IA générative et valeur NetApp

NetApp artificial intelligence solutions

NetApp
August 18, 2025

Sommaire

IA générative et valeur NetApp

La demande d'intelligence artificielle générative (IA) entraîne des bouleversements dans tous les secteurs, améliorant la créativité des entreprises et l'innovation des produits.

Abstrait

De nombreuses organisations utilisent l'IA générative pour créer de nouvelles fonctionnalités de produits, améliorer la productivité de l'ingénierie et prototyper des applications basées sur l'IA qui offrent de meilleurs résultats et expériences client. L'IA générative telle que les transformateurs pré-entraînés génératifs (GPT) utilise des réseaux neuronaux pour créer de nouveaux contenus, aussi divers que du texte, de l'audio et de la vidéo. Compte tenu de l'échelle extrême et des ensembles de données massifs impliqués dans les grands modèles linguistiques (LLM), il est essentiel de concevoir une infrastructure d'IA robuste qui tire parti des fonctionnalités de stockage de données convaincantes des options de déploiement sur site, hybrides et multicloud et réduit les risques associés à la mobilité des données, à la protection des données et à la gouvernance avant que les entreprises puissent concevoir des solutions d'IA. Cet article décrit ces considérations et les fonctionnalités NetApp AI correspondantes qui permettent une gestion transparente des données et un déplacement des données à travers le pipeline de données IA pour la formation, le recyclage, le réglage fin et l'inférence des modèles d'IA génératifs.

Résumé exécutif

Plus récemment, après le lancement de ChatGPT, un spin-off de GPT-3 en novembre 2022, de nouveaux outils d'IA utilisés pour générer du texte, du code, des images ou même des protéines thérapeutiques en réponse aux invites des utilisateurs ont acquis une renommée significative. Cela indique que les utilisateurs peuvent faire une demande en utilisant le langage naturel et que l'IA interprétera et générera du texte, tel que des articles d'actualité ou des descriptions de produits qui reflètent la demande de l'utilisateur ou produiront du code, de la musique, de la parole, des effets visuels et des ressources 3D à l'aide d'algorithmes formés sur des données déjà existantes. En conséquence, des expressions telles que « diffusion stable », « hallucinations », « ingénierie rapide » et « alignement des valeurs » apparaissent rapidement dans la conception des systèmes d'IA. Ces modèles d'apprentissage automatique (ML) auto-supervisés ou semi-supervisés deviennent largement disponibles en tant que modèles de base pré-entraînés (FM) via des fournisseurs de services cloud et d'autres fournisseurs d'IA, qui sont adoptés par divers établissements commerciaux dans tous les secteurs pour une large gamme de tâches de traitement du langage naturel (NLP) en aval. Comme l'affirment des cabinets d'analyse de recherche comme McKinsey : « L'impact de l'IA générative sur la productivité pourrait ajouter des milliers de milliards de dollars de valeur à l'économie mondiale. » Alors que les entreprises réinventent l'IA en tant que partenaires de réflexion des humains et que les FM élargissent simultanément ce que les entreprises et les institutions peuvent faire avec l'IA générative, les opportunités de gérer des volumes massifs de données continueront de croître. Ce document présente des informations introductives sur l'IA générative et les concepts de conception relatifs aux fonctionnalités NetApp qui apportent de la valeur aux clients NetApp , à la fois sur site et dans des environnements hybrides ou multicloud.

Alors, quel est l'intérêt pour les clients d'utiliser NetApp dans leurs environnements d'IA ? NetApp aide les organisations à faire face aux complexités créées par la croissance rapide des données et du cloud, la gestion multicloud et l'adoption de technologies de nouvelle génération, telles que l'IA. NetApp a combiné diverses fonctionnalités dans un logiciel de gestion de données intelligent et une infrastructure de stockage qui ont été bien équilibrés avec des performances élevées optimisées pour les charges de travail d'IA. Les solutions d'IA générative telles que les LLM doivent lire et traiter leurs ensembles de données sources du stockage vers la mémoire à plusieurs reprises pour favoriser l'intelligence. NetApp est un leader en matière de mobilité des données, de gouvernance des données et de technologies de sécurité des données dans l'écosystème Edge-to-Core-to-Cloud, permettant aux clients d'entreprise de créer des solutions d'IA à grande échelle. NetApp, avec un solide réseau de partenaires, aide les directeurs des données, les ingénieurs en IA,

