IA generativa y valor de NetApp

NetApp artificial intelligence solutions

NetApp
December 04, 2025

Tabla de contenidos

IA generativa y valor de NetApp

La demanda de inteligencia artificial (IA) generativa está generando disrupción en todas las industrias, mejorando la creatividad empresarial y la innovación de productos.

Abstracto

Muchas organizaciones están utilizando IA generativa para crear nuevas funciones de productos, mejorar la productividad de ingeniería y crear prototipos de aplicaciones impulsadas por IA que ofrecen mejores resultados y experiencias para los consumidores. La IA generativa, como los transformadores generativos preentrenados (GPT), utiliza redes neuronales para crear contenido nuevo, tan diverso como texto, audio y video. Dada la escala extrema y los conjuntos de datos masivos involucrados con los modelos de lenguaje grandes (LLM), es crucial diseñar una infraestructura de IA robusta que aproveche las atractivas características de almacenamiento de datos de las opciones de implementación locales, híbridas y multicloud, y reduzca los riesgos asociados con la movilidad de datos, la protección de datos y la gobernanza antes de que las empresas puedan diseñar soluciones de IA. En este documento se describen estas consideraciones y las capacidades de IA de NetApp correspondientes que permiten la gestión fluida de datos y el movimiento de datos a lo largo del flujo de datos de IA para entrenar, reentrenar, ajustar e inferir modelos generativos de IA.

Resumen ejecutivo

Más recientemente, tras el lanzamiento de ChatGPT, una escisión de GPT-3 en noviembre de 2022, las nuevas herramientas de IA utilizadas para generar texto, código, imágenes o incluso proteínas terapéuticas en respuesta a las indicaciones del usuario han ganado una fama significativa. Esto indica que los usuarios pueden realizar una solicitud usando lenguaje natural y la IA interpretará y generará texto, como artículos de noticias o descripciones de productos que reflejen la solicitud del usuario o producirán código, música, voz, efectos visuales y activos 3D utilizando algoritmos entrenados con datos ya existentes. Como resultado, frases como difusión estable, alucinaciones, ingeniería rápida y alineación de valores están surgiendo rápidamente en el diseño de sistemas de IA. Estos modelos de aprendizaje automático (ML) autosupervisados o semisupervisados se están volviendo ampliamente disponibles como modelos básicos entrenados previamente (FM) a través de proveedores de servicios en la nube y otras empresas proveedoras de IA, que están siendo adoptados por varios establecimientos comerciales en todas las industrias para una amplia gama de tareas posteriores de PNL (procesamiento del lenguaje natural). Como afirman firmas de análisis de investigación como McKinsey: "El impacto de la IA generativa en la productividad podría agregar billones de dólares en valor a la economía global". Mientras las empresas están reinventando la IA como un socio intelectual para los humanos y los gerentes de finanzas están ampliando simultáneamente lo que las empresas e instituciones pueden hacer con la IA generativa, las oportunidades para gestionar volúmenes masivos de datos seguirán creciendo. Este documento presenta información introductoria sobre la IA generativa y los conceptos de diseño en relación con las capacidades de NetApp que aportan valor a los clientes de NetApp , tanto en entornos locales como híbridos o multicloud.

Entonces, ¿qué beneficios obtienen los clientes al utilizar NetApp en sus entornos de IA?

NetApp ayuda a las organizaciones a afrontar las complejidades creadas por el rápido crecimiento de los datos y la nube, la gestión de múltiples nubes y la adopción de tecnologías de última generación, como la IA. NetApp ha combinado varias capacidades en un software de gestión de datos inteligente y una infraestructura de almacenamiento que están bien equilibradas con un alto rendimiento optimizado para cargas de trabajo de IA. Las soluciones de IA generativa como LLM necesitan leer y procesar sus conjuntos de datos de origen desde el almacenamiento a la memoria numerosas veces para fomentar la inteligencia.

