto	N. di GPU	Totale	Commento
squadra a.	2	4/4	1 carico di lavoro in coda
squadra a.	2	4/4	2 carichi di lavoro in coda
team-b	2	2/2	2 carichi di lavoro in coda
team-c	2	2/2	2 carichi di lavoro in coda

Vedere la seguente sequenza di comandi eseguiti:

```
$ runai submit a-3-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-a$
runai submit a-4-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-a$ runai
submit b-5-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-b$ runai
submit c-6-gg -i gcr.io/run-ai-demo/quickstart -g 2 -p team-c
```

A questo punto, si dovrebbero avere i seguenti stati:

Progetto	GPU allocate	Carichi di lavoro in coda
squadra a.	4/4	Due carichi di lavoro che richiedono due GPU ciascuno
team-b	2/2	Due carichi di lavoro che richiedono due GPU ciascuno
team-c	2/2	Due carichi di lavoro che richiedono due GPU ciascuno
team-d	8/8	Nessuno

Quindi, eliminare tutti i carichi di lavoro per `team-d`:

```
$ runai delete -p team-d d-1-gggg d-2-gg d-3-gg
```

Vedere la sezione "Equità nell'overquota", per discussioni sullo scenario di test.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare le seguenti risorse:

- Sistemi NVIDIA DGX
  - Sistema NVIDIA DGX-1https://www.nvidia.com/en-us/data-center/dgx-1/[]
  - NVIDIA V100 Tensor Core GPUhttps://www.nvidia.com/en-us/data-center/tesla-v100/[]
  - NVIDIA NGChttps://www.nvidia.com/en-us/gpu-cloud/[]
- Soluzione per l'orchestrazione di container Run:ai
  - Run:Introduzione al prodotto aihttps://docs.run.ai/home/components/[]
  - Esegui:documentazione di installazione aihttps://docs.run.ai/Administrator/Cluster-Setup/Installing-Run-AI-on-an-on-premise-Kubernetes-Cluster/[] https://docs.run.ai/Administrator/Researcher-Setup/Installing-the-Run-AI-Command-Line-Interface/
  - Invio di job in Run:ai CLIhttps://docs.run.ai/Researcher/Walkthroughs/Walkthrough-Launch-Unattended-Training-Workloads-/[] https://docs.run.ai/Researcher/Walkthroughs/Walkthrough-Start-and-Use-Interactive-Build-Workloads-/
  - Allocazione delle frazioni GPU in Run:ai CLIhttps://docs.run.ai/Researcher/Walkthroughs/Walkthrough-Using-GPU-Fractions/[]
- Piano di controllo ai di NetApp

    - Report tecnicohttps://www.netapp.com/us/media/tr-4798.pdf[]
    - Demo in formato brevehttps://youtu.be/gfr_sO27Rvo[]
    - Repository di GitHubhttps://github.com/NetApp/kubeflow_jupyter_pipeline[]
- Sistemi NetApp AFF
    - Scheda informativa su NetApp AFF Serie A.https://www.netapp.com/us/media/ds-3582.pdf[]
    - NetApp Flash Advantage per All Flash FAShttps://www.netapp.com/us/media/ds-3733.pdf[]
    - Raccolta di informazioni su ONTAP 9http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286[]
    - Report tecnico NetApp ONTAP FlexGroup Volumeshttps://www.netapp.com/us/media/tr-4557.pdf[]
- NetApp ONTAP ai
    - Guida alla progettazione di reti ONTAP ai con DGX-1 e Ciscohttps://www.netapp.com/us/media/nva-1121-design.pdf[]
    - Guida all'implementazione di ONTAP ai con DGX-1 e Cisco Networkinghttps://www.netapp.com/us/media/nva-1121-deploy.pdf[]
    - Guida alla progettazione di reti ONTAP ai con DGX-1 e Mellanoxhttp://www.netapp.com/us/media/nva-1138-design.pdf[]
    - Guida alla progettazione di ONTAP ai con DGX-2https://www.netapp.com/us/media/nva-1135-design.pdf[]

TR-4799-DESIGN: Architettura di riferimento ai di NetApp ONTAP per carichi di lavoro a guida autonoma

David Arnette e Sung-Han Lin, NetApp

La famiglia di sistemi NVIDIA DGX è la prima piattaforma al mondo di intelligenza artificiale integrata (ai) costruita appositamente per l'ai aziendale. I sistemi storage NetApp AFF offrono performance estreme e funzionalità di gestione dei dati del cloud ibrido leader di settore. NetApp e NVIDIA hanno collaborato per creare l'architettura di riferimento ai di NetApp ONTAP per offrire ai clienti una soluzione chiavi in mano per supportare i carichi di lavoro di ai e machine learning (ML) con performance, affidabilità e supporto di livello Enterprise.

"TR-4799-DESIGN: Architettura di riferimento ai di NetApp ONTAP per carichi di lavoro a guida autonoma"

TR-4811: Architettura di riferimento ai di NetApp ONTAP per il settore sanitario: Imaging diagnostico - progettazione della soluzione

Rick Huang, Sung-Han Lin, Sathish Thyagarajan, NetApp Jacci Cenci, NVIDIA

Questa architettura di riferimento offre linee guida per i clienti che creano un'infrastruttura di intelligenza artificiale (ai) utilizzando i sistemi NVIDIA DGX-2 e lo storage NetApp AFF per i casi di utilizzo nel settore sanitario. Include informazioni sui flussi di lavoro di alto livello utilizzati per lo sviluppo di modelli di deep learning (DL) per l'imaging diagnostico

medico, i casi di test validati e i risultati. Include anche consigli di dimensionamento per le implementazioni dei clienti.

"TR-4811: Architettura di riferimento ai di NetApp ONTAP per il settore sanitario: Imaging diagnostico - progettazione della soluzione"

TR-4807: Architettura di riferimento ai di NetApp ONTAP per i carichi di lavoro dei servizi finanziari - progettazione della soluzione

Karthikeyan Nagalingam, Sung-Han Lin, NetApp Jacci Cenci, NVIDIA

Questa architettura di riferimento offre linee guida per i clienti che stanno costruendo un'infrastruttura di intelligenza artificiale utilizzando i sistemi NVIDIA DGX-1 e lo storage NetApp AFF per i casi di utilizzo del settore finanziario. Include informazioni sui flussi di lavoro di alto livello utilizzati per lo sviluppo di modelli di deep learning per i test case e i risultati dei servizi finanziari. Include anche consigli di dimensionamento per le implementazioni dei clienti.

"TR-4807: Architettura di riferimento ai di NetApp ONTAP per i carichi di lavoro dei servizi finanziari - progettazione della soluzione"

Valore ai e NetApp generativo

La richiesta di intelligenza artificiale generativa (ai) sta causando interruzioni in tutti i settori, migliorando la creatività aziendale e l'innovazione di prodotto.

Autore: Satish Thyagarajan, NetApp

Astratto

Molte organizzazioni utilizzano l'ai generativa per creare nuove funzioni dei prodotti, migliorare la produttività tecnica e creare prototipi di applicazioni basate sull'intelligenza artificiale che offrono risultati migliori e migliori esperienze dei consumatori. L'intelligenza artificiale generativa come i trasformatori pre-addestrati generativi (GPT) utilizzano reti neurali per creare nuovi contenuti, diversi come testo, audio e video. Data la scalabilità estrema e gli enormi set di dati utilizzati con modelli di linguaggio di grandi dimensioni (LMS), è fondamentale progettare una solida infrastruttura ai che sfrutti le interessanti funzionalità di storage dei dati delle opzioni di implementazione on-premise, ibride e multicloud, riducendo i rischi associati alla mobilità dei dati, data Protection e governance prima che le aziende possano progettare soluzioni ai. In questo documento vengono descritte queste considerazioni e le corrispondenti funzionalità di NetApp® ai che consentono una gestione perfetta dei dati e lo spostamento dei dati nella pipeline di dati ai per la formazione, la riqualificazione, la messa a punto e la inferenza dei modelli di intelligenza artificiale generativi.

