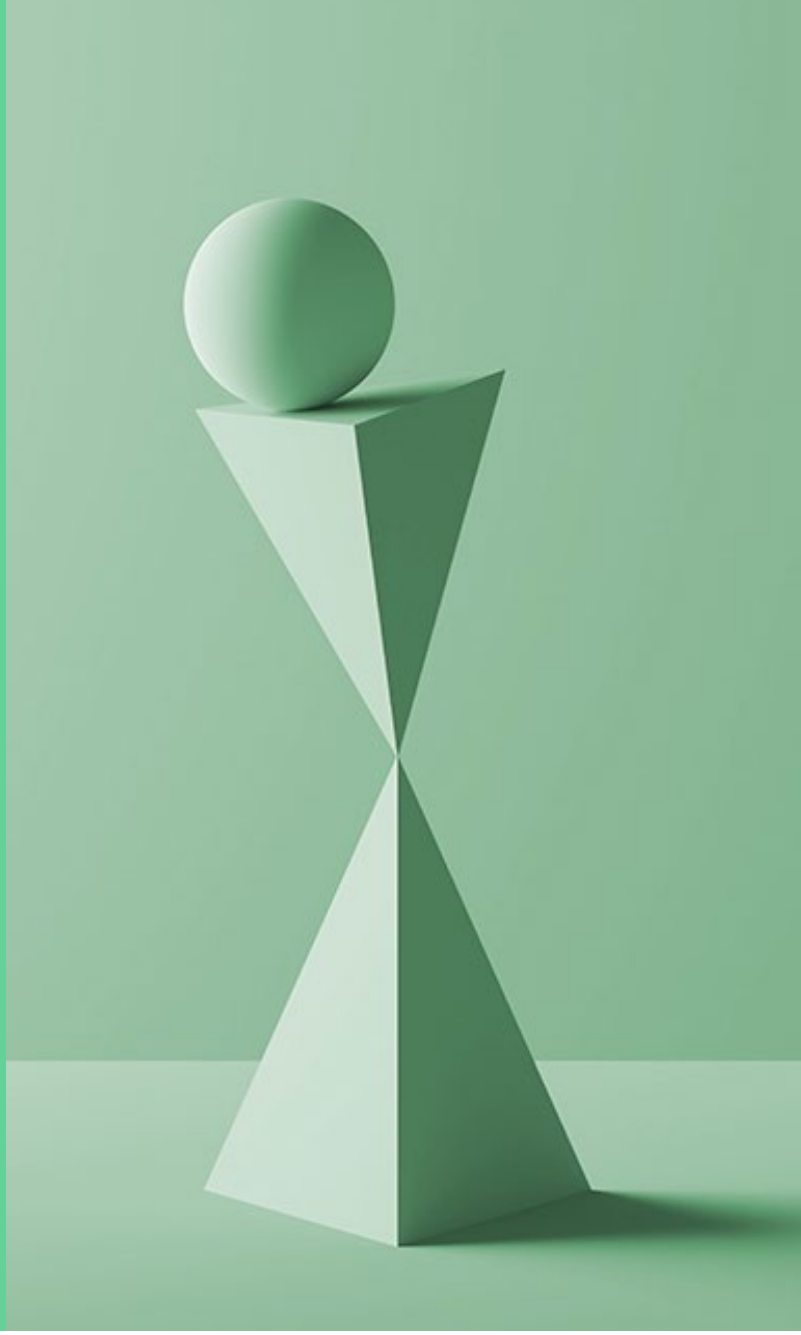
RESUMEN DE LA SOLUCIÓN

ONTAP AI

Simplifique, acelere e integre su canalización de datos de ML y DL con NetApp y NVIDIA



NetApp | NVIDIA

Retos de la infraestructura de IA

La inteligencia artificial (IA), el aprendizaje automático (ML) y el aprendizaje profundo (DL) permiten a las empresas detectar el fraude, mejorar las relaciones con los clientes, optimizar la cadena de suministros y ofrecer innovadores servicios y productos en un mercado cada vez más competitivo. La suya puede ser una de las muchas organizaciones que utilicen los nuevos métodos de IA para guiar la transformación digital y adquirir una ventaja competitiva. Para obtener el máximo beneficio de la IA, primero debe abordar varios retos clave.

