IA generativa e valor NetApp

NetApp artificial intelligence solutions

NetApp
December 04, 2025

Índice

IA generativa e valor NetApp

A demanda por inteligência artificial generativa (IA) está gerando disrupção em todos os setores, aumentando a criatividade empresarial e a inovação de produtos.

Resumo

Muitas organizações estão usando IA generativa para criar novos recursos de produtos, melhorar a produtividade da engenharia e prototipar aplicativos com tecnologia de IA que oferecem melhores resultados e experiências ao consumidor. A IA generativa, como os Transformadores Pré-treinados Generativos (GPT), usa redes neurais para criar novos conteúdos, tão diversos quanto texto, áudio e vídeo. Dada a escala extrema e os enormes conjuntos de dados envolvidos com grandes modelos de linguagem (LLMs), é crucial arquitetar uma infraestrutura de IA robusta que aproveite os recursos atraentes de armazenamento de dados das opções de implantação local, híbrida e multinuvem e reduza os riscos associados à mobilidade de dados, proteção de dados e governança antes que as empresas possam projetar soluções de IA. Este artigo descreve essas considerações e os recursos correspondentes do NetApp AI que permitem o gerenciamento e a movimentação de dados perfeitos no pipeline de dados de IA para treinamento, retreinamento, ajuste fino e inferência de modelos de IA generativos.

Sumário executivo

Mais recentemente, após o lançamento do ChatGPT, um spin-off do GPT-3 em novembro de 2022, novas ferramentas de IA usadas para gerar texto, código, imagem ou até mesmo proteínas terapêuticas em resposta a solicitações do usuário ganharam fama significativa. Isso indica que os usuários podem fazer uma solicitação usando linguagem natural e a IA interpretará e gerará texto, como artigos de notícias ou descrições de produtos que refletem a solicitação do usuário ou produzem código, música, fala, efeitos visuais e ativos 3D usando algoritmos treinados em dados já existentes. Como resultado, frases como Difusão Estável, Alucinações, Engenharia Rápida e Alinhamento de Valores estão surgindo rapidamente no design de sistemas de IA. Esses modelos de aprendizado de máquina (ML) autossupervisionados ou semissupervisionados estão se tornando amplamente disponíveis como modelos básicos pré-treinados (FM) por meio de provedores de serviços de nuvem e outros fornecedores de IA, que estão sendo adotados por vários estabelecimentos comerciais em todos os setores para uma ampla gama de tarefas de PNL (processamento de linguagem natural) posteriores. Conforme afirmam empresas de análise de pesquisa como a McKinsey: "O impacto da IA generativa na produtividade pode agregar trilhões de dólares em valor à economia global". Enquanto as empresas estão reinventando a IA como parceira de pensamento dos humanos e os FMs estão ampliando simultaneamente o que empresas e instituições podem fazer com IA generativa, as oportunidades para gerenciar grandes volumes de dados continuarão a crescer. Este documento apresenta informações introdutórias sobre IA generativa e os conceitos de design em relação aos recursos da NetApp que agregam valor aos clientes da NetApp , tanto em ambientes locais quanto híbridos ou multinuvem.

Então, o que os clientes ganham ao usar o NetApp em seus ambientes de IA?

A NetApp ajuda as organizações a lidar com as complexidades criadas pelo rápido crescimento de dados e nuvem, gerenciamento de várias nuvens e adoção de tecnologias de última geração, como IA. A NetApp combinou vários recursos em software de gerenciamento de dados inteligente e infraestrutura de armazenamento que foram bem equilibrados com alto desempenho otimizado para cargas de trabalho de IA. Soluções de IA generativas, como LLMs, precisam ler e processar seus conjuntos de dados de origem do armazenamento para a memória diversas vezes para promover a inteligência.