In sintesi

Più recentemente dopo il lancio di ChatGPT, uno spin-off di GPT-3 nel novembre 2022, nuovi strumenti di intelligenza artificiale utilizzati per generare testo, codice, immagine, o anche proteine terapeutiche in risposta alle richieste dell'utente hanno guadagnato una fama significativa. Ciò indica che gli utenti possono effettuare una richiesta utilizzando il linguaggio naturale e l'intelligenza artificiale interpreterà e genererà testo, come articoli di notizie o descrizioni di prodotti che riflettono la richiesta dell'utente o producono codice, musica,

voce, effetti visivi e risorse 3D utilizzando algoritmi addestrati su dati già esistenti. Di conseguenza, frasi come diffusione stabile, allucinazioni, Ingegneria prompt e allineamento dei valori stanno rapidamente emergendo nella progettazione dei sistemi ai. Questi modelli di machine learning (ML) con supervisione o supervisione automatica stanno diventando ampiamente disponibili come modelli di base preformati (FM) tramite fornitori di servizi cloud e altri fornitori di firma ai, che vengono adottati da vari istituti di attività di vari settori per un'ampia gamma di attività NLP (elaborazione del linguaggio naturale) a valle. Come asserito da aziende di analisti di ricerca come McKinsey – "l'impatto dell'intelligenza artificiale generativa sulla produttività potrebbe aggiungere migliaia di miliardi di dollari in valore all'economia globale". Mentre le aziende stanno reinventando l'ai come partner di pensiero per gli esseri umani e gli istituti finanziari si stanno ampliando contemporaneamente a ciò che le aziende e le istituzioni possono fare con l'ai generativa, le opportunità di gestire grandi volumi di dati continueranno a crescere. Questo documento presenta informazioni introduttive sull'ai generativa e i concetti di progettazione in relazione alle funzionalità NetApp che apportano valore ai clienti NetApp, sia on-premise che in ambienti ibridi o multicloud.

Quindi, cosa offre ai clienti l'utilizzo di NetApp nei propri ambienti di intelligenza artificiale? NetApp aiuta le organizzazioni a soddisfare le complessità create da una rapida crescita di dati e cloud, dalla gestione multicloud e dall'adozione di tecnologie di prossima generazione, come l'intelligenza artificiale. NetApp ha combinato varie funzionalità nel software di Intelligent data management e nell'infrastruttura storage che sono ben bilanciate con performance elevate ottimizzate per i workload ai. Le soluzioni ai generative come LLMS devono leggere ed elaborare i propri set di dati di origine dallo storage in memoria diverse volte per promuovere l'intelligenza. NetApp è stata leader nelle tecnologie di mobilità dei dati, governance dei dati e sicurezza dei dati nell'ecosistema edge-to-core-to-cloud, servendo i clienti aziendali a creare soluzioni ai su larga scala. NetApp, con una solida rete di partner, ha aiutato Chief data officer, tecnici ai, Enterprise Architect e data scientist nella progettazione di una pipeline di dati a flusso libero per la preparazione, la data Protection, e responsabilità di gestione dei dati strategiche del training e dell'inferenza dei modelli di ai, ottimizzando le performance e la scalabilità del ciclo di vita ai/ML. Le tecnologie e le funzionalità dei dati di NetApp, come NetApp® ONTAP ai® per la pipeline di dati deep learning, NetApp® SnapMirror® per il trasporto dei dati in modo semplice ed efficiente tra gli endpoint di storage, e NetApp® FlexCache® per il rendering in tempo reale quando il flusso di dati passa da batch a real-time e l'ingegneria dei dati avviene in tempi rapidi, apportano valore alla distribuzione di modelli di intelligenza artificiale generativa in tempo reale. Man mano che le aziende di qualsiasi tipo adottano nuovi strumenti ai, si trovano ad affrontare sfide legate ai dati, dall'edge al data center fino al cloud, che richiedono soluzioni ai scalabili, responsabili e spiegabili. In qualità di autorità nel campo dei dati per il cloud ibrido e multicloud, NetApp si impegna a creare una rete di partner e soluzioni congiunte che possano aiutare con tutti gli aspetti della creazione di una pipeline di dati e data Lake per il training generativo sui modelli di intelligenza artificiale (pre-training), la messa a punto, l'inferenza basata sul contesto e il monitoraggio del decadimento dei modelli di LLMS.

Che cos'è l'ai generativa?

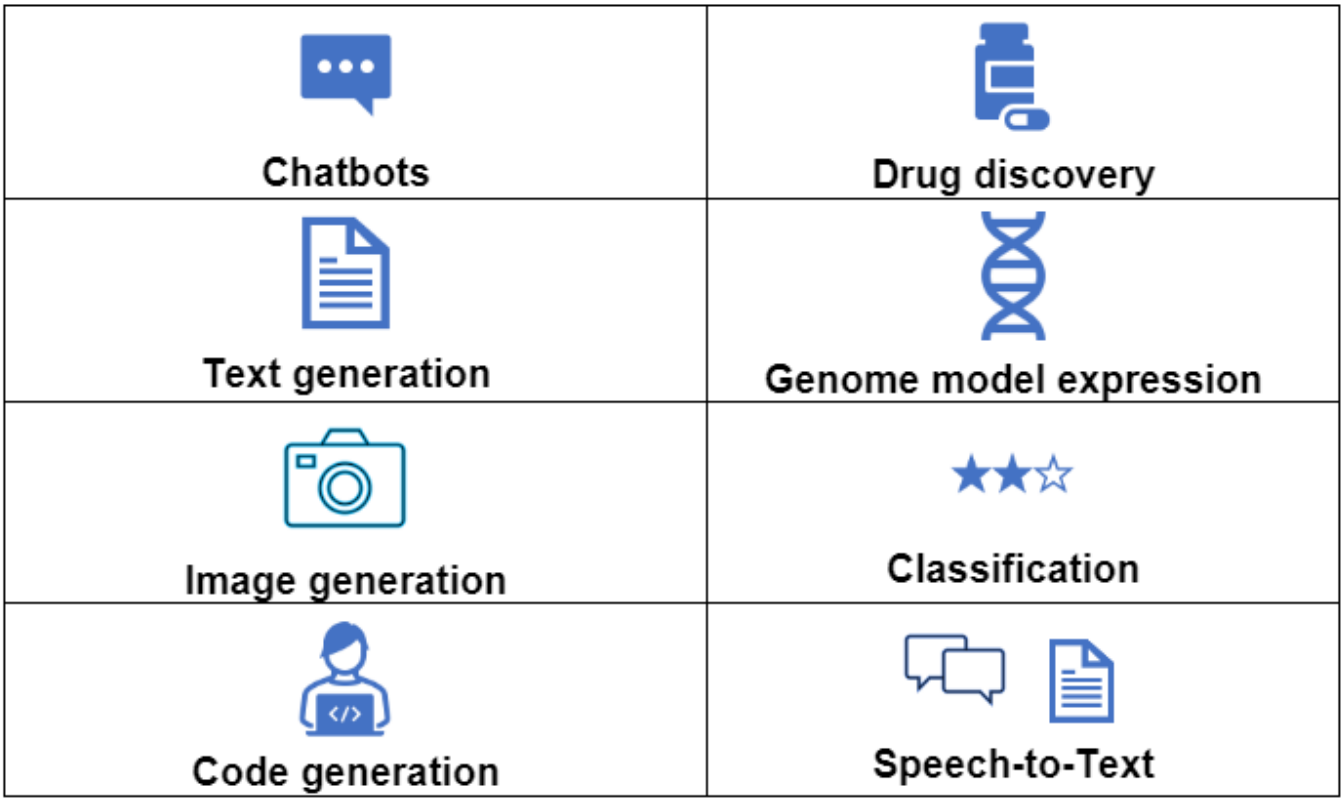
L'intelligenza artificiale generativa sta cambiando il modo in cui creiamo contenuti, generiamo nuovi concetti di progettazione ed esploriamo nuove composizioni. Illustra framework di rete neurali come GAN (Generative Adversarial Network), VAE (Variational AutoEncoder) e GPT (Generative Pre-Trained Transformers), che possono generare nuovi contenuti come testo, codice, immagini, audio, video, e dati sintetici. I modelli basati su trasformatori come la Chat-GPT di OpenAI, la Bard di Google, la FIORITURA del volto IMPETUOSO, e la lama di Meta sono emersi come la tecnologia fondamentale alla base di molti progressi nei modelli linguistici di grandi dimensioni. Allo stesso modo, Dall-e di OpenAI, Meta's CM3leon e Imagen di Google sono esempi di modelli di diffusione da testo a immagine che offrono ai clienti un grado senza precedenti di fotorealismo per creare nuove immagini complesse da zero o modificare immagini esistenti per generare immagini di alta qualità sensibili al contesto utilizzando l'aumento del set di dati e la sintesi da testo a immagine collegando la semantica testuale e visiva. Gli artisti digitali stanno iniziando ad applicare una combinazione di tecnologie di rendering come Nerf (Neural Radiance Field) con l'intelligenza artificiale generativa per convertire immagini 2D statiche in scene 3D coinvolgenti. In generale, le LLMS sono ampiamente caratterizzate da quattro parametri: (1) dimensione del modello (tipicamente in miliardi di parametri); (2) dimensione del set di dati di addestramento; (3) costo di addestramento, e (4) prestazione del modello dopo l'addestramento. Anche LLMS