les architectes d'entreprise et les scientifiques des données à concevoir un pipeline de données fluide pour la préparation des données, la protection des données et les responsabilités de gestion stratégique des données de la formation et de l'inférence des modèles d'IA, optimisant ainsi les performances et l'évolutivité du cycle de vie de l'IA/ML. Les technologies et fonctionnalités de données NetApp telles que NetApp ONTAP AI pour le pipeline de données d'apprentissage en profondeur, NetApp SnapMirror pour le transport transparent et efficace des données entre les points de terminaison de stockage et NetApp FlexCache pour le rendu en temps réel lorsque le flux de données passe du lot au temps réel et que l'ingénierie des données se produit à un moment opportun, apportent de la valeur au déploiement de modèles d'IA générative en temps réel. Alors que les entreprises de tous types adoptent de nouveaux outils d'IA, elles sont confrontées à des défis en matière de données, de la périphérie au centre de données en passant par le cloud, qui exigent des solutions d'IA évolutives, responsables et explicables. En tant qu'autorité en matière de données sur le cloud hybride et multi-cloud, NetApp s'engage à créer un réseau de partenaires et de solutions communes qui peuvent aider dans tous les aspects de la construction d'un pipeline de données et de lacs de données pour la formation de modèles d'IA génératifs (pré-formation), le réglage fin, l'inférence basée sur le contexte et la surveillance de la dégradation des modèles des LLM.

Qu'est-ce que l'IA générative ?

L'IA générative change la façon dont nous créons du contenu, générons de nouveaux concepts de conception et explorons de nouvelles compositions. Il illustre les cadres de réseaux neuronaux tels que le réseau antagoniste génératif (GAN), les autoencodeurs variationnels (VAE) et les transformateurs pré-entraînés génératifs (GPT), qui peuvent générer du nouveau contenu comme du texte, du code, des images, de l'audio, de la vidéo et des données synthétiques. Les modèles basés sur des transformateurs tels que Chat-GPT d'OpenAI, Bard de Google, BLOOM de Hugging Face et LLaMA de Meta sont devenus la technologie fondamentale qui sous-tend de nombreuses avancées dans les grands modèles linguistiques. De même, Dall-E d'OpenAI, CM3leon de Meta et Imagen de Google sont des exemples de modèles de diffusion de texte en image qui offrent aux clients un degré de photoréalisme sans précédent pour créer de nouvelles images complexes à partir de zéro ou modifier des images existantes pour générer des images contextuelles de haute qualité en utilisant l'augmentation de l'ensemble de données et la synthèse de texte en image reliant la sémantique textuelle et visuelle. Les artistes numériques commencent à appliquer une combinaison de technologies de rendu comme NeRF (Neural Radiance Field) avec l'IA générative pour convertir des images 2D statiques en scènes 3D immersives. En général, les LLM sont largement caractérisés par quatre paramètres : (1) la taille du modèle (généralement en milliards de paramètres) ; (2) la taille de l'ensemble de données de formation ; (3) le coût de la formation et (4) les performances du modèle après la formation. Les LLM se répartissent également principalement en trois architectures de transformateurs. (i) Modèles à encodeur uniquement. Par exemple BERT (Google, 2018) ; (ii) Modèles encodeur-décodeur, par exemple BART (Meta, 2020) et (iii) Modèles décodeur uniquement. Par exemple LLaMA (Meta, 2023), PaLM-E (Google, 2023). En fonction des besoins de l'entreprise, quelle que soit l'architecture choisie par l'entreprise, le nombre de paramètres de modèle (N) et le nombre de jetons (D) dans l'ensemble de données de formation déterminent généralement le coût de base de la formation (pré-formation) ou du réglage fin d'un LLM.

Cas d'utilisation en entreprise et tâches NLP en aval

Les entreprises de tous les secteurs découvrent de plus en plus le potentiel de l'IA pour extraire et produire de nouvelles formes de valeur à partir de données existantes pour les opérations commerciales, les ventes, le marketing et les services juridiques. Selon les informations de marché d'IDC (International Data Corporation) sur les cas d'utilisation et les investissements mondiaux en matière d'IA générative, la gestion des connaissances dans le développement de logiciels et la conception de produits sera la plus impactée, suivie de la création de scénarios pour le marketing et de la génération de code pour les développeurs. Dans le domaine de la santé, les organismes de recherche clinique innovent en médecine. Les modèles pré-entraînés comme ProteinBERT intègrent des annotations Gene Ontology (GO) pour concevoir rapidement des structures protéiques pour les médicaments médicaux, ce qui représente une étape importante dans la découverte de médicaments, la bioinformatique et la biologie moléculaire. Des entreprises de biotechnologie ont lancé des essais sur l'homme pour la médecine générative découverte par l'IA, qui vise à traiter des maladies comme la

fibrose pulmonaire (FPI), une maladie pulmonaire qui provoque des cicatrices irréversibles du tissu pulmonaire.