A NetApp é líder em tecnologias de mobilidade de dados, governança de dados e segurança de dados em todo o ecossistema da ponta ao núcleo e à nuvem, atendendo clientes empresariais que criam soluções de IA em escala. A NetApp, com uma forte rede de parceiros, tem ajudado diretores de dados, engenheiros de IA,

arquitetos corporativos e cientistas de dados no design de um pipeline de dados de fluxo livre para preparação de dados, proteção de dados e responsabilidades de gerenciamento estratégico de dados de treinamento e inferência de modelos de IA, otimizando o desempenho e a escalabilidade do ciclo de vida de IA/ML. As tecnologias e os recursos de dados da NetApp , como o NetApp ONTAP AI para pipeline de dados de aprendizado profundo, o NetApp SnapMirror para transportar dados de forma integrada e eficiente entre endpoints de armazenamento e o NetApp FlexCache para renderização em tempo real quando o fluxo de dados muda de lote para tempo real e a engenharia de dados acontece em tempo real, agregam valor à implantação de modelos de IA generativa em tempo real. À medida que empresas de todos os tipos adotam novas ferramentas de IA, elas enfrentam desafios de dados, da borda ao data center e à nuvem, que exigem soluções de IA escaláveis, responsáveis e explicáveis.

Como autoridade em dados em nuvem híbrida e múltipla, a NetApp está comprometida em construir uma rede de parceiros e soluções conjuntas que podem ajudar em todos os aspectos da construção de um pipeline de dados e data lakes para treinamento de modelos de IA generativos (pré-treinamento), ajuste fino, inferência baseada em contexto e monitoramento de decaimento de modelo de LLMs.

O que é IA Generativa?

A IA generativa está mudando a maneira como criamos conteúdo, geramos novos conceitos de design e exploramos novas composições. Ele ilustra estruturas de redes neurais como Rede Adversarial Generativa (GAN), Autocodificadores Variacionais (VAE) e Transformadores Pré-Treinados Generativos (GPT), que podem gerar novos conteúdos como texto, código, imagens, áudio, vídeo e dados sintéticos. Modelos baseados em transformadores, como o Chat-GPT da OpenAI, o Bard do Google, o BLOOM da Hugging Face e o LLaMA da Meta surgiram como a tecnologia fundamental que sustenta muitos avanços em grandes modelos de linguagem. Da mesma forma, o Dall-E da OpenAI, o CM3leon da Meta e o Imagen do Google são exemplos de modelos de difusão de texto para imagem que oferecem aos clientes um grau sem precedentes de fotorrealismo para criar imagens novas e complexas do zero ou editar imagens existentes para gerar imagens de alta qualidade com reconhecimento de contexto usando aumento de conjunto de dados e síntese de texto para imagem, vinculando semântica textual e visual. Artistas digitais estão começando a aplicar uma combinação de tecnologias de renderização como NeRF (Neural Radiance Field) com IA generativa para converter imagens 2D estáticas em cenas 3D imersivas. Em geral, os LLMs são amplamente caracterizados por quatro parâmetros: (1) Tamanho do modelo (normalmente em bilhões de parâmetros); (2) Tamanho do conjunto de dados de treinamento; (3) Custo do treinamento e (4) Desempenho do modelo após o treinamento. Os LLMs também se enquadram principalmente em três arquiteturas de transformadores. (i) Modelos somente com codificador. Ex.: BERT (Google, 2018); (ii) Modelos codificador-decodificador Ex.: BART (Meta, 2020) e (iii) Modelos somente decodificador. Por exemplo, LLaMA (Meta, 2023), PaLM-E (Google, 2023). Dependendo dos requisitos do negócio, independentemente da arquitetura que uma empresa escolher, o número de parâmetros do modelo (N) e o número de tokens (D) no conjunto de dados de treinamento geralmente determinam o custo base do treinamento (pré-treinamento) ou do ajuste fino de um LLM.

Casos de uso empresarial e tarefas de PNL posteriores

Empresas de todos os setores estão descobrindo cada vez mais o potencial da IA para extrair e produzir novas formas de valor de dados existentes para operações comerciais, vendas, marketing e serviços jurídicos. De acordo com a inteligência de mercado da IDC (International Data Corporation) sobre casos de uso e investimentos globais em IA generativa, a gestão do conhecimento no desenvolvimento de software e design de produtos será a mais impactada, seguida pela criação de histórias para marketing e geração de código para desenvolvedores. Na área da saúde, organizações de pesquisa clínica estão inovando na medicina. Modelos pré-treinados como o ProteinBERT incorporam anotações de Gene Ontology (GO) para projetar rapidamente estruturas de proteínas para medicamentos, representando um marco significativo na descoberta de medicamentos, bioinformática e biologia molecular. Empresas de biotecnologia iniciaram testes em humanos para medicamentos descobertos por IA generativa, que visam tratar doenças como fibrose pulmonar (FPI), uma doença pulmonar que causa cicatrizes irreversíveis no tecido pulmonar.

