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

NetApp Solutions

NetApp
April 18, 2024

This PDF was generated from https://docs.netapp.com/es-es/netapp-solutions/ai/ai-protopia-overview.html on April 18, 2024. Always check docs.netapp.com for the latest.

Tabla de contenidos

IA responsable e inferencia confidencial: IA de NetApp con transformación de imagen Protopía . . . 1
TR-4928: IA responsable e inferencia confidencial - NetApp AI con Protopía Image y transformación de datos . . . 1
Áreas de soluciones . . . 3
Información general de la tecnología . . . 5
Plan de pruebas y validación . . . 9
Configuración de prueba . . . 9
Procedimiento de prueba . . . 10
Comparación de precisión de inferencia . . . 24
Velocidad de ofuscación . . . 25
Conclusión . . . 25
Dónde encontrar información adicional y reconocimientos . . . 26

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.

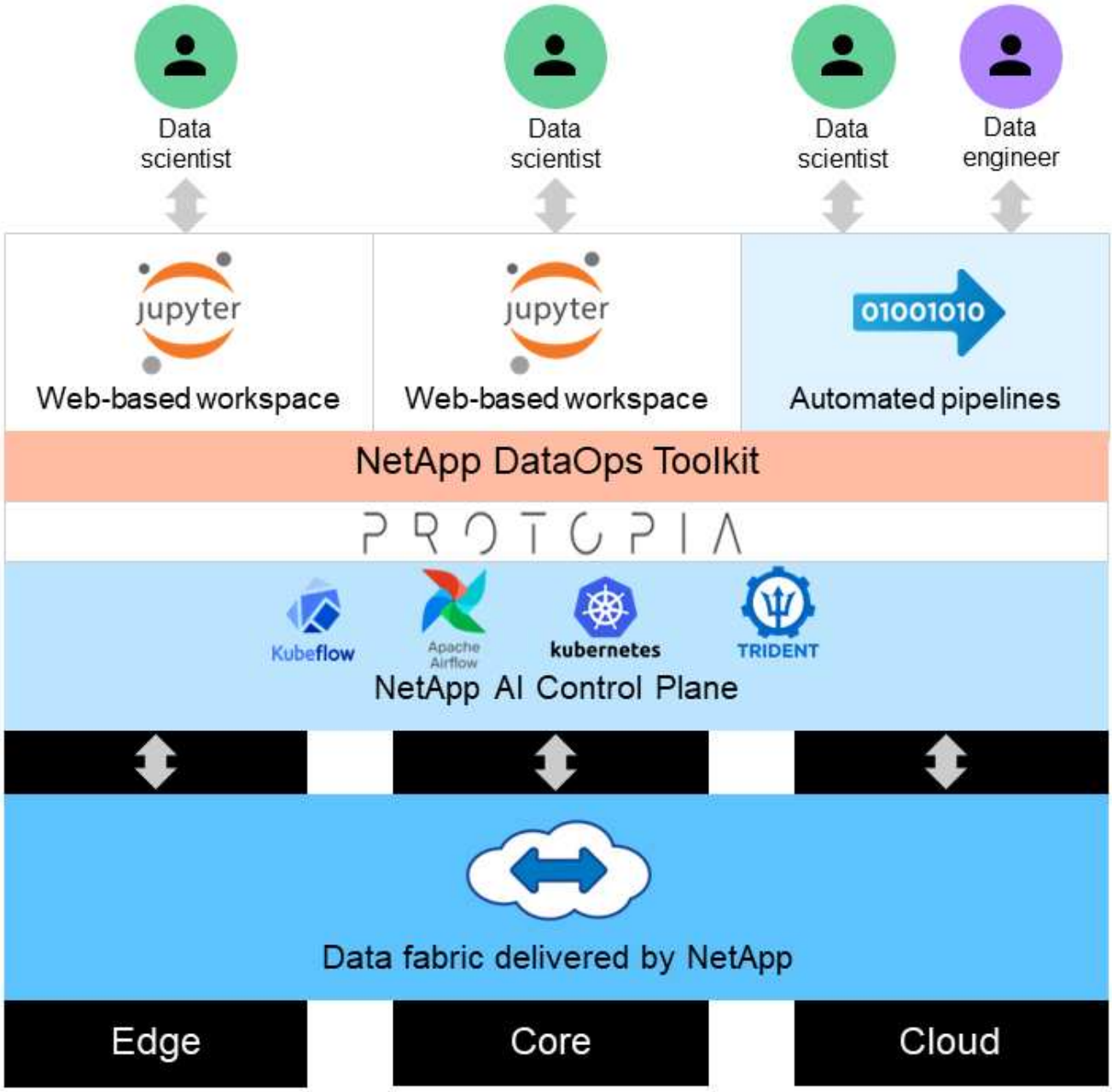
Público objetivo

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

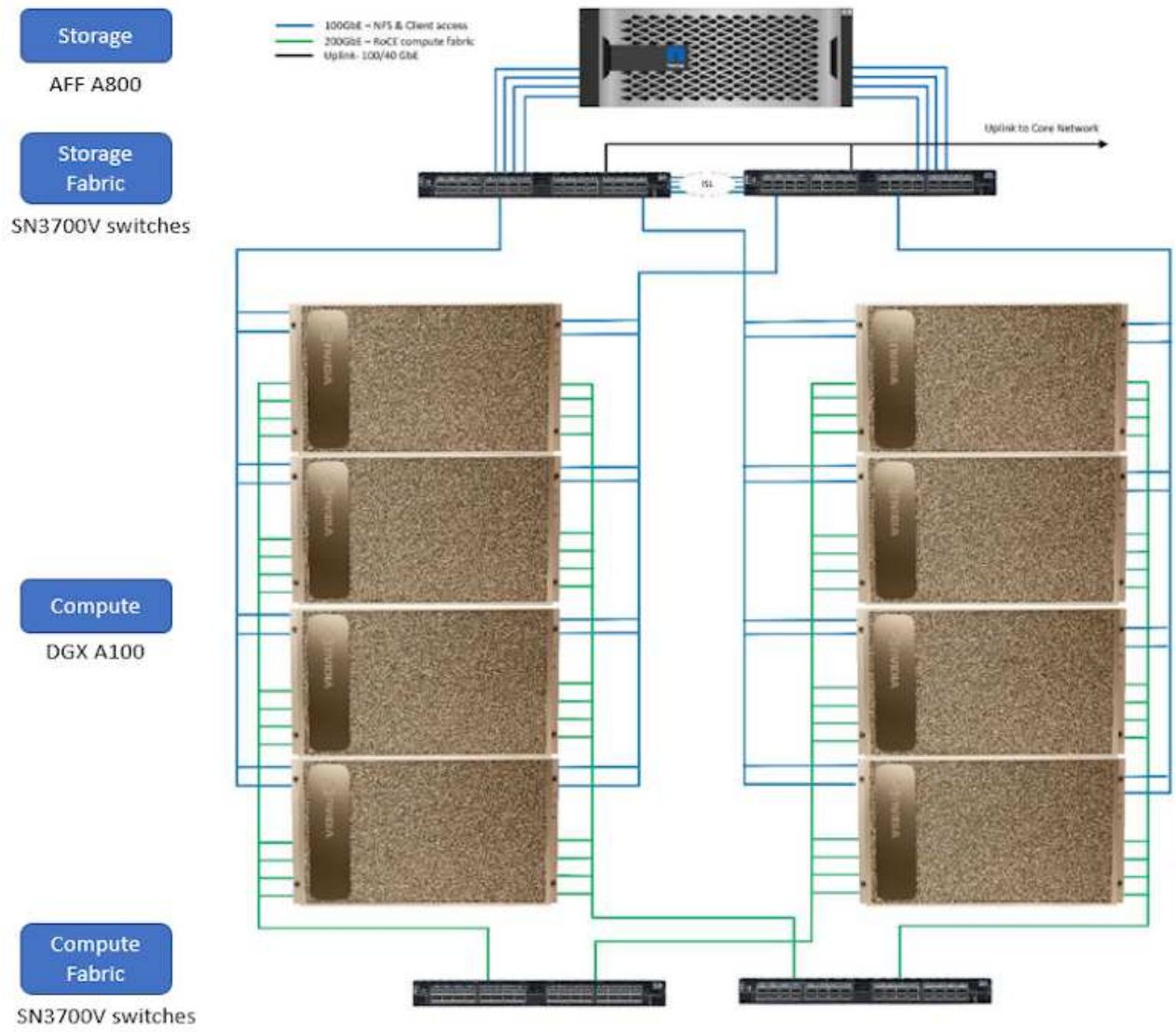
La arquitectura de referencia de ONTAP AI de NetApp, con sistemas DGX A100 y sistemas de almacenamiento conectados al cloud de NetApp, ha sido desarrollada y verificada por NetApp y NVIDIA. Proporciona a las organizaciones DE TI una arquitectura que ofrece las siguientes ventajas:

- 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.

