Casos de uso

NetApp Solutions

NetApp
April 25, 2024

Tabla de contenidos

Casos de uso

IA responsable e inferencia confidencial: IA de NetApp con transformación de imagen Protopía

TR-4928: IA responsable e inferencia confidencial - NetApp AI con Protopía Image y transformación de datos

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

Las interpretaciones visuales se han convertido en parte integral de la comunicación con la aparición de la captura de imágenes y el procesamiento de imágenes. La inteligencia artificial (IA) en el procesamiento de imágenes digitales brinda nuevas oportunidades de negocio, como en el campo médico para la identificación del cáncer y otras enfermedades, en el análisis visual geoespacial para estudiar peligros ambientales, en el reconocimiento de patrones, en el procesamiento de vídeos para combatir la delincuencia, etc. Sin embargo, esta oportunidad también viene con responsabilidades extraordinarias.

Cuantas más decisiones las organizaciones se pongan en manos de la IA, más aceptan riesgos relacionados con la privacidad y la seguridad de los datos, así como con cuestiones legales, éticas y reguladoras. La IA responsable permite aplicar una práctica que permite a las empresas y organizaciones gubernamentales crear confianza y gobernanza, algo crucial para la IA a escala en grandes empresas. Este documento describe una solución de inferencia de IA validada por NetApp en tres situaciones diferentes utilizando las tecnologías de gestión de datos de NetApp con el software de ofuscación de datos de Protopía con el fin de privatizar datos confidenciales y reducir riesgos y preocupaciones éticas.

Cada día, los consumidores y entidades de negocio generan millones de imágenes con diversos dispositivos digitales. La consiguiente explosión masiva de datos y cargas de trabajo informáticas hace que las empresas recurren a plataformas de cloud computing para obtener escalado y eficiencia. Mientras tanto, la preocupación por la privacidad de la información confidencial que contienen los datos de imágenes surge como consecuencia de la transferencia a un cloud público. La falta de garantías de seguridad y privacidad se convierte en la principal barrera para el despliegue de sistemas de IA que procesan imágenes.

Además, está el "derecho a borrar" Según el RGPD, el derecho de una persona a solicitar que una organización borre todos sus datos personales. También está la "Ley de Privacidad", que establece un código de prácticas de información justas. Las imágenes digitales, como las fotografías, pueden constituir datos personales en virtud del RGPD, que rige la forma en que deben recopilarse, procesarse y borrarse los datos. El no hacerlo es un incumplimiento del RGPD, que puede conllevar multas cuantiosas por incumplimiento de las normativas que pueden resultar seriamente perjudiciales para las organizaciones. Los principios de privacidad son uno de los pilares de la implementación de IA responsable que garantice la justicia en las predicciones de modelos de aprendizaje automático y aprendizaje profundo (DL) y reduce los riesgos asociados con la infracción de privacidad o el cumplimiento de normativas.

Este documento describe una solución de diseño validada en tres situaciones diferentes, con y sin la confusión de imágenes relevante para preservar la privacidad y poner en marcha una solución de IA responsable:

- **Escenario 1.** inferencia a petición dentro del cuaderno Juppyter.
- **Escenario 2.** inferencia por lotes en Kubernetes.

- **Escenario 3.** servidor de inferencia NVIDIA Triton.

Para esta solución, utilizamos el conjunto de datos y el punto de referencia de detección facial (FDDB), un conjunto de datos de regiones faciales diseñado para estudiar el problema de la detección facial sin restricciones, combinado con el marco de aprendizaje de máquinas PyTorch para la implementación de FaceBox. Este conjunto de datos contiene las anotaciones para 5171 caras en un conjunto de 2845 imágenes de varias resoluciones. Además, este informe técnico presenta algunas áreas de soluciones y casos de uso relevantes recopilados por clientes de NetApp e ingenieros de campo en situaciones en las que sea aplicable esta solución.

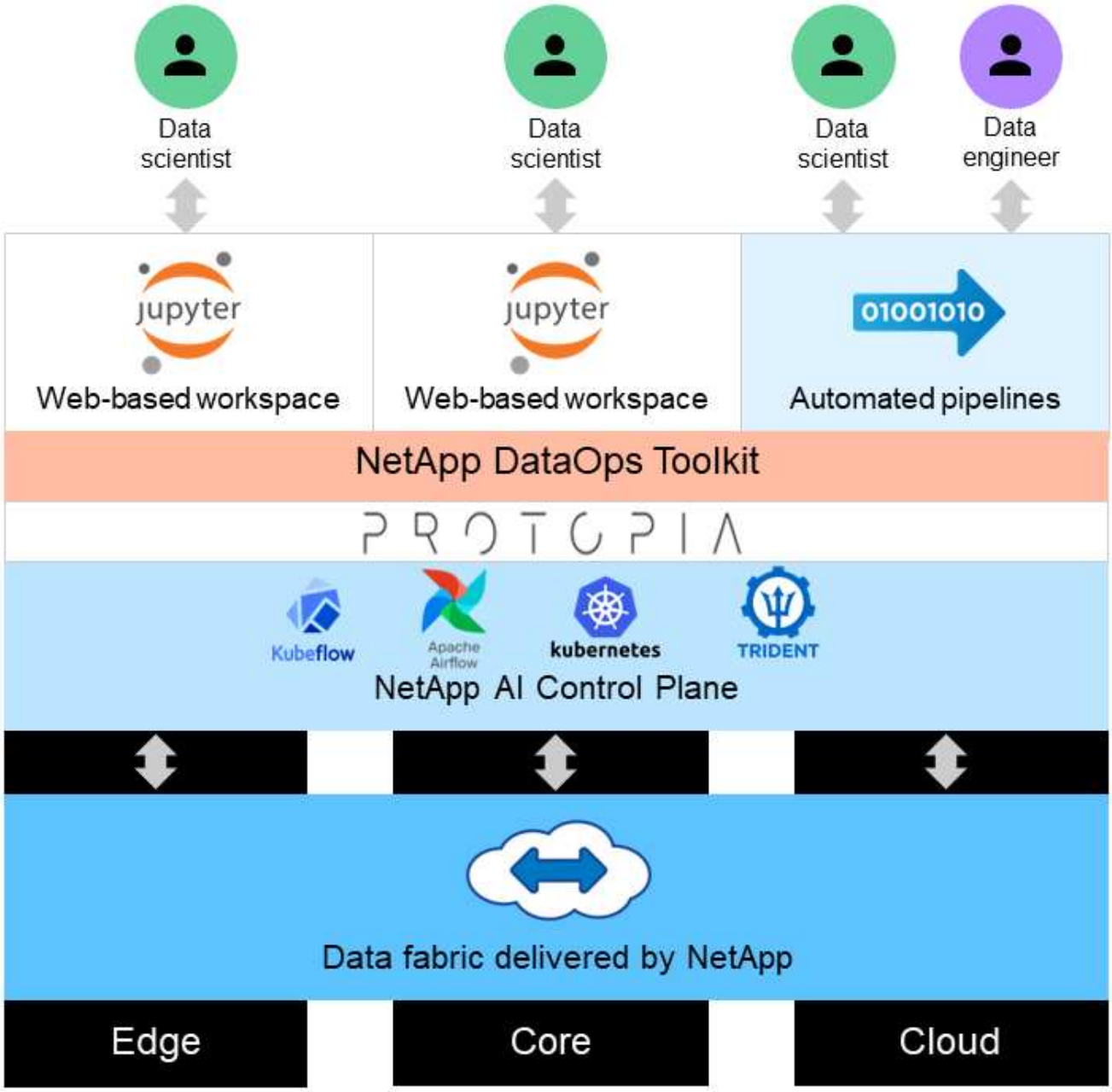
Público objetivo

Este informe técnico está dirigido a los siguientes destinatarios:

- Líderes empresariales y arquitectos empresariales que desean diseñar y poner en marcha una IA responsable y abordar cuestiones de protección y privacidad de datos relacionadas con el procesamiento de imágenes faciales en espacios públicos.
- Científicos de datos, ingenieros de datos, investigadores DE IA/aprendizaje automático (ML) y desarrolladores de sistemas de IA/ML que pretenden proteger y preservar la privacidad.
- Arquitectos empresariales que diseñan soluciones de confusión de datos para modelos y aplicaciones de IA/ML que cumplen con las normativas tales como RGPD, CCPA o las organizaciones gubernamentales (Ley de privacidad del Departamento de Defensa) y la Ley de privacidad del Departamento de Defensa (DoD).
- Científicos e ingenieros de IA buscan formas eficientes de poner en marcha el aprendizaje profundo (DL) y modelos de inferencia de IA/ML/DL que protegen la información confidencial.
- Los administradores de dispositivos periféricos y los administradores de servidor perimetral son responsables de la puesta en marcha y la gestión de modelos de inferencia perimetrales.

Arquitectura de la solución

Esta solución está diseñada para gestionar cargas de trabajo de IA de inferencia en lote y en tiempo real en grandes conjuntos de datos utilizando la potencia de procesamiento de las GPU junto con las CPU tradicionales. Esta validación demuestra la inferencia privacidad de ML y la gestión de datos óptima que necesitan las organizaciones que buscan una puesta en marcha de IA responsable. Esta solución proporciona una arquitectura adecuada para una plataforma Kubernetes de uno o varios nodos para el perímetro y el cloud computing interconectados con ONTAP AI de NetApp en el núcleo de las instalaciones, el kit de herramientas DataOPS de NetApp y el software de ofuscación por medio de las interfaces CLI y de Juppyter Lab. En la siguiente figura se muestra información general sobre la arquitectura lógica del Data Fabric con la tecnología de NetApp con el kit de herramientas de operaciones de datos y Protopía.



El software de ofuscación de Protopía se ejecuta sin problemas sobre el kit de herramientas DataOPS de NetApp y transforma los datos antes de salir del servidor de almacenamiento.

Áreas de soluciones

El procesamiento de imágenes digitales ofrece muchas ventajas, lo que permite a muchas organizaciones aprovechar al máximo los datos asociados con las representaciones visuales. Esta solución de NetApp y Protopia ofrece un diseño único de inferencia de IA para proteger y privatizar datos de IA/ML en TODO el ciclo de vida DE ML/DL. Permite a los clientes conservar la propiedad de datos confidenciales, utilizar modelos de puesta en marcha de cloud público o híbrido para escalar y eficiencia al eliminar las preocupaciones relacionadas con la privacidad y poner en marcha la inferencia de IA en el perímetro.

Inteligencia ambiental

Hay muchas maneras en que las industrias pueden aprovechar los análisis geoespaciales en las áreas de peligros ambientales. Los gobiernos y el departamento de obras públicas pueden obtener información práctica sobre la salud pública y las condiciones climáticas para asesorar mejor al público durante una pandemia o un desastre natural como los incendios forestales. Por ejemplo, puede identificar a un paciente con un COVID positivo en espacios públicos, como aeropuertos o hospitales, sin poner en peligro la privacidad del individuo afectado y alertar a las autoridades respectivas y al público cercano acerca de las medidas de seguridad necesarias.

Dispositivos portátiles

En el ejército y en campos de batalla, se puede utilizar la inferencia de IA en el borde como dispositivos portátiles para rastrear la salud de los soldados, monitorear el comportamiento de los conductores y alertar a las autoridades sobre la seguridad y los riesgos asociados de acercarse a los vehículos militares mientras se preserva y protege la privacidad de los soldados. El futuro de los militares está yendo de alta tecnología con el Internet de Battlefield Things (IoBT) y el Internet de las cosas militares (IoMT) por llevar equipos de combate que ayudan a los soldados a identificar a los enemigos y a actuar mejor en la batalla mediante la computación de avanzada. Proteger y preservar los datos visuales recopilados de dispositivos periféricos como aviones teledirigidos y engranajes portátiles es crucial para mantener a raya a los hackers y al enemigo.

Operaciones de evacuación no combatiente

Las operaciones de evacuación no combatiente (Neos) las lleva a cabo el Departamento de Defensa para ayudar a evacuar a ciudadanos y nacionales estadounidenses, personal civil del Departamento de Defensa y personas designadas (nación anfitriona (HN) y nacionales de terceros países (TCN) cuyas vidas están en peligro de un refugio seguro apropiado. Los controles administrativos vigentes utilizan en gran medida procesos manuales de detección de evacuados. Sin embargo, la precisión, la seguridad y la velocidad de la identificación del evacuado, el seguimiento del evacuado y la detección de amenazas podrían mejorarse utilizando herramientas de IA/ML altamente automatizadas combinadas con tecnologías de ofuscación por vídeo de IA/ML.

Salud e investigación biomédica

El procesamiento de imágenes se utiliza para diagnosticar patologías para la planificación quirúrgica a partir de imágenes 3D obtenidas de tomografía computarizada (TC) o resonancia magnética (RM). Las reglas de privacidad de HIPAA rigen cómo deben recopilarse, procesar y borrar los datos para toda la información personal e imágenes digitales como fotografías. Para que los datos se puedan calificar para que se puedan compartir con la normativa HIPAA Safe Harbor, es necesario eliminar las imágenes fotográficas de cara completa y cualquier imagen comparable. Las técnicas automatizadas como la desidentificación o-los algoritmos de decapado de cráneo utilizados para ocultar las características faciales de un individuo a partir de las imágenes estructurales de TC/RM se han convertido en una parte esencial del proceso de intercambio de datos para las instituciones de investigación biomédica.

Migración al cloud de los análisis de IA/ML

Los clientes empresariales han formado y puesto en marcha modelos de IA/ML en sus instalaciones. Por razones de escalado y eficiencia, estos clientes amplían su capacidad para trasladar las funciones de IA/ML a puestas en marcha de cloud público, híbrido o multicloud. Sin embargo, están ligados a qué datos se pueden exponer a otras infraestructuras. Las soluciones de NetApp se plantean una amplia gama de amenazas de ciberseguridad para las que es necesario "protección de datos" La evaluación de la seguridad y, cuando se combina con la transformación de datos de Protopía, minimizan los riesgos asociados con la migración de cargas de trabajo DE IA/ML de procesamiento de imágenes al cloud.

Para ver casos de uso adicionales para la computación perimetral y la inferencia de IA en otros sectores,

consulte "TR-4886 inferencia de IA en el perímetro" Y el blog de IA de NetApp, "Inteligencia frente a privacidad".

Información general de la tecnología

En esta sección se ofrece una descripción general de los distintos componentes técnicos necesarios para completar esta solución.

Protopía

Protopia AI ofrece una solución discreta y exclusiva de software para la inferencia confidencial en el mercado actual. La solución Protopia ofrece una protección sin igual para servicios de inferencia al minimizar la exposición de información confidencial. La IA solo se alimenta de la información en el registro de datos que es realmente esencial para realizar la tarea a mano y nada más. La mayoría de las tareas de inferencia no utilizan toda la información que existe en cada registro de datos. Independientemente de si su IA consume imágenes, voz, vídeo o incluso datos tabulares estructurados, Protopia solo ofrece lo que el servicio de inferencia necesita. La tecnología de núcleo patentada utiliza el ruido matemáticamente curado para transformar estocamente los datos y engardar la información que no necesita un servicio DE ML dado. Esta solución no enmascara los datos; más bien, cambia la representación de datos mediante el uso de ruido aleatorio curado.

La solución Protopia formula el problema de cambiar la representación como un método de maximización de turbación basado en gradiente que aún conserva la información pertinente en el espacio de características de entrada con respecto a la funcionalidad del modelo. Este proceso de detección se ejecuta como un pase de ajuste preciso al final del entrenamiento del modelo ML. Después de que el paso genera automáticamente un conjunto de distribuciones de probabilidad, una transformación de datos de baja sobrecarga aplica muestras de ruido de estas distribuciones a los datos, ocultando dichos datos antes de pasarlos al modelo para la inferencia.

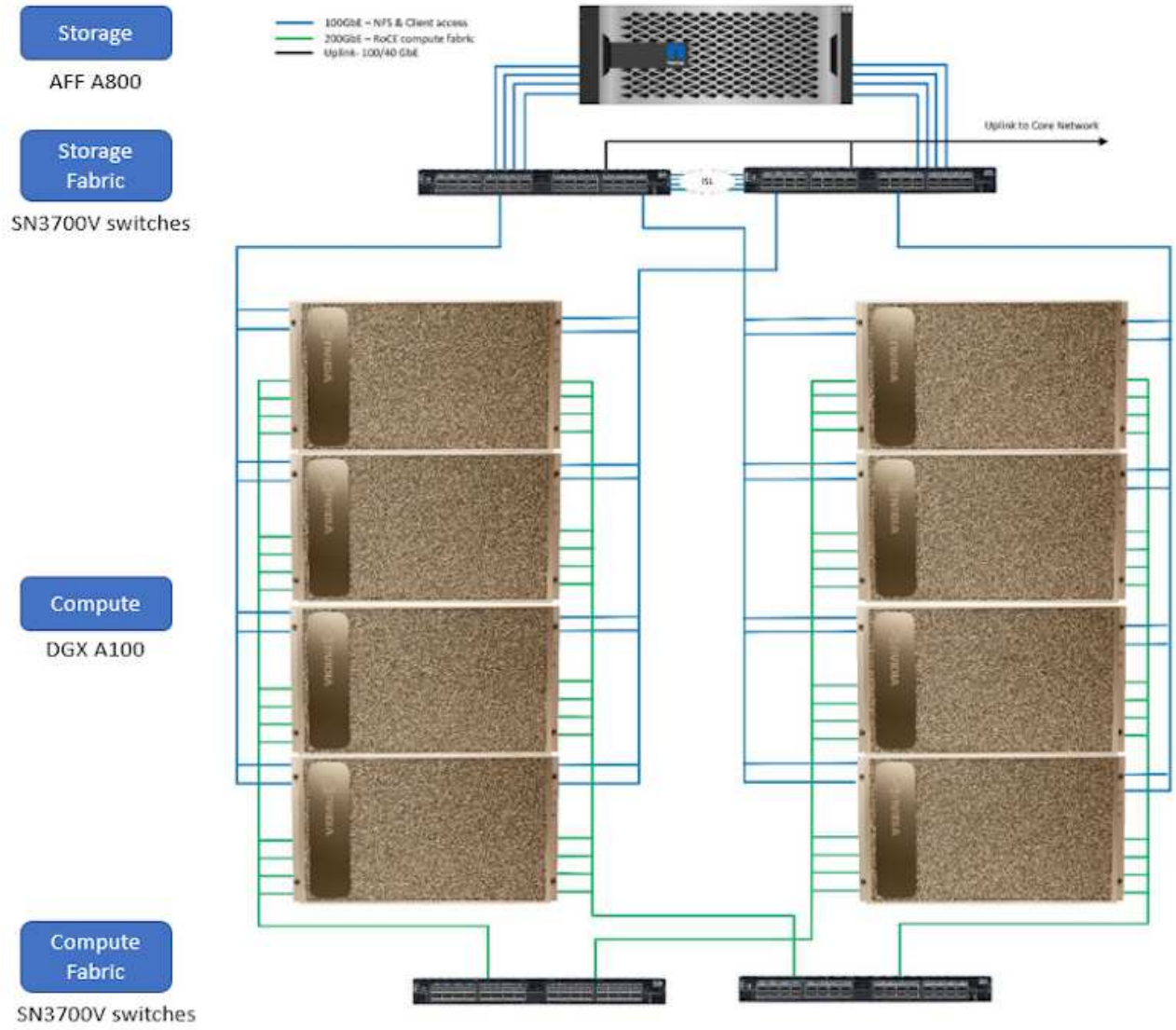
ONTAP AI de NetApp

La arquitectura de referencia de ONTAP AI de NetApp, con sistemas DGX A100 y sistemas de almacenamiento conectados al cloud de NetApp, ha sido desarrollada y verificada por NetApp y NVIDIA. Proporciona a las organizaciones DE TI una arquitectura que ofrece las siguientes ventajas:

- Elimina las complejidades de diseño
- Permite un escalado independiente de las capacidades de computación y almacenamiento
- Permite a los clientes empezar con poco y escalar sin problemas
- Ofrece opciones de almacenamiento para distintos niveles de rendimiento y coste

ONTAP AI integra perfectamente los sistemas DGX A100 y los sistemas de almacenamiento AFF A800 de NetApp con una red de vanguardia. ONTAP AI simplifica las puestas en marcha de IA eliminando complejidades y conjeturas de diseño. Los clientes pueden empezar con poco e ir creciendo de forma no disruptiva a la vez que gestionan de forma inteligente los datos desde el perímetro hasta el núcleo, pasando por el cloud.

En la siguiente figura, se muestran varias variaciones de la familia de soluciones de IA de ONTAP con sistemas DGX A100. El rendimiento del sistema AFF A800 se verifica con hasta ocho sistemas DGX A100. Al añadir pares de controladoras de almacenamiento al clúster ONTAP, la arquitectura puede escalarse a varios racks y admitir muchos sistemas DGX A100 y petabytes de capacidad de almacenamiento con rendimiento lineal. Este enfoque permite alterar de forma independiente las tasas de computación a almacenamiento en función del tamaño de los modelos de AP utilizados y de las métricas de rendimiento necesarias.



Para obtener más información sobre ONTAP AI, consulte "NVA-1153: ONTAP AI de NetApp con sistemas NVIDIA DGX A100 y switches Mellanox Spectrum Ethernet."

ONTAP de NetApp

ONTAP 9.11, la última generación del software de gestión del almacenamiento de NetApp, permite a las empresas modernizar su infraestructura y realizar la transición a un centro de datos preparado para el cloud. ONTAP ofrece las mejores capacidades de gestión de datos y permite la gestión y protección de los datos con un solo conjunto de herramientas, sin importar dónde residan. También puede mover los datos libremente a donde sea necesario: El perímetro, el núcleo o el cloud. ONTAP 9.11 incluye numerosas funciones que simplifican la gestión de datos, aceleran y protegen los datos esenciales y permiten disfrutar de funcionalidades de infraestructura de nueva generación en arquitecturas de cloud híbrido.

Kit de herramientas de operaciones de datos de NetApp

El kit de herramientas DataOps de NetApp es una biblioteca Python que facilita a los desarrolladores, científicos de datos, ingenieros de DevOps e ingenieros de datos la realización de varias tareas de gestión de datos, como el aprovisionamiento casi instantáneo de un nuevo volumen de datos o un espacio de trabajo JuppyterLab, el clonado casi instantáneo de un volumen de datos o un espacio de trabajo JuppyterLab. Y tomar instantáneas de un volumen de datos o espacio de trabajo JupyterLab para su trazabilidad o línea de

base. Esta biblioteca de Python puede funcionar como una utilidad de línea de comandos o una biblioteca de funciones que puede importar a cualquier programa de Python o a cualquier cuaderno de Jupyter.

Servidor de inferencia NVIDIA Triton

El servidor de inferencia de NVIDIA Triton es un software de servicio de inferencia de código abierto que ayuda a estandarizar la puesta en marcha y ejecución de modelos para ofrecer IA rápida y escalable en producción. Triton Invalidate Server optimiza la inferencia de IA al permitir a los equipos poner en marcha, ejecutar y escalar modelos de IA entrenados desde cualquier marco en cualquier infraestructura basada en GPU o CPU. El servidor de inferencia de Triton admite los principales marcos de trabajo, como TensorRT, NVIDIA TensorRT, PyTorch, MXNet, OpenVINO, etc. Triton se integra con Kubernetes para la orquestación y el escalado, que puede utilizar en las principales plataformas de inteligencia artificial y Kubernetes. También está integrada con muchas soluciones de software de MLOPS.

PyTorch

"PyTorch" Es un marco DE APRENDIZAJE AUTOMÁTICO de código abierto. Se trata de una biblioteca tensora optimizada para el aprendizaje profundo que utiliza GPU y CPU. El paquete PyTorch contiene estructuras de datos para tensores multidimensionales que proporcionan muchas utilidades para serializar eficazmente los tensores entre otras utilidades útiles. También tiene un par CUDA que permite ejecutar los cálculos tensores en una GPU de NVIDIA con capacidad de computación. En esta validación, utilizamos la biblioteca de OpenCV-Python (cv2) para validar nuestro modelo y aprovechar los conceptos de visión computarizada más intuitivos de Python.

Simplificar la gestión de los datos

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

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

- Compactación de datos inline y deduplicación expandida. La compactación de datos reduce el espacio perdido dentro de los bloques de almacenamiento, mientras que la deduplicación aumenta la capacidad efectiva de forma significativa. Esto es aplicable a los datos almacenados localmente y a los datos organizados en niveles en el cloud.
- Calidad de servicio (AQoS) mínima, máxima y adaptativa. Los controles granulares de calidad de servicio (QoS) ayudan a mantener los niveles de rendimiento para aplicaciones críticas en entornos altamente compartidos.
- FabricPool de NetApp. Proporciona la organización automática en niveles de datos fríos en opciones de almacenamiento en cloud privado como Amazon Web Services (AWS), Azure y la solución de almacenamiento StorageGRID de NetApp. Para obtener más información sobre FabricPool, consulte "TR-4598: Prácticas recomendadas de FabricPool".

Acelere y proteja sus datos

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

- Rendimiento y menor latencia. ONTAP ofrece la salida más alta posible con la menor latencia posible.
- Protección de datos. ONTAP ofrece capacidades integradas de protección de datos, con una administración común entre todas las plataformas.

- Cifrado de volúmenes de NetApp (NVE). ONTAP ofrece cifrado nativo en el nivel de volumen y permite la gestión de claves incorporada o externa.
- Multi-tenancy y autenticación multifactor. ONTAP permite compartir recursos de infraestructura con los niveles más altos de seguridad.

Infraestructura preparada para futuros retos

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

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

Control Astra de NetApp

La familia de productos Astra de NetApp ofrece servicios de gestión de datos para aplicaciones y almacenamiento para aplicaciones de Kubernetes en las instalaciones y en el cloud público, con la tecnología de gestión de datos y almacenamiento de NetApp. Le permite realizar fácilmente backups de aplicaciones Kubernetes, migrar datos a un clúster diferente y crear, de forma instantánea, clones de aplicaciones de trabajo. Si necesita gestionar aplicaciones de Kubernetes que se ejecutan en un cloud público, consulte la documentación de "Servicio de control Astra". Astra Control Service es un servicio gestionado por NetApp que proporciona gestión de datos para aplicaciones de clústeres de Kubernetes en Google Kubernetes Engine (GKE) y Azure Kubernetes Service (AKS).

Astra Trident de NetApp

Astra "Trident" De NetApp es un orquestador de almacenamiento dinámico de código abierto para Docker y Kubernetes que simplifica la creación, la gestión y el consumo de almacenamiento persistente. Trident, una aplicación nativa de Kubernetes, se ejecuta directamente dentro de un clúster de Kubernetes. Trident permite que los clientes implementen sin problemas imágenes de contenedores de DL en el almacenamiento de NetApp y proporciona una experiencia de clase empresarial para implementaciones de contenedores de IA. Los usuarios de Kubernetes (desarrolladores DE ML, científicos de datos, etc.) pueden crear, gestionar y automatizar la orquestación y el clonado para aprovechar las funcionalidades avanzadas de gestión de datos que se ofrecen con la tecnología de NetApp.

Copia y sincronización de NetApp BlueXP

"Copia y sincronización de BlueXP" Es un servicio de NetApp que ofrece una sincronización de datos rápida y segura. Ya tenga que transferir archivos entre recursos compartidos de archivos NFS o SMB en las instalaciones, 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, Google Cloud Storage, o IBM Cloud Object Storage, BlueXP Copy and Sync mueve los archivos a donde los necesites de forma rápida y segura. Una vez transferidos los datos, estarán completamente disponibles para su uso tanto en origen como en destino. BlueXP Copy y Syncc sincronizan continuamente los datos en

función de tu programación predefinida, moviendo solo los deltas, por lo que se reducen al mínimo el tiempo y el dinero que se invierten en la replicación de datos. Copia y sincronización de BlueXP es una herramienta de software como servicio (SaaS) extremadamente sencilla de configurar y utilizar. Las transferencias de datos activadas por BlueXP Copy and Sync se llevan a cabo por agentes de datos. Puedes poner en marcha agentes de datos de BlueXP Copy y Sync en AWS, Azure, Google Cloud Platform o en las instalaciones.

Clasificación de NetApp BlueXP

Impulsado por potentes algoritmos de IA, "Clasificación de NetApp BlueXP" proporciona controles automatizados y control de datos en todos sus datos. Puede localizar con facilidad el ahorro de costes, identificar problemas relacionados con el cumplimiento de normativas y la privacidad, y buscar oportunidades de optimización. La consola de clasificación de BlueXP le ofrece la información necesaria para identificar los datos duplicados y eliminar la redundancia, asignar datos personales, no personales y confidenciales, así como activar alertas para datos confidenciales y anomalías.

Plan de pruebas y validación

Para este diseño de solución, se validaron los tres siguientes supuestos:

- Una tarea de inferencia, con y sin Protopía ofuscación, dentro de un espacio de trabajo JupyterLab que fue orquestada mediante el kit de herramientas DataOPS de NetApp para Kubernetes.
- Una tarea de inferencia por lotes, con y sin ofuscación de Protopía, en Kubernetes con un volumen de datos orquestado mediante el kit de herramientas DataOps de NetApp para Kubernetes.
- Una tarea de inferencia mediante una instancia del servidor de inferencia de NVIDIA Triton orquestada mediante el kit de herramientas NetApp DataOPS para Kubernetes. Aplicamos la confusión Protopia a la imagen antes de llamar a la API de inferencia Triton para simular el requisito común de que cualquier dato que se transmita a través de la red debe ser ocultado. Este flujo de trabajo es aplicable para casos en los que los datos se recopilan en una zona de confianza, pero debe pasarse fuera de esa zona de confianza para la inferencia. Sin la ocultación de Protopia, no es posible implementar este tipo de flujo de trabajo sin que los datos confidenciales salgan de la zona de confianza.

Configuración de prueba

La siguiente tabla describe el entorno de validación del diseño de la solución.

Componente	Versión
Kubernetes	1.21.6
Controlador Astra Trident CSI de NetApp	22.01.0
Kit de herramientas Data OPS de NetApp para Kubernetes	2.3.0
Servidor de inferencia NVIDIA Triton	21.11-py3

Procedimiento de prueba

En esta sección se describen las tareas necesarias para completar la validación.

Requisitos previos

Para ejecutar las tareas descritas en esta sección, debe tener acceso a un host Linux o MacOS con las

siguientes herramientas instaladas y configuradas:

- Kubectl (se configura para acceder a un clúster de Kubernetes existente)
  - Se pueden encontrar instrucciones de instalación y configuración "aquí".
- Kit de herramientas Data OPS de NetApp para Kubernetes
  - Se pueden encontrar instrucciones de instalación "aquí".

Escenario 1 – inferencia bajo demanda en JuppyterLab

1. Cree un espacio de nombres de Kubernetes para las cargas de trabajo de inferencia de IA/ML.

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

2. Utilice el kit de herramientas DataOPS de NetApp para aprovisionar un volumen persistente para almacenar los datos en los que realizará la inferencia.

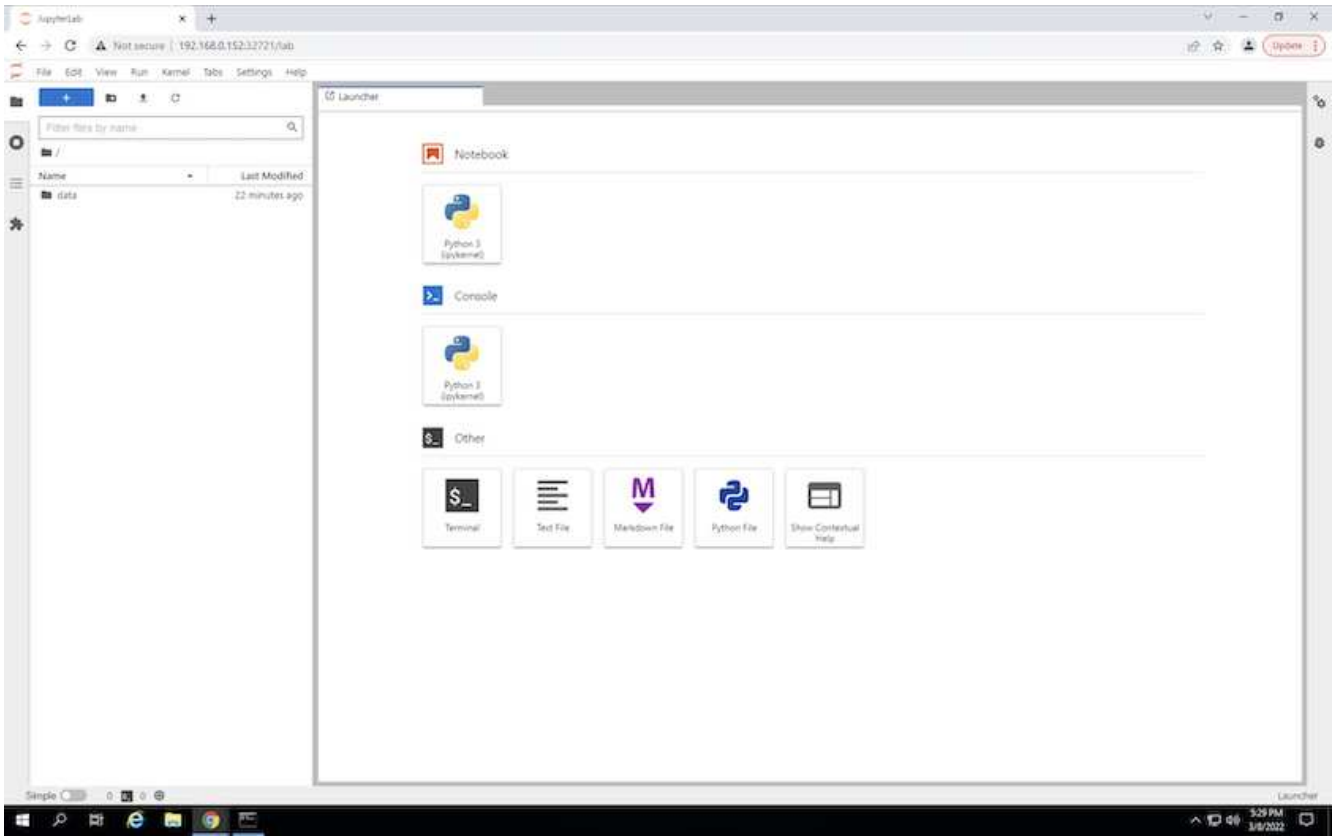
```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=inference-data --size=50Gi
Creating PersistentVolumeClaim (PVC) 'inference-data' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'inference-data' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully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

3. Use el kit de herramientas DataOPS de NetApp para crear un nuevo espacio de trabajo JuppyterLab. Monte el volumen persistente que se creó en el paso anterior mediante el `--mount- pvc` opción. Asigne las GPU de NVIDIA al espacio de trabajo según sea necesario mediante el `-- nvidia-gpu` opción.

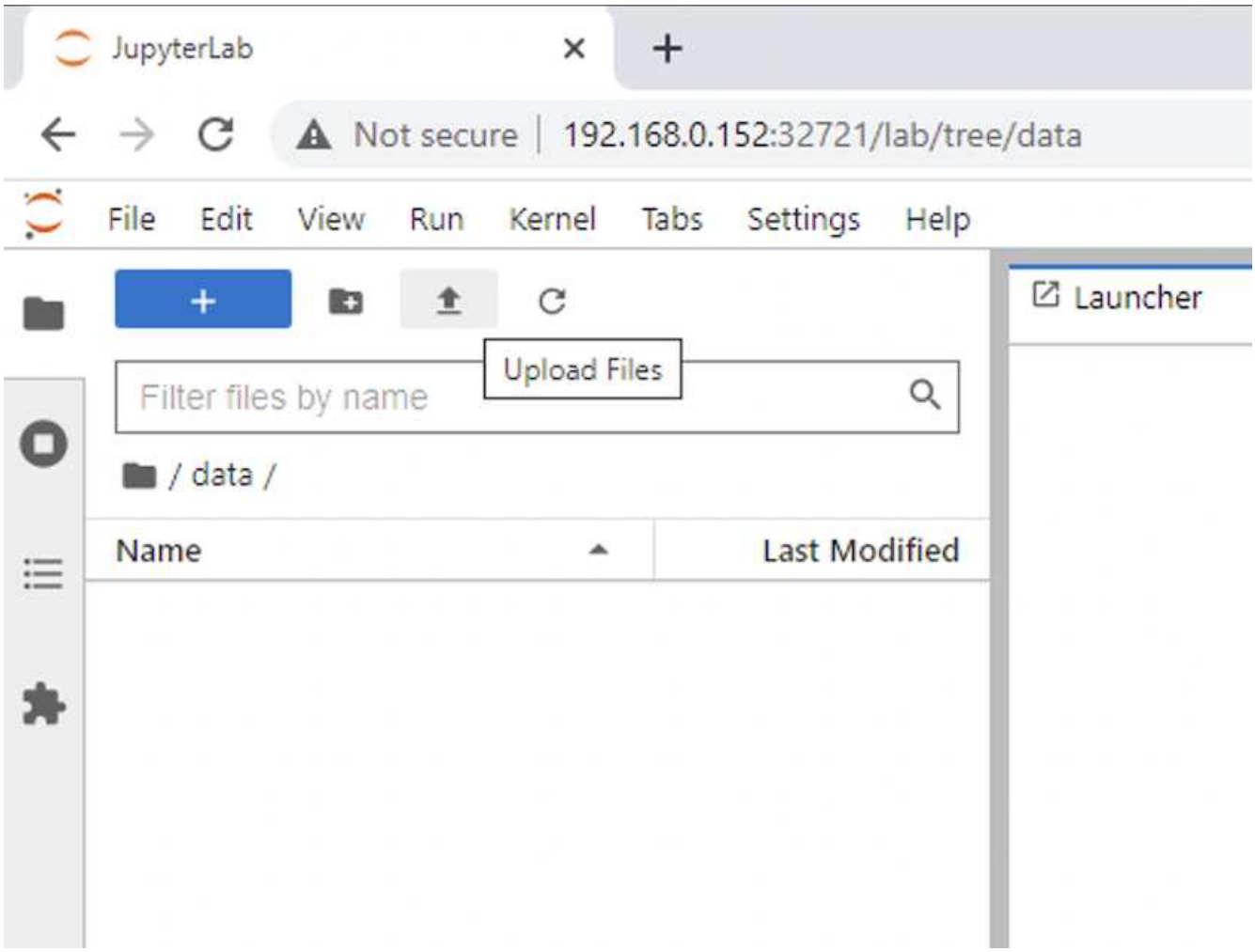
   En el siguiente ejemplo, el volumen persistente `inference-data` Está montado en el contenedor de espacio de trabajo JJupyterLab en `/home/jovyan/data`. Cuando utilice las imágenes del contenedor de Jupyter del proyecto oficial, `/home/jovyan` Se presenta como el directorio de nivel superior dentro de la interfaz Web JuppyterLab.

```
$ 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
```

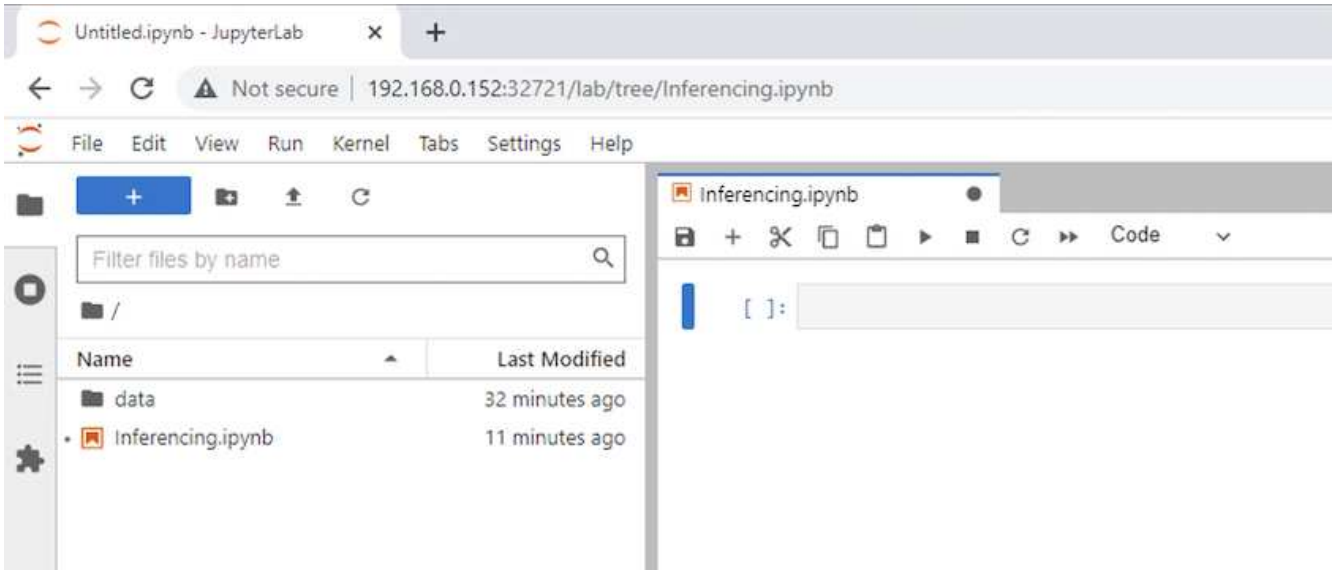
4. Acceda al espacio de trabajo JupyterLab utilizando la dirección URL especificada en la salida del `create jupyterlab` comando. El directorio de datos representa el volumen persistente que se montó en el espacio de trabajo.



5. Abra el `data` directory y cargue los archivos en los que se va a realizar la inferencia. Cuando se cargan archivos en el directorio de datos, se almacenan automáticamente en el volumen persistente que se montó en el espacio de trabajo. Para cargar archivos, haga clic en el icono cargar archivos, como se muestra en la siguiente imagen.



6. Vuelva al directorio de nivel superior y cree un nuevo portátil.



7. Agregue el código de inferencia al cuaderno. En el siguiente ejemplo, se muestra el código de inferencia para un caso de uso de detección de imagen.

Launcher | image-demo-pytorch.ipynb

Code | Python 3 (ipykernel)

STEP 3-1: Clean (Without obfuscation) detection

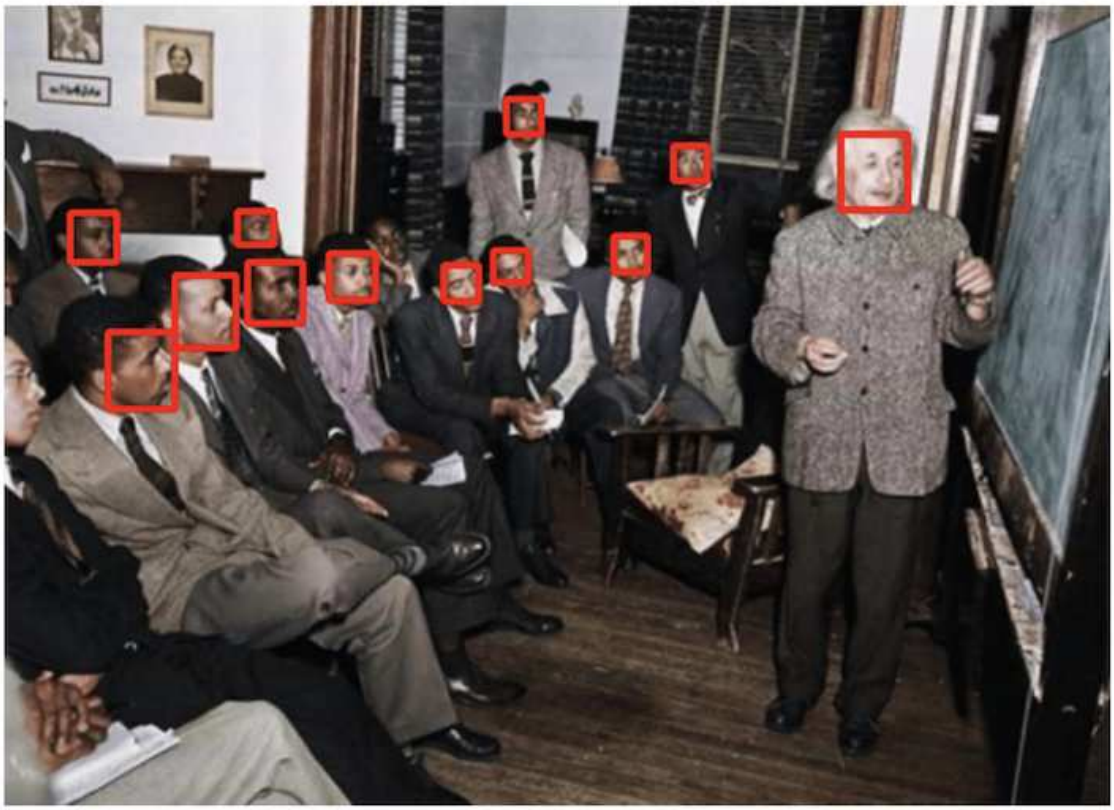
[9]:
```
# get current frame
frame = input_image

# 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. Agregue la ofuscación Protopia al código de inferencia. Protopia trabaja directamente con los clientes para proporcionar documentación específica para casos de uso y está fuera del alcance de este informe técnico. En el siguiente ejemplo se muestra el código de inferencia para un caso de uso de detección de imágenes con ofuscación Protopía agregada.

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)
```

Escenario 2: Inferencia por lotes en Kubernetes

1. Cree un espacio de nombres de Kubernetes para las cargas de trabajo de inferencia de IA/ML.

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

2. Utilice el kit de herramientas DataOPS de NetApp para aprovisionar un volumen persistente para almacenar los datos en los que realizará la inferencia.

```
$ 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. Rellene el nuevo volumen persistente con los datos en los que realizará la inferencia.

   Existen varios métodos para cargar datos en un PVC. Si actualmente sus datos están almacenados en una plataforma de almacenamiento de objetos compatible con S3, como StorageGRID de NetApp o Amazon S3, podrá utilizar "Funcionalidades de NetApp DataOPS Toolkit S3 Data mover". Otro método simple es crear un espacio de trabajo JJupyterLab y cargar archivos a continuación a través de la interfaz web JJupyterLab, como se indica en los pasos 3 a 5 de la sección “Escenario 1 – inferencia bajo demanda en JuppyterLab.”

4. Cree un trabajo de Kubernetes para la tarea de inferencia de lotes. El siguiente ejemplo muestra un trabajo de inferencia en lote para un caso de uso de detección de imagen. Este trabajo realiza la inferencia en cada imagen de un conjunto de imágenes y escribe métricas de precisión de inferencia para su colocación.

```
$ 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. Confirme que el trabajo de inferencia se completó correctamente.

```
$ 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. Agregue la ofuscación de Protopia a su trabajo de inferencia. Puede encontrar instrucciones específicas para casos de uso para agregar la ofuscación Protopia directamente desde Protopia, que está fuera del alcance de este informe técnico. El ejemplo siguiente muestra un trabajo de inferencia por lotes para un caso de uso de detección de cara con ofuscación Protopía agregada mediante un valor ALFA de 0.8. Este trabajo aplica la ofuscación Protopia antes de realizar la inferencia para cada imagen en un conjunto de imágenes y luego escribe las métricas de precisión de inferencia para el stdout.

   Hemos repetido este paso para los valores ALFA 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9 y 0.95. Puede ver los resultados en ""Comparación de precisión de inferencia.""

```
$ 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. Confirme que el trabajo de inferencia se completó correctamente.

```
$ 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
```

Escenario 3: Servidor de inferencia NVIDIA Triton

1. Cree un espacio de nombres de Kubernetes para las cargas de trabajo de inferencia de IA/ML.

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

2. Utilice el kit de herramientas DataOPS de NetApp para aprovisionar un volumen persistente y usarlo como repositorio de modelo para el servidor de inferencia NVIDIA Triton.

```
$ 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. Almacene su modelo en el nuevo volumen persistente en un "formato" Reconocida por el servidor de inferencia NVIDIA Triton.

   Existen varios métodos para cargar datos en un PVC. Un método simple es crear un espacio de trabajo JupyterLab y luego cargar archivos a través de la interfaz web JupyterLab, como se describe en los pasos 3 a 5 en “Escenario 1 – inferencia bajo demanda en JuppyterLab. ”

4. Utilice el kit de herramientas DataOPS de NetApp para poner en marcha una nueva instancia del servidor de inferencia NVIDIA Triton.

```
$ 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. Utilice el SDK del cliente Triton para realizar una tarea de inferencia. El siguiente extracto de código de Python utiliza el SDK del cliente de Triton Python para realizar una tarea de inferencia para un caso de uso de detección facial. En este ejemplo se llama a la API de Triton y se pasa una imagen para la inferencia. A continuación, el servidor de inferencia Triton recibe la solicitud, invoca el modelo y devuelve la salida de inferencia como parte de los resultados de la 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. Agregue la ofuscación Protopia al código de inferencia. Puede encontrar instrucciones específicas para casos de uso para agregar la ofuscación Protopia directamente desde Protopia; sin embargo, este proceso está fuera del alcance de este informe técnico. El ejemplo siguiente muestra el mismo código Python que se muestra en el paso anterior 5, pero con la ofuscación de Protopia agregada.

Tenga en cuenta que la confusión Protopia se aplica a la imagen antes de pasarla a la API de Triton. Así, la imagen no ofuscada nunca sale de la máquina local. Sólo la imagen oculta se pasa a través de la red. Este flujo de trabajo es aplicable para casos de uso en los que los datos se recopilan en una zona de confianza, pero luego debe pasarse fuera de esa zona de confianza para la inferencia. Sin la ocultación de Protopia, no es posible implementar este tipo de flujo de trabajo sin que haya datos confidenciales que salgan de la zona de confianza.