NetApp ha sido líder en tecnologías de movilidad de datos, gobernanza de datos y seguridad de datos en todo

el ecosistema del borde al núcleo y a la nube, ayudando a clientes empresariales a construir soluciones de IA a escala. NetApp, con una sólida red de socios, ha estado ayudando a los directores de datos, ingenieros de IA, arquitectos empresariales y científicos de datos en el diseño de un flujo de datos libre para la preparación de datos, la protección de datos y las responsabilidades de gestión estratégica de datos del entrenamiento y la inferencia de modelos de IA, optimizando el rendimiento y la escalabilidad del ciclo de vida de IA/ML. Las tecnologías y capacidades de datos de NetApp , como NetApp ONTAP AI para la canalización de datos de aprendizaje profundo, NetApp SnapMirror para transportar datos de manera fluida y eficiente entre puntos finales de almacenamiento, y NetApp FlexCache para la representación en tiempo real cuando el flujo de datos cambia de lote a tiempo real y la ingeniería de datos ocurre en el momento oportuno, aportan valor a la implementación de modelos de IA generativa en tiempo real. A medida que las empresas de todo tipo adoptan nuevas herramientas de IA, enfrentan desafíos de datos desde el borde hasta el centro de datos y la nube que exigen soluciones de IA escalables, responsables y explicables.

Como autoridad en datos sobre nubes híbridas y múltiples, NetApp se compromete a construir una red de socios y soluciones conjuntas que puedan ayudar con todos los aspectos de la construcción de una canalización de datos y lagos de datos para el entrenamiento de modelos de IA generativos (preentrenamiento), el ajuste fino, la inferencia basada en el contexto y el monitoreo de la descomposición del modelo de LLM.

¿Qué es la IA generativa?

La IA generativa está cambiando la forma en que creamos contenido, generamos nuevos conceptos de diseño y exploramos composiciones novedosas. Ilustra marcos de redes neuronales como la red generativa antagónica (GAN), los autocodificadores variacionales (VAE) y los transformadores generativos preentrenados (GPT), que pueden generar contenido nuevo como texto, código, imágenes, audio, video y datos sintéticos. Los modelos basados en transformadores como Chat-GPT de OpenAI, Bard de Google, BLOOM de Hugging Face y LLaMA de Meta han surgido como la tecnología fundamental que sustenta muchos avances en modelos de lenguaje de gran tamaño. Del mismo modo, Dall-E de OpenAI, CM3leon de Meta e Imagen de Google son ejemplos de modelos de difusión de texto a imagen que ofrecen a los clientes un grado de fotorrealismo sin precedentes para crear imágenes nuevas y complejas desde cero o editar imágenes existentes para generar imágenes de alta calidad que tengan en cuenta el contexto mediante el aumento de conjuntos de datos y la síntesis de texto a imagen que vincula la semántica textual y visual. Los artistas digitales están comenzando a aplicar una combinación de tecnologías de renderizado como NeRF (Neural Radiance Field) con IA generativa para convertir imágenes 2D estáticas en escenas 3D inmersivas. En general, los LLM se caracterizan por cuatro parámetros: (1) Tamaño del modelo (normalmente en miles de millones de parámetros); (2) Tamaño del conjunto de datos de entrenamiento; (3) Costo de entrenamiento, y (4) Rendimiento del modelo después del entrenamiento. Los LLM también se dividen principalmente en tres arquitecturas de transformadores. (i) Modelos de sólo codificador. Por ejemplo, BERT (Google, 2018); (ii) Modelos codificador-decodificador, por ejemplo, BART (Meta, 2020) y (iii) Modelos solo decodificador. Por ejemplo, LLaMA (Meta, 2023), PaLM-E (Google, 2023). Dependiendo del requisito comercial, independientemente de la arquitectura que elija una empresa, la cantidad de parámetros del modelo (N) y la cantidad de tokens (D) en el conjunto de datos de entrenamiento generalmente determinan el costo base del entrenamiento (preentrenamiento) o el ajuste de un LLM.

Casos de uso empresarial y tareas posteriores de PNL

Las empresas de todos los sectores están descubriendo cada vez más potencial para que la IA extraiga y produzca nuevas formas de valor a partir de datos existentes para operaciones comerciales, ventas, marketing y servicios legales. Según la información de mercado de IDC (International Data Corporation) sobre casos de uso e inversiones en IA generativa a nivel mundial, la gestión del conocimiento en el desarrollo de software y el diseño de productos será la más afectada, seguida por la creación de historias para marketing y la generación de código para desarrolladores. En el ámbito sanitario, las organizaciones de investigación clínica están abriendo nuevos caminos en la medicina. Los modelos preentrenados como ProteinBERT incorporan anotaciones de Gene Ontology (GO) para diseñar rápidamente estructuras de proteínas para

medicamentos médicos, lo que representa un hito significativo en el descubrimiento de fármacos, la bioinformática y la biología molecular. Las empresas de biotecnología han iniciado ensayos en humanos para una medicina generativa descubierta mediante inteligencia artificial, que tiene como objetivo tratar enfermedades como la fibrosis pulmonar (FPI), una enfermedad pulmonar que causa cicatrices irreversibles en el tejido pulmonar.