rientra principalmente in tre architetture di trasformatori. (i) modelli solo encoder. Ad es. BERT (Google, 2018); (II) codificatore-decodificatore, ad es. BART (Meta, 2020) e (III) modelli solo decoder. Ad esempio Llama (Meta, 2023), Palm-e (Google, 2023). A seconda del requisito aziendale, indipendentemente dall'architettura scelta da un'azienda per il numero di parametri del modello (N) e dal numero di token (D) nel set di dati di formazione, generalmente determinano il costo di base della formazione (pre-formazione) o l'ottimizzazione di un LLM.

Casi di utilizzo aziendali e attività NLP downstream

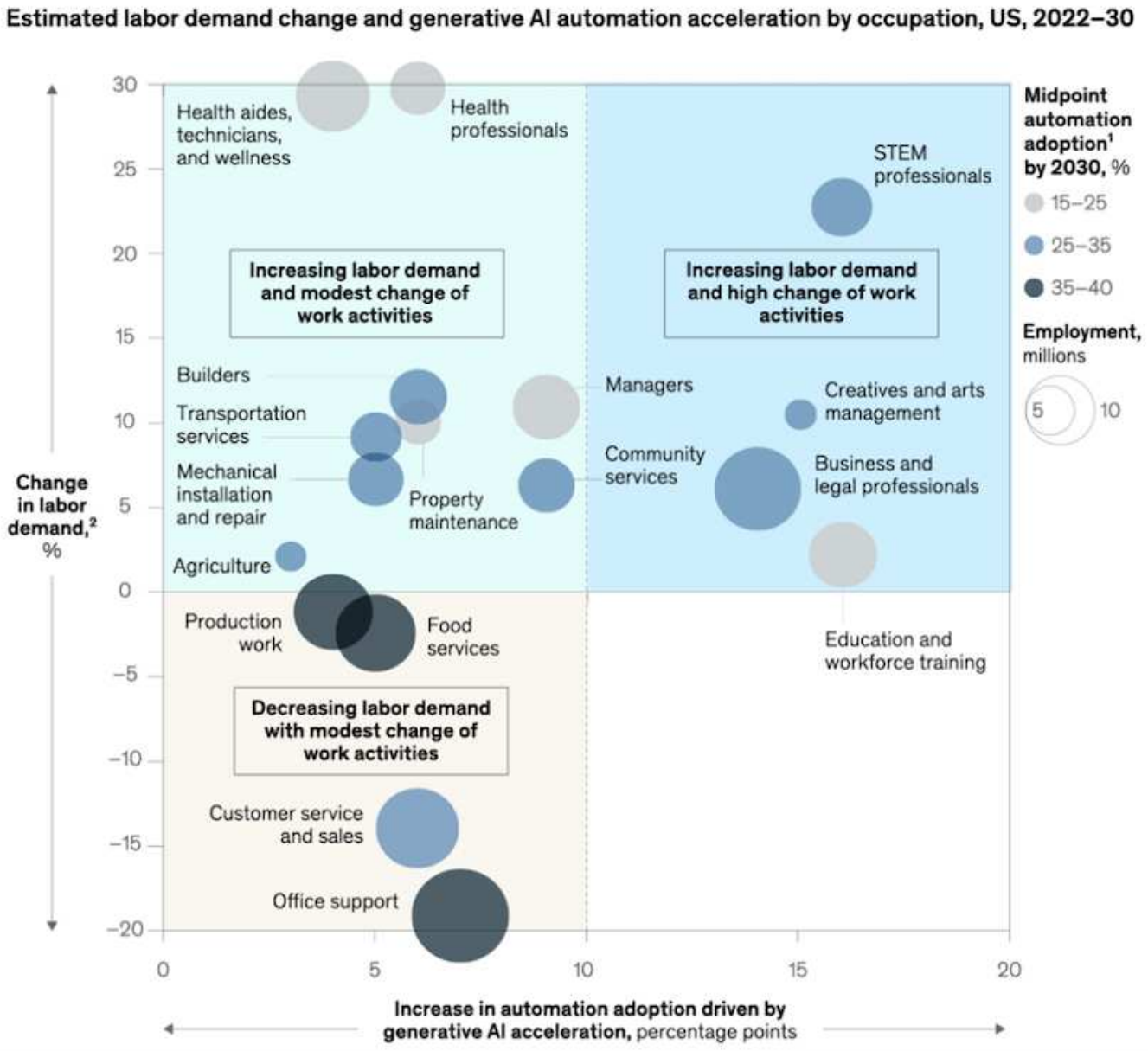
Le aziende di tutti i settori stanno scoprendo un potenziale sempre maggiore per l'ai, al fine di estrarre e produrre nuove forme di valore dai dati esistenti per operazioni di business, vendite, marketing e servizi legali. Secondo la market intelligence IDC (International Data Corporation) sui casi di utilizzo e gli investimenti dell'intelligenza artificiale generativa globale, la gestione della conoscenza nello sviluppo software e nella progettazione dei prodotti ha l'impatto maggiore, seguita dalla creazione di storyline per il marketing e la generazione di codice per gli sviluppatori. Nel settore sanitario, le organizzazioni di ricerca clinica stanno rivoluzionando la medicina. I modelli preformati come ProteinBERT incorporano annotazioni di gene ontology (GO) per progettare rapidamente strutture proteiche per i farmaci medici, rappresentando una pietra miliare significativa nella scoperta di farmaci, bioinformatica e biologia molecolare. Le aziende biotecnologiche hanno avviato studi clinici umani per la medicina generativa scoperta dall'intelligenza artificiale, che mira a trattare malattie come la fibrosi polmonare (IPF), una malattia polmonare che causa cicatrici irreversibili del tessuto polmonare.