Figura 1: Casos de uso que impulsionam a IA generativa

[Figura 1: Casos de uso que impulsionam a IA generativa]

O aumento na adoção da automação impulsionado pela IA generativa também está mudando a oferta e a demanda de atividades de trabalho para muitas ocupações. De acordo com a McKinsey, o mercado de trabalho dos EUA (diagrama abaixo) passou por uma rápida transição, que pode continuar apenas quando se considera o impacto da IA.

Fonte: McKinsey & Company

[Figura 2: Fonte: McKinsey Company]

Papel do armazenamento na IA generativa

Os LLMs dependem amplamente de aprendizado profundo, GPUs e computação. Entretanto, quando o buffer da GPU fica cheio, os dados precisam ser gravados rapidamente no armazenamento. Embora alguns modelos de IA sejam pequenos o suficiente para serem executados na memória, os LLMs exigem alto IOPS e armazenamento de alto rendimento para fornecer acesso rápido a grandes conjuntos de dados, especialmente se envolver bilhões de tokens ou milhões de imagens. Para um requisito típico de memória de GPU de um LLM, a memória necessária para treinar um modelo com 1 bilhão de parâmetros pode chegar a 80 GB com precisão total de 32 bits. Nesse caso, o LLaMA 2 da Meta, uma família de LLMs que varia em escala de 7 bilhões a 70 bilhões de parâmetros, pode exigir 70x80, aproximadamente 5.600 GB ou 5,6 TB de RAM de GPU. Além disso, a quantidade de memória necessária é diretamente proporcional ao número máximo de tokens que você deseja gerar. Por exemplo, se você deseja gerar saídas de até 512 tokens (cerca de 380 palavras), você precisa"512 MB" . Pode parecer insignificante, mas, se você quiser executar lotes maiores, isso começa a somar. Portanto, torna-se muito caro para organizações treinar ou ajustar LLMs na memória, tornando o armazenamento uma base fundamental para a IA generativa.

Três abordagens principais para LLMs

Para a maioria das empresas, com base nas tendências atuais, a abordagem para implantar LLMs pode ser condensada em três cenários básicos. Conforme descrito em um recente"Harvard Business Review" Artigo: (1) Treinamento (pré-treinamento) de um LLM do zero – caro e requer habilidades especializadas em IA/ML; (2) Ajuste fino de um modelo de base com dados empresariais – complexo, mas viável; (3) Uso de geração aumentada de recuperação (RAG) para consultar repositórios de documentos, APIs e bancos de dados vetoriais que contêm dados da empresa. Cada um deles tem compensações entre esforço, velocidade de iteração, custo-benefício e precisão do modelo em suas implementações, usadas para resolver diferentes tipos de problemas (diagrama abaixo).

Figura 3: Tipos de problemas

[Figura 3: Tipos de problemas]

Modelos de Fundação

Um modelo de fundação (FM), também conhecido como modelo base, é um grande modelo de IA (LLM) treinado em grandes quantidades de dados não rotulados, usando autossupervisão em escala, geralmente adaptado para uma ampla gama de tarefas de PNL posteriores. Como os dados de treinamento não são rotulados por humanos, o modelo emerge em vez de ser codificado explicitamente. Isso significa que o modelo pode gerar histórias ou uma narrativa própria sem ser explicitamente programado para isso. Portanto, uma característica importante da FM é a homogeneização, o que significa que o mesmo método é usado em muitos domínios. No entanto, com técnicas de personalização e ajuste fino, os FMs integrados aos produtos que aparecem hoje em dia não são bons apenas para gerar texto, texto para imagens e texto para código,

mas também para explicar tarefas específicas de domínio ou depurar código. Por exemplo, FMs como o Codex da OpenAI ou o Code Llama da Meta podem gerar código em várias linguagens de programação com base em descrições em linguagem natural de uma tarefa de programação. Esses modelos são proficientes em mais de uma dúzia de linguagens de programação, incluindo Python, C#, JavaScript, Perl, Ruby e SQL. Eles entendem a intenção do usuário e geram código específico que realiza a tarefa desejada, útil para desenvolvimento de software, otimização de código e automação de tarefas de programação.