Figure 1 : Cas d'utilisation de l'IA générative

[Figure 1 : Cas d'utilisation de l'IA générative]

L'augmentation de l'adoption de l'automatisation, induite par l'IA générative, modifie également l'offre et la demande d'activités professionnelles pour de nombreuses professions. Selon McKinsey, le marché du travail américain (diagramme ci-dessous) a connu une transition rapide, qui ne peut que se poursuivre si l'on prend en compte l'impact de l'IA.

Source : McKinsey & Company

[Figure 2 : Source : McKinsey Company]

Rôle du stockage dans l'IA générative

Les LLM s'appuient en grande partie sur l'apprentissage profond, les GPU et le calcul. Cependant, lorsque la mémoire tampon du GPU est pleine, les données doivent être écrites rapidement sur le stockage. Alors que certains modèles d'IA sont suffisamment petits pour s'exécuter en mémoire, les LLM nécessitent des IOPS élevées et un stockage à haut débit pour fournir un accès rapide à de grands ensembles de données, en particulier s'il s'agit de milliards de jetons ou de millions d'images. Pour une exigence de mémoire GPU typique d'un LLM, la mémoire nécessaire pour entraîner un modèle avec 1 milliard de paramètres pourrait aller jusqu'à 80 Go avec une précision totale de 32 bits. Dans ce cas, le LLaMA 2 de Meta, une famille de LLM dont l'échelle varie de 7 à 70 milliards de paramètres, peut nécessiter 70x80, environ 5600 Go ou 5,6 To de RAM GPU. De plus, la quantité de mémoire dont vous avez besoin est directement proportionnelle au nombre maximal de jetons que vous souhaitez générer. Par exemple, si vous souhaitez générer des sorties allant jusqu'à 512 jetons (environ 380 mots), vous avez besoin"512 Mo" . Cela peut sembler sans importance, mais si vous souhaitez exécuter des lots plus importants, cela commence à s'accumuler. Par conséquent, il est très coûteux pour les organisations de former ou d'affiner les LLM en mémoire, faisant ainsi du stockage une pierre angulaire de l'IA générative.

Trois approches principales pour les LLM

Pour la plupart des entreprises, sur la base des tendances actuelles, l'approche de déploiement des LLM peut être condensée en 3 scénarios de base. Comme décrit dans un récent"Harvard Business Review" article : (1) Former (pré-former) un LLM à partir de zéro – coûteux et nécessite des compétences expertes en IA/ML ; (2) Ajuster un modèle de base avec des données d'entreprise – complexe, mais faisable ; (3) Utiliser la génération augmentée par récupération (RAG) pour interroger les référentiels de documents, les API et les bases de données vectorielles qui contiennent des données d'entreprise. Chacune d'entre elles nécessite des compromis entre l'effort, la vitesse d'itération, la rentabilité et la précision du modèle dans leurs implémentations, utilisées pour résoudre différents types de problèmes (diagramme ci-dessous).

Figure 3 : Types de problèmes

[Figure 3 : Types de problèmes]

Modèles de fondation

Un modèle de fondation (FM), également connu sous le nom de modèle de base, est un grand modèle d'IA (LLM) formé sur de grandes quantités de données non étiquetées, utilisant l'auto-supervision à grande échelle, généralement adapté à une large gamme de tâches NLP en aval. Étant donné que les données de formation ne sont pas étiquetées par des humains, le modèle émerge plutôt que d'être explicitement codé. Cela signifie

que le modèle peut générer ses propres histoires ou son propre récit sans être explicitement programmé pour le faire. Une caractéristique importante de la FM est donc l'homogénéisation, ce qui signifie que la même méthode est utilisée dans de nombreux domaines. Cependant, grâce aux techniques de personnalisation et de réglage fin, les FM intégrés aux produits apparaissant de nos jours sont non seulement efficaces pour générer du texte, du texte en images et du texte en code, mais également pour expliquer des tâches spécifiques à un domaine ou déboguer du code. Par exemple, des FM comme Codex d'OpenAI ou Code Llama de Meta peuvent générer du code dans plusieurs langages de programmation en fonction des descriptions en langage naturel d'une tâche de programmation. Ces modèles maîtrisent plus d'une douzaine de langages de programmation, notamment Python, C#, JavaScript, Perl, Ruby et SQL. Ils comprennent l'intention de l'utilisateur et génèrent un code spécifique qui accomplit la tâche souhaitée, utile pour le développement de logiciels, l'optimisation du code et l'automatisation des tâches de programmation.