Figura 1: Casos de uso que impulsan la IA generativa

[Figura 1: Casos de uso que impulsan la IA generativa]

El aumento en la adopción de la automatización impulsada por la IA generativa también está cambiando la oferta y la demanda de actividades laborales para muchas ocupaciones. Según McKinsey, el mercado laboral de EE. UU. (diagrama a continuación) ha atravesado una rápida transición, que solo puede continuar si se tiene en cuenta el impacto de la IA.

Fuente: McKinsey & Company

[Figura 2: Fuente: McKinsey Company]

El papel del almacenamiento en la IA generativa

Los LLM se basan en gran medida en el aprendizaje profundo, las GPU y la computación. Sin embargo, cuando el búfer de la GPU se llena, los datos deben escribirse rápidamente en el almacenamiento. Si bien algunos modelos de IA son lo suficientemente pequeños para ejecutarse en la memoria, los LLM requieren IOPS altos y almacenamiento de alto rendimiento para brindar acceso rápido a grandes conjuntos de datos, especialmente si involucran miles de millones de tokens o millones de imágenes. Para un requisito típico de memoria de GPU de un LLM, la memoria necesaria para entrenar un modelo con mil millones de parámetros podría llegar hasta 80 GB con precisión completa de 32 bits. En ese caso, LLaMA 2 de Meta, una familia de LLM que varía en escala desde 7 mil millones a 70 mil millones de parámetros, puede requerir 70x80, aproximadamente 5600 GB o 5,6 TB de RAM de GPU. Además, la cantidad de memoria que necesitas es directamente proporcional al número máximo de tokens que quieras generar. Por ejemplo, si desea generar resultados de hasta 512 tokens (aproximadamente 380 palabras), necesita"512 MB" . Quizás parezca insignificante, pero si quieres procesar lotes más grandes, empieza a sumar. Por lo tanto, resulta muy costoso para las organizaciones capacitar o ajustar los LLM en memoria, convirtiendo así el almacenamiento en una piedra angular para la IA generativa.

Tres enfoques principales para los LLM

Para la mayoría de las empresas, según las tendencias actuales, el enfoque para implementar LLM se puede condensar en tres escenarios básicos. Como se describe en un artículo reciente"Revista de negocios de Harvard" Artículo: (1) Capacitación (preentrenamiento) de un LLM desde cero: costosa y requiere habilidades expertas en IA/ML; (2) Ajuste de un modelo de base con datos empresariales: complejo, pero factible; (3) Uso de generación aumentada por recuperación (RAG) para consultar repositorios de documentos, API y bases de datos vectoriales que contienen datos de la empresa. Cada uno de ellos implica compensaciones entre el esfuerzo, la velocidad de iteración, la relación coste-eficiencia y la precisión del modelo en sus implementaciones, utilizadas para resolver distintos tipos de problemas (diagrama a continuación).

Figura 3: Tipos de problemas

[Figura 3: Tipos de problemas]

Modelos de Fundación

Un modelo de fundación (FM), también conocido como modelo base, es un modelo de IA de gran tamaño

(LLM) entrenado en grandes cantidades de datos sin etiquetar, utilizando autosupervisión a escala, generalmente adaptado para una amplia gama de tareas de PNL posteriores. Dado que los datos de entrenamiento no están etiquetados por humanos, el modelo emerge en lugar de estar codificado explícitamente. Esto significa que el modelo puede generar historias o una narrativa propia sin estar programado explícitamente para hacerlo. Por lo tanto, una característica importante del FM es la homogeneización, lo que significa que se utiliza el mismo método en muchos dominios. Sin embargo, con las técnicas de personalización y ajuste, los FM integrados en los productos que aparecen hoy en día no solo son buenos para generar texto, texto a imágenes y texto a código, sino también para explicar tareas específicas del dominio o depurar código. Por ejemplo, FM como Codex de OpenAI o Code Llama de Meta pueden generar código en múltiples lenguajes de programación basándose en descripciones en lenguaje natural de una tarea de programación. Estos modelos son competentes en más de una docena de lenguajes de programación, incluidos Python, C#, JavaScript, Perl, Ruby y SQL. Entienden la intención del usuario y generan código específico que logra la tarea deseada, útil para el desarrollo de software, la optimización de código y la automatización de tareas de programación.

