DELL Technologies

Présentation de solution : Validated Design for Generative AI Inferencing with NVIDIA

Obtenez plus rapidement des informations à forte valeur ajoutée grâce à l'IA générative

Déployez rapidement une solution sur la pile complète pour l'inférence des grands modèles de langage grâce à l'intelligence artificielle générative (IA générative)

Améliorez la productivité et la qualité des informations

Cette architecture conjointe offre une conception modulaire et flexible prenant en charge une multitude de cas d'utilisation et d'exigences de calcul. Les composants peuvent être combinés et associés, et mis à l'échelle indépendamment selon vos besoins applicatifs.

Parmi les principaux exemples de cas d'utilisation d'inférence pris en charge :

Génération de langage naturel : Les modèles de génération peuvent être utilisés dans le cadre de tâches de génération de texte, comme la rédaction de documents ou de résumés, la génération de dialogues ou la création de contenu.

Chatbots et assistants virtuels : L'IA générative optimise les agents conversationnels, chatbots et assistants virtuels grâce à la génération de réponses en langage naturel basées sur des requêtes ou des instructions de l'utilisateur.

Développement de code : Bénéficiez d'une aide au développement logiciel avec des fonctionnalités comme la complétion de code, la possibilité de générer des tests unitaires ou une fonction de chat pour expliquer le code.

Générez des prédictions de délai de rentabilisation et des résultats plus rapides, et de meilleure qualité, tout en accélérant la prise de décision grâce à une puissante solution d'IA générative de Dell Technologies et NVIDIA. Cette solution conçue conjointement relève les défis de l'inférence comme la latence, la réactivité et les exigences de calcul, aidant à transformer les données de l'entreprise en résultats plus intelligents, à forte valeur ajoutée.

En s'appuyant sur des technologies innovantes, des services professionnels complets et un vaste écosystème de partenaires, votre organisation peut accélérer l'IA générative à l'échelle de l'entreprise. Aujourd'hui, les départements IT, les scientifiques des données, et les DevOps en IA peuvent proposer facilement une plateforme modulaire et évolutive pour l'inférence de l'IA générative et des grands modèles de langage (LLM).

- Créez de la valeur ajoutée avec une infrastructure sécurisée pour vos opérations stratégiques
- Mobilisez et faites évoluer les prédictions et informations de l'IA générative, du datacenter à la périphérie
- Améliorez la valeur IT grâce à des conseils stratégiques
- Dimensionnez votre infrastructure et consolidez tous vos besoins en matière d'inférence de l'IA

Accélérez les résultats avec une solution éprouvée

Créez rapidement une infrastructure sur site pour vos besoins applicatifs, avec une conception validée et une architecture de référence conçues pour simplifier l'adoption. En réduisant la complexité à chaque étape, vous pouvez maintenant obtenir plus d'informations et accélérer la prise de décision, tout en optimisant la productivité.

© 2023 Dell Inc. ou ses filiales.

NVIDIA

En savoir plus

- Consultez le Guide de conception
- AI InfoHub
- delltechnologies.com/ai
- Dell Technologies et NVIDIA

Qu'est-ce que l'inférence ?

Dans le domaine de l'IA, l'inférence désigne le processus consistant à utiliser un modèle entraîné pour générer des prédictions, prendre des décisions ou produire des résultats sur la base de données d'entrée. Cela consiste à appliquer les connaissances apprises et les schémas acquis pendant la phase d'entraînement du modèle à des données nouvelles, inconnues.

Lors de l'inférence, le modèle entraîné prend les données d'entrée et les traite via des algorithmes de calcul ou une architecture réseau neuronale pour produire un résultat ou une prédiction. Le modèle applique les paramètres appris, pondérations ou règles pour transformer les données d'entrée en informations ou actions pertinentes.

L'inférence est une étape cruciale dans le cycle de vie d'un système d'IA. Après l'entraînement d'un modèle sur des données, étiquetées ou non, afin d'apprendre des schémas et des corrélations, l'inférence permet au modèle de généraliser ses connaissances et de faire des prédictions ou de générer des réponses sur des données du monde réel ou inconnues.

Fournissez des résultats plus rapidement avec notre aide

Les experts des services Dell vous aident à exploiter la valeur de l'IA générative pour vos données plus rapidement, avec toute une gamme de services pour vous accompagner à chaque étape de votre parcours vers l'IA générative :

- **Stratégie :** élaborez votre feuille de route pour atteindre les objectifs d'innovation de vos parties prenantes IT et commerciales
- **Implémentation :** établissez votre plateforme en utilisant les conceptions Dell Validated Design pour mettre en œuvre le matériel et le logiciel d'inférence de l'IA générative
- **Adoption :** accélérez la valeur de vos cas d'utilisation d'IA générative en implémentant un modèle d'inférence préentraîné
- **Évolutivité :** gérez votre portefeuille de solutions d'innovation en matière d'IA générative avec des experts techniques résidents et des offres de formation pour développer les compétences de votre équipe en matière d'IA générative

Caractéristiques techniques

Les configurations Validated Design sont basées sur les serveurs Dell PowerEdge XE et au format rack les plus récents, optimisés par l'accélération de l'IA et équipés des tout derniers processeurs graphiques NVIDIA et de NVIDIA AI Enterprise, avec le logiciel Triton Inference Server et le framework NeMo. Le stockage Data Lake rapide et complet pour l'IA générative et les grands modèles de langage est fourni par les baies de stockage All-Flash ou hybride Dell PowerScale.

Calcul	Accélérateurs	Gestion de réseau	Logiciels	Stockage
Serveurs Dell PowerEdge R760xa	Processeurs graphiques NVIDIA A100 ou H100	Gestion de réseau NVIDIA, Dell PowerSwitch S5232F-ON ou S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ. NVIDIA AI Enterprise avec framework NeMo pour les grands modèles de langage (LLM) et Triton Inference Server ; NVIDIA Base Command Manager Essentials	Pris en charge par Dell PowerScale, ECS et ObjectScale

Dell Technologies et NVIDIA

Dell Technologies et NVIDIA travaillent ensemble pour exécuter et accélérer les charges applicatives d'IA générative, fournir du matériel et des logiciels validés par l'ingénierie pour accélérer les charges applicatives d'IA, de ML et de DL afin de répondre aux besoins des clients dans toutes les entreprises et tous les secteurs d'activité. Avec cette conception Validated Design pour l'inférence des grands modèles de langage (LLM), vous pouvez accélérer votre transformation numérique grâce à des données en temps réel qui améliorent la prise de décisions clés à grande échelle, avec des solutions optimisées pour accélérer le délai de rentabilisation de vos initiatives en matière d'IA.

En savoir plus sur les solutions Dell

Contacter un expert Dell Technologies

Afficher plus de ressources

Prenez part à la discussion avec #HashTag

© 2023 Dell Inc. ou ses filiales. Tous droits réservés. Dell et les autres marques citées sont des marques commerciales de Dell Inc. ou de ses filiales. SAP, SAP HANA, SAP S/4HANA et SAP Business One sont des marques déposées de SAP SE en Allemagne et dans d'autres pays. Les autres marques sont la propriété de leurs détenteurs respectifs.

DELL Technologies

NVIDIA