```
# 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)
```

Comparación de precisión de inferencia

Para esta validación, realizamos la inferencia para un caso de uso de detección de imágenes mediante un conjunto de imágenes en bruto. A continuación, realizamos la misma tarea de inferencia en el mismo conjunto de imágenes con la ofuscación Protopía agregada antes de la inferencia. Repetimos la tarea usando diferentes valores DE ALFA para el componente de ofuscación de Protopia. En el contexto de la ofuscación de Protopia, el valor ALFA representa la cantidad de ofuscación que se aplica, con un valor ALFA más alto que representa un nivel más alto de ofuscación. A continuación, comparamos la precisión de la inferencia en estas carreras diferentes.

En las dos tablas siguientes se ofrecen detalles sobre nuestro caso de uso y se resumen los resultados.

Protopia trabaja directamente con los clientes para determinar el valor ALFA adecuado para un caso de uso específico.

Componente	Detalles
Modelo	FaceBoxes (PyTorch) -
Conjunto de datos	Conjunto de datos FDDB

Ofuscación Protopia	ALFA	Precisión
No	N.A.	0.9337148153739079
Sí	0.05	0.9028766627325002
Sí	0.1	0.9024301009661478
Sí	0.2	0.9081836283186224
Sí	0.4	0.9073066107482036
Sí	0.6	0.8847816568680239
Sí	0.8	0.8841195749171925
Sí	0.9	0.8455427675252052
Sí	0.95	0.8455427675252052

Velocidad de ofuscación

Para esta validación, aplicamos la ofuscación de Protopia a una imagen de 1920 x 1080 píxeles cinco veces y medimos la cantidad de tiempo que tardó en completar el paso de ofuscación cada vez.

Utilizamos PyTorch ejecutándose en una única GPU NVIDIA V100 para aplicar la ofuscación y borramos la memoria caché de la GPU entre ejecuciones. El paso de ofuscación tomó 5,47 ms, 5.27 ms, 4,54 ms, 5.24 ms y 4,84 ms respectivamente para completar las cinco carreras. La velocidad media era de 5,072 ms.

Conclusión

Los datos existen en tres estados: En reposo, en tránsito y en computación. Una parte importante de cualquier servicio de inferencia de IA debe ser la protección de datos contra amenazas durante todo el proceso. La protección de datos durante la inferencia es vital, ya que el proceso puede exponer información privada tanto sobre clientes externos como sobre la empresa que proporciona el servicio de inferencia. Protopia AI es una solución no obstrusiva de solo software para la inferencia de IA confidencial en el mercado actual. Con Protopia, la IA solo recibe la información transformada de los registros de datos que es esencial para llevar a cabo la tarea de IA/ML disponible y nada más. Esta transformación estocástica no es una forma de enmascarar y se basa en cambiar matemáticamente la representación de los datos utilizando el ruido curado.

Los sistemas de almacenamiento de NetApp con funcionalidades de ONTAP ofrecen el mismo rendimiento o

incluso mejor que el almacenamiento SSD local y, combinados con el kit de herramientas DataOPS de NetApp, ofrecen las siguientes ventajas a científicos de datos, ingenieros de datos, desarrolladores DE IA/ML y responsables DE la toma DE decisiones TECNOLÓGICAS empresariales o de negocio:

- Uso compartido de datos sin esfuerzo entre sistemas de IA, análisis y otros sistemas de negocio cruciales. Este uso compartido de datos reduce la sobrecarga de la infraestructura, mejora el rendimiento y optimiza la gestión de datos en toda la empresa.
- Almacenamiento y cálculo escalables de forma independiente para minimizar los costes y mejorar el uso de recursos.
- Flujos de trabajo de desarrollo y puesta en marcha optimizados que utilizan copias Snapshot integradas y clones para espacios de trabajo de usuario instantáneos con gestión eficiente del espacio, control de versiones integrado y una puesta en marcha automatizada.
- Protección de datos de clase empresarial y regulación de datos para la recuperación ante desastres, la continuidad del negocio y los requisitos normativos.
- Invocación simplificada de las operaciones de gestión de datos; realice rápidamente copias Snapshot de los espacios de trabajo de científicos de datos para realizar backups y trazabilidad desde el kit de herramientas de DataOps de NetApp en los portátiles Jupyter.

La solución de NetApp y Protopía proporciona una arquitectura flexible de escalado horizontal ideal para puestas en marcha de inferencia de IA de clase empresarial. Permite la protección de datos y proporciona privacidad para información confidencial en la que los requisitos confidenciales de inferencia de IA pueden satisfacerse con prácticas responsables de IA tanto en puestas en marcha en el entorno local como en cloud híbrido.

Dónde encontrar información adicional y reconocimientos

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

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

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

- Almacenamiento persistente de NetApp para contenedores: Trident de NetApp

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

- Kit de herramientas de operaciones de datos de NetApp

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

- Almacenamiento persistente de NetApp para contenedores: Astra Trident de NetApp

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

- Protopía AI: Inferencia confidencial

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

- Copia y sincronización de NetApp BlueXP

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

- Servidor de inferencia NVIDIA Triton

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

- Documentación de NVIDIA Triton inferencias Server

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

- FaceBoxes en PyTorch

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

Reconocimientos

- Mark Cates, Director de producto, NetApp
- Sufian Ahmad, ingeniero técnico de marketing, NetApp
- Hadi Esmaeilzadeh, director técnico y profesor de Protopía AI

Análisis de la confianza con IA de NetApp

TR-4910: Análisis de sentimiento en las comunicaciones de los clientes con IA de NetApp

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

Este informe técnico proporciona directrices de diseño para que los clientes realicen análisis de opinión en un centro de soporte global de nivel empresarial usando las tecnologías de gestión de datos de NetApp con un marco de software de NVIDIA mediante la transferencia de aprendizaje e IA conversacional. Esta solución es aplicable a cualquier sector que desee obtener información de los clientes a partir de archivos de voz o texto grabados que representen registros de chat, correos electrónicos y otras comunicaciones de texto o audio. Hemos implementado una canalización integral para demostrar el reconocimiento automático de voz, el análisis de sensibilidad en tiempo real y el modelo de procesamiento de lenguaje natural y aprendizaje profundo: Funcionalidades de nueva formación en un clúster informático acelerado por GPU con almacenamiento all-flash conectado al cloud de NetApp. Se pueden formar y optimizar modelos lingüísticos masivos de vanguardia para realizar inferencia rápidamente con el centro de soporte global y crear una experiencia de cliente excepcional y evaluaciones de rendimiento de empleados objetivas y a largo plazo.

El análisis del sentimiento es un campo de estudio dentro del procesamiento del lenguaje natural (NLP) por el cual se extraen sentimientos positivos, negativos o neutrales del texto. Los sistemas de IA conversacionales han aumentado hasta alcanzar un nivel de integración casi global a medida que cada vez más personas acuden para interactuar con ellos. El análisis de confianza tiene una variedad de casos de uso, desde determinar el rendimiento de los empleados del centro de soporte en conversaciones con los llamantes y proporcionar respuestas automatizadas adecuadas al bot conversacional para predecir el precio de las acciones de una empresa basándose en las interacciones entre representantes de la empresa y la audiencia en llamadas trimestrales a ganancias. Además, el análisis de confianza puede utilizarse para determinar la visión del cliente sobre los productos, servicios o asistencia proporcionados por la Marca.

Esta solución integral utiliza modelos NLP para realizar un análisis de confianza de alto nivel que permita marcos analíticos del centro de soporte. Las grabaciones de audio se procesan en texto escrito y el sentimiento se extrae de cada frase de la conversación. Los resultados, agregados en un panel, se pueden elaborar para analizar los sentimientos de la conversación, tanto históricos como en tiempo real. Esta solución se puede generalizar a otras soluciones con modalidades de datos y necesidades de producción similares. Con los datos adecuados se pueden llevar a cabo otros casos de uso. Por ejemplo, las llamadas de beneficios de la empresa se pueden analizar para determinar la opinión utilizando la misma canalización de extremo a extremo. Otras formas de análisis de NLP, como el modelado de temas y el reconocimiento de entidades con nombre (TNM), también son posibles debido a la naturaleza flexible del ducto.

Estas implementaciones de IA fueron posibles gracias a NVIDIA RIVA, el kit de herramientas TAO de NVIDIA y el kit de herramientas DataOPS de NetApp trabajando juntos. Las herramientas de NVIDIA se utilizan para poner en marcha rápidamente soluciones de IA de gran rendimiento utilizando modelos y canalizaciones prediseñados. El kit de herramientas DataOPS de NetApp simplifica diversas tareas de gestión de datos para acelerar el desarrollo.

Valor para el cliente

Las empresas ven el valor de una herramienta de evaluación de empleados y de reacción del cliente para la conversación de texto, audio y vídeo para el análisis de opiniones. Los gerentes se benefician de la información presentada en el panel, permitiendo una evaluación de los empleados y la satisfacción del cliente basada en ambos lados de la conversación.

Además, el kit de herramientas DataOps de NetApp gestiona el versionado y la asignación de datos dentro de la infraestructura del cliente. Esto lleva a frecuentes actualizaciones de los análisis presentados en el panel de control sin crear costes rígidos de almacenamiento de datos.

Casos de uso

Debido al número de llamadas que procesa estos centros de soporte, la evaluación del rendimiento de la llamada puede llevar un tiempo considerable si se realiza manualmente. Los métodos tradicionales, como el conteo de bolsas de palabras y otros métodos, pueden lograr cierta automatización, pero estos métodos no capturan aspectos más matizados y un contexto semántico del lenguaje dinámico. Las técnicas de modelado de IA se pueden usar para realizar algunos de estos análisis más matices de forma automatizada. Además, con lo último en tecnología y herramientas de modelado preformadas publicadas por NVIDIA, AWS, Google y otras, puede ponerse en marcha una canalización integral con modelos complejos con una facilidad relativa personalizada.

Una canalización de extremo a extremo para el análisis de confianza del centro de soporte procesa archivos de audio en tiempo real a medida que los empleados converse con las personas que llaman. A continuación, estos archivos de audio se procesan para su uso en el componente voz a texto que los convierte en un formato de texto. Cada frase de la conversación recibe una etiqueta que indica el sentimiento (positivo, negativo o neutro).

El análisis de confianza puede proporcionar un aspecto esencial de las conversaciones para evaluar el rendimiento de la llamada. Estos sentimientos añaden un nivel adicional de profundidad a las interacciones entre empleados y personas que llaman. El panel de control de confianza asistido por IA proporciona a los administradores un seguimiento en tiempo real de la opinión dentro de una conversación, junto con un análisis retrospectivo de las llamadas pasadas del empleado.

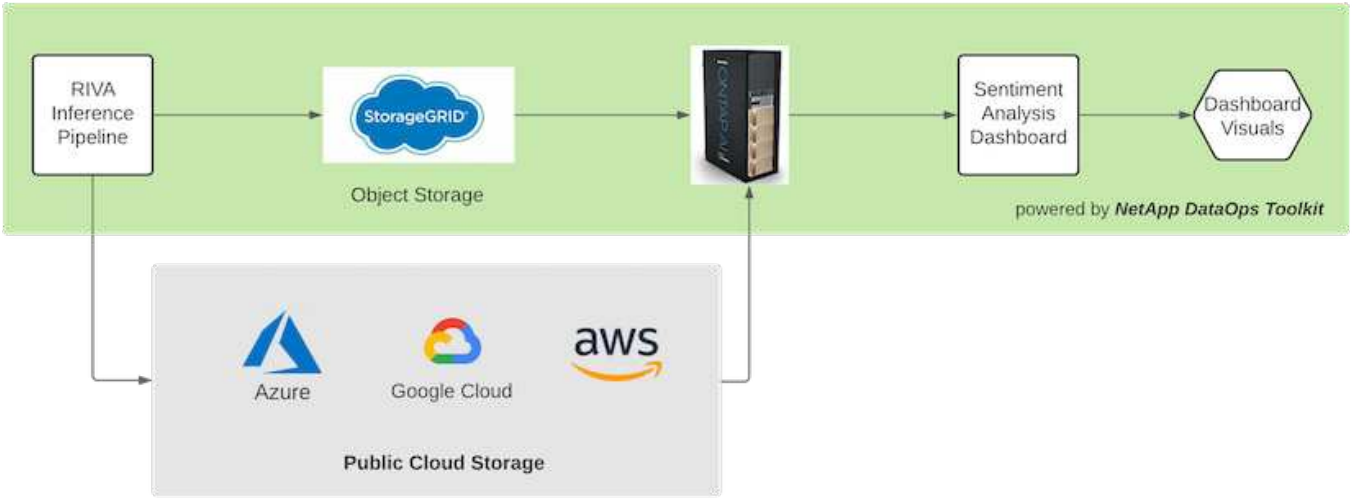
Existen herramientas prediseñadas que se pueden combinar de manera potente para crear rápidamente una canalización de IA completa para solucionar este problema. En este caso, la biblioteca NVIDIA RIVA se puede utilizar para realizar las dos tareas en serie: Transcripción de audio y análisis de sentimiento. El primero es un algoritmo de procesamiento de señales de aprendizaje supervisado y el segundo es un algoritmo de clasificación NLP de aprendizaje supervisado. Estos algoritmos listos para usar pueden ajustarse para cualquier caso de uso relevante con datos relevantes del negocio mediante el kit de herramientas TAO de NVIDIA. Esto lleva a que se estén creando soluciones más precisas y potentes por una fracción del coste y los recursos. Los clientes pueden incorporar la "Maxine de NVIDIA" Marco de trabajo para aplicaciones de videoconferencia aceleradas por GPU en el diseño del centro de soporte.

Los siguientes casos de uso son el núcleo de esta solución. En ambos casos de uso, se utiliza TAO Toolkit para realizar un ajuste preciso del modelo y RIVA para su implementación.

- Voz a texto
- Análisis de la confianza

Para analizar las interacciones del centro de soporte entre empleados y clientes, cada conversación del cliente en forma de llamadas de audio se puede realizar a través de la canalización para extraer sentimientos a nivel de frase. Esos sentimientos pueden ser verificados por un ser humano para justificar los sentimientos o ajustarlos según sea necesario. A continuación, los datos etiquetados se pasan al paso de ajuste fino para mejorar las predicciones de sentimientos. Si ya existen datos de confianza etiquetados, se puede acelerar el ajuste preciso del modelo. En cualquier caso, el ducto es generalizable a otras soluciones que requieren la ingestión de audio y la clasificación de frases.

Los resultados de opiniones de la IA se cargan en una base de datos cloud externa o en un sistema de almacenamiento gestionado por la empresa. Los resultados de la confianza se transfieren desde esta base de datos más grande al almacenamiento local para su uso dentro del panel de control que muestra el análisis de opinión de los administradores. La funcionalidad principal del panel es la de interactuar con el empleado del servicio de atención al cliente en tiempo real. Los gerentes pueden evaluar y proporcionar comentarios a los empleados durante sus llamadas con actualizaciones en vivo de la opinión de cada frase, así como una revisión histórica del desempeño pasado del empleado o de las reacciones del cliente.



La "Kit de herramientas de operaciones de datos de NetApp" Puede seguir gestionando sistemas de almacenamiento de datos incluso después de que la canalización de inferencia RIVA genere etiquetas de sentimiento. Estos resultados de la IA pueden cargarse en un sistema de almacenamiento de datos gestionado por el kit de herramientas DataOPS de NetApp. Los sistemas de almacenamiento de datos deben ser capaces de gestionar cientos de inserciones y seleccionar cada minuto. El sistema de almacenamiento del dispositivo local consulta el mayor almacenamiento de datos en tiempo real para su extracción. También es posible consultar a la instancia de almacenamiento de datos de mayor tamaño los datos históricos para mejorar aún más la experiencia de la consola. El kit de herramientas DataOPS de NetApp facilita ambos usos mediante el clonado rápido de los datos y la distribución entre todas las consolas que los utilizan.

Público objetivo

El público objetivo de la solución incluye los siguientes grupos:

- Gerentes de empleados
- Ingenieros/científicos de datos
- Administradores DE TECNOLOGÍA (en las instalaciones, en cloud o híbrida)

El seguimiento de los sentimientos a lo largo de las conversaciones es una herramienta valiosa para evaluar el rendimiento de los empleados. Mediante el panel de instrumentos de inteligencia artificial, los administradores pueden ver cómo los empleados y las personas que llaman cambian sus sentimientos en tiempo real, lo que permite evaluaciones en directo y sesiones de orientación. Además, las empresas pueden obtener información valiosa de los clientes que participan en conversaciones vocales, bots de chat de texto y videoconferencias. Dichos análisis de clientes utilizan las funcionalidades de procesamiento multimodal a escala con modernos modelos de IA y flujos de trabajo de vanguardia.

En cuanto a los datos, el centro de soporte procesa diariamente un gran número de archivos de audio. El kit de herramientas DataOPS de NetApp facilita esta tarea de gestión de datos tanto para la sintonización periódica de modelos como para consolas de análisis de opiniones.

Los administradores DE TECNOLOGÍA también se benefician del kit de herramientas de NetApp DataOps ya que permite mover datos con rapidez entre entornos de implementación y de producción. Los entornos y servidores de NVIDIA también deben gestionarse y distribuirse para permitir la inferencia en tiempo real.

Arquitectura

La arquitectura de esta solución de centro de soporte se basa en las herramientas prediseñadas de NVIDIA y el kit de herramientas Data OPS de NetApp. Las herramientas de NVIDIA se utilizan para poner en marcha rápidamente soluciones de IA de alto rendimiento utilizando modelos y canalizaciones prediseñados. El kit de herramientas DataOPS de NetApp simplifica diversas tareas de gestión de datos para acelerar el desarrollo.

Tecnología de soluciones

"RIVA DE NVIDIA" Es un SDK acelerado por GPU para crear aplicaciones de IA conversacionales multimodales que ofrecen rendimiento en tiempo real en las GPU. El kit de herramientas NVIDIA Train, Adapt y Optimize (TAO) proporciona un método más rápido y sencillo para acelerar la formación y crear rápidamente modelos de IA específicos para dominios con un alto rendimiento y precisión.

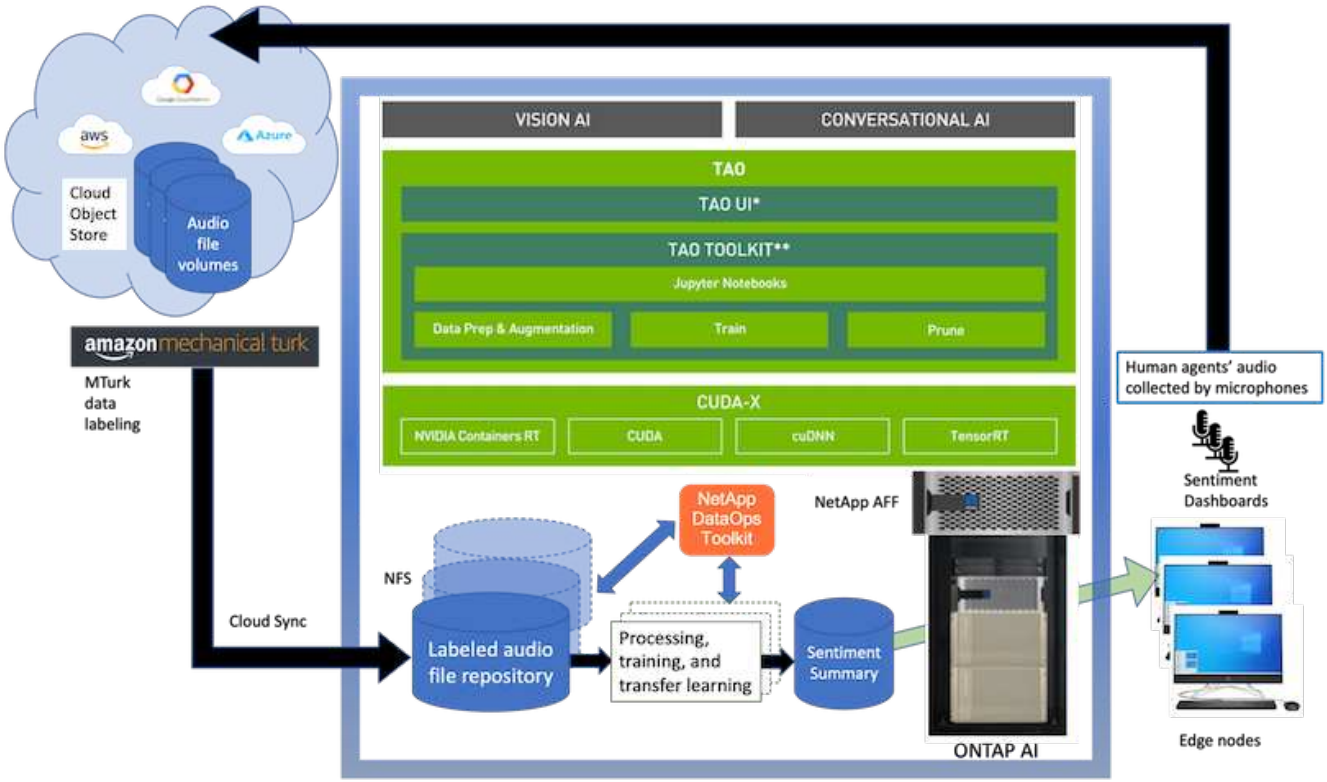
El kit de herramientas NetApp DataOPS es una biblioteca Python que facilita a los desarrolladores, científicos de datos, ingenieros de DevOps e ingenieros de datos la tarea de realizar varias tareas de gestión de datos. Esto incluye el aprovisionamiento casi instantáneo de un nuevo volumen de datos o espacio de trabajo JupyterLab, el clonado casi instantáneo de un volumen de datos o un espacio de trabajo JupyterLab y la creación casi instantánea de copias Snapshot de un volumen de datos o de un espacio de trabajo JupyterLab para su seguimiento y línea de base.

Diagrama arquitectónico

El siguiente diagrama muestra la arquitectura de la solución. Existen tres categorías de entorno principales: El cloud, el núcleo y el perímetro. Cada una de las categorías puede estar geográficamente dispersa. Por ejemplo, la nube contiene almacenes de objetos con archivos de audio en bloques de diferentes regiones, mientras que el núcleo puede contener centros de datos vinculados a través de una red de alta velocidad o Copia y sincronización de NetApp BlueXP. Los nodos EDGE denotan las plataformas de trabajo diario del agente humano individual, en las que hay disponibles micrófonos y herramientas de panel interactivas para visualizar la opinión y recopilar datos de audio de conversaciones con los clientes.

En los centros de datos acelerados por GPU, las empresas pueden utilizar NVIDIA "RIVA" Marco para crear aplicaciones de IA conversacionales, con las que el "Kit de herramientas Tao" Se conecta para la finetunización de modelos y el reciclaje mediante técnicas de aprendizaje en L de transferencia. Estas aplicaciones y flujos de trabajo de computación funcionan con la "Kit de herramientas de operaciones de datos de NetApp", Activación de las mejores capacidades de administración de datos que ONTAP tiene para ofrecer. El kit de herramientas permite a los equipos de datos corporativos realizar rápidamente prototipos de sus modelos con datos estructurados y no estructurados asociados a través de copias Snapshot y clones para llevar a cabo seguimientos, crear versiones, realizar pruebas A/B, proporcionando así seguridad, gobernabilidad, y cumplimiento de las normativas. Consulte la sección "Diseño del almacenamiento" para obtener más detalles.

Esta solución muestra los pasos detallados del procesamiento de archivos de audio, el entrenamiento de modelos NLP, el aprendizaje de transferencias y la administración de datos. La canalización integral resultante genera un resumen de sentimiento que se muestra en tiempo real en los paneles de control de los agentes de apoyo humano.



Requisitos de hardware

En la siguiente tabla se enumeran los componentes de hardware necesarios para implementar la solución. Los componentes de hardware que se usan en cualquier implementación particular de la solución pueden variar en función de las necesidades del cliente.

Pruebas de latencia de respuesta	Tiempo (milisegundos)
Procesamiento de datos	10
Inferencia	10

Estas pruebas de tiempo de respuesta se realizaron en más de 50,000 archivos de audio en 560 conversaciones. Cada archivo de audio tenía un tamaño de ~100 KB como MP3 y de ~1 MB cuando se convirtió a WAV. El paso de procesamiento de datos convierte MP3s en archivos WAV. Los pasos de inferencia convierten los archivos de audio en texto y extraen un sentimiento del texto. Estos pasos son todos independientes entre sí y pueden ser paralelizados para acelerar el proceso.

Teniendo en cuenta la latencia de transferencia de datos entre almacenes, los administradores deben poder ver actualizaciones del análisis de opinión en tiempo real en un segundo del final de la frase.

Hardware NVIDIA RIVA

Hardware subyacente	Requisitos
SO	Linux x86_64
Memoria GPU (ASR)	Modelos de streaming: Aprox. 5600 MB modelos sin secuencias: Aprox. 3100 MB
Memoria GPU (NLP)	~500 MB por modelo BERT

Hardware del kit de herramientas TAO de NVIDIA

Hardware subyacente	Requisitos
RAM del sistema	32 GB
RAM DE GPU	32 GB
CPU	8 núcleos
GPU	NVIDIA (A100, V100 y RTX 30x0)
SSD	100 GB

Sistema de almacenamiento flash

ONTAP 9 de NetApp

ONTAP 9.9, la última generación del software de gestión del almacenamiento de NetApp, permite a las empresas modernizar su infraestructura y realizar la transición a un centro de datos preparado para el cloud. ONTAP ofrece las mejores capacidades de gestión de datos y permite la gestión y protección de los datos con un solo conjunto de herramientas, sin importar dónde residan. También puede mover los datos libremente a donde sea necesario: El perímetro, el núcleo o el cloud. ONTAP 9.9 incluye numerosas funciones que simplifican la gestión de datos, aceleran y protegen los datos esenciales y permiten disfrutar de funcionalidades de infraestructura de nueva generación en arquitecturas de cloud híbrido.

Copia y sincronización de NetApp BlueXP

"Copia y sincronización de BlueXP" Es un servicio de NetApp que ofrece una sincronización de datos rápida y segura que permite transferir archivos entre recursos compartidos de archivos NFS o SMB en las instalaciones a cualquiera de los siguientes destinos:

- StorageGRID de NetApp
- ONTAP S3 de NetApp
- Cloud Volumes Service de NetApp
- Azure NetApp Files
- Simple Storage Service (Amazon S3)
- Sistema de archivos Elastic de Amazon (Amazon EFS)
- Azure Blob
- Google Cloud Storage
- Almacenamiento de objetos en cloud de IBM

BlueXP Copy and Sync mueve los archivos a donde los necesitas de una forma rápida y segura. Una vez transferidos los datos, estarán completamente disponibles para su uso tanto en el origen como en el destino. BlueXP Copy and Sync sincroniza los datos de forma continua en función de tu programación predefinida, moviendo solo los deltas, por lo que se reducen al mínimo el tiempo y el dinero que se invierten en la replicación de datos. BlueXP Copy and Sync es una herramienta de software como servicio (SaaS) fácil de configurar y usar. Las transferencias de datos activadas por BlueXP Copy and Sync se llevan a cabo por agentes de datos. Puedes poner en marcha agentes de datos de BlueXP Copy y Sync en AWS, Azure, Google Cloud Platform o en las instalaciones.

StorageGRID de NetApp

La suite de almacenamiento de objetos definida por software StorageGRID admite una amplia gama de casos de uso en entornos multicloud públicos, privados e híbridos sin problemas. Con innovaciones líderes del sector, StorageGRID de NetApp almacena, protege y preserva datos no estructurados para usos múltiples, incluida la gestión automatizada del ciclo de vida durante largos periodos de tiempo. Para obtener más información, consulte "StorageGRID de NetApp" sitio.

Requisitos de software

En la siguiente tabla se enumeran los componentes de software necesarios para implementar esta solución. Los componentes que se usan en cualquier implementación particular de la solución pueden variar en función de las necesidades del cliente.

Máquina host	Requisitos
RIVA (anteriormente JARVIS)	1.4.0
TAO Toolkit (antes Transfer Learning Toolkit)	3.0
ONTAP	9.9.1
SO DGX	5.1
DOTK	2.0.0

Software NVIDIA RIVA

De NetApp	Requisitos
Docker	>19.02 (con nvidia-docker instalado)>=19.03 si no se utiliza DGX
Controlador NVIDIA	Más de 465.19.01 418.40+, 440.33+, 450.51+ y 460.27+ para las GPU del centro de datos
So del contenedor	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
Servidor de inferencia Triton	2.9.0

Software NVIDIA TAO Toolkit

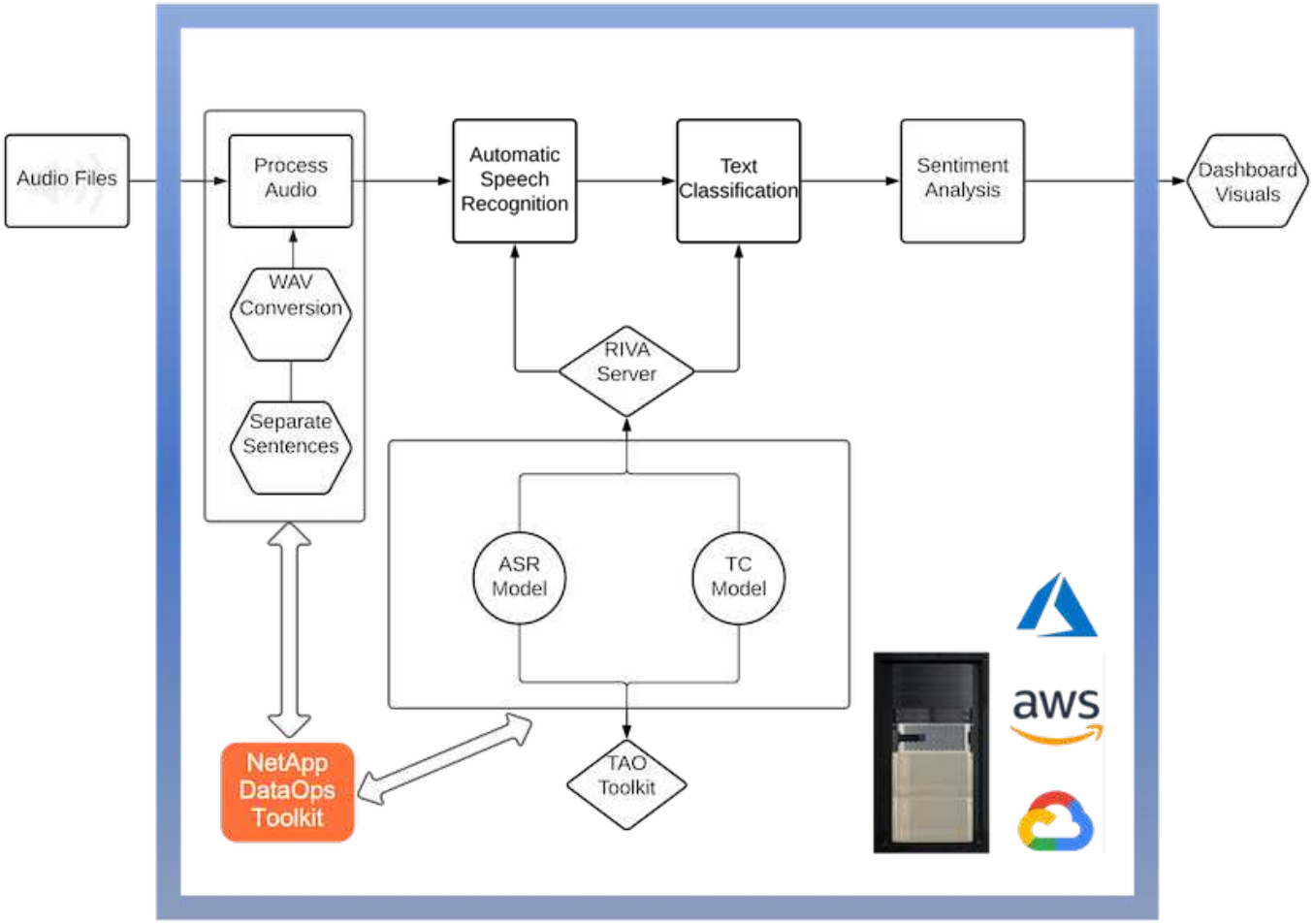
De NetApp	Requisitos
Sistema operativo Ubuntu 18.04 LTS	18.04
python	>=3.6.9
docker-ce	>19.03.5

De NetApp	Requisitos
docker-API	1.40
kit de herramientas de nvidia-container	>1.3.0-1
nvidia-container-runtime	3.4.0-1
nvidia-docker2	2.5.0-1
controlador nvidia	>455
python-pip	>21.06
nvidia-pyindex	Última versión

Detalles de casos de uso

Esta solución se aplica a los siguientes casos de uso:

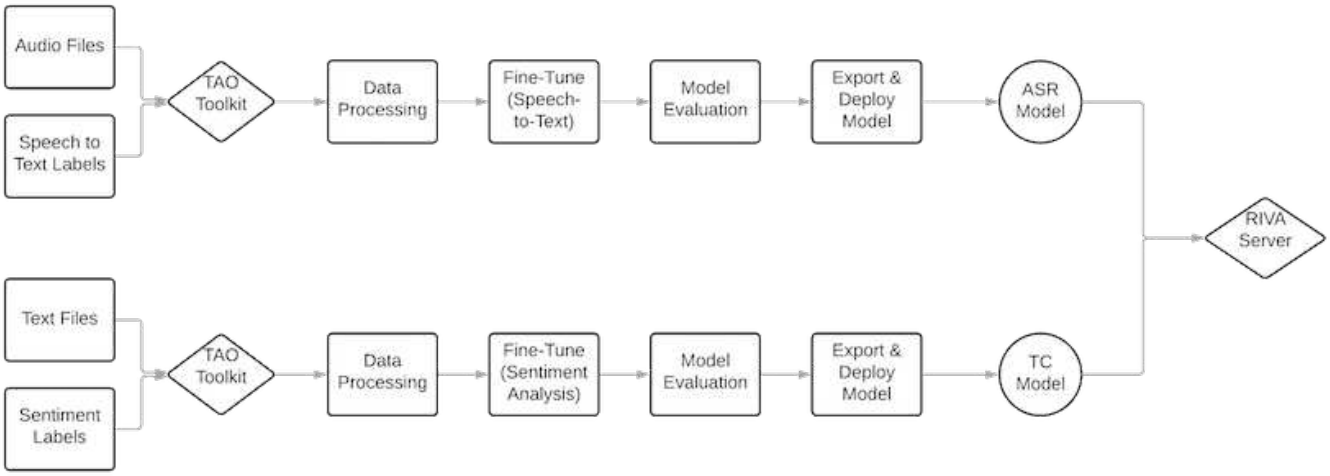
- Voz a texto
- Análisis de la confianza



El caso práctico de voz a texto comienza ingiriendo archivos de audio para los centros de soporte. Este audio se procesa entonces para ajustarse a la estructura requerida por RIVA. Si los archivos de audio aún no se han dividido en sus unidades de análisis, esto se debe hacer antes de pasar el audio a RIVA. Una vez procesado el archivo de audio, se pasa al servidor RIVA como una llamada API. El servidor emplea uno de los muchos

modelos que aloja y devuelve una respuesta. Este mensaje de voz a texto (parte del reconocimiento automático de voz) devuelve una representación de texto del audio. A partir de ahí, la canalización pasa a la parte del análisis de confianza.

Para el análisis de confianza, la salida de texto del reconocimiento automático de voz sirve como entrada a la clasificación de texto. Text Classification es el componente NVIDIA para clasificar el texto en cualquier número de categorías. Las categorías de sentimiento varían de positivo a negativo para las conversaciones del centro de apoyo. El rendimiento de los modelos se puede evaluar utilizando un conjunto de holdout para determinar el éxito del paso de ajuste fino.



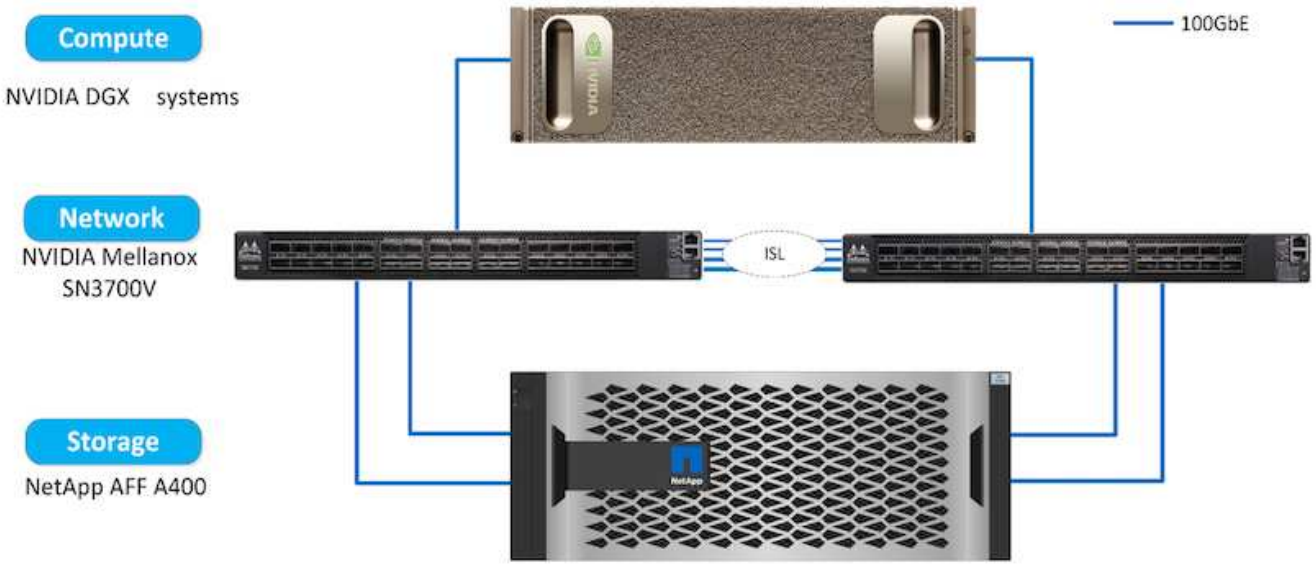
Una canalización similar se utiliza tanto para el análisis de voz a texto como para el análisis de sentimiento dentro del TAO Toolkit. La principal diferencia es el uso de etiquetas que se requieren para la afinación de los modelos. La canalización del kit de herramientas TAO comienza con el procesamiento de los archivos de datos. A continuación, los modelos preformados (procedentes de "Catálogo de NVIDIA NGC") se ajustan con precisión mediante los datos del centro de soporte. Los modelos ajustados con precisión se evalúan en función de sus mediciones de rendimiento correspondientes y, si tienen un rendimiento superior al de los modelos preformados, se implementan en EL servidor RIVA.

Consideraciones de diseño

En esta sección se describen las consideraciones de diseño para los distintos componentes de esta solución.

Diseño informático y de red

En función de las restricciones en la seguridad de los datos, todos los datos deben permanecer dentro de la infraestructura del cliente o en un entorno seguro.



Diseño del almacenamiento

El kit de herramientas Data OPS de NetApp sirve como servicio principal para gestionar los sistemas de almacenamiento. El kit de herramientas DataOps es una biblioteca Python que facilita a los desarrolladores, científicos de datos, ingenieros de DevOps e ingenieros de datos la tarea de gestión de datos realizar varias tareas de gestión de datos, como el aprovisionamiento casi instantáneo de un nuevo volumen de datos o un espacio de trabajo JJupyterLab, el clonado casi instantáneo de un volumen de datos o el espacio de trabajo JuppyterLab. Y una copia de Snapshot casi instantánea de un volumen de datos o un espacio de trabajo JupyterLab para poder seguir su seguimiento o crear una línea de base. Esta biblioteca de Python puede funcionar como una utilidad de línea de comandos o una biblioteca de funciones que se pueden importar a cualquier programa de Python o a cualquier ordenador portátil Jupyter.

Mejores prácticas de RIVA

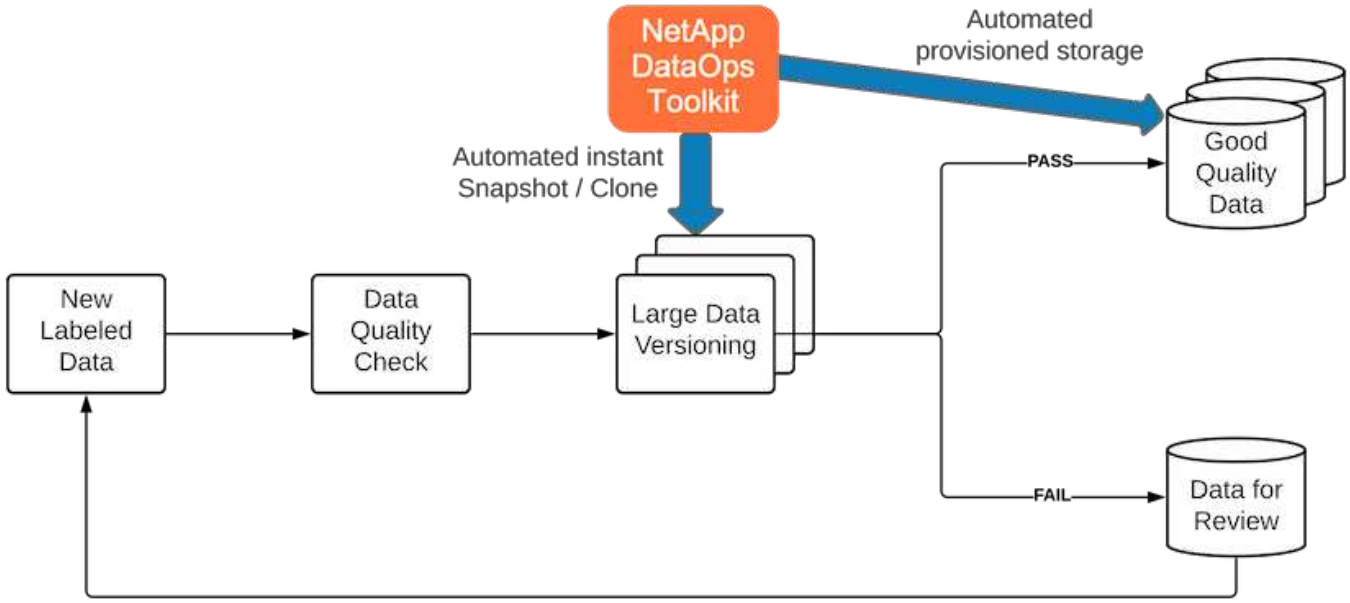
NVIDIA proporciona varias funciones generales "mejores prácticas de datos" Para utilizar RIVA:

- **Utilice formatos de audio sin pérdidas si es posible.** el uso de códecs con pérdida como MP3 puede reducir la calidad.
- **Aumentar los datos de entrenamiento.** la adición de ruido de fondo a los datos de entrenamiento de audio puede inicialmente disminuir la precisión y, sin embargo, aumentar la solidez.
- **Limite el tamaño del vocabulario si utiliza texto raspado.** muchas fuentes en línea contienen tipopos o pronombres auxiliares y palabras poco comunes. La eliminación de estos elementos puede mejorar el modelo de idioma.
- **Utilice una frecuencia de muestreo mínima de 16 kHz si es posible.** sin embargo, trate de no resampling, ya que al hacerlo se reduce la calidad de audio.

Además de estas mejores prácticas, los clientes deben priorizar la recopilación de un conjunto de datos de ejemplo representativo con etiquetas precisas para cada paso de la canalización. En otras palabras, el conjunto de datos de ejemplo debería reflejar proporcionalmente las características especificadas ejemplificadas en un conjunto de datos de destino. De forma similar, los anotadores del conjunto de datos tienen la responsabilidad de equilibrar la precisión y la velocidad del etiquetado, de modo que se maximice la calidad y la cantidad de los datos. Por ejemplo, esta solución de centro de soporte requiere archivos de audio, etiquetas de texto etiquetadas y etiquetas de sentimiento. La naturaleza secuencial de esta solución significa que los errores desde el principio de la canalización se propagan hasta el final Si los archivos de audio son de

mala calidad, las transcripciones de texto y las etiquetas de sentimiento también serán.

Esta propagación de errores también se aplica a los modelos entrenados en estos datos. Si las predicciones de sentimiento son 100% exactas pero el modelo de voz a texto tiene un rendimiento deficiente, entonces la canalización final está limitada por las transcripciones iniciales de audio a texto. Es esencial que los desarrolladores consideren el rendimiento de cada modelo individualmente y como un componente de una canalización mayor. En este caso en particular, el objetivo final es desarrollar una canalización que pueda predecir con precisión el sentimiento. Por lo tanto, la métrica general sobre la cual evaluar la canalización es la precisión de los sentimientos, que afecta directamente la transcripción del habla al texto.



El kit de herramientas DataOPS de NetApp complementa la canalización de comprobación de la calidad de los datos mediante su tecnología de clonado de datos casi instantánea. Cada archivo etiquetado debe evaluarse y compararse con los archivos etiquetados existentes. La distribución de estas comprobaciones de calidad a través de distintos sistemas de almacenamiento de datos garantiza una ejecución rápida y eficaz de estas comprobaciones.

Implementar el análisis de confianza del centro de soporte

La implementación de la solución consta de los siguientes componentes:

1. Kit de herramientas de operaciones de datos de NetApp
2. Configuración de NGC
3. Servidor NVIDIA RIVA
4. Kit de herramientas TAO de NVIDIA
5. Exportar modelos TAO a RIVA

Para realizar la implementación, lleve a cabo los siguientes pasos:

Kit de herramientas Data OPS de NetApp: Compatibilidad con el análisis de confianza del centro de soporte

Para utilizar la "Kit de herramientas de operaciones de datos de NetApp", lleve a cabo los siguientes pasos:

1. Instalación del kit de herramientas de PIP.

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

2. Configurar la gestión de datos

```
netapp_dataops_cli.py config
```

Configuración de NGC: Respaldar el análisis de confianza del centro

Para configurar "NVIDIA NGC", lleve a cabo los siguientes pasos:

1. Descargar el NGC.

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

2. Agregue su directorio actual a la ruta de acceso.

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

3. Debe configurar la CLI de NGC para su uso con el fin de poder ejecutar los comandos. Introduzca el siguiente comando, incluida su clave de API cuando se le solicite.

```
ngc config set
```

Para sistemas operativos que no están basados en Linux, visite "aquí".

NVIDIA RIVA Server: Análisis de opinión del centro de soporte

Para configurar "RIVA DE NVIDIA", lleve a cabo los siguientes pasos:

1. Descargar los archivos RIVA de NGC.

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

2. Inicialice LA configuración DE RIVA (`riva_init.sh`).
3. Inicie EL servidor RIVA (`riva_start.sh`).

4. Inicie EL cliente RIVA (`riva_start_client.sh`).
5. En EL cliente RIVA, instale la biblioteca de procesamiento de audio ("FFMPEG")

```
apt-get install ffmpeg
```

6. Inicie el "Juppyter" servidor.
7. Ejecute el portátil de canalización de inferencia DE RIVA.

Kit de herramientas TAO de NVIDIA: Análisis de opinión del centro de soporte

Para configurar NVIDIA TAO Toolkit, lleve a cabo los siguientes pasos:

1. Prepare y active una "entorno virtual" Para TAO Toolkit.
2. Instale el "paquetes requeridos".
3. Tire manualmente de la imagen utilizada durante el entrenamiento y ajuste preciso.

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

4. Inicie el "Juppyter" servidor.
5. Ejecute el cuaderno TAO de afinación fina.

Exportar modelos TAO a RIVA: Apoyar el análisis de confianza del centro

Para usar "Modelos TAO Toolkit EN RIVA", lleve a cabo los siguientes pasos:

1. Guarde los modelos en el cuaderno TAO de sintonización fina.
2. Copie los modelos TAO entrenados en el directorio del modelo RIVA.
3. Inicie EL servidor RIVA (`riva_start.sh`).

Obstáculos para la implementación

Estas son algunas cosas que debe tener en cuenta a medida que desarrolla su propia solución:

- El kit de herramientas Data OPS de NetApp se instala primero para garantizar que el sistema de almacenamiento de datos se ejecute de forma óptima.
- NVIDIA NGC debe instalarse antes de cualquier otra cosa porque autentica la descarga de imágenes y modelos.
- RIVA se debe instalar antes que TAO Toolkit. LA instalación DE RIVA configura el demonio docker para extraer imágenes según sea necesario.
- El DGX y el docker deben tener acceso a Internet para descargar los modelos.

Resultados de validación

Tal como se ha mencionado en la sección anterior, los errores se propagan por la canalización siempre que haya dos o más modelos de aprendizaje automático

ejecutándose de forma secuencial. Para esta solución, el sentimiento de la frase es el factor más importante a la hora de medir el nivel de riesgo de las acciones de la empresa. El modelo de voz a texto, aunque esencial para la canalización, sirve como unidad de preprocesamiento antes de que se puedan predecir los sentimientos. Lo que realmente importa es la diferencia en el sentimiento entre las frases de la verdad en el terreno y las oraciones predichas. Esto sirve como proxy para la velocidad de error de la palabra (WER). La precisión de voz a texto es importante, pero el WER no se utiliza directamente en la métrica de canalización final.

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

Estas métricas de sentimiento se pueden calcular para la puntuación F1, la recuperación y la precisión de cada frase. Los resultados se pueden agregar y mostrar dentro de una matriz de confusión, junto con los intervalos de confianza de cada métrica.

La ventaja de usar el aprendizaje de transferencia es un aumento en el rendimiento de los modelos por una fracción de los requisitos de datos, el tiempo de entrenamiento y el coste. Los modelos ajustados también deben compararse con sus versiones de referencia para garantizar que el aprendizaje de la transferencia mejore el rendimiento en lugar de deteriorarlo. En otras palabras, el modelo ajustado debería tener un mejor rendimiento en los datos del centro de soporte que el modelo preentrenado.

Evaluación de la canalización

Caso de prueba	Detalles
Número de prueba	Métrica de sentimiento de canalización
Probar los requisitos previos	Modelos ajustados para modelos de análisis de voz a texto y de sentimiento
Resultado esperado	La métrica de sentimiento del modelo ajustado funciona mejor que el modelo preentrenado original.

Métrica de sentimiento de canalización

1. Calcule la métrica de sentimiento para el modelo de referencia.
2. Calcule la métrica de sentimiento para el modelo ajustado.
3. Calcular la diferencia entre estas métricas.
4. Calcule la media de las diferencias entre todas las frases.

Vídeos y demostraciones

Existen dos portátiles que contienen la canalización de análisis de confianza: ""“Support-Center-Model-Transfer-Learning-and-Fine-Tuning.ipynb”" y.. ""“Support-Center-sentimiento-Analysis-Pipeline.ipynb”". Juntos, estos portátiles muestran cómo desarrollar una canalización para la incorporación de datos de centro de soporte y extraer sentimientos de cada frase mediante modelos de aprendizaje profundo de última generación ajustados a los datos del usuario.

Centro de apoyo - Análisis de confianza Pipeline.ipynb

Este cuaderno contiene la canalización RIVA de inferencia para la incorporación de audio, la conversión a texto y la extracción de sentimientos para su uso en un panel externo. El conjunto de datos se descarga y procesa automáticamente si aún no se ha hecho. La primera sección del bloc de notas es la voz a texto que controla la conversión de archivos de audio a texto. A continuación se muestra la sección Análisis de sentimientos que extrae sentimientos para cada frase de texto y muestra dichos resultados en un formato similar al panel propuesto.

> Este portátil debe ejecutarse antes del entrenamiento del modelo y del ajuste preciso porque el conjunto de datos MP3 debe descargarse y convertirse al formato correcto.

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

Centro de soporte - formación de modelos y afinación.ipynb

El entorno virtual del kit de herramientas de TAO debe configurarse antes de ejecutar el cuaderno (consulte la sección del kit de herramientas de TAO en la descripción general de comandos para obtener instrucciones de instalación).

Este cuaderno confía en el kit de herramientas TAO para ajustar los modelos de aprendizaje profundo en los datos de los clientes. Al igual que en el cuaderno anterior, este se separa en dos secciones para los

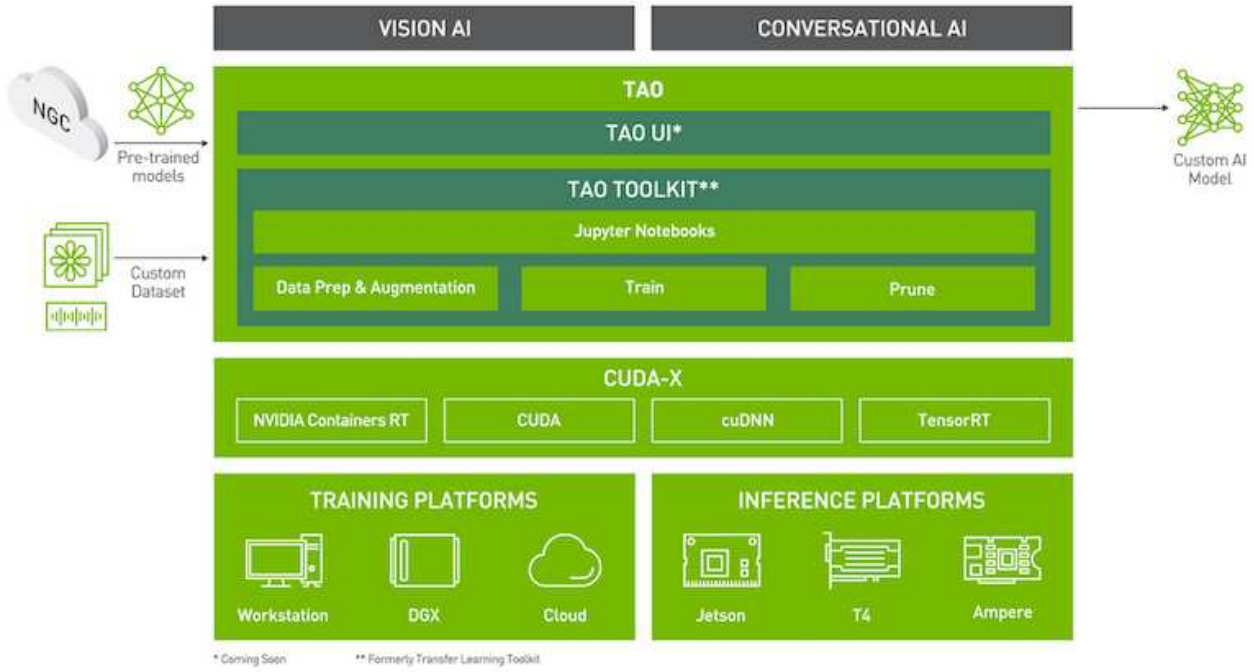
componentes de análisis de opinión y voz. Cada sección abarca el procesamiento de datos, la formación de modelos y el ajuste preciso, la evaluación de los resultados y la exportación de modelos. Por último, existe una sección final para implementar tanto sus modelos ajustados para su uso en RIVA.

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.

Conclusión

A medida que se considera cada vez más la experiencia de los clientes como un campo de batalla competitivo clave, un centro de soporte global aumentado por IA se convierte en un componente fundamental que las empresas de casi todos los sectores no pueden permitirse el lujo de descuidar. La solución propuesta en este informe técnico ha quedado demostrada para respaldar la prestación de estas experiencias excepcionales a los clientes, y el reto consiste ahora en garantizar que las empresas tomen acciones para modernizar su infraestructura de IA y sus flujos de trabajo.

Las mejores implementaciones de IA en servicio al cliente no consisten en sustituir a los agentes humanos. Más bien, la IA puede empoderarlos para crear experiencias de cliente excepcionales mediante análisis de sensibilidad en tiempo real, escalado de disputas y computación afectiva multimodal para detectar señales verbales, no verbales y faciales con las que los modelos de IA integrales pueden hacer recomendaciones a escala y complementar lo que podría carecer un agente humano individual. Asimismo, la IA puede proporcionar una mejor unión entre un cliente concreto con los agentes disponibles actualmente. Gracias a la IA, las empresas pueden extraer un valioso sentimiento de los clientes en cuanto a sus pensamientos e impresiones sobre los productos, servicios y la imagen de la Marca del proveedor.

La solución también se puede usar para construir datos de series temporales para que los agentes de soporte sirvan como métrica de evaluación del rendimiento objetivo. Las encuestas convencionales de satisfacción de clientes a menudo no tienen respuestas suficientes. Al recopilar la opinión de los empleados y los clientes a largo plazo, los empleadores pueden tomar decisiones informadas con respecto al desempeño de los agentes de apoyo.

La combinación de NetApp, SFL Scientific, marcos de orquestación de código abierto y NVIDIA reúne las tecnologías más recientes como servicios gestionados con una gran flexibilidad para acelerar la adopción de tecnología y mejorar el plazo de comercialización de las nuevas aplicaciones de IA/ML. Estos servicios avanzados se ofrecen en las instalaciones que se pueden transportar fácilmente para entornos nativos del cloud así como arquitecturas de puesta en marcha híbrida.

Dónde encontrar información adicional

Si quiere más información sobre el contenido de este documento, consulte los siguientes documentos o sitios web:

- Demostraciones interactivas en 3D

  "www.netapp.com/ai"

- Conecte directamente con un especialista en IA de NetApp

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

- Descripción de la solución NVDIA base Command Platform con NetApp

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

- NetApp para IA 10: Infografía buenas razones

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

- IA en la sanidad: Aprendizaje profundo para identificar las lesiones del COVID-19 en exploraciones pulmonares TC

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

- AI en la atención sanitaria: Monitorización del uso de la máscara facial en entornos sanitarios documento técnico

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

- IA en la sanidad: Informe técnico de imágenes de diagnóstico

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

- IA para minoristas: Inteligencia artificial de NetApp conversacional con NVIDIA RIVA

  "https://docs.netapp.com/us-en/netapp-solutions/ai/cainvidia_executive_summary.html"

- Resumen de la solución ONTAP AI de NetApp

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

- Resumen de la solución del kit de herramientas de operaciones de datos de NetApp

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

- Resumen de la solución del plano de control de IA de NetApp

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

- Libro electrónico "transformando el sector con los datos impulsan la IA"

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

- Resumen de la solución de IA EF-Series de NetApp

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

- Resumen de la solución de NetApp AI y Lenovo ThinkSystem para la inferencia de IA

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

- Resumen de la solución NetApp AI y Lenovo ThinkSystem para IA empresarial Y ML

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

- NetApp y NVIDIA: Redefiniendo lo que es posible con el vídeo de IA

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

Formación distribuida en Azure: Predicción de frecuencias mediante clic

TR-4904: Formación distribuida en Azure - predicción de frecuencias mediante clic

Rick Huang, Verron Martina, Muneer Ahmad, NetApp

El trabajo de un científico de datos debe centrarse en el entrenamiento y el ajuste de los modelos de aprendizaje automático y de inteligencia artificial (IA). Sin embargo, según una investigación de Google, los científicos de datos dedican aproximadamente el 80 % de su tiempo a averiguar cómo hacer que sus modelos funcionen con aplicaciones empresariales y se ejecuten a escala.

Para gestionar proyectos de IA y ML integrales, se necesita una comprensión más amplia de los componentes empresariales. Aunque DevOps ha retomado la definición, la integración y la puesta en marcha, estos tipos de componentes, LAS operaciones ML siguen un flujo similar que incluye proyectos de IA/ML. Para hacerse una idea de lo que puede tocar una canalización de IA/ML integral en la empresa, consulte la siguiente lista de componentes requeridos:

- Reducida
- Redes
- Oracle
- Sistemas de ficheros
- Contenedores
- Integración continua y canalización de puesta en marcha continua (CI/CD)
- Entorno de desarrollo integrado (IDE)
- Seguridad
- Políticas de acceso a los datos
- Hardware subyacente
- Cloud
- Virtualización
- Conjuntos de herramientas y bibliotecas de ciencia de datos

Público objetivo

El mundo de la ciencia de datos tiene múltiples disciplinas DE TECNOLOGÍA y negocio:

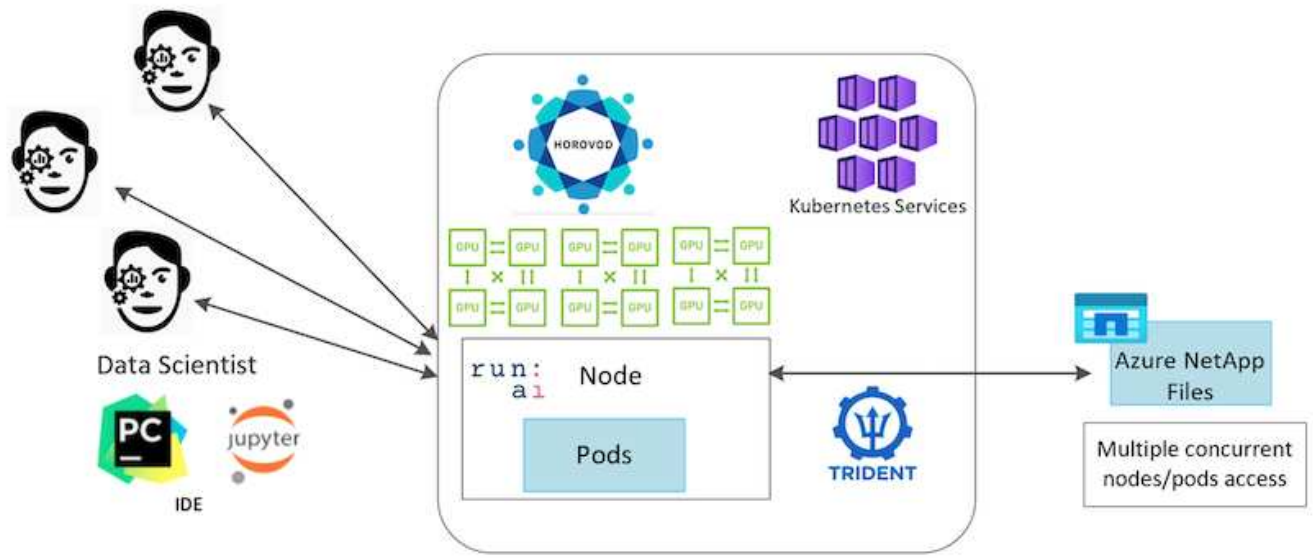
- El científico de datos necesita la flexibilidad para poder usar las herramientas y las bibliotecas de elección.
- El ingeniero de datos necesita saber cómo fluyen los datos y dónde residen.
- Un ingeniero de DevOps necesita herramientas para integrar nuevas aplicaciones de IA/ML en sus canalizaciones de CI/CD.
- Los administradores de cloud y arquitectos tienen que poder configurar y gestionar recursos de Azure.
- Los usuarios empresariales quieren tener acceso a aplicaciones de IA/ML.

En este informe técnico describimos cómo Azure NetApp Files, RAPIDS AI, Dink y Azure ayudan a cada uno de estos roles a aportar valor empresarial.

Descripción general de la solución

Esta solución sigue el ciclo de vida de una aplicación de IA/ML. Empezamos con la labor de los científicos de datos para definir los diferentes pasos necesarios para preparar datos y entrenar modelos. Al aprovechar RAPIDS on Dink, realizamos formación distribuida en el clúster de Azure Kubernetes Service (AKS) para reducir drásticamente el tiempo de entrenamiento en comparación con el método convencional de aprendizaje del kit de ciencias de Python. Para completar el ciclo completo, integramos la canalización con Azure NetApp Files.

Azure NetApp Files ofrece varios niveles de rendimiento. Los clientes pueden comenzar con un nivel estándar y escalar horizontalmente y verticalmente a un nivel de alto rendimiento de forma no disruptiva y sin necesidad de mover datos. Esta funcionalidad permite a los científicos de datos entrenar modelos a escala sin problemas de rendimiento y evitar silos de datos en el clúster, como se muestra en la siguiente figura.



Información general de la tecnología

Esta página proporciona una descripción general de la tecnología utilizada en esta solución.

Microsoft y NetApp

Desde mayo de 2019, Microsoft ha ofrecido un servicio de portales nativo de Azure para servicios de archivos NFS y SMB empresariales basados en la tecnología ONTAP de NetApp. Este desarrollo está impulsado por una asociación estratégica entre Microsoft y NetApp, y amplía aún más el alcance de los servicios de datos de ONTAP de primera calidad para Azure.

Azure NetApp Files

El servicio de Azure NetApp Files es un servicio de almacenamiento de ficheros de alto rendimiento y medida para empresas. Azure NetApp Files es compatible con cualquier tipo de carga de trabajo y está altamente disponible de manera predeterminada. Puede seleccionar los niveles de servicio y rendimiento, y configurar copias Snapshot a través del servicio. Azure NetApp Files es un servicio de primera parte de Azure para migrar y ejecutar las cargas de trabajo de archivo empresarial más exigentes en el cloud, incluidas bases de datos, SAP y aplicaciones de computación de alto rendimiento sin necesidad de modificar el código.

Esta arquitectura de referencia proporciona a las organizaciones DE TI las siguientes ventajas:

- Elimina las complejidades de diseño
- Permite un escalado independiente de las capacidades de computación y almacenamiento
- Permite a los clientes empezar con poco y escalar sin problemas
- Ofrece una gama de niveles de almacenamiento para distintos niveles de rendimiento y coste

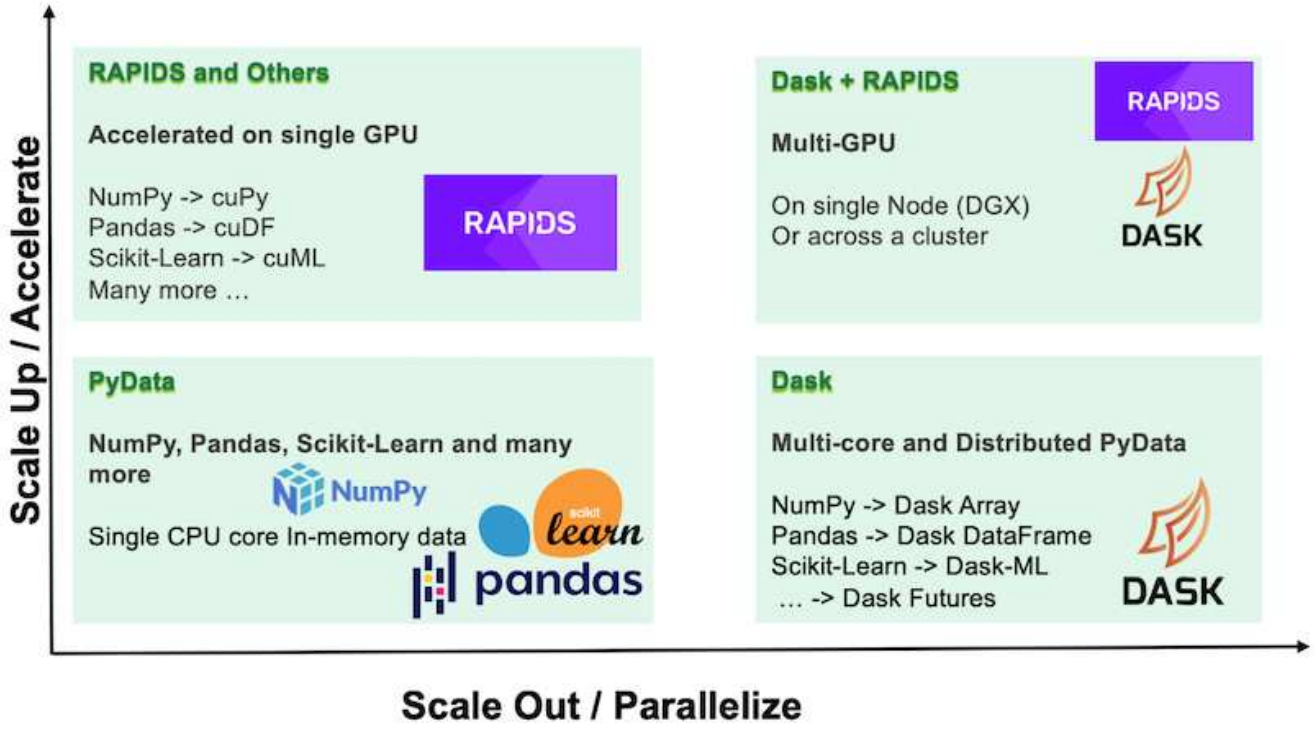
Información general sobre DASK y NVIDIA RAPIDS

DASK es una herramienta de computación paralela de código abierto que escala bibliotecas Python en varias máquinas y permite un procesamiento más rápido de grandes cantidades de datos. Proporciona una API similar a las bibliotecas de Python convencionales de un solo subproceso, como Pandas, numpy y scikit-Learn. Como resultado, los usuarios nativos de Python no se ven obligados a cambiar mucho su código

existente para utilizar recursos en el clúster.

NVIDIA RAPIDS es un paquete de bibliotecas de código abierto que permite ejecutar flujos de trabajo completos de APRENDIZAJE AUTOMÁTICO y análisis de datos en GPU. Junto con DASK, puede escalar con facilidad desde una estación de trabajo con GPU (escalado vertical) a clústeres de varios nodos y varias GPU (escalado horizontal).

Para poner en marcha DASK en un clúster, podría utilizar Kubernetes para la orquestación de recursos. También podría escalar verticalmente o reducir los nodos de trabajo según los requisitos del proceso, lo cual, a su vez, puede ayudarle a optimizar el consumo de recursos de clúster, como se muestra en la siguiente figura.



Requisitos de software

En la siguiente tabla se enumeran los requisitos de software necesarios para esta solución.

De NetApp	Versión
Servicio Azure Kubernetes	1.18.14
Imagen de contenedor DE RAPIDS y DASK	Repositorio: "Rapidsai/rapidsai" etiqueta: 0.17-cuda11.0-Runtime-ubuntu18.04
Trident de NetApp	20.01.1
Timón	3.0.0

Requisitos de recursos cloud

Esta página describe la configuración de los recursos de cloud para Azure NetApp Files.

Configure Azure NetApp Files

Configure Azure NetApp Files como se describe en "Inicio rápido: Configure Azure NetApp Files y cree un volumen NFS".

Puede pasar más allá de la sección “Crear volumen NFS para Azure NetApp Files” porque va a crear volúmenes a través de Trident. Antes de continuar, realice los siguientes pasos:

1. Regístrese para Azure NetApp Files y el proveedor de recursos de NetApp (a través de la shell de Azure) ("enlace").
2. Crear una cuenta en Azure NetApp Files ("enlace").
3. Configurar un pool de capacidad (un estándar o Premium de 4 TB como mínimo, según sus necesidades) ("enlace").la siguiente tabla enumera los requisitos de configuración de red para configurar en la nube. El clúster de DASK y Azure NetApp Files deben estar en la misma red virtual de Azure (vnet) o en una vnet con conexión entre iguales.

Recursos	Tipo/versión
Servicio Azure Kubernetes	1.18.14
Nodo de agente	3x Standard_DS2_v2
Nodo GPU	3x Standard_NC63_v3
Azure NetApp Files	Pool de capacidad estándar
Capacidad en TB	4

Resumen de casos de uso de predicción de velocidad mediante clic

Este caso de uso se basa en el público disponible "Terabyte haga clic en registros" conjunto de datos de "Laboratorio Criteo AI". Con los recientes avances en las plataformas Y aplicaciones DE ML, ahora se presta mucha atención al aprendizaje a escala. La tasa de clics (CTR) se define como el número medio de clics-throughs por cien impresiones de anuncios en línea (expresado como porcentaje). Se ha adoptado ampliamente como métrica clave en diversos mercados verticales del sector y casos de uso, incluidos el marketing digital, el comercio minorista, el comercio electrónico y los proveedores de servicios. Algunos ejemplos de uso de CTR como una métrica importante para el tráfico potencial de clientes son los siguientes:

- **Marketing digital:** in "Google Analytics", CTR se puede usar para medir cómo se están realizando las palabras clave, los anuncios y los listados libres de un anunciante o comerciante. Un CTR alto es una buena indicación de que los usuarios encuentran sus anuncios y listados útiles y relevantes. CTR también contribuye a la CTR esperada de su palabra clave, que es un componente de "Clasificación de anuncios".
- **Comercio electrónico:** además de apalancar "Google Analytics", hay por lo menos algunas estadísticas de visitantes en un fondo de comercio electrónico. Aunque estas estadísticas pueden no parecer útiles a primera vista, suelen ser fáciles de leer y pueden ser más precisas que otras informaciones. Los conjuntos de datos de primera parte compuestos por estas estadísticas son de propiedad y, por lo tanto, los más relevantes para los vendedores, compradores y plataformas de comercio electrónico. Estos conjuntos de datos se pueden utilizar para establecer pruebas de rendimiento, comparar los resultados con el año pasado y el pasado construyendo una serie temporal para un análisis más profundo.
- **Retail:** los minoristas tradicionales pueden correlacionar el número de visitantes y el número de clientes

con el CTR. El número de clientes se puede ver desde su historial de puntos de venta. El CTR de los sitios web de los minoristas o del tráfico de anuncios puede resultar en las ventas mencionadas. Los programas de fidelidad son otro caso de uso, ya que los clientes redirigidos de anuncios en línea u otros sitios web podrían unirse para obtener recompensas. Los minoristas pueden conseguir clientes a través de programas de fidelidad y comportamientos récord de historias de ventas para crear un sistema de recomendaciones que no solo predice las conductas de compra de los consumidores en diferentes categorías, sino que también personaliza las cupones y reduce la pérdida de clientes.

- **Proveedores de servicios:** las empresas de telecomunicaciones y los proveedores de servicios de Internet tienen una abundancia de datos de telemetría de usuarios de primera parte para casos de uso de IA, ML y analítica perspicaz. Por ejemplo, las telecomunicaciones pueden aprovechar los registros diarios de historial de dominios de navegación web de sus suscriptores móviles para ajustar los modelos existentes y producir una segmentación de público actualizada, predecir el comportamiento de los clientes y colaborar con los anunciantes para colocar anuncios en tiempo real para una mejor experiencia en línea. En este flujo de trabajo de marketing basado en datos, CTR es una métrica importante para reflejar las conversiones.

En el contexto del marketing digital, "Criteo Terabyte haga clic en registros" Son ahora el conjunto de datos de referencia a la hora de evaluar la escalabilidad de las plataformas Y algoritmos ML. Al predecir la tarifa de clic, un anunciante puede seleccionar a los visitantes que tienen más probabilidades de responder a los anuncios, analizar su historial de navegación y mostrar los anuncios más relevantes basados en los intereses del usuario.

La solución proporcionada en este informe técnico destaca las siguientes ventajas:

- Las ventajas de Azure NetApp Files en formación distribuida o a gran escala
- RÁPIDO procesamiento de datos habilitado para CUDA (cuDF, cúpula, etc.) y algoritmos ML (cuML)
- El marco informático paralelo de DASK para la formación distribuida

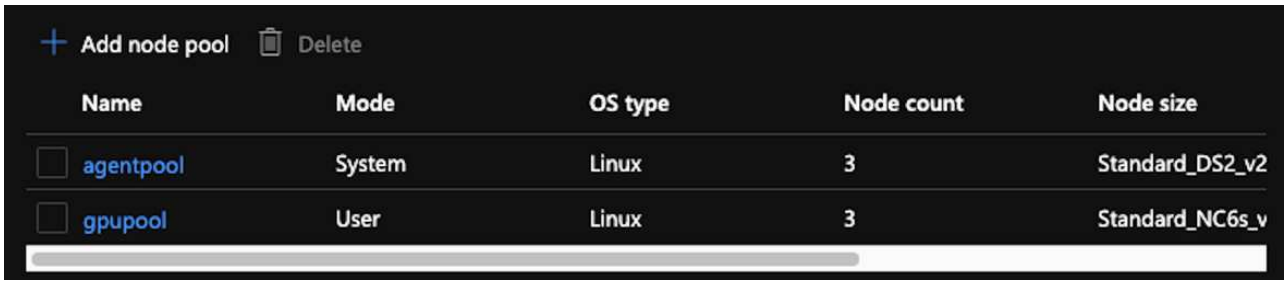
Un flujo de trabajo integral basado en RAPIDS AI y Azure NetApp Files demuestra la drástica mejora del tiempo de entrenamiento de los modelos de bosques aleatorios en dos órdenes de magnitud. Esta mejora es significativa en comparación con el enfoque convencional de pandas al tratar los registros de clic del mundo real con 45GB de datos tabulares estructurados (en promedio) cada día. Esto equivale a un DataFrame que contiene aproximadamente veinte mil millones de filas. Demostraremos la configuración del entorno de clúster, la instalación de marcos y bibliotecas, la carga y el procesamiento de datos, la formación convencional frente a la distribuida, la visualización y supervisión, y compararemos los resultados fundamentales de tiempo de ejecución completo en este informe técnico.

Configuración

Instalar y configurar el clúster AKS

Para instalar y configurar el clúster AKS, consulte la página web "Cree un clúster de AKS" y, a continuación, realice los siguientes pasos:

1. Al seleccionar el tipo de nodo (nodos System [CPU] o worker [GPU]), seleccione lo siguiente:
   a. Los nodos del sistema principal deben ser DS2v2 estándar (`agentpool` tres nodos predeterminados).
   b. A continuación, agregue el pool Standard_NC66s_v3 del nodo de trabajo (tres nodos como mínimo) para el grupo de usuarios (para nodos GPU) denominado `gpupool`.



2. La puesta en marcha tarda entre 5 y 10 minutos. Una vez finalizado, haga clic en Connect to Cluster.
3. Para conectarse al clúster AKS recién creado, instale lo siguiente desde su entorno local (portátil/pc):
   a. La herramienta de línea de comandos de Kubernetes que utiliza "Instrucciones proporcionadas para su SO específico"
   b. El CLI de Azure tal como se describe en el documento, "Instale la CLI de Azure"
4. Para acceder al clúster AKS desde el terminal, introduzca `az login` e introduzca las credenciales.
5. Ejecute los dos comandos siguientes:

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

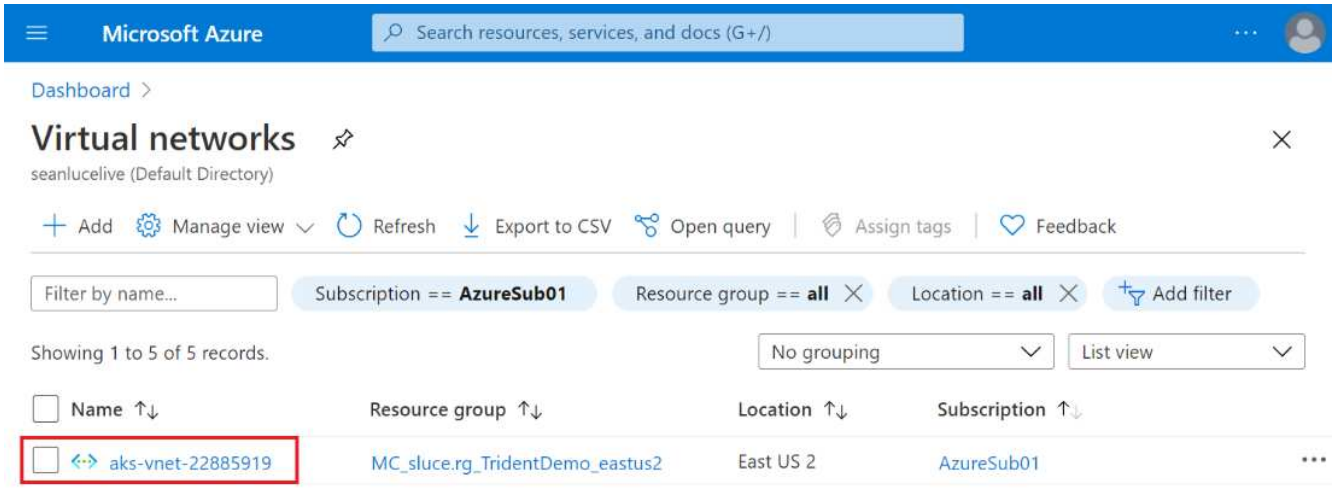
6. Introduzca `Azure CLI: kubectl get nodes`.
7. Si los seis nodos están en funcionamiento, como se muestra en el siguiente ejemplo, su clúster AKS estará listo y conectado a su entorno local

```
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 ~ %
```

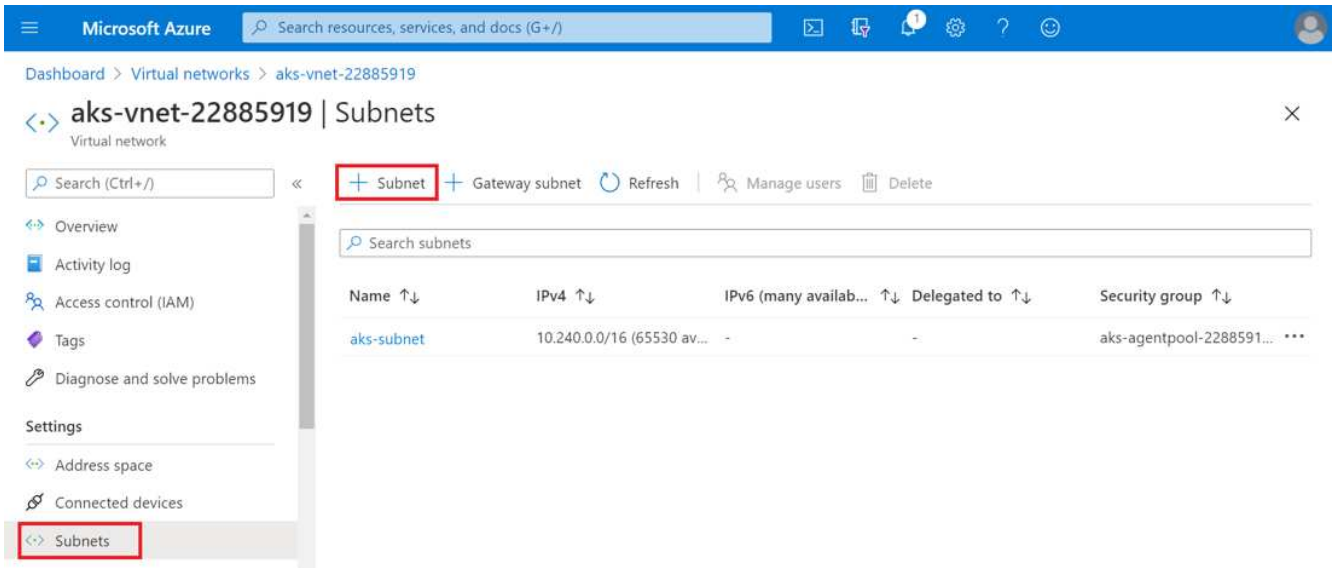
Cree una subred delegada para Azure NetApp Files

Para crear una subred delegada para Azure NetApp Files, lleve a cabo los siguientes pasos:

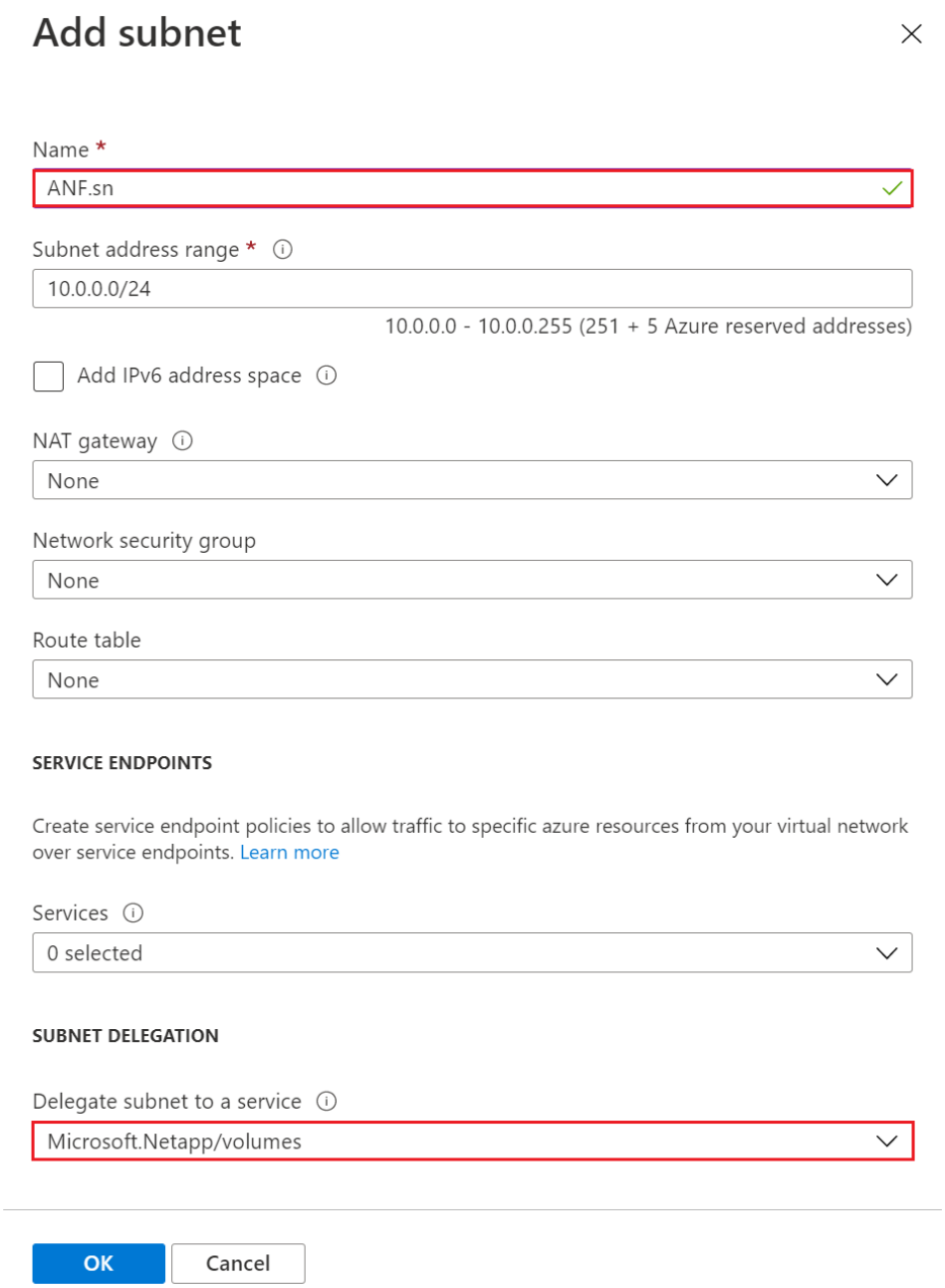
1. Acceda a Virtual Networks dentro del portal de Azure. Busque la red virtual que acaba de crear. Debe tener un prefijo como `aks-vnet`.
2. Haga clic en el nombre de la vnet.



3. Haga clic en subredes y, a continuación, en +Subnet en la barra de herramientas superior.



4. Proporcione la subred con un nombre como `ANF.sn` Y, en el encabezado Delegación de subred, seleccione `Microsoft.Netapp/volumes`. No cambie nada más. Haga clic en Aceptar.



Los volúmenes Azure NetApp Files se asignan al clúster de aplicaciones y se consumen como reclamaciones de volúmenes persistentes (RVP) en Kubernetes. A su vez, este proceso le proporciona la flexibilidad para asignarlos a diferentes servicios, tales como portátiles Juppyter, funciones sin servidor, etc.

Los usuarios de servicios pueden consumir almacenamiento desde la plataforma de muchas maneras. Como este informe técnico trata sobre NFSS, los principales beneficios de Azure NetApp Files son:

- Brindar a los usuarios la capacidad de usar copias de Snapshot.
- Permitir a los usuarios almacenar grandes cantidades de datos en volúmenes de Azure NetApp Files.
- Utilización de las ventajas en el rendimiento de los volúmenes Azure NetApp Files cuando se ejecutan sus modelos en conjuntos de archivos de gran tamaño.

AKS vnet y Azure NetApp Files vnet

Para conectar el AKS vnet al vnet de Azure NetApp Files, lleve a cabo los siguientes pasos:

1. Introduzca redes virtuales en el campo de búsqueda.
2. Seleccione `vnet aks-vnet-name`. Haga clic en él e introduzca los peerings en el campo de búsqueda.
3. Haga clic en +Agregar.
4. Introduzca los siguientes descriptores:
   a. El nombre del enlace de relación de paridad es `aks-vnet-name_to_anf`.
   b. SubscriptionId y Azure NetApp Files vnet como partner de vnet peering.
   c. Deje todas las secciones que no sean asteriscos con los valores predeterminados.
5. Haga clic en Añadir.

Para obtener más información, consulte "Crear, cambiar o eliminar una conexión de red virtual entre iguales".

Instale Trident

Para instalar Trident con Helm, lleve a cabo los siguientes pasos:

1. Instale Helm (para obtener instrucciones de instalación, visite "origen").
2. Descargue y extraiga el instalador de Trident 20.01.1.

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

3. Cambie el directorio a. `trident-installer`.