Ajuste fino, especificidad del dominio y reentrenamiento

Una de las prácticas comunes con la implementación de LLM luego de la preparación y el preprocesamiento de datos es seleccionar un modelo previamente entrenado que haya sido entrenado en un conjunto de datos grande y diverso. En el contexto del ajuste fino, este puede ser un modelo de lenguaje grande de código abierto como"Llama de Meta 2" entrenado con 70 mil millones de parámetros y 2 billones de tokens. Una vez seleccionado el modelo previamente entrenado, el siguiente paso es ajustarlo a los datos específicos del dominio. Esto implica ajustar los parámetros del modelo y entrenarlo con los nuevos datos para adaptarse a un dominio y una tarea específicos. Por ejemplo, BloombergGPT, un LLM propietario capacitado en una amplia gama de datos financieros al servicio de la industria financiera.

Los modelos específicos de dominio diseñados y entrenados para una tarea específica generalmente tienen mayor precisión y desempeño dentro de su alcance, pero baja transferibilidad a otras tareas o dominios. Cuando el entorno empresarial y los datos cambian durante un período, la precisión de predicción del FM podría comenzar a disminuir en comparación con su desempeño durante las pruebas. Es en este momento cuando el reentrenamiento o ajuste del modelo se vuelve crucial.

El reentrenamiento de modelos en IA/ML tradicional se refiere a la actualización de un modelo de ML implementado con nuevos datos, generalmente realizado para eliminar dos tipos de desviaciones que ocurren. (1) Deriva de concepto: cuando el vínculo entre las variables de entrada y las variables de destino cambia con el tiempo, dado que la descripción de lo que queremos predecir cambia, el modelo puede producir predicciones inexactas. (2) Desviación de datos: ocurre cuando las características de los datos de entrada cambian, como cambios en los hábitos o el comportamiento del cliente a lo largo del tiempo y, por lo tanto, la incapacidad del modelo para responder a dichos cambios.

De manera similar, la capacitación se aplica a los FM/LLM, sin embargo, puede ser mucho más costosa (en millones de dólares), por lo que no es algo que la mayoría de las organizaciones podrían considerar. Se encuentra bajo investigación activa y aún está emergiendo en el ámbito de LLMOps. Por lo tanto, en lugar de volver a entrenar, cuando se produce un deterioro del modelo en los FM ajustados, las empresas pueden optar por ajustarlo nuevamente (mucho más barato) con un conjunto de datos más nuevo. Para una perspectiva de costos, a continuación se muestra un ejemplo de una tabla de precios de modelos de Azure-OpenAI Services. Para cada categoría de tarea, los clientes pueden ajustar y evaluar modelos en conjuntos de datos específicos.

Fuente: Microsoft Azure

[Fuente: Microsoft Azure]

Ingeniería rápida e inferencia

La ingeniería rápida se refiere a los métodos efectivos de cómo comunicarse con los LLM para realizar las tareas deseadas sin actualizar los pesos del modelo. Tan importante como es el entrenamiento y el ajuste del modelo de IA para las aplicaciones de PNL, es igualmente importante la inferencia, donde los modelos entrenados responden a las indicaciones del usuario. Los requisitos del sistema para la inferencia generalmente se basan mucho más en el rendimiento de lectura del sistema de almacenamiento de IA que alimenta datos desde los LLM a las GPU, ya que necesita poder aplicar miles de millones de parámetros de modelo almacenados para producir la mejor respuesta.

LLMOps, Monitoreo de Modelos y Almacenes de Vectores

Al igual que las operaciones de aprendizaje automático tradicionales (MLOps), las operaciones de modelos de lenguaje grandes (LLMOps) también requieren la colaboración de científicos de datos e ingenieros de DevOps con herramientas y mejores prácticas para la gestión de LLM en entornos de producción. Sin embargo, el flujo de trabajo y la pila tecnológica para los LLM pueden variar de algunas maneras. Por ejemplo, las canalizaciones LLM creadas utilizando marcos como LangChain unen múltiples llamadas API LLM a puntos finales de incrustación externos, como almacenes de vectores o bases de datos vectoriales. El uso de un punto final de integración y un almacén de vectores para conectores posteriores (como una base de datos vectorial) representa un avance significativo en la forma en que se almacenan y acceden los datos. A diferencia de los modelos ML tradicionales que se desarrollan desde cero, los LLM a menudo se basan en el aprendizaje por transferencia, ya que estos modelos comienzan con FM que se ajustan con nuevos datos para mejorar el rendimiento en un dominio más específico. Por lo tanto, es fundamental que LLMOps brinde capacidades de gestión de riesgos y monitoreo del deterioro del modelo.