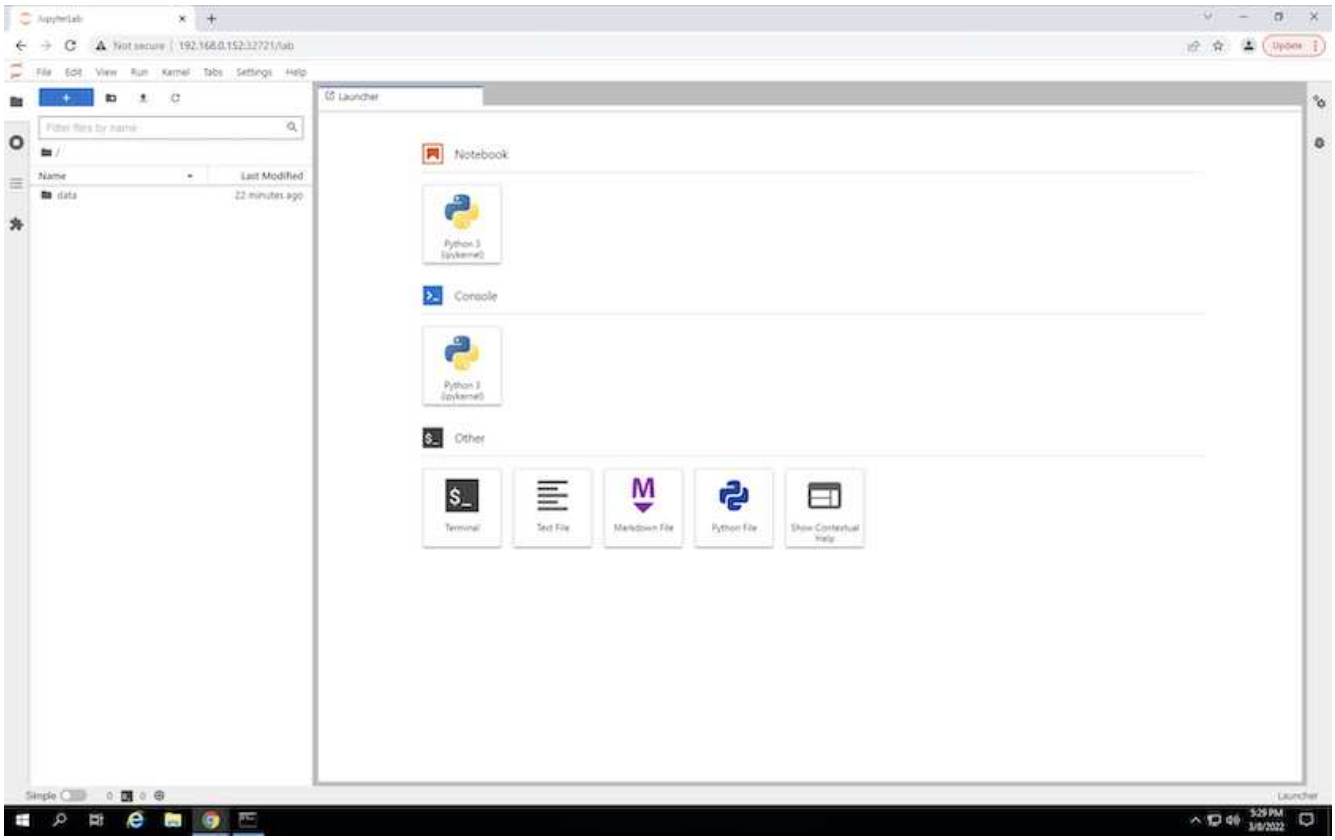
```
$ netapp_dataops_k8s_cli.py create volume --namespace=inference --pvc
-name=inference-data --size=50Gi
Creating PersistentVolumeClaim (PVC) 'inference-data' in namespace
'inference'.
PersistentVolumeClaim (PVC) 'inference-data' created. Waiting for
Kubernetes to bind volume to PVC.
Volume successfully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

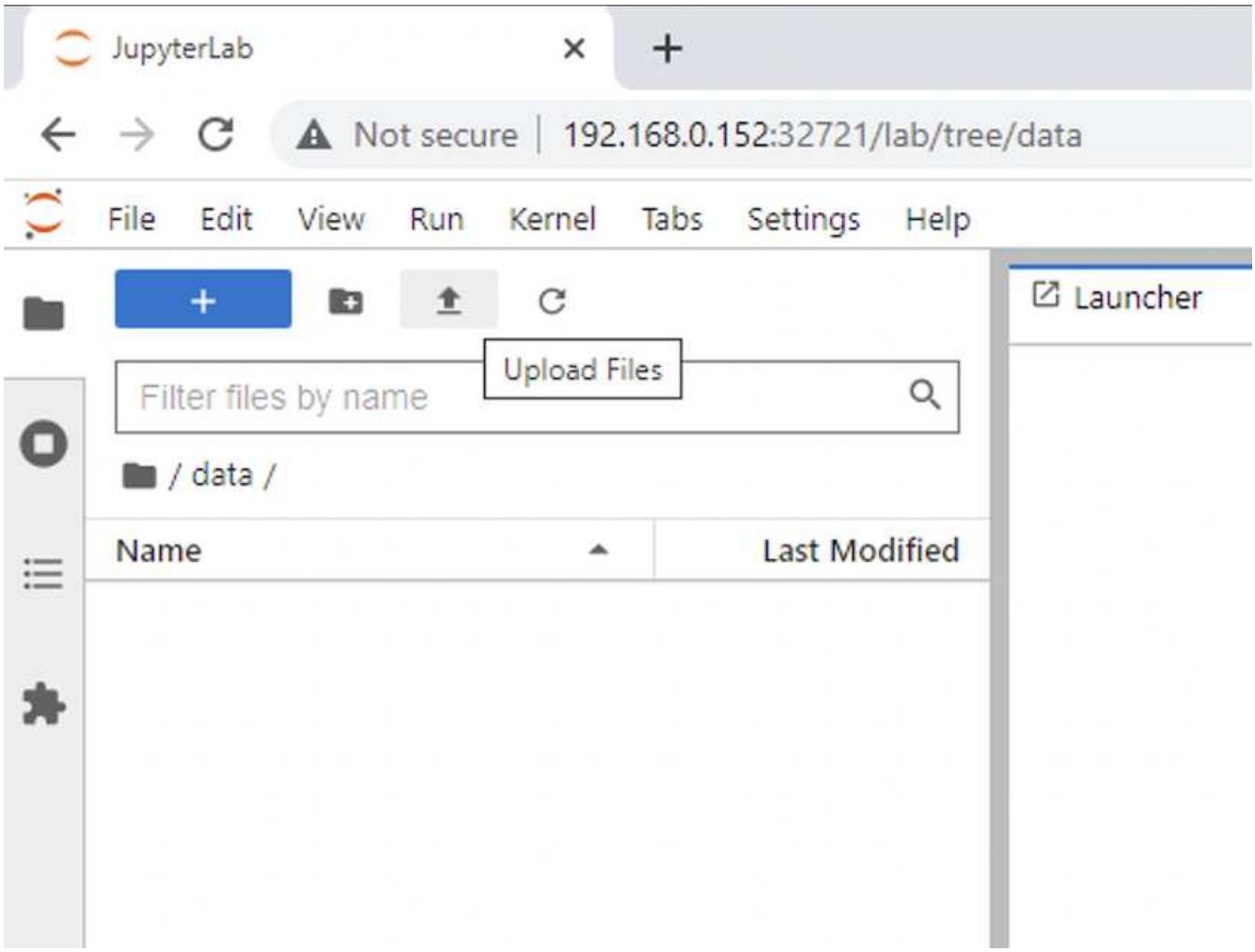