Réglage fin, spécificité du domaine et recyclage

L'une des pratiques courantes avec le déploiement de LLM après la préparation et le prétraitement des données consiste à sélectionner un modèle pré-entraîné qui a été formé sur un ensemble de données volumineux et diversifié. Dans le contexte d'un réglage fin, cela peut être un modèle de langage open source de grande taille tel que"Le lama de Meta 2" formé sur 70 milliards de paramètres et 2 000 milliards de jetons. Une fois le modèle pré-entraîné sélectionné, l'étape suivante consiste à l'affiner sur les données spécifiques au domaine. Cela implique d'ajuster les paramètres du modèle et de l'entraîner sur les nouvelles données pour s'adapter à un domaine et à une tâche spécifiques. Par exemple, BloombergGPT, un LLM propriétaire formé sur un large éventail de données financières au service du secteur financier. Les modèles spécifiques à un domaine, conçus et formés pour une tâche spécifique, ont généralement une précision et des performances supérieures dans leur champ d'application, mais une faible transférabilité entre d'autres tâches ou domaines. Lorsque l'environnement commercial et les données changent au cours d'une période donnée, la précision de prédiction du FM peut commencer à diminuer par rapport à ses performances lors des tests. C'est à ce moment-là que le recyclage ou le réglage fin du modèle devient crucial. Le recyclage de modèles dans l'IA/ML traditionnel fait référence à la mise à jour d'un modèle ML déployé avec de nouvelles données, généralement effectuée pour éliminer deux types de dérives qui se produisent. (1) Dérive conceptuelle – lorsque le lien entre les variables d'entrée et les variables cibles change au fil du temps, puisque la description de ce que nous voulons prédire change, le modèle peut produire des prédictions inexactes. (2) Dérive des données – se produit lorsque les caractéristiques des données d'entrée changent, comme les changements dans les habitudes ou le comportement des clients au fil du temps et donc l'incapacité du modèle à répondre à ces changements. De la même manière, la reconversion s'applique aux FM/LLM, mais elle peut être beaucoup plus coûteuse (en millions de dollars), et n'est donc pas quelque chose que la plupart des organisations pourraient envisager. Il fait l'objet de recherches actives, toujours en émergence dans le domaine des LLMOps. Ainsi, au lieu de procéder à une nouvelle formation, lorsque la dégradation du modèle se produit dans les FM affinés, les entreprises peuvent opter pour un nouveau réglage fin (beaucoup moins cher) avec un ensemble de données plus récent. Pour une perspective de coût, vous trouverez ci-dessous un exemple de tableau de prix modèle d'Azure-OpenAI Services. Pour chaque catégorie de tâches, les clients peuvent affiner et évaluer les modèles sur des ensembles de données spécifiques.

Source : Microsoft Azure

[Source : Microsoft Azure]

Ingénierie rapide et inférence

L'ingénierie rapide fait référence aux méthodes efficaces permettant de communiquer avec les LLM pour effectuer les tâches souhaitées sans mettre à jour les poids du modèle. Aussi important que soit l'entraînement et le réglage fin des modèles d'IA pour les applications PNL, l'inférence est tout aussi importante, lorsque les modèles entraînés répondent aux invites de l'utilisateur. Les exigences système pour l'inférence sont généralement beaucoup plus axées sur les performances de lecture du système de stockage d'IA qui alimente les données des LLM vers les GPU, car il doit être capable d'appliquer des milliards de paramètres de modèle

stockés pour produire la meilleure réponse.

LLMOps, surveillance des modèles et magasins de vecteurs

Tout comme les opérations d'apprentissage automatique traditionnelles (MLOps), les opérations de modèles de langage volumineux (LLMOps) nécessitent également la collaboration de scientifiques des données et d'ingénieurs DevOps avec des outils et des meilleures pratiques pour la gestion des LLM dans les environnements de production. Cependant, le flux de travail et la pile technologique des LLM peuvent varier de certaines manières. Par exemple, les pipelines LLM créés à l'aide de frameworks tels que LangChain enchaînent plusieurs appels d'API LLM vers des points de terminaison d'intégration externes tels que des magasins vectoriels ou des bases de données vectorielles. L'utilisation d'un point de terminaison d'intégration et d'un magasin vectoriel pour les connecteurs en aval (comme pour une base de données vectorielle) représente une évolution significative dans la manière dont les données sont stockées et consultées. Contrairement aux modèles ML traditionnels qui sont développés à partir de zéro, les LLM s'appuient souvent sur l'apprentissage par transfert, car ces modèles commencent avec des FM qui sont affinés avec de nouvelles données pour améliorer les performances dans un domaine plus spécifique. Il est donc crucial que les LLMOps fournissent des capacités de gestion des risques et de surveillance de la dégradation des modèles.