Figura 1: Casi d'utilizzo che promuovono l'ai generativa



Gli aumenti nell'adozione dell'automazione promossi dall'intelligenza artificiale generativa stanno anche modificando l'offerta e la domanda di attività lavorative per molte occupazioni. Come per McKinsey, il mercato del lavoro statunitense (diagramma sotto) ha attraversato una rapida transizione, che potrebbe continuare solo quando si tiene conto dell'impatto dell'ai.

Fonte: McKinsey & Company



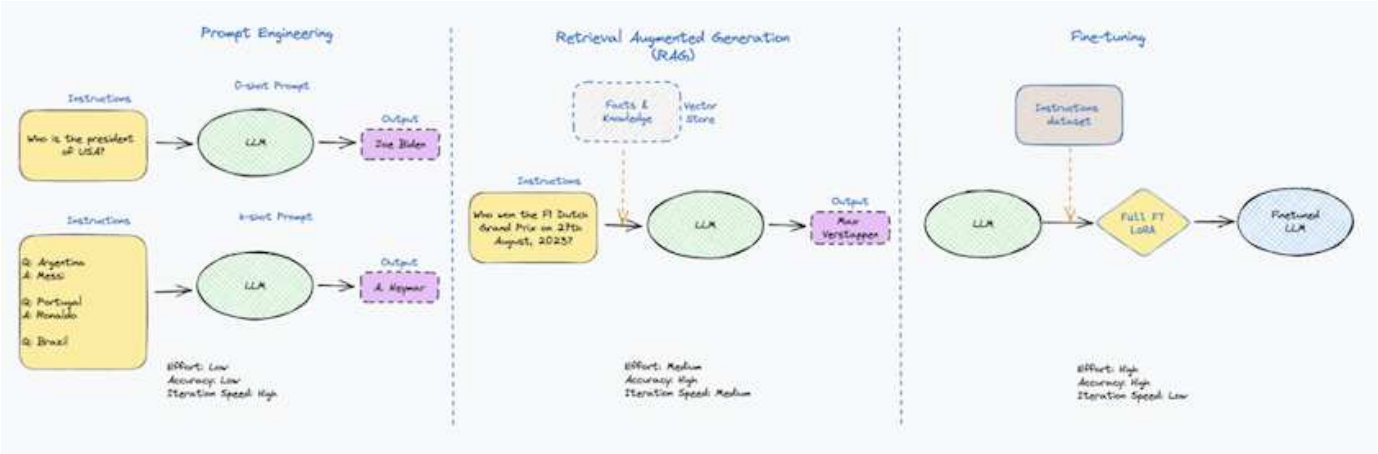
Ruolo dello storage nell'ai generativa

LLMS si basa in gran parte su deep learning, GPU e calcolo. Tuttavia, quando il buffer della GPU si riempie, i dati devono essere scritti rapidamente nella memoria. Alcuni modelli di ai sono abbastanza piccoli da essere eseguiti in memoria, mentre LLMS richiede IOPS elevati e storage ad alto throughput per fornire un rapido accesso a grandi set di dati, specialmente se coinvolge miliardi di token o milioni di immagini. Per un tipico requisito di memoria GPU di un LLM, la memoria necessaria per addestrare un modello con 1 miliardi di parametri potrebbe arrivare a 80GB a precisione completa a 32 bit. In tal caso, Llama 2 di Meta, una famiglia di LLMS in scala da 7 miliardi a 70 miliardi di parametri, può richiedere 70x80, circa 5600GB o 5,6TB di RAM GPU. Inoltre, la quantità di memoria necessaria è direttamente proporzionale al numero massimo di token che si desidera generare. Ad esempio, se si desidera generare output fino a 512 token (circa 380 parole), è necessario "512MB". Può sembrare inconsequenziale, ma se si desidera eseguire batch più grandi inizia a sommarsi. Pertanto, rendendo molto costoso per le organizzazioni che si occupano della formazione o della messa a punto di LLMS in memoria, rendendo così lo storage un caposaldo per l'ai generativa.

Tre approcci primari al LLMS

Per la maggior parte delle aziende, in base alle tendenze attuali, l'approccio alla distribuzione di LLMS può essere sintetizzato in 3 scenari di base. Come descritto in un recente ""Harvard Business Review"" articolo: (1) formazione (pre-formazione) un LLM da zero – costoso e richiede competenze ai/ML avanzate; (2) ottimizzazione di un modello di base con dati aziendali – complessi ma fattibili; (3) utilizzo della generazione aumentata di recupero (RAG) per eseguire query su archivi di documenti, API e database vettoriali che contengono dati aziendali. Ognuno di questi ha dei compromessi tra sforzo, velocità di iterazione, efficienza dei costi e accuratezza del modello nelle proprie implementazioni, utilizzati per risolvere diversi tipi di problemi (diagramma sotto).

Figura 3: Tipi di problemi



Modelli di base

Un modello di fondazione (FM) noto anche come modello di base è un modello di ai di grandi dimensioni (LLM) addestrato su grandi quantità di dati non marcati, utilizzando l'auto-supervisione su larga scala, generalmente adattato per un'ampia gamma di compiti NLP a valle. Poiché i dati di addestramento non sono etichettati dagli esseri umani, il modello emerge piuttosto che essere codificato esplicitamente. Ciò significa che il modello può generare storie o una propria narrazione senza essere esplicitamente programmato per farlo. Una caratteristica importante di FM è quindi l'omogeneizzazione, il che significa che lo stesso metodo viene utilizzato in molti domini. Tuttavia, con tecniche di personalizzazione e ottimizzazione, i sistemi FMS integrati nei prodotti che appaiono in questi giorni non sono solo efficaci nel generare testo, testo-immagini e testo-codice, ma anche per spiegare attività specifiche del dominio o codice di debug. Ad esempio, FMS come il Codex di OpenAI o il Codice Llama di Meta possono generare codice in più linguaggi di programmazione in base alle descrizioni del linguaggio naturale di un task di programmazione. Questi modelli sono esperti in più di una dozzina di linguaggi di programmazione, tra cui Python, C#, JavaScript, Perl, Ruby, e SQL. Essi comprendono l'intento dell'utente e generano codice specifico che esegue l'attività desiderata utile per lo sviluppo del software, l'ottimizzazione del codice e l'automazione delle attività di programmazione.