Las integraciones que corren a cargo del propio usuario suelen ser complejas. Montar e integrar componentes genéricos de software, redes, almacenamiento y computación destinados al aprendizaje profundo y al aprendizaje automático puede aumentar aún más la complejidad y prolongar los tiempos de puesta en marcha. Como resultado, se acaban desperdiciando valiosos recursos de datos en este proceso de integración de sistemas.

Alcanzar un rendimiento previsible y escalable no es fácil. En las prácticas recomendadas de aprendizaje profundo, se sugiere que las organizaciones comiencen con los recursos justos y los vayan ampliando a medida que lo necesiten. Tradicionalmente, se han empleado el almacenamiento DAS y la computación para introducir datos en los flujos de trabajo de IA. Sin embargo, el escalado con almacenamiento tradicional puede provocar interrupciones y tiempos de inactividad en las operaciones que estén en curso en ese momento.

Las interrupciones afectan a la productividad de los científicos de datos. La infraestructura de ML y DL consta de múltiples interdependencias entre el hardware y el software. Para mantener en funcionamiento una infraestructura, hace falta contar con una gran experiencia en IA para toda la pila. Los tiempos de inactividad o una ralentización del rendimiento de la IA pueden desencadenar una serie de reacciones en cadena que reduciría la productividad de los desarrolladores y provocaría un descontrol en los gastos operativos.

La solución

Ahora, puede cumplir totalmente la promesa de IA, ML y DL. Simplifique, acelere e integre su canalización de datos con la arquitectura demostrada de ONTAP® AI de NetApp®, que está impulsada por los sistemas NVIDIA DGX™ y el almacenamiento all-flash conectado a cloud de NetApp. Optimice el flujo de datos con total confianza y acelere los análisis, la formación y la inferencia con Data Fabric, desde el perímetro al núcleo y a cloud.

Ventajas clave

Reducir el riesgo con soluciones validadas y flexibles

- Consiga que todo funcione con mayor rapidez eliminando conjeturas y complejidades del diseño.
- Optimice la configuración y la puesta en marcha con soluciones preconfiguradas disponibles.

Ofrecer el rendimiento y la escalabilidad adecuados

- Empiece con lo justo y crezca sin interrupciones.
- Acelere los resultados con una solución de alto rendimiento.

Crear una canalización de datos integrada

- Gestione con inteligencia sus datos con una canalización integrada que abarque desde el perímetro hasta el núcleo y cloud.
- Ponga en marcha una solución que está respaldada con experiencia en IA y opciones de soporte sencillas.

Unificar cargas de trabajo de IA

- Elimine silos de infraestructura.
- Responda con flexibilidad a las demandas del negocio.

ONTAP AI de NetApp es una de las primeras pilas de infraestructura convergente en incorporar el sistema NVIDIA DGX A100, el primer sistema de IA de 5 petaflops del mundo, y los switches Ethernet de alto rendimiento de NVIDIA Mellanox. Obtiene unas cargas de trabajo de IA unificadas, una puesta en marcha simplificada y un rápido retorno de la inversión.

«El aprendizaje profundo está revolucionando prácticamente todos los mercados en los que trabajamos. Lo aplicamos en diversos mercados, lo que impulsa el arte de lo posible. ONTAP AI de NetApp, impulsado por los sistemas NVIDIA DGX y el almacenamiento all-flash de NetApp, está simplificando y acelerando la canalización de datos para el aprendizaje profundo».

Tim Ensor, director de Inteligencia artificial Cambridge Consultants

Figura 1) Arquitecturas de ONTAP AI con DGX A100; configuraciones de dos, cuatro y ocho nodos.