Risques et éthique à l'ère de l'IA générative

« ChatGPT – C'est astucieux mais ça continue à dire des bêtises. » – MIT Tech Review. Le principe du « garbage in » et du « garbage out » a toujours été un défi en informatique. La seule différence avec l'IA générative est qu'elle excelle à rendre les déchets hautement crédibles, ce qui conduit à des résultats inexacts. Les LLM ont tendance à inventer des faits pour s'adapter au récit qu'ils construisent. Par conséquent, les entreprises qui voient l'IA générative comme une excellente opportunité de réduire leurs coûts avec des équivalents IA doivent détecter efficacement les deep fakes, réduire les biais et diminuer les risques pour maintenir les systèmes honnêtes et éthiques. Un pipeline de données fluide doté d'une infrastructure d'IA robuste qui prend en charge la mobilité des données, la qualité des données, la gouvernance des données et la protection des données via un cryptage de bout en bout et des garde-fous d'IA est essentiel à la conception de modèles d'IA génératifs responsables et explicables.

Scénario client et NetApp

Figure 3 : Flux de travail d'apprentissage automatique/modèle de langage volumineux

[Figure 3 : Flux de travail d'apprentissage automatique/modèle de langage volumineux]

Sommes-nous en train de nous entraîner ou de peaufiner ? La question de savoir s'il faut (a) former un modèle LLM à partir de zéro, affiner un FM pré-entraîné ou utiliser RAG pour récupérer des données à partir de référentiels de documents en dehors d'un modèle de base et augmenter les invites, et (b) soit en exploitant des LLM open source (par exemple, Llama 2) ou des FM propriétaires (par exemple, ChatGPT, Bard, AWS Bedrock) est une décision stratégique pour les organisations. Chaque approche présente un compromis entre rentabilité, gravité des données, opérations, précision du modèle et gestion des LLM.

En tant qu'entreprise, NetApp intègre l'IA en interne dans sa culture de travail et dans son approche de la conception et de l'ingénierie des produits. Par exemple, la protection autonome contre les ransomwares de NetApp est construite à l'aide de l'IA et de l'apprentissage automatique. Il permet une détection précoce des anomalies du système de fichiers pour aider à identifier les menaces avant qu'elles n'affectent les opérations. Deuxièmement, NetApp utilise l'IA prédictive pour ses opérations commerciales telles que les prévisions de ventes et d'inventaire et les chatbots pour aider les clients dans les services d'assistance produit du centre d'appels, les spécifications techniques, la garantie, les manuels de service, etc. Troisièmement, NetApp apporte de la valeur client au pipeline de données d'IA et au flux de travail ML/LLM via des produits et des solutions au service des clients qui créent des solutions d'IA prédictives telles que la prévision de la demande,

l'imagerie médicale, l'analyse des sentiments et des solutions d'IA génératives comme les GAN pour la détection d'anomalies d'images industrielles dans le secteur manufacturier et la détection de la lutte contre le blanchiment d'argent et de la fraude dans les services bancaires et financiers avec des produits et des fonctionnalités NetApp tels que NetApp ONTAP AI, NetApp SnapMirror et NetApp FlexCache.