Ajuste fino, especificidade de domínio e retreinamento

Uma das práticas comuns na implantação do LLM após a preparação e o pré-processamento de dados é selecionar um modelo pré-treinado que foi treinado em um conjunto de dados grande e diversificado. No contexto de ajustes finos, isso pode ser um modelo de linguagem de código aberto de grande porte, como… "Lhama de Meta 2" Treinado com 70 bilhões de parâmetros e 2 trilhões de tokens. Depois que o modelo pré-treinado for selecionado, o próximo passo é ajustá-lo nos dados específicos do domínio. Isso envolve ajustar os parâmetros do modelo e treiná-lo com os novos dados para se adaptar a um domínio e tarefa específicos. Por exemplo, a BloombergGPT, uma LLM proprietária treinada em uma ampla gama de dados financeiros que atendem ao setor financeiro.

Modelos específicos de domínio projetados e treinados para uma tarefa específica geralmente têm maior precisão e desempenho dentro de seu escopo, mas baixa transferibilidade entre outras tarefas ou domínios. Quando o ambiente de negócios e os dados mudam ao longo de um período, a precisão da previsão do FM pode começar a diminuir quando comparada ao seu desempenho durante os testes. É nesse momento que o retreinamento ou ajuste fino do modelo se torna crucial.

O retreinamento de modelo em IA/ML tradicional refere-se à atualização de um modelo de ML implantado com novos dados, geralmente realizado para eliminar dois tipos de desvios que ocorrem. (1) Desvio de conceito – quando a ligação entre as variáveis de entrada e as variáveis de destino muda ao longo do tempo, uma vez que a descrição do que queremos prever muda, o modelo pode produzir previsões imprecisas. (2) Desvio de dados – ocorre quando as características dos dados de entrada mudam, como mudanças nos hábitos ou comportamento do cliente ao longo do tempo e, portanto, a incapacidade do modelo de responder a tais mudanças.

De forma semelhante, a reciclagem se aplica a FMs/LLMs, porém pode ser muito mais custosa (em milhões de dólares), portanto, não é algo que a maioria das organizações consideraria. Está sob pesquisa ativa, ainda emergindo no campo do LLMOps. Então, em vez de retreinar, quando ocorre decadência do modelo em FMs ajustados, as empresas podem optar por fazer um novo ajuste fino (muito mais barato) com um conjunto de dados mais novo. Para uma perspectiva de custo, listamos abaixo um exemplo de uma tabela de preços de modelo do Azure-OpenAI Services. Para cada categoria de tarefa, os clientes podem ajustar e avaliar modelos em conjuntos de dados específicos.

Fonte: Microsoft Azure

[Fonte: Microsoft Azure]

Engenharia rápida e inferência

Engenharia rápida refere-se aos métodos eficazes de como se comunicar com LLMs para executar tarefas desejadas sem atualizar os pesos do modelo. Tão importante quanto o treinamento e o ajuste fino do modelo de IA para aplicações de PNL, a inferência é igualmente importante, onde os modelos treinados respondem às solicitações do usuário. Os requisitos do sistema para inferência geralmente dependem muito mais do desempenho de leitura do sistema de armazenamento de IA que alimenta dados de LLMs para as GPUs, pois ele precisa ser capaz de aplicar bilhões de parâmetros de modelo armazenados para produzir a melhor resposta.

LLMOps, Monitoramento de Modelos e Vectorstores

Assim como as operações de aprendizado de máquina tradicionais (MLOps), as operações de modelos de grande linguagem (LLMOps) também exigem a colaboração de cientistas de dados e engenheiros de DevOps com ferramentas e práticas recomendadas para o gerenciamento de LLMs em ambientes de produção. No entanto, o fluxo de trabalho e a pilha de tecnologia para LLMs podem variar de algumas maneiras. Por exemplo, pipelines LLM criados usando estruturas como LangChain encadeiam diversas chamadas de API LLM para endpoints de incorporação externos, como vectorstores ou bancos de dados de vetores. O uso de um endpoint de incorporação e um vectorstore para conectores downstream (como um banco de dados vetorial) representa um desenvolvimento significativo na forma como os dados são armazenados e acessados. Ao contrário dos modelos tradicionais de ML que são desenvolvidos do zero, os LLMs geralmente dependem da aprendizagem por transferência, pois esses modelos começam com FMs que são ajustados com novos dados para melhorar o desempenho em um domínio mais específico. Portanto, é crucial que o LLMOps forneça recursos de gerenciamento de risco e monitoramento de decaimento do modelo.

