

Formación de Azure distribuida: Detección de carriles

NetApp Solutions

NetApp April 25, 2024

This PDF was generated from https://docs.netapp.com/es-es/netapp-solutions/ai/runai-ld_solution_overview.html on April 25, 2024. Always check docs.netapp.com for the latest.

Tabla de contenidos

TI	R-4896: Formación distribuida en Azure: Detección de carriles - diseño de soluciones	. 1
	Descripción general de la solución	. 1
	Detección de carriles: Formación distribuida con RUN:Al	. 4
	Conclusión	29
	Información adicional	29

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

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

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.

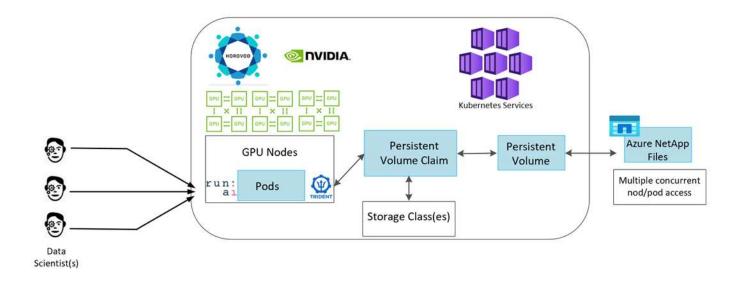
Tecnología de soluciones

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:Al

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: Al para orquestar la asignación y la gestión de recursos. RUN: Al 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:

6. Introduzca este comando en la CLI de Azure:

kubectl get nodes



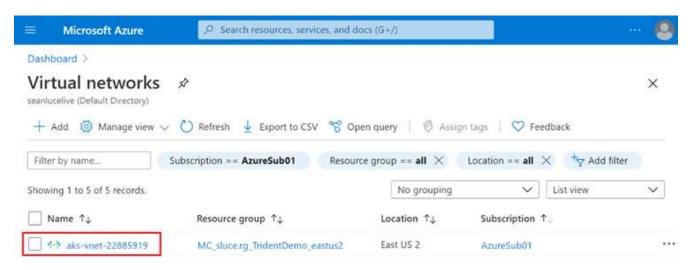
Si los seis nodos están en funcionamiento como se ve aquí, su clúster de AKS estará listo y conectado a su entorno local.

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 ~ %				

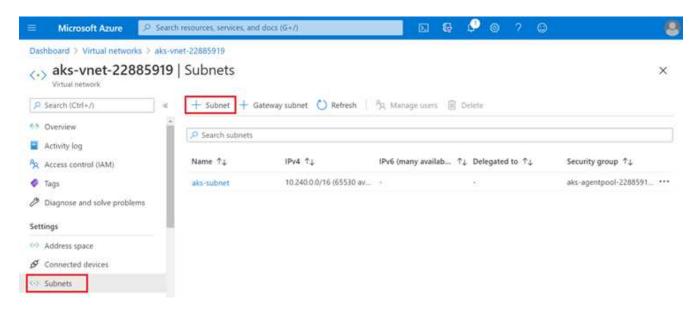
Cree una subred delegada para Azure NetApp Files

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.

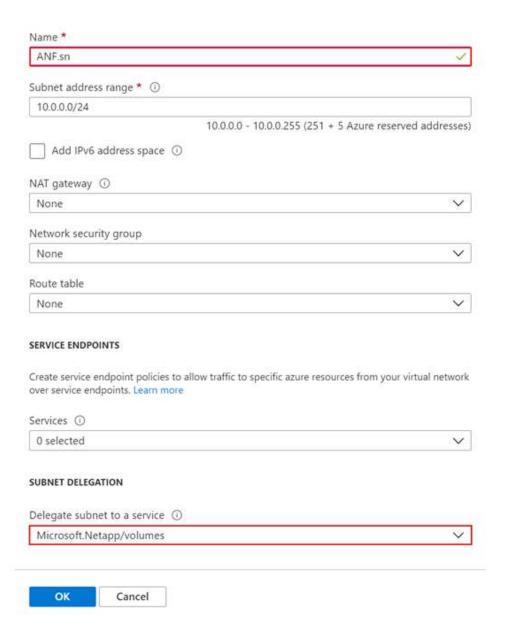


2. Haga clic en subredes y seleccione +Subnet en la barra de herramientas superior.



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.