Messa a punto, specificità di dominio e riqualificazione

Una delle procedure comuni per l'implementazione di LLM dopo la preparazione dei dati e la pre-elaborazione dei dati consiste nella scelta di un modello pre-addestrato che sia stato addestrato su un insieme di dati ampio e diversificato. Nel contesto della messa a punto di precisione, questo può essere un modello di linguaggio open-source di grandi dimensioni come "Llama di meta 2" formazione basata su 70 miliardi di parametri e 2 trilioni di token. Una volta selezionato il modello pre-addestrato, il passo successivo è quello di ottimizzarlo sui dati specifici del dominio. Ciò implica la regolazione dei parametri del modello e la formazione sui nuovi dati per adattarli a un dominio e a un'attività specifici. Ad esempio, BloombergGPT, un LLM proprietario addestrato su un'ampia gamma di dati finanziari al servizio del settore finanziario. I modelli specifici del dominio progettati e addestrati per un'attività specifica hanno generalmente una maggiore precisione e prestazioni all'interno del

loro ambito, ma una bassa trasferibilità tra altre attività o domini. Quando l'ambiente aziendale e i dati cambiano nel corso di un certo periodo, l'accuratezza di previsione del FM potrebbe iniziare a diminuire rispetto alle prestazioni durante il test. Ciò avviene quando la riqualificazione o la messa a punto del modello diventa cruciale. Il riaddestramento dei modelli nelle tecniche ai/ML tradizionali si riferisce all'aggiornamento di un modello ML implementato con nuovi dati, generalmente eseguito per eliminare due tipi di derive che si verificano. (1) deriva del concetto – quando il collegamento tra le variabili di input e le variabili di target cambia nel tempo, poiché la descrizione di ciò che vogliamo prevedere le modifiche, il modello può produrre previsioni imprecise. (2) deriva dei dati: Si verifica quando cambiano le caratteristiche dei dati inseriti, come variazioni delle abitudini o del comportamento dei clienti nel tempo e, di conseguenza, l'incapacità del modello di reagire a tali cambiamenti. In un modo simile, la riqualificazione si applica a FMS/LLMS, tuttavia può essere molto più costoso (in milioni di dollari), quindi non qualcosa che la maggior parte delle organizzazioni potrebbe prendere in considerazione. È sotto ricerca attiva, ancora emergente nel regno di LLMOps. Quindi, invece di un nuovo training, quando il decadimento dei modelli si verifica nei sistemi FMS ottimizzati, le aziende possono optare per una nuova messa a punto (molto più economica) con un set di dati più recente. Dal punto di vista dei costi, di seguito è riportato un esempio di tabella prezzi modello di Azure-OpenAI Services. Per ogni categoria di attività, i clienti possono mettere a punto e valutare i modelli su set di dati specifici.

Fonte: Microsoft Azure

Model	**Per 1000 token**
Text-Ada	$0.0001
GPT-3.5 Turbo	$0.003
GPT-4	$0.06
Text-Davinci	$0.02
Model	**Per 100 images**
Dall-E	$2

Progettazione del prompt e inferenza

Il termine "progettazione rapida" si riferisce ai metodi efficaci per comunicare con LLMS per eseguire le attività desiderate senza aggiornare i pesi del modello. Tanto importante quanto il training e la messa a punto dei modelli di intelligenza artificiale sono per le applicazioni NLP, l'inferenza è ugualmente importante, laddove i modelli addestrati rispondono alle richieste degli utenti. I requisiti di sistema per l'inferenza sono generalmente molto più sulle performance in lettura del sistema storage ai, che invia i dati da LLMS alle GPU quando ha bisogno di applicare miliardi di parametri di modello memorizzati per produrre la migliore risposta.

LLMOps, Model Monitoring e Vectorstores

Come le operazioni MLOps (Machine Learning Ops) tradizionali, anche le operazioni LLMOps (Large Language Model Operations) richiedono la collaborazione di data scientist e tecnici DevOps con tool e Best practice per la gestione di LLMS negli ambienti di produzione. Tuttavia, il flusso di lavoro e lo stack tecnico per LLMS possono variare in alcuni modi. Ad esempio, le pipeline LLM create utilizzando framework come la stringa LangChain insieme a chiamate API LLM multiple verso endpoint esterni di incorporazione, come vectorstores o database vettoriali. L'utilizzo di un endpoint e di un vectorstore incorporati per i connettori downstream (come un database vettoriale) rappresenta un significativo sviluppo nel modo in cui i dati vengono memorizzati e accessibili. Rispetto ai tradizionali modelli ML sviluppati da zero, i sistemi LLMS spesso si

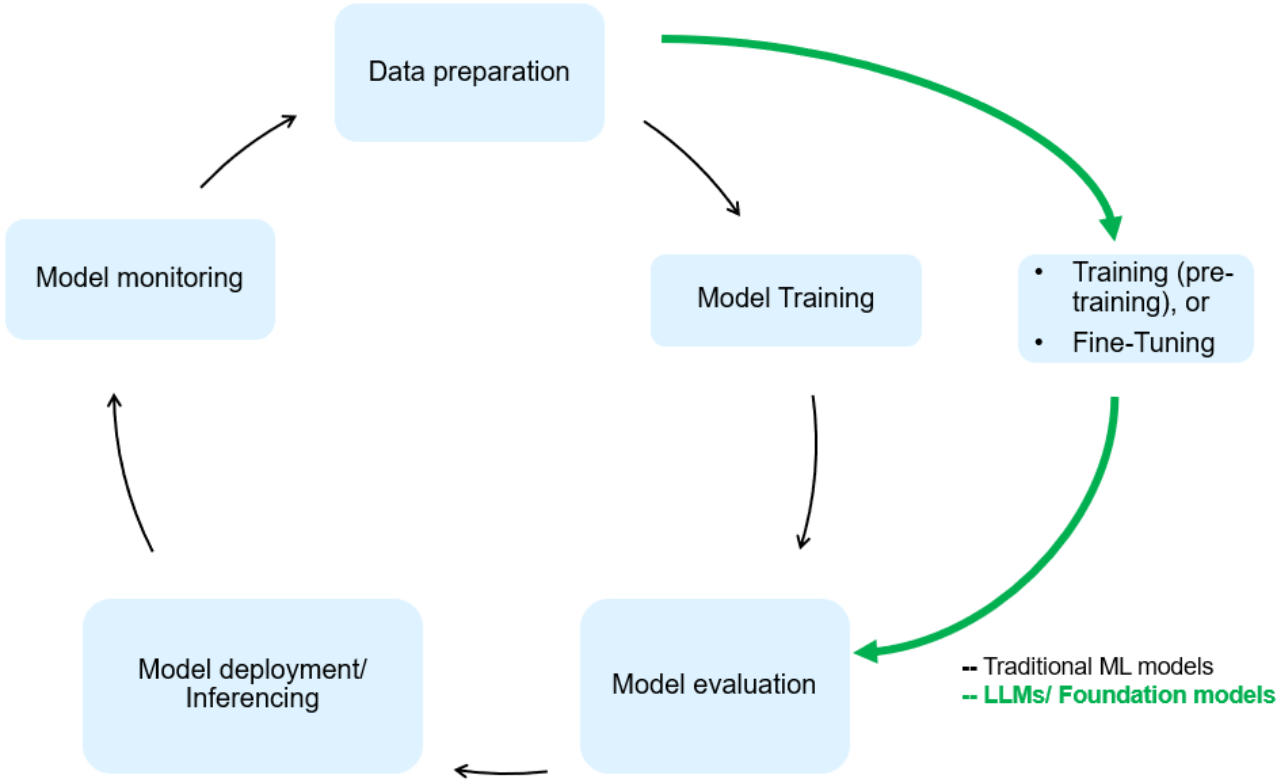
basano sull'apprendimento dei trasferimenti, poiché questi modelli iniziano con FMS che vengono ottimizzati con nuovi dati per migliorare le prestazioni in un dominio più specifico. Pertanto, è fondamentale che LLMOps fornisca le funzionalità di gestione del rischio e di monitoraggio del decadimento del modello.