Riesgos y ética en la era de la IA generativa

"ChatGPT: es ingenioso, pero aun así dice tonterías". – MIT Tech Review. Basura que entra, basura que sale: ese ha sido siempre el desafío en informática. La única diferencia con la IA generativa es que se destaca por hacer que la basura sea altamente creíble, lo que conduce a resultados inexactos. Los LLM son propensos a inventar hechos que se ajusten a la narrativa que están construyendo. Por lo tanto, las empresas que ven la IA generativa como una gran oportunidad para reducir sus costos con equivalentes de IA necesitan detectar eficientemente las falsificaciones profundas, reducir los sesgos y disminuir los riesgos para mantener los sistemas honestos y éticos. Una tubería de datos de flujo libre con una infraestructura de IA robusta que respalde la movilidad de los datos, la calidad de los datos, la gobernanza de los datos y la protección de los datos mediante cifrado de extremo a extremo y barandillas de IA es fundamental en el diseño de modelos de IA generativos responsables y explicables.

Escenario del cliente y NetApp

Figura 3: Flujo de trabajo de aprendizaje automático/modelo de lenguaje grande

[Figura 3: Flujo de trabajo de aprendizaje automático/modelo de lenguaje grande]

¿Estamos entrenando o afinando? La cuestión de si (a) entrenar un modelo LLM desde cero, ajustar un FM entrenado previamente o usar RAG para recuperar datos de repositorios de documentos fuera de un modelo base y aumentar las indicaciones, y (b) aprovechar LLM de código abierto (por ejemplo, Llama 2) o FM propietarios (por ejemplo, ChatGPT, Bard, AWS Bedrock) es una decisión estratégica para las organizaciones. Cada enfoque implica un equilibrio entre la relación coste-eficiencia, la gravedad de los datos, las operaciones, la precisión del modelo y la gestión de los LLM.

NetApp , como empresa, adopta la IA internamente en su cultura de trabajo y en su enfoque de los esfuerzos de diseño e ingeniería de productos. Por ejemplo, la protección autónoma contra ransomware de NetApp está construida utilizando inteligencia artificial y aprendizaje automático. Proporciona detección temprana de

anomalías del sistema de archivos para ayudar a identificar amenazas antes de que afecten las operaciones. En segundo lugar, NetApp utiliza IA predictiva para sus operaciones comerciales, como pronósticos de ventas e inventario, y chatbots para ayudar a los clientes en servicios de soporte de productos del centro de llamadas, especificaciones técnicas, garantía, manuales de servicio y más. En tercer lugar, NetApp aporta valor al cliente en el flujo de trabajo de ML/LLM y en la canalización de datos de IA mediante productos y soluciones que prestan servicios a clientes que crean soluciones de IA predictivas, como previsión de la demanda, imágenes médicas, análisis de sentimientos y soluciones de IA generativa como GAN para la detección de anomalías en imágenes industriales en el sector manufacturero y la detección de fraude y lavado de dinero en servicios bancarios y financieros con productos y capacidades de NetApp como NetApp ONTAP AI, NetApp SnapMirror y NetApp FlexCache.

Capacidades de NetApp

El movimiento y la gestión de datos en aplicaciones de IA generativa, como chatbots, generación de código, generación de imágenes o expresión de modelos genómicos, pueden abarcar el borde, el centro de datos privado y el ecosistema híbrido de múltiples nubes. Por ejemplo, un robot de inteligencia artificial en tiempo real que ayuda a un pasajero a mejorar su boleto de avión a clase ejecutiva desde una aplicación de usuario final expuesta a través de API de modelos previamente entrenados como ChatGPT no puede lograr esa tarea por sí solo, ya que la información del pasajero no está disponible públicamente en Internet. La API requiere acceso a la información personal del pasajero y a la información del boleto de la aerolínea, que puede existir en un ecosistema híbrido o multicloud. Un escenario similar podría aplicarse a científicos que comparten una molécula de fármaco y datos de pacientes a través de una aplicación de usuario final que utiliza LLM para realizar ensayos clínicos en todo el proceso de descubrimiento de fármacos que involucra de una a muchas instituciones de investigación biomédica. Los datos confidenciales que se transmiten a los FM o LLM pueden incluir información personal identificable, información financiera, información de salud, datos biométricos, datos de ubicación, datos de comunicaciones, comportamiento en línea e información legal. En un evento de renderizado en tiempo real, ejecución inmediata e inferencia de borde, hay un movimiento de datos desde la aplicación del usuario final a los puntos finales de almacenamiento a través de modelos LLM propietarios o de código abierto a un centro de datos local o plataformas de nube pública. En todos estos escenarios, la movilidad y la protección de datos son cruciales para las operaciones de IA que involucran LLM que dependen de grandes conjuntos de datos de entrenamiento y del movimiento de dichos datos.