   En el siguiente ejemplo, el volumen persistente `inference-data` Está montado en el contenedor de espacio de trabajo JJupyterLab en `/home/jovyan/data`. Cuando utilice las imágenes del contenedor de Jupyter del proyecto oficial, `/home/jovyan` Se presenta como el directorio de nivel superior dentro de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

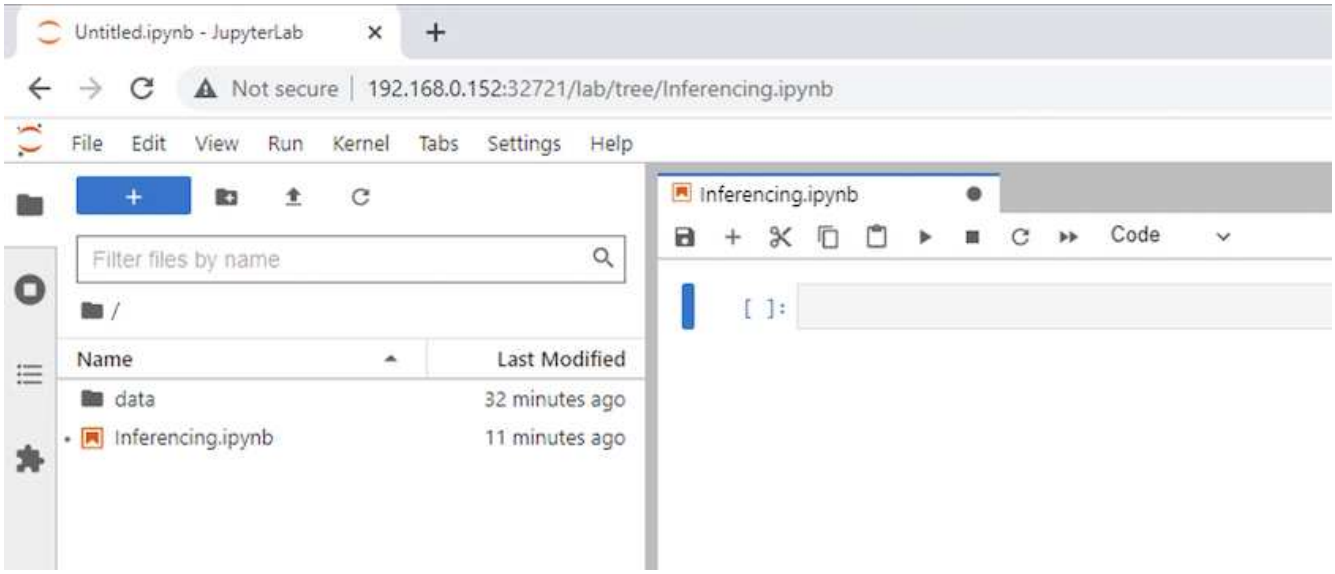