Capacités NetApp

Le mouvement et la gestion des données dans les applications d'IA générative telles que le chatbot, la génération de code, la génération d'images ou l'expression de modèles de génome peuvent s'étendre à l'écosystème périphérique, au centre de données privé et à l'écosystème multicloud hybride. Par exemple, un robot IA en temps réel aidant un passager à surclasser son billet d'avion en classe affaires à partir d'une application utilisateur final exposée via des API de modèles pré-entraînés tels que ChatGPT ne peut pas accomplir cette tâche par lui-même puisque les informations sur le passager ne sont pas accessibles au public sur Internet. L'API nécessite l'accès aux informations personnelles du passager et aux informations sur le billet de la compagnie aérienne qui peuvent exister dans un écosystème hybride ou multicloud. Un scénario similaire pourrait s'appliquer aux scientifiques partageant une molécule de médicament et des données de patients via une application d'utilisateur final qui utilise des LLM pour réaliser des essais cliniques dans le cadre de la découverte de médicaments impliquant un à plusieurs établissements de recherche biomédicale. Les données sensibles transmises aux FM ou aux LLM peuvent inclure des informations personnelles identifiables, des informations financières, des informations sur la santé, des données biométriques, des données de localisation, des données de communication, des données de comportement en ligne et des informations juridiques. Dans un tel cas de rendu en temps réel, d'exécution rapide et d'inférence de périphérie, il y a un mouvement de données de l'application de l'utilisateur final vers les points de terminaison de stockage via des modèles LLM open source ou propriétaires vers un centre de données sur site ou des plateformes de cloud public. Dans tous ces scénarios, la mobilité et la protection des données sont cruciales pour les opérations d'IA impliquant des LLM qui s'appuient sur de grands ensembles de données de formation et sur le mouvement de ces données.

Figure 4 : IA générative - Pipeline de données LLM

[Figure 4 : Pipeline de données génératif AI-LLM]

Le portefeuille d'infrastructures de stockage, de données et de services cloud de NetApp est alimenté par un logiciel de gestion de données intelligent.

Préparation des données : Le premier pilier de la pile technologique LLM est en grande partie inchangé par rapport à l'ancienne pile ML traditionnelle. Le prétraitement des données dans le pipeline d'IA est nécessaire pour normaliser et nettoyer les données avant la formation ou le réglage fin. Cette étape inclut des connecteurs pour ingérer des données où qu'elles résident sous la forme d'un niveau Amazon S3 ou dans des systèmes de stockage sur site tels qu'un magasin de fichiers ou un magasin d'objets comme NetApp StorageGRID.

- NetApp ONTAP* est la technologie fondamentale qui sous-tend les solutions de stockage critiques de NetApp dans le centre de données et le cloud. ONTAP inclut diverses fonctionnalités et capacités de gestion et de protection des données, notamment une protection automatique contre les ransomwares contre les cyberattaques, des fonctionnalités de transport de données intégrées et des capacités d'efficacité de stockage pour une gamme d'architectures sur site, hybrides, multiclouds dans des situations de stockage NAS, SAN, objet et défini par logiciel (SDS) des déploiements LLM.
- NetApp ONTAP AI* pour la formation de modèles d'apprentissage en profondeur. NetApp ONTAP prend en charge le stockage direct GPU NVIDIA avec l'utilisation de NFS sur RDMA pour les clients NetApp avec cluster de stockage ONTAP et nœuds de calcul NVIDIA DGX. Il offre des performances rentables pour lire et traiter les ensembles de données sources du stockage vers la mémoire à plusieurs reprises afin de favoriser l'intelligence, permettant aux organisations de bénéficier d'un accès à la formation, au réglage fin et à la mise à l'échelle des LLM.

- NetApp FlexCache* est une fonctionnalité de mise en cache à distance qui simplifie la distribution de fichiers et met en cache uniquement les données lues activement. Cela peut être utile pour la formation, le recyclage et le réglage fin du LLM, apportant de la valeur aux clients ayant des exigences commerciales telles que le rendu en temps réel et l'inférence LLM.
- NetApp SnapMirror* est une fonctionnalité ONTAP qui réplique les snapshots de volume entre deux systèmes ONTAP . Cette fonctionnalité transfère de manière optimale les données en périphérie vers votre centre de données sur site ou vers le cloud. SnapMirror peut être utilisé pour déplacer des données de manière sécurisée et efficace entre les clouds sur site et les clouds hyperscaler, lorsque les clients souhaitent développer une IA générative dans des clouds avec RAG contenant des données d'entreprise. Il transfère efficacement uniquement les modifications, économisant ainsi la bande passante et accélérant la réplication, apportant ainsi des fonctionnalités essentielles de mobilité des données lors des opérations de formation, de recyclage et de réglage fin des FM ou des LLM.
- NetApp SnapLock* apporte une capacité de disque immuable sur les systèmes de stockage basés sur ONTAP pour le contrôle de version des ensembles de données. L'architecture microcore est conçue pour protéger les données client avec le moteur FPolicy Zero Trust. NetApp garantit la disponibilité des données client en résistant aux attaques par déni de service (DoS) lorsqu'un attaquant interagit avec un LLM d'une manière particulièrement gourmande en ressources.
- NetApp Cloud Data Sense* permet d'identifier, de cartographier et de classer les informations personnelles présentes dans les ensembles de données d'entreprise, d'appliquer des politiques, de répondre aux exigences de confidentialité sur site ou dans le cloud, d'améliorer la posture de sécurité et de se conformer aux réglementations.
- Classification NetApp BlueXP*, optimisée par Cloud Data Sense. Les clients peuvent automatiquement analyser, catégoriser et agir sur les données de l'ensemble du parc de données, détecter les risques de sécurité, optimiser le stockage et accélérer les déploiements cloud. Il combine des services de stockage et de données via son plan de contrôle unifié. Les clients peuvent utiliser des instances GPU pour le calcul et des environnements multicloud hybrides pour la hiérarchisation du stockage à froid et pour les archives et les sauvegardes.
- Dualité fichier-objet NetApp *. NetApp ONTAP permet un accès à double protocole pour NFS et S3. Avec cette solution, les clients peuvent accéder aux données NFS des blocs-notes Amazon AWS SageMaker via des buckets S3 de NetApp Cloud Volumes ONTAP. Cela offre une flexibilité aux clients qui ont besoin d'un accès facile à des sources de données hétérogènes avec la possibilité de partager des données à partir de NFS et de S3. Par exemple, le réglage fin des FM comme les modèles de génération de texte Llama 2 de Meta sur SageMaker avec accès aux buckets d'objets fichiers.