Add subnet ×



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

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:

- 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. j son 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:

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

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:

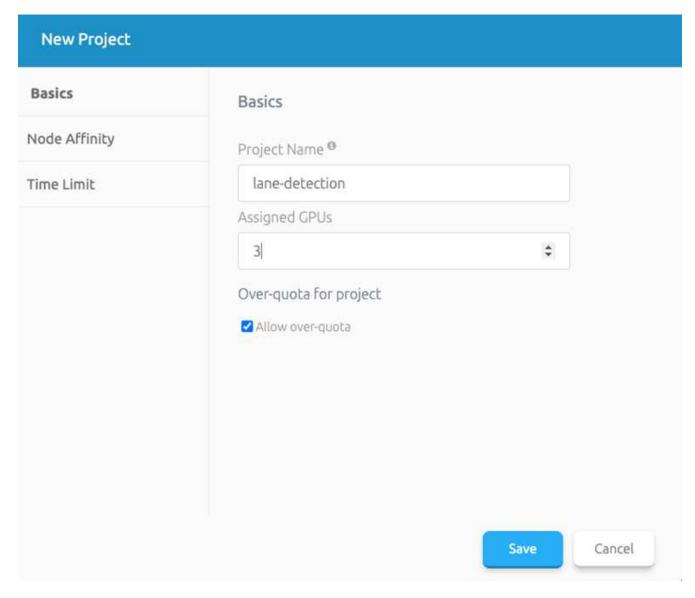
```
kubectl get volumesnapshotclass
```

El resultado debería ser como el ejemplo siguiente:

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, lanedetection) por lo tanto, la cuenta de servicio predeterminada que pertenece a runai-lanedetection 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:Al 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: Al 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: Al 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:Al:

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 projec	t 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
				6% 24.3M 20m36s
751600K	 	. <u>.</u>	 	6% 26.3M 20m36s

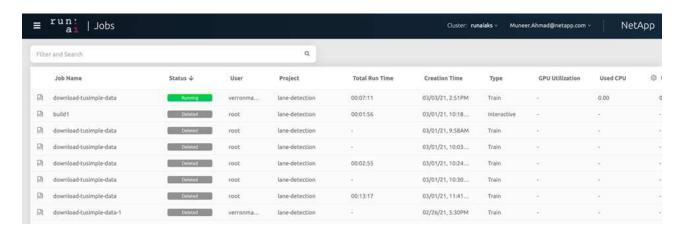
e. Enumere la pvc creado. Utilice esto pvc comando para la formación en el siguiente paso.

```
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: Al 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: Al 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
```

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):

pvc-download-tusimple-data-0 Bound pvc-bb03b74d-2c17-40c4-a445-79f3de8d16d5 100Gi RWX

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

d. Envíe el TRABAJO RUN: Al 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 estab	lecer 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

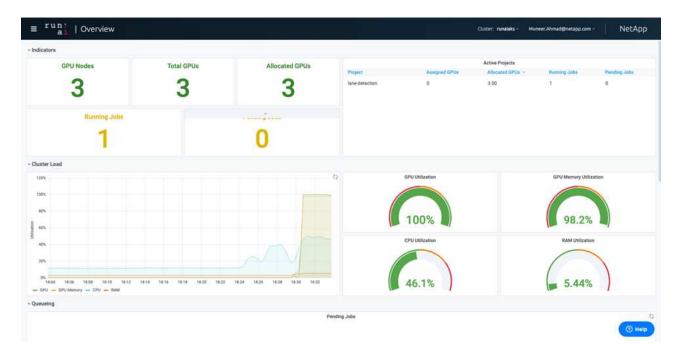
NAME STATUS AGE NODE IMAGE TYPE PROJECT USER GPUs Allocated (Requested) PODS
SERVICE URL(S)
download-tusimple-data Succeeded 1d muneer7589/download-tusimple:1.0 Train lane-detection verronmartina - (0) 0 (0)
dist-lane-detection-training Init:0/1 2m <multiple> muneer7589/dist-lane-detection:3.1 Train lane-detection root 3 (3) 4 (0)
```

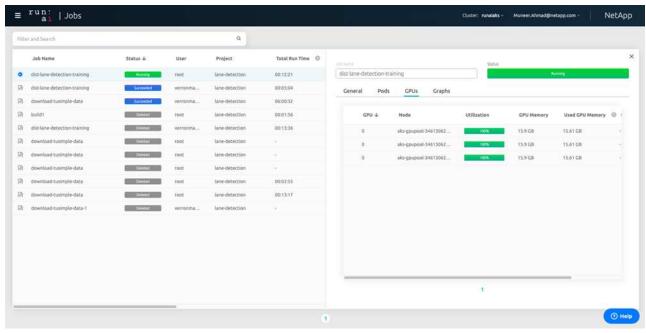
f. Registros de trabajos enviados:

```
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]style="style="color: blue;">style="style="style="color: blue;">style="st

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

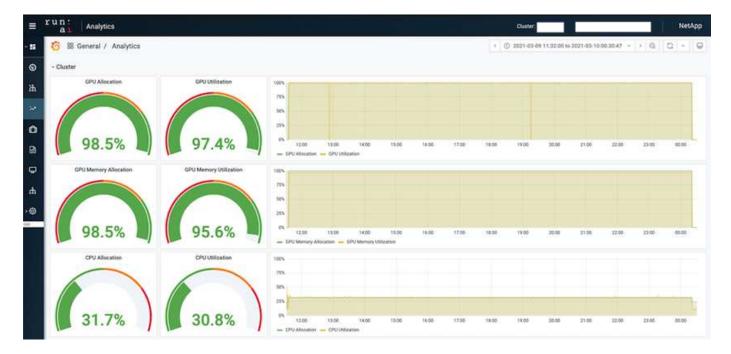
Para mostrar la escalabilidad lineal de la solución, se han realizado pruebas de rendimiento para dos supuestos: Una GPU y tres GPU. La asignación de GPU, la utilización de la GPU y la memoria, se han capturado diferentes métricas de uno y tres nodos durante el entrenamiento en el conjunto de datos de 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".

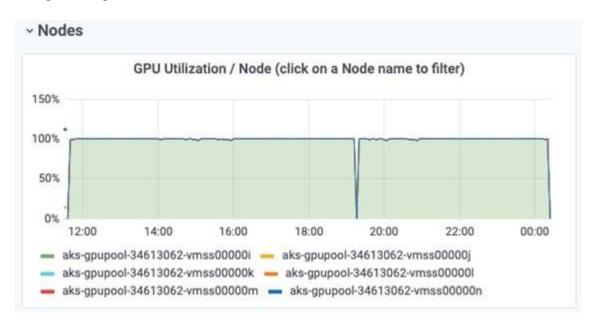
El tiempo de procesamiento en una GPU era de 12 horas y 45 minutos. El tiempo de procesamiento en tres GPU en tres nodos era de aproximadamente 4 horas y 30 minutos.

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.

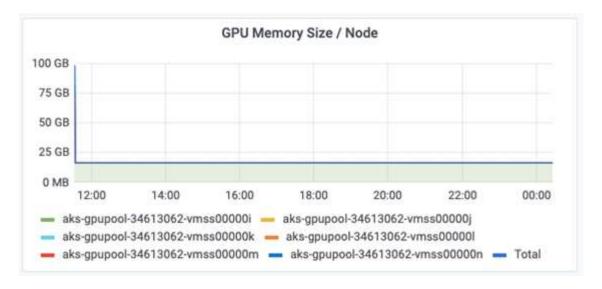
La siguiente figura muestra la asignación de 1 GPU y la utilización de memoria.



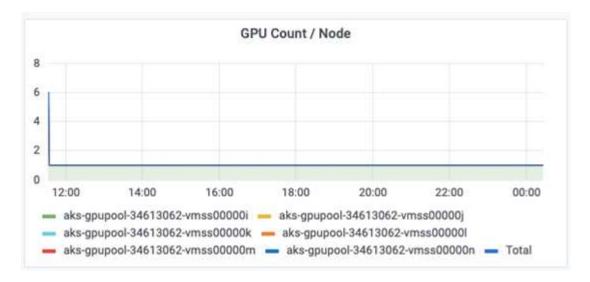
La siguiente figura ilustra el uso de GPU de un solo nodo.