Reducir el riesgo con soluciones validadas y flexibles
La innovación en inteligencia artificial se desarrolla a un ritmo vertiginoso, de manera que diseñar infraestructuras de IA eficaces se convierte en todo un desafío. Con ONTAP AI, podrá olvidarse de pruebas y conjeturas y empezará a trabajar antes con una arquitectura de referencia demostrada en el sector. Del mismo modo, si elige una solución integrada preconfigurada que sea fácil de adquirir y poner en marcha, eliminará la complejidad del diseño y la gestión.

La solución integrada ONTAP AI está disponible en cuatro opciones preconfiguradas con posibilidad de ampliar la capacidad y un software avanzado opcional. Esta solución integrada también reduce la complejidad al incluir instalación *in situ* y soporte completo con un único número al que llamar para solucionar problemas que van de la notificación de incidentes a su resolución.

Ofrecer el rendimiento y la escalabilidad adecuados
Las rutinas de formación del aprendizaje profundo demandan cantidades ingentes de capacidad informática. Al simplificar la formación en imágenes, se reducen los costes totales de computación a la vez que se aceleran la productividad y la innovación de IA.

Creado con la nueva arquitectura Ampere de NVIDIA, el sistema DGX A100 ofrece un rendimiento de formación hasta seis veces superior al de la generación anterior. Obtiene unos resultados equivalentes a los de un centro de datos de infraestructura de computación para análisis, formación e inferencia, ahora consolidado en un único sistema. El sistema DGX A100 requiere la vigésima quinta parte del espacio y la vigésima parte de la potencia de los sistemas de CPU por una décima parte del coste.

La inversión en tecnología de vanguardia requiere de un almacenamiento igualmente innovador que pueda gestionar miles de imágenes de formación por segundo. Necesita una solución de servicios de datos de alto rendimiento que se adapte a las cargas de trabajo de formación en aprendizaje profundo más exigentes.

Con el almacenamiento all-flash de NetApp, puede esperar obtener más de 2 GB/s de rendimiento sostenido (5 GB/s en el punto álgido). Además, hay menos de 1 milisegundo de latencia y la GPU opera a más del 95 % de la utilización. Un solo sistema AFF A800 de NetApp admite un rendimiento de 25 GB/s en lecturas secuenciales y una tasa de 1 millón de IOPS en pequeñas lecturas aleatorias, con latencias de menos de 500 microsegundos para cargas de trabajo NAS.

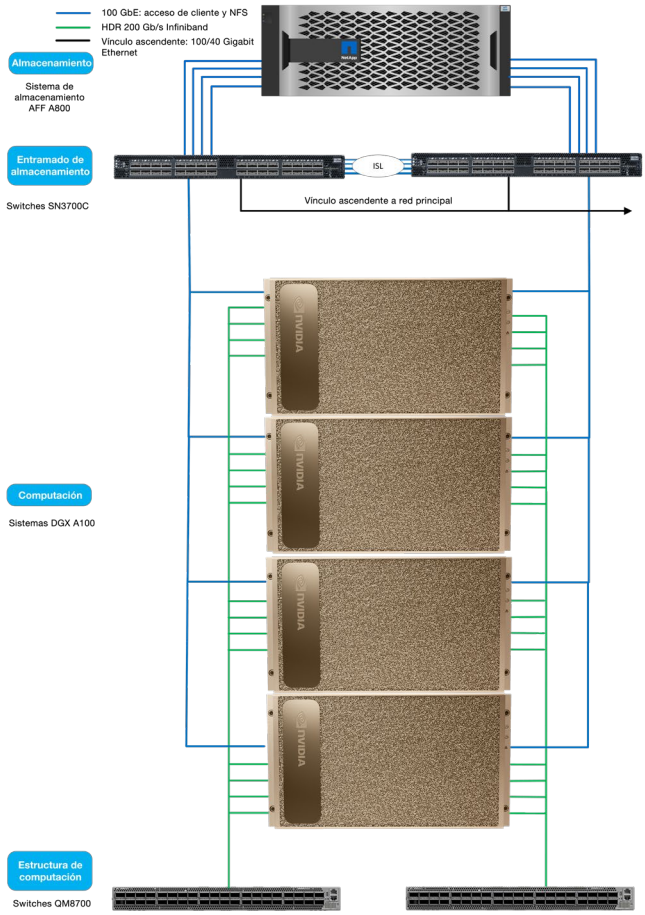


Figura 2) Configuración de cuatro nodos de ONTAP AI con switches Mellanox Spectrum de 100 GbE.