```
$cd trident-installer
```

4. Copiar `tridentctl` a un directorio del sistema `$PATH`.

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

5. Instale Trident en el clúster Kubernetes (K8s) con Helm ("origen"):
   a. Cambie el directorio a `helm` directorio.

   ```
   $cd helm
   ```

   b. Instale Trident.

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

c. Comprobar el estado de los pods de Trident.

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

Si todos los pods están ya en funcionamiento, se instala Trident y se puede avanzar.

6. Configurar el back-end de Azure NetApp Files y la clase de almacenamiento para AKS.

a. Cree un principio de Azure Service.

El principal del servicio es cómo Trident se comunica con Azure para manipular sus recursos de Azure NetApp Files.

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

El resultado debería ser como el ejemplo siguiente:

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

7. Cree un archivo del back-end json de Trident, nombre de ejemplo `anf-backend.json`.

8. Con el editor de texto preferido, complete los siguientes campos dentro del `anf-backend.json` archivo:

```
{
    "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. Sustituya los siguientes campos:
   - `subscriptionID`. Su ID de suscripción de Azure.
   - `tenantID`. Su ID de inquilino de Azure de la salida de `az ad sp` en el paso anterior.
   - `clientID`. Su AppID desde la salida de `az ad sp` en el paso anterior.
   - `clientSecret`. Su contraseña de la salida de `az ad sp` en el paso anterior.
10. Indique a Trident que cree el back-end de Azure NetApp Files en la `trident` espacio de nombres con `anf-backend.json` como archivo de configuración:

```
$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. Cree una clase de almacenamiento. Los usuarios de Kubernetes aprovisionan volúmenes con RVP que especifican una clase de almacenamiento por nombre. Indique a K8S que cree una clase de almacenamiento `azurenetappfiles` Que hace referencia al back-end de Trident creado en el paso anterior.
12. Cree una AYLMA (`anf-storage-class.yaml`) archivo para la clase de almacenamiento y copiar.

```
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. Compruebe que la clase de almacenamiento se ha creado.

```
kubectl get sc azurenetappfiles
```

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

Configurar Dask con LA implementación DE RAPIDS en AKS con Helm

Para configurar el despliegue de Dask con RAPIDS en AKS con Helm, lleve a cabo los siguientes pasos:

1. Cree un espacio de nombres para instalar DASK con RAPIDS.

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

2. Crear una RVP para almacenar el conjunto de datos con velocidad de clic:
   a. Guarde el siguiente contenido de YAML en un archivo para crear un PVC.

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

   b. Aplique el archivo YAML al clúster de Kubernetes.

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

3. Clone el `rapidsai git` repositorio ("https://github.com/rapidsai/helm-chart").

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

4. Modificar `values.yaml` E incluya el PVC creado anteriormente para los trabajadores y el espacio de trabajo Juppyter.
   a. Vaya a la `rapidsai` directorio del repositorio.

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

   b. Actualice el `values.yaml` Coloque el archivo y monte el volumen con la RVP.

```
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. Vaya al directorio principal del repositorio e implemente Dask con tres nodos de trabajo en AKS utilizando

Helm.

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

Niveles de rendimiento de Azure NetApp Files

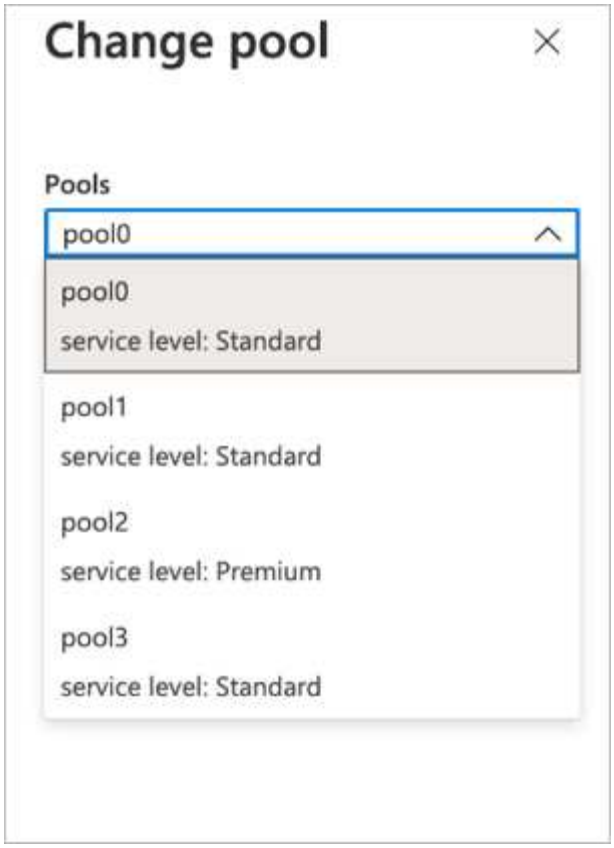
Es posible cambiar el nivel de servicio de un volumen existente si se mueve el volumen a otro pool de capacidad que utiliza el nivel de servicio que se desea para el volumen. Esta solución permite a los clientes comenzar con un conjunto de datos pequeño y un número reducido de GPU en el nivel estándar y escalar horizontalmente o verticalmente hasta el nivel Premium a medida que aumenta la cantidad de datos y las GPU. El nivel Premium ofrece cuatro veces el rendimiento por terabyte como nivel estándar, así como la escalabilidad vertical se realiza sin tener que mover datos para cambiar el nivel de servicio de un volumen.

Cambie dinámicamente el nivel de servicio de un volumen

Para cambiar de forma dinámica el nivel de servicio de un volumen, complete los pasos siguientes:

1. En la página Volumes, haga clic con el botón derecho en el volumen cuyo nivel de servicio desea cambiar. Seleccione Cambiar pool.

2. En la ventana Cambiar pool, seleccione el pool de capacidad al que desea mover el volumen.



3. Haga clic en Aceptar.

Automatice el cambio en el nivel de rendimiento

Existen las siguientes opciones para automatizar los cambios en el nivel de rendimiento:

- El cambio de nivel de servicio dinámico sigue en la vista previa pública en este momento y no está activado de forma predeterminada. Para habilitar esta función en la suscripción a Azure, consulte esta documentación sobre cómo "Cambie dinámicamente el nivel de servicio de un volumen".
- Se proporcionan comandos de cambio de pool de volúmenes de la interfaz de línea de comandos de Azure en "documentación de cambio de pool de volúmenes" y en el ejemplo siguiente:

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

- PowerShell: El "Set-AzNetAppFilesVolumePool" Cambia el pool de un volumen Azure NetApp Files y se muestra en el ejemplo siguiente:

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

Haga clic en el procesamiento de datos de predicción de velocidad y el entrenamiento de modelos

Bibliotecas para el procesamiento de datos y el entrenamiento de modelos

En la tabla siguiente se enumeran las bibliotecas y los marcos que se utilizaron para generar esta tarea. Todos estos componentes se han integrado completamente con los controles de seguridad y acceso basados en roles de Azure.

Bibliotecas/marco de trabajo	Descripción
CuML DASK	Para QUE EL ML funcione en la GPU, el "Biblioteca de cuML" Ofrece acceso al paquete cuML DE RAPIDS con DASK. RAPIDS cuML implementa algoritmos DE ML más conocidos, como los métodos de clustering, reducción de dimensiones y regresión, con implementaciones basadas en GPU de alto rendimiento que ofrecen una velocidad de hasta 100 veces superior a los métodos basados en CPU.
DASK cuDF	CuDF incluye varias otras funciones que admiten la extracción, transformación y carga (ETL) acelerada por GPU, como la subconfiguración de datos, transformaciones, codificación en caliente, etc. El equipo DE RAPIDS mantiene un "biblioteca dask-cudf" Eso incluye métodos auxiliares para usar DASK y cuDF.
Formación en Scikit	Scikit-Learn proporciona docenas de algoritmos y modelos de aprendizaje automático integrados, llamados estimadores. Cada uno "estimator" se puede ajustar a algunos datos mediante su "encajar" método.

Hemos utilizado dos cuadernos para construir los gasoductos ML para su comparación; uno es el método convencional de curscikit-aprender de pandas, y el otro es el entrenamiento distribuido con RAPIDS y Dink. Cada portátil se puede probar individualmente para ver el rendimiento en términos de tiempo y escala. Cubrimos cada bloc de notas individualmente para demostrar las ventajas de la formación distribuida con RAPIDS y Dink.

Cargue Criteo haga clic en el día 15 de los registros en pandas y entrena un cikit-aprende el modelo de bosque aleatorio

En esta sección se describe cómo utilizamos Pandas y DASK DataFrames para cargar datos Click Logs del conjunto de datos Criteo Terabyte. El caso de uso es relevante en la publicidad digital para intercambios de anuncios para crear perfiles de usuarios al predecir si se hará clic en anuncios o si el intercambio no está utilizando un modelo exacto en una canalización automatizada.

Se cargaron los datos del día 15 desde el conjunto de datos Click Logs, sumando 45 GB. Ejecutar la siguiente celda en el portátil Jupyter `CTR-PandasRF-collated.ipynb` Crea un DataFrame de pandas que contiene los primeros 50 millones de filas y genera un modelo de bosque aleatorio cikit-aprender.

```
%%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)
```

Para realizar la predicción utilizando un modelo de bosque aleatorio entrenado, ejecute el siguiente párrafo en este cuaderno. Tomamos las últimas filas de un millón del día 15 como conjunto de pruebas para evitar cualquier duplicación. La celda también calcula la precisión de la predicción, definida como el porcentaje de ocurrencias que el modelo predice con precisión si un usuario hace clic o no en un anuncio. Para revisar cualquier componente desconocido en este cuaderno, consulte "documentación oficial de scikit-aprender".

```
# 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))
```

Cargar día 15 en DASK y entrenar un modelo de bosque aleatorio DASK cuML

De una manera similar a la sección anterior, cargue Criteo Click Logs Day 15 en Pandas y entrena un cikit-aprende el modelo de bosque aleatorio. En este ejemplo, realizamos la carga de DataFrame con DASK cuDF y entrenamos un modelo de bosque aleatorio en DASK cuML. Hemos comparado las diferencias en el tiempo de formación y el escalado en la sección "“Comparación del tiempo de formación”."

criteo_dask_RF.ipynb

Este portátil importa `numpy`, `cuml`, y lo necesario `dask` bibliotecas, como se muestra en el siguiente ejemplo:

```
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
```

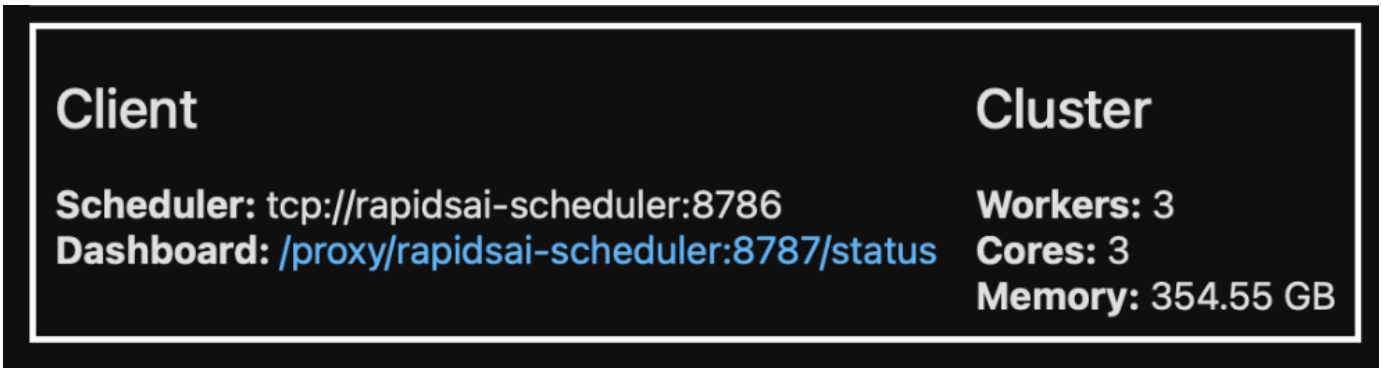
Inicie cliente DASK().

```
client = Client()
```

Si su clúster está configurado correctamente, puede ver el estado de los nodos de trabajo.

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

En nuestro clúster AKS, se muestra el siguiente estado:



Tenga en cuenta que DASK emplea el paradigma de ejecución lenta: En lugar de ejecutar el código de procesamiento al instante, DASK crea en su lugar un gráfico cíclico dirigido (DAG) de ejecución. DAG contiene un conjunto de tareas y sus interacciones que cada trabajador necesita ejecutar. Este diseño significa que las tareas no se ejecutan hasta que el usuario le indique a DASK que las ejecute de una forma u otra. Con DASK tiene tres opciones principales:

- **Call comput() en un DataFrame.** esta llamada procesa todas las particiones y, a continuación, devuelve los resultados al planificador para la agregación final y conversión a cuDF DataFrame. Esta opción debe usarse con moderación y sólo en resultados muy reducidos a menos que el nodo del programador se quede sin memoria.
- **Call persistent() en un DataFrame.** esta llamada ejecuta el gráfico, pero, en lugar de devolver los resultados al nodo del planificador, los mantiene en la memoria a través del clúster para que el usuario pueda reutilizar estos resultados intermedios en la canalización sin necesidad de volver a ejecutar el mismo procesamiento.
- **Call head() en un DataFrame.** al igual que con cuDF, esta llamada devuelve 10 registros al nodo del planificador. Esta opción se puede utilizar para comprobar rápidamente si el DataFrame contiene el formato de salida deseado o si los propios registros tienen sentido, en función del procesamiento y cálculo.

Por lo tanto, a menos que el usuario llama a cualquiera de estas acciones, los trabajadores se sientan inactivos esperando que el programador inicie el procesamiento. Este paradigma de ejecución perezosa es común en marcos informáticos modernos en paralelo y distribuidos como Apache Spark.

En el siguiente párrafo se entrena un modelo de bosque aleatorio mediante el uso de DASK cuML para computación acelerada por GPU distribuida y se calcula la precisión de predicción del modelo.

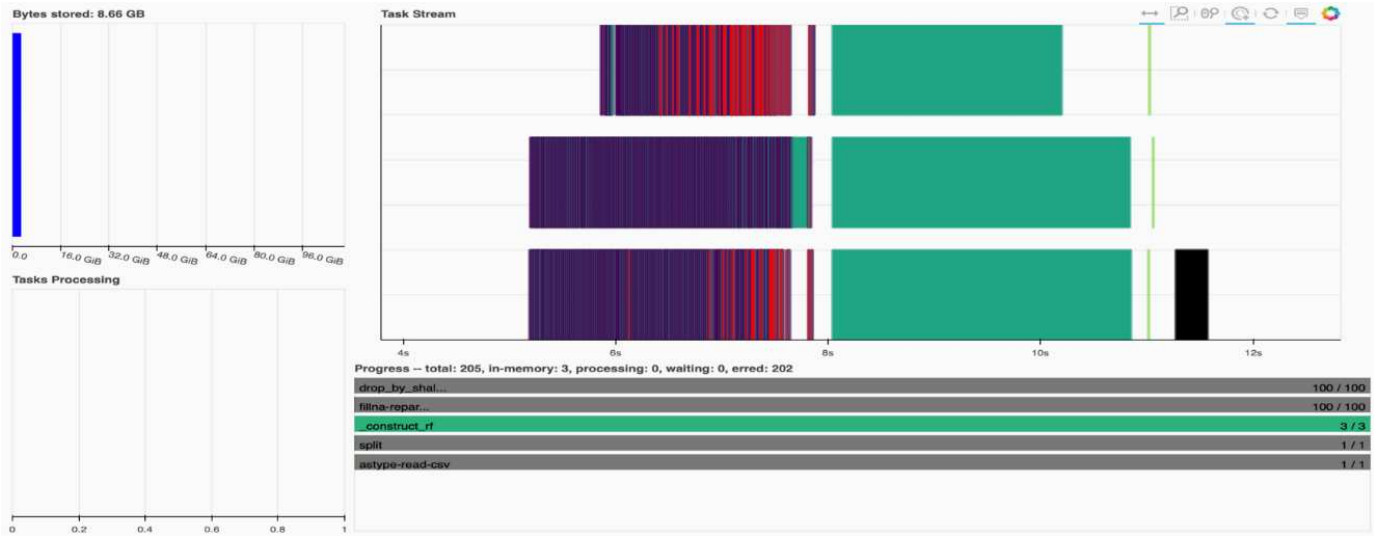
```
Adsf
# Random Forest building parameters
n_streams = 8 # optimization
max_depth = 10
n_bins = 16
n_trees = 10
cuml_model = cumlDaskRF(max_depth=max_depth, n_estimators=n_trees,
n_bins=n_bins, n_streams=n_streams, verbose=True, client=client)
cuml_model.fit(gdf_sliced_small, Y)
# Model prediction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

Supervisión de tarea mediante el panel de control de flujos de tareas nativo

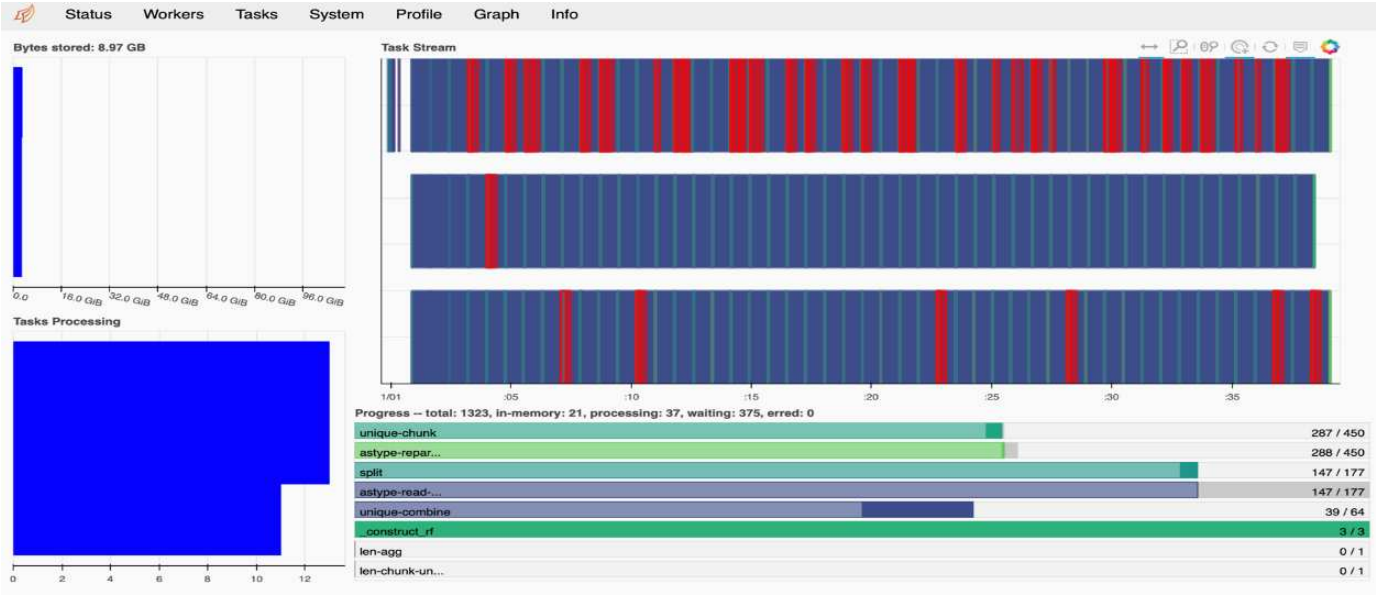
La "Planificador distribuido DASK" proporciona comentarios en directo de dos formas:

- Un panel interactivo que contiene muchos trazados y tablas con información en directo
- Una barra de progreso adecuada para uso interactivo en consolas o portátiles

En nuestro caso, la siguiente figura muestra cómo puede supervisar el progreso de la tarea, incluidos los bytes almacenados, el flujo de tareas con un desglose detallado del número de flujos y el progreso por los nombres de tareas con las funciones asociadas ejecutadas. En nuestro caso, debido a que tenemos tres nodos de trabajo, hay tres partes principales del flujo y los códigos de color denotan diferentes tareas dentro de cada flujo.



Tiene la opción de analizar tareas individuales y examinar el tiempo de ejecución en milisegundos o identificar cualquier obstáculo o impedimento. Por ejemplo, la siguiente figura muestra los flujos de tareas para la etapa de ajuste del modelo de bosque aleatorio. Se están ejecutando muchas más funciones, incluido el fragmento único para el procesamiento de DataFrame, _construct_rf para ajustar el bosque aleatorio, etc. La mayor parte del tiempo se ha empleado en operaciones DataFrame debido al gran tamaño (45GB) de un día de datos de los registros de clic de Criteo.



Comparación del tiempo de entrenamiento

Esta sección compara el tiempo de entrenamiento del modelo utilizando pandas convencionales en comparación con el DASK. Para Pandas, cargamos una cantidad menor de datos debido a la naturaleza del tiempo de procesamiento más lento, para evitar que se desbordara la memoria. Por lo tanto, interpolamos los resultados para ofrecer una comparación justa.

La siguiente tabla muestra la comparación del tiempo de entrenamiento bruto cuando hay significativamente menos datos utilizados para el modelo de bosque aleatorio de pandas (50 millones de filas de 20 mil millones por día 15 del conjunto de datos). Esta muestra sólo utiliza menos del 0.25% de todos los datos disponibles. Mientras que para DASK-cuML entrenamos el modelo de bosque aleatorio en las 20 mil millones de filas disponibles. Los dos enfoques dieron lugar a un tiempo de capacitación comparable.

Enfoque	Tiempo de entrenamiento
Scikit-Learn: Usando sólo 50 m de filas en el día 15 como datos de entrenamiento	47 minutos y 21 segundos
RAPIDS-Dask: Utilizando todas las filas 20B del día 15 como datos de entrenamiento	1 hora, 12 minutos y 11 segundos

Si interpolamos los resultados del tiempo de entrenamiento linealmente, como se muestra en la siguiente tabla, hay una ventaja significativa a utilizar el entrenamiento distribuido con DASK. Tomaría el enfoque convencional de Pandas scikit-Learn 13 días para procesar y entrenar 45GB de datos para un solo día de registros tecleo, mientras que EL enfoque RAPIDS-DASk procesa la misma cantidad de datos 262.39 veces más rápido.

Enfoque	Tiempo de entrenamiento
Scikit-Learn: Usando todas las filas 20B en el día15 como datos de entrenamiento	13 días, 3 horas, 40 minutos y 11 segundos
RAPIDS-Dask: Utilizando todas las filas 20B del día 15 como datos de entrenamiento	1 hora, 12 minutos y 11 segundos

En la tabla anterior, puede ver que usando RAPIDS con Dink para distribuir el procesamiento de datos y el entrenamiento de modelos en varias instancias de GPU, el tiempo de ejecución es significativamente más corto en comparación con el procesamiento convencional de Pandas DataFrame con el entrenamiento de modelos scikit-Learn. Este marco permite un escalado vertical y horizontal en el cloud, así como en las instalaciones, en un clúster multinodo con varias GPU.

Monitorizar Dink y RAPIDS con Prometheus y Grafana

Una vez que todo se pone en marcha, ejecute inferencias sobre nuevos datos. Los modelos predicen si un usuario hace clic en un anuncio basado en actividades de navegación. Los resultados de la predicción se almacenan en un cuDF de DASK. Puede supervisar los resultados con Prometheus y visualizar en paneles Grafana.

Para obtener más información, consulte este tema "RAPIDS AI Media Post".

Creación de versiones de conjuntos de datos y modelos con el kit de herramientas de operaciones de datos de NetApp

El kit de herramientas DataOPS de NetApp para Kubernetes abstrae los recursos de almacenamiento y las cargas de trabajo de Kubernetes hasta el nivel de espacio de trabajo de ciencia de datos. Estas funciones se presentan en una interfaz sencilla y fácil de usar diseñada para científicos e ingenieros de datos. Utilizando la forma familiar de un programa de Python, el kit de herramientas permite a científicos e ingenieros de datos aprovisionar y destruir espacios de trabajo de JuppyterLab en cuestión de segundos. Estas áreas de trabajo pueden contener terabytes o incluso petabytes de capacidad de almacenamiento, lo que permite a los científicos de datos almacenar todos sus conjuntos de datos de entrenamiento directamente en sus espacios de trabajo de proyectos. Han pasado los días de gestionar los espacios de trabajo y los volúmenes de datos por separado.

Para obtener más información, visite el Kit de herramientas "Repositorio de GitHub".

Portátiles Juppyter para referencias

Existen dos cuadernos Juppyter asociados a este informe técnico:

- **"CTR-PandasRF-collated.ipynb."** Este cuaderno carga el día 15 desde el conjunto de datos de registros Criteo Terabyte Click, procesa y formatea datos en un DataFrame de Pandas, entrena un modelo de bosque aleatorio Scikit-Learn, realiza predicción y calcula la precisión.
- **"criteo_dask_RF.ipynb."** Este cuaderno carga el día 15 desde el conjunto de datos de registros Criteo Terabyte Click, procesa y formatea datos en un CuDF DASK, entrena un modelo de bosque aleatorio DASK cuML, realiza predicción y calcula la precisión. Al aprovechar varios nodos de trabajo con GPU, este método de procesamiento y entrenamiento de datos distribuidos y modelos es altamente eficiente. Cuantos más datos procese, mayor será el ahorro de tiempo que se consigue con el método DE ML convencional. Puede implementar este portátil en el cloud, en las instalaciones o en un entorno híbrido en el que el clúster de Kubernetes contenga recursos informáticos y de almacenamiento en diferentes ubicaciones, siempre y cuando su configuración de red permita el movimiento libre de datos y la distribución de modelos.

Conclusión

Azure NetApp Files, RAPIDS y Dink aceleran y simplifican la puesta en marcha del procesamiento y formación DE ML a gran escala gracias a la integración con herramientas de orquestación como Docker y Kubernetes. Al unificar la canalización de datos completa, esta solución reduce la latencia y la complejidad inherentes de muchas cargas de trabajo informáticas avanzadas, y permite salvar en la práctica la brecha entre el desarrollo y las operaciones. Los científicos de datos pueden ejecutar consultas en grandes conjuntos de datos y compartir de forma segura datos y modelos algorítmicos con otros usuarios durante la fase de entrenamiento.

Cuando cree sus propias canalizaciones de IA/ML, configurar la integración, la gestión, la seguridad y la accesibilidad de los componentes en una arquitectura es una tarea ardua. Dar a los desarrolladores acceso y control de su entorno presenta otro conjunto de retos.

Al crear un modelo de entrenamiento distribuido completo y una canalización de datos en el cloud, demostramos dos órdenes de mejora de magnitud en el tiempo de finalización total de los flujos de trabajo en comparación con un enfoque convencional de código abierto que no utilizaba el procesamiento de datos acelerado por GPU y los marcos informáticos.

La combinación de NetApp, Microsoft, marcos de orquestación de código abierto y NVIDIA reúne las tecnologías más recientes como servicios gestionados con una gran flexibilidad para acelerar la adopción de tecnología y mejorar el plazo de comercialización de las nuevas aplicaciones de IA/ML. Estos servicios avanzados se ofrecen en un entorno nativo del cloud que se puede realizar fácilmente en las instalaciones y en arquitecturas de puesta en marcha híbrida.

Dónde encontrar información adicional

Si quiere obtener más información sobre el contenido de este documento, consulte los siguientes recursos:

- Azure NetApp Files.:
  - Página de arquitectura de soluciones para Azure NetApp Files

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

- Almacenamiento persistente de Trident para contenedores:
  - Azure NetApp Files y Trident

    "https://netapptrident.readthedocs.io/en/stablev20.07/kubernetes/operations/tasks/backends/anf.html"

- DASK y RAPIDS:
  - DASK

    "https://docs.dask.org/en/latest/"

  - Instalar el DASK

    "https://docs.dask.org/en/latest/install.html"

  - API DASK

"https://docs.dask.org/en/latest/api.html"

- Aprendizaje automático DASK

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

- DASK Distributed Diagnostics

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

- Marco Y herramientas DE ML:
  - TensorFlow: Un marco de aprendizaje automático de código abierto para todos

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

  - Docker

    "https://docs.docker.com"

  - Kubernetes

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

  - Kubeflow

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

  - Servidor de portátiles Juppyter

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

TR-4896: Formación distribuida en Azure: Detección de carriles - diseño de soluciones

Muneer Ahmad y Verron Martina, NetApp Ronen dar, RUN:AI

Desde mayo de 2019, Microsoft proporciona un servicio de portal nativo de Azure para servicios de archivos NFS y SMB empresariales basados en la tecnología ONTAP de NetApp. Este desarrollo está impulsado por una asociación estratégica entre Microsoft y NetApp, y amplía aún más el alcance de los servicios de datos de ONTAP de primera calidad para Azure.

NetApp, un proveedor líder de servicios de datos en el cloud, se ha Unido para EJECUTARSE: IA, una empresa que virtualiza la infraestructura de IA, para permitir una experimentación de IA más rápida con un uso completo de la GPU. Esta alianza permite a los equipos acelerar la IA ejecutando muchos experimentos en paralelo, con un acceso rápido a los datos y aprovechando recursos informáticos ilimitados. EJECUTAR: La IA permite el uso completo de la GPU al automatizar la asignación de recursos y la arquitectura contrastada de Azure NetApp Files permite que cada experimento se ejecute a la máxima velocidad al eliminar las obstrucciones de la canalización de datos.

NetApp y EJECUTE: La IA ha Unido sus fuerzas para ofrecer a sus clientes una plataforma preparada para el futuro en su viaje a la IA en Azure. Desde el análisis y la computación de alto rendimiento (HPC) hasta las

decisiones autónomas (en las que los clientes pueden optimizar sus inversiones EN TECNOLOGÍA pagando solo por lo que necesitan y cuando lo necesitan), la alianza entre NetApp Y RUN: La IA ofrece una única experiencia unificada en el cloud de Azure.

Descripción general de la solución

En esta arquitectura, lo primordial es la parte de cálculo más intensivo del proceso de formación distribuida de detección de carriles POR inteligencia artificial o aprendizaje automático (ML). La detección de carriles es una de las tareas más importantes en la conducción autónoma, que ayuda a guiar los vehículos por la localización de las marcas de carril. Los componentes estáticos, como las marcas de carril, guían al vehículo a conducir por la carretera de forma interactiva y segura.

Los enfoques basados en la red neuronal convolucional (CNN) han empujado la comprensión y segmentación de la escena a un nuevo nivel. Aunque no funciona bien para objetos con estructuras largas y regiones que podrían ser ocluidas (por ejemplo, postes, sombra en el carril, etc.). La Red neuronal convolucional Espacial (SCNN) generaliza la CNN a un rico nivel espacial. Permite la propagación de información entre neuronas en la misma capa, lo que lo hace más adecuado para objetos estructurados como carriles, polos o camiones con oclusiones. Esta compatibilidad se debe a que la información espacial se puede reforzar y conserva la suavidad y la continuidad.

Es necesario inyectar miles de imágenes de escenas en el sistema para permitir que el modelo aprenda y distinga los diversos componentes del conjunto de datos. Estas imágenes incluyen el tiempo, día o noche, carreteras de varios carriles y otras condiciones de tráfico.

Para la formación, es necesario disponer de buena calidad y cantidad de datos. Una única GPU o varias GPU pueden tardar entre días y semanas para completar el entrenamiento. El entrenamiento con distribución de datos puede acelerar el proceso mediante varias GPU de varios nodos. Horovod es uno de esos marcos que concede entrenamiento distribuido pero la lectura de datos en clústeres de GPU puede ser un obstáculo. Azure NetApp Files proporciona un rendimiento elevado y una latencia baja constante ultrarrápida que proporciona funcionalidades de escalado horizontal y escalado vertical, de tal modo que se utilicen las GPU en la mejor capacidad computacional. Nuestros experimentos verificaron que todas las GPU del clúster se utilizan de media más del 96 % para entrenar la detección de carriles con SCNN.

Público objetivo

La ciencia de datos incorpora varias disciplinas EN TECNOLOGÍA y negocio, por lo que múltiples personas forman parte de nuestro público objetivo:

- Los científicos de datos necesitan la flexibilidad necesaria para utilizar las herramientas y las bibliotecas que prefieran.
- Los ingenieros de datos necesitan saber cómo fluyen los datos y dónde residen.
- Expertos en casos de uso de conducción autónoma.
- Administradores de cloud y arquitectos para configurar y gestionar recursos de cloud (Azure).
- Un ingeniero de DevOps necesita herramientas para integrar nuevas aplicaciones de IA/ML en sus canalizaciones de integración continua y de puesta en marcha continua (CI/CD).
- Los usuarios empresariales quieren tener acceso a aplicaciones de IA/ML.

En este documento, describimos cómo Azure NetApp Files, EJECUTA: IA y Microsoft Azure ayudan a cada uno de estos roles a aportar valor empresarial.

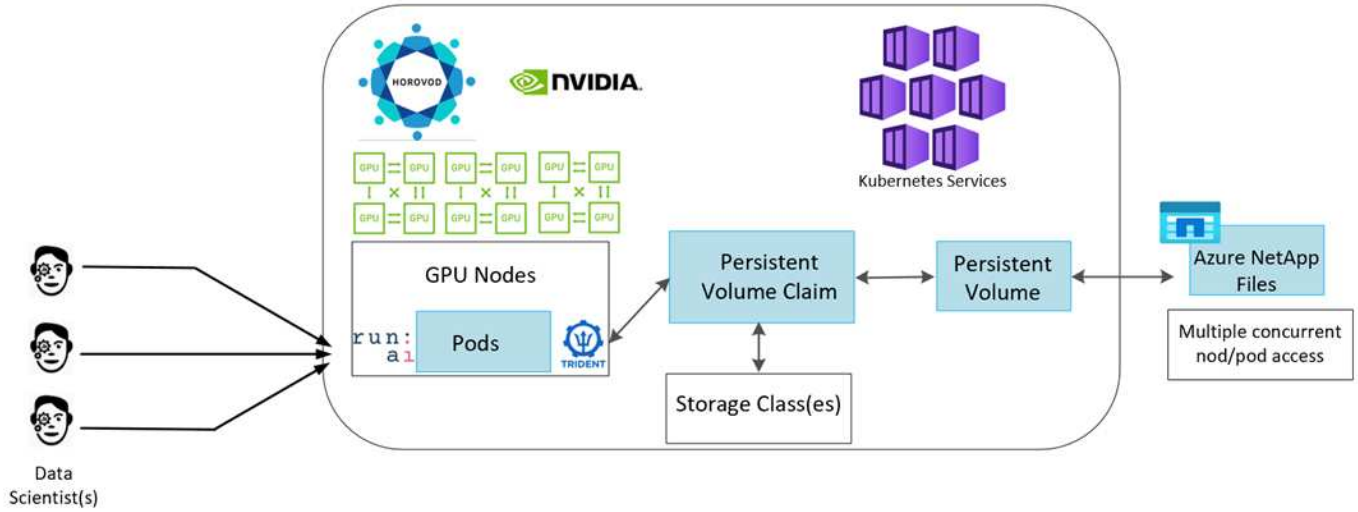
Tecnología de soluciones

Esta sección abarca los requisitos tecnológicos del caso práctico de detección de carriles al implementar una solución de formación distribuida a escala que se ejecuta completamente en el cloud de Azure. La siguiente figura muestra información general sobre la arquitectura de la solución.

Los elementos utilizados en esta solución son:

- Azure Kubernetes Service (AKS)
- SKU de Azure Compute con GPU de NVIDIA
- Azure NetApp Files
- EJECUCIÓN: IA
- Trident de NetApp

Los vínculos a todos los elementos mencionados aquí se enumeran en el "Información adicional" sección.



Requisitos de servicios y recursos cloud

En la siguiente tabla se enumeran los componentes de hardware necesarios para implementar la solución. Los componentes cloud que se usan en cualquier implementación de la solución pueden variar en función de las necesidades del cliente.

Cloud	Cantidad
AKS	Un mínimo de tres nodos de sistema y tres nodos de trabajo de GPU
Nodos del sistema de SKU de máquinas virtuales (VM)	Tres Standard_DS2_v2
Nodos de trabajo de GPU de VM SKU	Tres Standard_NC63_v3
Azure NetApp Files	Nivel estándar de 4 TB

Requisitos de software

En la siguiente tabla se enumeran los componentes de software necesarios para implementar la solución. Los

componentes de software que se usan en cualquier implementación de la solución pueden variar en función de las necesidades del cliente.

De NetApp	Versión u otra información
AKS - versión Kubernetes	1.18.14
EJECUCIÓN: CLI DE IA	v2.2.25
EJECUTAR:versión del operador de Kubernetes de orquestación de IA	1.0.109
Horovod	0.21.2
Trident de NetApp	20.01.1
Timón	3.0.0

Detección de carriles: Formación distribuida con RUN:AI

En esta sección se ofrecen detalles sobre la configuración de la plataforma para la realización de formación distribuida de detección de carriles a escala mediante EL orquestador de IA. Se trata de la instalación de todos los elementos de la solución y de la ejecución del trabajo de formación distribuido en dicha plataforma. EL versionado DE ML se completa utilizando SnapshotTM de NetApp vinculado A RUN: Experimentos de IA para conseguir la reproducibilidad de los datos y los modelos. LA creación de versiones DE ML desempeña un papel fundamental en el seguimiento de modelos, el intercambio de trabajo entre miembros del equipo, la reproducibilidad de resultados, la implementación de nuevas versiones de modelos a la producción y la procedencia de los datos. El control de versiones DE NetApp ML (Snapshot) puede capturar versiones puntuales de los datos, modelos entrenados y registros asociados con cada experimento. Cuenta con un amplio soporte en la API que facilita la integración con la plataforma DE IA RUN: Solo tiene que activar un evento basado en el estado de entrenamiento. También hay que capturar el estado de todo el experimento sin cambiar nada en el código o los contenedores que se ejecutan sobre Kubernetes (K8s).

Por último, este informe técnico se resume en la evaluación del rendimiento en varios nodos habilitados para GPU en toda la serie AKS.

Formación distribuida para el caso de uso de detección de carriles mediante el conjunto de datos TuSimple

En este informe técnico, la formación distribuida se realiza en el conjunto de datos TuSimple para la detección de carriles. Horovod se utiliza en el código de entrenamiento para realizar el entrenamiento con datos distribuidos en varios nodos de GPU simultáneamente en el clúster de Kubernetes a través de AKS. El código se presenta como imágenes de contenedor para la descarga y el procesamiento de datos de TuSimple. Los datos procesados se almacenan en volúmenes persistentes asignados por el complemento Trident de NetApp. Para el entrenamiento, se crea una imagen de contenedor más y utiliza los datos almacenados en volúmenes persistentes creados durante la descarga de los datos.

Para enviar el trabajo de datos y entrenamiento, use RUN: AI para orquestar la asignación y la gestión de recursos. RUN: AI le permite realizar operaciones de interfaz de paso de mensajes (MPI) que son necesarias para Horovod. Este diseño permite que varios nodos de GPU se comuniquen entre sí para actualizar los

pesos de entrenamiento después de cada lote de entrenamiento. También permite supervisar el entrenamiento a través de la interfaz de usuario y la CLI, para facilitar la supervisión del progreso de los experimentos.

NetApp Snapshot se integra dentro del código de entrenamiento y captura el estado de los datos y el modelo entrenado para cada experimento. Esta funcionalidad le permite realizar un seguimiento de la versión de los datos y el código que se usan y del modelo entrenado asociado generado.

Instalación e instalación de AKS

Para la configuración e instalación del clúster de AKS, vaya a. "Cree un clúster de AKS". A continuación, siga estos pasos:

1. Al seleccionar el tipo de nodos (ya sean nodos del sistema (CPU) o de trabajadores (GPU), seleccione lo siguiente:

   a. Añada el nodo del sistema principal llamado `agentpool` en la `Standard_DS2_v2` tamaño. Utilice los tres nodos predeterminados.

   b. Agregar nodo de trabajo `gpupool` con `the Standard_NC6s_v3` el tamaño del pool. Utilice un mínimo de tres nodos para los nodos GPU.

   La implementación tarda entre 5 y 10 minutos.

2. Una vez completada la implementación, haga clic en Connect to Cluster. Para conectarse al clúster AKS recién creado, instale la herramienta de línea de comandos Kubernetes desde su entorno local (portátil/PC). Visite "Instalar herramientas" Para instalarlo según el sistema operativo.
3. "Instale el CLI de Azure en su entorno local".
4. Para acceder al clúster AKS desde el terminal, primero introduzca `az login` y coloque las credenciales.
5. Ejecute los dos comandos siguientes:

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

6. Introduzca este comando en la CLI de Azure:

```
kubectl get nodes
```

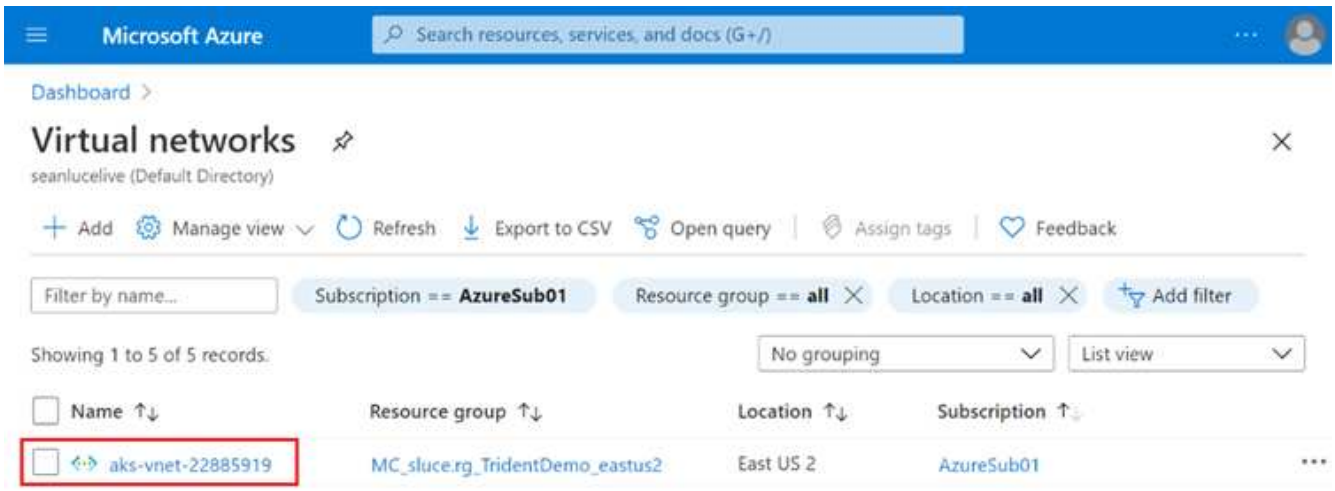
   Si los seis nodos están en funcionamiento como se ve aquí, su clúster de AKS estará listo y conectado a su entorno local.

```
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 ~ %
```

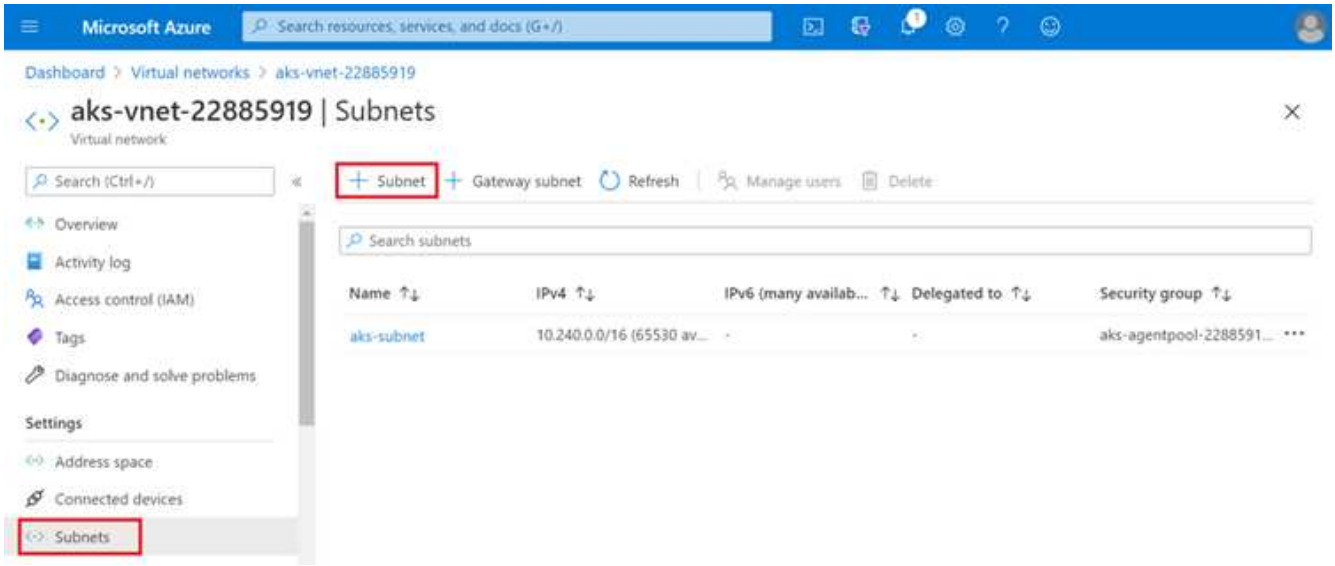
Cree una subred delegada para Azure NetApp Files

Para crear una subred delegada para Azure NetApp Files, siga esta serie de pasos:

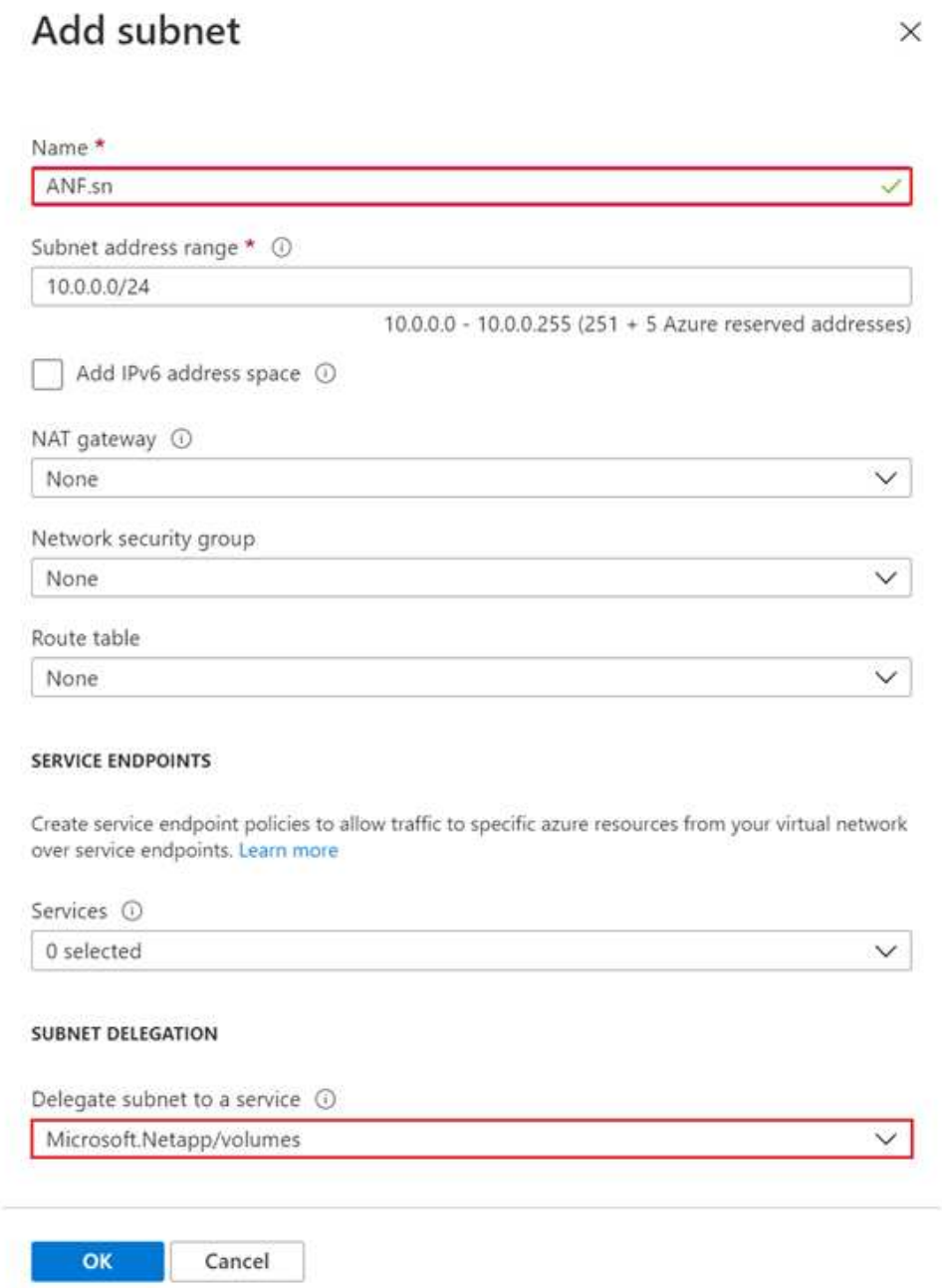
1. Acceda a redes virtuales en el portal de Azure. Busque la red virtual que acaba de crear. Debería tener un prefijo como el de los robles-vnet, como se ve aquí. Haga clic en el nombre de la red virtual.



2. Haga clic en subredes y seleccione +Subnet en la barra de herramientas superior.



3. Proporcione la subred con un nombre como `ANF.sn` Y en el encabezado Delegación de subred, seleccione Microsoft.NetApp/volumes. No cambie nada más. Haga clic en Aceptar.



Los volúmenes Azure NetApp Files se asignan al clúster de aplicaciones y se consumen como reclamaciones de volúmenes persistentes (RVP) en Kubernetes. A su vez, esta asignación nos proporciona la flexibilidad para asignar volúmenes a diferentes servicios, ya sea ordenadores portátiles Jupyter, funciones sin servidor, etc.

Los usuarios de servicios pueden consumir almacenamiento desde la plataforma de muchas maneras. Las principales ventajas de Azure NetApp Files son:

- Ofrece a los usuarios la capacidad de usar copias Snapshot.
- Permite a los usuarios almacenar grandes cantidades de datos en volúmenes de Azure NetApp Files.
- Obtenga las ventajas en el rendimiento de los volúmenes de Azure NetApp Files cuando ejecute sus

modelos en conjuntos de archivos de gran tamaño.

Configuración de Azure NetApp Files

Para completar la configuración de Azure NetApp Files, primero debe configurarlo como se describe en "Inicio rápido: Configure Azure NetApp Files y cree un volumen NFS".

Sin embargo, puede omitir los pasos para crear un volumen NFS para Azure NetApp Files a medida que creará volúmenes a través de Trident. Antes de continuar, asegúrese de que dispone de:

1. "Registrado para Azure NetApp Files y el proveedor de recursos de NetApp (a través de Azure Cloud Shell)".
2. "Se creó una cuenta en Azure NetApp Files".
3. "Configure un pool de capacidad" (Como mínimo, 4 TIB Standard o Premium, según sus necesidades).

Agrupación de la red virtual de AKS y la red virtual de Azure NetApp Files

A continuación, conecte la red virtual AKS (vnet) con Azure NetApp Files vnet siguiendo estos pasos:

1. En el cuadro de búsqueda de la parte superior del portal de Azure, escriba redes virtuales.
2. Haga clic en vnet aks- vnet-name y, a continuación, escriba peerings en el campo de búsqueda.
3. Haga clic en +Agregar e introduzca la información proporcionada en la siguiente tabla:

Campo	Valor o descripción
Nombre de enlace de relación entre iguales	aks-vnet-name_to_anf
SubscriptionId	Suscripción de la red virtual de Azure NetApp Files a la que se está creando una relación de paridad
Partner de vnet peering	Red virtual de Azure NetApp Files

> Deje todas las secciones que no sean asteriscos por defecto

4. Haga clic en AGREGAR o en Aceptar para agregar la conexión a la red virtual.

Si desea más información, visite "Crear, cambiar o eliminar una conexión de red virtual entre iguales".

Trident

Trident es un proyecto de código abierto que NetApp mantiene para el almacenamiento persistente en contenedores para aplicaciones. Trident se ha implementado como una controladora de aprovisionamiento externa que se ejecuta como un "pod", supervisando volúmenes y automatizando totalmente el proceso de aprovisionamiento.

Trident de NetApp permite una integración sin problemas con K8S mediante la creación y el montaje de volúmenes persistentes para almacenar conjuntos de datos de entrenamiento y modelos entrenados. Esta funcionalidad facilita a los científicos e ingenieros de datos el uso de K8 sin los problemas de almacenar y gestionar manualmente conjuntos de datos. Trident también elimina la necesidad que tienen los científicos de datos de aprender a gestionar nuevas plataformas de datos a medida que integra las tareas relacionadas con la gestión de datos a través de la integración lógica de las API.

Instale Trident

Para instalar el software Trident, realice los pasos siguientes:

1. "Primero instale el timón".
2. Descargue y extraiga el instalador de 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. Cambie el directorio a. `trident-installer`.

```
cd trident-installer
```

4. Copiar `tridentctl` a un directorio del sistema `$PATH`.

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

5. Instale Trident en el clúster K8s con Helm:
   a. Cambiar el directorio al directorio del timón.

   ```
   cd helm
   ```

   b. Instale Trident.

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

   c. Compruebe el estado de Trident pods de la manera habitual K8s:

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

   d. Si todos los pods están ya en funcionamiento, se instala Trident y se podrá seguir avanzando.

Configure el back-end de Azure NetApp Files y la clase de almacenamiento

Para configurar el back-end de Azure NetApp Files y la clase de almacenamiento, complete los siguientes pasos:

1. Vuelva al directorio inicial.

```
cd ~
```

2. Clone el "repositorio de proyectos" `lane-detection-SCNN-horovod`.
3. Vaya a la `trident-config` directorio.

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

4. Crear un principio de servicio de Azure (el principio del servicio es cómo Trident se comunica con Azure para acceder a sus recursos de Azure NetApp Files).

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

El resultado debería ser como el ejemplo siguiente:

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

5. Cree Trident `backend json` archivo.
6. Con el editor de texto preferido, complete los siguientes campos de la tabla siguiente dentro de la `anf-backend.json` archivo.

Campo	Valor
SubscriptionId	Su ID de suscripción de Azure
ID de tenantID	Su ID de inquilino de Azure (de la salida de az ad sp en el paso anterior)
ID del Cliente	Su AppID (de la salida de az ad sp en el paso anterior)
ClientSecret	Su contraseña (de la salida de az ad sp en el paso anterior)

El archivo debería tener el siguiente ejemplo:

```
{
    "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. Indique a Trident que cree el back-end de Azure NetApp Files en la `trident` espacio de nombres, utilizar `anf-backend.json` como el archivo de configuración de la siguiente manera:

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

8. Cree la clase de almacenamiento:

   a. Los usuarios de K8 aprovisionan volúmenes mediante el uso de EVs que especifican una clase de almacenamiento por nombre. Indique a K8S que cree una clase de almacenamiento `azurenetappfiles` Que hará referencia al back-end de Azure NetApp Files creado en el paso anterior utilizando lo siguiente:

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

   b. Compruebe que la clase de almacenamiento se haya creado mediante el siguiente comando:

   ```
   kubectl get sc azurenetappfiles
   ```

   El resultado debería ser como el ejemplo siguiente:

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

Ponga en marcha y configure componentes snapshot para volúmenes en AKS

Si el clúster no viene preinstalado con los componentes Snapshot de volumen correctos, puede realizar la

instalación manual de estos componentes ejecutando los siguientes pasos:

> AKS 1.18.14 no tiene una controladora Snapshot instalada previamente.

1. Instale los CRD de la versión beta de instantánea utilizando los siguientes comandos:

```
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. Instale el controlador Snapshot con los siguientes documentos de 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. Configurar K8s `volumesnapshotclass`: Antes de crear una instantánea de volumen, a. "clase de snapshot de volumen" debe estar configurado. Cree una clase de copia Snapshot de volumen para Azure NetApp Files y utilícela para crear versiones ML con la tecnología Snapshot de NetApp. Cree `volumesnapshotclass netapp-csi-snapclass` y configúrelo como predeterminado "volumesnapshotclass "como tal:

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

   El resultado debería ser como el ejemplo siguiente:

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

4. Compruebe que la clase de copia Snapshot de volumen se haya creado con el siguiente comando:

```
kubectl get volumesnapshotclass
```

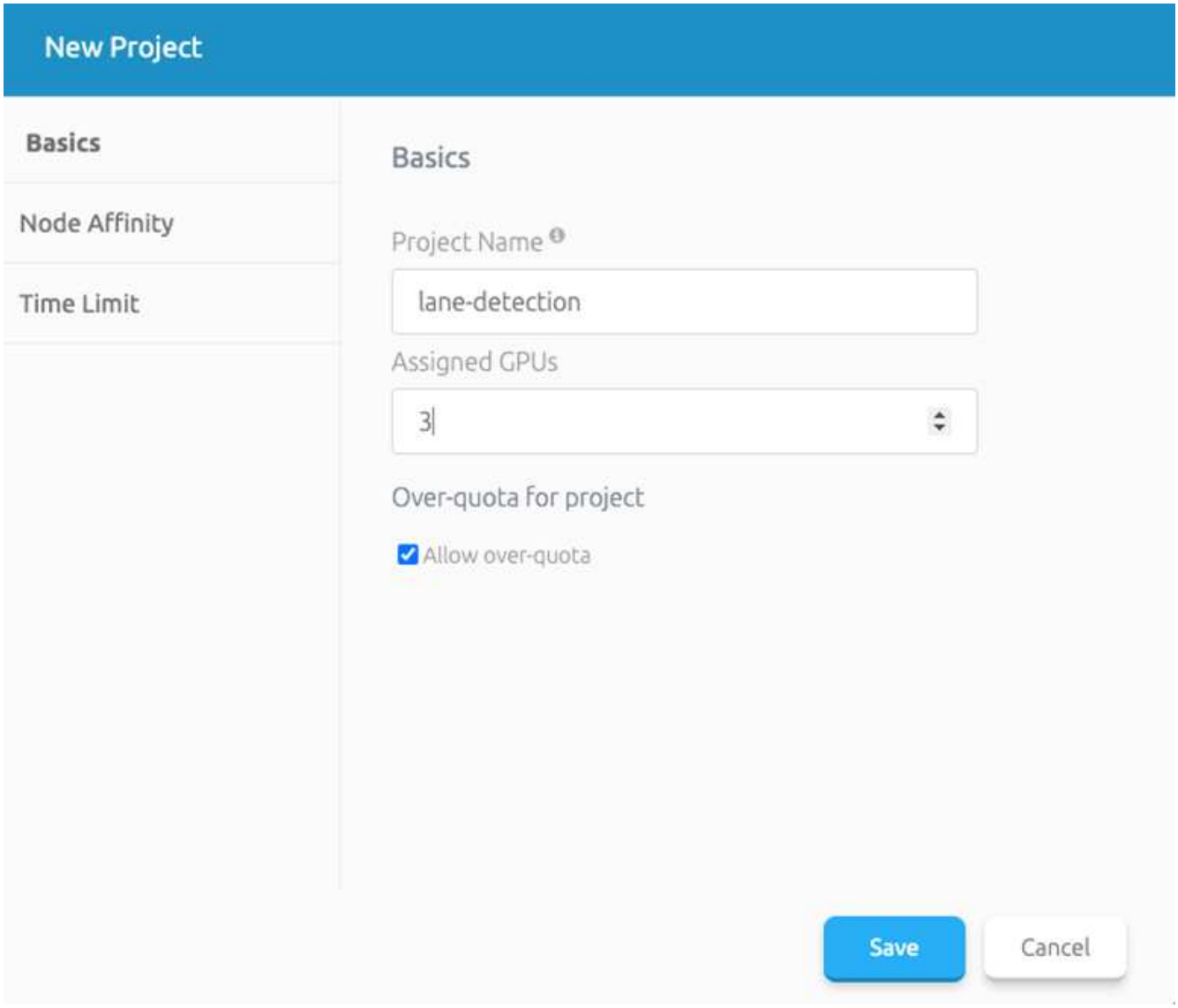
   El resultado debería ser como el ejemplo siguiente:

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

EJECUCIÓN:instalación de IA

Para instalar RUN:AI, realice los siguientes pasos:

1. "Instalar EJECUTAR:clúster AI en AKS".
2. Vaya a app.runai.ai, haga clic en Crear nuevo proyecto y asigne un nombre a la detección de carriles. Creará un espacio de nombres en un clúster K8s a partir de `runai-` seguido del nombre del proyecto. En este caso, el espacio de nombres creado sería la detección de pistas en ejecución.



3. "INSTALAR RUN:AI CLI".
4. En el terminal, establezca la detección de carriles como UNA EJECUCIÓN predeterminada: Proyecto de IA mediante el siguiente comando:

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

El resultado debería ser como el ejemplo siguiente:

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

5. Crear ClusterRole y ClusterRoleBinding para el espacio de nombres del proyecto (por ejemplo, `lane-detection)` por lo tanto, la cuenta de servicio predeterminada que pertenece a `runai-lane-detection` el espacio de nombres tiene permiso para ejecutar `volumesnapshot` operaciones durante la ejecución de trabajos:

   a. Enumere los espacios de nombres para comprobarlo `runai-lane-detection` existe usando este comando:

   ```
   kubectl get namespaces
   ```

   El resultado debería aparecer como el ejemplo siguiente:

```
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. Crear función de clúster `netappsnapshot` Y ClusterRoleBinding `netappsnapshot` con los siguientes comandos:

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

Descargue y procese el conjunto de datos de TuSimple como RUN:AI job

El proceso para descargar y procesar el conjunto de datos TuSimple COMO UNA EJECUCIÓN: El trabajo de IA es opcional. Se trata de los siguientes pasos:

1. Cree y empuje la imagen del docker, o omita este paso si desea utilizar una imagen del docker existente (por ejemplo, `muneer7589/download-tusimple:1.0)`

   a. Cambie al directorio principal:

```
cd ~
```

b. Vaya al directorio de datos del proyecto `lane-detection-SCNN-horovod`:

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

c. Modificar `build_image.sh` script de shell y cambiar el repositorio de docker a la suya. Por ejemplo, sustituir `muneer7589` con el nombre de repositorio de docker. También puede cambiar el nombre y LA ETIQUETA de la imagen del docker (por ejemplo `download-tusimple` y.. `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. Ejecute el script para crear la imagen de docker y empújela al repositorio de docker mediante los siguientes comandos:

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

2. Envíe el TRABAJO RUN: AI para descargar, extraer, preprocesar y almacenar el conjunto de datos de detección de carriles TuSimple en un `pvc`Creado dinámicamente por Trident de NetApp:

   a. Use los siguientes comandos para enviar LA EJECUCIÓN: AI job:

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

b. Introduzca la información de la siguiente tabla para enviar EL trabajo RUN:AI:

Campo	Valor o descripción
-name	Nombre del trabajo
-pvc	PVC del formato [StorageClassName]:Size:ContainerMountPath en el envío de trabajos anterior, está creando un PVC basado en demanda usando Trident con archivos azurenetappfiles de la clase de almacenamiento. La capacidad de volumen persistente aquí es 100Gi y está montada en path /mnt.
-imagen	Imagen de Docker que se utilizará al crear el contenedor para este trabajo

El resultado debería ser como el ejemplo siguiente:

```
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. Enumere los trabajos ENVIADOS RUN:AI.

```
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. Compruebe los registros de trabajos enviados.

```
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
```

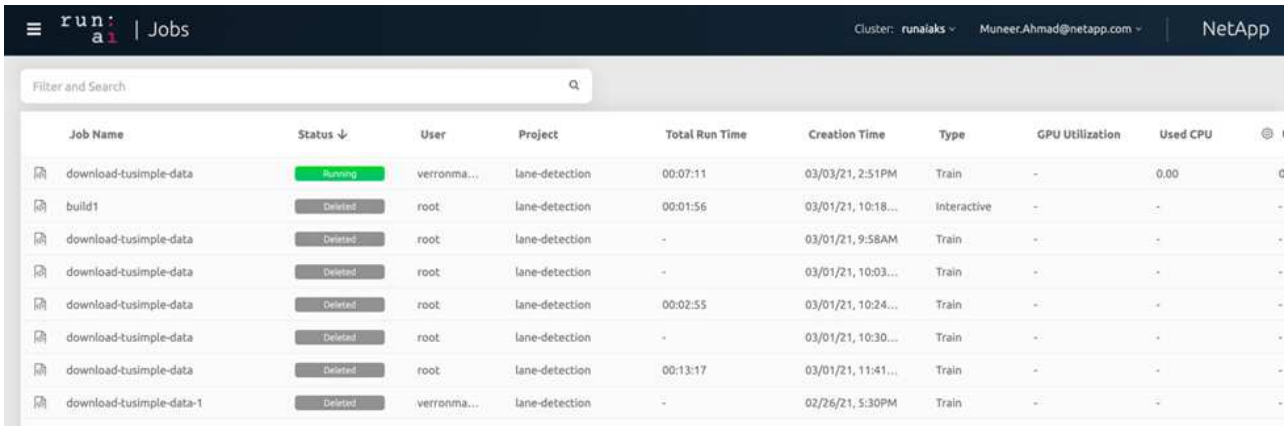
e. Enumere la `pvc` creado. Utilice esto `pvc` comando para la formación en el siguiente paso.

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

El resultado debería ser como el ejemplo siguiente:

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

a. Compruebe el trabajo EN RUN: AI UI (o. `app.run.ai`).



Realice la formación de detección de carriles distribuidos con Horovod

El entrenamiento de detección de carriles distribuidos con Horovod es un proceso opcional. Sin embargo, estos son los pasos implicados:

1. Cree y empuje la imagen del docker o omita este paso si desea utilizar la imagen del docker existente (por ejemplo, `muneer7589/dist-lane-detection:3.1)`:

   a. Cambie al directorio inicial.

   ```
   cd ~
   ```

   b. Vaya al directorio del proyecto `lane-detection-SCNN-horovod.`

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

   c. Modifique el `build_image.sh` script de shell y cambie el repositorio de docker a la suya (por ejemplo, reemplace `muneer7589` con el nombre del repositorio del docker). También puede cambiar el nombre y LA ETIQUETA de la imagen del docker (`dist-lane-detection` y.. `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. Ejecute la secuencia de comandos para crear la imagen de docker y empújela al repositorio de docker.

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

2. Enviar LA CARRERA: AI job para llevar a cabo la formación distribuida (MPI):

a. Uso de envío DE LA EJECUCIÓN: La inteligencia artificial para la creación automática de la RVP en el paso anterior (para la descarga de datos) solo le permite tener acceso a RWO, que no permite que varios POD o nodos accedan al mismo RVP para el entrenamiento distribuido. Actualice el modo de acceso a ReadWriteMany y utilice el parche Kubernetes para hacerlo.

b. En primer lugar, ejecute el siguiente comando para obtener el nombre del volumen de la RVP:

```
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. Aplicar patches al modo de acceso de volumen y actualización a ReadWriteMany (reemplace el nombre del volumen por el suyo en el siguiente comando):

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

d. Envíe el TRABAJO RUN: AI MPI para la ejecución del trabajo de formación distribuida utilizando la información de la tabla siguiente:

```
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	Valor o descripción
nombre	Nombre del trabajo de formación distribuido
gran km	Monte un dispositivo /dev/sm de gran tamaño. Es un sistema de archivos compartidos montado en la RAM y proporciona memoria compartida lo suficientemente grande para que varios trabajadores de CPU procesen y carguen lotes en la RAM de la CPU.
procesos	Número de procesos de formación distribuidos
gpu	Número de GPU/procesos que se van a asignar para la tarea en esta tarea, hay tres procesos de trabajo de GPU (--process=3), cada uno asignado con una única GPU (--gpu 1)
rvp	Utilice el volumen persistente existente (pvc-download-tusimple-data-0) creado por el trabajo anterior (download-tusimple-data) y se monta en la ruta /mnt
imagen	Imagen de Docker que se utilizará al crear el contenedor para este trabajo
Defina las variables de entorno que se van a establecer en el contenedor	
USE_WORKERS	Al establecer el argumento en true, se activa la carga de datos multiproceso
NÚM_TRABAJADORES	Número de procesos de trabajo del cargador de datos
TAMAÑO_LOTE	Tamaño de lote de entrenamiento
VALOR_USO	Establecer el argumento en true permite la validación
VAL_BATCH_SIZE	Tamaño del lote de validación

Campo	Valor o descripción
ENABLE_SNAPSHOT	Establecer el argumento en true permite tomar instantáneas de datos y modelos entrenados para el control de versiones EN ML
NOMBRE_PVC	Nombre de la rvp de la que se va a realizar una instantánea. En la presentación de trabajos anterior, está tomando una instantánea de pvc-download-tusimple-data-0, que consta de conjuntos de datos y modelos entrenados

El resultado debería ser como el ejemplo siguiente:

```
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. Enumera el trabajo enviado.

```
runai list jobs
```

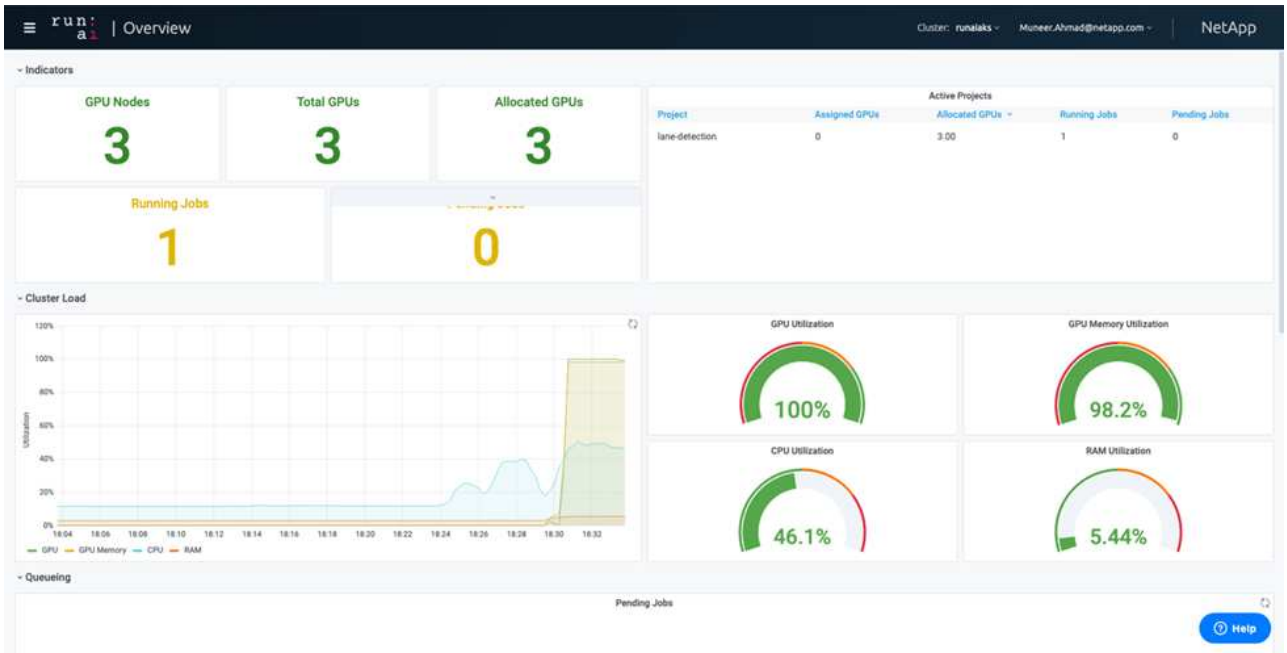
```
NAME                          STATUS     AGE  NODE        IMAGE                              TYPE   PROJECT         USER           GPUs Allocated (Requested)  PODs
 SERVICE URL(S)
download-tusimple-data        Succeeded  1d                 muneer7589/download-tusimple:1.0   Train  lane-detection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

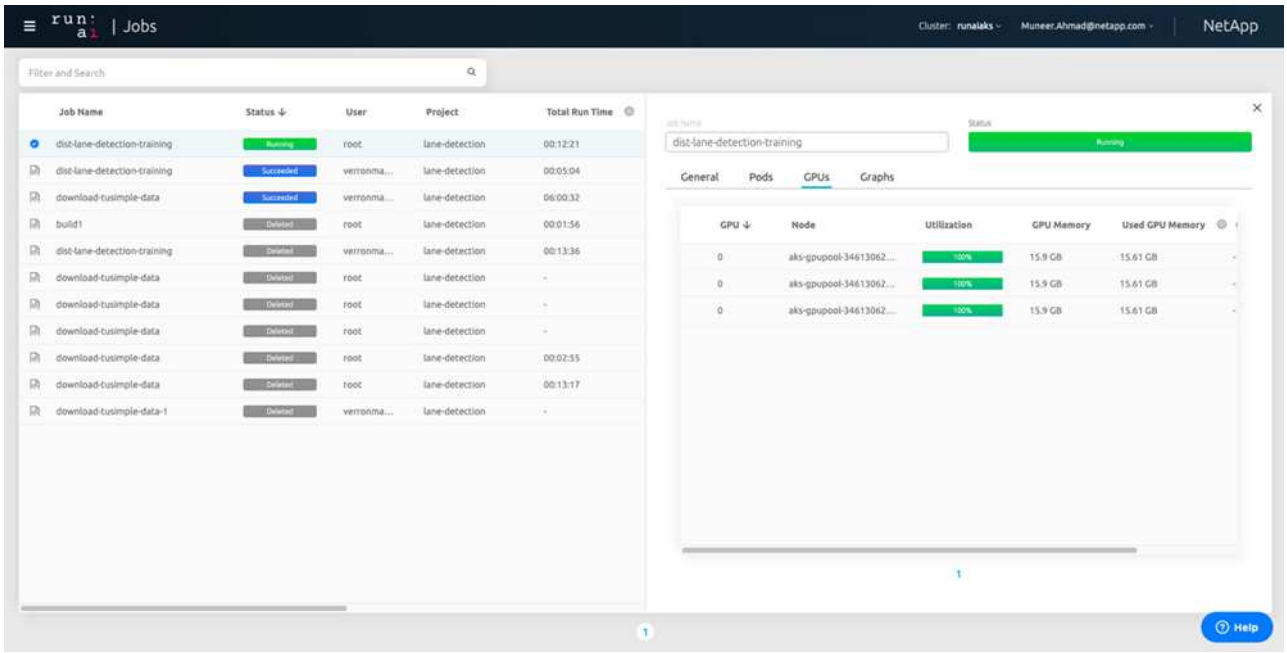
f. Registros de trabajos enviados:

```
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. Compruebe la tarea de entrenamiento EN EJECUCIÓN: Interfaz gráfica de usuario de IA (o app.runai.ai): RUN: Consola de IA, como se puede ver en las siguientes figuras. La primera figura detalla tres GPU asignadas para el trabajo de entrenamiento distribuido, repartidos en tres nodos en AKS, y la segunda EJECUCIÓN: Trabajos de IA:





h. Una vez finalizada la formación, compruebe la copia de Snapshot de NetApp que se creó y vinculado con RUN: Trabajo de IA.

```
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
```

Restaure datos de la copia Snapshot de NetApp

Para restaurar datos de la copia Snapshot de NetApp, complete los siguientes pasos:

1. Cambie al directorio inicial.

```
cd ~
```

2. Vaya al directorio del proyecto `lane-detection-SCNN-horovod`.

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

3. Modificar `restore-snaphot-pvc.yaml` y actualícelo `dataSource` `name` Campo de la copia Snapshot desde la que desea restaurar datos. También puede cambiar el nombre del PVC al que se restaurarán los datos, en este ejemplo su `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. Cree una nueva RVP mediante `restore-snapshot-pvc.yaml`.