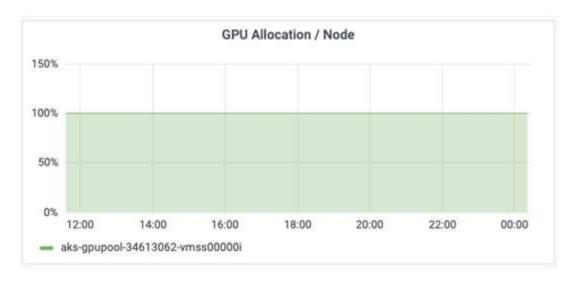
La siguiente figura ilustra el tamaño de la memoria de un solo nodo (16 GB).



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



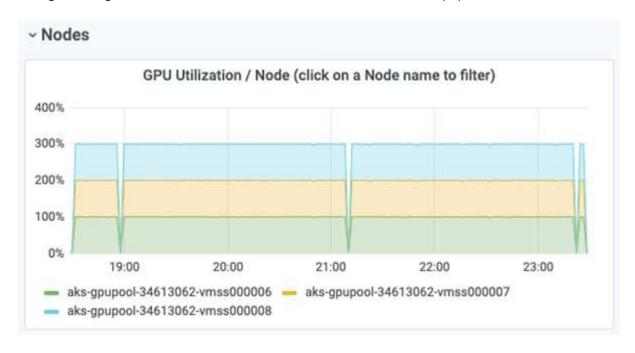
La siguiente figura muestra la asignación de GPU de un solo nodo (%).



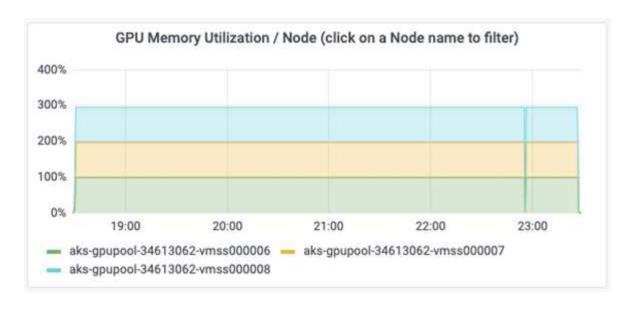
La siguiente figura muestra tres GPU en tres nodos: Asignación de GPU y memoria.



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 (%).



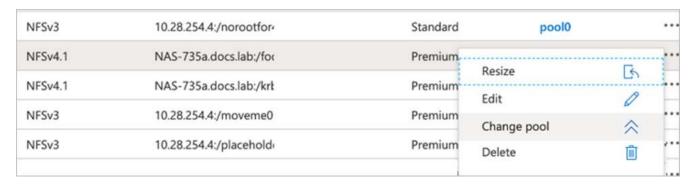
Niveles de servicio de Azure NetApp Files

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.

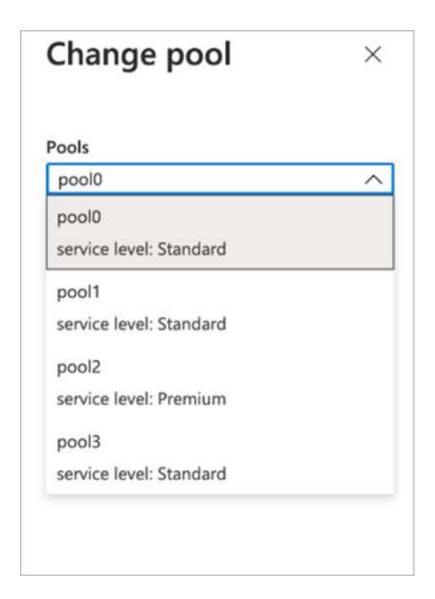
Cambie dinámicamente el nivel de servicio de un volumen

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 lA 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/"

Información de copyright

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

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

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

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

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

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

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

Información de la marca comercial

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