Con la arquitectura escalable en racks de NetApp, su organización puede pasar de decenas de terabytes a decenas de petabytes con el almacenamiento all-flash. Asimismo, con ONTAP FlexGroup de NetApp, hasta 20 petabytes de un solo espacio de nombres pueden manejar más de 400 000 millones de archivos.

Crear una canalización de datos integrada que abarque desde el perímetro hasta el núcleo y cloud

ONTAP AI utiliza su Data Fabric para unificar la gestión de datos en toda su canalización con una plataforma única. Use las mismas herramientas para controlar y proteger con seguridad los datos que se encuentran en movimiento, en uso o en descanso, y cumpla los requisitos de cumplimiento de normativas con confianza. Si surge un problema en su entorno de aprendizaje profundo, puede confiar en nuestro modelo de soporte contrastado para ayudarle a resolver los problemas y proporcionarle la orientación adecuada.

Unificar cargas de trabajo de IA

Ahora, su organización puede eliminar los silos de infraestructuras que están infrautilizadas o que quitan recursos a las cargas de trabajo de IA. Con ONTAP AI, obtiene una solución universal de infraestructura de IA que está desarrollada en sistemas DGX A100. La solución consolida el análisis, la formación y la inferencia en una plataforma que responde con flexibilidad para satisfacer las demandas del negocio. También consigue un mejor TCO que con las arquitecturas heredadas.

NetApp y NVIDIA: una colaboración que impulsa la innovación

En el centro de ONTAP AI, está el sistema DGX A100, un elemento básico universal para la IA de centro de datos que es compatible con la formación, la inferencia, la ciencia de datos y otras cargas de trabajo de alto rendimiento. Cada sistema DGX A100 cuenta con ocho GPU A100 Tensor Core de NVIDIA y procesadores dobles AMD EPYC™ de 2.ª generación. También integra las últimas interconexiones de adaptadores de alta velocidad ConnectX-6 con conectividad para Ethernet e InfiniBand de 100/200 Gb NVIDIA Mellanox.

Mediante la nueva tecnología de GPU (MIG) multiinstancia de NVIDIA, es posible acelerar múltiples cargas de trabajo más pequeñas al crear particiones del sistema DGX A100 de hasta 56 instancias por sistema. Gracias a esta aceleración, su organización podrá asignar un rendimiento de GPU de manera eficiente en ONTAP AI. Sus equipos de expertos en datos de toda la empresa pueden iterar con mayor rapidez, automatizar la reproducibilidad y entregar proyectos de IA con hasta tres meses de antelación y una mayor calidad.

Los sistemas AFF de NetApp mantienen el flujo de datos en los procesos de aprendizaje profundo y aprendizaje automático con el almacenamiento all-flash más flexible y más rápido del sector, que incluye las primeras tecnologías NVMe integrales del mundo. El sistema AFF A800 es capaz de suministrar datos a los sistemas DGX con una velocidad hasta cuatro veces superior a la de las soluciones de la competencia.[1]

La solución de ONTAP AI viene integrada en los switches Ethernet Mellanox Spectrum. Estos switches ofrecen la baja latencia, la alta densidad, el alto rendimiento y la eficiencia de potencia que demandan los entornos de IA.

1. Rendimiento de lectura de hasta 300 GB/s por clúster all-flash frente a los 75 GB/s de uno de los principales competidores.