4. Acceda al espacio de trabajo JupyterLab utilizando la dirección URL especificada en la salida del `create jupyterlab` comando. El directorio de datos representa el volumen persistente que se montó en el espacio de trabajo.



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

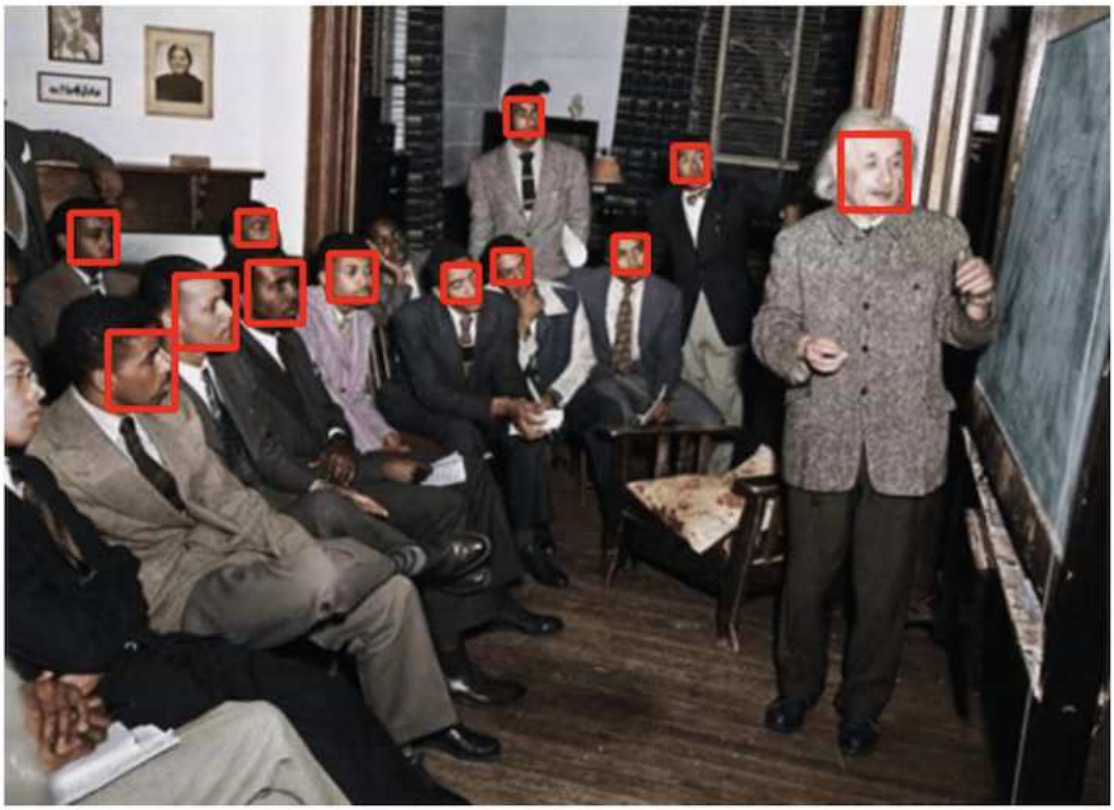
[9]:
```
# get current frame
frame = input_image

# 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

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.