Riscos e Ética na era da IA Generativa

"ChatGPT – É inteligente, mas ainda fala bobagens." – MIT Tech Review. Lixo que entra e lixo que sai sempre foi um caso desafiador na computação. A única diferença com a IA generativa é que ela se destaca em tornar o lixo altamente confiável, levando a resultados imprecisos. Os LLMs tendem a inventar fatos para se adequar à narrativa que estão construindo. Portanto, as empresas que veem a IA generativa como uma grande oportunidade para reduzir seus custos com equivalentes de IA precisam detectar falsificações profundas com eficiência, reduzir vieses e diminuir riscos para manter os sistemas honestos e éticos. Um pipeline de dados de fluxo livre com uma infraestrutura de IA robusta que ofereça suporte à mobilidade de dados, qualidade de dados, governança de dados e proteção de dados por meio de criptografia de ponta a ponta e proteções de IA é essencial no design de modelos de IA generativos responsáveis e explicáveis.

Cenário do cliente e NetApp

Figura 3: Fluxo de trabalho do modelo de aprendizado de máquina/linguagem grande

[Figura 3: Fluxo de trabalho do modelo de aprendizado de máquina/linguagem grande]

Estamos treinando ou ajustando? A questão de (a) treinar um modelo LLM do zero, ajustar um FM pré-treinado ou usar o RAG para recuperar dados de repositórios de documentos fora de um modelo de base e aumentar os prompts e (b) aproveitar LLMs de código aberto (por exemplo, Llama 2) ou FMs proprietários (por exemplo, ChatGPT, Bard, AWS Bedrock) é uma decisão estratégica para as organizações. Cada abordagem tem um equilíbrio entre custo-benefício, gravidade dos dados, operações, precisão do modelo e gerenciamento de LLMs.

A NetApp , como empresa, adota a IA internamente em sua cultura de trabalho e em sua abordagem aos esforços de design e engenharia de produtos. Por exemplo, a proteção autônoma contra ransomware da NetApp é criada usando IA e aprendizado de máquina. Ele fornece detecção antecipada de anomalias no sistema de arquivos para ajudar a identificar ameaças antes que elas afetem as operações. Em segundo lugar, a NetApp usa IA preditiva para suas operações comerciais, como previsão de vendas e estoque, e chatbots para auxiliar os clientes em serviços de suporte a produtos de call center, especificações técnicas, garantia, manuais de serviço e muito mais. Em terceiro lugar, a NetApp agrega valor ao cliente para o pipeline de dados de IA e o fluxo de trabalho de ML/LLM por meio de produtos e soluções que atendem clientes que criam soluções de IA preditivas, como previsão de demanda, imagens médicas, análise de sentimentos e soluções de IA generativas, como GANs para detecção de anomalias em imagens industriais no setor de manufatura e combate à lavagem de dinheiro e detecção de fraudes em serviços bancários e financeiros com produtos e recursos da NetApp , como NetApp ONTAP AI, NetApp SnapMirror e NetApp FlexCache.

Recursos do NetApp

A movimentação e o gerenciamento de dados em aplicativos de IA generativa, como chatbot, geração de código, geração de imagens ou expressão de modelo de genoma, podem abranger a borda, o data center privado e o ecossistema multicloud híbrido. Por exemplo, um bot de IA em tempo real que ajuda um passageiro a atualizar sua passagem aérea para a classe executiva a partir de um aplicativo de usuário final exposto por meio de APIs de modelos pré-treinados, como o ChatGPT, não consegue realizar essa tarefa sozinho, pois as informações do passageiro não estão disponíveis publicamente na internet. A API requer acesso às informações pessoais do passageiro e às informações da passagem da companhia aérea, que podem existir em um ecossistema híbrido ou multinuvem. Um cenário semelhante pode se aplicar a cientistas que compartilham uma molécula de medicamento e dados de pacientes por meio de um aplicativo de usuário final que usa LLMs para realizar testes clínicos em descobertas de medicamentos envolvendo instituições de pesquisa biomédica de um para muitos. Dados confidenciais que são passados para FMs ou LLMs podem incluir PII, informações financeiras, informações de saúde, dados biométricos, dados de localização, dados de comunicação, comportamento online e informações legais. Em um evento de renderização em tempo real, execução rápida e inferência de borda, há movimentação de dados do aplicativo do usuário final para endpoints de armazenamento por meio de modelos LLM proprietários ou de código aberto para um data center local ou plataformas de nuvem pública. Em todos esses cenários, a mobilidade e a proteção de dados são cruciais para as operações de IA envolvendo LLMs que dependem de grandes conjuntos de dados de treinamento e movimentação desses dados.