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

   El resultado debería ser como el ejemplo siguiente:

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

5. Si desea utilizar los datos recién restaurados para el entrenamiento, el envío de trabajos sigue siendo el mismo que antes; sólo sustituya el `PVC_NAME` con el restaurado `PVC_NAME` al enviar el trabajo de formación, como se muestra en los siguientes comandos:

```
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"
```

Evaluación del rendimiento

Para mostrar la escalabilidad lineal de la solución, se han realizado pruebas de rendimiento para dos supuestos: Una GPU y tres GPU. La asignación de GPU, la utilización de la GPU y la memoria, se han capturado diferentes métricas de uno y tres nodos durante el entrenamiento en el conjunto de datos de detección de carriles TuSimple. Los datos se incrementan cinco veces sólo por analizar la utilización de los recursos durante los procesos de entrenamiento.

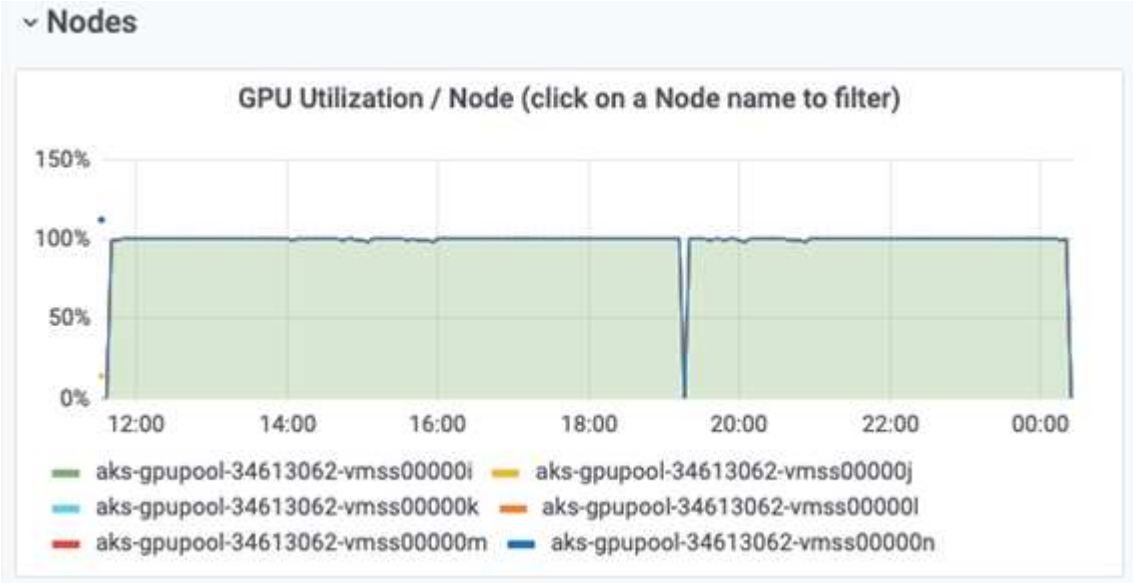
La solución permite a los clientes comenzar con un conjunto de datos pequeño y unas pocas GPU. Cuando aumentan la cantidad de datos y la demanda de GPU, los clientes pueden escalar horizontalmente de forma dinámica los terabytes del nivel estándar y escalar rápidamente hasta el nivel Premium para obtener el cuádruple de rendimiento por terabyte sin necesidad de mover datos. Este proceso se explica más detalladamente en la sección, "Niveles de servicio de Azure NetApp Files".

El tiempo de procesamiento en una GPU era de 12 horas y 45 minutos. El tiempo de procesamiento en tres GPU en tres nodos era de aproximadamente 4 horas y 30 minutos.

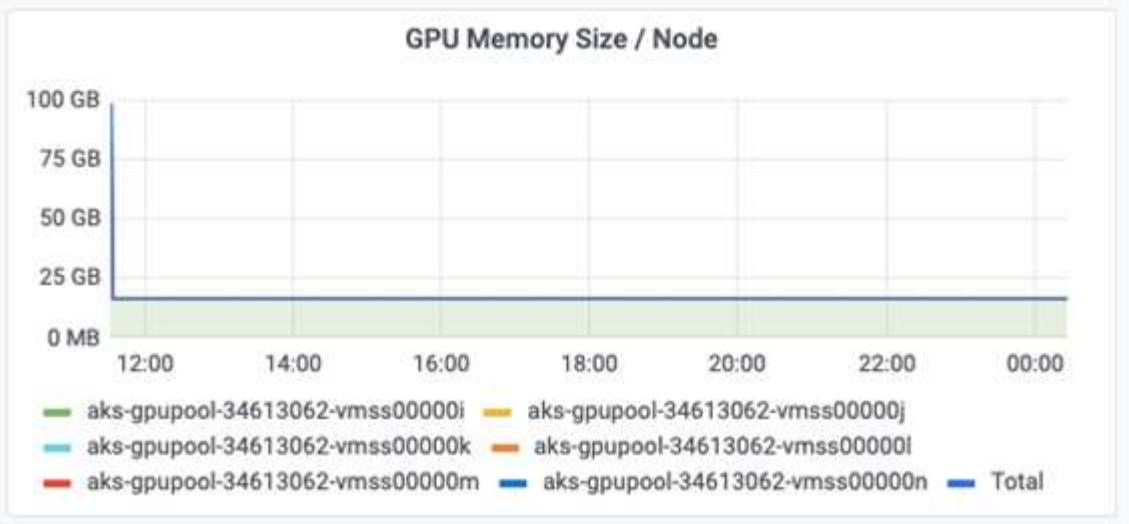
Las cifras que se muestran a lo largo del resto de este documento muestran ejemplos de rendimiento y escalabilidad basados en las necesidades empresariales individuales.

La siguiente figura muestra la asignación de 1 GPU y la utilización de memoria.

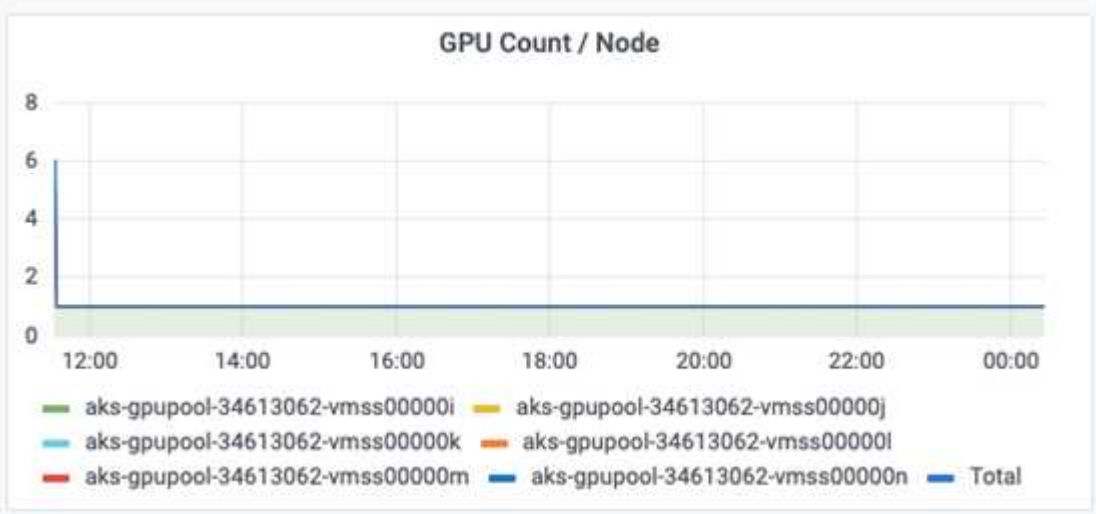
La siguiente figura ilustra el uso de GPU de un solo nodo.



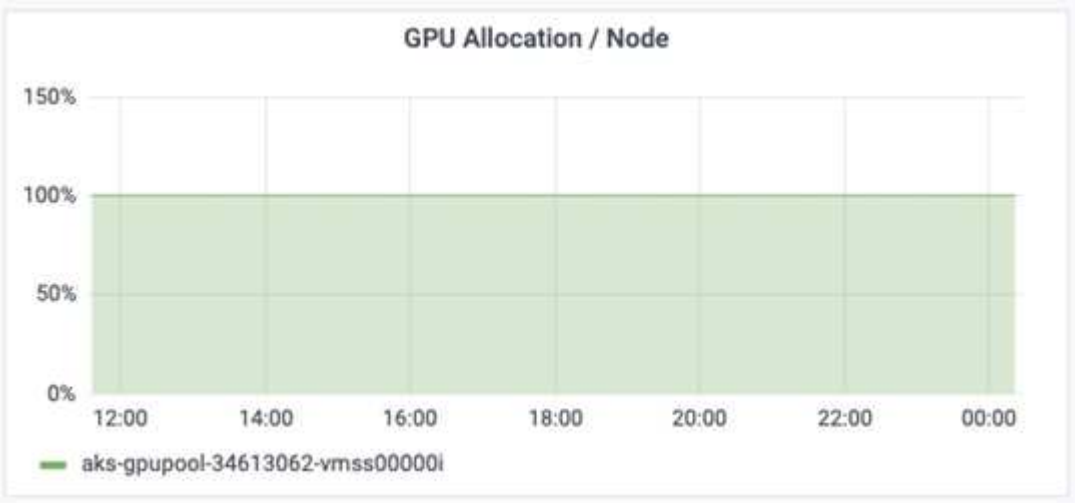
La siguiente figura ilustra el tamaño de la memoria de un solo nodo (16 GB).



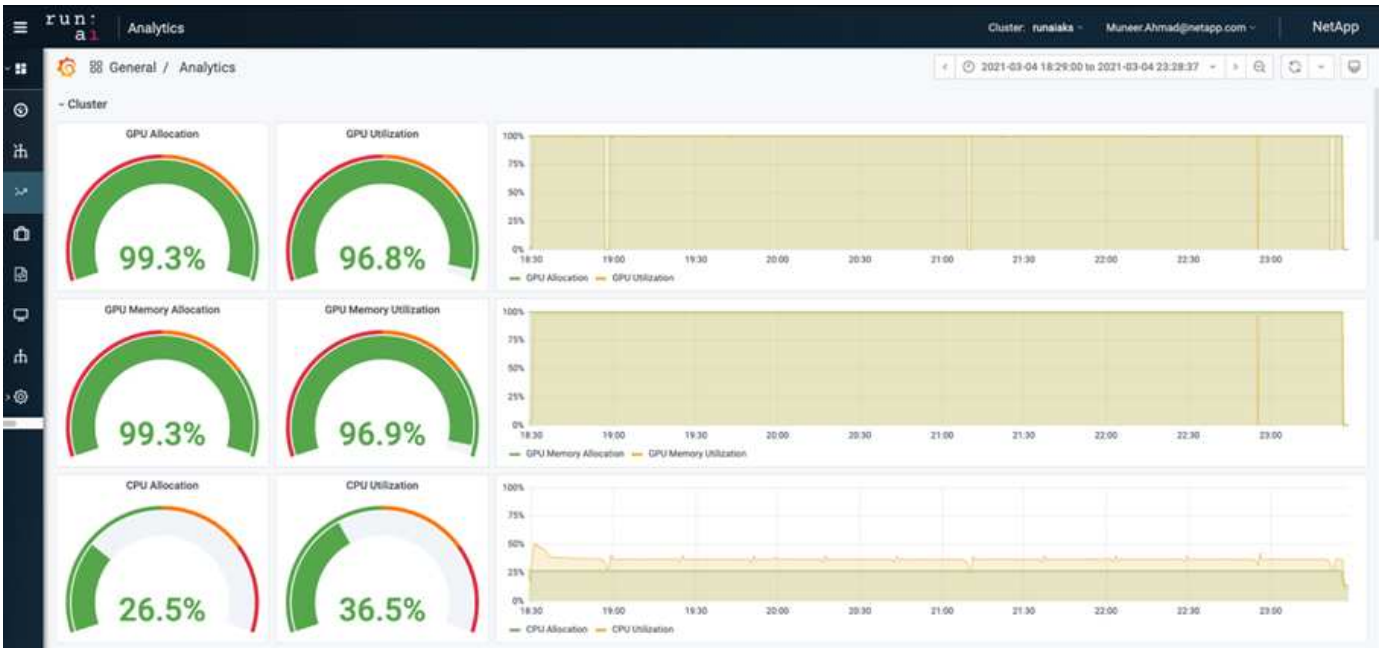
La siguiente figura muestra el número de GPU de nodo único (1).



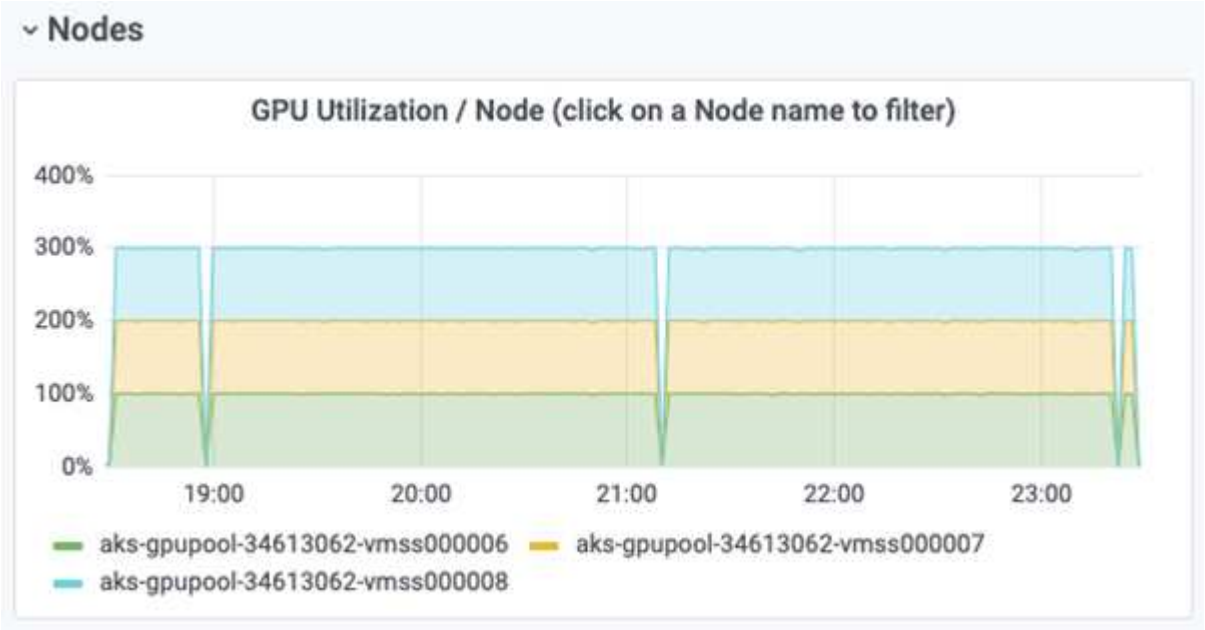
La siguiente figura muestra la asignación de GPU de un solo nodo (%).



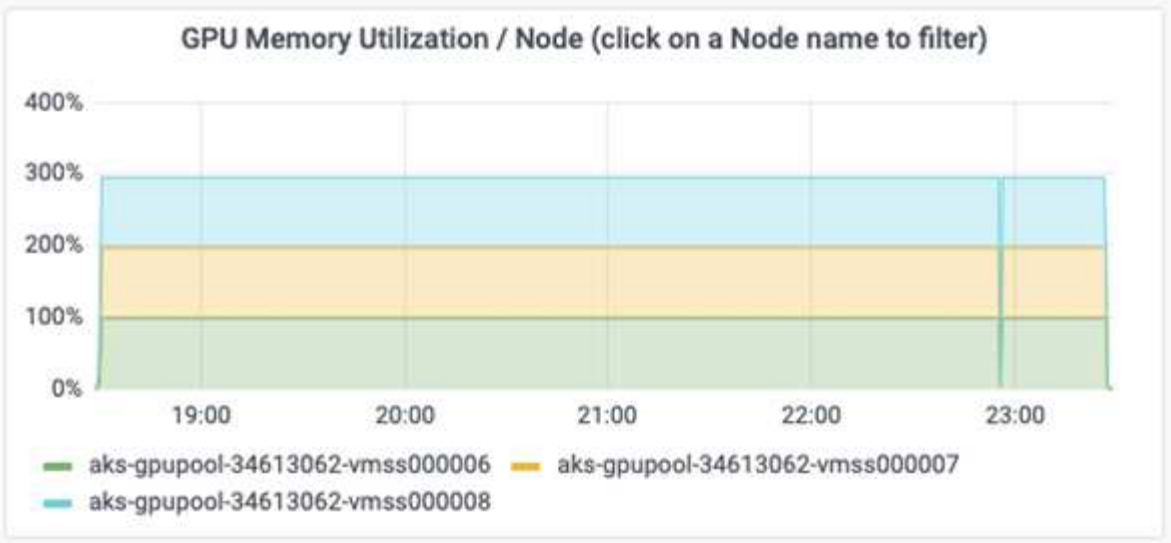
La siguiente figura muestra tres GPU en tres nodos: Asignación de GPU y memoria.



La siguiente figura muestra tres GPU en la utilización de tres nodos (%).



La siguiente figura muestra tres GPU en una utilización de la memoria de tres nodos (%).



Niveles de servicio de Azure NetApp Files

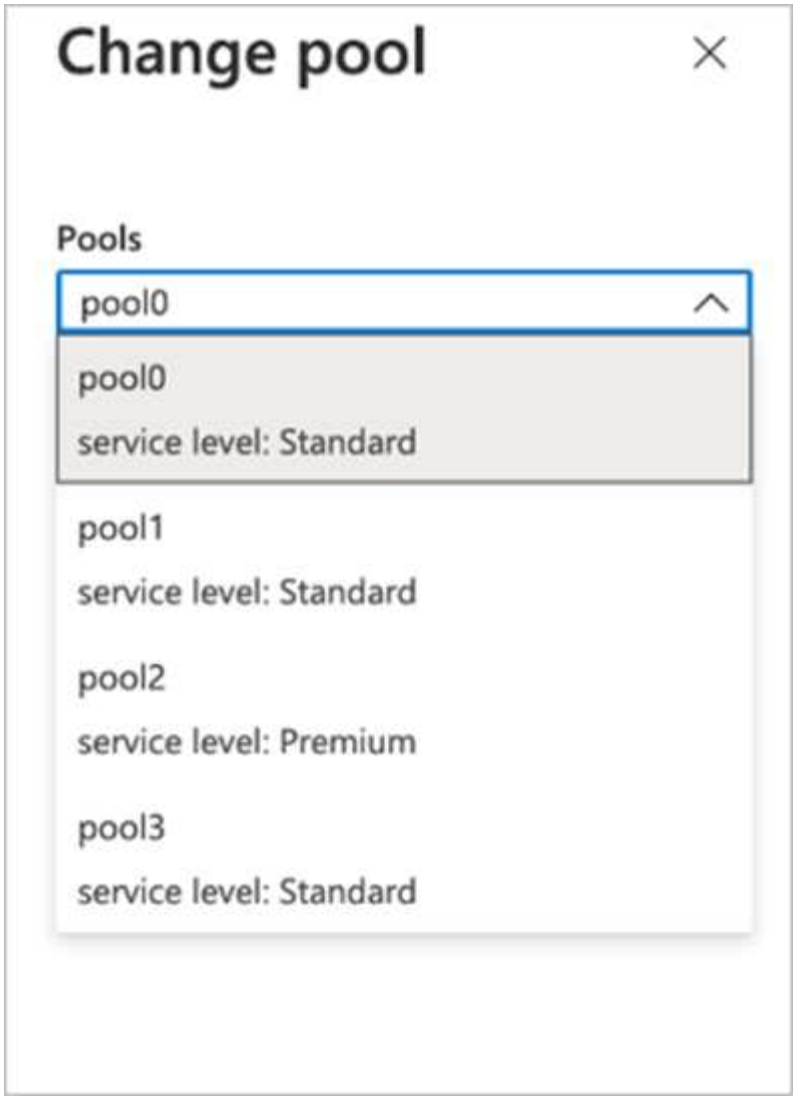
Es posible cambiar el nivel de servicio de un volumen existente si se mueve el volumen a otro pool de capacidad que utiliza "nivel de servicio" se desea para el volumen. Este cambio de nivel de servicio existente del volumen no requiere la migración de los datos. Además, no afecta el acceso al volumen.

Cambie dinámicamente el nivel de servicio de un volumen

Para cambiar el nivel de servicio de un volumen, siga estos pasos:

1. En la página Volumes, haga clic con el botón derecho en el volumen cuyo nivel de servicio desea cambiar. Seleccione Cambiar pool.

2. En la ventana Cambiar pool, seleccione el pool de capacidad al que desea mover el volumen. A continuación, haga clic en Aceptar.



Automatizar el cambio de nivel de servicio

El cambio de nivel de servicio dinámico se encuentra actualmente en la vista previa pública, pero no está habilitado de forma predeterminada. Para activar esta función en la suscripción a Azure, siga estos pasos proporcionados en el documento " "Cambie dinámicamente el nivel de servicio de un volumen"."

- También puede utilizar los siguientes comandos para Azure: CLI. Para obtener más información sobre cómo cambiar el tamaño del pool de Azure NetApp Files, visite "volumen de archivos de az netapparchivo: Gestione los recursos de volúmenes del Azure NetApp Files (ANF)".

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

- La `set- aznetappfilesvolumepool` El cmdlet que se muestra aquí puede cambiar el pool de un volumen Azure NetApp Files. Para obtener más información sobre el cambio del tamaño del pool de volúmenes y Azure PowerShell, visite "Cambiar el pool de un volumen Azure NetApp Files".

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

Conclusión

NetApp and RUN: IA se ha asociado en la creación de este informe técnico para mostrar las funcionalidades únicas de la Azure NetApp Files junto con LA EJECUCIÓN: Plataforma de IA para simplificar la orquestación de las cargas de trabajo de IA. Este informe técnico proporciona una arquitectura de referencia para optimizar el proceso de canalizaciones de datos y orquestación de cargas de trabajo para la formación en la detección de carriles distribuidos.

En conclusión, con respecto a la formación distribuida a escala (especialmente en un entorno de cloud público), la orquestación de recursos y el componente de almacenamiento son una parte fundamental de la solución. Asegurarse de que la gestión de datos nunca afecta al procesamiento de varias GPU, por lo que se traduce en una utilización óptima de los ciclos de la GPU. Por lo tanto, hacer que el sistema sea lo más rentable posible para fines de capacitación distribuidos a gran escala.

El Data Fabric que ofrece NetApp supera el reto al permitir a los científicos e ingenieros de datos conectarse entre sí tanto en las instalaciones como en el cloud disponer de datos síncronos sin realizar ninguna intervención manual. En otras palabras, el tejido de datos suaviza el proceso de gestión del flujo de trabajo de IA expandiéndose por varias ubicaciones. También facilita la disponibilidad de datos basada en demanda al acercar los datos a la computación y realizar análisis, formación y validación donde y cuando sea necesario. Esta funcionalidad no solo permite la integración de datos, sino también la protección y seguridad de toda la canalización de datos.

Información adicional

Si quiere más información sobre el contenido de este documento, consulte los siguientes documentos o sitios web:

- Conjunto de datos: TuSimple

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

- Arquitectura de redes de aprendizaje profundo: Red neuronal convolucional espacial

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

- Marco de entrenamiento de aprendizaje profundo distribuido: Horovod

  "https://horovod.ai/"

- EJECUCIÓN: Solución de orquestación de contenedores de IA: EJECUCIÓN: Introducción del producto de IA

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

- EJECUCIÓN: Documentación de instalación de IA

  "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"

- Envío de trabajos EN EJECUCIÓN: Interfaz de línea de comandos de IA

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

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

- Recursos de cloud para Azure: Azure NetApp Files

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

- Servicio Azure Kubernetes

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

- SKU de Azure VM

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

- Azure VM con SKU de GPU

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

- Trident de NetApp

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

- Data Fabric con tecnología de NetApp

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

- Documentación de productos de NetApp

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

TR-4841: Sistema operativo de IA para el cloud híbrido con almacenamiento en caché de datos

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

El crecimiento exponencial de los datos y el crecimiento exponencial DE LOS ML y la IA se han convergido para crear una economía de zettabytes con retos exclusivos de desarrollo e implementación.

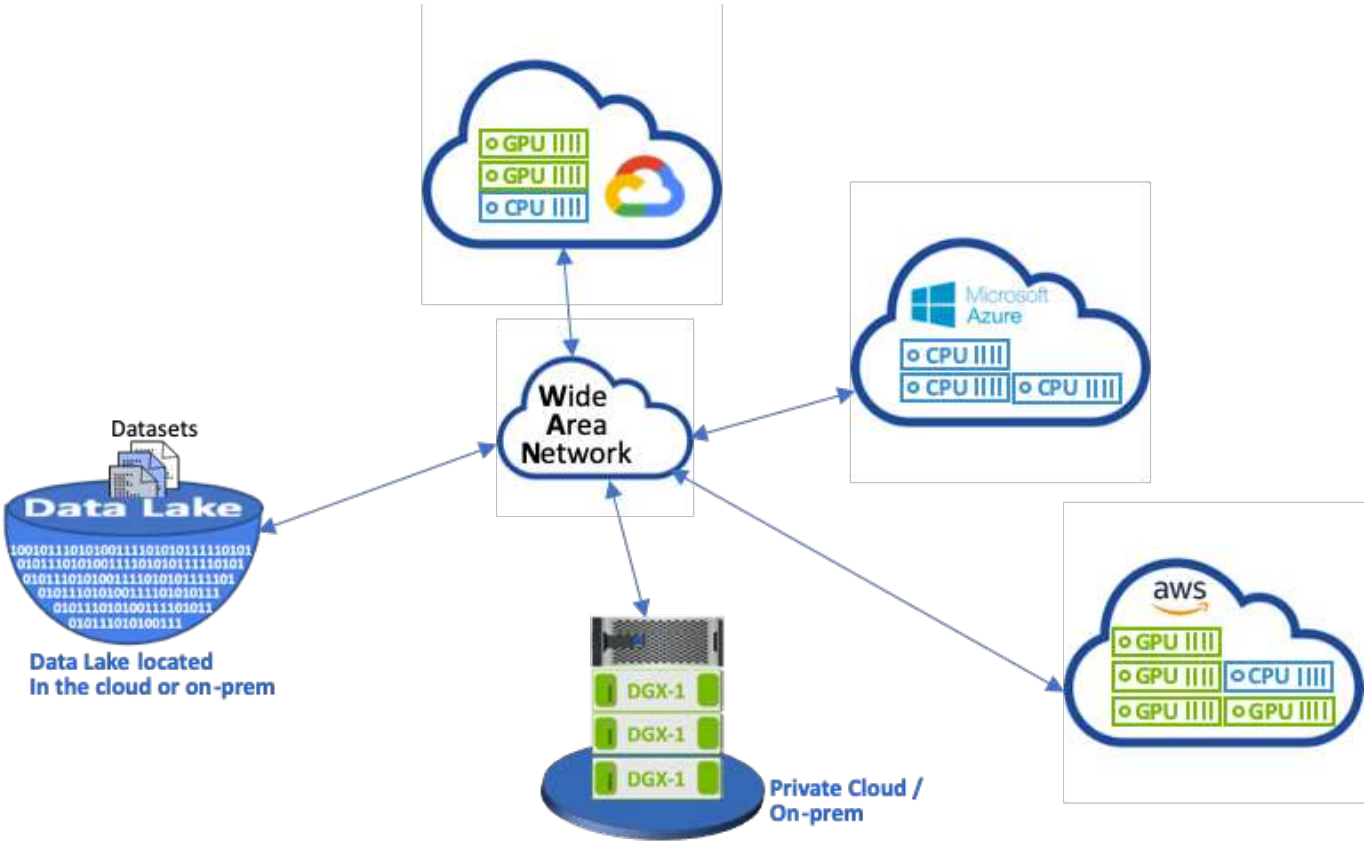
A pesar de que los modelos DE ML consumen gran cantidad de datos y requieren un almacenamiento de datos de alto rendimiento cerca de los recursos informáticos, en la práctica no es tan sencillo implementar este modelo, especialmente con instancias de cloud híbrido y computación elástica. Normalmente, se almacenan cantidades masivas de datos en lagos de datos de bajo coste, en los que los recursos informáticos

de IA de alto rendimiento, como las GPU, no pueden acceder a ellos de forma eficiente. Este problema se agrava en una infraestructura de cloud híbrido, en la que algunas cargas de trabajo funcionan en el cloud y otras están ubicadas en las instalaciones o en un entorno diferente de informática de alto rendimiento.

En este documento, presentamos una nueva solución que permite a los profesionales DE TECNOLOGÍA y a los ingenieros de datos crear una verdadera plataforma de IA de cloud híbrido con un concentrador de datos con topología que permite a los científicos de datos crear de forma instantánea y automática una caché de sus conjuntos de datos cerca de sus recursos informáticos, dondequiera que se encuentren. Como resultado, no solo se puede realizar el entrenamiento con modelos de alto rendimiento, sino que se obtienen ventajas adicionales, como la colaboración de varios profesionales de la IA, que tienen acceso inmediato a las cachés de los conjuntos de datos, las versiones y los linajes dentro de un concentrador de versiones del conjunto de datos.

Descripción general de casos de uso y declaración de problemas

Los conjuntos de datos y las versiones de conjuntos de datos normalmente se encuentran en un lago de datos, como el almacenamiento basado en objetos StorageGRID de NetApp, que ofrece un coste reducido y otras ventajas operativas. Los científicos de datos extraen estos conjuntos de datos y los ingenieros en varios pasos para prepararlos para el entrenamiento con un modelo específico, a menudo creando varias versiones en el proceso. Al siguiente paso, el científico de datos debe seleccionar recursos informáticos optimizados (GPU, instancias de CPU de gama alta, un clúster local, etc.) para ejecutar el modelo. En la siguiente figura, se muestra la falta de proximidad del conjunto de datos en un entorno de computación DE ML.



Sin embargo, deben ejecutarse varios experimentos de entrenamiento en paralelo en distintos entornos informáticos, cada uno de los cuales requiere una descarga del conjunto de datos del lago de datos, lo cual es

un proceso costoso y largo. No se garantiza la proximidad del conjunto de datos a entorno de computación (especialmente para un cloud híbrido). Además, otros miembros del equipo que llevan a cabo sus propios experimentos con el mismo conjunto de datos deben pasar por el mismo arduo proceso. Más allá del obvio y lento acceso a los datos, los retos incluyen las dificultades para seguir las versiones de los conjuntos de datos, compartir conjuntos de datos, colaboración y reproducibilidad.

Requisitos del cliente

Los requisitos del cliente pueden variar para lograr una ejecución DE ML de alto rendimiento mientras se usan los recursos de forma eficiente. Por ejemplo, los clientes pueden requerir lo siguiente:

- Acceso rápido a los conjuntos de datos desde cada instancia de computación que ejecuta el modelo de entrenamiento sin incurrir en costosas descargas y complejidades de acceso a los datos
- El uso de cualquier instancia de computación (GPU o CPU) en el cloud o en las instalaciones, sin preocuparse de la ubicación de los conjuntos de datos
- Aumenta la eficiencia y la productividad ejecutando múltiples experimentos de entrenamiento en paralelo con diferentes recursos informáticos en el mismo conjunto de datos sin retrasos innecesarios y latencia de los datos
- Costes mínimos de instancias de computación
- Reproducibilidad mejorada con herramientas para mantener registros de los conjuntos de datos, su linaje, versiones y otros detalles de metadatos
- Colaboración y uso compartido mejorados para que cualquier miembro autorizado del equipo pueda acceder a los conjuntos de datos y realizar experimentos

Para implementar el almacenamiento de conjuntos de datos en caché con el software de gestión de datos ONTAP de NetApp, los clientes deben realizar las siguientes tareas:

- Configure y establezca el almacenamiento NFS más cercano a los recursos de computación.
- Determinar qué conjunto de datos y versión almacenar en caché.
- Supervise la memoria total comprometida con los conjuntos de datos almacenados en caché y cuánto almacenamiento NFS está disponible para las confirmaciones adicionales de la caché (por ejemplo, gestión de caché).
- La caducidad de los conjuntos de datos de la caché si no se han utilizado en un momento determinado. El valor predeterminado es un día; hay otras opciones de configuración disponibles.

Descripción general de la solución

En esta sección se revisa una canalización de ciencia de datos convencional y sus inconvenientes. También presenta la arquitectura de la solución de almacenamiento en caché del conjunto de datos propuesta.

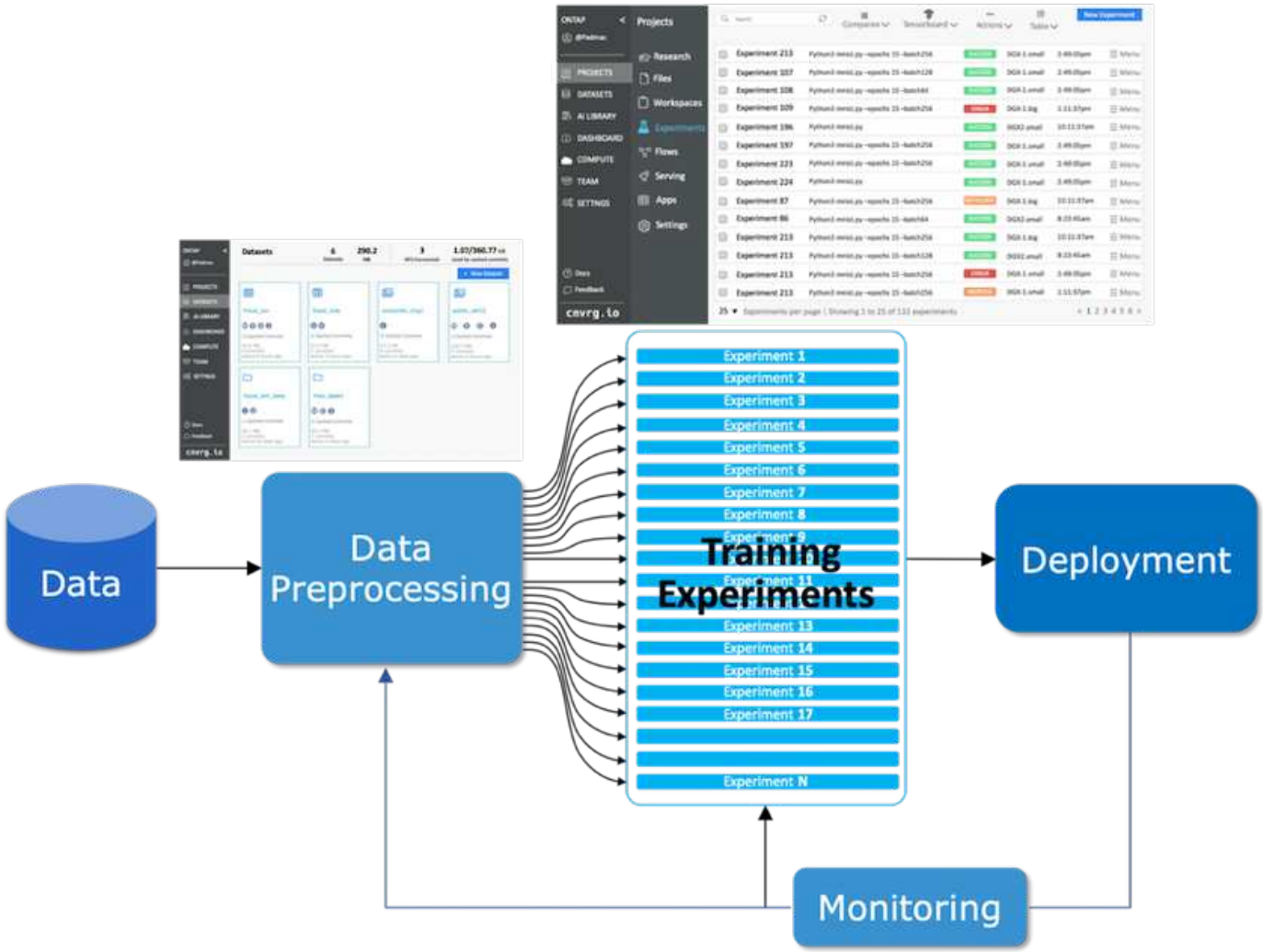
Canalización convencional de ciencia de datos e inconvenientes

Una secuencia típica de desarrollo e implementación de modelos ML implica pasos iterativos que incluyen lo siguiente:

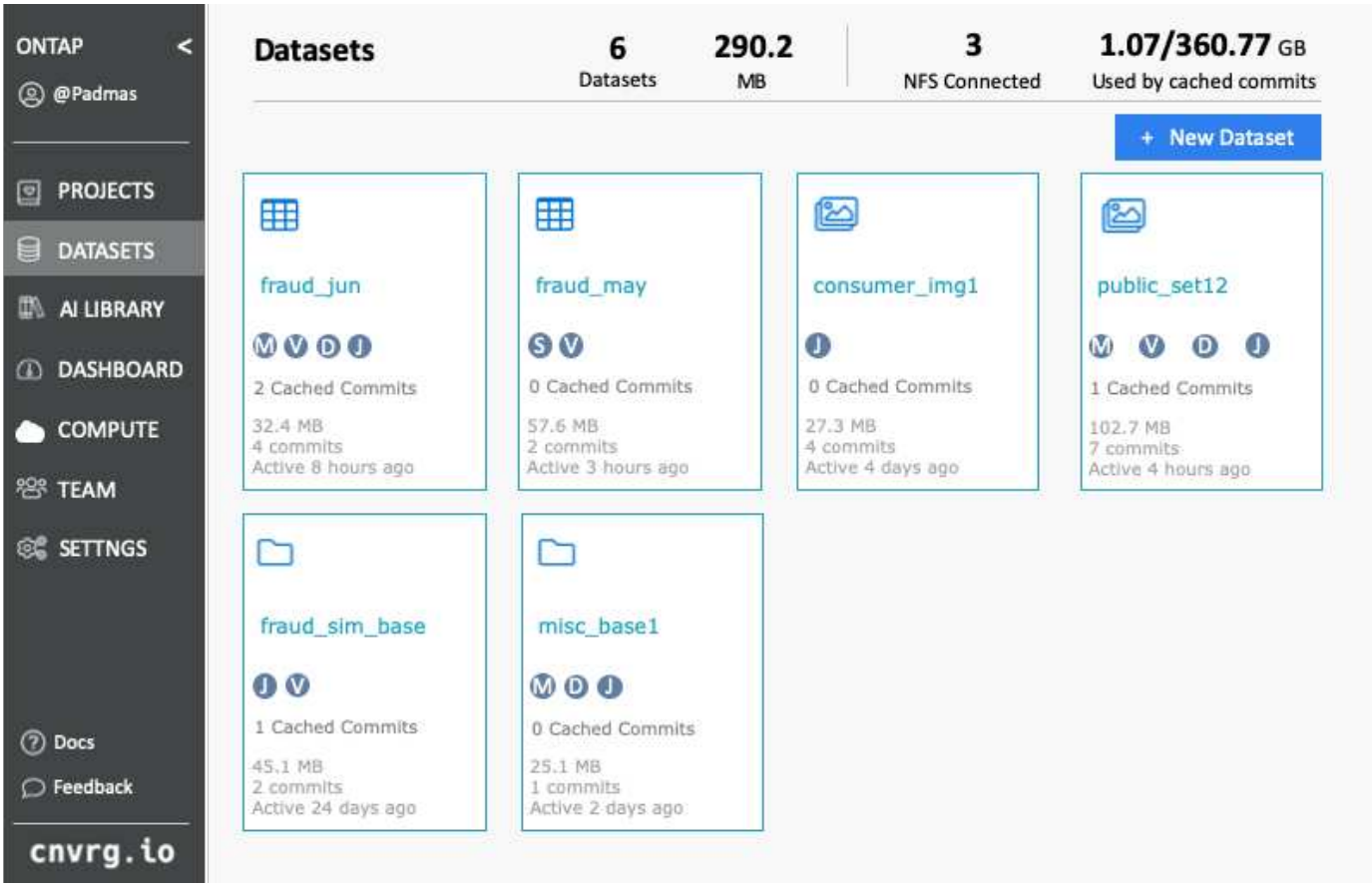
- Ingestión de datos
- Procesamiento previo de los datos (crear varias versiones de los conjuntos de datos)
- Ejecución de múltiples experimentos que implican la optimización de hiperparámetros, diferentes modelos,

etc.

- Puesta en marcha
- Monitoringcnvrg.io ha desarrollado una plataforma completa para automatizar todas las tareas, desde la investigación hasta la implementación. En la siguiente figura se muestra una pequeña muestra de capturas de pantalla del panel relativas a la canalización.



Es muy común tener múltiples conjuntos de datos en juego desde los repositorios públicos y los datos privados. Además, es probable que cada conjunto de datos tenga varias versiones como resultado de la limpieza del conjunto de datos o la ingeniería de funciones. Se necesita un panel que proporcione un concentrador de conjuntos de datos y un concentrador de versiones para garantizar que las herramientas de colaboración y consistencia estén disponibles para el equipo, como se puede ver en la siguiente figura.



El siguiente paso de la canalización es el entrenamiento, que requiere varias instancias paralelas de modelos de entrenamiento, cada uno asociado con un conjunto de datos y una determinada instancia de computación. La vinculación de un conjunto de datos con un cierto experimento con una determinada instancia informática es un reto, ya que es posible que algunos experimentos se realicen mediante instancias de GPU de Amazon Web Services (AWS), mientras que otras instancias de DGX-1 o DGX-2 se llevan a cabo en las instalaciones. Es posible que se ejecuten otros experimentos en servidores de CPU en GCP, mientras que la ubicación del conjunto de datos no está muy cerca de los recursos informáticos que realizan el entrenamiento. Una proximidad razonable tendría conectividad de 10 GbE completa o más baja latencia desde el almacenamiento del conjunto de datos a la instancia de computación.

Es una práctica común que los científicos de datos descarguen el conjunto de datos a la instancia de computación que realice el entrenamiento y ejecute el experimento. Sin embargo, este enfoque presenta varios problemas posibles:

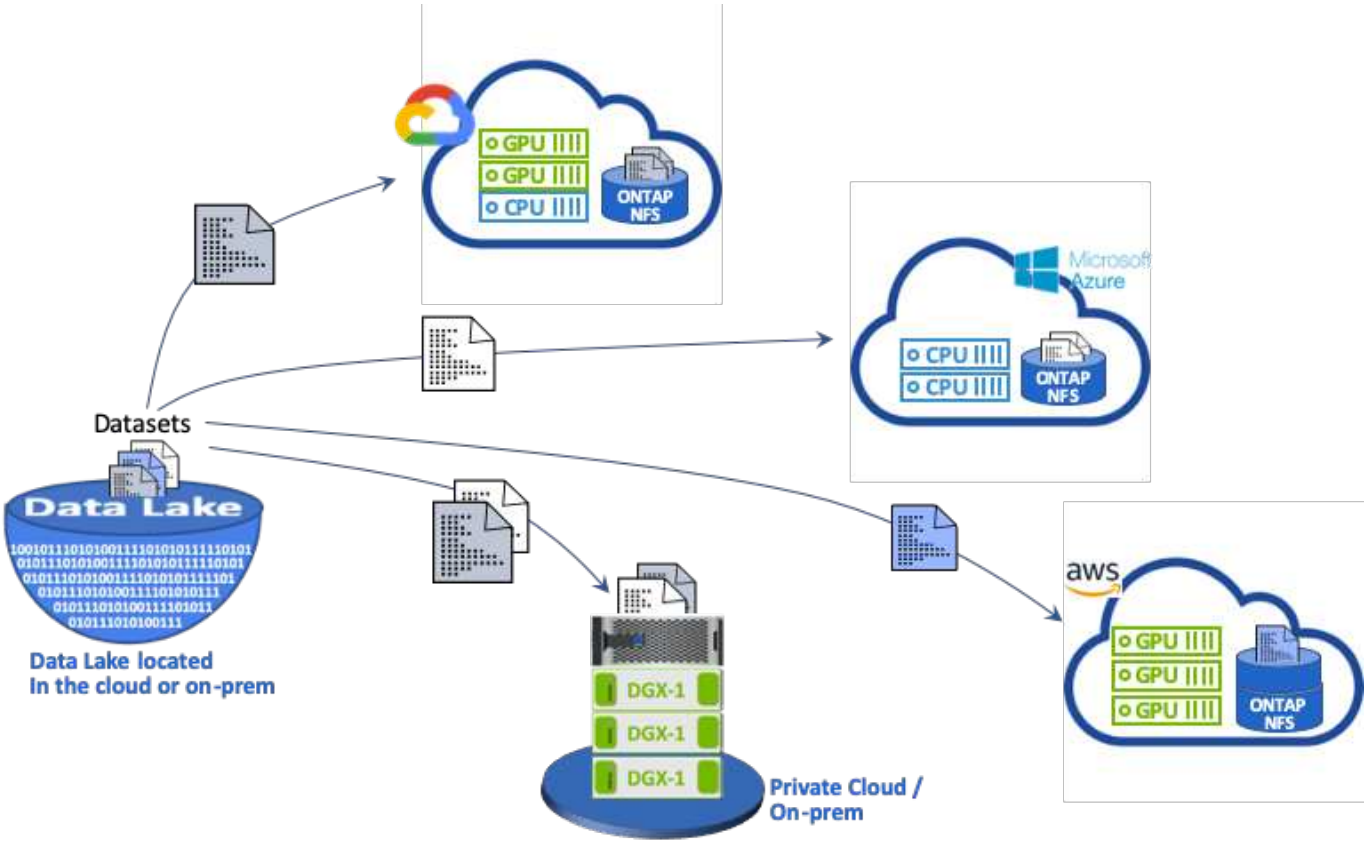
- Cuando el científico de datos descarga el conjunto de datos en una instancia informática, no hay garantías de que el almacenamiento informático integrado tenga un alto rendimiento (un ejemplo de un sistema de alto rendimiento sería la solución NVMe A800 de ONTAP AFF).
- Cuando el conjunto de datos descargado reside en un nodo de computación, el almacenamiento puede convertirse en un cuello de botella cuando se ejecutan los modelos distribuidos en varios nodos (a diferencia del almacenamiento distribuido de alto rendimiento de ONTAP de NetApp).
- La siguiente iteración del experimento de entrenamiento podría realizarse en una instancia de computación diferente debido a conflictos de cola o prioridades, creando de nuevo una distancia significativa de la red desde el conjunto de datos hasta la ubicación de computación.
- Otros miembros del equipo que ejecutan experimentos de entrenamiento en el mismo clúster informático no pueden compartir este conjunto de datos; cada uno realiza la descarga (costosa) del conjunto de datos desde una ubicación arbitraria.

- Si se necesitan otros conjuntos de datos o versiones del mismo conjunto de datos para las siguientes tareas de entrenamiento, los científicos de datos deben volver a realizar la descarga (costosa) del conjunto de datos en la instancia informática que realiza la versión training.NetApp y cnvrg.io han creado una nueva solución de almacenamiento en caché de conjuntos de datos que elimina estos obstáculos. Esta solución crea una ejecución acelerada de la canalización DE ML mediante el almacenamiento en caché de conjuntos de datos activos en el sistema de almacenamiento de alto rendimiento de ONTAP. Con NFS de ONTAP, los conjuntos de datos se almacenan en caché una vez (y solo una) en una estructura de datos con tecnología de NetApp (como AFF A800), que se encuentra junto con el sistema informático. Dado que el almacenamiento de alta velocidad NFS de ONTAP de NetApp puede dar servicio a nodos de computación DE varios ML, el rendimiento de los modelos de entrenamiento se optimiza, lo que permite ahorrar costes, productividad y eficiencia operativa a la organización.

Arquitectura de la solución

Esta solución de NetApp y cnvrg.io proporciona almacenamiento en caché de conjuntos de datos, como se muestra en la siguiente figura. El almacenamiento en caché de conjuntos de datos permite a los científicos de datos elegir la versión deseada de conjuntos de datos o conjuntos de datos y moverlos a la caché NFS de ONTAP, que se encuentra cerca del clúster de computación DE ML. Ahora, el científico de datos puede realizar varios experimentos sin incurrir en retrasos ni descargas. Además, todos los ingenieros de colaboración pueden utilizar el mismo conjunto de datos con el clúster de computación conectado (con la libertad de elegir cualquier nodo) sin descargar adicionales del lago de datos. A los científicos de datos se les ofrece un panel que realiza un seguimiento y supervisa todos los conjuntos de datos y versiones y ofrece una vista de los conjuntos de datos que se almacenan en caché.

La plataforma cnvrg.io detecta automáticamente conjuntos de datos antiguos que no se han utilizado durante un cierto tiempo y los desaloja de la caché, por lo que mantiene un espacio libre de la caché NFS para conjuntos de datos utilizados con mayor frecuencia. Es importante tener en cuenta que el almacenamiento en caché del conjunto de datos con ONTAP funciona en el cloud y en las instalaciones, por lo que proporciona la máxima flexibilidad.



Conceptos y componentes

En esta sección se tratan conceptos y componentes asociados al almacenamiento en caché de datos en un flujo DE trabajo DE ML.

Aprendizaje automático

LA ML se está convirtiendo rápidamente en un factor esencial para muchas empresas y organizaciones de todo el mundo. Por lo tanto, los equipos DE TECNOLOGÍA y DevOps se enfrentan ahora al reto de estandarizar cargas de trabajo DE ML y aprovisionar cloud, recursos informáticos en las instalaciones y recursos informáticos híbridos que dan soporte a los flujos de trabajo dinámicos e intensivos que requieren las tareas de ML y las canalizaciones.

Aprendizaje automático basado en contenedores y Kubernetes

Los contenedores son instancias aisladas del espacio de usuario que se ejecutan sobre un kernel de sistema operativo host compartido. La adopción de contenedores está aumentando rápidamente. Los contenedores ofrecen muchos de los mismos beneficios de uso de pruebas de espacio que las máquinas virtuales (VM). Sin embargo, debido a que se eliminan las capas de hipervisor y de sistema operativo «guest» de las que dependen las máquinas virtuales, los contenedores son mucho más ligeros.

Los contenedores también permiten el empaquetado eficiente de dependencias de aplicaciones, tiempos de ejecución, etc. directamente en una aplicación. El formato de embalaje de contenedor más utilizado es el contenedor Docker. Una aplicación que se haya contenedor en el formato de contenedor Docker se puede ejecutar en cualquier máquina que pueda ejecutar contenedores Docker. Esto es cierto incluso si las dependencias de la aplicación no están presentes en la máquina, porque todas las dependencias están empaquetadas en el propio contenedor. Para obtener más información, visite la "Sitio web de Docker".

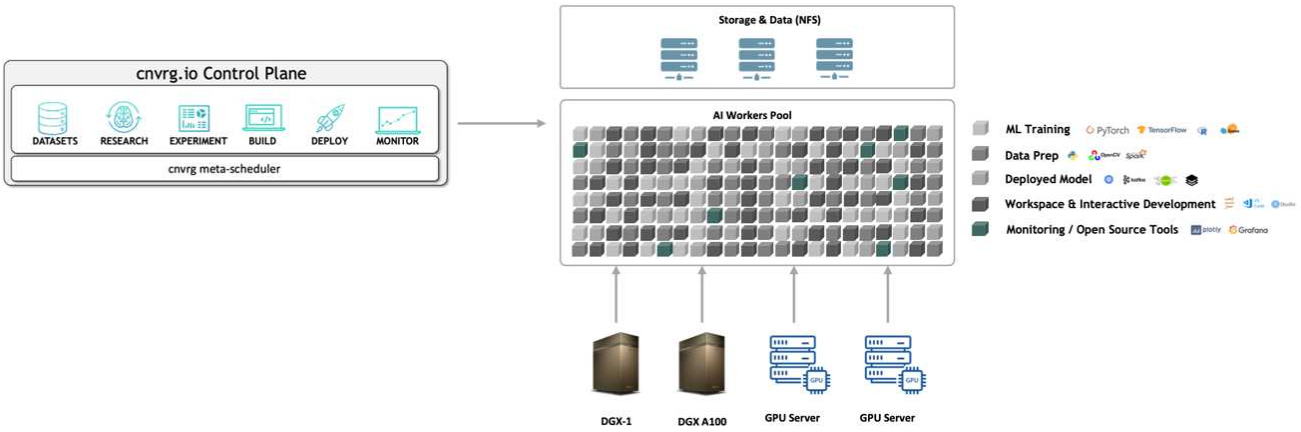
Kubernetes, el popular orquestador de contenedores, permite a los científicos de datos lanzar trabajos y canalizaciones flexibles basados en contenedores. También permite a los equipos de infraestructura gestionar y supervisar cargas de trabajo DE ML en un único entorno gestionado y nativo del cloud. Para obtener más información, visite la "Sitio web de Kubernetes".

cnvrg.io

Cnvrg.io es un sistema operativo de IA que transforma la forma en la que las empresas gestionan, escalan y aceleran la IA y el desarrollo científico de datos de la investigación a la producción. La plataforma de código primero está desarrollada por científicos de datos para científicos de datos y ofrece flexibilidad para ejecutarse en las instalaciones o en el cloud. Gracias a la gestión de modelos, MLOPS y soluciones continuas DE ML, cnvrg.io aporta una tecnología de primera línea a los equipos de ciencia de datos para que puedan dedicar menos tiempo a DevOps y centrarse en la auténtica magia, gracias a los algoritmos. Desde que utiliza cnvrg.io, los equipos de distintos sectores han obtenido más modelos de producción, lo que da como resultado un aumento del valor empresarial.

Cnvrg.io Meta-Scheduler

cnvrg. i/o tiene una arquitectura única que permite A LOS DEPARTAMENTOS DE TECNOLOGÍA e ingenieros conectar distintos recursos informáticos al mismo plano de control y que cnvrg.io gestiona tareas DE ML en todos los recursos. Esto significa que puede conectar varios clústeres de Kubernetes en las instalaciones, servidores de VM y cuentas de cloud, y ejecutar cargas de trabajo DE ML en todos los recursos, como se muestra en la siguiente figura.



Cnvrg.io almacenamiento en caché de datos

cnvrg.io permite a los científicos de datos definir versiones de conjuntos de datos calientes y fríos con su tecnología de almacenamiento en caché de datos. De forma predeterminada, los conjuntos de datos se almacenan en una base de datos de almacenamiento de objetos centralizada. A continuación, los científicos de datos pueden almacenar en caché una versión de datos específica en el recurso de computación seleccionado para ahorrar tiempo en la descarga y, por tanto, aumentar EL desarrollo DE ML y la productividad. Los conjuntos de datos que se almacenan en la caché y no se utilizan durante unos días se borran automáticamente del NFS seleccionado. El almacenamiento en caché y el borrado de la caché se pueden realizar con un solo clic; no se requiere codificación, NI trabajo de DevOps.

Cnvrg.io fluye y canalizaciones ML

Cnvrg.io fluye es una herramienta para construir tuberías DE PRODUCCIÓN ML. Cada componente de un flujo es un script/código que se ejecuta en una computación seleccionada con una imagen de Docker base. Este diseño permite a los científicos e ingenieros de datos crear una única canalización que puede ejecutar tanto en las instalaciones como en el cloud. cnvrg.io garantiza que los datos, los parámetros y los artefactos se mueven entre los diferentes componentes. Además, se supervisa y se sigue cada flujo para obtener ciencia de datos reproducibles al 100%.

NÚCLEO cnvrg.io

El NÚCLEO cnvrg.io es una plataforma gratuita para que la comunidad de ciencia de datos pueda ayudar a los científicos de datos a centrarse más en la ciencia de datos y menos en DevOps. La infraestructura flexible DE CORE aporta a los científicos de datos el control de usar cualquier idioma, marco de IA o entorno informático, ya sea en las instalaciones o en el cloud, para poder hacer lo que mejor hacen o crear algoritmos. El NÚCLEO cnvrg.io se puede instalar fácilmente con un único comando en cualquier clúster de Kubernetes.

ONTAP AI de NetApp

ONTAP AI es una arquitectura de referencia de centro de datos para cargas de trabajo DE APRENDIZAJE profundo (DL) y ML que utiliza sistemas de almacenamiento AFF de NetApp y sistemas DGX de NVIDIA con GPU Tesla V100. ONTAP AI se basa en el protocolo de archivos NFS estándar del sector en Ethernet de 100 GB y proporciona a los clientes una infraestructura DE APRENDIZAJE PROFUNDO DE alto rendimiento QUE utiliza tecnologías estándar para el centro de datos para reducir los gastos generales de implementación y administración. Con una red y protocolos estandarizados, ONTAP AI se integra en entornos de cloud híbrido a la vez que mantiene la coherencia y la simplicidad operativas. Como solución de infraestructura prevalidada, ONTAP AI reduce el tiempo y el riesgo de la puesta en marcha y la sobrecarga de la administración de forma significativa, lo que permite a los clientes lograr una rentabilidad de la inversión más rápida.

DeepOps de NVIDIA

DeepOps es un proyecto de código abierto de NVIDIA que, con Ansible, automatiza la puesta en marcha de clústeres de servidores de GPU de acuerdo con las prácticas recomendadas. DeepOps es modular y se puede utilizar para realizar varias tareas de puesta en marcha. En este documento y en el ejercicio de validación descrito, DeepOps se utiliza para poner en marcha un clúster de Kubernetes que consta de nodos de trabajo de servidor GPU. Para obtener más información, visite la "Sitio web DeepOps".

Trident de NetApp

Trident es un orquestador de almacenamiento de código abierto desarrollado y mantenido por NetApp que simplifica en gran medida la creación, la gestión y el consumo de almacenamiento persistente para cargas de trabajo de Kubernetes. Trident, en sí misma, una aplicación nativa de Kubernetes, se ejecuta directamente en un clúster de Kubernetes. Con Trident, los usuarios de Kubernetes (desarrolladores, científicos de datos, administradores de Kubernetes, etc.) pueden crear, gestionar e interactuar con volúmenes de almacenamiento persistente en el formato Kubernetes estándar, con el que ya están familiarizados. Al mismo tiempo, pueden aprovechar las funciones avanzadas de gestión de datos de NetApp y un Data Fabric con tecnología de NetApp. Trident elimina las complejidades del almacenamiento persistente y facilita el consumo. Para obtener más información, visite la "Sitio web de Trident".

StorageGRID de NetApp

StorageGRID de NetApp es una plataforma de almacenamiento de objetos definida por software diseñada para satisfacer estas necesidades proporcionando un almacenamiento sencillo y similar al cloud a el que los usuarios pueden acceder mediante el protocolo S3. StorageGRID es un sistema de escalado horizontal diseñado para admitir varios nodos en sitios conectados a Internet, independientemente de la distancia. Con el motor de políticas inteligente de StorageGRID, los usuarios pueden elegir objetos de codificación de borrado en todos los sitios para lograr resiliencia geográfica o replicación de objetos entre sitios remotos para minimizar la latencia de acceso WAN. StorageGRID proporciona un excelente lago de datos de almacenamiento de objetos primarios de cloud privado en esta solución.

Cloud Volumes ONTAP de NetApp

El software de gestión de datos Cloud Volumes ONTAP de NetApp proporciona control, protección y eficiencia para los datos de usuarios con la flexibilidad de proveedores de cloud público como AWS, Google Cloud Platform y Microsoft Azure. Cloud Volumes ONTAP es un software para la gestión de datos nativo del cloud, integrado en el software de almacenamiento ONTAP de NetApp, que proporciona a los usuarios una plataforma de almacenamiento universal superior que cubre sus necesidades de datos en el cloud. Disponer de un mismo software de almacenamiento en el cloud y en las instalaciones proporciona a los usuarios el valor de una estructura de datos sin necesidad de formar al personal INFORMÁTICO en todos los métodos nuevos para gestionar los datos.

Para los clientes interesados en modelos de puesta en marcha de cloud híbrido, Cloud Volumes ONTAP puede proporcionar las mismas funcionalidades y un rendimiento líder en la mayoría de clouds públicos para proporcionar una experiencia de usuario fluida y coherente en cualquier entorno.

Requisitos de hardware y software

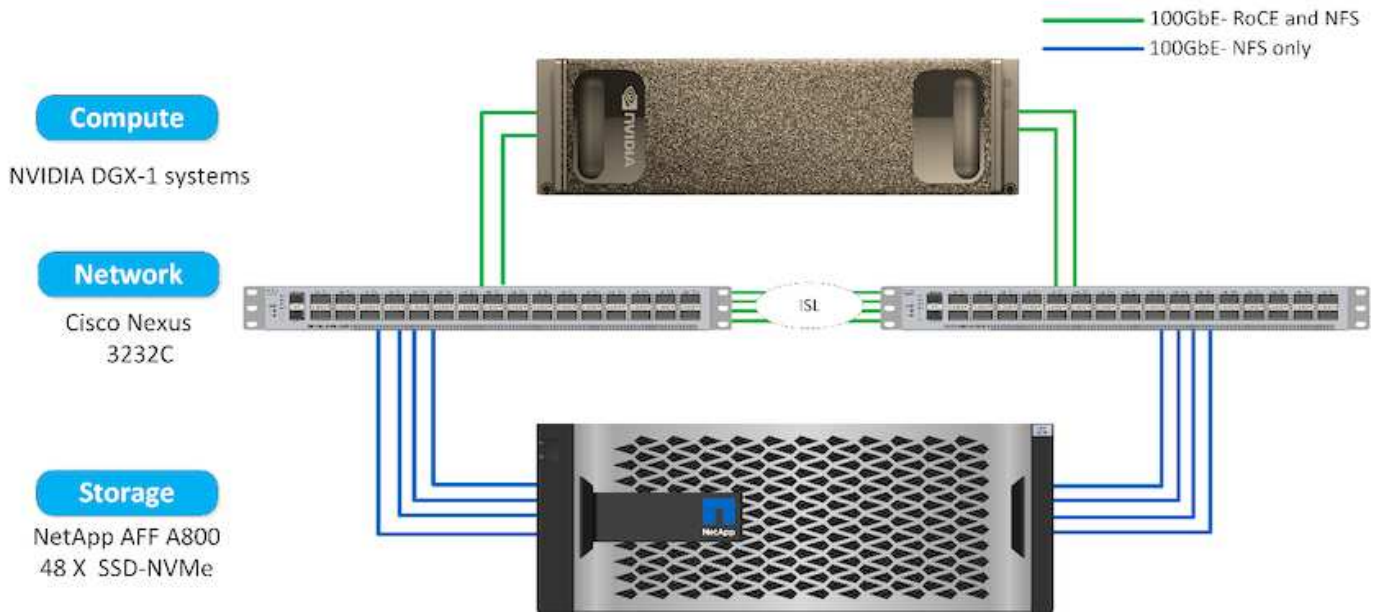
En esta sección se tratan los requisitos tecnológicos de la solución de IA de ONTAP.

Requisitos de hardware

Aunque los requisitos de hardware dependen de cargas de trabajo específicas de los clientes, es posible poner en marcha IA de ONTAP a cualquier escala para la ingeniería de datos, el entrenamiento de modelos y

la inferencia de producción desde una única GPU hasta configuraciones a escala de rack para operaciones DE APRENDIZAJE AUTOMÁTICO o aprendizaje profundo a gran escala. Si quiere más información sobre ONTAP AI, consulte "Sitio web de IA de ONTAP".

Esta solución se validó utilizando un sistema DGX-1 para computación, un sistema de almacenamiento AFF A800 de NetApp y Cisco Nexus 3232C para conectividad de red. El AFF A800 utilizado en esta validación puede admitir hasta 10 sistemas DGX-1 para la mayoría de cargas de trabajo DE ML/DL. En la siguiente figura se muestra la topología ONTAP AI utilizada para entrenar el modelo en esta validación.