Figura 4: Canal de datos de IA generativa - LLM

[Figura 4: Canal de datos generativo de IA-LLM]

La cartera de infraestructura de almacenamiento, datos y servicios en la nube de NetApp está impulsada por un software de gestión de datos inteligente.

Preparación de datos: El primer pilar de la pila tecnológica de LLM prácticamente no se ha visto afectado por la antigua pila tradicional de ML. El preprocesamiento de datos en la canalización de IA es necesario para normalizar y limpiar los datos antes del entrenamiento o el ajuste. Este paso incluye conectores para ingerir datos donde sea que residan en forma de un nivel de Amazon S3 o en sistemas de almacenamiento locales, como un almacén de archivos o un almacén de objetos como NetApp StorageGRID.

- NetApp ONTAP* es la tecnología fundamental que sustenta las soluciones de almacenamiento críticas de NetApp en el centro de datos y la nube. ONTAP incluye varias funciones y capacidades de protección y administración de datos, incluida protección automática contra ransomware contra ciberataques, funciones de transporte de datos integradas y capacidades de eficiencia de almacenamiento para una variedad de arquitecturas, desde locales, híbridas, multicloud en NAS, SAN, objetos y situaciones de almacenamiento definido por software (SDS) de implementaciones LLM.
- NetApp ONTAP AI * para entrenamiento de modelos de aprendizaje profundo. NetApp ONTAP admite el almacenamiento directo de GPU NVIDIA con el uso de NFS sobre RDMA para clientes de NetApp con clúster de almacenamiento ONTAP y nodos de cómputo NVIDIA DGX. Ofrece un rendimiento rentable

para leer y procesar conjuntos de datos de origen desde el almacenamiento a la memoria numerosas veces para fomentar la inteligencia, lo que permite a las organizaciones capacitación, ajuste y escalamiento del acceso a los LLM.

- NetApp FlexCache* es una capacidad de almacenamiento en caché remoto que simplifica la distribución de archivos y almacena en caché solo los datos leídos activamente. Esto puede ser útil para la capacitación, el reentrenamiento y el ajuste de LLM, aportando valor a los clientes con requisitos comerciales como representación en tiempo real e inferencia de LLM.
- NetApp SnapMirror* es una función de ONTAP que replica instantáneas de volumen entre dos sistemas ONTAP . Esta función transfiere datos de manera óptima desde el borde a su centro de datos local o a la nube. SnapMirror se puede utilizar para mover datos de forma segura y eficiente entre nubes locales y nubes hiperescalables cuando los clientes desean desarrollar IA generativa en nubes con RAG que contienen datos empresariales. Transfiere eficientemente solo los cambios, ahorrando ancho de banda y acelerando la replicación, aportando así características esenciales de movilidad de datos durante las operaciones de entrenamiento, reentrenamiento y ajuste de FM o LLM.
- NetApp SnapLock* brinda capacidad de disco inmutable en sistemas de almacenamiento basados en ONTAP para el control de versiones de conjuntos de datos. La arquitectura de micronúcleos está diseñada para proteger los datos del cliente con el motor FPolicy Zero Trust. NetApp garantiza que los datos de los clientes estén disponibles al resistir ataques de denegación de servicio (DoS) cuando un atacante interactúa con un LLM de una manera que consume especialmente recursos.
- NetApp Cloud Data Sense* ayuda a identificar, mapear y clasificar la información personal presente en conjuntos de datos empresariales, implementar políticas, cumplir con los requisitos de privacidad en las instalaciones o en la nube, ayudar a mejorar la postura de seguridad y cumplir con las regulaciones.