Figura 4: Pipeline de dados de IA generativa - LLM

[Figura 4: Pipeline de dados generativos de IA-LLM]

O portfólio de infraestrutura de armazenamento, dados e serviços de nuvem da NetApp é alimentado por software de gerenciamento de dados inteligente.

Preparação de dados: O primeiro pilar da pilha de tecnologia do LLM permanece praticamente intocado pela pilha de ML tradicional mais antiga. O pré-processamento de dados no pipeline de IA é necessário para normalizar e limpar os dados antes do treinamento ou ajuste fino. Esta etapa inclui conectores para ingerir dados onde quer que eles residam na forma de uma camada do Amazon S3 ou em sistemas de armazenamento locais, como um armazenamento de arquivos ou um armazenamento de objetos como o NetApp StorageGRID.

- NetApp ONTAP* é a tecnologia fundamental que sustenta as soluções de armazenamento crítico da NetApp no data center e na nuvem. O ONTAP inclui vários recursos e capacidades de gerenciamento e proteção de dados, incluindo proteção automática contra ransomware contra ataques cibernéticos, recursos integrados de transporte de dados e capacidades de eficiência de armazenamento para uma variedade de arquiteturas, desde locais, híbridas, multiclouds em NAS, SAN, objetos e situações de armazenamento definido por software (SDS) de implantações de LLM.
- NetApp ONTAP AI* para treinamento de modelos de aprendizado profundo. O NetApp ONTAP oferece suporte ao NVIDIA GPU Direct Storage com o uso de NFS sobre RDMA para clientes NetApp com cluster de armazenamento ONTAP e nós de computação NVIDIA DGX. Ele oferece um desempenho econômico para ler e processar conjuntos de dados de origem do armazenamento para a memória diversas vezes para promover a inteligência, permitindo que as organizações tenham acesso a treinamento, ajuste fino e dimensionamento para LLMs.
- NetApp FlexCache* é um recurso de cache remoto que simplifica a distribuição de arquivos e armazena em cache apenas os dados lidos ativamente. Isso pode ser útil para treinamento, reciclagem e ajuste fino de LLM, agregando valor aos clientes com requisitos de negócios, como renderização em tempo real e inferência de LLM.
- NetApp SnapMirror* é um recurso ONTAP que replica instantâneos de volume entre quaisquer dois sistemas ONTAP . Esse recurso transfere dados de ponta de forma otimizada para seu data center local

ou para a nuvem. O SnapMirror pode ser usado para mover dados de forma segura e eficiente entre nuvens locais e hiperescaláveis, quando os clientes desejam desenvolver IA generativa em nuvens com RAG contendo dados empresariais. Ele transfere com eficiência apenas as alterações, economizando largura de banda e acelerando a replicação, trazendo, assim, recursos essenciais de mobilidade de dados durante as operações de treinamento, retreinamento e ajuste fino de FMs ou LLMs.