Le service * NetApp Cloud Sync* offre un moyen simple et sécurisé de migrer des données vers n'importe quelle cible, dans le cloud ou sur site. Cloud Sync transfère et synchronise de manière transparente les données entre le stockage sur site ou dans le cloud, le NAS et les magasins d'objets.

- NetApp XCP* est un logiciel client qui permet des migrations de données rapides et fiables de n'importe quel système vers NetApp et de NetApp vers NetApp . XCP offre également la possibilité de déplacer efficacement des données en masse à partir des systèmes de fichiers Hadoop HDFS vers ONTAP NFS, S3 ou StorageGRID et les analyses de fichiers XCP offrent une visibilité sur le système de fichiers.
- NetApp DataOps Toolkit* est une bibliothèque Python qui permet aux scientifiques des données, aux DevOps et aux ingénieurs de données d'effectuer facilement diverses tâches de gestion des données, telles que le provisionnement, le clonage ou la capture instantanée quasi instantanée d'un volume de données ou d'un espace de travail JupyterLab, soutenus par un stockage NetApp évolutif hautes performances.

Sécurité des produits NetApp. Les LLM peuvent révéler par inadvertance des données confidentielles dans leurs réponses, ce qui constitue une préoccupation pour les RSSI qui étudient les vulnérabilités associées aux applications d'IA exploitant les LLM. Comme le souligne l'OWASP (Open Worldwide Application Security Project), les problèmes de sécurité tels que l'empoisonnement des données, la fuite de données, le déni de

service et les injections rapides dans les LLM peuvent avoir un impact sur les entreprises en raison de l'exposition des données à des accès non autorisés au service des attaquants. Les exigences de stockage des données doivent inclure des contrôles d'intégrité et des instantanés immuables pour les données structurées, semi-structurées et non structurées. Les instantanés NetApp et SnapLock sont utilisés pour le contrôle de version des ensembles de données. Il apporte un contrôle d'accès strict basé sur les rôles (RBAC), ainsi que des protocoles sécurisés et un cryptage standard du secteur pour sécuriser les données au repos et en transit. Cloud Insights et Cloud Data Sense offrent ensemble des fonctionnalités pour vous aider à identifier de manière médico-légale la source de la menace et à hiérarchiser les données à restaurer.

* ONTAP AI avec DGX BasePOD *

L'architecture de référence NetApp ONTAP AI avec NVIDIA DGX BasePOD est une architecture évolutive pour les charges de travail d'apprentissage automatique (ML) et d'intelligence artificielle (IA). Pour la phase de formation critique des LLM, les données sont généralement copiées du stockage de données vers le cluster de formation à intervalles réguliers. Les serveurs utilisés dans cette phase utilisent des GPU pour paralléliser les calculs, créant ainsi un énorme appétit pour les données. Répondre aux besoins bruts en bande passante d'E/S est essentiel pour maintenir une utilisation élevée du GPU.

* ONTAP AI avec NVIDIA AI Enterprise*

NVIDIA AI Enterprise est une suite de logiciels d'IA et d'analyse de données cloud native de bout en bout, optimisée, certifiée et prise en charge par NVIDIA pour s'exécuter sur VMware vSphere avec les systèmes certifiés NVIDIA. Ce logiciel facilite le déploiement, la gestion et la mise à l'échelle simples et rapides des charges de travail d'IA dans l'environnement cloud hybride moderne. NVIDIA AI Enterprise, optimisé par NetApp et VMware, offre une gestion des charges de travail et des données d'IA de niveau entreprise dans un package simplifié et familier.