Rischi ed etica nell'era dell'intelligenza artificiale generativa

"ChatGPT – è un'impresa intelligente, ma continua a generare assurdità."– MIT Tech Review. L'immondizia in–garbage out è sempre stato il caso più impegnativo con il computing. L'unica differenza con l'intelligenza artificiale generativa è che eccelle nel rendere la spazzatura altamente credibile, portando a risultati imprecisi. Le LLM sono inclini a inventare fatti che si adattino alla narrativa che sta costruendo. Pertanto, le aziende che considerano l'intelligenza artificiale generativa come una grande opportunità per ridurre i costi con equivalenti di intelligenza artificiale devono rilevare in modo efficiente i punti deboli, ridurre i pregiudizi e ridurre i rischi per mantenere i sistemi onesti ed etici. Una pipeline di dati a flusso libero con una solida infrastruttura ai che supporta mobilità dei dati, qualità dei dati, governance dei dati e data Protection tramite crittografia end-to-end e barriere all'ai è eminente nella progettazione di modelli ai responsabili e utilizzabili generativi.

Scenario cliente e NetApp

Figura 3: Flusso di lavoro del modello di apprendimento automatico/linguaggio di grandi dimensioni



Stiamo addestrando o perfezionando? la domanda se (a) addestrare un modello LLM da zero, mettere a punto un FM pre-addestrato, o utilizzare RAG per recuperare i dati da archivi di documenti al di fuori di un modello di base e aumentare i prompt, e (b) sfruttare LLMS open-source (ad esempio, Llama 2) o FMS proprietario (ad esempio, ChatGPT, Bard, AWS Bedrock) è una decisione strategica per le organizzazioni. Ogni approccio ha un compromesso tra efficienza dei costi, gravità dei dati, operazioni, accuratezza del modello e gestione di LLMS.

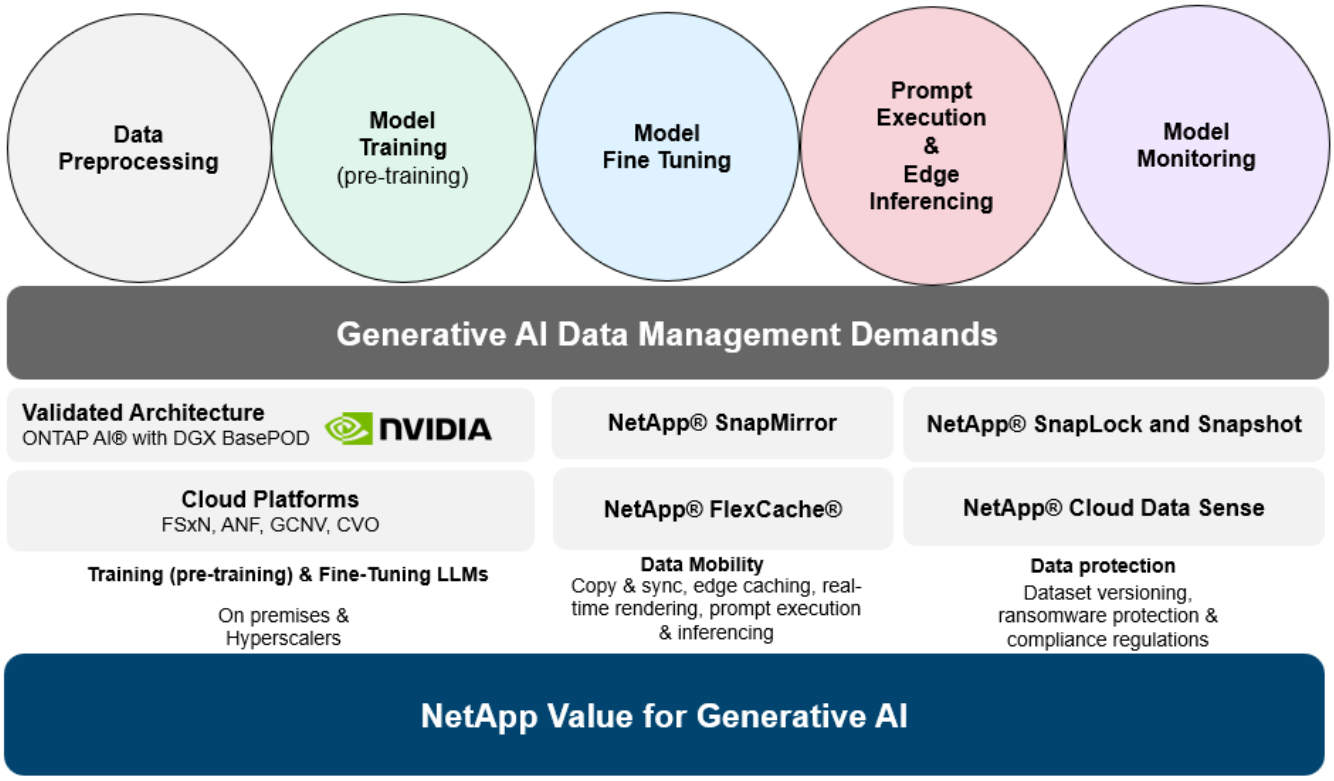
NetApp come azienda abbraccia l'ai internamente nella propria cultura lavorativa e nell'approccio alle attività di progettazione e progettazione dei prodotti. Ad esempio, la protezione autonoma da ransomware di NetApp è costruita utilizzando ai e machine learning. Permette di rilevare tempestivamente le anomalie del file system

per identificare le minacce prima che abbiano un impatto sulle operazioni. In secondo luogo, NetApp utilizza l'ai predittiva per le proprie operazioni di business, come le previsioni di vendite e inventario e i chatbot, per assistere i clienti nei servizi di supporto ai prodotti per call center, nelle specifiche tecniche, nella garanzia, nei manuali di assistenza e altro ancora. Terzo, NetApp porta il valore del cliente nella pipeline di dati ai e nel flusso di lavoro ML/LLM tramite prodotti e soluzioni, che offrono ai clienti la possibilità di creare soluzioni ai predittive come previsioni della domanda, imaging medico, analisi del sentimento, e soluzioni di intelligenza artificiale generativa come Gans per il rilevamento delle anomalie delle immagini industriali nel settore manifatturiero e anti-riciclaggio di denaro e rilevamento delle frodi nei servizi bancari e finanziari con prodotti e funzionalità NetApp come NetApp® ONTAP ai®, NetApp® SnapMirror® e NetApp® FlexCache®.