- O NetApp SnapLock* traz capacidade de disco imutável em sistemas de armazenamento baseados em ONTAP para controle de versão de conjuntos de dados. A arquitetura microcore foi projetada para proteger os dados do cliente com o mecanismo FPolicy Zero Trust. A NetApp garante que os dados do cliente estejam disponíveis resistindo a ataques de negação de serviço (DoS) quando um invasor interage com um LLM de uma forma que consome muitos recursos.
- O NetApp Cloud Data Sense* ajuda a identificar, mapear e classificar informações pessoais presentes em conjuntos de dados empresariais, promulgar políticas, atender a requisitos de privacidade no local ou na nuvem, ajudar a melhorar a postura de segurança e cumprir regulamentações.
- Classificação NetApp BlueXP*, fornecida pelo Cloud Data Sense. Os clientes podem escanear, analisar, categorizar e agir automaticamente sobre dados em todo o acervo de dados, detectar riscos de segurança, otimizar o armazenamento e acelerar implantações na nuvem. Ele combina serviços de armazenamento e dados por meio de seu plano de controle unificado. Os clientes podem usar instâncias de GPU para computação e ambientes multicloud híbridos para camadas de armazenamento a frio e para arquivos e backups.
- Dualidade arquivo-objeto do NetApp *. O NetApp ONTAP permite acesso de protocolo duplo para NFS e S3. Com esta solução, os clientes podem acessar dados NFS de notebooks Amazon AWS SageMaker por meio de buckets S3 do NetApp Cloud Volumes ONTAP. Isso oferece flexibilidade aos clientes que precisam de acesso fácil a fontes de dados heterogêneas com a capacidade de compartilhar dados do NFS e do S3. Por exemplo, ajuste fino de FMs como os modelos de geração de texto Llama 2 da Meta no SageMaker com acesso a buckets de arquivo-objeto.

O serviço * NetApp Cloud Sync* oferece uma maneira simples e segura de migrar dados para qualquer destino, na nuvem ou no local. O Cloud Sync transfere e sincroniza dados perfeitamente entre armazenamento local ou em nuvem, NAS e armazenamentos de objetos.

- NetApp XCP* é um software cliente que permite migrações de dados rápidas e confiáveis de qualquer para NetApp e de NetApp para NetApp . O XCP também oferece a capacidade de mover dados em massa de forma eficiente dos sistemas de arquivos Hadoop HDFS para o ONTAP NFS, S3 ou StorageGRID, e a análise de arquivos do XCP fornece visibilidade do sistema de arquivos.
- NetApp DataOps Toolkit* é uma biblioteca Python que simplifica para cientistas de dados, DevOps e engenheiros de dados a execução de diversas tarefas de gerenciamento de dados, como provisionamento, clonagem ou criação de snapshots quase instantâneos de um volume de dados ou espaço de trabalho do JupyterLab, apoiados por armazenamento NetApp de alto desempenho escalável.

Segurança do produto da NetApp. Os LLMs podem inadvertidamente revelar dados confidenciais em suas respostas, o que é uma preocupação para os CISOs que estudam as vulnerabilidades associadas a aplicativos de IA que utilizam LLMs. Conforme descrito pelo OWASP (Open Worldwide Application Security Project), problemas de segurança como envenenamento de dados, vazamento de dados, negação de serviço e injeções imediatas em LLMs podem impactar empresas, desde a exposição de dados até o acesso não autorizado por invasores. Os requisitos de armazenamento de dados devem incluir verificações de integridade e instantâneos imutáveis para dados estruturados, semiestruturados e não estruturados. NetApp Snapshots e SnapLock estão sendo usados para controle de versão de conjuntos de dados. Ele traz controle de acesso rigoroso baseado em funções (RBAC), bem como protocolos seguros e criptografia padrão do setor para proteger dados em repouso e em trânsito. O Cloud Insights e o Cloud Data Sense juntos oferecem recursos para ajudar você a identificar forensemente a origem da ameaça e priorizar quais dados restaurar.

* ONTAP AI com DGX BasePOD*

A arquitetura de referência de IA do NetApp ONTAP com NVIDIA DGX BasePOD é uma arquitetura escalável para cargas de trabalho de aprendizado de máquina (ML) e inteligência artificial (IA). Para a fase crítica de treinamento dos LLMs, os dados normalmente são copiados do armazenamento de dados para o cluster de treinamento em intervalos regulares. Os servidores usados nesta fase usam GPUs para paralelizar cálculos, criando um enorme apetite por dados. Atender às necessidades de largura de banda de E/S bruta é crucial para manter alta utilização da GPU.

* ONTAP AI com NVIDIA AI Enterprise*

O NVIDIA AI Enterprise é um conjunto completo e nativo em nuvem de software de IA e análise de dados que é otimizado, certificado e suportado pela NVIDIA para ser executado no VMware vSphere com sistemas certificados pela NVIDIA. Este software facilita a implantação, o gerenciamento e o dimensionamento simples e rápidos de cargas de trabalho de IA no ambiente de nuvem híbrida moderno. O NVIDIA AI Enterprise, com tecnologia NetApp e VMware, oferece carga de trabalho de IA de nível empresarial e gerenciamento de dados em um pacote simplificado e familiar.