Un Data Fabric de NetApp ofrece las mejores gestión de datos e integración en cloud para ayudarle a acelerar el aprendizaje profundo a la vez que gestiona y protege los datos cruciales. ONTAP ofrece un ratio sin igual de reducción de datos general de 22:1 y un TCO hasta un 54 % inferior en comparación con el almacenamiento DAS.

El sistema DGX A100 cuenta con la tecnología de pila de software de NVIDIA DGX, que incluye software optimizado para cargas de trabajo de ciencia de datos e inteligencia artificial. Obtiene un rendimiento maximizado que permite a su empresa alcanzar antes el retorno de la inversión en una infraestructura de IA.

El plano de control de IA de NetApp ayuda a simplificar la gestión de datos de IA al integrar Kubernetes y Kubeflow con un Data Fabric de NetApp. Esta solución integrada le proporciona una disponibilidad y una portabilidad de datos óptimas desde el perímetro hasta el núcleo y cloud. El kit de herramientas de DataOps de NetApp es una mejora para el plano de control de IA, pues es una biblioteca de Python con la que los científicos e ingenieros de datos pueden realizar numerosas tareas de gestión de datos con mayor facilidad. Por ejemplo, pueden aprovisionar un nuevo volumen de datos, clonar un volumen de datos de forma instantánea y crear una copia de un volumen de datos con Snapshot™ de NetApp para rastrearlo o marcarlo como referencia.

Disponer de las herramientas adecuadas es crucial para triunfar. Es por ello que ONTAP AI está validado con los software líderes de operaciones de aprendizaje automático (MLOps), incluidos Domino Data Lab e Iguazio, entre otros. Sus equipos pueden usar herramientas que ya conozcan para maximizar el valor de su entorno de IA y acelerar los plazos que se necesitan para obtener información.

Componentes de la solución

- Sistemas DGX A100 de NVIDIA
- Sistema de almacenamiento NetApp AFF A-Series con ONTAP 9
- NVIDIA Mellanox Spectrum SN3700C, NVIDIA Mellanox Quantum QM8700 y/o NVIDIA Mellanox Spectrum SN3700-V
- Pila de software NVIDIA DGX
- Plano de control de IA de NetApp
- Kit de herramientas de DataOps de NetApp

Arquitecturas de referencia

NetApp ha lanzado las siguientes arquitecturas de referencia basadas en ONTAP AI, dirigidas a casos de uso de sectores específicos:

- Arquitectura de referencia de ONTAP AI para la sanidad: imagen de diagnóstico
- Arquitectura de referencia de ONTAP AI para cargas de trabajo de conducción autónoma: diseño de soluciones
- Arquitectura de referencia de ONTAP AI para cargas de trabajo de servicios financieros: diseño de soluciones

---

Acerca de NetApp

En un sector lleno de generalistas, NetApp es un especialista. Nos centramos en una cosa: ayudar a que su empresa aproveche al máximo sus datos. NetApp incorpora a cloud los servicios de datos de clase empresarial en los que confía, y lleva la sencilla flexibilidad de cloud al centro de datos. Nuestras soluciones líderes del sector funcionan en diversos entornos del cliente y en los clouds públicos más grandes del mundo.

Como empresa de software centrado en datos y orientado a cloud, solo NetApp puede ayudar a crear su Data Fabric exclusivo, a simplificar y conectar su cloud, y a proporcionar con seguridad los datos, los servicios y las aplicaciones correctos a las personas adecuadas en cualquier momento y lugar. www.netapp.com/es

Acerca de NVIDIA

La invención de la GPU en 1999 por parte de NVIDIA desencadenó el crecimiento del mercado de videojuegos en PC, redefinió los gráficos de ordenador modernos y revolucionó la computación paralela. Más recientemente, el aprendizaje profundo de las GPU ha puesto en marcha la inteligencia artificial moderna (la próxima era de la informática) donde la unidad de procesamiento gráfico actúa como el cerebro de ordenadores, robots y vehículos autónomos que pueden percibir y comprender el mundo a su alrededor.
Más información en www.nvidia.com.

NetApp