Funzionalità di NetApp

Lo spostamento e la gestione dei dati in applicazioni ai generative come chatbot, generazione di codice, generazione di immagini o espressione del modello del genoma possono estendersi a livello di edge, data center privato ed ecosistema multicloud ibrido. Ad esempio, un ai-bot in tempo reale che aiuti un passeggero ad aggiornare il proprio biglietto aereo alla classe business da un'app per l'utente finale esposta tramite API di modelli preformati come ChatGPT non può raggiungere tale compito da solo, poiché le informazioni sul passeggero non sono pubblicamente disponibili su Internet. L'API richiede l'accesso alle informazioni personali del passeggero e alle informazioni sui biglietti da parte del vettore aereo che potrebbe esistere in un ecosistema ibrido o multicloud. Uno scenario simile potrebbe essere applicato agli scienziati che condividono una molecola di farmaco e i dati del paziente tramite un'applicazione per l'utente finale che utilizza LLMS per eseguire sperimentazioni cliniche attraverso la scoperta di farmaci che coinvolgono istituti di ricerca biomedica uno a molti. I dati sensibili che vengono trasmessi a FMS o LLMS possono includere PII, informazioni finanziarie, informazioni sanitarie, dati biometrici, dati di posizione, dati di comunicazione, comportamento online e informazioni legali. In questo caso di rendering in real-time, esecuzione tempestiva e inferenza dell'edge, si verifica uno spostamento dei dati dall'app dell'utente finale agli endpoint di storage attraverso modelli LLM open source o proprietari, verso un data center on-premise o piattaforme di cloud pubblico. In tutti questi scenari, la mobilità e la protezione dei dati sono fondamentali per le operazioni ai che coinvolgono i sistemi LLMS che si basano su grandi set di dati di training e lo spostamento di tali dati.

Figura 4: Pipeline di dati ai generativa - LLM



Il portfolio NetApp di infrastruttura storage, servizi dati e cloud si basa sul software di Intelligent data management.

Data Preparation: Il primo pilastro dello stack tecnologico LLM non viene in gran parte toccato dal vecchio stack ML tradizionale. Il preprocessing dei dati nella pipeline ai è necessario per normalizzare e pulire i dati prima del training o del tuning. Questo passaggio include connettori per acquisire i dati ovunque si trovino sotto forma di Tier Amazon S3 o in sistemi storage on-premise come un file store o un archivio di oggetti come NetApp StorageGRID.

NetApp® ONTAP è la tecnologia di base che è alla base delle soluzioni di storage critiche di NetApp nei data center e nel cloud. ONTAP include varie funzionalità e funzionalità di gestione e protezione dei dati, tra cui protezione automatica dal ransomware contro gli attacchi informatici, funzionalità di trasporto dei dati integrate e funzionalità di efficienza dello storage per una serie di architetture da on-premise, ibride, multicloud in NAS, SAN, a oggetti, e situazioni di Software Defined Storage (SDS) delle implementazioni di LLM.

NetApp® ONTAP ai® per la formazione di modelli di apprendimento approfondito. NetApp® ONTAP® supporta NVIDIA GPU Direct Storage™ con l'utilizzo di NFS su RDMA per i clienti NetApp con cluster di storage ONTAP e nodi di calcolo NVIDIA DGX . Offre performance efficienti in termini di costi per leggere ed elaborare i set di dati di origine dallo storage in memoria numerose volte per promuovere l'intelligence, consentendo alle organizzazioni con training, messa a punto e scalabilità dell'accesso a LLMS.

NetApp® FlexCache® è una funzionalità di caching remoto che semplifica la distribuzione dei file e memorizza nella cache solo i dati attivamente letti. Ciò può essere utile per la formazione LLM, la riqualificazione e l'ottimizzazione, offrendo valore ai clienti con esigenze aziendali quali il rendering in tempo reale e l'inferenza LLM.

NetApp® SnapMirror è una funzione ONTAP che replica gli snapshot di volume tra due sistemi ONTAP. Questa funzionalità trasferisce i dati in maniera ottimale a livello di edge, nel data center on-premise o nel cloud. SnapMirror può essere utilizzato per spostare i dati in modo sicuro ed efficiente tra cloud on-premise e hyperscaler, quando i clienti desiderano sviluppare l'ai generativa nei cloud con RAG contenente i dati aziendali. Trasferisce in modo efficiente solo le modifiche, risparmiando larghezza di banda e velocizzando la replica, offrendo così funzionalità essenziali di mobilità dei dati durante le operazioni di formazione, riaddestramento e ottimizzazione di FMS o LLMS.

NetApp® SnapLock offre funzionalità disco immutabili sui sistemi di storage basati su ONTAP per la versione del set di dati. L'architettura del microcore è progettata per proteggere i dati dei clienti con il motore FPolicy™ Zero Trust. NetApp garantisce che i dati dei clienti siano disponibili resistendo agli attacchi DOS (Denial of Service) quando un utente malintenzionato interagisce con un LLM in modo particolarmente dispendioso in termini di risorse.

NetApp® Cloud Data Sense aiuta a identificare, mappare e classificare le informazioni personali presenti nei set di dati aziendali, attuare policy, soddisfare i requisiti di privacy on-premise o nel cloud, migliorare la sicurezza e rispettare le normative.

Classificazione NetApp® BlueXP™, basata su Cloud Data Sense. I clienti possono analizzare, categorizzare e agire automaticamente sui dati nel patrimonio dati, rilevare i rischi di sicurezza, ottimizzare lo storage e accelerare le implementazioni del cloud. Combina storage e servizi dati tramite il suo piano di controllo unificato, i clienti possono utilizzare istanze GPU per il calcolo e ambienti multicloud ibridi per il tiering cold storage e per archivi e backup.

Dualità file-oggetto NetApp. NetApp ONTAP consente un accesso dual-Protocol per NFS e S3. Con questa soluzione, i clienti possono accedere ai dati NFS dai notebook Amazon AWS SageMaker tramite bucket S3 di NetApp Cloud Volumes ONTAP. Ciò offre flessibilità ai clienti che necessitano di un facile accesso a origini dati eterogenee con la capacità di condividere i dati sia da NFS che da S3. Ad esempio, è possibile ottimizzare FMS come i modelli di generazione del testo Llama 2 di Meta su SageMaker con accesso ai bucket file-

oggetto.

Il servizio NetApp® Cloud Sync offre un modo semplice e sicuro per migrare i dati a qualsiasi destinazione, nel cloud o in sede. Cloud Sync trasferisce e sincronizza perfettamente i dati tra storage on-premise o cloud, NAS e archivi di oggetti.

NetApp XCP è un software client che consente migrazioni di dati da NetApp a NetApp rapide e affidabili. XCP offre anche la funzionalità di spostare in modo efficiente i dati in blocco dai file system Hadoop HDFS in ONTAP NFS, S3 o StorageGRID e le analitiche dei file XCP garantiscono visibilità nel file system.

NetApp® DataOps Toolkit è una libreria Python che semplifica l'esecuzione di varie attività di gestione dei dati da parte di data scientist, DevOps e data engineer, come il provisioning quasi istantaneo, il cloning o la creazione di snapshot di un volume di dati o di uno spazio di lavoro JupyterLab supportato da storage NetApp scale-out ad alte prestazioni.

Sicurezza dei prodotti NetApp. LLMS potrebbe rivelare inavvertitamente dati riservati nelle proprie risposte, quindi una preoccupazione per i CISO che studiano le vulnerabilità associate alle applicazioni ai che sfruttano LLMS. Come delineato da OWASP (Open Worldwide Application Security Project), problemi di sicurezza come avvelenamento dei dati, perdita di dati, negazione del servizio e rapide iniezioni all'interno di LLMS possono avere un impatto sulle aziende dall'esposizione dei dati agli attacchi degli utenti non autorizzati. I requisiti di archiviazione dei dati devono includere controlli di integrità e snapshot immutabili per dati strutturati, semi-strutturati e non strutturati. Le Snapshot di NetApp e SnapLock sono utilizzate per il controllo delle versioni del set di dati. Porta un rigido role-based access control (RBAC), così come protocolli sicuri e crittografia standard di settore per la protezione dei dati a riposo e in transito. Cloud Insights e Cloud Data Sense insieme offrono funzionalità che consentono di identificare l'origine della minaccia in modo forense e di assegnare priorità ai dati da ripristinare.