Plateformes Cloud 1P

Les offres de stockage cloud entièrement gérées sont disponibles nativement sur Microsoft Azure sous le nom d' Azure NetApp Files (ANF), sur AWS sous le nom d' Amazon FSx for NetApp ONTAP (FSx ONTAP) et sur Google sous le nom de Google Cloud NetApp Volumes (GNCV). 1P est un système de fichiers géré et hautes performances qui permet aux clients d'exécuter des charges de travail d'IA hautement disponibles avec une sécurité des données améliorée dans les clouds publics, pour affiner les LLM/FM avec des plates-formes ML natives du cloud comme AWS SageMaker, Azure-OpenAI Services et Vertex AI de Google.

Suite de solutions partenaires NetApp

En plus de ses principaux produits, technologies et capacités de données, NetApp collabore également étroitement avec un solide réseau de partenaires d'IA pour apporter une valeur ajoutée aux clients.

- Les garde-fous NVIDIA * dans les systèmes d'IA servent de garanties pour garantir l'utilisation éthique et responsable des technologies d'IA. Les développeurs d'IA peuvent choisir de définir le comportement des applications basées sur LLM sur des sujets spécifiques et les empêcher de s'engager dans des discussions sur des sujets indésirables. Guardrails, une boîte à outils open source, offre la possibilité de connecter un LLM à d'autres services, de manière transparente et sécurisée, pour créer des systèmes conversationnels LLM fiables, sûrs et sécurisés.

Domino Data Lab fournit des outils polyvalents de niveau entreprise pour créer et produire une IA générative - rapide, sûre et économique, où que vous soyez dans votre parcours d'IA. Avec la plateforme Enterprise MLOps de Domino, les scientifiques des données peuvent utiliser leurs outils préférés et toutes leurs données, former et déployer facilement des modèles n'importe où et gérer les risques et les coûts de manière efficace, le tout à partir d'un seul centre de contrôle.

Modzy pour Edge AI. NetApp et Modzy se sont associés pour fournir une IA à grande échelle à tout type de données, y compris les images, l'audio, le texte et les tableaux. Modzy est une plateforme MLOps pour le déploiement, l'intégration et l'exécution de modèles d'IA, qui offre aux scientifiques des données les capacités de surveillance des modèles, de détection des dérives et d'explicabilité, avec une solution intégrée pour une inférence LLM transparente.

Run:AI et NetApp se sont associés pour démontrer les capacités uniques de la solution NetApp ONTAP AI avec la plate-forme de gestion de cluster Run:AI pour simplifier l'orchestration des charges de travail d'IA. Il divise et joint automatiquement les ressources GPU, conçues pour faire évoluer vos pipelines de traitement de données vers des centaines de machines avec des cadres d'intégration intégrés pour Spark, Ray, Dask et Rapids.

Conclusion

L'IA générative ne peut produire des résultats efficaces que lorsque le modèle est formé sur des volumes importants de données de qualité. Bien que les LLM aient franchi des étapes remarquables, il est essentiel de reconnaître leurs limites, les défis de conception et les risques associés à la mobilité et à la qualité des données. Les LLM s'appuient sur des ensembles de données de formation volumineux et disparates provenant de sources de données hétérogènes. Les résultats inexacts ou biaisés générés par les modèles peuvent mettre en danger les entreprises et les consommateurs. Ces risques peuvent correspondre à des contraintes pour les LLM émergeant potentiellement des défis de gestion des données associés à la qualité des données, à la sécurité des données et à la mobilité des données. NetApp aide les organisations à faire face aux complexités créées par la croissance rapide des données, la mobilité des données, la gestion multicloud et l'adoption de l'IA. Une infrastructure d'IA à grande échelle et une gestion efficace des données sont essentielles pour définir le succès des applications d'IA telles que l'IA générative. Il est essentiel que les clients couvrent tous les scénarios de déploiement sans compromettre la capacité d'expansion selon les besoins des entreprises tout en maintenant la rentabilité, la gouvernance des données et les pratiques éthiques d'IA sous contrôle. NetApp travaille constamment pour aider ses clients à simplifier et à accélérer leurs déploiements d'IA.

Informations sur le copyright

Informations sur les marques commerciales