Para ampliar esta solución a un cloud público, Cloud Volumes ONTAP puede ponerse en marcha junto con los recursos informáticos de la GPU del cloud e integrarse en un tejido de datos del cloud híbrido que permita a los clientes utilizar los recursos que sean apropiados para una carga de trabajo en concreto.

Requisitos de software

En la siguiente tabla se muestran las versiones de software específicas que se utilizan en la validación de esta solución.

Componente	Versión
Ubuntu	18.04.4 LTS
SO DGX DE NVIDIA	4.4.0
DeepOps de NVIDIA	20.02.1
Kubernetes	1.15
Timón	3.1.0
cnvrg.io	3.0.0
ONTAP de NetApp	9.6P4

Para esta validación de soluciones, se puso en marcha Kubernetes como un clúster de un solo nodo en el sistema DGX-1. Para las puestas en marcha a gran escala, se deben poner en marcha nodos maestros de Kubernetes independientes para proporcionar una alta disponibilidad de los servicios de gestión y reservar valiosos recursos DGX para las cargas de trabajo DE APRENDIZAJE AUTOMÁTICO y aprendizaje profundo.

Detalles de validación y puesta en marcha de la solución

En las siguientes secciones se tratan los detalles de la validación y puesta en marcha de la solución.

Puesta en marcha de IA de ONTAP

La puesta en marcha de ONTAP AI requiere la instalación y configuración de hardware de almacenamiento, computación y redes. Este documento no cubre las instrucciones específicas para la puesta en marcha de la infraestructura de IA de ONTAP. Para obtener información detallada sobre la implementación, consulte "NVA-1121-PUESTA en MARCHA: ONTAP AI de NetApp, con tecnología NVIDIA".

Para esta validación de soluciones, se creó y montó un solo volumen en el sistema DGX-1. A continuación, ese punto de montaje se montó en los contenedores para que los datos sean accesibles para el entrenamiento. Para puestas en marcha a gran escala, Trident de NetApp automatiza la creación y el montaje de volúmenes para eliminar la sobrecarga administrativa y permitir la gestión de recursos por parte del usuario final.

Puesta en marcha de Kubernetes

Para poner en marcha y configurar su clúster de Kubernetes con NVIDIA DeepOPS, realice las siguientes tareas desde un host de salto de implementación:

1. Descargue NVIDIA DeepOps siguiendo las instrucciones de "Página de Inicio" En el sitio de NVIDIA DeepOps GitHub.
2. Implemente Kubernetes en su clúster siguiendo las instrucciones que se indican en "Guía de puesta en marcha de Kubernetes" En el sitio de NVIDIA DeepOps GitHub.

> Para que la puesta en marcha de DeepOps Kubernetes funcione, debe haber el mismo usuario en todos los nodos maestro y de trabajador de Kubernetes.

Si la implementación falla, cambie el valor de `kubectl_localhost` a falso in `deepops/config/group_vars/k8s-cluster.yml` y repita el paso 2. La `Copy kubectl binary to ansible host` tarea, que sólo se ejecuta cuando el valor de `kubectl_localhost` Es cierto que se basa en el módulo Fetch Ansible, que tiene problemas conocidos con el uso de memoria. Estos problemas de uso de la memoria a veces pueden provocar un error en la tarea. Si la tarea falla debido a un problema de memoria, el resto de la operación de implementación no se completa correctamente.

Si la implementación se completa correctamente después de cambiar el valor de `kubectl_localhost` para `false`, a continuación, debe copiar manualmente la `kubectl binary` Desde un nodo maestro de Kubernetes a un host de salto de implementación. Puede encontrar la ubicación del `kubectl binary` en un nodo maestro específico ejecutando el `which kubectl` comando directamente en ese nodo.

Puesta en marcha de cnvrg.io

Ponga en marcha EL NÚCLEO cnvrg mediante Helm

Timón es la forma más sencilla de poner en marcha cnvrg rápidamente mediante cualquier clúster, en las instalaciones, MiniClube o en cualquier clúster de cloud (como AKS, EKS y GKE). En esta sección se describe cómo se instaló cnvrg en una instancia de las instalaciones (DGX-1) con Kubernetes instalado.

Requisitos previos

Antes de completar la instalación, debe instalar y preparar las siguientes dependencias en su equipo local:

- Kubectl
- Timón 3.x
- Clúster de Kubernetes 1.15 o posterior

Despliegue mediante Helm

1. Para descargar los gráficos cnvrg del timón más actualizados, ejecute el siguiente comando:

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

2. Antes de poner en marcha cnvrg, necesita la dirección IP externa del clúster y el nombre del nodo en el que se va a implementar cnvrg. Para poner en marcha cnvrg en un clúster de Kubernetes en las instalaciones, ejecute el siguiente comando:

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

3. Ejecute el `helm install` comando. Todos los servicios y sistemas se instalan automáticamente en el clúster. El proceso puede tardar hasta 15 minutos.
4. La `helm install` command puede tardar hasta 10 minutos. Cuando finalice la implementación, vaya a la dirección URL de su cnvrg recién implementado o agregue el nuevo clúster como un recurso dentro de su organización. La `helm` Command le informa de la URL correcta.

```
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. Cuando el estado de todos los contenedores se ejecuta o se completa, cnvrg se ha implementado correctamente. Debería ser similar a la siguiente salida de ejemplo:

```
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
```

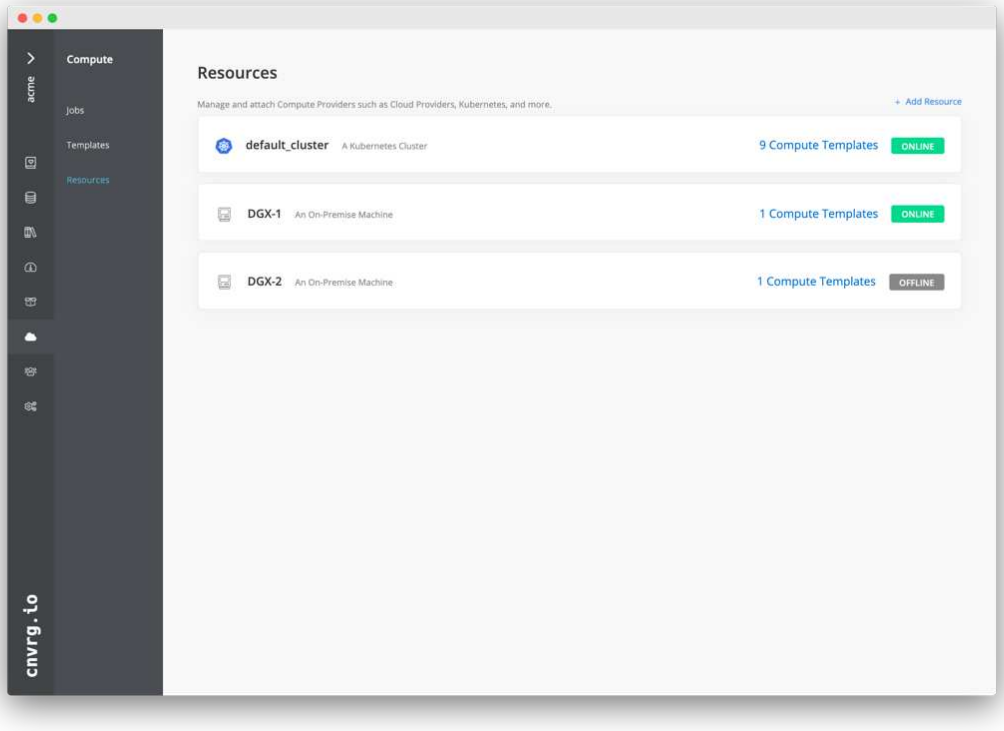
Formación de modelos de visión computarizada con ResNet50 y el conjunto de datos de rayos X torácicos

Cnvrg.io AI OS se puso en marcha en una configuración de Kubernetes en una arquitectura de IA ONTAP de NetApp basada en el sistema NVIDIA DGX. Para la validación, se utilizó el conjunto de datos de radiografía de tórax de los NIH que consistía en imágenes desidentificadas de radiografías de tórax. Las imágenes estaban en formato PNG. Los datos fueron proporcionados por el Centro clínico de los NIH y están disponibles a través del "Sitio de descarga de NIH". Utilizamos una muestra de 250 GB de los datos con 627, 615 imágenes en 15 clases.

El conjunto de datos se cargó en la plataforma cnvrg y se almacenó en caché en una exportación NFS desde el sistema de almacenamiento AFF A800 de NetApp.

Configure los recursos de computación

La arquitectura cnvrg y la funcionalidad de programación de metadatos permiten que los ingenieros y profesionales DE TECNOLOGÍA adjunte diferentes recursos informáticos a una única plataforma. En nuestra configuración, utilizamos el mismo cnvrg de clúster que se puso en marcha para ejecutar cargas de trabajo de aprendizaje profundo. Si necesita conectar clústeres adicionales, utilice la GUI, tal como se muestra en la siguiente captura de pantalla.



Cargar datos

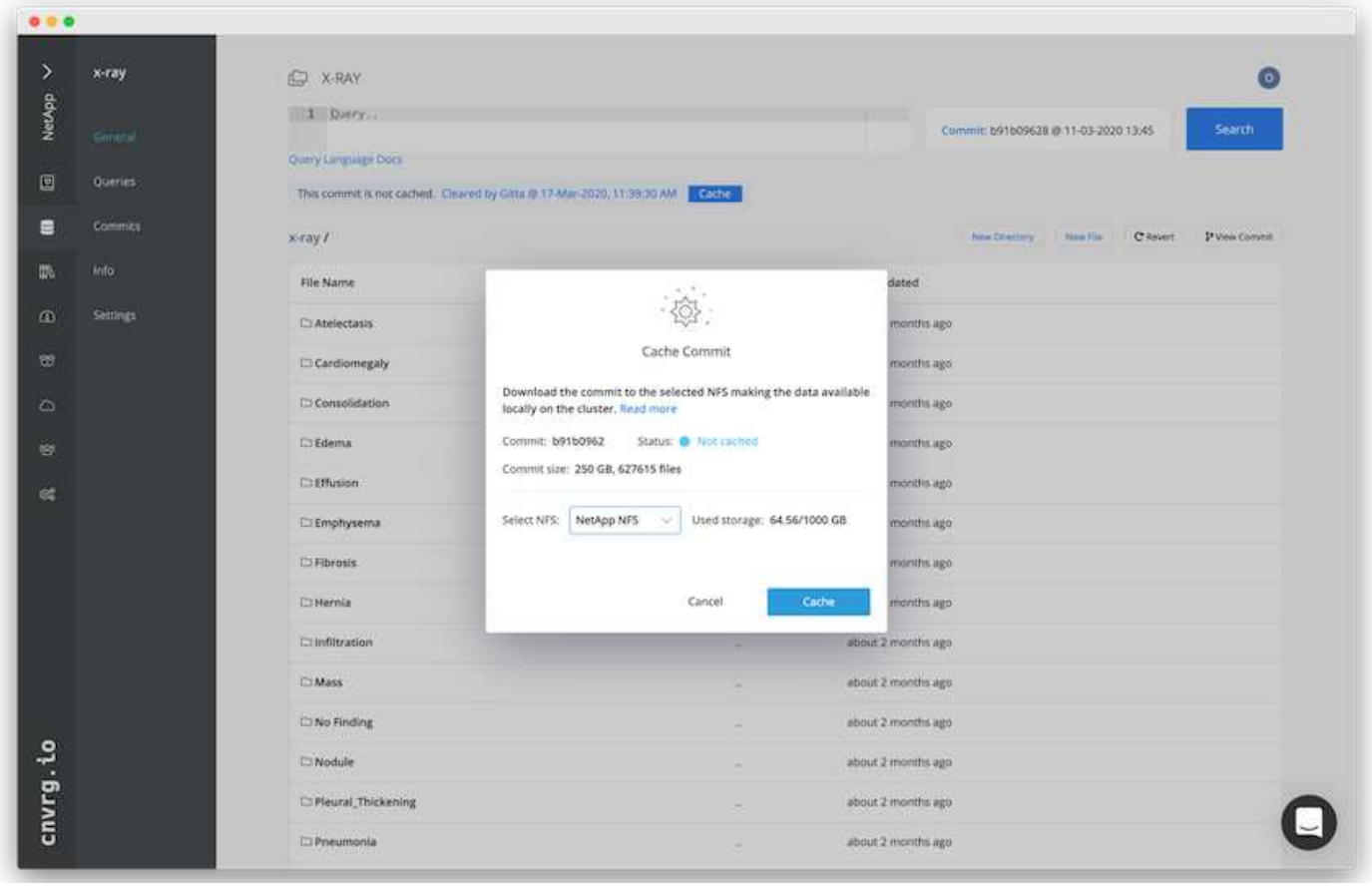
Para cargar datos en la plataforma cnvrg, puede utilizar la interfaz gráfica de usuario o la interfaz de línea de comandos cnvrg. En el caso de conjuntos de datos de gran tamaño, NetApp recomienda el uso de la CLI porque es una herramienta sólida, escalable y fiable capaz de gestionar un gran número de archivos.

Para cargar datos, realice los siguientes pasos:

1. Descargue el "CLI de cnvrg".
2. desplácese hasta el directorio de rayos x.
3. Inicialice el conjunto de datos en la plataforma con el `cnvrg data init` comando.
4. Cargue todo el contenido del directorio en el lago de datos central con el `cnvrg data sync` Command.una vez que los datos se cargan en el almacén de objetos central (StorageGRID, S3 u otros), puede navegar por la GUI. La siguiente figura muestra un archivo PNG cargado de imagen de fibrosis torácica. Además, cnvrg introduce los datos de modo que cualquier modelo que cree pueda reproducirse en la versión de los datos.

Datos de Cach

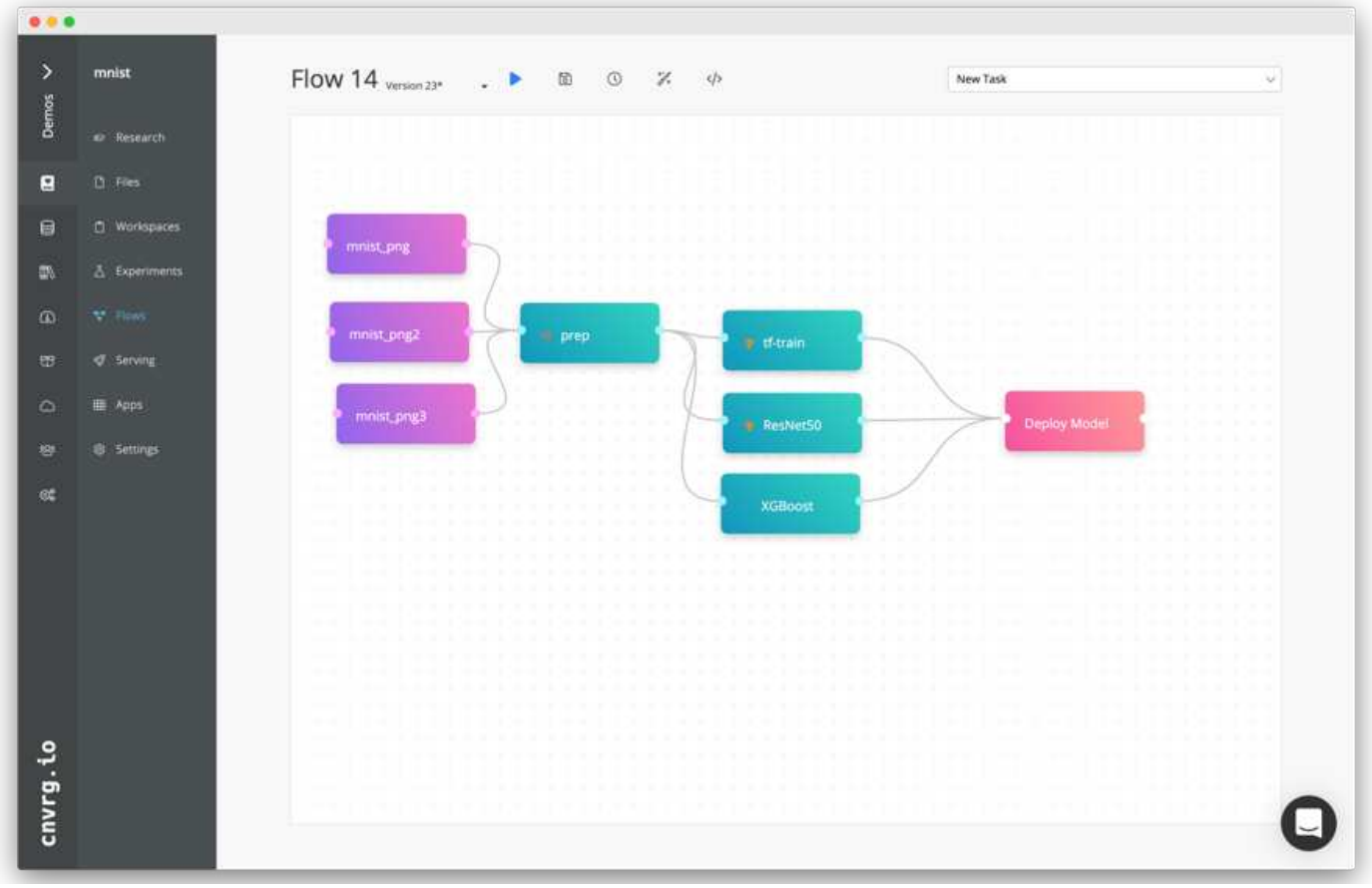
Para realizar el entrenamiento con más rapidez y evitar la descarga de más de 600 000 archivos para cada entrenamiento y experimento, utilizamos la función de almacenamiento en caché de datos después de cargar los datos inicialmente en el almacén de objetos del lago de datos central.



Después de que los usuarios hagan clic en caché, cnvrg descarga los datos de su registro específico del almacén de objetos remoto y los almacena en caché en el volumen NFS de ONTAP. Una vez que finalice, los datos estarán disponibles para el entrenamiento instantáneo. Además, si los datos no se utilizan durante unos días (para el entrenamiento de modelos o la exploración, por ejemplo), cnvrg borra automáticamente la memoria caché.

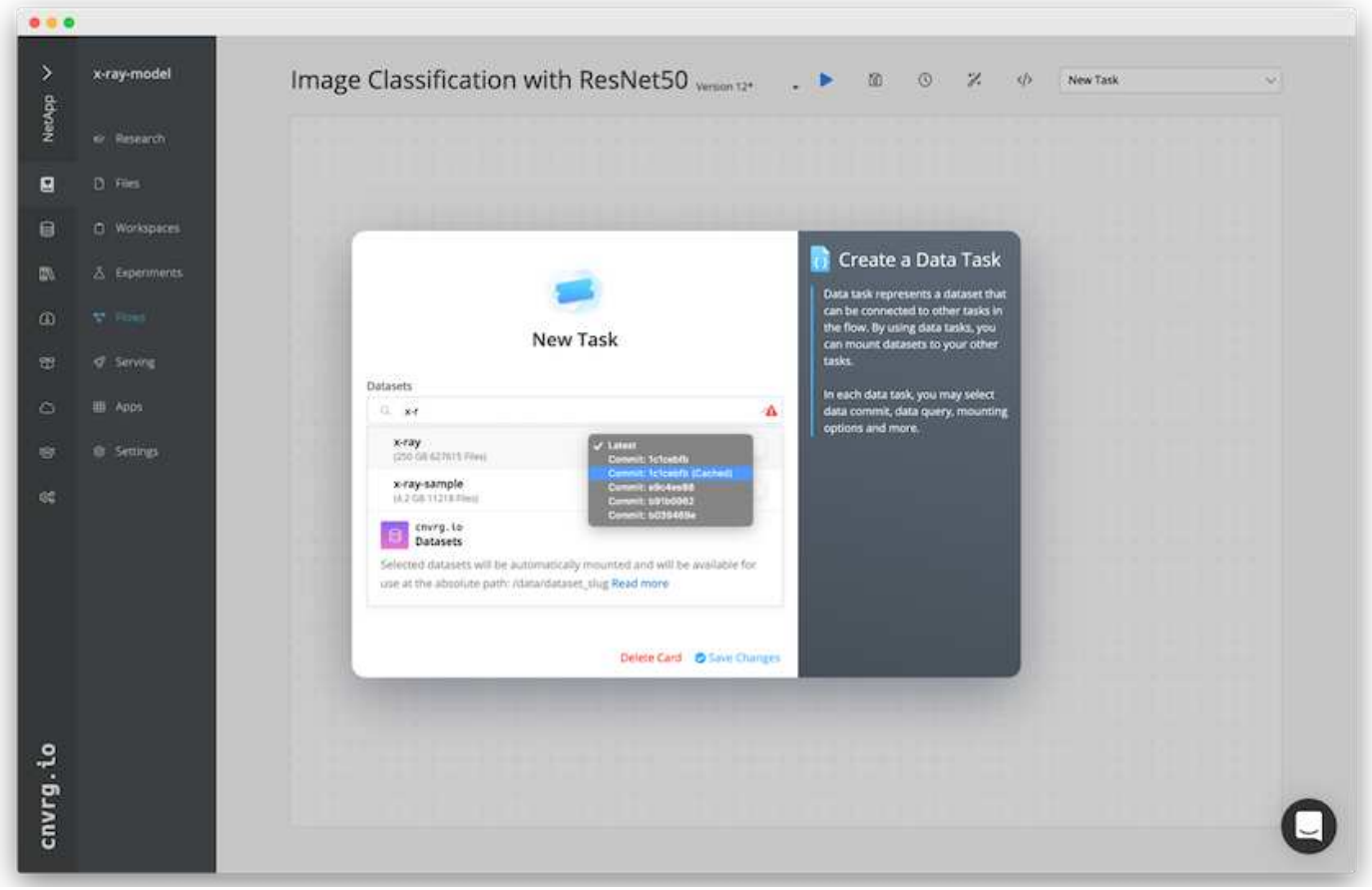
Cree un ML de canalización con datos en caché

Cnvrg fluye le permite construir fácilmente tuberías ML de producción. Los flujos son flexibles, pueden funcionar para cualquier tipo de caso de uso DE ML y pueden crearse a través de la GUI o el código. Cada componente de un flujo puede ejecutarse en un recurso de computación diferente con una imagen de Docker diferente, por lo que es posible crear canalizaciones de APRENDIZAJE AUTOMÁTICO optimizadas y de cloud híbrido.



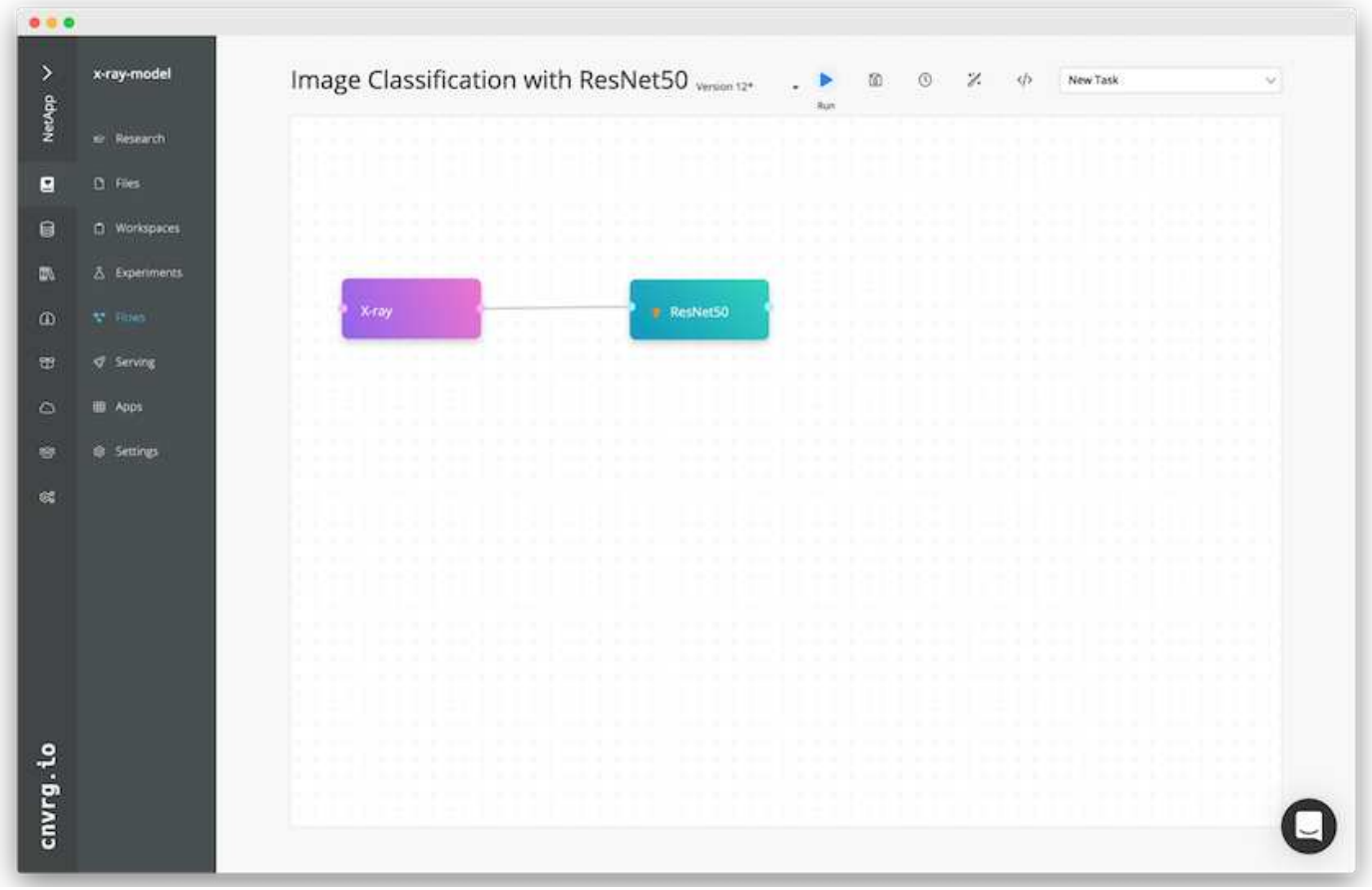
Cómo construir el flujo de la radiografía de tórax: Ajuste de los datos

Hemos añadido nuestro conjunto de datos a un flujo recién creado. Al agregar el conjunto de datos, puede seleccionar la versión específica (Commit) e indicar si desea la versión en caché. En este ejemplo, hemos seleccionado la confirmación almacenada en caché.



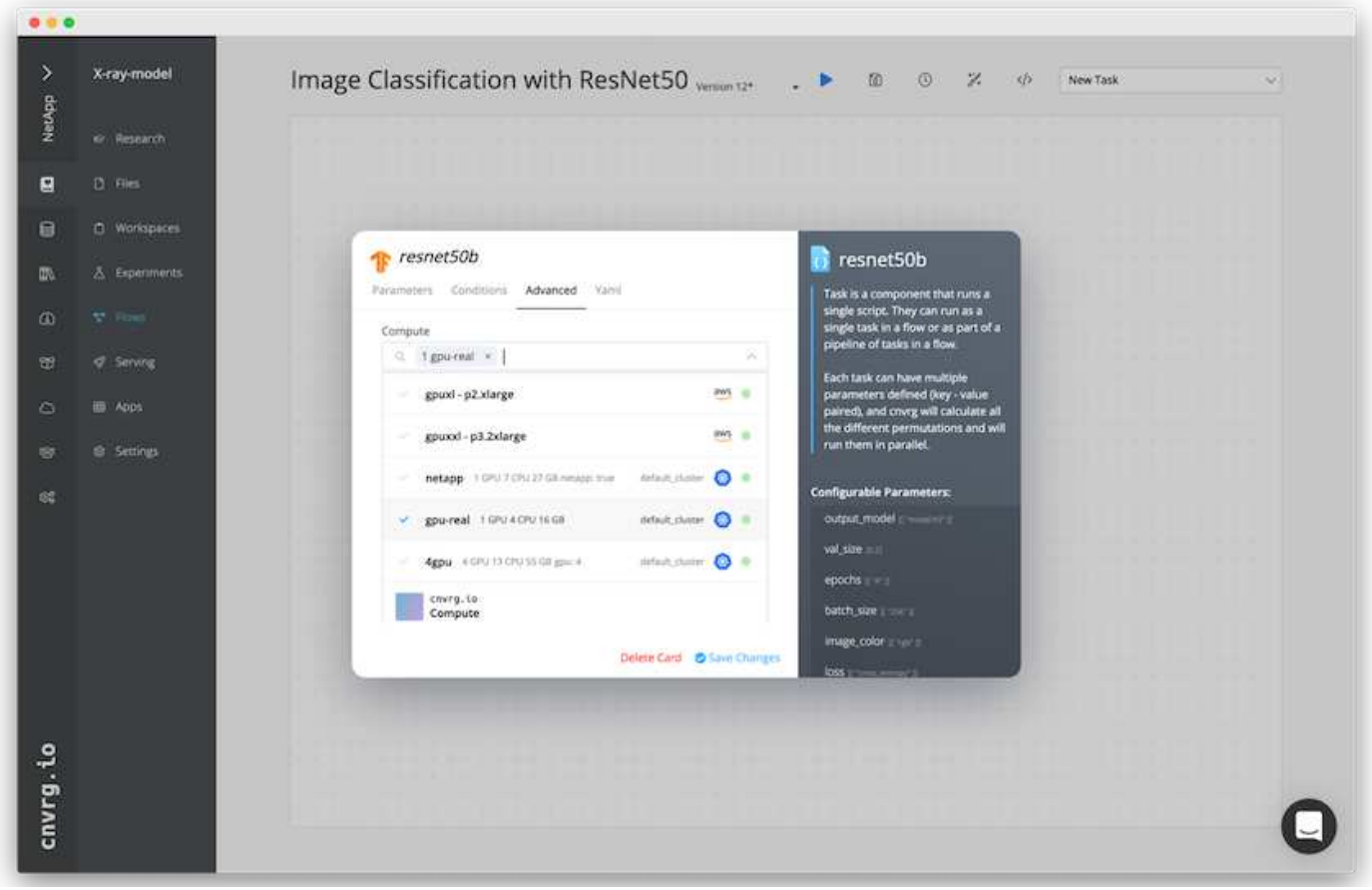
Cómo construir el flujo de rayos X del tórax: Ajuste del modelo de entrenamiento: ResNet50

En la canalización, puede agregar cualquier tipo de código personalizado que desee. En cnvrg, también existe la biblioteca de IA, una colección de componentes DE ML reutilizables. En la biblioteca de IA existen algoritmos, scripts, orígenes de datos y otras soluciones que se pueden usar en cualquier flujo DE APRENDIZAJE profundo o DE ML. En este ejemplo, hemos seleccionado el módulo ResNet50 premontado. Se utilizaron parámetros predeterminados como batch_size:128, épocas:10 y más. Estos parámetros pueden verse en los documentos de la Biblioteca de IA. La siguiente captura de pantalla muestra el nuevo flujo con el conjunto de datos de rayos X conectado a ResNet50.



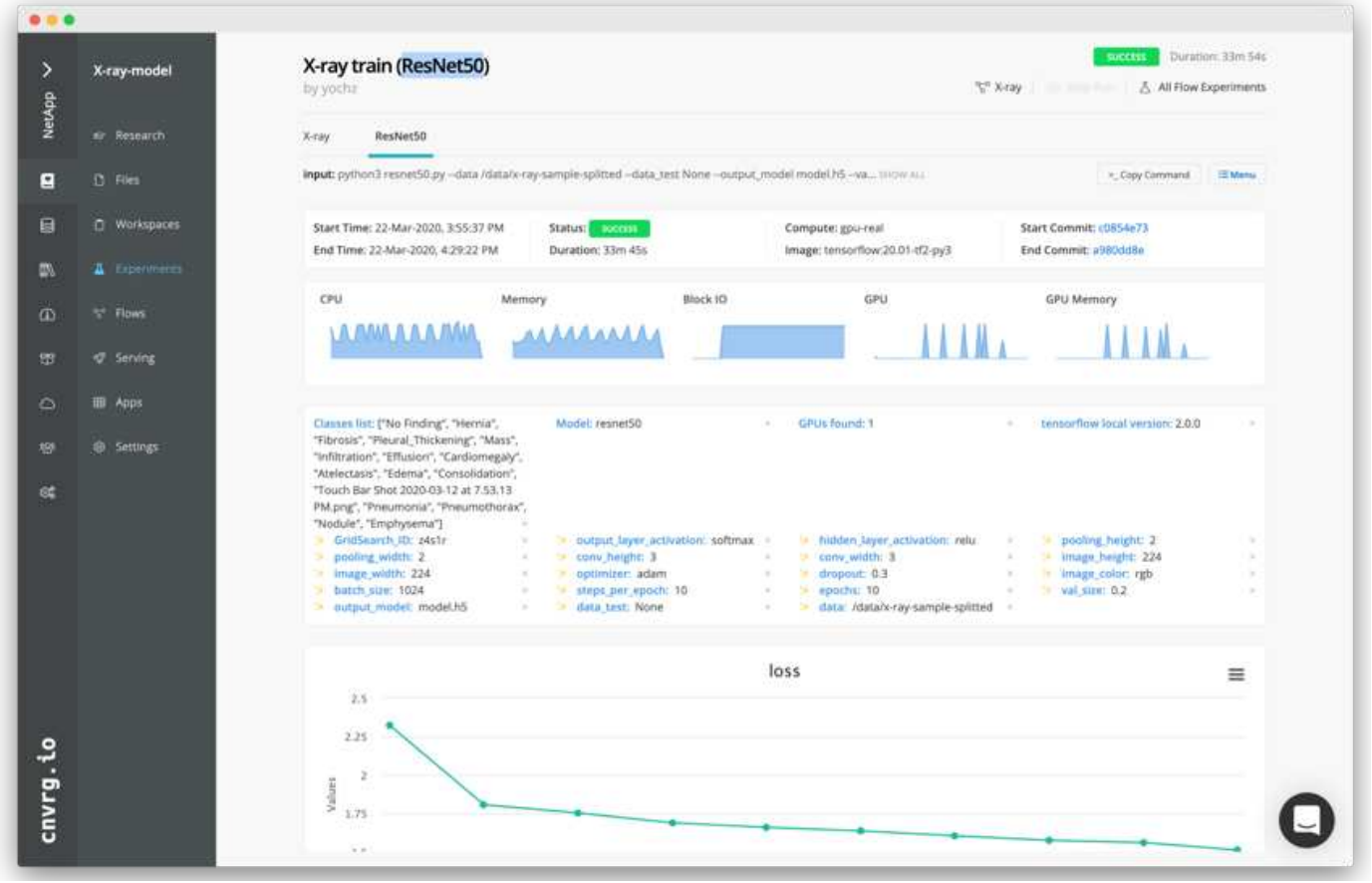
Defina el recurso de computación para ResNet50

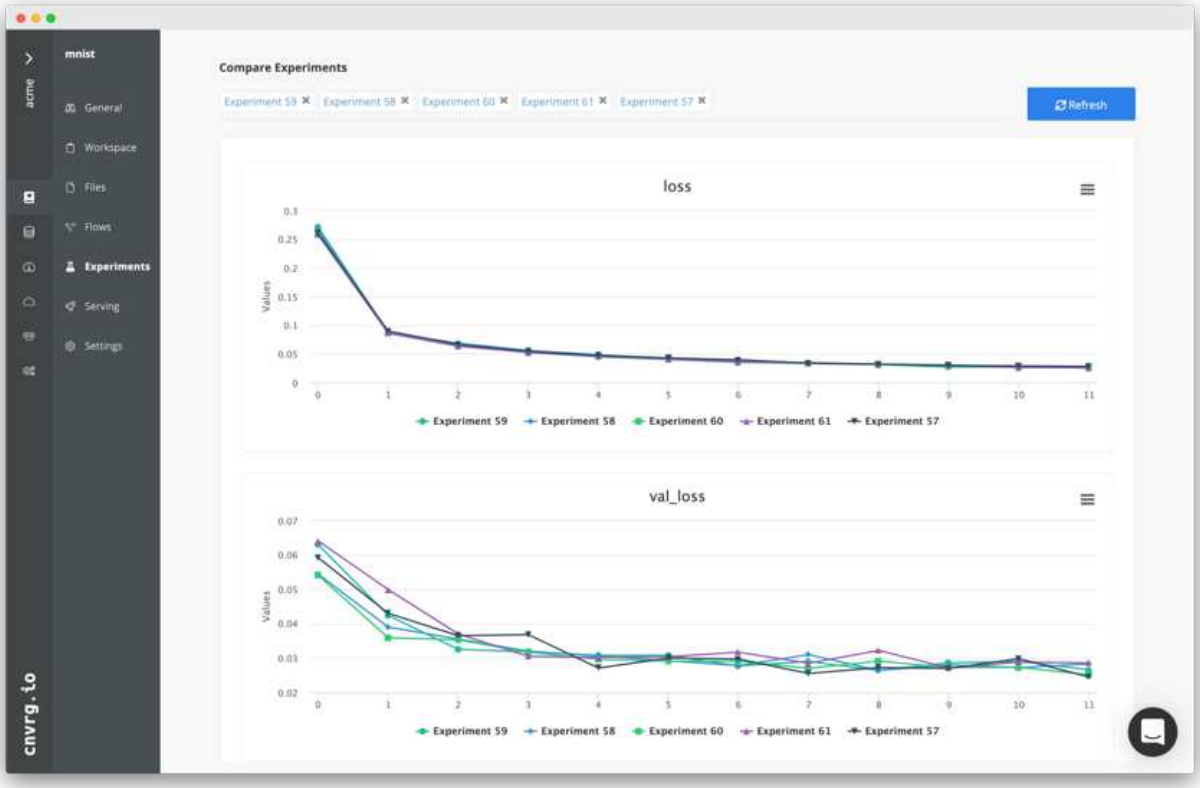
Cada algoritmo o componente en flujos cnvrg puede ejecutarse en una instancia de computación diferente, con una imagen de Docker diferente. En nuestra configuración, queríamos ejecutar el algoritmo de entrenamiento en los sistemas DGX de NVIDIA con la arquitectura de IA ONTAP de NetApp. En la siguiente figura, hemos seleccionado `gpu-real`, que es una plantilla de cálculo y una especificación para nuestro clúster local. También creamos una cola de plantillas y seleccionamos varias plantillas. De esta manera, si el `gpu-real` no se puede asignar el recurso (si, por ejemplo, otros científicos de datos lo están utilizando), puede habilitar la explosión automática en el cloud añadiendo una plantilla de proveedor de cloud. La siguiente captura de pantalla muestra el uso de gpu-real como nodo de computación para ResNet50.



Seguimiento y seguimiento de resultados

Después de ejecutar un flujo, cnvrg activa el motor de seguimiento y supervisión. Cada ejecución de un flujo se documenta y actualiza automáticamente en tiempo real. Hiperparámetros, métricas, uso de recursos (utilización de GPU, etc.), versión de código, artefactos, registros Y así sucesivamente están automáticamente disponibles en la sección experimentos, como se muestra en las dos capturas de pantalla siguientes.


X-ray-model
X-ray train (ResNet50)
by yochz
SUCCESS
Duration: 33m 54s
X-ray
All Flow Experiments
X-ray
ResNet50
Research
Files
Workspaces
Experiments
Flows
Serving
Apps
Settings
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
Model: resnet50
GPUs found: 1
tensorflow local version: 2.0.0
loss
Values
cnvrg.io


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

Conclusión

NetApp y cnvrg.io se han asociado para ofrecer a los clientes una solución completa de gestión de datos para el desarrollo de software ML y DL. ONTAP AI proporciona almacenamiento y computación de alto rendimiento para cualquier escala de operación y el software cnvrg.io optimiza los flujos de trabajo de ciencia de datos y mejora la utilización de recursos.

Reconocimientos

- Mike Oglesby, Ingeniero Técnico de Marketing de NetApp
- Santosh Rao, director técnico sénior, NetApp

Dónde encontrar información adicional

Si quiere obtener más información sobre el contenido de este documento, consulte los siguientes recursos:

- Nvrg.io ("https://cnvrg.io"):
  - NÚCLEO Cnvrg (plataforma DE ML libre)

    https://cnvrg.io/platform/core
  - Documentos de Cnvrg

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

    https://www.nvidia.com/en-us/data-center/dgx-1/
  - GPU de núcleo tensor NVIDIA Tesla V100

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

    https://www.nvidia.com/en-us/gpu-cloud/
- Sistemas AFF de NetApp:
  - Especificaciones técnicas de AFF

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

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

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

- Informe técnico de NetApp FlexGroup

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

- Almacenamiento persistente de NetApp para contenedores:
  - Trident de NetApp

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

- Matriz de interoperabilidad de NetApp:
  - Herramienta de matriz de interoperabilidad de NetApp

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

- Redes de IA de ONTAP:
  - Switches Cisco Nexus 3232C

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

  - Switches Mellanox Spectrum serie 2000

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

- Marco Y herramientas DE ML:
  - DALÍ

    https://github.com/NVIDIA/DALI

  - TensorFlow: Un marco de aprendizaje automático de código abierto para todos

    https://www.tensorflow.org/

  - Horovod: El marco de Uber para el aprendizaje automático distribuido de código abierto para TensorFlow

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

  - Habilitación de GPU en el ecosistema Container Runtime

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

  - Docker

    https://docs.docker.com

  - Kubernetes

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

  - DeepOps de NVIDIA

    https://github.com/NVIDIA/deepops

  - Kubeflow

http://www.kubeflow.org/

- Servidor de portátiles Juppyter

  http://www.jupyter.org/

- Conjuntos de datos y pruebas:
  - Conjunto de datos de radiografía de tórax de los NIH

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

  - Xiaosong Wang, Yifan Peng, le Lu, Zhiyong Lu, Mohammadadi Bagheri, Ronald Summers, ChestX-ray8: Base de datos de rayos X en el pecho a escala de hospital y puntos de referencia sobre clasificación y localización de enfermedades comunes del tórax, IEEE CVPR, págs 3462-3471, 2017TR-4841-0620

Informe técnico TR-4732: Big Data Analytics datos en inteligencia artificial

Karthikeyan Nagalingam, NetApp

Este documento describe cómo trasladar los datos de análisis de Big Data y los datos de computación de alto rendimiento a IA. La IA procesa datos de NFS a través de exportaciones NFS, mientras que los clientes suelen tener sus datos de IA en una plataforma de análisis de Big Data, como HDFS, Blob o S3, así como plataformas HPC como GPFS. Este documento proporciona directrices para trasladar datos de análisis de Big Data y datos de HPC a IA mediante NetApp XCP y NIPAM. También hablamos de las ventajas empresariales que supone trasladar datos de Big Data y de HPC a IA.

Conceptos y componentes

Almacenamiento de análisis de Big Data

Los análisis de Big Data son el principal proveedor de almacenamiento para HDFS. Un cliente suele utilizar un sistema de archivos compatible con Hadoop (HCFS), como almacenamiento blob de Windows Azure, MapR File System (MapR-FS) y almacenamiento de objetos S3.

Sistema de archivos paralelos general

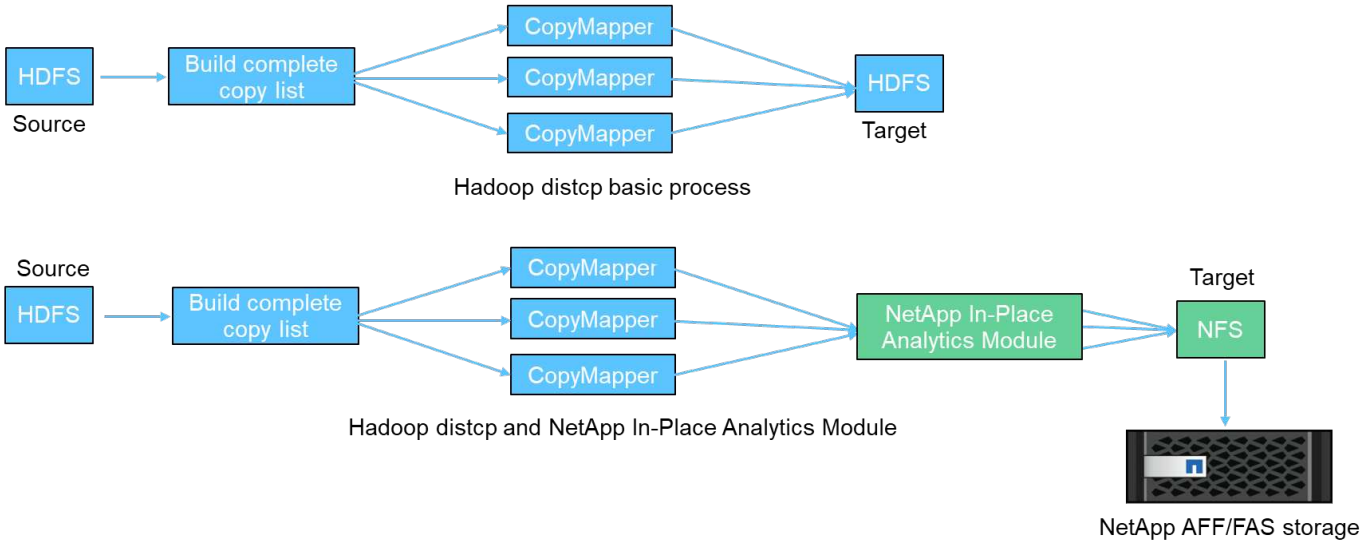
GPFS de IBM es un sistema de archivos empresariales que ofrece una alternativa a HDFS. GPFS proporciona flexibilidad para que las aplicaciones decidan el tamaño del bloque y el diseño de replicación, lo que proporciona un buen rendimiento y eficiencia.

Módulo de análisis in situ de NetApp

El módulo de análisis in situ (NIPAM, in situ) de NetApp sirve como controlador para que los clústeres de Hadoop accedan a datos NFS. Consta de cuatro componentes: Un pool de conexión, un InputStream NFS, una caché de gestión de archivos y un OutputStream NFS. Para obtener más información, consulte "TR-4382: Módulo de análisis in situ de NetApp."

Copia distribuida de Hadoop

La copia distribuida de Hadoop (DistCp) es una herramienta de copia distribuida que se usa para las tareas de adaptación entre clústeres y dentro de clústeres de gran tamaño. Esta herramienta utiliza MapReduce para la distribución de datos, el manejo de errores y los informes. Amplía la lista de archivos y directorios e introduce las tareas de asignación para copiar los datos de la lista de origen. La siguiente imagen muestra la operación DistCp en HDFS y no HDFS.



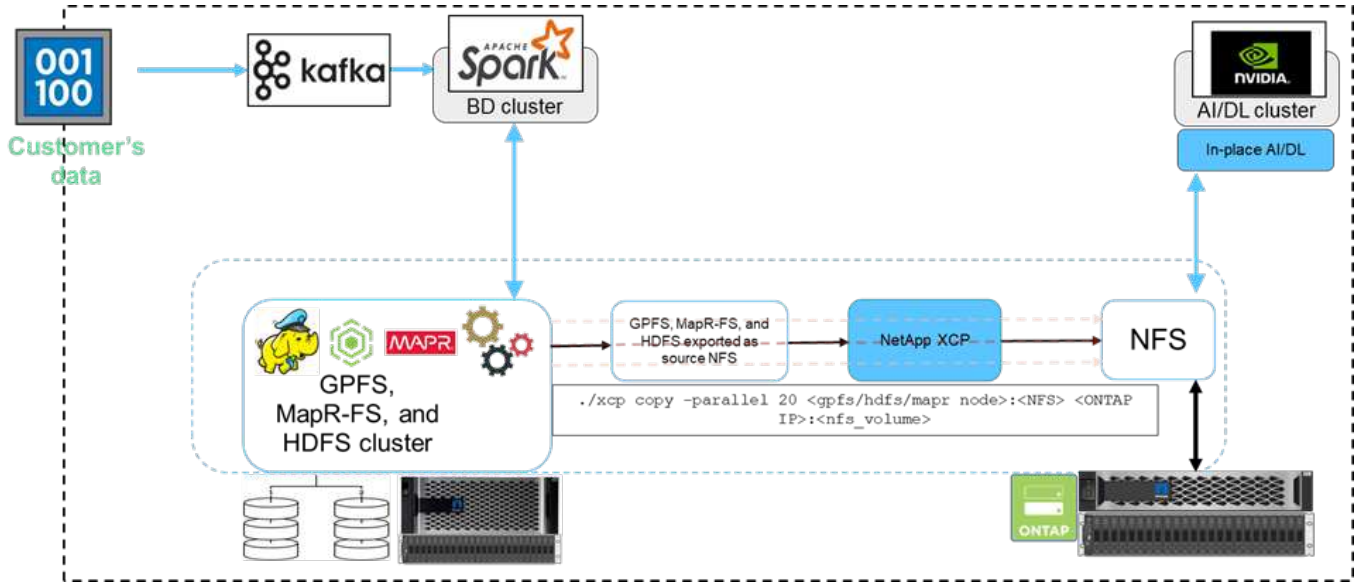
Hadoop DistCp mueve datos entre los dos sistemas HDFS sin necesidad de utilizar un controlador adicional. NetApp proporciona el controlador para sistemas que no son HDFS. En un destino NFS, NIPAM proporciona el controlador para copiar datos que Hadoop DistCp utiliza para comunicarse con destinos NFS al copiar datos.

Cloud Volumes Service de NetApp

Cloud Volumes Service de NetApp es un servicio de archivos nativo del cloud con un rendimiento extremo. Este servicio ayuda a sus clientes a acelerar el plazo de comercialización mediante el rápido aumento y reducción de los recursos, así como el uso de las funciones de NetApp para mejorar la productividad y reducir el tiempo de inactividad del personal. Cloud Volumes Service es la alternativa adecuada para la recuperación ante desastres y sus backups en el cloud, ya que reduce el espacio global del centro de datos y consume menos almacenamiento en cloud público nativo.

XCP de NetApp

NetApp XCP es un software cliente que permite una migración de datos de cualquiera a NetApp y de NetApp a NetApp rápida y fiable. Esta herramienta está diseñada para copiar una gran cantidad de datos NAS no estructurados de cualquier sistema NAS a una controladora de almacenamiento de NetApp. La herramienta de migración XCP utiliza un motor de transmisión de E/S multicanal y de varios canales que puede procesar muchas solicitudes en paralelo, como la migración de datos, listas de archivos o directorios y la creación de informes de espacio. Esta es la herramienta de migración de datos de NetApp predeterminada. Puede utilizar XCP para copiar datos de un clúster de Hadoop y de una HPC al almacenamiento NFS de NetApp. El siguiente diagrama muestra la transferencia de datos de un clúster Hadoop y HPC a un volumen NFS de NetApp mediante XCP.



Copia y sincronización de NetApp BlueXP

La copia y sincronización de BlueXP de NetApp es un software como servicio de replicación de datos híbrida que transfiere y sincroniza datos de NFS, S3 y CIFS sin problemas y de forma segura entre el almacenamiento on-premises y el almacenamiento en la nube. Este software se utiliza para migración de datos, archivado, colaboración, análisis, etc. Una vez transferidos los datos, BlueXP Copy y Sync sincronizan continuamente los datos entre el origen y el destino. De cara al futuro, luego transfiere el delta. También protege los datos dentro de su propia red, en el cloud o en las instalaciones. Este software se basa en el modelo de pago por uso, que ofrece una solución rentable y ofrece capacidades de supervisión y generación de informes para su transferencia de datos.

Inferencia de IA en el perímetro - NetApp con Lenovo ThinkSystem - Diseño de la solución

TR-4886: Inferencia de IA en el Edge - NetApp con Lenovo ThinkSystem - Diseño de la solución

Sathish Thyagarajan, NetApp Miroslav Hodak, Lenovo

Resumen

Varios supuestos de aplicaciones emergentes, como los sistemas avanzados de asistencia al conductor (ADAS), el sector 4.0, las ciudades inteligentes y el Internet de las cosas (IoT), requieren el procesamiento de flujos de datos continuos con una latencia cercana a cero. Este documento describe una arquitectura de computación y almacenamiento para poner en marcha la inferencia de inteligencia artificial (IA) basada en GPU en controladoras de almacenamiento de NetApp y servidores Lenovo ThinkSystem en un entorno perimetral que satisface estos requisitos. Este documento también proporciona datos de rendimiento para las pruebas de rendimiento de inferencia MLPerf estándares del sector, por lo que evalúa diversas tareas de inferencia en servidores periféricos equipados con GPU T4 de NVIDIA. Investigamos el rendimiento de escenarios de inferencia multisecuencia, sin conexión y con múltiples flujos, y mostramos que la arquitectura con un sistema de almacenamiento en red compartido rentable tiene un alto rendimiento y proporciona un punto central para la gestión de modelos y datos en servidores periféricos múltiples.

Introducción

Las empresas están generando cada vez más volúmenes masivos de datos en el extremo de la red. Con el fin de obtener el máximo valor a partir de sensores inteligentes y datos del Internet de las cosas, las organizaciones buscan una solución de transmisión de eventos en tiempo real que permita el uso de la informática perimetral. Así, pues, cada vez se realizan más tareas informáticas exigentes fuera del perímetro y fuera de los centros de datos. La inferencia de IA es uno de los motores de esta tendencia. Los servidores periféricos proporcionan suficiente potencia computacional para estas cargas de trabajo, sobre todo cuando se utilizan aceleradores, pero el almacenamiento limitado suele ser un problema, especialmente en entornos con varios servidores. En este documento mostramos cómo puede poner en marcha un sistema de almacenamiento compartido en un entorno perimetral y cómo beneficia a las cargas de trabajo de inferencia de IA sin perjudicar el rendimiento.

Este documento describe una arquitectura de referencia para la inferencia de IA en el perímetro. Combina múltiples servidores Lenovo ThinkSystem Edge con un sistema de almacenamiento de NetApp para crear una solución fácil de poner en marcha y gestionar. Está previsto ser una guía de referencia para implementaciones prácticas en diversas situaciones, como la planta de fábrica con múltiples cámaras y sensores industriales, sistemas de punto de venta (POS) en transacciones de venta al por menor o sistemas de autoconducción completa (FSD) que identifican anomalías visuales en vehículos autónomos.

Este documento recoge las pruebas y validación de una configuración de computación y almacenamiento que consta de Lenovo ThinkSystem SE350 Edge Server y un sistema de almacenamiento AFF y EF-Series de NetApp de gama básica. Las arquitecturas de referencia proporcionan una solución eficiente y rentable para puestas en marcha de IA, a la vez que proporcionan servicios de datos completos, protección de datos integrada, escalabilidad fluida y almacenamiento de datos conectado al cloud con ONTAP de NetApp y el software para la gestión de datos SANtricity de NetApp.

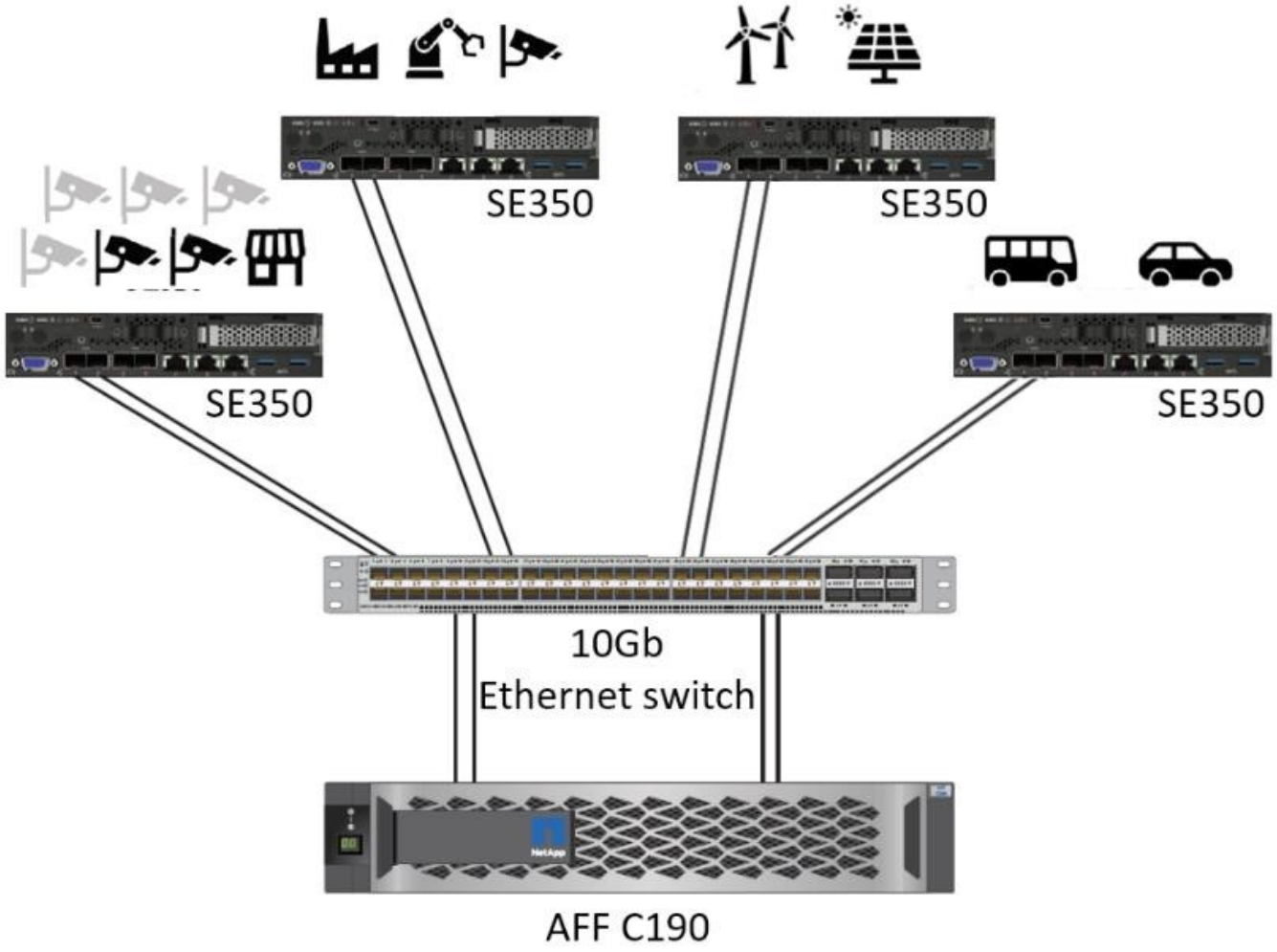
Público objetivo

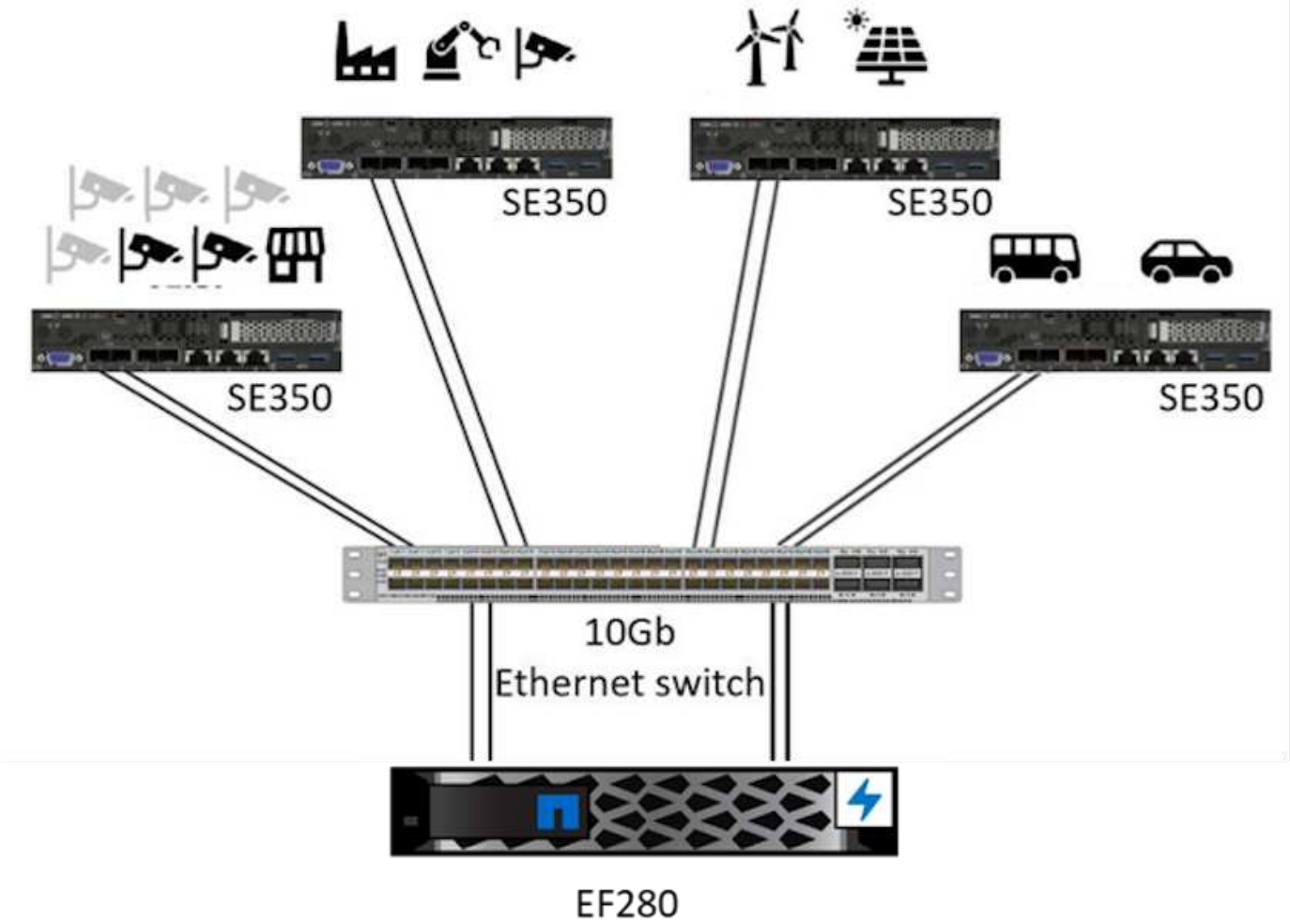
Este documento está dirigido a los siguientes destinatarios:

- Líderes de negocio y arquitectos empresariales que quieren aprovechar la IA en el perímetro.
- Científicos de datos, ingenieros de datos, investigadores DE IA/aprendizaje automático y desarrolladores de sistemas de IA.
- Arquitectos empresariales que diseñan soluciones para el desarrollo de modelos y aplicaciones de IA/ML.
- Científicos e ingenieros de IA que buscan formas eficientes de poner en marcha modelos DE aprendizaje profundo (DL) Y ML.
- Los administradores de dispositivos periféricos y los administradores de servidor perimetral son responsables de la puesta en marcha y la gestión de modelos de inferencia perimetrales.

Arquitectura de la solución

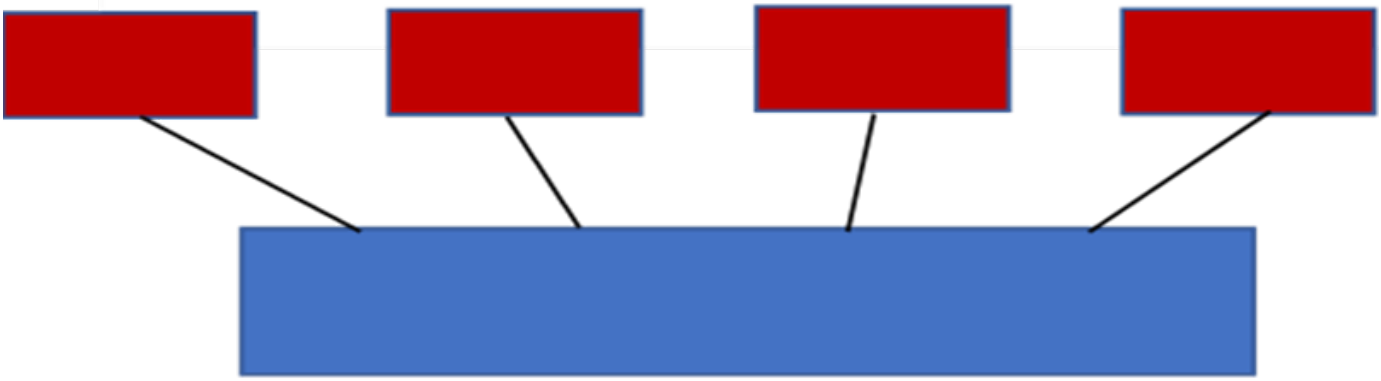
Este servidor Lenovo ThinkSystem y la solución de almacenamiento ONTAP o SANtricity de NetApp están diseñados para gestionar la inferencia de IA en grandes conjuntos de datos mediante la potencia de procesamiento de GPU junto con CPU tradicionales. Esta validación demuestra un alto rendimiento y una gestión de datos óptima con una arquitectura que utiliza uno o varios servidores de borde Lenovo SR350 interconectados con un único sistema de almacenamiento AFF de NetApp, como se muestra en las siguientes dos figuras.


SE350
SE350
SE350
SE350
10Gb
Ethernet switch
AFF C190



La descripción general de la arquitectura lógica de la siguiente figura muestra las funciones de los elementos de computación y almacenamiento de esta arquitectura. Específicamente, muestra lo siguiente:

- Dispositivos de computación periféricos que realizan inferencia en los datos que recibe de cámaras, sensores, etc.
- Un elemento de almacenamiento compartido que ofrece diversos objetivos:
  - Ofrece una ubicación central para los modelos de inferencia y otros datos que necesitan para realizar la inferencia. Los servidores informáticos acceden al almacenamiento directamente y utilizan los modelos de inferencia en toda la red sin necesidad de copiarlos de forma local.
  - Los modelos actualizados se empujan aquí.
  - Archiva los datos de entrada que los servidores perimetrales reciben para analizarlos posteriormente. Por ejemplo, si los dispositivos periféricos están conectados a las cámaras, el elemento de almacenamiento mantiene los vídeos capturados por las cámaras.



rojo	azul
Sistema de computación Lenovo	Sistema de almacenamiento AFF de NetApp
Dispositivos periféricos que realizan inferencia en las entradas de cámaras, sensores, etc.	Almacenamiento compartido que mantiene los modelos de inferencia y los datos de los dispositivos periféricos para su posterior análisis.

Esta solución de NetApp y Lenovo ofrece las siguientes ventajas clave:

- GPU acelera la computación en el extremo.
- Implementación de varios servidores periféricos, respaldados y gestionado desde un almacenamiento compartido.
- Protección de datos sólida para cumplir con los objetivos de punto de recuperación (RPO) y los objetivos de tiempo de recuperación (RTO) bajos sin pérdida de datos.
- Gestión de datos optimizada con copias Snapshot y clones de NetApp para optimizar los flujos de trabajo de desarrollo.

Cómo utilizar esta arquitectura

Este documento valida el diseño y el rendimiento de la arquitectura propuesta. Sin embargo, no hemos probado ciertas piezas a nivel de software, como la gestión y sincronización de datos de contenedores, cargas de trabajo o modelos con cloud o centros de datos en las instalaciones, ya que son específicas de un escenario de puesta en marcha. Aquí existen varias opciones.

En el nivel de gestión de contenedores, la gestión de contenedores de Kubernetes es una buena opción y es compatible con una versión totalmente ascendente (Canonical) o con una versión modificada adecuada para implementaciones empresariales (Red Hat). La "Plano de control de IA de NetApp" Que utiliza Trident de NetApp y los recientemente añadidos "Kit de herramientas de operaciones de datos de NetApp" Proporciona trazabilidad, funciones de gestión de datos, interfaces y herramientas integradas para que los científicos e ingenieros de datos se integren con el sistema de almacenamiento de NetApp. Kubeflow, el kit DE herramientas ML para Kubernetes, proporciona funcionalidades de IA adicionales junto con compatibilidad para el control de versiones de modelos y KFServing en varias plataformas como TensorFlow Serving o NVIDIA Triton inferpensado Server. Otra opción es la plataforma NVIDIA EGX, que proporciona gestión de las cargas de trabajo junto con acceso a un catálogo de contenedores de inferencia de IA habilitados para GPU. Sin embargo, es posible que estas opciones requieran gran esfuerzo y experiencia para ponerlas en producción y es posible que requieran la ayuda de un proveedor de software independiente (ISV) o un asesor de terceros.

La ventaja clave de la inferencia de IA y la computación perimetral es la capacidad de que los dispositivos calculen, procesen y analicen datos con un alto nivel de calidad sin latencia. Hay demasiados ejemplos de casos de uso de computación perimetral que describir en este documento, pero aquí hay algunos ejemplos destacados:

Automóviles: Vehículos autónomos

La ilustración clásica de la informática avanzada se encuentra en los sistemas avanzados de asistencia al conductor (ADAS) en vehículos autónomos (AV). La IA en vehículos sin conductor debe procesar rápidamente una gran cantidad de datos procedentes de cámaras y sensores para garantizar su seguridad. Tomar demasiado tiempo para interpretar entre un objeto y un humano puede significar la vida o la muerte, por lo tanto poder procesar los datos lo más cerca posible del vehículo es crucial. En este caso, uno o varios servidores de computación periféricos se encarga de las entradas de cámaras, RADAR, LiDAR y otros sensores, mientras que el almacenamiento compartido contiene modelos de inferencia y almacena datos de entrada de los sensores.

Atención sanitaria: Monitorización de pacientes

Uno de los mayores impactos de la IA y la informática perimetral es su capacidad para mejorar la supervisión continua de pacientes para enfermedades crónicas, tanto en las unidades de cuidados intensivos como en las unidades de cuidados intensivos (UCI). Los datos de los dispositivos periféricos que supervisan los niveles de insulina, la respiración, la actividad neurológica, el ritmo cardíaco y las funciones gastrointestinales requieren un análisis instantáneo de los datos que deben ser objeto de acciones inmediatas porque hay poco tiempo para actuar y salvar la vida de alguien.

Venta al por menor: Pago sin cajero

La computación perimetral puede ayudar a los minoristas a reducir el tiempo de salida y aumentar el tráfico de pies. Los sistemas sin cajero admiten varios componentes, como los siguientes:

- Autenticación y acceso. Conectar el comprador físico a una cuenta validada y permitir el acceso al espacio de venta al por menor.
- Supervisión de inventario. Utilizar sensores, etiquetas RFID y sistemas de visión computarizada para confirmar la selección o deselección de artículos por parte de los compradores.

  Aquí, cada uno de los servidores perimetrales gestiona cada contador de retirada y el sistema de almacenamiento compartido sirve como punto de sincronización central.

Servicios financieros: Seguridad humana en quioscos y prevención del fraude

Las organizaciones bancarias utilizan la IA y la informática perimetral para innovar y crear experiencias bancarias personalizadas. Los quioscos interactivos, mediante el análisis de datos en tiempo real y la inferencia de IA, permiten ahora a los cajeros automáticos no sólo ayudar a los clientes a retirar el dinero, sino también supervisar de forma proactiva los quioscos a través de las imágenes capturadas con las cámaras para identificar el riesgo para la seguridad humana o el comportamiento fraudulento. En este escenario, los servidores periféricos informáticos y los sistemas de almacenamiento compartido se conectan a quioscos y cámaras interactivos para ayudar a los bancos a recopilar y procesar datos con modelos de inferencia de IA.

Fabricación: Industria 4.0

La cuarta revolución industrial (Industry 4.0) ha comenzado, junto con tendencias emergentes como Smart Factory e impresión 3D. Con el fin de prepararse para un futuro impulsado por los datos, la comunicación y el

Internet de las cosas (M2M) a gran escala están integrados para una mayor automatización sin necesidad de intervención humana. La fabricación ya está altamente automatizada y añadir características de IA es una continuación natural de la tendencia a largo plazo. La IA permite automatizar operaciones que se pueden automatizar con la ayuda de la visión computarizada y otras funcionalidades de IA. Puede automatizar el control de calidad o las tareas que se basan en la visión humana o en la toma de decisiones para realizar análisis más rápidos de materiales en líneas de ensamblaje en plantas de fabricación para ayudar a las plantas de fabricación a cumplir con los estándares ISO requeridos de gestión de la seguridad y la calidad. Aquí, cada servidor perimetral informático está conectado a una matriz de sensores que supervisan el proceso de fabricación, y cuando es necesario, los modelos de inferencia actualizados se ven empujados al almacenamiento compartido.

Telecomunicaciones: Detección de óxido, inspección de torre y optimización de la red

La industria de las telecomunicaciones utiliza técnicas de IA y visión informática para procesar imágenes que detectan automáticamente el óxido e identifican las torres celulares que contienen corrosión y, por lo tanto, requieren una inspección adicional. El uso de imágenes con drones y modelos de IA para identificar regiones distintas de una torre para analizar la oxidación, las grietas superficiales y la corrosión ha aumentado en los últimos años. La demanda continúa creciendo para tecnologías de IA que permiten inspeccionar eficientemente la infraestructura de telecomunicaciones y las torres de células, evaluar periódicamente la degradación y reparar rápidamente cuando sea necesario.

Además, otro caso de uso emergente en las telecomunicaciones es el uso de algoritmos de IA y ML para predecir patrones de tráfico de datos, detectar dispositivos compatibles con 5G y automatizar y aumentar la gestión energética de múltiples entradas y salidas múltiples (MIMO). El hardware de MIMO se utiliza en las torres de radio para aumentar la capacidad de la red; sin embargo, esto viene con costos de energía adicionales. Los modelos ML para “modo de suspensión MIMO” implementados en las zonas de células pueden predecir el uso eficiente de las radios y ayudar a reducir los costes de consumo de energía para los operadores de redes móviles (MNO). Las soluciones de computación avanzada y de inferencia de IA ayudan a las MNO a reducir la cantidad de datos transmitidos hacia los centros de datos, reducir su TCO, optimizar las operaciones de red y mejorar el rendimiento general de los usuarios finales.

Información general de la tecnología

Esta sección describe la base tecnológica de esta solución de IA.

Sistemas AFF de NetApp

Los sistemas de almacenamiento AFF de NetApp más innovadores permiten que las puestas en marcha de inferencia de IA en el perímetro satisfagan los requisitos de almacenamiento de la empresa con un rendimiento líder del sector, una flexibilidad superior, integración con el cloud y la mejor gestión de datos de su clase. Los sistemas AFF de NetApp han sido diseñados específicamente para flash y ayudan a acelerar, gestionar y proteger los datos esenciales para la empresa.

- Los sistemas de almacenamiento AFF de NetApp de gama básica se basan en hardware FAS2750 y medios flash SSD
- Dos controladoras en configuración de alta disponibilidad

Los sistemas de almacenamiento C190 de gama básica de AFF admiten las siguientes funciones:

- Un número máximo de unidades de estado sólido de 960 GB
- Dos configuraciones posibles:
  - Ethernet (10 GbE): 4 puertos 10GBASE-T (RJ-45)
  - Unificado (FC de 16 GB o 10 GbE): 4 puertos de adaptador de destino unificado 2 (UTA2)
- Una capacidad efectiva máxima de 50,5 TB

> En el caso de cargas de trabajo NAS, un único sistema AFF C190 de gama básica admite un rendimiento de 4,4 GB/s para lecturas secuenciales y 230 000 IOPS para lecturas aleatorias pequeñas con latencias de 1 ms o menos.

AFF A220 de NetApp

NetApp también ofrece otros sistemas de almacenamiento de gama básica que proporcionan un mayor rendimiento y escalabilidad para las puestas en marcha a gran escala. En el caso de cargas de trabajo NAS, un único sistema AFF A220 de gama básica admite:

- Rendimiento de 6,2 Gbps para lecturas secuenciales
- 3375 000 IOPS para lecturas aleatorias pequeñas con latencias de 1 ms o menos
- Recuento máximo de unidades de SSD 144 TB, 3,8 TB o 7,6 TB
- AFF A220 es escalable a más de 1 PB de capacidad efectiva

AFF A250 de NetApp

- La capacidad efectiva máxima es de 35 PB con el escalado horizontal máximo de 2-24 nodos (12 parejas de alta disponibilidad).
- Ofrece un aumento del rendimiento de ≥ 45 % con respecto a AFF A220
- 440 000 IOPS lecturas aleatorias a 1 ms
- Basado en el último lanzamiento de ONTAP de NetApp: ONTAP 9.8

- Aprovecha dos Ethernet de 25 GB para alta disponibilidad e interconexión de clúster

Sistemas E-Series EF de NetApp

EF-Series es una familia de cabinas DE almacenamiento SAN all-flash de gama básica y media que pueden acelerar el acceso a sus datos y ayudarle a obtener valor de ella más rápido con el software SANtricity de NetApp. Estos sistemas ofrecen almacenamiento flash SAS y NVMe y le proporcionan IOPS asequibles hasta extremas, tiempos de respuesta inferiores a 100 microsegundos y un ancho de banda de hasta 44 Gbps, lo cual los convierte en ideales para cargas de trabajo mixtas y aplicaciones exigentes como la inferencia de IA y la computación de alto rendimiento (HPC).

La siguiente figura muestra el sistema de almacenamiento EF280 de NetApp.

EF280 de NetApp

- Compatibilidad con FC de 32 GB/16 GB, iSCSI de 25 GB/10 GB y SAS de 12 GB
- La capacidad efectiva máxima es 96 unidades que totalizan 1,5 PB
- Rendimiento de 10 Gbps (lecturas secuenciales)
- 300K IOPS (lectura aleatoria)
- El sistema EF280 de NetApp es la cabina all-flash (AFA) más económica de la cartera de NetApp

EF300 de NetApp

- 24 unidades SSD NVMe para una capacidad total de 367 TB
- Opciones de expansión con un total de 240 HDD NL-SAS, 92 SSD SAS o una combinación de ambos
- 100 GB NVMe/IB, NVMe/roce, Iser/IB y SRP/IB
- NVME DE 32 GB/FC, FCP
- ISCSI de 25 GB
- 20 Gbps (lecturas secuenciales)
- 670K IOPS (lecturas aleatorias)

Para obtener más información, consulte "Especificaciones técnicas de las cabinas all-flash EF-Series de NetApp EF600, F300, EF570 y EF280".

ONTAP 9 de NetApp

ONTAP 9.8.1, la última generación del software de gestión del almacenamiento de NetApp, permite a las empresas modernizar su infraestructura y realizar la transición a un centro de datos preparado para el cloud. ONTAP ofrece las mejores capacidades de gestión de datos y permite la gestión y protección de los datos con un solo conjunto de herramientas, sin importar dónde residan. También puede mover los datos libremente a donde sea necesario: El perímetro, el núcleo o el cloud. ONTAP 9.8.1 incluye numerosas funciones que simplifican la gestión de datos, aceleran y protegen los datos esenciales y permiten disfrutar de funcionalidades de infraestructura de nueva generación en arquitecturas de cloud híbrido.

Simplificar la gestión de los datos

La gestión de los datos es crucial para las operaciones TECNOLÓGICAS empresariales, de modo que se utilicen recursos apropiados para las aplicaciones y conjuntos de datos. ONTAP incluye las siguientes funciones para facilitar y simplificar las operaciones, y reducir el coste total de las operaciones:

- **Compactación de datos inline y deduplicación expandida.** la compactación de datos reduce el espacio perdido dentro de los bloques de almacenamiento, y la deduplicación aumenta significativamente la capacidad efectiva. Esto es aplicable a los datos almacenados localmente y a los datos organizados en niveles en el cloud.
- **Calidad de servicio (AQoS) mínima, máxima y adaptativa.** los controles de calidad de servicio (QoS) granulares ayudan a mantener los niveles de rendimiento para aplicaciones críticas en entornos altamente compartidos.
- **FabricPool de NetApp.** esta función ofrece una organización automática por niveles de datos inactivos a opciones de almacenamiento en cloud privado o público, como la solución de almacenamiento Amazon Web Services (AWS), Azure y StorageGRID de NetApp. Para obtener más información sobre FabricPool, consulte "TR-4598".

Acelere y proteja sus datos

ONTAP 9 ofrece niveles superiores de rendimiento y protección de datos, y amplía estas capacidades de las siguientes maneras:

- **Rendimiento y menor latencia.** ONTAP ofrece el rendimiento más alto posible con la menor latencia posible.
- **Protección de datos.** ONTAP ofrece capacidades integradas de protección de datos con administración común en todas las plataformas.
- **Cifrado de volumen de NetApp (NVE).** ONTAP ofrece cifrado nativo a nivel de volumen con compatibilidad para gestión de claves incorporada y externa.
- **Multitenancy y autenticación multifactor.** ONTAP permite compartir recursos de infraestructura con los niveles más altos de seguridad.

Infraestructura preparada para futuros retos

ONTAP 9 ayuda a satisfacer las exigentes y siempre cambiantes necesidades de la empresa con las siguientes funciones:

- **Escalado sencillo y operaciones no disruptivas.** ONTAP admite la adición sin interrupciones de capacidad a las controladoras existentes y a los clústeres de escalado horizontal. Los clientes pueden empezar a utilizar tecnologías punteras como NVMe y FC 32 GB, sin necesidad de realizar costosas migraciones de datos y sin cortes.
- **Conexión en cloud.** ONTAP es el software de gestión del almacenamiento con mejor conexión en cloud,

con opciones de almacenamiento definido por software (ONTAP Select) e instancias nativas del cloud (Cloud Volumes Service de NetApp) en todos los clouds públicos.

- **Integración con aplicaciones emergentes.** ONTAP ofrece servicios de datos de clase empresarial para plataformas y aplicaciones de última generación, como vehículos autónomos, ciudades inteligentes e Industria 4.0, utilizando la misma infraestructura que soporta las aplicaciones empresariales existentes.

SANtricity de NetApp

SANtricity de NetApp está diseñado para ofrecer un rendimiento líder del sector, fiabilidad y simplicidad en las cabinas all-flash EF-Series y flash híbrido. Logre el máximo rendimiento y aprovechamiento de sus cabinas all-flash EF-Series y flash híbridas de E-Series para aplicaciones de cargas de trabajo pesadas, como análisis de datos, videovigilancia y backup y recuperación de datos. Con SANtricity, los ajustes de configuración, el mantenimiento, la expansión de la capacidad y otras tareas se pueden realizar mientras el almacenamiento sigue online. SANtricity también ofrece una protección de datos superior, supervisión proactiva y seguridad certificada: Todo accesible a través de la sencilla interfaz integrada de System Manager. Para obtener más información, consulte "Especificaciones técnicas del software SANtricity para E-Series de NetApp".

Rendimiento optimizado

El software SANtricity optimizado para el rendimiento ofrece datos —con una alta tasa de IOPS, un alto rendimiento y una baja latencia— a todas sus aplicaciones de análisis de datos, videovigilancia y backup. Acelere el rendimiento en aplicaciones con baja latencia y una alta tasa de IOPS, y en aplicaciones con requisitos de amplio ancho de banda y un rendimiento alto.

Maximice el tiempo de actividad

Complete todas las tareas de gestión mientras el almacenamiento sigue en línea. Modifique las configuraciones, realice tareas de mantenimiento o amplíe la capacidad sin interrumpir el flujo de I/O. Proporcione los mayores niveles de fiabilidad posibles con funciones automatizadas, opciones de configuración en línea, la tecnología de pools de discos dinámicos (DPP) de vanguardia, entre otras.

Esté tranquilo

El software SANtricity ofrece una protección de datos superior, supervisión proactiva y seguridad certificada: Todo a través de la sencilla interfaz incluida de System Manager. Simplifique las tareas de gestión del almacenamiento. Obtenga la flexibilidad que necesita para realizar un mejor ajuste de todos los sistemas de almacenamiento E-Series. Gestione su sistema E-Series de NetApp en todo momento y en cualquier lugar. Nuestra interfaz integrada basada en web optimiza el flujo de trabajo de gestión.

Trident de NetApp

"Trident" De NetApp es un orquestador de almacenamiento dinámico de código abierto para Docker y Kubernetes que simplifica la creación, la gestión y el consumo de almacenamiento persistente. Trident, una aplicación nativa de Kubernetes, se ejecuta directamente dentro de un clúster de Kubernetes. Trident permite que los clientes implementen sin problemas imágenes de contenedores de DL en el almacenamiento de NetApp y proporciona una experiencia de clase empresarial para implementaciones de contenedores de IA. Los usuarios de Kubernetes (como desarrolladores DE ML y científicos de datos) pueden crear, gestionar y automatizar la orquestación y el clonado para aprovechar las funcionalidades de gestión de datos avanzadas de NetApp, impulsadas por la tecnología de NetApp.

Copia y sincronización de NetApp BlueXP

"Copia y sincronización de BlueXP" Es un servicio de NetApp que ofrece una sincronización de datos rápida y segura. Ya tenga que transferir archivos entre recursos compartidos de archivos NFS o SMB en las

instalaciones, 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, Google Cloud Storage, o IBM Cloud Object Storage, BlueXP Copy and Sync mueve los archivos a donde los necesites de forma rápida y segura. Una vez transferidos los datos, estarán completamente disponibles para su uso tanto en origen como en destino. BlueXP Copy and Sync sincroniza los datos de forma continua en función de tu programación predefinida, moviendo solo los deltas, por lo que se reducen al mínimo el tiempo y el dinero que se invierten en la replicación de datos. Copia y sincronización de BlueXP es una herramienta de software como servicio (SaaS) extremadamente sencilla de configurar y usar. Las transferencias de datos activadas por BlueXP Copy and Sync se llevan a cabo por agentes de datos. Puedes poner en marcha agentes de datos de BlueXP Copy y Sync en AWS, Azure, Google Cloud Platform o en las instalaciones.

Servidores Lenovo ThinkSystem

Los servidores Lenovo ThinkSystem incluyen hardware, software y servicios innovadores que resuelven los desafíos actuales de los clientes y ofrecen un enfoque de diseño modular, evolutivo y adecuado para su propósito para afrontar los desafíos del futuro. Estos servidores se capitalizan en las mejores tecnologías estándar del sector, junto con innovaciones diferenciadas de Lenovo, para proporcionar la mayor flexibilidad posible en servidores x86.

Las ventajas clave de la implementación de servidores Lenovo ThinkSystem incluyen:

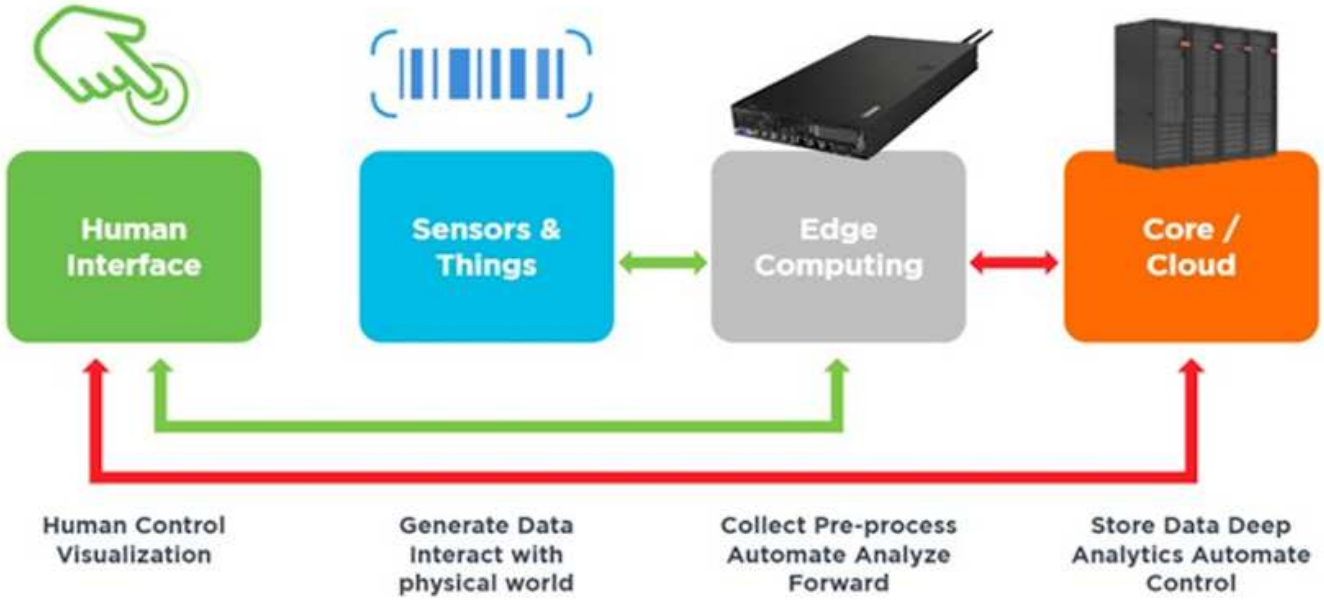
- Diseños modulares y altamente escalables que crecen a medida que lo hace su negocio
- Resiliencia líder en el sector para ahorrar horas de costosos tiempos de inactividad no programados
- Tecnologías flash rápidas para reducir las latencias, acelerar los tiempos de respuesta y gestionar los datos de forma más inteligente en tiempo real

En el ámbito de la IA, Lenovo está adoptando un enfoque práctico para ayudar a las empresas a comprender y adoptar las ventajas DEL APRENDIZAJE AUTOMÁTICO y la IA para sus cargas de trabajo. Los clientes de Lenovo pueden explorar y evaluar las ofertas de IA de Lenovo en los centros de innovación de IA de Lenovo para comprender por completo el valor de su caso de uso en particular. Con el fin de mejorar la rentabilidad de la inversión, este enfoque centrado en el cliente proporciona a los clientes una prueba de concepto para las plataformas de desarrollo de soluciones que están listas para usar y optimizadas para la IA.

Servidor Lenovo ThinkSystem SE350 Edge

La computación perimetral permite analizar los datos de dispositivos de IoT en el extremo de la red antes de enviarlos al centro de datos o al cloud. El sistema ThinkSystem SE350 de Lenovo, como se muestra en la siguiente figura, está diseñado para los requisitos únicos de implementación en el perímetro, con un enfoque en flexibilidad, conectividad, seguridad y capacidad de gestión remota en un factor de forma compacto y reforzado con el medio ambiente.

Con el procesador Intel Xeon D con la flexibilidad necesaria para admitir la aceleración de las cargas de trabajo de IA perimetral, el SE350 se ha diseñado específicamente para afrontar los retos de las implementaciones de servidores en una gran variedad de entornos fuera del centro de datos.



Rendim. MLPerf

MLPerf es el conjunto de pruebas de rendimiento líder del sector para evaluar el rendimiento de la IA. Cubre muchas áreas de IA aplicada, incluida la clasificación de imágenes, la detección de objetos, las imágenes médicas y el procesamiento del lenguaje natural (NLP). En esta validación, hemos utilizado cargas de trabajo de inferencia v0.7, que es la última iteración de la inferencia MLPerf al finalizar esta validación. La "Inferencia del rendimiento ML0,7" la suite incluye cuatro nuevas pruebas de rendimiento para centros de datos y sistemas periféricos:

- **BERT.** representación de encoder bidireccional de Transformers (BERT) ajustada para responder preguntas utilizando el conjunto de datos de escuadrón.
- **DLRM.** el modelo de recomendación de aprendizaje profundo (DLRM) es un modelo de personalización y recomendación que se entrena para optimizar las tarifas de clic (CTR).
- **3D U-Net.** la arquitectura 3D U-Net está entrenada en el conjunto de datos de segmentación del tumor cerebral (Brats).

- **RNN-T.** el transductor de red neural recurrente (RNN-T) es un modelo de reconocimiento automático de voz (ASR) que se entrena en un subconjunto de LibriSpeech. Los resultados y el código de la inferencia de MLPerf están disponibles y se liberan públicamente bajo la licencia de Apache. La inferencia MLPerf tiene una división Edge, que admite los siguientes escenarios:
- **Flujo único.** este escenario imita a los sistemas en los que la capacidad de respuesta es un factor crítico, como las consultas de IA sin conexión realizadas en smartphones. Las consultas individuales se envían al sistema y se registran los tiempos de respuesta. como resultado, se indica la latencia del percentil 90 de todas las respuestas.
- **MultiStream.** este punto de referencia es para sistemas que procesan la entrada de varios sensores. Durante la prueba, las consultas se envían a un intervalo de tiempo fijo. Se impone una limitación de calidad de servicio (latencia máxima permitida). La prueba informa del número de flujos que el sistema puede procesar mientras cumple la restricción QoS.
- **Fuera de línea.** este es el escenario más sencillo que cubre las aplicaciones de procesamiento por lotes y la métrica es el procesamiento en muestras por segundo. Todos los datos están disponibles para el sistema y el punto de referencia mide el tiempo que tarda en procesar todas las muestras.

Lenovo ha publicado puntuaciones de inferencia MLPerf para SE350 con T4, el servidor utilizado en este documento. Consulte los resultados en "https://mlperf.org/inference-results-0-7/" En la sección “Edge, Closed Division”, en la entrada #0.7-145.

Plan de pruebas

Este documento sigue a la inferencia MLPerf v0.7 "codificación", Inferencia MLPerf v1.1 "codificación", y. "bases de datos". Ejecutamos pruebas de rendimiento MLPerf diseñadas para la inferencia en el perímetro tal y como se definen en la tabla siguiente.

Zona	Tarea	Modelo	Conjunto de datos	Tamaño de QSL	Calidad	Limitación de latencia de múltiples flujos
Visión	Clasificación de imágenes	Resnet50v1.5	ImageNET (224 x 224)	1024	99% del FP32	50 ms
Visión	Detección de objetos (grande)	SSD-ResNet34	COCO (1200x1200)	64	99% del FP32	66 ms
Visión	Detección de objetos (pequeño)	SSD-MobileNetsv1	COCO (300 x 300)	256	99% del FP32	50 ms
Visión	Segmentación de imagen médica	UNET 3D	Brats 2019 (224x224x160)	16	99% y 99.9% del FP32	n.a.
Voz	Voz a texto	RNNT	Llibrispeech dev-Clean	2513	99% del FP32	n.a.
Idioma	Procesamiento de idiomas	BERT	Escuadra v1.1	10833	99% del FP32	n.a.

En la siguiente tabla se presentan los escenarios de referencia de Edge.

Zona	Tarea	Escenarios
Visión	Clasificación de imágenes	Flujo único, sin conexión, multisecuencia
Visión	Detección de objetos (grande)	Flujo único, sin conexión, multisecuencia
Visión	Detección de objetos (pequeño)	Flujo único, sin conexión, multisecuencia
Visión	Segmentación de imagen médica	Flujo único, sin conexión
Voz	Voz a texto	Flujo único, sin conexión
Idioma	Procesamiento de idiomas	Flujo único, sin conexión

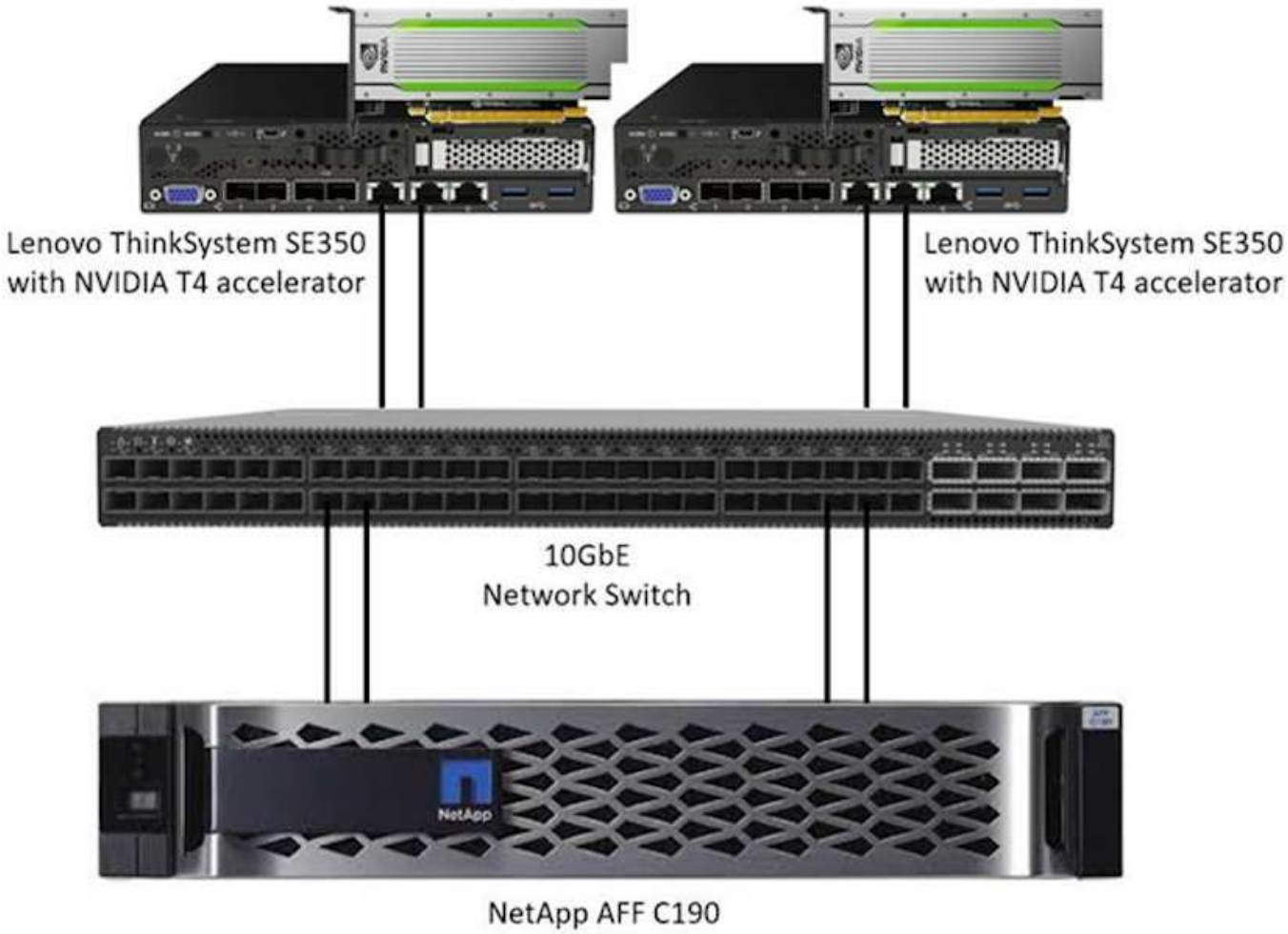
Realizamos estas pruebas de rendimiento utilizando la arquitectura de almacenamiento en red desarrollada en esta validación y comparamos los resultados de las ejecuciones locales en servidores perimetrales que antes se enviaban al MLPerf. La comparación consiste en determinar cuánto impacto tiene el almacenamiento compartido en el rendimiento de la inferencia.

Configuración de prueba

La figura siguiente muestra la configuración de prueba. Utilizamos el sistema de almacenamiento AFF C190 de NetApp y dos servidores Lenovo ThinkSystem SE350 (cada uno con un acelerador NVIDIA T4). Estos componentes se conectan a través de un switch de red de 10 GbE. El almacenamiento en red contiene conjuntos de datos de validación y pruebas y modelos preentrenados. Los servidores proporcionan una funcionalidad computacional y se accede al almacenamiento mediante un protocolo NFS.

En este apartado se describen las configuraciones probadas, la infraestructura de red, el servidor SE350 y los detalles de aprovisionamiento de almacenamiento. En la siguiente tabla se enumeran los componentes básicos para la arquitectura de la solución.

Componentes de la solución	Detalles
Servidores Lenovo ThinkSystem	• 2 servidores SE350 con una tarjeta GPU NVIDIA T4
	• Cada servidor contiene una CPU Intel Xeon D-2123IT con cuatro núcleos físicos que funcionan a 2,20 GHz y 128 GB de RAM
Sistema de almacenamiento AFF de gama básica de NetApp (par de alta disponibilidad)	• Software ONTAP 9 de NetApp • 24 unidades SSD de 960 GB • Protocolo NFS • Un grupo de interfaces por controladora, con cuatro direcciones IP lógicas para puntos de montaje



La siguiente tabla enumera la configuración del almacenamiento: AFF C190 con 2 ranuras de 24 unidades 2RU.

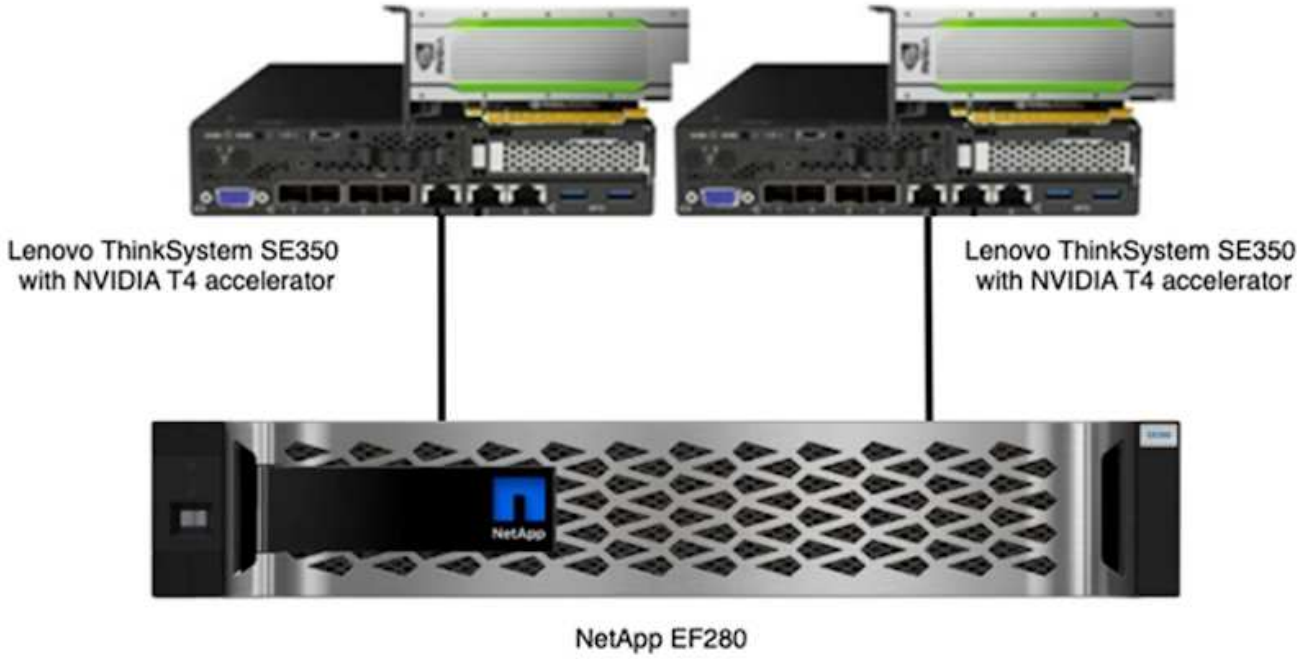
Controladora	Agregado	Volumen FlexGroup	Agregatesize	Volumen	Punto de montaje del sistema operativo
Control1	Agr1	/Netaptenovo_AI_fg	8,42 TIB	15 TB	/netapp_lenovo_fg
Control2	Agr2		8,42 TIB		

La carpeta /netappLenovo_AI_fg contiene los conjuntos de datos utilizados para la validación del modelo.

La siguiente figura muestra la configuración de prueba. Utilizamos el sistema de almacenamiento EF280 de NetApp y dos servidores Lenovo ThinkSystem SE350 (cada uno con un acelerador NVIDIA T4). Estos componentes se conectan a través de un switch de red de 10 GbE. El almacenamiento en red contiene conjuntos de datos de validación y pruebas y modelos preentrenados. Los servidores proporcionan una funcionalidad computacional y se accede al almacenamiento mediante un protocolo NFS.

La siguiente tabla enumera la configuración de almacenamiento para EF280.

Controladora	Grupo de volúmenes	Volumen	Volumen	Tamaño DDPsize	Método de conexión
Control1	DDP1	Volumen 1	8,42 TIB	16 TB	SE350-1 a iSCSI LUN 0
Control2		Volumen 2	8,42 TIB		SE350-2 a iSCSI LUN 1



Procedimiento de prueba

Este apartado describe los procedimientos de prueba utilizados para validar esta solución.

Configuración de inferencia de IA y sistema operativo

Para AFF C190, utilizamos Ubuntu 18.04 con controladores NVIDIA y docker con soporte para GPU de NVIDIA y usamos MLPerf "codificación" Disponible como parte de la presentación de Lenovo a la inferencia MLPerf v0.7.

Para EF280, utilizamos Ubuntu 20.04 con controladores NVIDIA y docker con soporte para las GPU de NVIDIA y MLPerf "codificación" Disponible como parte de la presentación de Lenovo a la inferencia MLPerf v1.1.

Para configurar la inferencia de IA, siga estos pasos:

1. Descargue los conjuntos de datos que requieren registro, el conjunto de datos ImageNET 2012 Validation, el conjunto de datos Criteo Terabyte y el conjunto de entrenamiento Brats 2019 y, a continuación, descomprima los archivos.
2. Cree un directorio de trabajo con al menos 1 TB y defina una variable ambiental `MLPERF_SCRATCH_PATH` referencia al directorio.

Debe compartir este directorio en el almacenamiento compartido para el caso de uso del almacenamiento de red o en el disco local cuando realice pruebas con datos locales.

3. Ejecute la Marca `prebuild` comando, que crea e inicia el contenedor docker para las tareas de inferencia necesarias.

   Los siguientes comandos se ejecutan desde el contenedor docker en ejecución:

   - Descargue modelos de IA preformados para tareas de inferencia de MLPerf: `make download_model`
   - Descargue conjuntos de datos adicionales que se pueden descargar gratuitamente: `make download_data`
   - Preprocesar los datos: Make `preprocess_data`
   - Ejecución: `make build`.
   - Cree motores de inferencia optimizados para la GPU en servidores informáticos: `make generate_engines`
   - Para ejecutar cargas de trabajo de inferencia, ejecute el siguiente (un comando):