Plataformas em Nuvem 1P

As ofertas de armazenamento em nuvem totalmente gerenciadas estão disponíveis nativamente no Microsoft Azure como Azure NetApp Files (ANF), na AWS como Amazon FSx for NetApp ONTAP (FSx ONTAP) e no Google como Google Cloud NetApp Volumes (GNCV). 1P é um sistema de arquivos gerenciado e de alto desempenho que permite aos clientes executar cargas de trabalho de IA de alta disponibilidade com segurança de dados aprimorada em nuvens públicas, para ajustar LLMs/FMs com plataformas de ML nativas da nuvem, como AWS SageMaker, Azure-OpenAI Services e Vertex AI do Google.

Conjunto de soluções para parceiros da NetApp

Além de seus principais produtos de dados, tecnologias e recursos, a NetApp também colabora estreitamente com uma rede robusta de parceiros de IA para agregar valor aos clientes.

- Os Guardrails da NVIDIA * em sistemas de IA servem como salvaguardas para garantir o uso ético e responsável das tecnologias de IA. Os desenvolvedores de IA podem escolher definir o comportamento de aplicativos com tecnologia LLM em tópicos específicos e impedi-los de se envolver em discussões sobre tópicos indesejados. Guardrails, um kit de ferramentas de código aberto, oferece a capacidade de conectar um LLM a outros serviços de forma integrada e segura para criar sistemas de conversação LLM confiáveis, seguros e protegidos.

Domino Data Lab fornece ferramentas versáteis e de nível empresarial para criar e produzir IA generativa - rápida, segura e econômica, onde quer que você esteja em sua jornada de IA. Com a plataforma Enterprise MLOps da Domino’s, os cientistas de dados podem usar as ferramentas preferidas e todos os seus dados, treinar e implantar modelos facilmente em qualquer lugar e gerenciar riscos e custos de forma eficaz — tudo a partir de um único centro de controle.

Modzy para Edge AI. A NetApp e a Modzy fizeram uma parceria para fornecer IA em escala para qualquer tipo de dado, incluindo imagens, áudio, texto e tabelas. Modzy é uma plataforma MLOps para implantação, integração e execução de modelos de IA, oferecendo aos cientistas de dados recursos de monitoramento de modelos, detecção de desvios e explicabilidade, com uma solução integrada para inferência LLM perfeita.

A **Run:AI** e a NetApp fizeram uma parceria para demonstrar os recursos exclusivos da solução NetApp ONTAP AI com a plataforma de gerenciamento de cluster Run:AI para simplificar a orquestração de cargas de trabalho de IA. Ele divide e une automaticamente os recursos da GPU, projetados para dimensionar seus pipelines de processamento de dados para centenas de máquinas com estruturas de integração integradas

para Spark, Ray, Dask e Rapids.

Conclusão

A IA generativa pode produzir resultados eficazes somente quando o modelo é treinado em grandes quantidades de dados de qualidade. Embora os LLMs tenham alcançado marcos notáveis, é fundamental reconhecer suas limitações, desafios de design e riscos associados à mobilidade e à qualidade dos dados. Os LLMs dependem de conjuntos de dados de treinamento grandes e díspares de fontes de dados heterogêneas. Resultados imprecisos ou tendenciosos gerados pelos modelos podem colocar empresas e consumidores em risco. Esses riscos podem corresponder a restrições para LLMs que surgem potencialmente de desafios de gerenciamento de dados associados à qualidade de dados, segurança de dados e mobilidade de dados. A NetApp ajuda as organizações a lidar com as complexidades criadas pelo rápido crescimento de dados, mobilidade de dados, gerenciamento de várias nuvens e adoção de IA. A infraestrutura de IA em escala e o gerenciamento eficiente de dados são cruciais para definir o sucesso de aplicações de IA como a IA generativa. É essencial que os clientes cubram todos os cenários de implantação sem comprometer a capacidade de expansão conforme necessário pelas empresas, mantendo a eficiência de custos, a governança de dados e as práticas éticas de IA sob controle. A NetApp trabalha constantemente para ajudar os clientes a simplificar e acelerar suas implantações de IA.

Informações sobre direitos autorais

Informações sobre marcas comerciais