Clasificación * NetApp BlueXP*, impulsada por Cloud Data Sense. Los clientes pueden escanear, analizar, categorizar y actuar automáticamente sobre los datos en todo el patrimonio de datos, detectar riesgos de seguridad, optimizar el almacenamiento y acelerar las implementaciones en la nube. Combina servicios de almacenamiento y datos a través de su plano de control unificado. Los clientes pueden usar instancias de GPU para computación y entornos multicloud híbridos para niveles de almacenamiento en frío y para archivos y copias de seguridad.

*Dualidad archivo-objeto de NetApp *. NetApp ONTAP permite el acceso de protocolo dual para NFS y S3. Con esta solución, los clientes pueden acceder a datos NFS desde notebooks de Amazon AWS SageMaker a través de buckets S3 desde NetApp Cloud Volumes ONTAP. Esto ofrece flexibilidad a los clientes que necesitan acceso fácil a fuentes de datos heterogéneas con la capacidad de compartir datos tanto de NFS como de S3. Por ejemplo, ajustar FMs como los modelos de generación de texto Llama 2 de Meta en SageMaker con acceso a depósitos de objetos de archivo.

El servicio * NetApp Cloud Sync* ofrece una forma sencilla y segura de migrar datos a cualquier destino, en la nube o en las instalaciones. Cloud Sync transfiere y sincroniza sin problemas datos entre almacenamiento local o en la nube, NAS y almacenes de objetos.

- NetApp XCP * es un software cliente que permite migraciones de datos rápidas y confiables de cualquier plataforma a NetApp y de NetApp a NetApp . XCP también proporciona la capacidad de mover datos masivos de manera eficiente desde sistemas de archivos Hadoop HDFS a ONTAP NFS, S3 o StorageGRID y el análisis de archivos XCP proporciona visibilidad del sistema de archivos.
- NetApp DataOps Toolkit * es una biblioteca de Python que permite a los científicos de datos, DevOps e ingenieros de datos realizar fácilmente diversas tareas de administración de datos, como aprovisionamiento, clonación o captura de instantáneas casi instantáneas de un volumen de datos o un espacio de trabajo de JupyterLab respaldados por almacenamiento NetApp de escalamiento horizontal de alto rendimiento.

Seguridad del producto de NetApp. Los LLM pueden revelar inadvertidamente datos confidenciales en sus respuestas, lo que preocupa a los CISO que estudian las vulnerabilidades asociadas con las aplicaciones de

IA que aprovechan los LLM. Como lo describe OWASP (Proyecto Abierto de Seguridad de Aplicaciones Mundiales), los problemas de seguridad como envenenamiento de datos, fuga de datos, denegación de servicio e inyecciones rápidas dentro de los LLM pueden afectar a las empresas desde la exposición de datos hasta el acceso no autorizado por parte de atacantes. Los requisitos de almacenamiento de datos deben incluir controles de integridad e instantáneas inmutables para datos estructurados, semiestructurados y no estructurados. Se utilizan instantáneas de NetApp y SnapLock para el control de versiones de conjuntos de datos. Ofrece un estricto control de acceso basado en roles (RBAC), así como protocolos seguros y cifrado estándar de la industria para proteger los datos en reposo y en tránsito. Cloud Insights y Cloud Data Sense juntos ofrecen capacidades para ayudarlo a identificar forensemente la fuente de la amenaza y priorizar qué datos restaurar.

* ONTAP AI con DGX BasePOD *

La arquitectura de referencia de IA de NetApp ONTAP con NVIDIA DGX BasePOD es una arquitectura escalable para cargas de trabajo de aprendizaje automático (ML) e inteligencia artificial (IA). Para la fase crítica de entrenamiento de los LLM, los datos normalmente se copian del almacenamiento de datos al clúster de entrenamiento a intervalos regulares. Los servidores que se utilizan en esta fase utilizan GPU para paralelizar los cálculos, lo que crea un enorme apetito por los datos. Satisfacer las necesidades de ancho de banda de E/S sin procesar es crucial para mantener una alta utilización de la GPU.

* ONTAP AI con NVIDIA AI Enterprise *

NVIDIA AI Enterprise es una suite integral de software de inteligencia artificial y análisis de datos nativa de la nube, optimizada, certificada y respaldada por NVIDIA para ejecutarse en VMware vSphere con sistemas certificados NVIDIA. Este software facilita la implementación, la gestión y el escalamiento simples y rápidos de cargas de trabajo de IA en el entorno de nube híbrida moderno. NVIDIA AI Enterprise, con tecnología de NetApp y VMware, ofrece gestión de datos y cargas de trabajo de IA de nivel empresarial en un paquete simplificado y familiar.