```
make run_harness RUN_ARGS="--benchmarks=<BENCHMARKS>
--scenarios=<SCENARIOS>"
```

Se ejecuta la inferencia de IA

Se ejecutaron tres tipos de ejecuciones:

- Inferencia de IA de un único servidor con almacenamiento local
- Inferencia de IA de un único servidor con el almacenamiento en red
- Inferencia de la IA en varios servidores con el almacenamiento en red

Resultados de la prueba

Se llevaron a cabo una multitud de pruebas para evaluar el rendimiento de la arquitectura propuesta.

Existen seis cargas de trabajo diferentes (clasificación de imágenes, detección de objetos [pequeño], detección de objetos [grande], imágenes médicas, voz a texto, Y procesamiento de lenguaje natural [NLP]), que se puede ejecutar en tres escenarios diferentes: Sin conexión, flujo único y multisecuencia.

El último escenario se implementa sólo para la clasificación de imágenes y la detección de objetos.

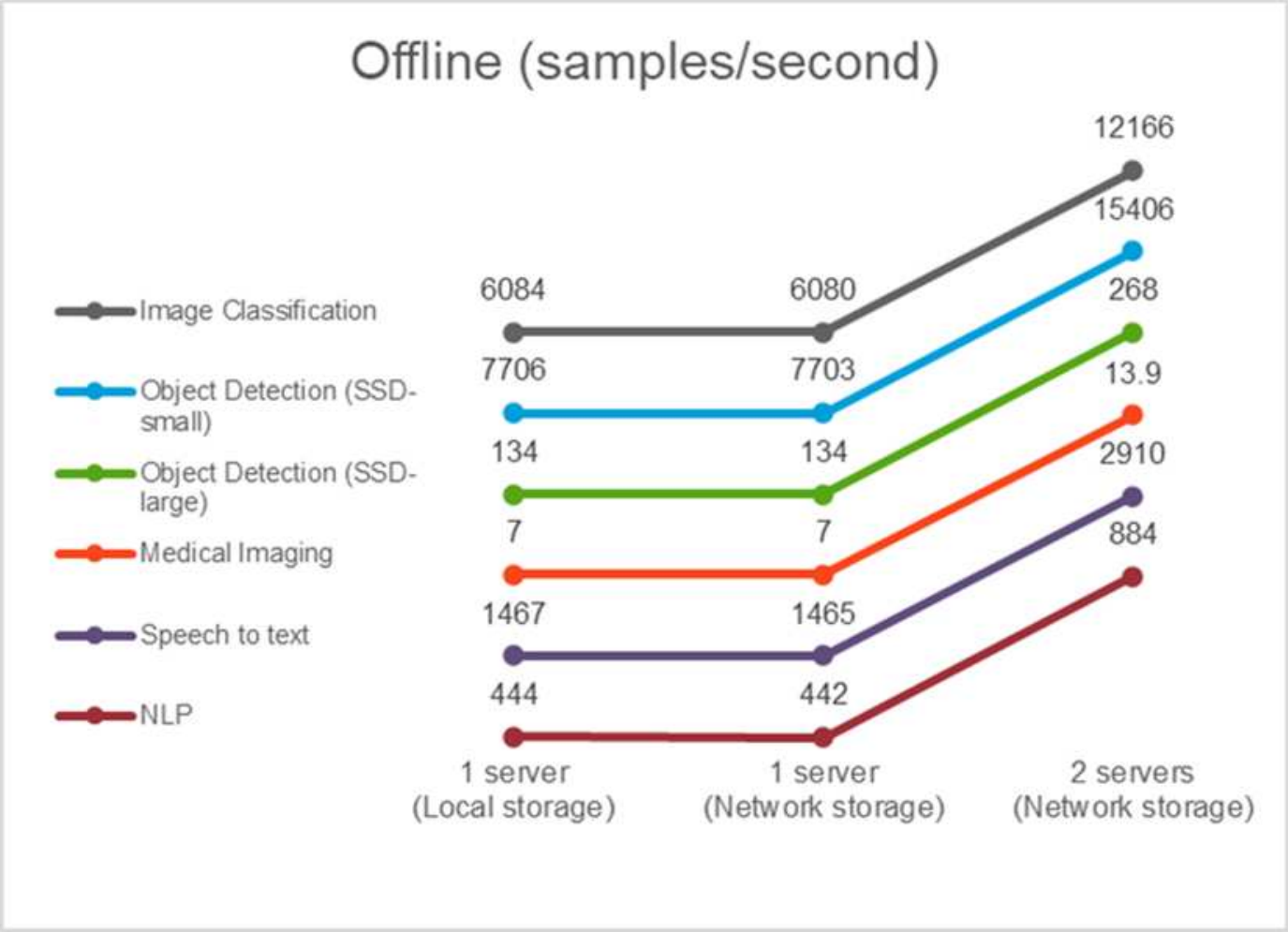
Esto permite 15 cargas de trabajo posibles, todas ellas probadas en tres configuraciones diferentes:

- Un único servidor/almacenamiento local
- Almacenamiento único de servidor/red
- Almacenamiento en red/multiservidor

Los resultados se describen en las siguientes secciones.

Inferencia de la IA en un escenario sin conexión para AFF

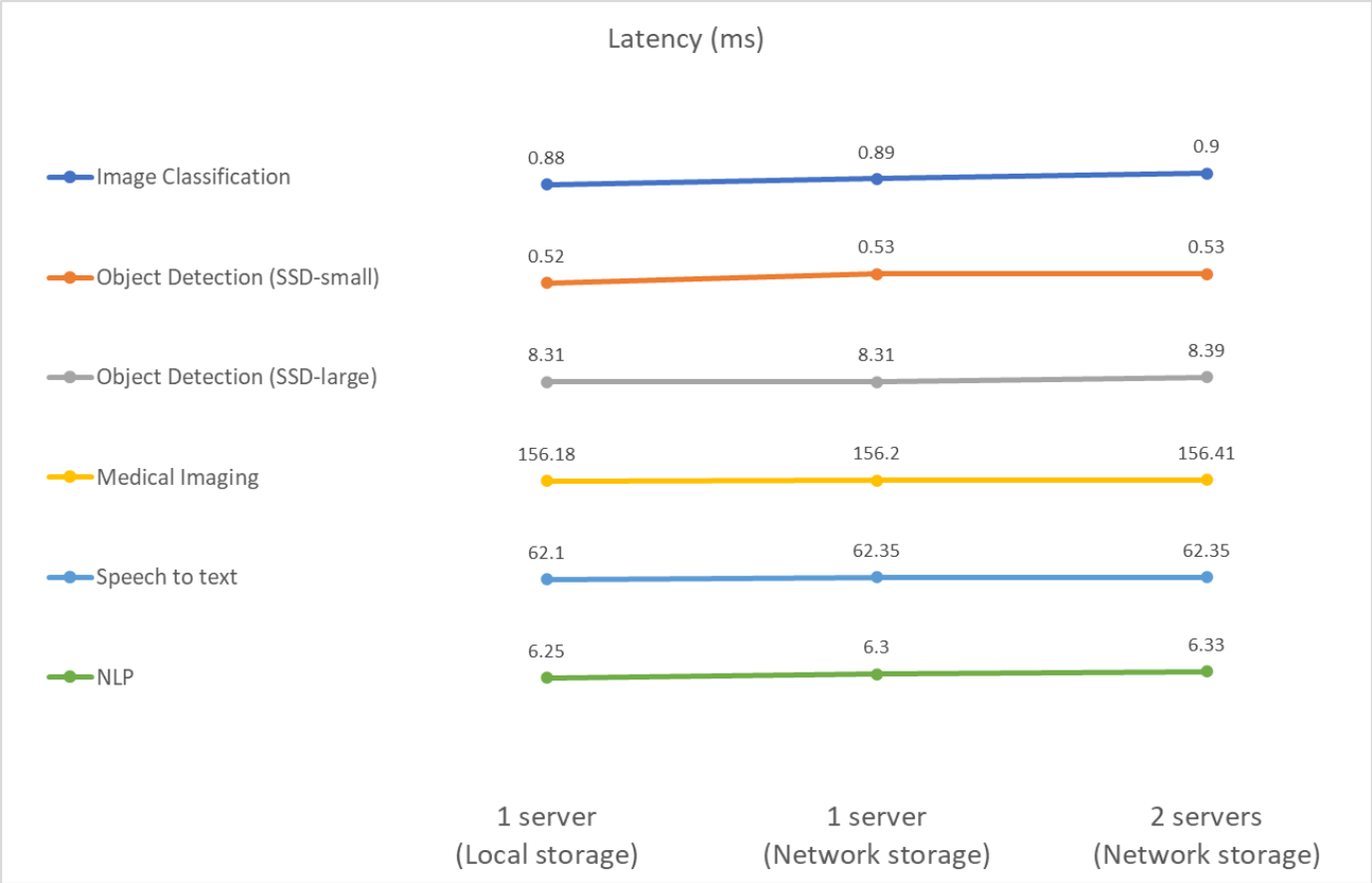
En este escenario, todos los datos estaban disponibles para el servidor y se midió el tiempo que se tardaba en procesar todas las muestras. Reportamos los anchos de banda en muestras por segundo como resultados de las pruebas. Cuando se utilizaron más de un servidor informático, hemos indicado el total de ancho de banda total en todos los servidores. En la siguiente figura se muestran los resultados de los tres casos de uso. Para el caso de dos servidores, reportamos un ancho de banda combinado de ambos servidores.



Los resultados muestran que el almacenamiento en red no afecta de manera negativa al rendimiento; el cambio es mínimo y, en algunas tareas, no se encuentra ninguno. Cuando se añade el segundo servidor, el ancho de banda total se duplica exactamente o, en el peor de los casos, el cambio es inferior al 1%.

Inferencia de IA en un escenario de flujo único para AFF

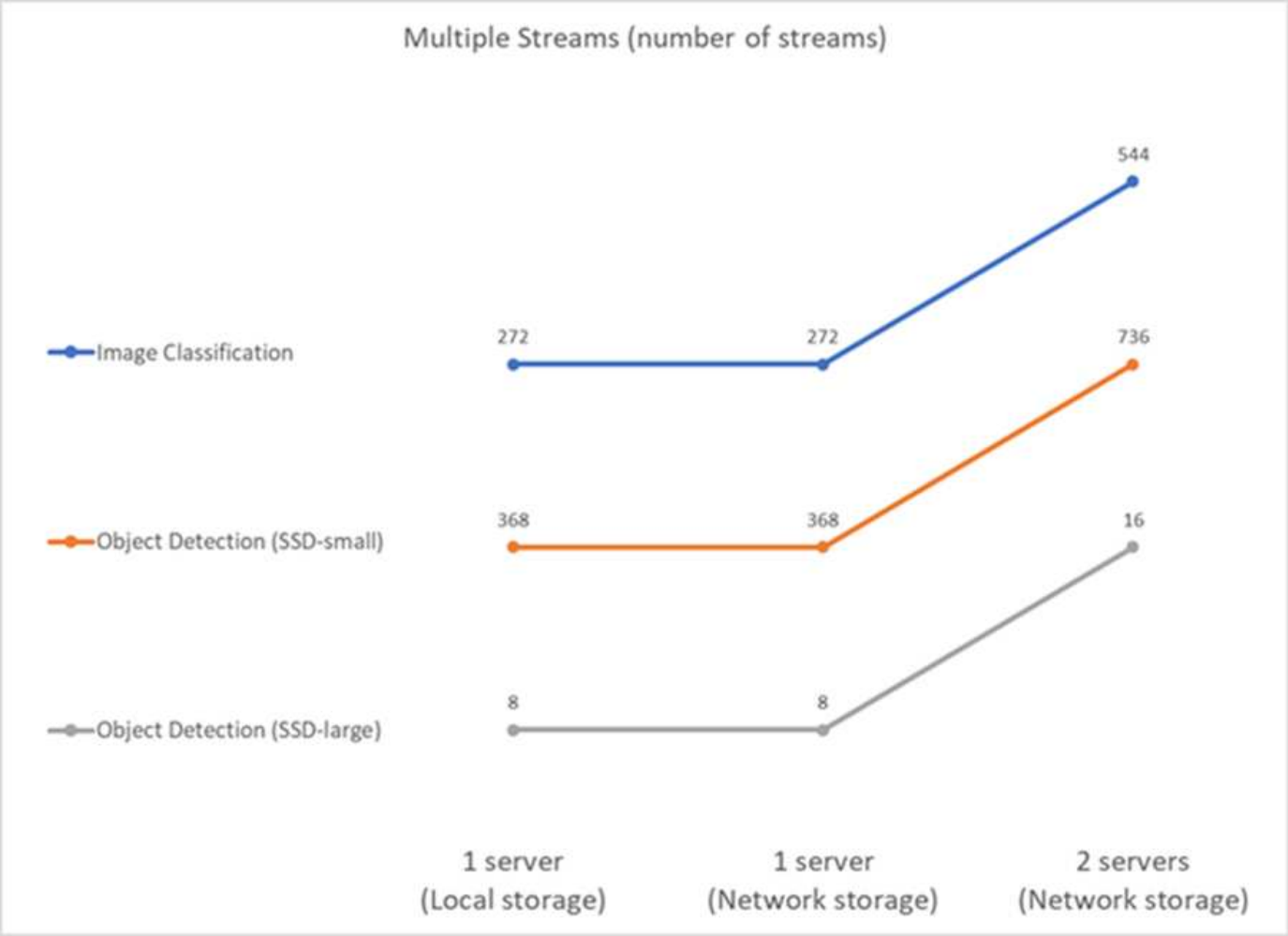
Esta prueba mide la latencia. Para el caso de varios servidores informáticos, reportamos la latencia media. Los resultados del conjunto de tareas se indican en la siguiente figura. Para el caso de dos servidores, hemos registrado la latencia media de ambos servidores.



Los resultados, de nuevo, muestran que el almacenamiento de red es suficiente para manejar las tareas. La diferencia entre el almacenamiento local y el de red en un caso de servidor es mínima o ninguna. De igual modo, cuando dos servidores utilizan el mismo almacenamiento, la latencia de ambos servidores se mantiene igual o se modifica en una cantidad muy pequeña.

Inferencia de IA en un escenario de transmisión múltiple para AFF

En este caso, el resultado es el número de flujos que el sistema puede manejar mientras se satisface la restricción QoS. Por lo tanto, el resultado siempre es un entero. En más de un servidor, reportamos el número total de flujos sumados en todos los servidores. No todas las cargas de trabajo admiten este escenario, pero hemos ejecutado las que lo hacen. Los resultados de nuestras pruebas se resumen en la siguiente figura. Para el caso de dos servidores, reportamos el número combinado de flujos de ambos servidores.



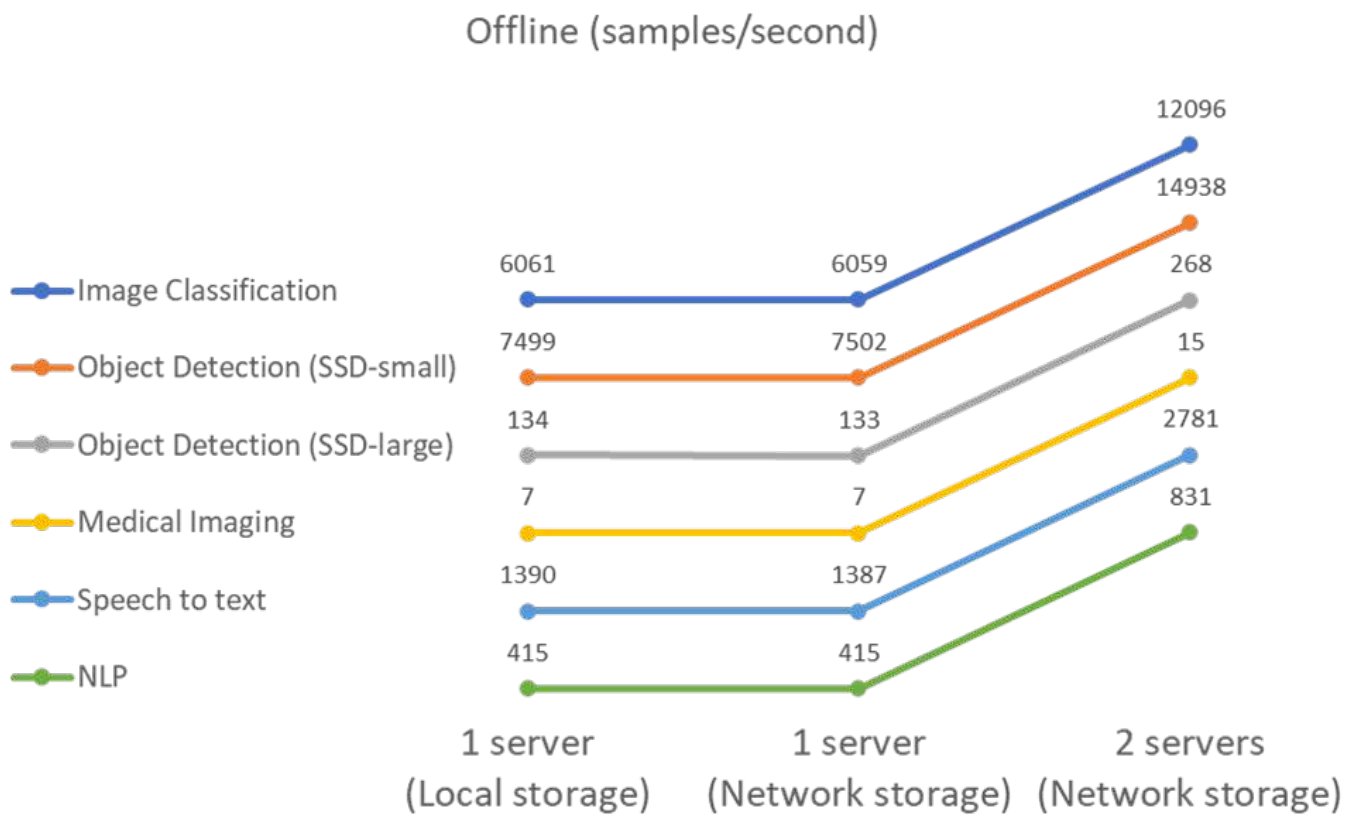
Los resultados muestran un rendimiento perfecto de la configuración; el almacenamiento local y en red ofrecen los mismos resultados y al añadir el segundo servidor se duplica el número de flujos que puede gestionar la configuración propuesta.

Resultados de la prueba para EF

Se llevaron a cabo una multitud de pruebas para evaluar el rendimiento de la arquitectura propuesta. Existen seis cargas de trabajo diferentes (clasificación de imágenes, detección de objetos [pequeño], detección de objetos [grande], imágenes médicas, voz a texto, Y procesamiento de lenguaje natural [NLP]), que se ejecutaron en dos escenarios diferentes: Offline y de flujo único. Los resultados se describen en las siguientes secciones.

Inferencia de la IA en un escenario sin conexión para EF

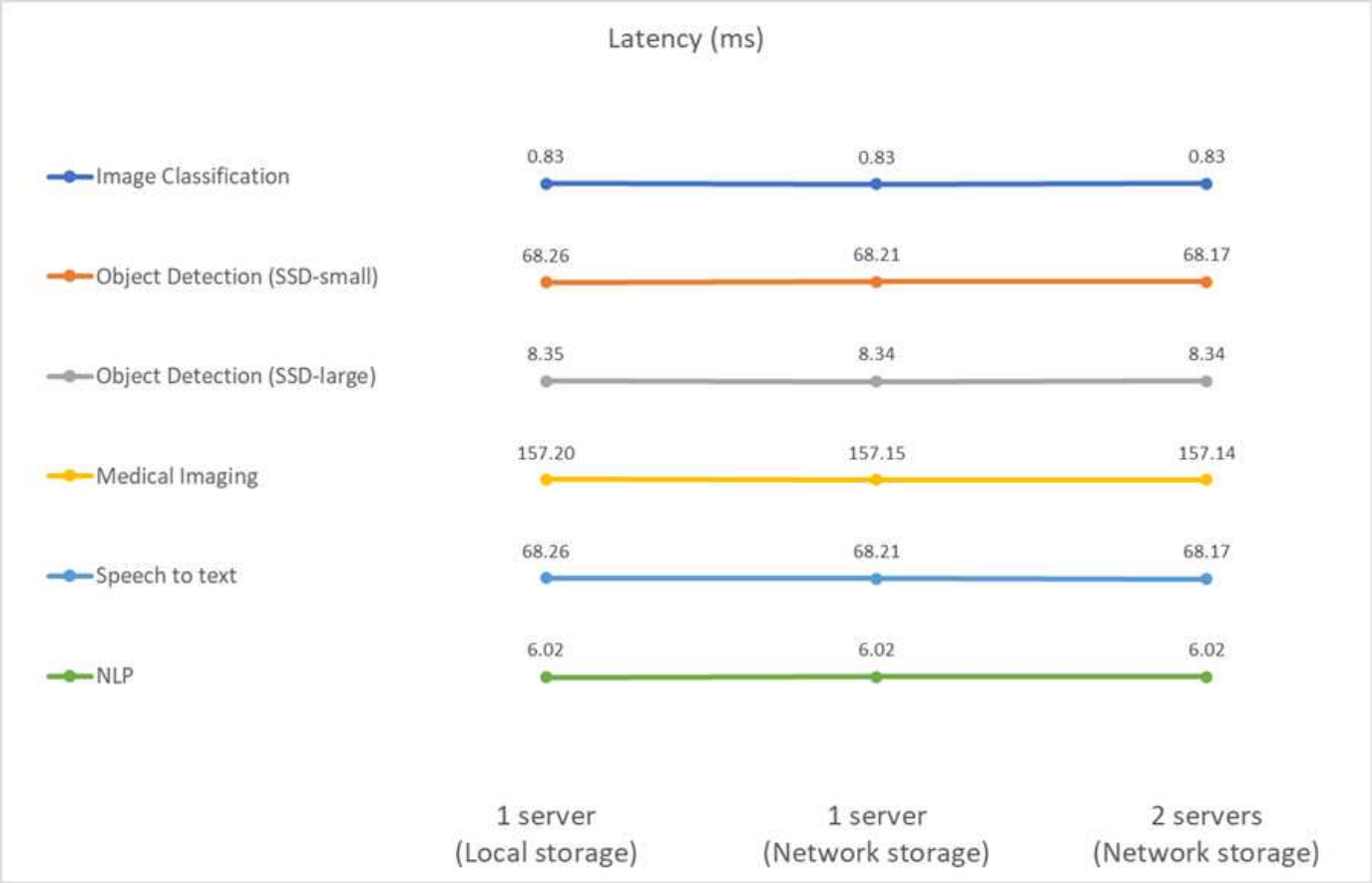
En este escenario, todos los datos estaban disponibles para el servidor y se midió el tiempo que se tardaba en procesar todas las muestras. Reportamos los anchos de banda en muestras por segundo como resultados de las pruebas. En las ejecuciones de un solo nodo reportamos la media de ambos servidores, mientras que en dos ejecuciones de servidor reportamos un ancho de banda total resumido en todos los servidores. Los resultados de los casos de uso se muestran en la siguiente figura.



Los resultados muestran que el almacenamiento en red no afecta de manera negativa al rendimiento; el cambio es mínimo y, en algunas tareas, no se encuentra ninguno. Cuando se añade el segundo servidor, el ancho de banda total se duplica exactamente o, en el peor de los casos, el cambio es inferior al 1%.

Inferencia de IA en un escenario de flujo único para EF

Esta prueba mide la latencia. En todos los casos, reportamos una latencia media en todos los servidores involucrados en las ejecuciones. Los resultados para el conjunto de tareas se dan.



Los resultados muestran de nuevo que el almacenamiento en red es suficiente para manejar las tareas. La diferencia entre el almacenamiento local y el de red en un caso de servidor es mínima o ninguna. De igual modo, cuando dos servidores utilizan el mismo almacenamiento, la latencia de ambos servidores se mantiene igual o se modifica en una cantidad muy pequeña.

Opciones de ajuste de tamaño de arquitectura

Puede ajustar la configuración utilizada para la validación y adaptarla a otros casos prácticos.

Servidor de computación

Utilizamos una CPU Intel Xeon D-2123IT, que es el nivel más bajo de CPU compatible con SE350, con cuatro núcleos físicos y 60 W TDP. Aunque el servidor no admite la sustitución de CPU, se puede pedir con una CPU más potente. La CPU más alta admitida es Intel Xeon D-2183IT con 16 núcleos, 100 W con 2,20 GHz. Esto aumenta considerablemente la capacidad computacional de la CPU. Mientras que la CPU no era un cuello de botella para ejecutar las cargas de trabajo de inferencia en sí, ayuda con el procesamiento de datos y otras tareas relacionadas con la inferencia. Actualmente, NVIDIA T4 es la única GPU disponible para casos de uso periféricos; por lo tanto, actualmente, no es posible actualizar o degradar el GPU.

Almacenamiento compartido

Para las pruebas y la validación, se utilizó el sistema AFF C190 de NetApp, que tiene una capacidad de almacenamiento máxima de 50,5 TB, una rendimiento de 4,4 GB/s para lecturas secuenciales y 230 000 IOPS para pequeñas lecturas aleatorias, con los fines de este documento y se ha demostrado ser apto para cargas de trabajo de inferencia periférica.

No obstante, si se requiere más capacidad de almacenamiento o velocidades de red más rápidas, deberá utilizar el o el sistema AFF A220 de NetApp "AFF A250 de NetApp" sistemas de almacenamiento. Además, el sistema EF280 de NetApp, que tiene una capacidad máxima de 1,5 PB, 10 Gbps de ancho de banda, también se utilizó para la validación de esta solución. Si prefiere más capacidad de almacenamiento con un mayor ancho de banda, "EF300 de NetApp" puede utilizarse.

Conclusión

La automatización impulsada por IA y la informática perimetral es un enfoque líder que ayuda a las organizaciones empresariales a lograr la transformación digital y maximizar la eficiencia y la seguridad operativas. En el caso de los entornos periféricos, los datos se procesan con mayor rapidez, porque no tienen que desplazarse hacia y desde un centro de datos. Por lo tanto, el coste asociado al envío de datos a los centros de datos o al cloud se ve disminuido. La reducción de la latencia y el aumento de la velocidad pueden ser beneficiosas cuando las empresas deben tomar decisiones casi en tiempo real mediante modelos de inferencia de IA implementados en el perímetro.

Los sistemas de almacenamiento de NetApp ofrecen el mismo rendimiento o mejor que el almacenamiento SSD local y ofrecen las siguientes ventajas a los científicos de datos, ingenieros de datos, desarrolladores de IA/ML y responsables de la toma de decisiones EMPRESARIALES o TECNOLÓGICAS:

- Uso compartido de datos sin esfuerzo entre sistemas de IA, análisis y otros sistemas de negocio cruciales. Este uso compartido de datos reduce la sobrecarga de la infraestructura, mejora el rendimiento y optimiza la gestión de datos en toda la empresa.
- Almacenamiento y cálculo escalables de forma independiente para minimizar los costes y mejorar el uso de recursos.
- Flujos de trabajo de desarrollo y puesta en marcha optimizados que utilizan copias Snapshot integradas y clones para espacios de trabajo de usuario instantáneos con gestión eficiente del espacio, control de versiones integrado y una puesta en marcha automatizada.
- Protección de datos de clase empresarial para la recuperación ante desastres y la continuidad del negocio. La solución de NetApp y Lenovo presentada en este documento es una arquitectura flexible y de escalado horizontal ideal para puestas en marcha de inferencia de IA en el perímetro.

Reconocimientos

- J.J. Falkanger, Sr. Director de Soluciones HPC e IA de Lenovo
- Dave Arnette, ingeniero técnico de marketing, NetApp
- Joey Parnell, responsable técnico de soluciones de IA de E-Series de NetApp
- Cody Harryman, ingeniero de control de calidad, NetApp

Dónde encontrar información adicional

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

- Página de producto de las cabinas AFF a-Series de NetApp

  "https://www.netapp.com/data-storage/aff-a-series/"

- Software de gestión de datos ONTAP de NetApp: Biblioteca de información de ONTAP 9

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

- TR-4727: Introducción a EF-Series de NetApp

  https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf

- Especificaciones técnicas del software SANtricity para E-Series de NetApp

  https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf

- Almacenamiento persistente de NetApp para contenedores: Trident de NetApp

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

- Rendim. MLPerf
  - "https://mlcommons.org/en/"
  - "http://www.image-net.org/"
  - "https://mlcommons.org/en/news/mlperf-inference-v11/"
- Copia y sincronización de NetApp BlueXP

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

- Prueba de rendimiento de TensorFlow

  "https://github.com/tensorflow/benchmarks"

- Servidor Lenovo ThinkSystem SE350 Edge

  "https://lenovopress.com/lp1168"

- Sistema de almacenamiento flash unificado Lenovo ThinkSystem DM5100F

  "https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"

Artículo técnico WP-7328: La IA conversacional de NetApp con NVIDIA Jarvis

Rick Huang, Sung-han Lin, NetApp Davide Onofrío, NVIDIA

La familia de sistemas NVIDIA DGX está compuesta por los primeros sistemas basados en inteligencia artificial (IA) integrados del mundo, creados expresamente para la IA empresarial. Los sistemas de almacenamiento AFF de NetApp proporcionan un rendimiento extremo y funcionalidades de gestión de datos de cloud híbrido líderes en el sector. NetApp y NVIDIA se han asociado para crear la arquitectura de referencia de IA ONTAP de NetApp, una solución lista para usar para cargas de trabajo de IA y aprendizaje automático (ML) que ofrece rendimiento, fiabilidad y soporte de clase empresarial.

Este whitepaper proporciona instrucciones direccionales a los clientes para crear sistemas de IA conversacionales que respaldan diferentes casos de uso en diversos sectores verticales. Incluye información

sobre la implementación del sistema mediante NVIDIA Jarvis. Las pruebas se llevaron a cabo con una estación DGX de NVIDIA y un sistema de almacenamiento AFF A220 de NetApp.

El público objetivo de la solución incluye los siguientes grupos:

- Arquitectos empresariales que diseñan soluciones para el desarrollo de modelos de IA y software para casos de uso de IA conversacionales como un asistente de ventas al por menor virtual
- Científicos de datos que buscan formas eficientes de lograr los objetivos de desarrollo del modelado del lenguaje
- Ingenieros de datos encargados de mantener y procesar datos de texto, como preguntas de clientes y transcripciones de diálogo
- Responsables de la toma de decisiones tecnológicas y ejecutivas y líderes de negocio interesados en transformar la experiencia de IA conversacional y lograr el plazo de comercialización más rápido de iniciativas de IA

Descripción general de la solución

NetApp ONTAP AI y copia y sincronización de BlueXP

La arquitectura de ONTAP AI de NetApp, impulsada por los sistemas NVIDIA DGX y los sistemas de almacenamiento conectados al cloud de NetApp, fue desarrollada y verificada por NetApp y NVIDIA. Esta arquitectura de referencia proporciona a las organizaciones DE TI las siguientes ventajas:

- Elimina las complejidades de diseño
- Permite un escalado independiente de las capacidades de computación y almacenamiento
- Permite a los clientes empezar con poco y escalar sin problemas
- Ofrece una gama de opciones de almacenamiento para diferentes funcionalidades de rendimiento y rentabilidad.ONTAP AI integra a la perfección los sistemas de almacenamiento DGX y AFF A220 de NetApp con redes de vanguardia. Los sistemas ONTAP AI y DGX de NetApp simplifican las puestas en marcha de IA al eliminar complejidades y conjeturas en la fase de diseño. Los clientes pueden empezar poco a poco y aumentar sus sistemas de forma ininterrumpida, a la vez que gestionan de forma inteligente datos entre el perímetro, el núcleo y el cloud.

La función Copy and Sync de NetApp BlueXP te permite mover datos fácilmente a través de diversos protocolos, ya sea entre dos unidades NFS, dos unidades CIFS o una unidad de archivos y almacenamiento de Amazon S3, Amazon Elastic File System (EFS) o Azure Blob. El funcionamiento activo-activo permite seguir trabajando tanto con el origen como con el objetivo al mismo tiempo, sincronizando de forma incremental los cambios de datos cuando sea necesario. Al permitirte mover y sincronizar gradualmente los datos entre cualquier sistema de origen y de destino, ya sea on-premises o en la nube, BlueXP Copy and Sync ofrece una amplia variedad de nuevas formas de utilizar los datos. Es muy fácil llegar a la migración de datos entre sistemas locales, integración en el cloud y migración al cloud, o colaboración y análisis de datos. La siguiente figura muestra los orígenes y destinos disponibles.

En sistemas de IA conversacional, los desarrolladores pueden aprovechar la función Copy and Sync de BlueXP para archivar el historial de conversaciones desde el cloud a los centros de datos y permitir el entrenamiento sin conexión de modelos de procesamiento del lenguaje natural (NLP). Gracias a los modelos de formación que permiten reconocer más intenciones, el sistema de IA conversacional estará mejor equipado para gestionar preguntas más complejas por parte de los usuarios finales.

NVIDIA Jarvis Multimodal Framework

"NVIDIA Jarvis" Es un marco integral para crear servicios de IA que puedan conversación. Incluye los siguientes servicios optimizados para GPU:

- Reconocimiento automático de voz (ASR)
- Comprensión del lenguaje natural (NLU)
- Integración con servicios logísticos específicos de dominio
- Texto a voz (TTS)
- Los servicios basados en Jarvis utilizan modelos de aprendizaje profundo de vanguardia para hacer frente a la compleja y desafiante tarea de IA conversacional en tiempo real. Para permitir la interacción natural y en tiempo real con un usuario final, los modelos deben completar el cálculo en menos de 300 milisegundos. Las interacciones naturales son difíciles, y requieren la integración sensorial multimodal. Los gasoductos son también complejos y requieren coordinación entre los servicios mencionados.

Jarvis es un marco de aplicaciones totalmente acelerado para la creación de servicios de IA conversacionales multimodales que usan una canalización de aprendizaje profundo integral. El marco Jarvis incluye modelos de IA conversacionales, herramientas y servicios integrales optimizados para tareas de habla, visión y NLU. Además de los servicios de IA, Jarvis le permite fusionar simultáneamente la visión, el audio y otras entradas de sensores para proporcionar capacidades como conversaciones multiusuario y multicontexto en aplicaciones como asistentes virtuales, diarización multiusuario y asistentes de centros de llamadas.

NVIDIA Nemo

"NVIDIA Nemo" Es un kit de herramientas Python de código abierto para crear, formar y ajustar modelos de IA de última generación acelerados por GPU y conversacionales que usan interfaces de programación de aplicaciones (API) fáciles de usar. Nemo ejecuta una computación de precisión mixta mediante núcleos tensores en GPU de NVIDIA y puede escalarse verticalmente hasta varias GPU fácilmente para ofrecer el máximo rendimiento de entrenamiento posible. Nemo se utiliza para crear modelos para aplicaciones ASR, NLP y TTS en tiempo real, como transcripciones de videollamadas, asistentes inteligentes de vídeo y soporte de centros de llamadas automatizados en diferentes sectores verticales, incluyendo servicios sanitarios,

financieros, minoristas y telecomunicaciones.

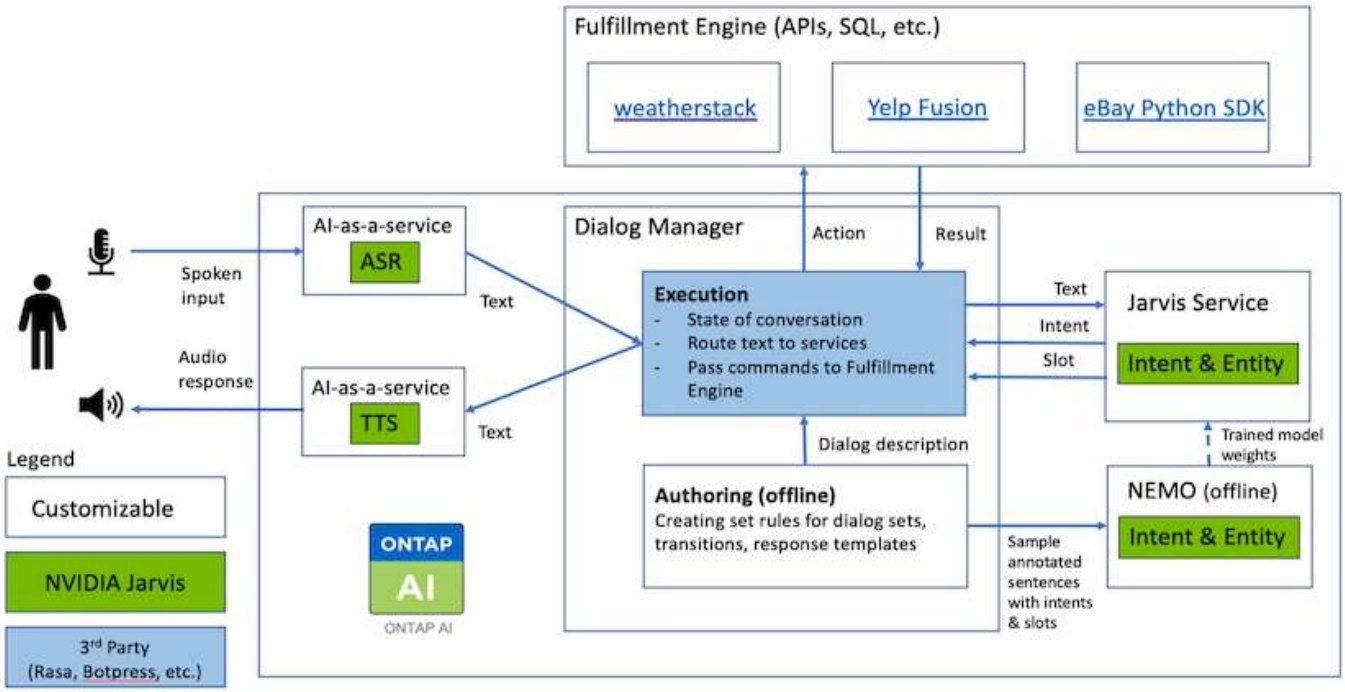
Utilizamos Nemo para entrenar modelos que reconocen intentos complejos de preguntas de usuario en el historial de conversaciones archivadas. Esta capacitación amplía las capacidades del asistente virtual al por menor más allá de lo que Jarvis apoya como se ofrece.

Resumen del caso de uso de venta al por menor

Con NVIDIA Jarvis, creamos un asistente de venta al por menor virtual que acepta comentarios o comentarios y responde a preguntas sobre el tiempo, puntos de interés y precios de inventario. El sistema de IA conversacional puede recordar el flujo de conversaciones, por ejemplo, formular una pregunta de seguimiento si el usuario no especifica la ubicación para el clima o los puntos de interés. El sistema también reconoce entidades complejas como “comida tailandesa” o “memoria portátil”. Entiende preguntas de lenguaje natural como “¿lloverá la próxima semana en los Ángeles?” Una demostración del asistente virtual al por menor se puede encontrar en "Personalizar Estados y flujos para el caso de uso de comercio minorista".

Tecnología de soluciones

La siguiente figura ilustra la arquitectura del sistema de IA conversacional propuesta. Puede interactuar con el sistema con la señal de voz o con la entrada de texto. Si se detecta una entrada hablada, Jarvis AI-as-service (AIaaS) realiza ASR para producir texto para Dialog Manager. Dialog Manager recuerda los estados de la conversación, enruta el texto a los servicios correspondientes y pasa los comandos al motor de ejecución. El servicio Jarvis NLP toma texto, reconoce intentos y entidades, y envía esos intentos y ranuras de entidad de vuelta a Dialog Manager, que luego envía Acción al motor de cumplimiento. El motor de ejecución consta de API o bases de datos SQL de terceros que responden a las consultas de los usuarios. Después de recibir el resultado de Logística Engine, Dialog Manager enruta el texto a Jarvis TTS AIaaS para producir una respuesta de audio para el usuario final. Podemos archivar el historial de conversaciones, anotar frases con intentos y ranuras para la formación de Nemo, de forma que el servicio de NLP mejore a medida que más usuarios interactúan con el sistema.



Requisitos de hardware

Esta solución se validó con un sistema de almacenamiento DGX Station y uno AFF A220. Jarvis requiere una GPU T4 o V100 para realizar cálculos profundos de la red neuronal.

En la siguiente tabla se enumeran los componentes de hardware necesarios para implementar la solución según se ha probado.

Hardware subyacente	Cantidad
GPU T4 O V100	1
Estación DGX de NVIDIA	1

Requisitos de software

En la siguiente tabla se enumeran los componentes de software necesarios para implementar la solución según se ha probado.

De NetApp	Versión u otra información
Software de gestión de datos ONTAP de NetApp	9.6
Firmware de switch Cisco NX-OS	7.0(3)I6(1)
SO DGX DE NVIDIA	4.0.4 - Ubuntu 18.04 LTS
NVIDIA Jarvis Framework	EA v0.2
NVIDIA Nemo	nvcr.io/nvidia/nemo:v0.10
Plataforma contenedora Docker	18.06.1-ce [e68fc7a]

Descripción general

En esta sección se ofrece información detallada sobre la implantación del asistente de venta al por menor virtual.

Despliegue de Jarvis

Puede registrarse para "Programa Jarvis Early Access" Para obtener acceso a contenedores Jarvis en NVIDIA GPU Cloud (NGC). Después de recibir credenciales de NVIDIA, puede implementar Jarvis siguiendo los pasos siguientes:

1. Firma a NGC.
2. Establezca la organización en NGC: `ea-2-jarvis`.
3. Localice Jarvis EA v0.2 activos: Contenedores Jarvis están en `Private Registry` > `Organization Containers`.
4. Seleccione Jarvis: Desplácese a. `Model Scripts` y haga clic en `Jarvis Quick Start`
5. Compruebe que todos los activos funcionan correctamente.
6. Busque la documentación para crear sus propias aplicaciones: Los archivos PDF se pueden encontrar en `Model Scripts` > `Jarvis Documentation` > `File Browser`.

Personalizar Estados y flujos para el caso de uso de comercio minorista

Puede personalizar los Estados y flujos de Dialog Manager para sus casos de uso específicos. En nuestro ejemplo de venta al por menor, tenemos los siguientes cuatro archivos yaml para dirigir la conversación según diferentes intentos.

S la siguiente lista de nombres de archivo y descripción de cada archivo:

- `main_flow.yml`: Define los principales flujos y estados de conversación y dirige el flujo a los otros tres archivos yaml cuando sea necesario.
- `retail_flow.yml`: Contiene estados relacionados con preguntas sobre puntos de interés o minoristas. El sistema proporciona la información de la tienda más cercana o el precio de un artículo determinado.
- `weather_flow.yml`: Contiene estados relacionados con las preguntas sobre el clima. Si no se puede determinar la ubicación, el sistema hace una pregunta de seguimiento para aclarar.
- `error_flow.yml`: Trata los casos en los que las intenciones del usuario no entran en los tres archivos yaml anteriores. Después de mostrar un mensaje de error, el sistema vuelve a enruta para aceptar preguntas de usuario.las siguientes secciones contienen las definiciones detalladas de estos archivos 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

```
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

```yaml
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
```

Conéctese a API de terceros como motor de cumplimiento

Conectamos las siguientes API de terceros como motor de cumplimiento de normativas

para responder a las preguntas:

- "API de WeatherStack": regresa el clima, la temperatura, las precipitaciones y la nieve en un lugar determinado.
- "API de Fusion yelp": devuelve la información de la tienda más cercana en una ubicación determinada.
- "SDK de Python de eBay": devuelve el precio de un artículo determinado.