ONTAP ai con DGX BasePOD

L'architettura di riferimento NetApp® ONTAP® ai con NVIDIA DGX BasePOD è un'architettura scalabile per i workload di machine learning (ML) e intelligenza artificiale (ai). Per la fase di addestramento critico di LLMS, i dati vengono generalmente copiati dalla memoria dati nel cluster di addestramento a intervalli regolari. I server utilizzati in questa fase utilizzano le GPU per parallelizzare i calcoli, creando un enorme appetito per i dati. Soddisfare le esigenze di larghezza di banda i/o raw è fondamentale per mantenere un elevato utilizzo della GPU.

ONTAP ai con 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 ai Enterprise, basata su NetApp e VMware, offre gestione dei dati e dei workload ai di livello Enterprise in un pacchetto semplice e familiare.

1P piattaforme cloud

Le offerte di cloud storage completamente gestite sono disponibili nativamente su Microsoft Azure as Azure NetApp Files (ANF), su AWS come Amazon FSX per NetApp ONTAP (FSxN) e su Google come Google Cloud NetApp Volumes (GNCV). 1P è un file system gestito e dalle performance elevate che consente ai clienti di eseguire carichi di lavoro ai altamente disponibili con maggiore sicurezza dei dati nei cloud pubblici, per ottimizzare LLMS/FMS con piattaforme ML native del cloud come AWS SageMaker, Azure-OpenAI Services e Vertex ai di Google.

Suite di soluzioni per partner NetApp

Oltre ai suoi principali prodotti, tecnologie e funzioni per i dati, NetApp collabora inoltre in stretta collaborazione con una solida rete di partner ai per offrire valore aggiunto ai clienti.

NVIDIA Guardrails nei sistemi di intelligenza artificiale funge da salvaguardia per garantire l'uso etico e responsabile delle tecnologie di intelligenza artificiale. Gli sviluppatori di IA possono scegliere di definire il comportamento delle applicazioni basate su LLM su argomenti specifici e impedire loro di avviare discussioni su argomenti indesiderati. Guardrails, un toolkit open-source, consente di collegare un LLM ad altri servizi in modo semplice e sicuro per creare sistemi di conversazione LLM affidabili, sicuri e sicuri.

Domino Data Lab fornisce strumenti di livello aziendale versatili per la creazione e la produzione di intelligenza artificiale generativa, veloci, sicuri ed economici, ovunque vi troviate nel vostro viaggio di intelligenza artificiale. Con la piattaforma MLOps Enterprise di Domino, i data scientist possono utilizzare strumenti preferiti e tutti i loro dati, addestrare e implementare i modelli in modo semplice ovunque e gestire i rischi in modo economico, il tutto da un unico centro di controllo.

Modzy per Edge ai. NetApp® e Modzy hanno collaborato per offrire ai su larga scala a qualsiasi tipo di dati, tra cui immagini, audio, testo e tabelle. Modzy è una piattaforma MLOps per l'implementazione, l'integrazione e l'esecuzione di modelli ai, offre ai data scientist le funzionalità di monitoring dei modelli, rilevamento di deriva e spiegabilità, con una soluzione integrata per un'inferenza LLM perfetta.

Run:ai e NetApp hanno collaborato per dimostrare le funzionalità uniche della soluzione NetApp ONTAP ai con la piattaforma di gestione dei cluster Run:ai per semplificare l'orchestrazione dei workload ai. Suddivide e unisce automaticamente le risorse GPU, progettate per scalare le pipeline di elaborazione dati a centinaia di macchine con framework di integrazione integrati per Spark, Ray, Dask e Rapids.

Conclusione

L'intelligenza artificiale generativa può produrre risultati efficaci solo quando il modello è addestrato su una serie di dati di qualità. Sebbene LLMS abbia raggiunto importanti traguardi, è fondamentale riconoscerne i limiti, le sfide di progettazione e i rischi associati alla mobilità e alla qualità dei dati. I sistemi LLMS si basano su set di dati di training eterogenei e di grandi dimensioni provenienti da fonti eterogenee di dati. I risultati imprecisi o parziali generati dai modelli possono mettere a repentaglio sia le aziende che i consumatori. Questi rischi possono corrispondere a vincoli per LLMS che possono emergere potenzialmente da problemi di gestione dei dati associati alla qualità dei dati, alla sicurezza dei dati e alla mobilità dei dati. NetApp aiuta le organizzazioni a soddisfare le complessità create dalla rapida crescita dei dati, dalla mobilità dei dati, dalla gestione multicloud e dall'adozione dell'ai. L'infrastruttura ai su larga scala e la gestione efficiente dei dati sono fondamentali per definire il successo delle applicazioni ai come l'ai generativa. Sono clienti critici che coprono tutti gli scenari di implementazione, senza compromettere la capacità di espandersi quando le aziende hanno bisogno di mantenere sotto controllo l'efficienza dei costi, la governance dei dati e pratiche etiche di ai. NetApp lavora costantemente per aiutare i clienti a semplificare e accelerare le proprie implementazioni di IA.

TR-4785: Implementazione dell'ai con NetApp e-Series e BeeGFS

Nagalakshmi Raju, Daniel Landes, Nathan Swartz, Amine Bennani, NetApp

Le applicazioni di intelligenza artificiale (ai), machine learning (ML) e deep learning (DL) implicano grandi set di dati e calcoli elevati. Per eseguire con successo questi workload, è necessaria un'infrastruttura agile che consenta di scalare perfettamente i nodi di storage e di calcolo. Questo report include le fasi per l'esecuzione di un modello di

training ai in una modalità distribuita, che consente una scalabilità perfetta dei nodi di calcolo e storage. Il report include anche diverse metriche delle performance per mostrare come una soluzione che combina lo storage NetApp e-Series con il file system parallelo BeeGFS offre una soluzione semplice, flessibile e conveniente per i carichi di lavoro ai.

"TR-4785: Implementazione dell'ai con NetApp e-Series e BeeGFS"

NVA-1150-DESIGN: Guida alla progettazione dei sistemi Quantum StorNext con NetApp e-Series

Ryan Rodine, NetApp

Questo documento fornisce informazioni dettagliate su come progettare una soluzione di file system parallelo StorNext con i sistemi storage NetApp e-Series. Questa soluzione copre l'array all-flash NetApp EF280, l'array NVMe all-flash NetApp EF300, l'array NVMe all-flash EF600 e il sistema ibrido NetApp E5760. Offre una caratterizzazione delle performance basata sul benchmark Frametest, uno strumento ampiamente utilizzato per i test nel settore dei media e dell'intrattenimento.

"NVA-1150-DESIGN: Guida alla progettazione dei sistemi Quantum StorNext con NetApp e-Series"

NVA-1150-DEPLOY: Guida all'implementazione dei sistemi Quantum StorNext con NetApp e-Series

Ryan Rodine, NetApp

Questo documento fornisce informazioni dettagliate su come implementare una soluzione di file system parallelo StorNext con i sistemi storage NetApp e-Series. Questa soluzione copre l'array all-flash NetApp EF280, l'array NVMe all-flash NetApp EF300, l'array NVMe all-flash NetApp EF600 e il sistema ibrido NetApp E5760. Offre una caratterizzazione delle performance basata sul benchmark Frametest, uno strumento ampiamente utilizzato per i test nel settore dei media e dell'intrattenimento.

"NVA-1150-DEPLOY: Guida all'implementazione dei sistemi Quantum StorNext con NetApp e-Series"