Plataformas en la nube 1P

Las ofertas de almacenamiento en la nube totalmente administradas están disponibles de forma nativa en Microsoft Azure como Azure NetApp Files (ANF), en AWS como Amazon FSx for NetApp ONTAP (FSx ONTAP) y en Google como Google Cloud NetApp Volumes (GNCV). 1P es un sistema de archivos administrado y de alto rendimiento que permite a los clientes ejecutar cargas de trabajo de IA de alta disponibilidad con seguridad de datos mejorada en nubes públicas, para ajustar LLM/FM con plataformas de ML nativas de la nube como AWS SageMaker, Azure-OpenAI Services y Vertex AI de Google.

Suite de soluciones para socios de NetApp

Además de sus principales productos, tecnologías y capacidades de datos, NetApp también colabora estrechamente con una sólida red de socios de IA para brindar valor agregado a los clientes.

- NVIDIA Guardrails * en los sistemas de IA sirven como medidas de protección para garantizar el uso ético y responsable de las tecnologías de IA. Los desarrolladores de IA pueden elegir definir el comportamiento de las aplicaciones impulsadas por LLM en temas específicos y evitar que participen en discusiones sobre temas no deseados. Guardrails, un kit de herramientas de código abierto, brinda la posibilidad de conectar un LLM a otros servicios de manera segura y sin inconvenientes para construir sistemas conversacionales LLM confiables y seguros.

Domino Data Lab ofrece herramientas versátiles y de nivel empresarial para crear y producir IA generativa: de manera rápida, segura y económica, dondequiera que se encuentre en su recorrido hacia la IA. Con la plataforma Enterprise MLOps de Domino, los científicos de datos pueden usar las herramientas preferidas y

todos sus datos, entrenar e implementar modelos fácilmente en cualquier lugar y administrar el riesgo y los costos de manera efectiva, todo desde un centro de control.

Modzy para Edge AI. NetApp y Modzy se han asociado para brindar IA a escala para cualquier tipo de datos, incluidas imágenes, audio, texto y tablas. Modzy es una plataforma MLOps para implementar, integrar y ejecutar modelos de IA, que ofrece a los científicos de datos las capacidades de monitoreo de modelos, detección de desviaciones y explicabilidad, con una solución integrada para una inferencia LLM fluida.

Run:AI y NetApp se han asociado para demostrar las capacidades únicas de la solución NetApp ONTAP AI con la plataforma de gestión de clústeres Run:AI para simplificar la orquestación de cargas de trabajo de IA. Divide y une automáticamente los recursos de la GPU, diseñado para escalar sus canales de procesamiento de datos a cientos de máquinas con marcos de integración integrados para Spark, Ray, Dask y Rapids.

Conclusión

La IA generativa solo puede producir resultados efectivos cuando el modelo se entrena con grandes cantidades de datos de calidad. Si bien los LLM han alcanzado hitos notables, es fundamental reconocer sus limitaciones, los desafíos de diseño y los riesgos asociados con la movilidad y la calidad de los datos. Los LLM se basan en conjuntos de datos de entrenamiento grandes y dispares provenientes de fuentes de datos heterogéneas. Los resultados inexactos o sesgados generados por los modelos pueden poner en peligro tanto a las empresas como a los consumidores. Estos riesgos pueden corresponder a limitaciones para los LLM que surgen potencialmente de los desafíos de gestión de datos asociados con la calidad de los datos, la seguridad de los datos y la movilidad de los datos. NetApp ayuda a las organizaciones a afrontar las complejidades creadas por el rápido crecimiento de los datos, la movilidad de los datos, la gestión de múltiples nubes y la adopción de IA. La infraestructura de IA a gran escala y la gestión eficiente de datos son cruciales para definir el éxito de las aplicaciones de IA como la IA generativa. Es fundamental que los clientes cubran todos los escenarios de implementación sin comprometer la capacidad de expansión según lo necesiten las empresas y manteniendo al mismo tiempo la rentabilidad, la gobernanza de datos y las prácticas éticas de IA bajo control. NetApp trabaja constantemente para ayudar a los clientes a simplificar y acelerar sus implementaciones de IA.

Información de copyright

Información de la marca comercial