Demostración del asistente de venta al por menor de NetApp

Grabamos un vídeo de demostración del Asistente para minoristas de NetApp (NARA).

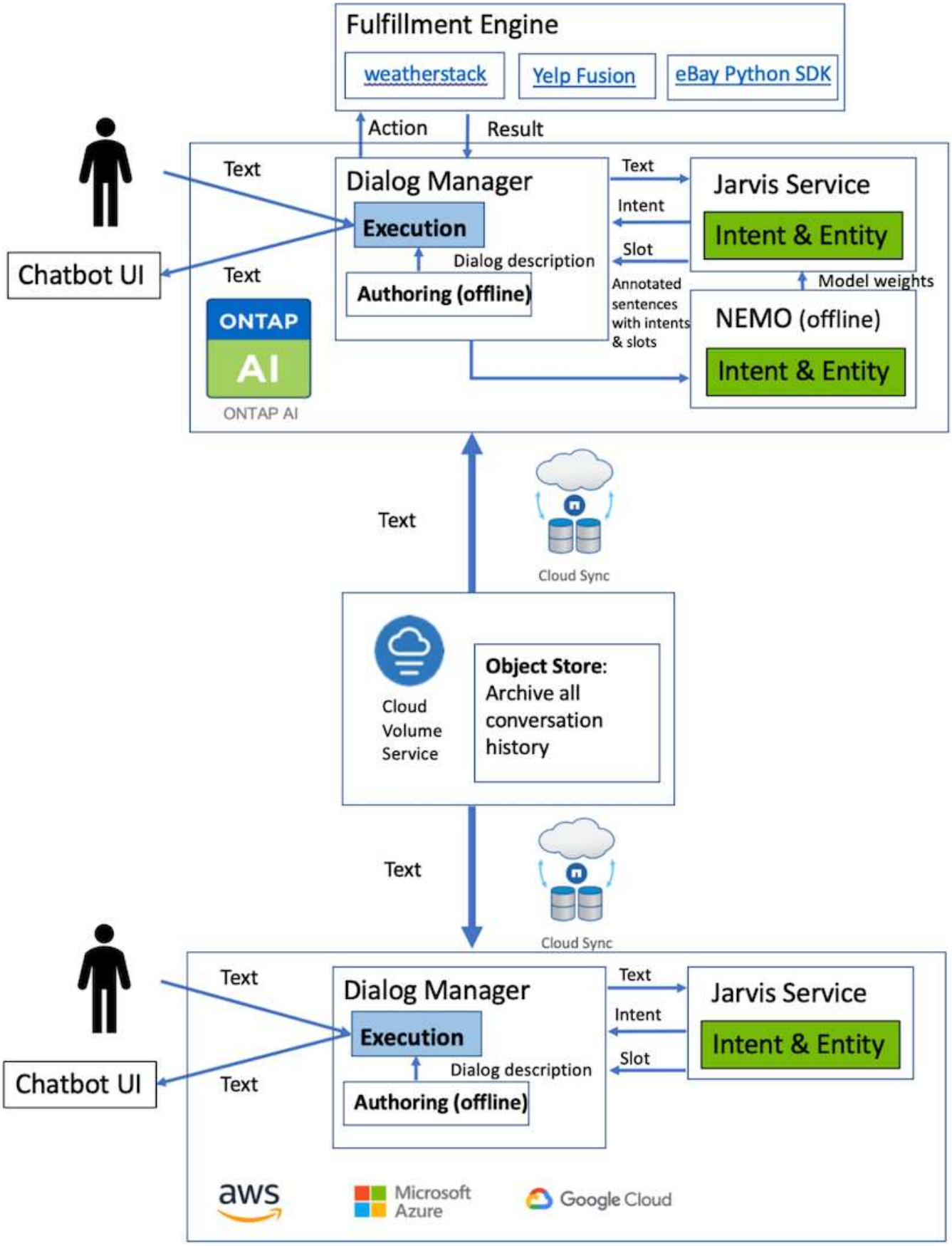
Demostración en vídeo de NARA

Demostración en vídeo de NARA

Utiliza la copia y sincronización de NetApp BlueXP para archivar el historial de conversaciones

Al volcar el historial de conversaciones en un archivo CSV una vez al día, podemos aprovechar la copia y sincronización de BlueXP para descargar los archivos de registro en un almacenamiento local. La siguiente figura muestra la arquitectura que indica que

Jarvis ha puesto en marcha on-premises y en nubes públicas, mientras utiliza la función Copy y Sync de BlueXP para enviar el historial de conversaciones para la formación de Nemo. En la sección encontrará más información sobre la formación de Nemo "Expanda modelos de intención utilizando 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

Expanda modelos de intención utilizando Nemo Training

NVIDIA Nemo es un kit de herramientas creado por NVIDIA para crear aplicaciones de IA conversacionales. Este kit de herramientas incluye colecciones de módulos preentrenados para ASR, NLP y TTS, lo que permite a investigadores y científicos de datos componer fácilmente arquitecturas complejas de redes neuronales y centrarse más en el diseño de sus propias aplicaciones.

Como se muestra en el ejemplo anterior, NARA sólo puede manejar un tipo limitado de preguntas. Esto se debe a que el modelo NLP pre-entrenado sólo entrena en este tipo de preguntas. Si queremos permitir QUE NARA pueda gestionar una gama más amplia de preguntas, debemos volver a formar este con nuestros propios conjuntos de datos. Por lo tanto, aquí mostramos cómo podemos utilizar Nemo para ampliar el modelo NLP para satisfacer los requisitos. Comenzamos convirtiendo el registro recolectado de NARA en el formato de Nemo, y luego entrenamos con el conjunto de datos para mejorar el modelo NLP.

Modelo

Nuestro objetivo es permitir A NARA ordenar los elementos según las preferencias del usuario. Por ejemplo, podemos pedir A NARA que sugiera el restaurante de sushi con mejor calificación o que quiera QUE NARA busque los vaqueros con el precio más bajo. Para ello, utilizamos el modelo de detección de intención y relleno de ranuras proporcionado en Nemo como modelo de entrenamiento. Este modelo permite A NARA comprender la intención de buscar preferencias.

Preparación de datos

Para entrenar el modelo, recopilamos el conjunto de datos para este tipo de preguntas y lo convertimos al formato Nemo. Aquí enumeramos los archivos que utilizamos para entrenar el modelo.

dict.intents.csv

Este archivo enumera todos los intentos que queremos que el Nemo entienda. Aquí tenemos dos intentos principales y una intención que sólo se utiliza para categorizar las preguntas que no encajan en ninguno de los intentos primarios.

```
price_check
find_the_store
unknown
```

dict.slots.csv

En este archivo se enumeran todas las ranuras que podemos etiquetar en nuestras preguntas de formación.

```
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.tsv

Este es el conjunto de datos de entrenamiento principal. Cada línea comienza con la pregunta que sigue a la lista de la categoría de intención en el archivo dict.intent.csv. La etiqueta se enumera a partir de cero.

train_slots.tsv

```
20 46 24 25 6 32 6
52 52 24 6
23 52 14 40 52 25 6 32 6
…
```

Entrenar el modelo

```
docker pull nvcr.io/nvidia/nemo:v0.10
```

A continuación, utilizamos el siguiente comando para iniciar el contenedor. En este comando, limitamos el contenedor para usar una única GPU (ID de GPU = 1), ya que se trata de un ejercicio de entrenamiento de poco peso. También mapeamos nuestro espacio de trabajo local /Workspace/nemo/ a la carpeta dentro del contenedor /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
```

Dentro del contenedor, si queremos empezar desde el modelo ORIGINAL BERT pre-entrenado, podemos usar el siguiente comando para iniciar el procedimiento de entrenamiento. data_dir es el argumento para establecer la ruta de los datos de entrenamiento. dir_trabajo le permite configurar dónde desea almacenar los archivos de punto de control.

```
cd examples/nlp/intent_detection_slot_tagging/
python joint_intent_slot_with_bert.py \
    --data_dir /nemo/training_data\
    --work_dir /nemo/log
```

Si contamos con nuevos conjuntos de datos de entrenamiento y queremos mejorar el modelo anterior, podemos utilizar el siguiente comando para continuar desde el punto que hemos detenido. checkpoint_dir lleva la ruta a la carpeta de puntos de control anteriores.

```
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
```

Inferencia del modelo

Se debe validar el rendimiento del modelo entrenado después de una serie determinada de épocas. El siguiente comando nos permite probar la consulta una por una. Por ejemplo, en este comando, queremos comprobar si nuestro modelo puede identificar adecuadamente la intención de la consulta `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
```

A continuación, se muestra la salida de la inferencia. En el resultado, podemos ver que nuestro modelo entrenado puede predecir correctamente la intención find_the_store, y devolver las palabras clave en las que estamos interesados. Con estas palabras clave, permitimos A LA NARA buscar lo que los usuarios desean y realizar una búsqueda más 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
```

Conclusión

Un auténtico sistema de IA conversacional se involucra en un diálogo similar al humano, comprende el contexto y proporciona respuestas inteligentes. Estos modelos de IA a menudo son enormes y muy complejos. Con las GPU de NVIDIA y el almacenamiento de

NetApp, se pueden formar y optimizar modelos masivos de lenguaje de última generación para ejecutar inferencia rápidamente. Se trata de una importante medida para poner fin a la compensación entre un modelo de IA que es rápido frente a uno que es grande y complejo. Los modelos de comprensión del lenguaje optimizados para GPU se pueden integrar en las aplicaciones de IA en sectores como el sanitario, el minorista y los servicios financieros. Esto hace posible que asistentes de voz digitales avanzados en altavoces inteligentes y líneas de servicio al cliente. Estos sistemas de IA conversacionales de alta calidad permiten a las empresas de sectores verticales proporcionar servicios personalizados que antes eran impensables al tratar con los clientes.

Jarvis permite el despliegue de casos de uso como asistentes virtuales, avatares digitales, sensor multimodal Fusion (CV con fusible ASR/NLP/TTS) o cualquier caso de uso independiente ASR/NLP/TTS/CV, como la transcripción. Hemos creado un asistente de venta al por menor virtual que puede responder a preguntas relacionadas con el tiempo, los puntos de interés y los precios del inventario. También demostramos cómo mejorar las funcionalidades de comprensión del lenguaje natural del sistema de IA conversacional archivando el historial de conversaciones con BlueXP Copy y Sync y entrenando modelos de Nemo en datos nuevos.

Reconocimientos

Los autores agradecen las contribuciones que nuestros estimados colegas de NVIDIA han realizado a este documento técnico: Davide Onofrío, Alex Qi, Sicong Ji, Marty Jain y Robert Sohigian. Los autores también desean reconocer las contribuciones de miembros clave del equipo de NetApp: Santosh Rao, David Arnette, Michael Oglesby, Brent Davis, Andy Sayare, Erik Mulder, y Mike McNamara.

Nuestro más sincero agradecimiento a todos aquellos que nos ofrecieron sus valiosos comentarios y su experiencia para la creación de este informe.

Dónde encontrar información adicional

Si quiere obtener más información sobre el contenido de este documento, consulte los siguientes recursos:

- NVIDIA DGX Station, GPU V100, GPU Cloud
  - Estación DGX de NVIDIAhttps://www.nvidia.com/en-us/data-center/dgx-station/["https://www.nvidia.com/en-us/data-center/dgx-station/"^]
  - GPU de núcleo tensor NVIDIA V100https://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/"^]
- NVIDIA Jarvis Multimodal Framework
  - NVIDIA Jarvishttps://developer.nvidia.com/nvidia-jarvis["https://developer.nvidia.com/nvidia-jarvis"^]
  - Acceso temprano 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"^]

    - Guía para desarrolladoreshttps://nvidia.github.io/NeMo/["https://nvidia.github.io/NeMo/"^]
- Sistemas AFF de NetApp
    - Especificaciones técnicas de AFF a-Series de NetApphttps://www.netapp.com/us/media/ds-3582.pdf["https://www.netapp.com/us/media/ds-3582.pdf"^]
    - Ventaja de NetApp Flash para All Flash FAShttps://www.netapp.com/us/media/ds-3733.pdf["https://www.netapp.com/us/media/ds-3733.pdf"^]
    - Biblioteca de información de ONTAP 9http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286["http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286"^]
    - Informe técnico de NetApp ONTAP FlexGroup Volumeshttps://www.netapp.com/us/media/tr-4557.pdf["https://www.netapp.com/us/media/tr-4557.pdf"^]
- ONTAP AI de NetApp
    - Guía de diseño de ONTAP AI con DGX-1 y Cisco Networkinghttps://www.netapp.com/us/media/nva-1121-design.pdf["https://www.netapp.com/us/media/nva-1121-design.pdf"^]
    - Guía de puesta en marcha de ONTAP AI con DGX-1 y Cisco Networkinghttps://www.netapp.com/us/media/nva-1121-deploy.pdf["https://www.netapp.com/us/media/nva-1121-deploy.pdf"^]
    - Guía de diseño de ONTAP AI con DGX-1 y Mellanox Networkinghttp://www.netapp.com/us/media/nva-1138-design.pdf["http://www.netapp.com/us/media/nva-1138-design.pdf"^]
    - Guía de diseño de 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: Solución de orquestación de NetApp con Run:AI

Rick Huang, David Arnette, Sung-han Lin, NetApp Yaron Goldberg, Run:AI

Los sistemas de almacenamiento AFF de NetApp proporcionan un rendimiento extremo y funcionalidades de gestión de datos de cloud híbrido líderes en el sector. NetApp y Run:AI se han asociado para demostrar las funcionalidades únicas de la solución de IA ONTAP de NetApp para cargas de trabajo de inteligencia artificial (IA) y aprendizaje automático (ML) que ofrecen rendimiento, fiabilidad y soporte de clase empresarial. Ejecutar:la orquestación de la IA de las cargas de trabajo de IA añade una programación basada en Kubernetes y una plataforma de uso de recursos para ayudar a los investigadores a gestionar y optimizar la utilización de la GPU. Junto con los sistemas DGX de NVIDIA, la solución combinada de NetApp, NVIDIA y Run:AI proporciona una pila de infraestructura diseñada específicamente para cargas de trabajo de IA empresariales. Este informe técnico proporciona instrucciones direccionales a los clientes que crean sistemas de IA conversacionales que respaldan diversos casos prácticos y mercados verticales del sector. Incluye información sobre la puesta en marcha de Run:AI y un sistema de almacenamiento AFF A800 de NetApp y sirve como arquitectura de referencia para obtener el método más sencillo de poner en marcha con éxito iniciativas de IA de forma rápida.

El público objetivo de la solución incluye los siguientes grupos:

- Los arquitectos empresariales que diseñan soluciones para el desarrollo de modelos de IA y software para casos prácticos basados en Kubernetes, como microservicios en contenedores
- Los científicos de datos que buscan formas eficientes de alcanzar objetivos de desarrollo de modelos eficientes en un entorno de clúster con varios equipos y proyectos
- Ingenieros de datos a cargo del mantenimiento y la ejecución de modelos de producción
- Responsables de la toma de decisiones de TECNOLOGÍA y ejecutivos y líderes empresariales que desean crear la experiencia óptima de utilización de recursos de clúster de Kubernetes y lograr el plazo de comercialización más rápido posible gracias a las iniciativas de IA

Descripción general de la solución

Plano de control de la IA y la IA con ONTAP de NetApp

La arquitectura de IA de ONTAP de NetApp, desarrollada y verificada por NetApp y NVIDIA, cuenta con la tecnología de sistemas NVIDIA DGX y los sistemas de almacenamiento conectados al cloud de NetApp. Esta arquitectura de referencia proporciona a las organizaciones DE TI las siguientes ventajas:

- Elimina las complejidades de diseño
- Permite un escalado independiente de las capacidades de computación y almacenamiento
- Permite a los clientes empezar con poco y escalar sin problemas
- Ofrece opciones de almacenamiento para distintos niveles de rendimiento y coste

ONTAP AI de NetApp integra perfectamente los sistemas DGX y los sistemas de almacenamiento AFF A800 de NetApp con una red de vanguardia. Los sistemas ONTAP AI y DGX de NetApp simplifican las puestas en marcha de IA al eliminar complejidades y conjeturas en la fase de diseño. Los clientes pueden empezar poco a poco y aumentar sus sistemas de forma ininterrumpida, a la vez que gestionan de forma inteligente datos entre el perímetro, el núcleo y el cloud.

El plano de control de IA de NetApp es una solución de gestión de experimentos y datos de IA, ML y aprendizaje profundo (DL) para científicos de datos e ingenieros de datos. A medida que las organizaciones aumentan el uso de la IA, deben hacer frente a numerosos retos, como la escalabilidad de la carga de trabajo y la disponibilidad de los datos. El plano de control de IA de NetApp responde a estos retos mediante funcionalidades como clonar rápidamente un espacio de nombres de datos del mismo modo que una repo de Git y definir e implementar flujos de trabajo de entrenamiento de IA que incorporen la creación casi instantánea de bases de datos y modelos para la trazabilidad y versionado. Con el plano de control de IA de NetApp, puede replicar datos sin problemas entre sitios y regiones y aprovisionar rápidamente espacios de trabajo de los portátiles Juppyter con acceso a conjuntos de datos masivos.

Ejecutar:Plataforma de IA para orquestación de cargas de trabajo de IA

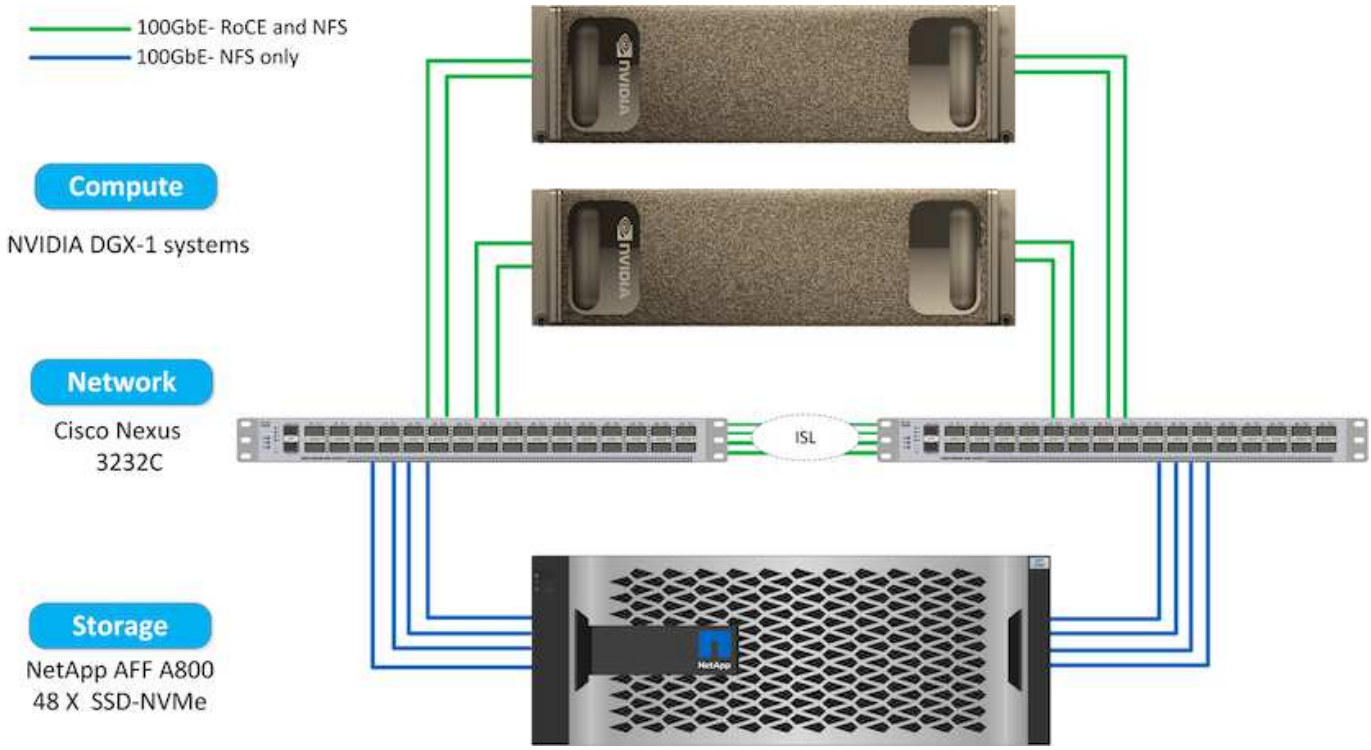
Ejecución:la IA ha creado la primera plataforma de orquestación y virtualización del mundo para la infraestructura de IA. Al abstraer las cargas de trabajo del hardware subyacente, Run:AI crea un pool compartido de recursos de GPU que se pueden aprovisionar de forma dinámica, lo que permite una orquestación eficiente de las cargas de trabajo de IA y un uso optimizado de las GPU. Los científicos de datos pueden consumir sin problemas grandes cantidades de potencia de GPU para mejorar y acelerar sus investigaciones mientras los equipos DE TECNOLOGÍA conservan el control centralizado entre sitios y la visibilidad en tiempo real sobre el aprovisionamiento de recursos, la cola y el uso. La plataforma Run:AI se ha creado sobre Kubernetes, por lo que permite una integración sencilla con los flujos de trabajo existentes de tecnología y ciencia de datos.

La plataforma Run:AI ofrece los siguientes beneficios:

- **Plazo de innovación más rápido.** con los mecanismos de puesta en cola, priorización y agrupación de recursos de IA junto con un sistema de almacenamiento de NetApp, los investigadores se eliminan de los problemas de gestión de la infraestructura y pueden centrarse exclusivamente en la ciencia de datos. Ejecución:los clientes de NetApp y IA aumentan la productividad ejecutando tantas cargas de trabajo como necesiten sin cuellos de botella en la canalización de datos o en la computación.
- **Aumento de la productividad del equipo.** Ejecutar:los algoritmos de justicia de IA garantizan que todos los usuarios y equipos obtengan su parte justa de los recursos. Es posible predefinir políticas sobre proyectos de prioridad y la plataforma permite la asignación dinámica de recursos de un usuario o equipo a otro, lo que ayuda a los usuarios a obtener un acceso puntual a los recursos de la GPU codiciados.
- **Uso mejorado de la GPU.** el Planificador Run:AI permite a los usuarios utilizar fácilmente GPU fraccionarias, GPU enteros y varios nodos de GPU para el entrenamiento distribuido en Kubernetes. De esta forma, las cargas de trabajo de IA se ejecutan según sus necesidades, no en función de la capacidad. Los equipos de ciencia de datos pueden realizar más experimentos de IA en la misma infraestructura.

Tecnología de soluciones

Esta solución se implementó con un sistema AFF A800 de NetApp, dos servidores DGX-1 y dos switches Cisco Nexus 3232C 100 GbE. Cada servidor DGX-1 está conectado a los switches Nexus mediante cuatro conexiones 100 GbE que se utilizan para las comunicaciones entre GPU. Para ello se utilizan accesos remotos directos a la memoria (RDMA) sobre Ethernet convergente (roce). En estos enlaces también se producen las comunicaciones IP tradicionales para el acceso al almacenamiento NFS. Cada controladora de almacenamiento está conectado a los switches de red mediante cuatro enlaces 100 GbE. En la siguiente figura se muestra la arquitectura de la solución de IA de ONTAP utilizada en este informe técnico para todos los escenarios de pruebas.



Hardware utilizado en esta solución

Esta solución se validó con la arquitectura de referencia ONTAP AI dos nodos DGX-1 y un sistema de almacenamiento AFF A800. Consulte "NVA-1121" para obtener más detalles sobre la infraestructura utilizada en esta validación.

En la siguiente tabla se enumeran los componentes de hardware necesarios para implementar la solución según se ha probado.

Hardware subyacente	Cantidad
Sistemas DGX-1	2
A800 de AFF	1
Switches Nexus 3232C	2

Requisitos de software

Esta solución se validó con una puesta en marcha de Kubernetes básica con el operador Run:AI instalado. Kubernetes se puso en marcha usando el "DeepOps de NVIDIA" motor de puesta en marcha, que implementa todos los componentes necesarios para un entorno listo para la producción. DeepOps ya se implementa automáticamente "Trident de NetApp" Para la integración del almacenamiento persistente con el entorno k8S, se crearon las clases de almacenamiento predeterminadas, de modo que los contenedores aprovechan el almacenamiento del sistema de almacenamiento A800 de AFF. Para obtener más información sobre Trident con Kubernetes en ONTAP AI, consulte "TR-4798".

En la siguiente tabla se enumeran los componentes de software necesarios para implementar la solución según se ha probado.

De NetApp	Versión u otra información
Software de gestión de datos ONTAP de NetApp	9.6p4
Firmware de switch Cisco NX-OS	7.0(3)I6(1)
SO DGX DE NVIDIA	4.0.4 - Ubuntu 18.04 LTS
La versión de Kubernetes	1.17
Versión de Trident	20.04.0
Ejecución: CLI de IA	v2.1.13
Ejecutar:versión del operador de Kubernetes de orquestación de IA	1.0.39
Plataforma contenedora Docker	18.06.1-ce [e68fc7a]

Requisitos de software adicionales para la ejecución: Se puede encontrar IA en "Ejecutar:requisitos previos del clúster de GPU de IA".

Óptimo uso de clúster y de la GPU con Run:AI

En las siguientes secciones se ofrecen detalles sobre la instalación de Run:AI, escenarios de prueba y resultados realizados en esta validación.

Validamos el funcionamiento y el rendimiento de este sistema mediante herramientas estándar del sector,

incluidas las pruebas de rendimiento TensorFlow. El conjunto de datos ImageNET se utilizó para entrenar ResNet-50, que es un famoso modelo de red neuronal convolucional (CNN) DL para la clasificación de imágenes. RESNET-50 ofrece un resultado de entrenamiento preciso con un tiempo de procesamiento más rápido, lo que nos permitió generar una demanda suficiente sobre el almacenamiento.

Ejecutar:instalación de AI

Para instalar Run:AI, lleve a cabo los siguientes pasos:

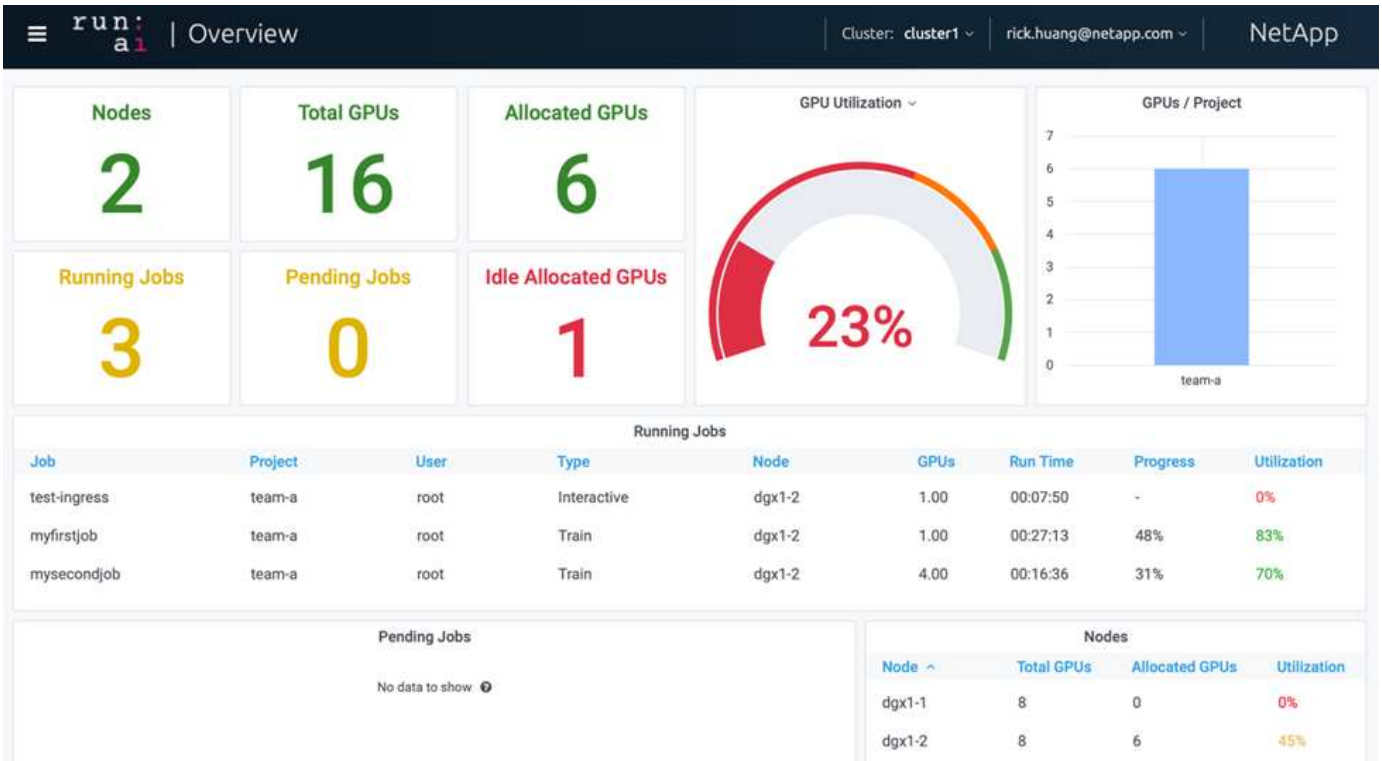
1. Instale el clúster de Kubernetes con DeepOps y configure la clase de almacenamiento predeterminada de NetApp.
2. Prepare los nodos de GPU:
   a. Compruebe que los controladores de NVIDIA están instalados en los nodos de la GPU.
   b. Compruebe que `nvidia-docker` se instala y configura como el tiempo de ejecución de docker predeterminado.
3. Ejecución de instalación:AI:
   a. Inicie sesión en el "Ejecución: IU de administración de IA" para crear el clúster.
   b. Descargue el creado `runai-operator-<clustername>.yaml` archivo.
   c. Aplique la configuración del operador al clúster de Kubernetes.

```
kubectl apply -f runai-operator-<clustername>.yaml
```

4. Compruebe la instalación:
   a. Vaya a. "https://app.run.ai/".
   b. Vaya a la consola Overview.
   c. Compruebe que el número de GPU de la parte superior derecha refleja el número esperado de GPU y los nodos de GPU forman parte de la lista de servidores. Para obtener más información acerca de la puesta en marcha de Run:IA, consulte "Instalar Run:IA en un clúster de Kubernetes en las instalaciones" y.. "Instalación de la CLI Run:AI".

Ejecutar:Paneles de IA y vistas

Después de instalar Run:AI en su clúster de Kubernetes y configurar los contenedores correctamente, verá las siguientes consolas y vistas en "https://app.run.ai" en su navegador, como se muestra en la siguiente figura.



Hay 16 GPU en total en el clúster proporcionados por dos nodos DGX-1. Puede ver el número de nodos, el total de GPU disponibles, las GPU asignadas con cargas de trabajo, el número total de trabajos en ejecución, los trabajos pendientes y las GPU asignadas inactivas. En el lado derecho, el diagrama de barras muestra las GPU por proyecto, que resume cómo usan los distintos equipos el recurso de clúster. En el medio se encuentra la lista de trabajos actualmente en ejecución con detalles de trabajo, incluidos el nombre del trabajo, el proyecto, el usuario, el tipo de trabajo, El nodo en el que se ejecuta cada trabajo, el número de GPU asignados para ese trabajo, el tiempo de ejecución actual del trabajo, el progreso del trabajo en porcentaje y el uso de la GPU para ese trabajo. Tenga en cuenta que el clúster está infrautilizado (uso de la GPU al 23%) porque solo hay tres trabajos en ejecución enviados por un único equipo (`team-a`).

En la siguiente sección, mostramos cómo crear varios equipos en la pestaña proyectos y asignar GPU para cada equipo con el fin de maximizar el uso del clúster y gestionar los recursos cuando hay muchos usuarios por clúster. Los escenarios de prueba imitan entornos empresariales en los que los recursos de memoria y GPU se comparten entre cargas de trabajo de entrenamiento, inferencia e interactivas.

Creación de proyectos para equipos de ciencia de datos y asignación de GPU

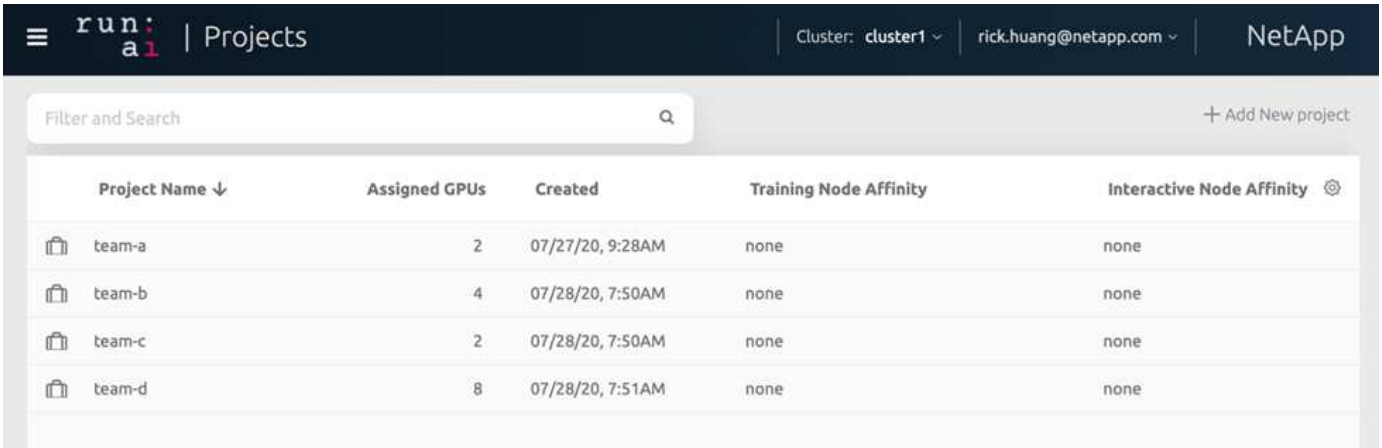
Los investigadores pueden enviar cargas de trabajo a través de la CLI Run:AI, Kubeflow o procesos similares. Para agilizar la asignación de recursos y crear prioridades, Run:AI introduce el concepto de proyectos. Los proyectos son entidades de cuota que asocian un nombre de proyecto con la asignación y las preferencias de la GPU. Se trata de una forma sencilla y cómoda de gestionar varios equipos de ciencia de datos.

Un investigador que presenta una carga de trabajo debe asociar un proyecto con una solicitud de carga de trabajo. El programador Run:AI compara la solicitud con las asignaciones actuales y el proyecto, y determina si la carga de trabajo puede asignarse recursos o si debe permanecer en estado pendiente.

Como administrador del sistema, puede establecer los siguientes parámetros en la ficha Ejecutar:proyectos AI:

- **Proyectos modelo.** establecer un proyecto por usuario, establecer un proyecto por equipo de usuarios y establecer un proyecto por proyecto de organización real.
- * Cuotas del proyecto.* cada proyecto está asociado con una cuota de GPU que se pueden asignar para este proyecto al mismo tiempo. Se trata de una cuota garantizada en el sentido de que se garantiza que los investigadores que utilizan este proyecto obtengan este número de GPU, independientemente del estado del clúster. Por lo general, la suma de la asignación de proyectos debe ser igual al número de GPU del clúster. Más allá de eso, un usuario de este proyecto puede recibir un exceso de cuota. Mientras no se utilicen las GPU, un investigador que usa este proyecto puede obtener más GPU. Demostramos escenarios de prueba de exceso de cuota y consideraciones de equidad en "Utilización óptima del clúster gracias a la asignación de GPU por encima de una cuota", "Equidad en la asignación de recursos básicos", y. "Justicia por exceso de cuotas".
- Cree un proyecto nuevo, actualice un proyecto existente y elimine un proyecto existente.
- **Limite los trabajos para que se ejecuten en grupos de nodos específicos**. Puede asignar proyectos específicos para que se ejecuten solo en nodos específicos. Esto resulta útil cuando el equipo de proyecto necesita hardware especializado, por ejemplo, con suficiente memoria. Como alternativa, un equipo de proyecto podría ser el propietario de hardware específico que se adquirió con un presupuesto especializado o cuando sea necesario dirigir las cargas de trabajo de creación o interactivas para trabajar en hardware más débil y dirigir la formación de mayor duración o las cargas de trabajo sin supervisión a nodos más rápidos. Para ver los comandos para agrupar nodos y establecer la afinidad de un proyecto específico, consulte "Ejecute:Documentación de IA".
- **Limitar la duración de los trabajos interactivos**. Los investigadores se olvidan frecuentemente de cerrar trabajos interactivos. Esto podría conducir a una pérdida de recursos. Algunas organizaciones prefieren limitar la duración de los trabajos interactivos y cerrarlos automáticamente.

La siguiente figura muestra la vista proyectos con cuatro equipos creados. A cada equipo se le asigna un número diferente de GPU para dar cuenta de diferentes cargas de trabajo, con un número total de GPU igual al de las GPU disponibles totales en un clúster que consta de dos DGX-1s.



Enviar trabajos en ejecución:AI CLI

En esta sección se ofrecen detalles sobre los comandos básicos Run:AI que puede utilizar para ejecutar cualquier trabajo con Kubernetes. Se divide en tres partes según el tipo de carga de trabajo. Las cargas de trabajo de IA/ML/AP se pueden dividir en dos tipos genéricos:

- **Sesiones de formación desatendida**. Con estos tipos de cargas de trabajo, el científico de datos prepara una carga de trabajo de ejecución automática y la envía para su ejecución. Durante la ejecución, el cliente puede examinar los resultados. Este tipo de carga de trabajo se utiliza a menudo en la producción o

cuando el desarrollo de modelos se encuentra en una etapa en la que no se requiere intervención humana.

- **Sesiones interactivas de construcción**. Con este tipo de cargas de trabajo, el científico de datos abre una sesión interactiva con Bash, Juppyter Notebook, PyCharm remoto o IDE similares y accede directamente a los recursos de la GPU. Incluimos un tercer escenario para ejecutar cargas de trabajo interactivas con puertos conectados a fin de revelar un puerto interno al usuario del contenedor.

Cargas de trabajo de formación desatendida

Después de configurar los proyectos y asignar GPU, puede ejecutar cualquier carga de trabajo de Kubernetes usando el siguiente comando en la línea de comandos:

```
$ runai project set team-a runai submit hyper1 -i gcr.io/run-ai-
demo/quickstart -g 1
```

Este comando inicia un trabajo de entrenamiento desatendido para el equipo a con una asignación de una única GPU. El trabajo se basa en una imagen de docker de muestra, `gcr.io/run-ai-demo/quickstart`. Nosotros nombramos el trabajo `hyper1`. A continuación, puede supervisar el progreso del trabajo ejecutando el siguiente comando:

```
$ runai list
```

En la siguiente figura se muestra el resultado del `runai list` comando. Los Estados típicos que puede ver incluyen los siguientes:

- `ContainerCreating`. El contenedor docker se está descargando del repositorio en la nube.
- `Pending`. El trabajo está a la espera de ser programado.
- `Running`. El trabajo se está ejecutando.

```
~> 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
```

Para obtener un estado adicional en su trabajo, ejecute el siguiente comando:

```
$ runai get hyper1
```

Para ver los registros del trabajo, ejecute el `runai logs <job-name>` comando:

```
$ runai logs hyper1
```

En este ejemplo, debería ver el registro de una sesión DL en ejecución, incluyendo la época de entrenamiento actual, ETA, valor de la función de pérdida, precisión y tiempo transcurrido para cada paso.

Puede ver el estado del clúster en la interfaz de usuario Run:AI en "https://app.run.ai/". En Paneles >

Descripción general, puede supervisar el uso de la GPU.

Para detener esta carga de trabajo, ejecute el siguiente comando:

```
$ runai delte hyper1
```

Este comando detiene la carga de trabajo de entrenamiento. Puede verificar esta acción ejecutando `runai list` de nuevo. Para obtener información detallada, consulte "iniciar cargas de trabajo de formación sin supervisión".

Cargas de trabajo de compilación interactiva

Después de configurar proyectos y asignar GPU, puede ejecutar una carga de trabajo de compilación interactiva utilizando el siguiente comando en la línea de comandos:

```
$ runai submit build1 -i python -g 1 --interactive --command sleep --args
infinity
```

El trabajo se basa en un pitón de imagen de Docker de muestra. Nombramos el edificio de trabajo 1.

> La `-- interactive` indicador significa que el trabajo no tiene inicio ni fin Es responsabilidad del investigador cerrar el trabajo. El administrador puede definir un límite de tiempo para los trabajos interactivos después de que el sistema los termine.

La `--g 1` El indicador asigna una única GPU a este trabajo. El comando y el argumento proporcionados son `--command sleep—args infinity`. Debe proporcionar un comando o el contenedor se inicia y, a continuación, sale inmediatamente.

Los siguientes comandos funcionan de forma similar a los comandos descritos en Cargas de trabajo de formación desatendida:

- `runai list`: Muestra el nombre, el estado, la edad, el nodo, la imagen, Proyecto, usuario y GPU para trabajos.
- `runai get build1`: Muestra el estado adicional en la generación de trabajos 1.
- `runai delete build1`: Detiene la carga de trabajo interactiva build1.para obtener un shell bash en el contenedor, el siguiente comando:

```
$ runai bash build1
```

Esto proporciona un shell directo en el equipo. A continuación, los científicos de datos pueden desarrollar o afinar sus modelos dentro del contenedor.

Puede ver el estado del clúster en la interfaz de usuario Run:AI en "https://app.run.ai". Para obtener información detallada, consulte "iniciar y utilizar cargas de trabajo de compilación interactivas".

Cargas de trabajo interactivas con puertos conectados

Como extensión de las cargas de trabajo de compilación interactiva, puede revelar puertos internos al usuario del contenedor al iniciar un contenedor con la CLI Run:AI. Esto resulta útil para entornos cloud, trabajar con Juppyter Notebooks o conectarse a otros microservicios. "Entrada" Permite el acceso a los servicios de Kubernetes desde fuera del clúster de Kubernetes. Puede configurar el acceso mediante la creación de una colección de reglas que definan qué conexiones entrantes llegan a qué servicios.

Para mejorar la gestión del acceso externo a los servicios de un clúster, sugerimos que instalen los administradores de clúster "Entrada" Y configurar LoadBalancer.

Para utilizar Ingress como tipo de servicio, ejecute el siguiente comando para establecer el tipo de método y los puertos al enviar la carga de trabajo:

```
$ 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 vez que el contenedor se haya iniciado correctamente, ejecute `runai list` para ver la `SERVICE URL(S)` Con el que acceder al Cuaderno de Jupyter. La dirección URL está compuesta por el punto final de entrada, el nombre del trabajo y el puerto. Por ejemplo, consulte https://10.255.174.13/test-ingress-8888.

Para obtener información detallada, consulte "iniciar una carga de trabajo de compilación interactiva con puertos conectados".

Se obtiene un uso elevado del clúster

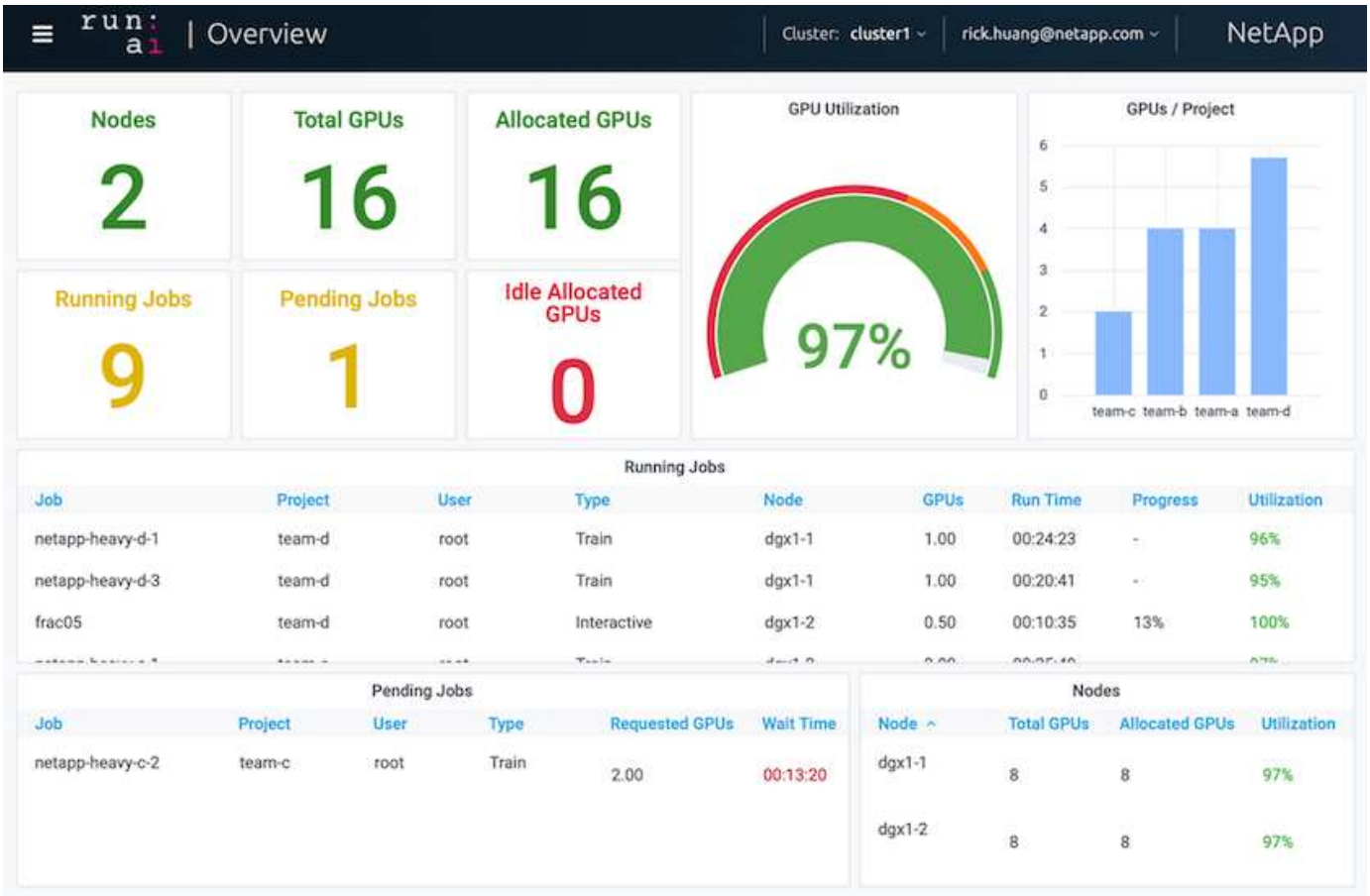
En esta sección, emulamos un escenario realista en el que cuatro equipos de ciencia de datos envían sus propias cargas de trabajo para demostrar la solución de orquestación Run:AI que logra un uso elevado del clúster mientras mantiene la priorización y el equilibrio de los recursos de la GPU. Empezamos utilizando la prueba de rendimiento ResNet-50 descrita en el apartado "RESNET-50 con resumen de la prueba de rendimiento de conjunto de datos ImageNET":

```
$ 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
```

Ejecutamos la misma prueba de rendimiento ResNet-50 como en la "NVA-1121". Usamos la bandera `--local-image` para contenedores que no residen en el repositorio de docker público. Montamos los directorios `/mnt` y.. `/tmp` En el nodo DGX-1 del host a. `/mnt` y.. `/tmp` al contenedor, respectivamente. El conjunto de datos se encuentra en NetApp AFFA800 con el `dataset_dir` argumento que apunta al directorio. Ambas `--num_devices=1` y.. `-g 1` Significa que asignamos una GPU para este trabajo. El primero es un argumento para el `run.py` script, mientras que el último es un indicador para el `runai submit` comando.

La siguiente figura muestra un panel de información general del sistema con un uso del 97 % de la GPU y las

dieciséis GPU disponibles asignadas. Puede ver fácilmente cuántas GPU se asignan a cada equipo en el gráfico de barras GPU/proyecto. El panel trabajos en ejecución muestra los nombres de los trabajos en ejecución actuales, el proyecto, el usuario, el tipo, el nodo, Las GPU consumidas, tiempo de ejecución, progreso y detalles de uso. Se muestra una lista de las cargas de trabajo que están en cola con el tiempo de espera en trabajos pendientes. Finalmente, el recuadro Nodes ofrece cifras de GPU y utilización de nodos DGX-1 individuales en el clúster.



Asignación de GPU fraccionaria para las cargas de trabajo menos exigentes o interactivas

Cuando investigadores y desarrolladores trabajan en sus modelos, ya sea en las etapas de desarrollo, ajuste de hiperparámetros o depuración, estas cargas de trabajo suelen requerir menos recursos computacionales. Por lo tanto, es más eficiente aprovisionar GPU y memoria fraccionarias de modo que la misma GPU se pueda asignar simultáneamente a otras cargas de trabajo. Ejecutar:la solución de orquestación de IA proporciona un sistema de uso compartido de GPU fraccionario para las cargas de trabajo en contenedores en Kubernetes. El sistema admite cargas de trabajo que ejecutan programas CUDA y se adapta especialmente a tareas de IA ligeras como la inferencia y la creación de modelos. El sistema de GPU fraccionaria proporciona a los equipos de ciencia de datos e ingeniería de IA la capacidad de ejecutar varias cargas de trabajo simultáneamente en una única GPU. De este modo, las empresas pueden ejecutar más cargas de trabajo, como visión informática, reconocimiento de voz y procesamiento de lenguaje natural en el mismo hardware, con lo que se reducen los costes.

Ejecutar:el sistema de GPU fraccionaria de IA crea de manera efectiva GPU lógicas virtualizadas con su propia memoria y espacio de computación que los contenedores pueden utilizar y acceder como si fueran procesadores independientes. De este modo, es posible ejecutar varias cargas de trabajo en contenedores en paralelo y en la misma GPU sin interferir entre sí. La solución es transparente, sencilla y portátil y no requiere ningún cambio en los contenedores en sí.

Una usecasa típica podría ver dos a ocho trabajos ejecutándose en la misma GPU, lo que significa que podría trabajar ocho veces más con el mismo hardware.

Para el trabajo `frac05` perteneciente al proyecto `team-d` En la siguiente figura podemos comprobar que el número de GPU asignadas era de 0.50. Esto es verificado por el `nvidia-smi` Comando, que muestra que la memoria de la GPU disponible para el contenedor era de 16,255 MB: La mitad de las 32 GB por GPU V100 en el 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 |
+-----------------------------------------------------------------------------+
```

Utilización óptima del clúster gracias a la asignación de GPU por encima de una cuota

En esta sección y en las secciones "Equidad en la asignación de recursos básicos", y. "Justicia por exceso de cuotas", Hemos diseñado escenarios de pruebas avanzados para demostrar las capacidades de orquestación Run:AI para la administración de cargas de trabajo complejas, la programación preventiva automática y el aprovisionamiento de GPU con exceso de cuota. Hemos hecho esto para lograr un uso elevado de los recursos de clúster y optimizar la productividad de los equipos científicos de datos a nivel empresarial en un entorno de IA de ONTAP.

Para estas tres secciones, defina los siguientes proyectos y cuotas:

Proyecto	Cuota
equipo a	4
equipo-b	2
equipo-c	2
equipo d	8

Además, utilizamos los siguientes contenedores para estas tres secciones:

- Portátil Jupyter: `jupyter/base-notebook`
- Ejecutar:AI Quickstart: `gcr.io/run-ai-demo/quickstart`

Definimos los siguientes objetivos para este escenario de prueba:

- Muestre la simplicidad del aprovisionamiento de recursos y cómo los recursos se abstraen de los usuarios
- Mostrar cómo los usuarios pueden aprovisionar fácilmente fracciones de una GPU y un número entero de GPU
- Muestre cómo el sistema elimina los cuellos de botella de computación al permitir que equipos o usuarios hagan uso de su cuota de recursos si hay GPU libres en el clúster
- Muestre cómo se eliminan los cuellos de botella en la canalización de datos utilizando la solución de NetApp cuando se ejecutan tareas con un gran procesamiento como el contenedor de NetApp
- Muestra cómo se ejecutan los diversos tipos de contenedores mediante el sistema
  - Portátil Jupyter
  - Ejecute:contenedor de IA
- Muestra una alta utilización cuando el clúster está lleno

Para obtener información detallada sobre la secuencia de comandos real ejecutada durante la prueba, consulte "Detalles de las pruebas para la sección 4.8".

Cuando se envían las 13 cargas de trabajo, puede ver una lista con los nombres de los contenedores y las GPU asignadas, como se muestra en la siguiente figura. Contamos con siete cursos de formación y seis trabajos interactivos, que simulan cuatro equipos de ciencia de datos, cada uno con sus propios modelos en ejecución o en desarrollo. Para trabajos interactivos, los desarrolladores individuales están utilizando Jupyter Notebooks para escribir o depurar su código. Por lo tanto, es adecuado aprovisionar fracciones de GPU sin usar demasiados recursos de clúster.

```
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
```

Los resultados de este escenario de prueba muestran lo siguiente:

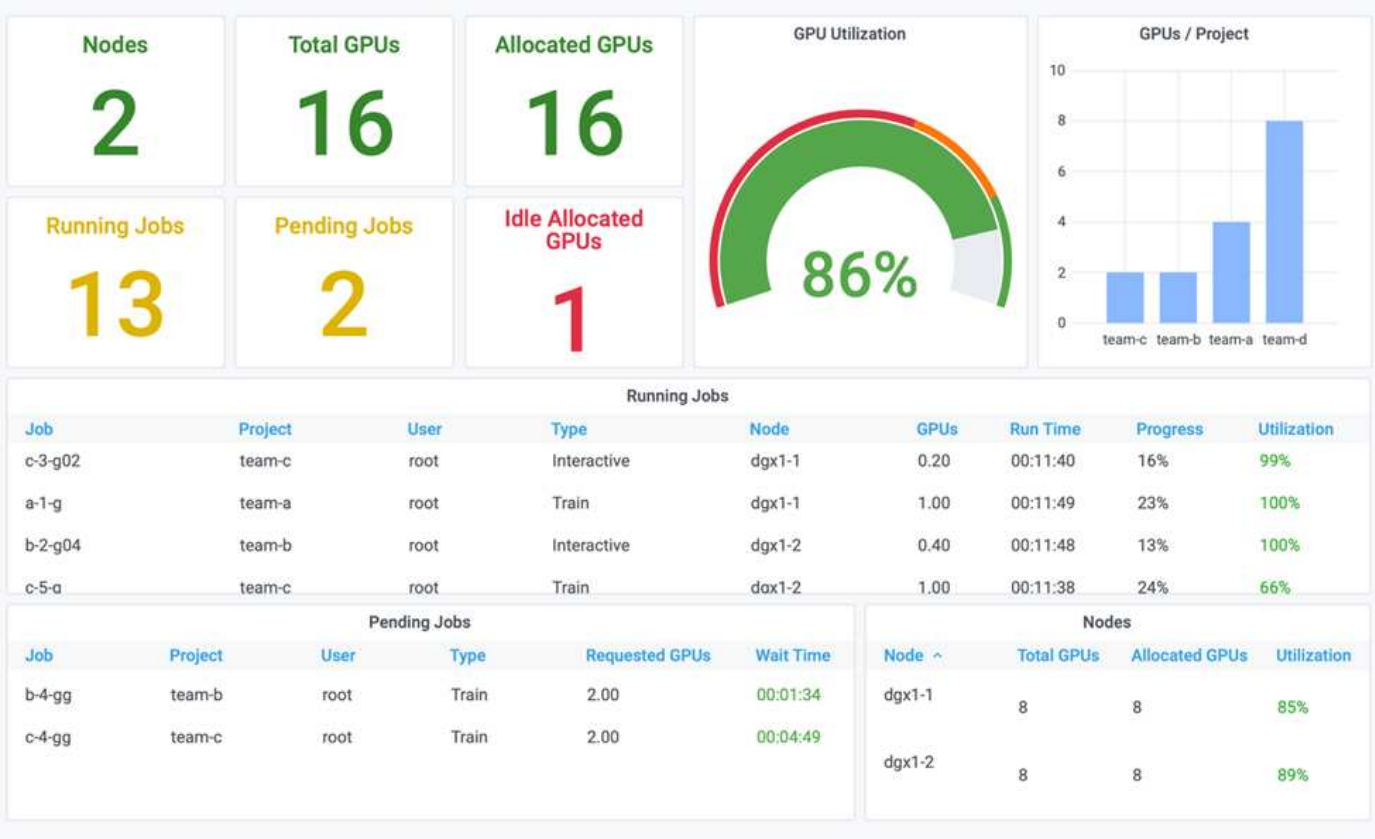
- El clúster debe estar lleno: Se utilizan 16/16 GPU.
- Uso elevado de clúster.
- Más experimentos que las GPU debido a la asignación fraccionaria.
- `team-d` no está utilizando toda su cuota; por lo tanto, `team-b` y.. `team-c` Puede utilizar más GPU para sus experimentos, lo que acelera el plazo de innovación.

Equidad en la asignación de recursos básicos

En esta sección, lo mostramos cuando `team-d` Solicita más GPU (están por debajo de su cuota), el sistema coloca en pausa las cargas de trabajo de `team-b` y.. `team-c` y los mueve a un estado pendiente de manera equitativa.

Para obtener información detallada, incluidos los envíos de trabajos, las imágenes contenedoras utilizadas y las secuencias de comandos ejecutadas, consulte la sección "Detalles de las pruebas para la sección 4.9".

La siguiente figura muestra el uso resultante del clúster, las GPU asignadas por equipo y los trabajos pendientes debido al equilibrio de carga automático y a la programación preventiva. Podemos observar que cuando el número total de GPU solicitado por todas las cargas de trabajo de equipo supera el total de GPU disponibles en el clúster, el algoritmo de integridad interno de Run:AI pone en pausa un trabajo por cada una `team-b` y.. `team-c` porque han cumplido con su cuota de proyecto. Esto proporciona una utilización elevada del clúster global, mientras que los equipos de ciencia de datos siguen trabajando con las limitaciones de recursos definidas por un administrador.



Los resultados de este escenario de prueba demuestran lo siguiente:

- **Equilibrio de carga automático.** el sistema equilibra automáticamente la cuota de las GPU, de modo que cada equipo utiliza ahora su cuota. Las cargas de trabajo en pausa pertenecen a equipos que se

encontraban por encima de su cuota.

- **Pausa de uso compartido justo.** el sistema elige detener la carga de trabajo de un equipo que estaba por encima de su cuota y luego detener la carga de trabajo del otro equipo. Ejecutar:la IA tiene algoritmos internos de justicia.

Justicia por exceso de cuotas

En esta sección, ampliamos el escenario en el que varios equipos envían cargas de trabajo y superan su cuota. De esta manera, mostramos cómo el algoritmo de equidad de Run:AI asigna recursos de clúster según la proporción de cuotas predefinidas.

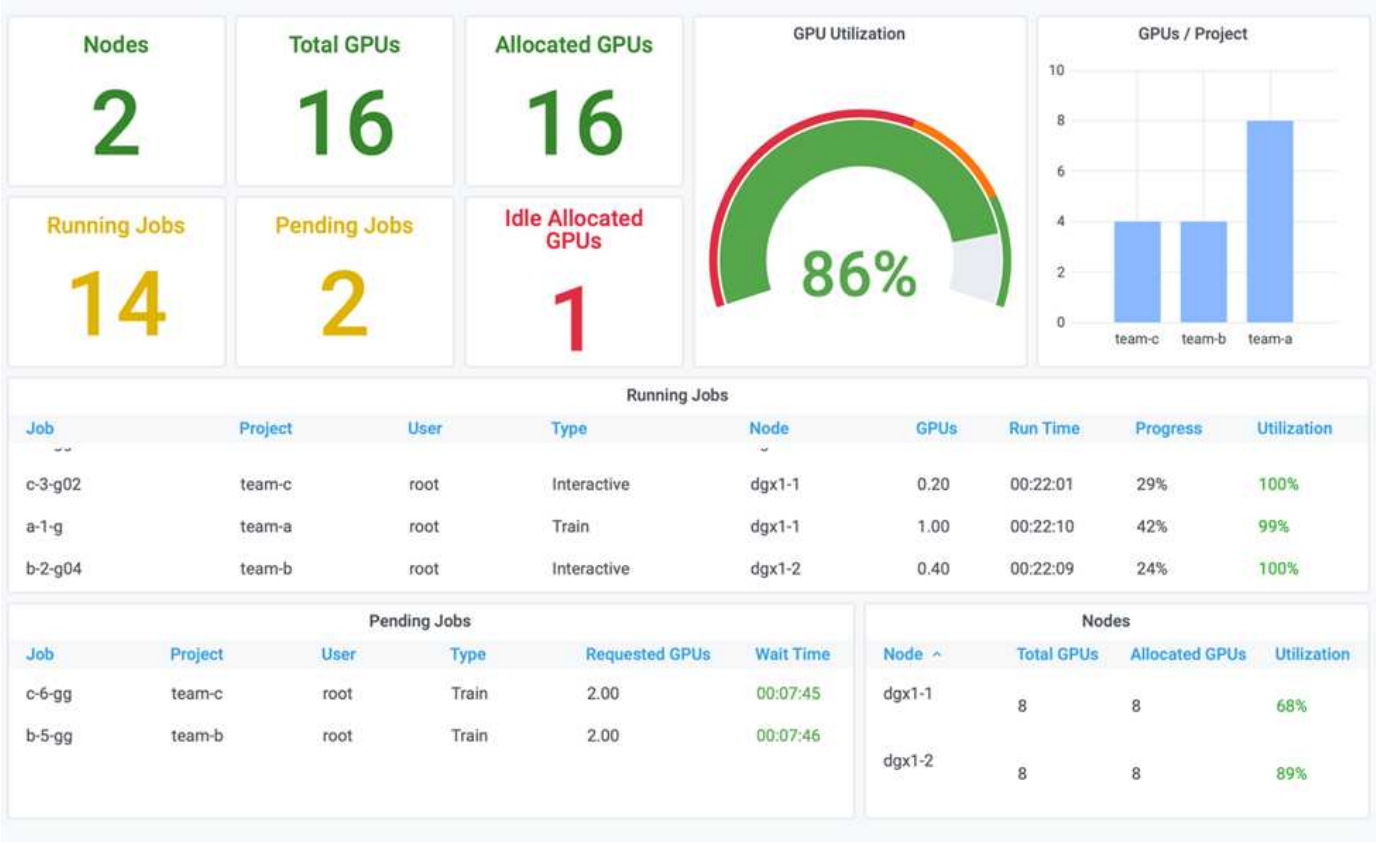
Objetivos de este escenario de prueba:

- Muestra el mecanismo de cola cuando varios equipos solicitan GPU a través de su cuota.
- Muestre cómo distribuye el sistema una cuota del clúster entre varios equipos que se encuentran por encima de su cuota en función de la proporción entre sus cuotas, de modo que el equipo con la cuota mayor obtenga una parte mayor de la capacidad de reserva.

Al final de "Equidad en la asignación de recursos básicos", hay dos cargas de trabajo en cola: una para `team-b` y uno para `team-c`. En esta sección, ponemos en cola las cargas de trabajo adicionales.

Para obtener información detallada, incluidos los envíos de trabajos, las imágenes contenedoras utilizadas y las secuencias de comandos ejecutadas, consulte "Detalles de la prueba para la sección 4.10".

Cuando todos los trabajos se someten según la sección "Detalles de la prueba para la sección 4.10", el panel del sistema muestra eso `team-a`, `team-b`, y. `team-c` Todos tienen más GPU que su cuota predefinida. `team-a` Ocupa cuatro GPU más que su cuota de software predefinida (cuatro), mientras que `team-b` y.. `team-c` Cada uno ocupa dos GPU más que su cuota blanda (dos). La proporción de GPU asignadas por encima de una cuota es igual a la de su cuota predefinida. Esto se debe a que el sistema utilizó la cuota preconfigurada como referencia de prioridad y se aprovisionó de manera acorde cuando varios equipos soliciten más GPU, por lo que superó su cuota. Este equilibrio de carga automático proporciona justicia y priorización cuando los equipos de ciencia de datos empresariales están involucrados activamente en el desarrollo y la producción de modelos de IA.



Los resultados de este escenario de prueba muestran lo siguiente:

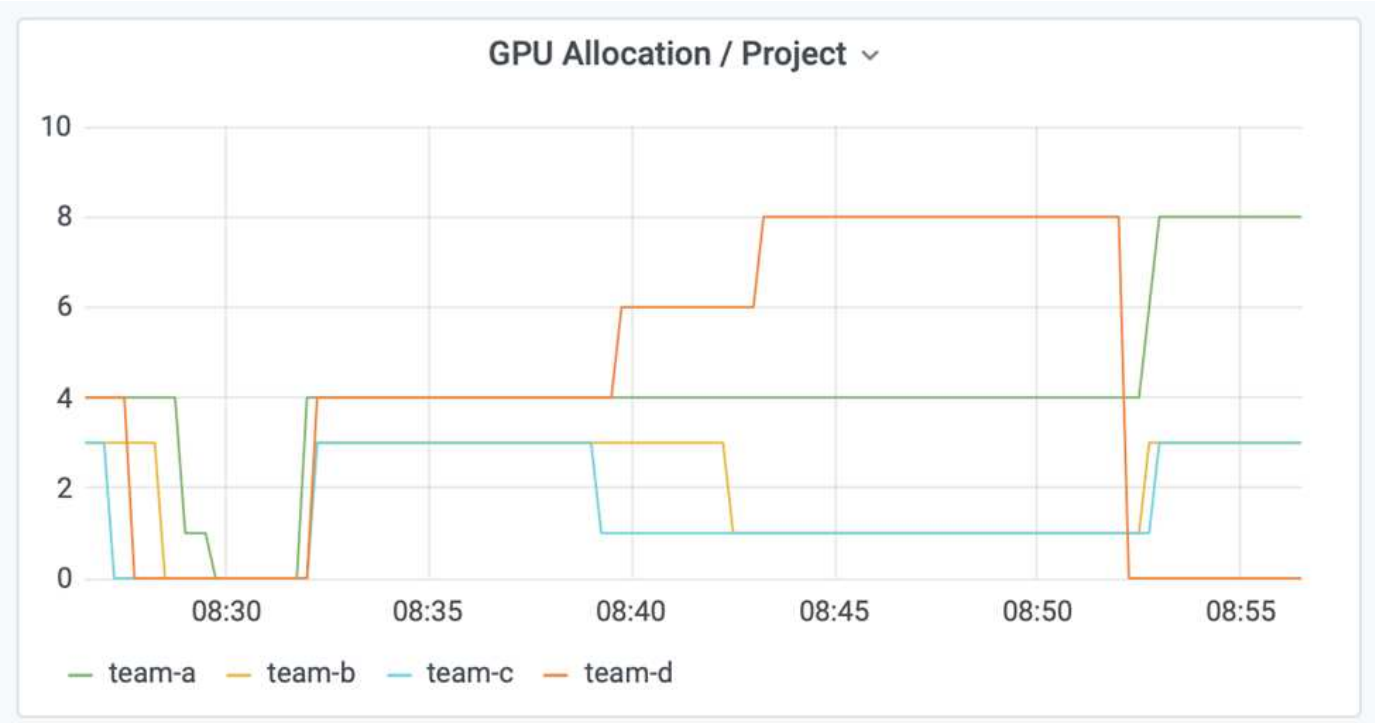
- El sistema empieza a deponer las cargas de trabajo de otros equipos en cola.
- El orden de la depuesta en cola se decide de acuerdo a los algoritmos de justicia, tal que `team-b` y.. `team-c` Obtenga la misma cantidad de GPU de cuota superior (ya que tienen una cuota similar) y. `team-a` Obtiene una cantidad doble de GPU, ya que su cuota es dos veces superior a la de `team-b` y.. `team-c`.
- Toda la asignación se realiza automáticamente.

Por lo tanto, el sistema debe estabilizarse en los siguientes estados:

Proyecto	GPU asignadas	Comentar
equipo a	8/4	Cuatro GPU a lo largo de la cuota. Cola vacía.
equipo-b	4/2	Dos GPU por encima de la cuota. Una carga de trabajo en cola.
equipo-c	4/2	Dos GPU por encima de la cuota. Una carga de trabajo en cola.
equipo d	0/8	No utiliza GPU en absoluto, sin cargas de trabajo en cola.

La siguiente figura muestra la asignación de GPU por proyecto a lo largo del tiempo en el panel Run:AI Analytics de las secciones "Utilización óptima del clúster gracias a la asignación de GPU por encima de una cuota", "Equidad en la asignación de recursos básicos", y. "Justicia por exceso de cuotas". Cada línea de la figura indica el número de GPU aprovisionadas para un equipo de ciencias de datos en cualquier momento. Podemos observar que el sistema asigna dinámicamente las GPU de acuerdo con las cargas de trabajo

enviadas. Esto permite a los equipos revisar las cuotas cuando hay GPU disponibles en el clúster y, después, adelantarse a los trabajos según sea necesario, antes de alcanzar, por fin, un estado estable para los cuatro equipos.



Guardar datos en un volumen persistente aprovisionado por Trident

Trident de NetApp es un proyecto de código abierto totalmente compatible diseñado para ayudarle a satisfacer las sofisticadas demandas de persistencia de sus aplicaciones en contenedores. Puede leer y escribir datos en un volumen persistente (VP) de Kubernetes aprovisionado por Trident con la ventaja añadida de la organización en niveles de datos, el cifrado, la tecnología Snapshot de NetApp, el cumplimiento de normativas y el alto rendimiento que ofrece el software de gestión de datos ONTAP de NetApp.

Reutilizar EVs en un espacio de nombres existente

En el caso de proyectos de IA de mayor tamaño, es posible que sea más eficiente que diferentes contenedores para leer y escribir datos en el mismo PV de Kubernetes. Para reutilizar una solicitud de volumen persistente de Kubernetes (PVC), el usuario ya debe haber creado una RVP. Consulte "Documentación de Trident de NetApp" Para obtener más detalles sobre la creación de una RVP. A continuación se muestra un ejemplo de reutilización de un PVC existente:

```
$ runai submit pvc-test -p team-a --pvc test:/tmp/pvc1mount -i gcr.io/run-
ai-demo/quickstart -g 1
```

Ejecute el siguiente comando para ver el estado del trabajo `pvc-test` para el proyecto `team-a`:

```
$ runai get pvc-test -p team-a
```

Debe ver el montaje PV /tmp/pvc1montado a. `team-a` trabajo `pvc-test`. De este modo, se pueden leer varios contenedores desde el mismo volumen, lo que resulta útil cuando hay varios modelos competidores en desarrollo o producción. Los científicos de datos pueden crear un conjunto de modelos y, a continuación, combinar los resultados de las predicciones por voto mayoritario u otras técnicas.

Utilice lo siguiente para acceder al shell del contenedor:

```
$ runai bash pvc-test -p team-a
```

A continuación, puede comprobar el volumen montado y acceder a los datos dentro del contenedor.

Esta capacidad de reutilización de RVP funciona con volúmenes FlexVol de NetApp y volúmenes ONTAP FlexGroup de NetApp, lo que permite a los ingenieros de datos disfrutar de opciones de gestión de datos más flexibles y sólidas para aprovechar su Data Fabric con tecnología de NetApp.

Conclusión

NetApp y Run:AI se han asociado en este informe técnico para mostrar las exclusivas funcionalidades de la solución ONTAP AI de NetApp junto con la plataforma Run:AI para simplificar la orquestación de las cargas de trabajo de IA. Los pasos anteriores proporcionan una arquitectura de referencia para optimizar el proceso de canalizaciones de datos y orquestación de cargas de trabajo para el aprendizaje profundo. Animamos a los clientes que deseen poner en marcha estas soluciones a NetApp y a Run:AI para obtener más información.

Detalles de las pruebas para la sección 4.8

Esta sección contiene los detalles de las pruebas de la sección "Utilización óptima del clúster gracias a la asignación de GPU por encima de una cuota".

Enviar trabajos en el siguiente orden:

Proyecto	Imagen	N.o de GPU	Total	Comentar
equipo a	Juppyter	1	1/4	–
equipo a	NetApp	1	2/4	–
equipo a	Ejecución: IA	2	4/4	Usando toda su cuota
equipo-b	Ejecución: IA	0.6	0.6/2	GPU fraccionaria
equipo-b	Ejecución: IA	0.4	1/2	GPU fraccionaria
equipo-b	NetApp	1	2/2	–
equipo-b	NetApp	2	4/2	Dos sobre cuota
equipo-c	Ejecución: IA	0.5	0.5/2	GPU fraccionaria
equipo-c	Ejecución: IA	0.3	0.8/2	GPU fraccionaria
equipo-c	Ejecución: IA	0.2	1/2	GPU fraccionaria

Proyecto	Imagen	N.o de GPU	Total	Comentar
equipo-c	NetApp	2	3/2	Uno sobre cuotas
equipo-c	NetApp	1	4/2	Dos sobre cuota
equipo d	NetApp	4	4/8	Utilizando la mitad de su cuota

Estructura de comandos:

```
$ runai submit <job-name> -p <project-name> -g <#GPUs> -i <image-name>
```

Secuencia de comandos real utilizada en la prueba:

```
$ 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
```

En este punto, debe tener los siguientes estados:

Proyecto	GPU asignadas	Cargas de trabajo en cola
equipo a	4/4 (cuota suave/asignación real)	Ninguno
equipo-b	4/2	Ninguno
equipo-c	4/2	Ninguno
equipo d	4/8	Ninguno

Consulte la sección "Mayor uso de clúster con la asignación de GPU over-uota" para discusiones sobre el escenario de pruebas en curso.

Detalles de las pruebas para la sección 4.9

Esta sección contiene detalles de la prueba para la sección "Equidad en la asignación de recursos básicos".

Enviar trabajos en el siguiente orden:

Proyecto	N.o de GPU	Total	Comentar
equipo d	2	6/8	La carga de trabajo Team-b/c se pone en pausa y se mueve a. `pending`.
equipo d	2	8/8	Las cargas de trabajo de otros equipos (b/c) se pausan y pasan a. `pending`.

Consulte la siguiente secuencia de comandos ejecutada:

```
$ 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
```

En este punto, debe tener los siguientes estados:

Proyecto	GPU asignadas	Cargas de trabajo en cola
equipo a	4/4	Ninguno
equipo-b	2/2	Ninguno
equipo-c	2/2	Ninguno
equipo d	8/8	Ninguno

Consulte la sección "Equidad en la asignación de recursos básicos" para un debate sobre el escenario de pruebas en curso.

Detalles de las pruebas para la sección 4.10

Esta sección contiene detalles de la prueba para la sección "Justicia por exceso de cuotas".

Enviar trabajos en el siguiente orden para `team-a`, `team-b`, y. `team-c`:

Proyecto	N.o de GPU	Total	Comentar
equipo a	2	4/4	1 carga de trabajo en cola

Proyecto	N.o de GPU	Total	Comentar
equipo a	2	4/4	2 cargas de trabajo en cola
equipo-b	2	2/2	2 cargas de trabajo en cola
equipo-c	2	2/2	2 cargas de trabajo en cola

Consulte la siguiente secuencia de comandos ejecutada:

```
$ 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
```

En este punto, debe tener los siguientes estados:

Proyecto	GPU asignadas	Cargas de trabajo en cola
equipo a	4/4	Dos cargas de trabajo que solicitan GPU dos cada una
equipo-b	2/2	Dos cargas de trabajo que solicitan dos GPU cada una
equipo-c	2/2	Dos cargas de trabajo que solicitan dos GPU cada una
equipo d	8/8	Ninguno

A continuación, elimine todas las cargas de trabajo para `team-d`:

```
$ runai delete -p team-d d-1-gggg d-2-gg d-3-gg
```

Consulte la sección "Justicia por exceso de cuotas", para las discusiones sobre el escenario de pruebas en curso.

Dónde encontrar información adicional

Si quiere obtener más información sobre el contenido de este documento, consulte los siguientes recursos:

- Sistemas DGX de NVIDIA
  - Sistema DGX-1 de NVIDIAhttps://www.nvidia.com/en-us/data-center/dgx-1/[]
  - GPU de núcleo tensor NVIDIA V100https://www.nvidia.com/en-us/data-center/tesla-v100/[]
  - NVIDIA NGChttps://www.nvidia.com/en-us/gpu-cloud/[]

- Ejecute:solución de orquestación de contenedores de IA
  - Ejecución: Introducción de producto de IAhttps://docs.run.ai/home/components/[]
  - Ejecución:documentación de instalación de IAhttps://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/
  - Enviar trabajos en Ejecutar: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-/
  - Asignar fracciones de GPU en la CLI de Run:AIhttps://docs.run.ai/Researcher/Walkthroughs/Walkthrough-Using-GPU-Fractions/[]
- Plano de control de IA de NetApp
  - Informe técnicohttps://www.netapp.com/us/media/tr-4798.pdf[]
  - Demostración en formato brevehttps://youtu.be/gfr_sO27Rvo[]
  - Repositorio de GitHubhttps://github.com/NetApp/kubeflow_jupyter_pipeline[]
- Sistemas AFF de NetApp
  - Especificaciones técnicas de AFF a-Series de NetApphttps://www.netapp.com/us/media/ds-3582.pdf[]
  - Ventaja de NetApp Flash para All Flash FAShttps://www.netapp.com/us/media/ds-3733.pdf[]
  - Biblioteca de información de ONTAP 9http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286[]
  - Informe técnico de NetApp ONTAP FlexGroup Volumeshttps://www.netapp.com/us/media/tr-4557.pdf[]
- ONTAP AI de NetApp
  - Guía de diseño de ONTAP AI con DGX-1 y Cisco Networkinghttps://www.netapp.com/us/media/nva-1121-design.pdf[]
  - Guía de puesta en marcha de ONTAP AI con DGX-1 y Cisco Networkinghttps://www.netapp.com/us/media/nva-1121-deploy.pdf[]
  - Guía de diseño de ONTAP AI con DGX-1 y Mellanox Networkinghttp://www.netapp.com/us/media/nva-1138-design.pdf[]
  - Guía de diseño de ONTAP AI con DGX-2https://www.netapp.com/us/media/nva-1135-design.pdf[]

TR-4799-DESIGN: Arquitectura de referencia ONTAP AI de NetApp para cargas de trabajo de conducción autónoma

David Arnette y Sung-han Lin, NetApp

La familia de sistemas NVIDIA DGX es la primera plataforma de inteligencia artificial (IA) integrada del mundo, específicamente diseñada para IA empresarial. Los sistemas de almacenamiento AFF de NetApp proporcionan un rendimiento extremo y funcionalidades de gestión de datos de cloud híbrido líderes en el sector. NetApp y NVIDIA se han asociado para crear la arquitectura de referencia de IA ONTAP de NetApp con el fin de ofrecer a los clientes una solución lista para usar que respalda cargas de trabajo DE IA y aprendizaje automático (ML) con rendimiento, fiabilidad y soporte de nivel empresarial.

"TR-4799-DESIGN: Arquitectura de referencia ONTAP AI de NetApp para cargas de trabajo de conducción

autónoma"

TR-4811: Arquitectura de referencia de IA ONTAP de NetApp para el sector sanitario: Imágenes de diagnóstico - diseño de la solución

Rick Huang, Sung-han Lin, Sathish Thyagarajan, NetApp Jacci Cenci, NVIDIA

Esta arquitectura de referencia ofrece directrices para los clientes que crean una infraestructura de inteligencia artificial (IA) con sistemas NVIDIA DGX-2 y almacenamiento AFF de NetApp para casos de uso de sanidad. Incluye información sobre los flujos de trabajo de alto nivel utilizados en el desarrollo de modelos de aprendizaje profundo (DL) para imágenes de diagnóstico médico, casos de prueba validados y resultados. También incluye recomendaciones para el dimensionamiento de las puestas en marcha de los clientes.

"TR-4811: Arquitectura de referencia de IA ONTAP de NetApp para el sector sanitario: Imágenes de diagnóstico - diseño de la solución"

TR-4807: Arquitectura de referencia de IA de ONTAP de NetApp para cargas de trabajo de servicios financieros - Diseño de la solución

Karthikeyan Nagalingam, Sung-han Lin, NetApp Jacci Cenci, NVIDIA

Esta arquitectura de referencia ofrece directrices a los clientes que crean una infraestructura de inteligencia artificial con los sistemas NVIDIA DGX-1 y el almacenamiento AFF de NetApp para casos de uso con sectores financieros. Por ejemplo, incluye información sobre los flujos de trabajo de alto nivel utilizados en el desarrollo de modelos de aprendizaje profundo para los casos de pruebas y los resultados de servicios financieros. También incluye recomendaciones para el dimensionamiento de las puestas en marcha de los clientes.

"TR-4807: Arquitectura de referencia de IA de ONTAP de NetApp para cargas de trabajo de servicios financieros - Diseño de la solución"

IA generativa y valor de NetApp

Autor: Sathish Thyagarajan, NetApp

Sinopsis

La demanda de inteligencia artificial (IA) generativa está impulsando la disrupción en todos los sectores y mejorando la creatividad empresarial y la innovación de productos. Muchas organizaciones usan la IA generativa para crear nuevas características de productos, mejorar la productividad de ingeniería y crear prototipos de aplicaciones impulsadas por IA que ofrezcan mejores resultados y experiencias de los consumidores. La IA generativa, como los Transformadores Generativos Pre-entrenados (GPT), utilizan redes

neuronales para crear nuevo contenido, tan diverso como texto, audio y vídeo. Dada la escala extrema y los conjuntos de datos masivos que participan en los modelos de lenguaje grande (LLM), es crucial diseñar una infraestructura de IA robusta que aproveche las atractivas funciones de almacenamiento de datos de las opciones de puesta en marcha locales, híbridas y multicloud, y reducir los riesgos asociados a la movilidad de datos, la gobernanza y la protección de datos antes de que las empresas puedan diseñar soluciones de IA. En este documento se describen estas consideraciones y las capacidades de IA de NetApp® correspondientes que permiten una gestión de datos y un movimiento de datos sin interrupciones en la canalización de datos de IA para el entrenamiento, el reciclaje, el ajuste y la inferencia de modelos de IA generativos.

Resumen ejecutivo

Más recientemente, después del lanzamiento de ChatGPT, un spin-off de GPT-3 en noviembre de 2022, las nuevas herramientas de IA utilizadas para generar texto, código, imagen o incluso proteínas terapéuticas en respuesta a las indicaciones de los usuarios han ganado fama significativa. Esto indica que los usuarios pueden realizar una solicitud utilizando lenguaje natural y la IA interpretará y generará texto, como artículos de noticias o descripciones de productos que reflejen la solicitud del usuario o produzcan código, música, voz, efectos visuales y activos 3D mediante algoritmos entrenados en datos ya existentes. Como resultado, frases como la difusión estable, alucinaciones, ingeniería de avisos y alineación de valores están emergiendo rápidamente en el diseño de sistemas de IA. Estos modelos de aprendizaje automático (ML) autosupervisados o semisupervisados están cada vez más disponibles como modelos de base preformados (FM) a través de los proveedores de servicios cloud y otros proveedores de firma de IA, que están adoptando diversos establecimientos empresariales de diferentes sectores para una amplia gama de tareas descendentes de procesamiento del lenguaje natural (NLP). Como afirman firmas de analistas de investigación como McKinsey: “El impacto de la IA generativa en la productividad podría agregar billones de dólares en valor a la economía global”. Mientras que las empresas se están replanteando la IA como partners de pensamiento para los humanos y FMS se están ampliando a la vez a lo que las empresas e instituciones pueden hacer con la IA generativa, las oportunidades para gestionar volúmenes masivos de datos seguirán creciendo. Este documento presenta información introductoria sobre IA generativa y los conceptos de diseño con relación a las funcionalidades de NetApp que aportan valor a los clientes de NetApp, tanto en entornos híbridos como en las instalaciones o multicloud.

Por lo tanto, ¿qué es lo que los clientes utilizan NetApp en sus entornos de IA? NetApp ayuda a las organizaciones a satisfacer las complejidades creadas por el rápido crecimiento de datos y nube, la gestión de múltiples nubes y la adopción de tecnologías de nueva generación, como la IA. NetApp ha combinado varias funcionalidades en software de gestión inteligente de datos e infraestructuras de almacenamiento que han estado bien equilibradas con un alto rendimiento optimizado para cargas de trabajo de IA. Las soluciones de IA generativas como las LLM necesitan leer y procesar sus conjuntos de datos de origen desde el almacenamiento a la memoria numerosas veces para fomentar la inteligencia. NetApp ha sido una empresa líder en tecnologías de movilidad de datos, gobernanza de datos y seguridad de datos en el ecosistema del perímetro al núcleo y al cloud, sirviendo a los clientes empresariales a crear soluciones de IA a escala. NetApp, con una sólida red de partners ha ayudado a directores de datos, ingenieros de IA, arquitectos empresariales y científicos de datos en el diseño de una canalización de datos libre para la preparación, la protección de datos así como las responsabilidades estratégicas de gestión de datos de la formación e inferencia de modelos de IA, optimizando el rendimiento y la escalabilidad del ciclo de vida de la IA/ML. Tecnologías y capacidades de datos de NetApp, como NetApp® ONTAP AI® para la canalización de datos de aprendizaje profundo, NetApp® SnapMirror® para transportar datos de forma fluida y eficiente entre puntos finales de almacenamiento, y NetApp® FlexCache® para el renderizado en tiempo real Cuando el flujo de datos cambia de batch a tiempo real y la ingeniería de datos ocurre en un momento oportuno, aporta valor a la implementación de modelos de IA generativa en tiempo real. A medida que las empresas de todo tipo adoptan nuevas herramientas de IA, deben hacer frente a desafíos en materia de datos desde el perímetro hasta el centro de datos y el cloud que demandan soluciones de IA escalables, responsables y explicables. Como referente en datos para cloud híbrido y multicloud, NetApp se compromete a crear una red de partners y soluciones conjuntas que pueden ayudar en todos los aspectos de la construcción de una canalización de datos y lagos de datos para el entrenamiento de modelos de IA generativos (entrenamiento previo), el ajuste,

la inferencia basada en el contexto y la supervisión del deterioro de modelos de las LL.

¿Qué es la IA generativa?

La IA generativa está cambiando la forma en que creamos contenido, generamos nuevos conceptos de diseño y exploramos composiciones novedosas. Ilustra los marcos de redes neuronales como la Red Adversario Generativa (GAN), los Autocodificadores Variacionales (VAE) y los Transformadores Generativos Pre-entrenados (GPT), que pueden generar contenido nuevo, como texto, código, imágenes, audio, vídeo, etc. y datos sintéticos. Los modelos basados en transformadores como Chat-GPT de OpenAI, Bard de Google, abrazar la FLORACIÓN DE Face y la llama de Meta han surgido como la tecnología fundamental que sustenta muchos avances en modelos lingüísticos grandes. Del mismo modo, Dall-E de OpenAI, Meta's CM3leon y Google's Imagen son ejemplos de modelos de difusión de texto a imagen que ofrecen a los clientes un grado de fotorealismo sin precedentes para crear imágenes nuevas y complejas desde cero o editar imágenes existentes para generar imágenes contextuales de alta calidad utilizando el aumento de conjuntos de datos y la síntesis de texto a imagen enlazando textuales y semántica visual. Los artistas digitales están empezando a aplicar una combinación de tecnologías de renderizado como Nerf (Neural Radiance Field) con IA generativa para convertir imágenes 2D estáticas en escenas 3D inmersivas. En general, las LLM se caracterizan ampliamente por cuatro parámetros: (1) Tamaño del modelo (normalmente en miles de millones de parámetros); (2) Tamaño del conjunto de datos de entrenamiento; (3) Costo de entrenamiento y (4) Rendimiento del modelo después del entrenamiento. Los LLMs también se dividen principalmente en tres arquitecturas de transformadores. (i) Modelos solo para codificadores. Por ejemplo, BERT (Google, 2018); (ii) Encoder-Decoder, por ejemplo, BART (Meta, 2020) y (iii) modelos solo Decoder. Por ejemplo, llama (Meta, 2023), Palm-E (Google, 2023). Según los requisitos del negocio, independientemente de la arquitectura que elija una compañía el número de parámetros del modelo (N) y el número de tokens (D) del conjunto de datos de entrenamiento generalmente determinan el coste base de entrenamiento (entrenamiento previo) o el ajuste de un LLM.

Casos prácticos empresariales y tareas de NLP descendentes

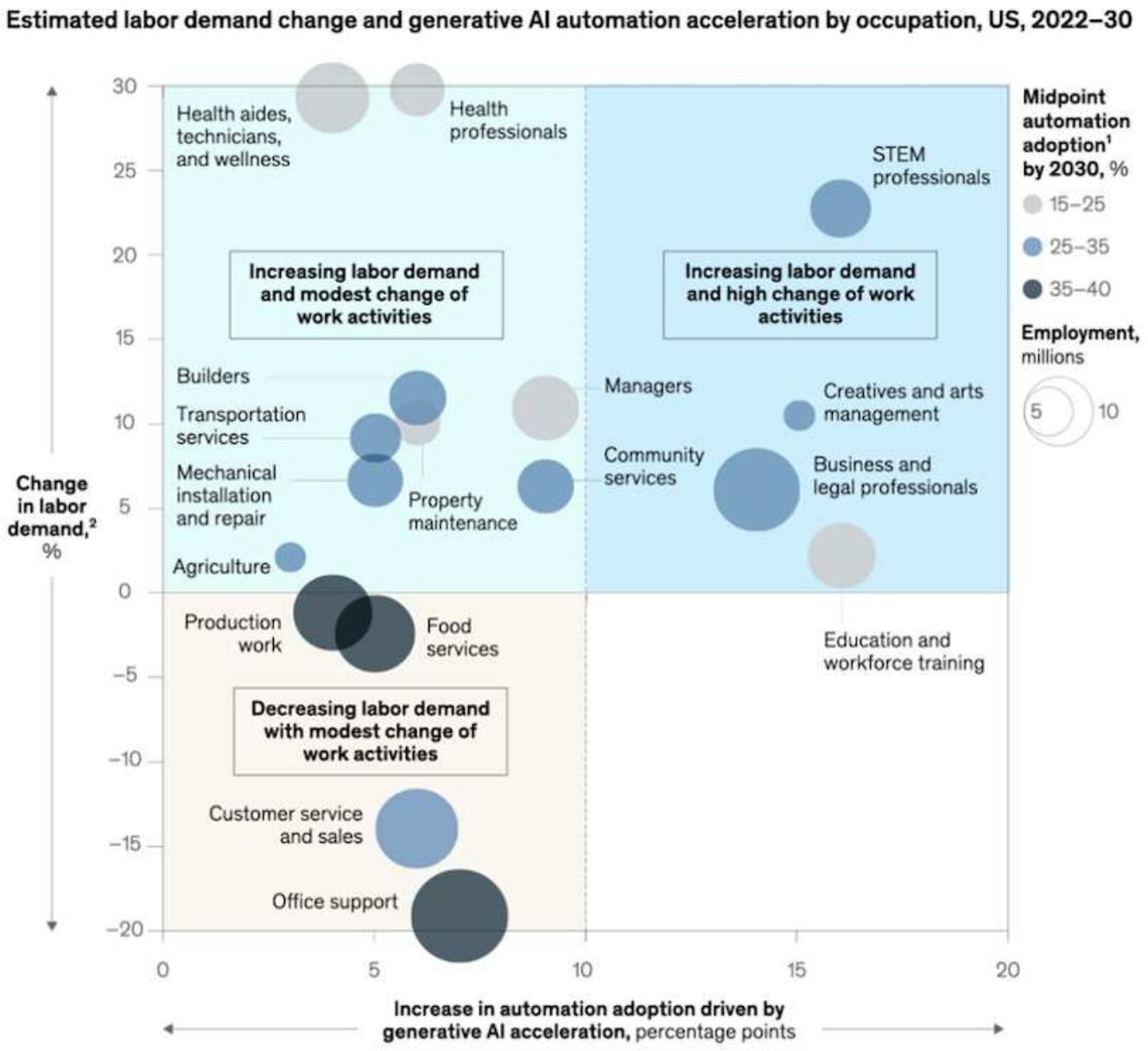
Las empresas de todos los sectores están descubriendo cada vez más potencial para que la IA extraiga y produzca nuevas formas de valor a partir de los datos existentes para operaciones empresariales, ventas, marketing y servicios legales. Según la inteligencia de mercado de IDC (International Data Corporation) sobre casos de uso e inversiones de IA generativa global, la gestión del conocimiento en el desarrollo de software y el diseño de productos será la más afectada, seguida de la creación de historias para el marketing y la generación de código para los desarrolladores. En el sector sanitario, las organizaciones de investigación clínica están abriendo nuevos caminos en la medicina. Los modelos preentrenados como ProteinBERT incorporan anotaciones de Ontología Genética (GO) para diseñar rápidamente estructuras de proteínas para medicamentos médicos, lo que representa un hito significativo en el descubrimiento de fármacos, la bioinformática y la biología molecular. Las empresas de biotecnología han iniciado ensayos en humanos para la medicina generalizada descubierta por IA, que tiene como objetivo tratar enfermedades como la fibrosis pulmonar (FPI), una enfermedad pulmonar que causa cicatrices irreversibles del tejido pulmonar.

Figura 1: Casos prácticos que impulsan la IA generativa

Chatbots	Drug discovery
Text generation	Genome model expression
Image generation	Classification
Code generation	Speech-to-Text

El aumento en la adopción de la automatización impulsado por la IA generativa también está cambiando el suministro y la demanda de las actividades del trabajo para muchas ocupaciones. Según McKinsey, el mercado laboral estadounidense (diagrama a continuación) ha pasado por una rápida transición, que solo puede continuar teniendo en cuenta el impacto de la IA.

Fuente: McKinsey & Company



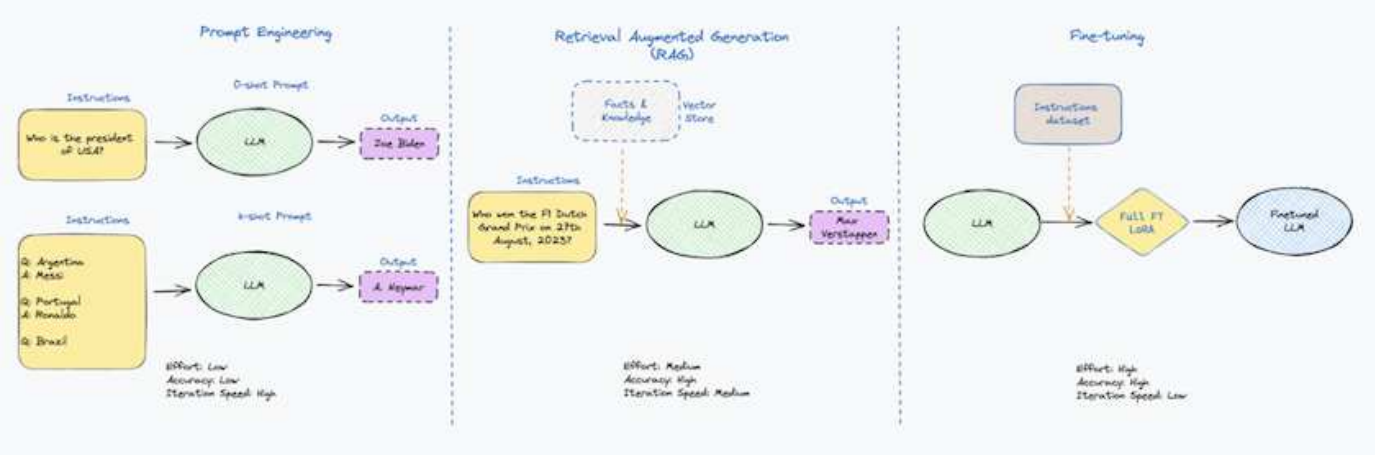
Función del almacenamiento en IA generativa

Las LLMs dependen en gran medida del aprendizaje profundo, las GPU y la computación. Sin embargo, cuando el búfer de la GPU se llena, los datos deben escribirse rápidamente en el almacenamiento. Mientras que algunos modelos de IA son lo suficientemente pequeños para ejecutarse en la memoria, los LLM requieren un alto volumen de IOPS y un almacenamiento de alto rendimiento para proporcionar un acceso rápido a grandes conjuntos de datos, especialmente si se trata de miles de millones de tokens o millones de imágenes. Para un requisito de memoria de GPU típico de un LLM, la memoria necesaria para entrenar un modelo con 1 millones de parámetros podría llegar a 80GB @32-bit de precisión completa. En ese caso, la llama 2 de Meta, una familia de LLMs que varían en escala de 7 mil millones a 70 mil millones de parámetros, puede requerir 70x80, aprox 5600GB o 5,6TB GB de RAM de GPU. Además, la cantidad de memoria que necesita es directamente proporcional al número máximo de tokens que desea generar. Por ejemplo, si desea generar salidas de hasta 512 tokens (alrededor de 380 palabras), necesita "512MB". Puede parecer intrascendente, pero, si quieres correr lotes más grandes, comienza a sumar. Por lo tanto, lo hace muy caro para las organizaciones que entrenan o ajustan los LLM en la memoria, lo que convierte al almacenamiento en la piedra angular para la IA generativa.

Tres enfoques principales para LLMs

Para la mayoría de las empresas, basado en las tendencias actuales, el enfoque para implementar LLMs se puede condensar en 3 escenarios básicos. Como se describe en un reciente "“Harvard Business Review”" artículo: (1) Capacitación (pre-entrenamiento) Un LLM desde cero – costoso y requiere habilidades expertas de IA/ML; (2) Ajuste de un modelo de base con datos empresariales – complejos, pero factibles; (3) Uso de la generación aumentada de recuperación (RAG) para consultar repositorios de documentos, API y bases de datos vectoriales que contienen datos de la empresa. Cada uno de ellos tiene desventajas entre el esfuerzo, la velocidad de iteración, la rentabilidad y la precisión de los modelos en sus implementaciones, que se utilizan para resolver diferentes tipos de problemas (diagrama a continuación).

Figura 3: Tipos de problemas



Modelos de base

Un modelo de base (FM) también conocido como modelo base es un modelo de IA de gran tamaño (LLM) entrenado sobre grandes cantidades de datos sin etiquetar, utilizando la autosupervisión a escala y, por lo general, adaptado a una gran variedad de tareas posteriores del PLN. Dado que los datos de entrenamiento no son etiquetados por los humanos, el modelo emerge en lugar de ser explícitamente codificado. Esto significa que el modelo puede generar historias o una narrativa propia sin estar explícitamente programado para hacerlo. Por lo tanto, una característica importante de FM es la homogeneización, lo que significa que el mismo método se utiliza en muchos dominios. Sin embargo, con las técnicas de personalización y ajuste, los FMS integrados en los productos que aparecen en estos días no solo son buenos para generar texto, texto a imágenes y texto a código, sino también para explicar tareas específicas de dominio o depurar código. Por ejemplo, FMS como el Codex de OpenAI o el Code Llama de Meta pueden generar código en varios lenguajes de programación basados en descripciones de lenguaje natural de una tarea de programación. Estos modelos son competentes en más de una docena de lenguajes de programación, incluyendo Python, C#, JavaScript, Perl, Ruby, y SQL. Entienden la intención del usuario y generan código específico que logra la tarea deseada útil para el desarrollo de software, la optimización de código y la automatización de tareas de programación.

Ajuste fino, especificidad de dominio y nueva formación

Una de las prácticas comunes con la implementación de LLM después de la preparación de datos y el procesamiento previo de los datos es seleccionar un modelo previamente entrenado que se haya entrenado en un conjunto de datos grande y diverso. En el contexto del ajuste fino, puede tratarse de un modelo de lenguaje grande de código abierto como "Meta’s Llama 2" entrenado en 70 000 millones de parámetros y 2 billones de tokens. Una vez seleccionado el modelo preentrenado, el siguiente paso es ajustarlo en los datos específicos del dominio. Esto implica ajustar los parámetros del modelo y entrenarlo en los nuevos datos para adaptarse a un dominio y tarea específicos. Por ejemplo, BloombergGPT, un LLM propietario entrenado en una amplia gama de datos financieros que sirven a la industria financiera. Los modelos específicos de dominio diseñados y entrenados para una tarea específica generalmente tienen mayor precisión y rendimiento dentro

de su alcance, pero baja transferibilidad entre otras tareas o dominios. Cuando el entorno empresarial y los datos cambian durante un período, la precisión de predicción del FM podría comenzar a disminuir en comparación con su rendimiento durante las pruebas. Esto ocurre cuando resulta crucial volver a entrenar o ajustar el modelo. El reentrenamiento de modelos en IA/ML tradicional hace referencia a actualizar un modelo DE ML puesto en marcha con nuevos datos, por lo general se realizan para eliminar dos tipos de derivas que ocurren. (1) Derivación del concepto: Cuando el vínculo entre las variables de entrada y las variables de destino cambia con el tiempo, ya que la descripción de lo que queremos predecir los cambios, el modelo puede producir predicciones inexactas. (2) Derivación de datos: Se produce cuando las características de los datos de entrada cambian, como los cambios en los hábitos o el comportamiento del cliente a lo largo del tiempo y, por lo tanto, la incapacidad del modelo para responder a dichos cambios. De manera similar, la recapacitación se aplica a FMS/LLMs, sin embargo, puede ser mucho más costosa (en millones de dólares), por lo tanto, no es algo que la mayoría de las organizaciones puedan considerar. Está bajo investigación activa, todavía emergiendo en el reino de LLMOps. Por lo tanto, en lugar de volver a entrenarse, cuando se produce la decadencia de modelos en FMS ajustados, las empresas pueden optar por el ajuste de nuevo (mucho más barato) con un conjunto de datos más nuevo. Para una perspectiva de costes, a continuación se muestra un ejemplo de una tabla de precios de modelo de Azure-OpenAI Services. Para cada categoría de tareas, los clientes pueden ajustar y evaluar modelos en conjuntos de datos específicos.

Fuente: 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

Ingeniería de pedidos e inferencia

La ingeniería rápida se refiere a los métodos efectivos de cómo comunicarse con LLMs para realizar las tareas deseadas sin actualizar los pesos del modelo. Tan importante como el entrenamiento y el ajuste de los modelos de IA es para las aplicaciones de PLN, la inferencia es igualmente importante, cuando los modelos entrenados responden a las indicaciones de los usuarios. Los requisitos del sistema de inferencia suelen estar mucho más en el rendimiento de lectura del sistema de almacenamiento de IA que alimenta datos desde los LLM a las GPU, ya que necesitan poder aplicar miles de millones de parámetros del modelo almacenados para producir la mejor respuesta.

LLMOps, Model Monitoring y Vectorstores

Al igual que las operaciones tradicionales de aprendizaje automático (MLOps), las operaciones de modelos de lenguaje grande (LLMOps) también requieren la colaboración de científicos de datos e ingenieros de DevOps con herramientas y prácticas recomendadas para la gestión de LLM en entornos de producción. Sin embargo, el flujo de trabajo y la pila de tecnología para LLMs podrían variar de alguna manera. Por ejemplo, LLM Pipelines creados usando marcos como LangChain unen varias llamadas de API LLM a puntos finales de incrustación externos, como vectorstores o bases de datos vectoriales. El uso de un punto final de incrustación y un almacén de vectores para conectores descendentes (como una base de datos vectorial)

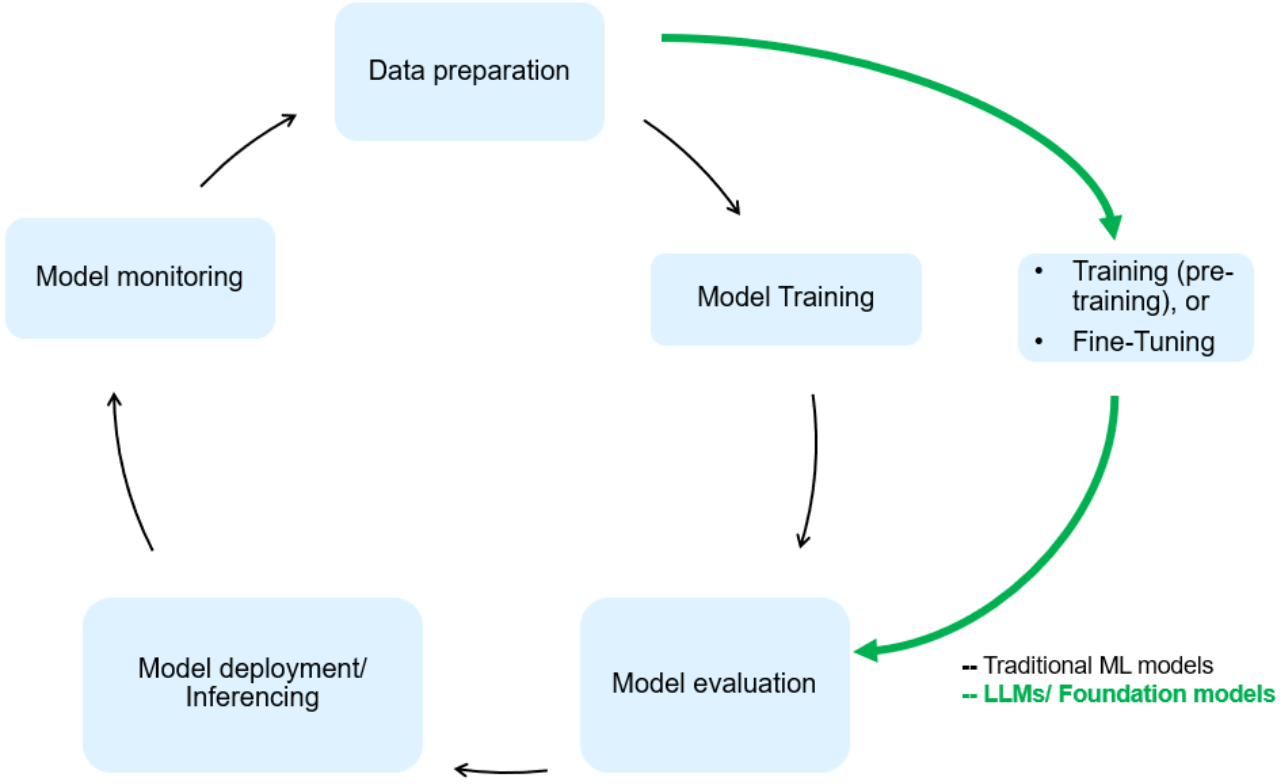
representa un desarrollo significativo en la forma en que se almacenan y se accede a los datos. A diferencia de los modelos tradicionales DE ML que se desarrollan desde cero, las LLM a menudo dependen del aprendizaje de transferencia, ya que estos modelos comienzan con FMS que están ajustados con nuevos datos para mejorar el rendimiento en un dominio más específico. Por lo tanto, es crucial que LLMOps ofrezca las capacidades de gestión de riesgos y monitoreo de deterioro de modelos.

Riesgos y Ética en la Era de la IA Generativa

"ChatGPT – Es elegante, pero todavía arroja tonterías."– MIT Tech Review. La basura en-basura, siempre ha sido el caso difícil con la informática. La única diferencia con la IA generativa es que destaca en hacer que la basura sea altamente creíble, lo que lleva a resultados imprecisos. Las LLMs son propensas a inventar hechos para adaptarse a la narrativa que está construyendo. Por lo tanto, las empresas que ven la IA generativa como una gran oportunidad para reducir sus costes con equivalentes de IA necesitan detectar de manera eficiente falsificaciones profundas, reducir los sesgos y reducir los riesgos para mantener los sistemas honestos y éticos. Una canalización de datos con flujo libre con una sólida infraestructura de IA que admita la movilidad de datos, la calidad de los datos, la gobernanza y la protección de datos a través de cifrado integral y barreras de IA es eminente en el diseño de modelos de IA generativos responsables y explicables.

Situación de cliente y NetApp

Figura 3: Aprendizaje automático/flujo de trabajo de modelos de lenguaje grande



*¿Estamos entrenando o ajustando? * La pregunta de si (a) entrenar un modelo LLM desde cero, ajustar un FM pre-entrenado, o usar RAG para recuperar datos de repositorios de documentos fuera de un modelo de base y aumentar las indicaciones, y (b) ya sea mediante el aprovechamiento de LLM de código abierto (por ejemplo, Llama 2) o FMS propietario (por ejemplo, ChatGPT, Bard, AWS Bedrock) es una decisión estratégica para las organizaciones. Cada enfoque tiene una compensación entre la rentabilidad, la gravedad de los datos, las operaciones, la precisión de los modelos y la gestión de LLMs.

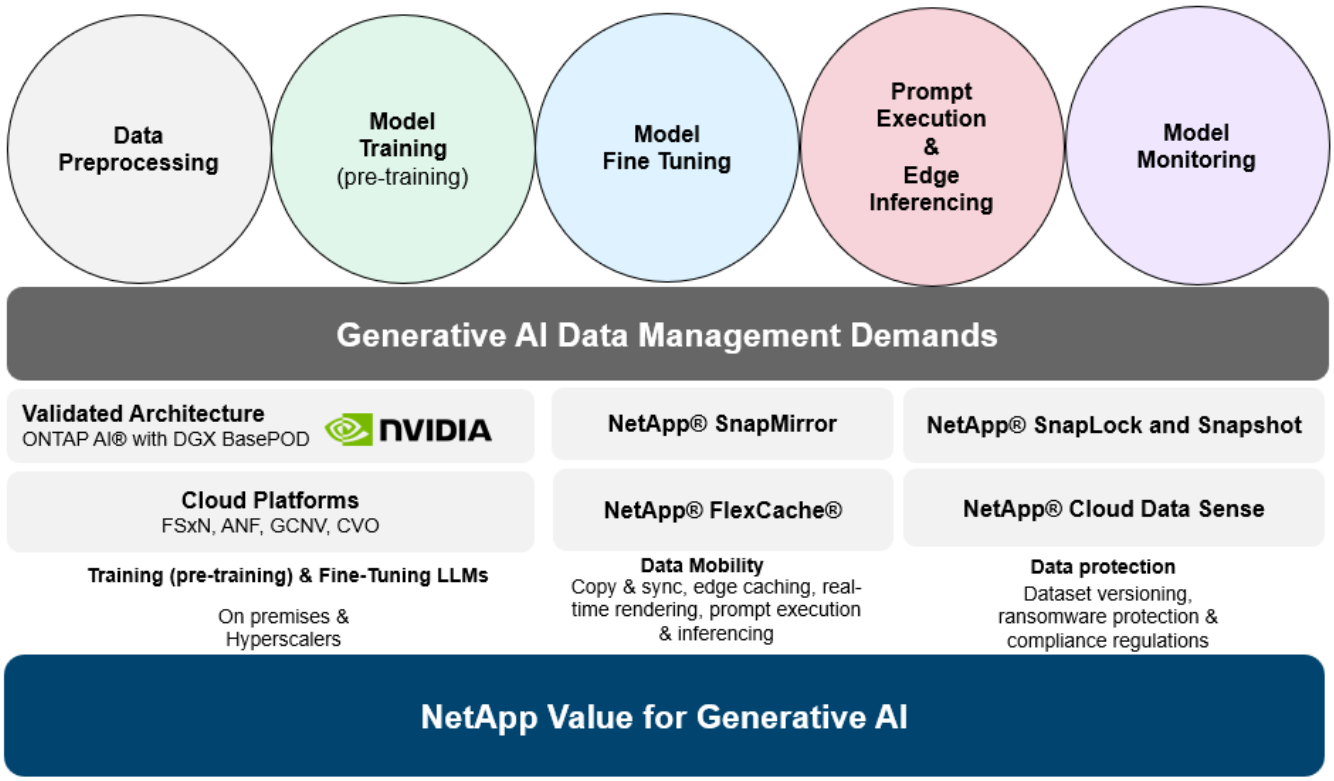
NetApp como empresa adopta la IA internamente en su cultura de trabajo y en su enfoque del diseño de

productos y los esfuerzos de ingeniería. Por ejemplo, la protección autónoma frente al ransomware de NetApp se crea con la IA y el aprendizaje automático. Proporciona la detección temprana de las anomalías en el sistema de archivos para ayudar a identificar las amenazas antes de que afecten a las operaciones. En segundo lugar, NetApp utiliza la IA predictiva para sus operaciones empresariales, como la previsión de ventas e inventario y los bots conversacionales, para ayudar a los clientes en los servicios de soporte de productos de centro de llamadas, especificaciones técnicas, garantía y manuales de servicio, etc. En tercer lugar, NetApp aporta valor al cliente a la canalización de datos de IA y al flujo de trabajo DE ML/LLM a través de productos y soluciones que sirven a los clientes que crean soluciones de IA predictivas como previsión de demanda, imágenes médicas, análisis de sentimientos, y soluciones de IA generativas como Gans para la detección de anomalías en imágenes industriales en el sector de fabricación y la detección de fraudes y lavado de dinero en servicios bancarios y financieros con productos y capacidades de NetApp como NetApp® ONTAP AI®, NetApp® SnapMirror® y NetApp® FlexCache®.

Funcionalidades de NetApp

El movimiento y la gestión de los datos en aplicaciones de IA generativa, como bot conversacional, generación de código, generación de imágenes o expresión de modelos genómicos, puede abarcar el perímetro, los centros de datos privados y el ecosistema multicloud híbrido. Por ejemplo, un robot de IA en tiempo real que ayuda a un pasajero a actualizar su boleto de avión a clase empresarial desde una aplicación de usuario final expuesta a través de API de modelos pre-entrenados como ChatGPT no puede lograr esa tarea por sí mismo, ya que la información del pasajero no está disponible públicamente en Internet. La API requiere acceso a la información personal del pasajero y a la información del boleto de la aerolínea que puede existir en un ecosistema híbrido o multicloud. Un escenario similar podría aplicarse a los científicos que comparten una molécula de fármaco y datos de pacientes a través de una aplicación de usuario final que utiliza LLMs para llevar a cabo ensayos clínicos en el descubrimiento de fármacos que involucran a instituciones de investigación biomédica de uno a varios. Los datos confidenciales que se pasan a FMS o LLMs pueden incluir información PII, información financiera, información de salud, datos biométricos, datos de ubicación, etc. datos de comunicaciones, comportamiento en línea e información legal. En tal caso de representación en tiempo real, ejecución rápida e inferencia periférica, hay movimiento de datos desde la aplicación de usuario final a extremos de almacenamiento a través de modelos LLM propios o de código abierto, hasta un centro de datos en las instalaciones o en plataformas de cloud público. En todos estos escenarios, la movilidad y la protección de datos son cruciales para las operaciones de IA que involucran LLM que dependen de grandes conjuntos de datos de entrenamiento y del movimiento de estos datos.

Figura 4: IA generativa: Canalización de datos de LLM



La cartera de infraestructura de almacenamiento, datos y servicios cloud de NetApp cuenta con la tecnología del software inteligente para la gestión de datos.

Preparación de datos: El primer pilar de la pila de tecnología LLM está en gran parte intacto de la pila de ML tradicional más antigua. El preprocesamiento de datos en la canalización de IA es necesario para normalizar y limpiar los datos antes del entrenamiento o ajuste. En este paso se incluyen conectores para ingerir datos dondequiera que estén en forma de un nivel Amazon S3 o en sistemas de almacenamiento en las instalaciones, como un almacén de archivos o de objetos como NetApp StorageGRID.

NetApp® ONTAP es la tecnología fundamental que sustenta las soluciones de almacenamiento críticas de NetApp en el centro de datos y la nube. ONTAP incluye varias funciones y funcionalidades de gestión y protección de datos, como la protección automática frente a ransomware contra ciberataques, funciones integradas de transporte de datos y funcionalidades de eficiencia del almacenamiento para una gama de arquitecturas en las instalaciones, híbridas y multiclouds en NAS, SAN, objetos, y las situaciones de almacenamiento definido por software (SDS) de las implementaciones de LLM.

NetApp® ONTAP AI® para el entrenamiento de modelos de aprendizaje profundo. NetApp® ONTAP® es compatible con NVIDIA GPU Direct Storage™ con el uso de NFS sobre RDMA para clientes de NetApp con el clúster de almacenamiento de ONTAP y nodos de computación NVIDIA DGX. Ofrece un rendimiento rentable para leer y procesar conjuntos de datos de origen desde el almacenamiento a la memoria numerosas veces para fomentar la inteligencia, lo que permite a las organizaciones formar, ajustar y escalar el acceso a las LLM.

NetApp® FlexCache® es una capacidad de almacenamiento en caché remoto que simplifica la distribución de archivos y almacena en caché solo los datos leídos activamente. Puede ser útil para la formación de LLM, reformaciones y ajustes precisos, lo que ofrece valor a los clientes con requisitos de negocio como la renderización en tiempo real e la inferencia de LLM.

NetApp® SnapMirror es una función de ONTAP que replica instantáneas de volumen entre dos sistemas ONTAP cualquiera. Esta función transfiere de forma óptima los datos que se encuentren en el perímetro a su centro de datos ubicado en las instalaciones o al cloud. SnapMirror puede utilizarse para mover datos de

forma segura y eficiente entre instalaciones y clouds de proveedores a hiperescala, cuando los clientes quieran desarrollar una IA generativa en clouds con RAG que contenga datos empresariales. Transfiere eficientemente solo los cambios, ahorrando ancho de banda y acelerando la replicación, ofreciendo así funciones esenciales de movilidad de datos durante las operaciones de entrenamiento, re-entrenamiento y ajuste de FMS o LLM.

NetApp® SnapLock aporta capacidad de disco inmutable en los sistemas de almacenamiento basados en ONTAP para el control de versiones de conjuntos de datos. La arquitectura microcore está diseñada para proteger los datos de los clientes con el motor FPolicy™ Zero Trust. NetApp garantiza que los datos de los clientes estén disponibles al resistir los ataques de denegación de servicio (DoS) cuando un atacante interactúa con un LLM de una forma que consume particularmente recursos.

NetApp® Cloud Data Sense ayuda a identificar, mapear y clasificar la información personal presente en los conjuntos de datos empresariales, promulgar políticas, cumplir con los requisitos de privacidad en las instalaciones o en la nube, ayudar a mejorar la postura de seguridad y cumplir con las regulaciones.

Clasificación NetApp® BlueXP™, con tecnología Cloud Data Sense. Los clientes pueden escanear, analizar, categorizar y actuar automáticamente sobre los datos en todo el conjunto de datos, detectar riesgos de seguridad, optimizar el almacenamiento y acelerar las puestas en marcha de cloud. Combina servicios de datos y almacenamiento a través de su plano de control unificado, los clientes pueden utilizar instancias de GPU para computación y entornos de multicloud híbridos para la organización en niveles de almacenamiento frío, así como archivos y backups.

NetApp dualidad de archivo-objeto. NetApp ONTAP permite el acceso de protocolo doble para NFS y S3. Con esta solución, los clientes pueden acceder a los datos NFS desde los portátiles de Amazon AWS SageMaker a través de bloques de S3 TB desde NetApp Cloud Volumes ONTAP. Esto ofrece flexibilidad a los clientes que necesitan un acceso fácil a fuentes de datos heterogéneas con la capacidad de compartir datos desde NFS y S3. Por ejemplo, afinar FMS como los modelos de generación de texto Llama 2 de Meta en SageMaker con acceso a cubos de objetos de archivo.

El servicio NetApp® Cloud Sync ofrece una forma sencilla y segura de migrar datos a cualquier destino, en la nube o en las instalaciones. Cloud Sync transfiere y sincroniza sin problemas los datos entre el almacenamiento en las instalaciones o en el cloud, NAS y almacenes de objetos.

NetApp XCP es un software cliente que permite migraciones de datos de cualquiera a NetApp y de NetApp a NetApp rápidas y fiables. XCP también proporciona la capacidad de mover datos masivos de forma eficiente desde sistema de archivos Hadoop HDFS a NFS de ONTAP, S3 o StorageGRID, y los análisis de archivos XCP proporcionan visibilidad sobre el sistema de archivos.

NetApp® DataOps Toolkit es una biblioteca de Python que facilita a los científicos de datos, DevOps e ingenieros de datos la realización de varias tareas de gestión de datos, como el aprovisionamiento, la clonación o la captura casi instantánea de un volumen de datos o un espacio de trabajo JupyterLab respaldado por un almacenamiento NetApp escalable de alto rendimiento.

Seguridad de los productos de NetApp. Los LLMs pueden revelar inadvertidamente datos confidenciales en sus respuestas, por lo tanto, una preocupación para los CISO que estudian las vulnerabilidades asociadas con las aplicaciones de IA que aprovechan los LLMs. Como se describe en OWASP (Open Worldwide Application Security Project), los problemas de seguridad como el envenenamiento de datos, la fuga de datos, la denegación de servicio y las inyecciones rápidas dentro de LLMs pueden afectar a las empresas de la exposición de los datos al acceso no autorizado que presta servicios a los atacantes. Los requisitos de almacenamiento de datos deben incluir comprobaciones de la integridad y snapshots inmutables de datos estructurados, semiestructurados y no estructurados. Las copias Snapshot de NetApp y SnapLock se utilizan para el control de versiones de conjuntos de datos. Proporciona un estricto control de acceso basado en roles (RBAC), así como protocolos seguros y cifrado estándar del sector para proteger los datos en reposo y en tránsito. Cloud Insights y Cloud Data Sense juntos ofrecen funcionalidades para ayudarte a identificar de

forma forense el origen de la amenaza y priorizar los datos que debes restaurar.

ONTAP AI con DGX BasePOD

La arquitectura de referencia de IA NetApp® ONTAP® con NVIDIA DGX BasePOD es una arquitectura escalable para cargas de trabajo de aprendizaje automático (ML) e inteligencia artificial (IA). Durante la crucial fase de formación de los LLM, suelen copiarse datos a intervalos periódicos del almacenamiento de datos al clúster de formación. Los servidores empleados en esta fase utilizan las GPU para paralelizar los cálculos, lo que crea un tremendo apetito de datos. Satisfacer las necesidades brutas de ancho de banda E/S es esencial para poder mantener un uso de GPU elevado.

ONTAP AI con NVIDIA AI Enterprise

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

1P plataformas en la nube

Las ofertas de almacenamiento en la nube totalmente gestionadas están disponibles de forma nativa en Microsoft Azure como Azure NetApp Files (ANF), en AWS como Amazon FSx para NetApp ONTAP (FSxN) y en Google como Google Cloud NetApp Volumes (GNCV). 1P es un sistema de archivos gestionado y de alto rendimiento que permite a los clientes ejecutar cargas de trabajo de IA de alta disponibilidad con seguridad de datos mejorada en nubes públicas, para ajustar LLMs/FMS con plataformas de ML nativas en la nube como AWS SageMaker, Azure-OpenAI Services y Vertex AI de Google.

Suite de soluciones para partners de NetApp

Además de sus funciones, tecnologías y productos de datos básicos, NetApp también colabora estrechamente con una sólida red de partners de IA con el fin de ofrecer valor añadido a los clientes.

Las Guardrails de NVIDIA en los sistemas de IA sirven como salvaguardas para garantizar el uso ético y responsable de las tecnologías de IA. Los desarrolladores de IA pueden elegir definir el comportamiento de las aplicaciones impulsadas por LLM en temas específicos e impedir que participen en discusiones sobre temas no deseados. Guardrails, un kit de herramientas de código abierto, proporciona la capacidad de conectar un LLM a otros servicios, de forma fluida y segura para crear sistemas conversacionales LLM confiables, seguros y seguros.

Domino Data Lab proporciona herramientas versátiles de nivel empresarial para crear y producir IA generativa: Rápida, segura y económica, dondequiera que se encuentre en su viaje a IA. Con la plataforma Enterprise MLOps de Domino, los científicos de datos pueden utilizar las herramientas preferidas y todos sus datos, entrenar e implementar modelos fácilmente en cualquier lugar y administrar riesgos y de manera rentable, todo desde un único centro de control.

Modzy para Edge AI. NetApp® y Modzy se han asociado para ofrecer IA a escala a cualquier tipo de datos, incluidas imágenes, audio, texto y tablas. Modzy es una plataforma MLOps para implementar, integrar y ejecutar modelos de IA, ofrece a los científicos de datos las capacidades de supervisión de modelos, detección de deriva y explicabilidad, con una solución integrada para una inferencia LLM perfecta.

Run:AI y NetApp se han asociado para demostrar las capacidades únicas de la solución de IA de NetApp ONTAP con la plataforma de gestión de clústeres Run:AI para simplificar la orquestación de cargas de trabajo de IA. Divide y une automáticamente los recursos de GPU, diseñados para escalar sus canalizaciones de

procesamiento de datos a cientos de máquinas con marcos de integración integrados para Spark, Ray, DASK y Rapids.

Conclusión

La IA generativa solo puede producir resultados efectivos cuando se entrena el modelo en remas de datos de calidad. Si bien las LLMs han logrado hitos notables, es fundamental reconocer sus limitaciones, desafíos de diseño y riesgos asociados con la movilidad de datos y la calidad de los datos. Las LLM dependen de conjuntos de datos de entrenamiento de gran tamaño y dispares procedentes de orígenes de datos heterogéneos. Los resultados inexactos o los resultados sesgados generados por los modelos pueden poner en peligro tanto a las empresas como a los consumidores. Estos riesgos pueden corresponder a las restricciones que surgen para las LLM potencialmente de los retos de gestión de datos asociados con la calidad, la seguridad de los datos y la movilidad de los mismos. NetApp ayuda a las organizaciones a cumplir las complejidades creadas por el rápido crecimiento de los datos, la movilidad de datos, la gestión multicloud y la adopción de IA. La infraestructura de IA a escala y la gestión de datos eficiente son cruciales para definir el éxito de las aplicaciones de IA como la IA generativa. Es crucial que los clientes cubran todas las situaciones de puesta en marcha sin comprometer su capacidad de expansión a medida que las empresas necesitan mantener el control de la rentabilidad, el gobierno de los datos y las prácticas éticas de IA. NetApp trabaja constantemente para ayudar a los clientes a simplificar y acelerar sus puestas en marcha de IA.

TR-4785: Puesta en marcha de IA con E-Series de NetApp y BeeGFS

Nagalakshmi Raju, Daniel Landas, Nathan Swartz, Amine Bennani, NetApp

Las aplicaciones de inteligencia artificial (IA), aprendizaje automático (ML, por sus siglas en inglés) y aprendizaje profundo (DL, por sus siglas en inglés) implican conjuntos de datos de gran tamaño y cálculos elevados. Para ejecutar estas cargas de trabajo con éxito, necesita una infraestructura ágil que le permita escalar horizontalmente tanto los nodos de almacenamiento como los de computación sin problemas. Este informe incluye los pasos para poner en marcha un modelo de entrenamiento de IA en un modo distribuido, que permite un escalado horizontal fluido de los nodos de computación y almacenamiento. El informe también incluye varios indicadores de rendimiento para mostrar cómo una solución, que combina el almacenamiento E-Series de NetApp con el sistema de archivos en paralelo BeeGFS, proporciona una solución flexible, rentable y sencilla para cargas de trabajo de IA.

"TR-4785: Puesta en marcha de IA con E-Series de NetApp y BeeGFS"

Diseño de NVA-1150: Quantum StorNext con la guía de diseño de sistemas E-Series de NetApp

Ryan Rodine, NetApp

Este documento proporciona información detallada sobre cómo diseñar una solución de sistema de archivos paralelo StorNext con los sistemas de almacenamiento E-Series de NetApp. Esta solución abarca la cabina all-flash EF280 de NetApp, la cabina NVMe all-flash EF300 de NetApp, la cabina NVMe all-flash EF600 y el sistema híbrido E5760 de

NetApp. Ofrece una caracterización del rendimiento basada en las pruebas comparativas de Frametest, una herramienta que se utiliza ampliamente para realizar pruebas en el sector del entretenimiento y los medios de comunicación.

"Diseño de NVA-1150: Quantum StorNext con la guía de diseño de sistemas E-Series de NetApp"

NVA-1150-PUESTA en MARCHA: Quantum StorNext con la guía de puesta en marcha de sistemas E-Series de NetApp

Ryan Rodine, NetApp

Este documento proporciona detalles sobre cómo poner en marcha una solución de sistema de archivos paralelo StorNext con sistemas de almacenamiento E-Series de NetApp. Esta solución abarca la cabina all-flash EF280 de NetApp, la cabina NVMe all-flash EF300 de NetApp, la cabina NVMe all-flash EF600 de NetApp y el sistema híbrido E5760 de NetApp. Ofrece una caracterización del rendimiento basada en las pruebas comparativas de Frametest, una herramienta que se utiliza ampliamente para realizar pruebas en el sector del entretenimiento y los medios de comunicación.

"NVA-1150-PUESTA en MARCHA: Quantum StorNext con la guía de puesta en marcha de sistemas E-Series de NetApp"