L'IA générative dans l'entreprise

Une infrastructure de production évolutive et modulaire pour l'artificiel Grands modèles de langage d'intelligence

mai 2023

H19611

Papier blanc

Abstrait

Ce livre blanc présente une vue d'ensemble de l'IA générative et présente le projet Helix, une collaboration entre Dell Technologies et NVIDIA visant à permettre des solutions d'IA générative full-stack hautes performances, évolutives et modulaires pour les grands modèles de langage dans l'entreprise.

Solutions technologiques Dell

Contenu

Introduction 5

Résumé exécutif ..... 5 À propos de ce document 6

Public 6

Contexte et concepts de l'IA générative 6

Contexte 6

Définition et aperçu 7

Évolution 7

Modèles de transformateurs 8

Caractéristiques de la charge de travail 8

Types de charges de travail ..... 9 Types de sorties 11

Défis commerciaux et techniques ..... 12 Propriété du contenu ..... 12 Qualité des données ..... 12 Complexité du modèle ..... 12 Considérations éthiques 12

Durabilité ..... 12 Conformité réglementaire 12

Avantages 13

Avantages de l'IA générative 13

Avantages Dell et NVIDIA 13

Cas d'utilisation 14

Architecture des solutions Dell et NVIDIA 15

Architecture de haut niveau 15

Esprit de modularité ..... 15 Modules architecturaux 17

Évolutivité 18

Sécurité 18

Considérations sur les composants d'infrastructure pour l'IA 19

Calcul 19

Accélérateurs 19

Stockage 19

Réseautage 19

Résumé 19

Contenu

Infrastructure Dell et composants logiciels..... .....20 Serveurs Dell
PowerEdge ....20 Stockage de fichiers
Dell..... ....21 Stockage d'objets
Dell..... ....21 Réseau Dell
PowerSwitch..... ....22 Dell OpenManage
Enterprise..... ....23 Dell OpenManage Enterprise Power
Manager ....23 Dell
CloudIQ 24
Services Dell 24

Infrastructure NVIDIA et composants logiciels ....25 Accélérateurs
NVIDIA 25
Logiciel NVIDIA IA 26

Configurations du système ....29 Inférence de
grand modèle ....29 Personnalisation grand
modèle..... ....30 Formation sur grand
modèle..... 30
Résumé 31

Conclusion..... 32
L'avantage de l'IA générative ....32 Nous
apprécions vos commentaires 32

Les références..... 33
Documentation Dell Technologies..... .....33 Documentation
NVIDIA 33

Introduction

Exécutif résumé

La croissance des applications et des cas d'utilisation de l'intelligence artificielle (IA) est stupéfiante, avec des impacts sur presque toutes les facettes de la vie professionnelle et personnelle. L'IA générative, la branche de l'IA conçue pour générer de nouvelles données, images, codes ou autres types de contenu que les humains ne programment pas explicitement, devient particulièrement percutante et influente.

Selon un analyste, la taille du marché mondial de l'IA générative était déjà estimée à 10,79 milliards de dollars en 2022. Elle devrait approcher les 118 milliards de dollars d'ici 2032, avec une croissance annuelle composée (TCAC) de 27 % entre 2023 et 20321.

Outre une myriade d'autres applications, les cas d'utilisation incluent :

- Agents conversationnels et chatbots pour le service client
- Création de contenu audio et visuel
- Programmation de logiciels
- Sécurité, détection des fraudes et renseignements sur les menaces
- Interaction et traduction en langage naturel

Rares sont les domaines de l'entreprise et de la société qui ne sont pas touchés d'une manière ou d'une autre par cette technologie.

Bien que les modèles publics d'IA générative tels que ChatGPT, Google Bard AI, DALL-E et d'autres offres plus spécialisées soient intrigants, il existe des préoccupations légitimes quant à leur utilisation dans l'entreprise. Ces préoccupations incluent la propriété des résultats, qui englobe les questions d'exactitude, de véracité et d'attribution de la source.

Par conséquent, il existe un besoin impérieux pour les entreprises de développer leurs propres modèles linguistiques étendus (LLM), formés sur des ensembles de données propriétaires ou développés et affinés à partir de modèles pré-entraînés connus.

Dell Technologies et NVIDIA

Dell Technologies et NVIDIA ont ouvert la voie en proposant des innovations conjointes pour l'IA et le calcul haute performance. Nous collaborons activement dans ce nouvel espace pour permettre aux clients de créer et d'exploiter des modèles d'IA génératifs pour l'entreprise.

- Dell Technologies dispose d'une infrastructure de pointe qui comprend de puissants serveurs avec accélération de l'unité de traitement graphique (GPU) NVIDIA, systèmes de stockage de données, mise en réseau, gestion des systèmes, conceptions de référence et années d'expérience dans l'aide aux entreprises dans leurs initiatives d'IA.

- NVIDIA possède les principales solutions d'accélération GPU et de mise en réseau de bout en bout, logiciel de gestion de cluster, logiciel NVIDIA AI Enterprise, état de l'art,

1 Recherche sur la présence (https://www.precedenceresearch.com/generative-ai-market)

des modèles de base pré-entraînés, y compris le framework NeMo, et l'expertise nécessaire pour créer, personnaliser et exécuter une IA générative.

Nous travaillons désormais en partenariat sur un nouveau projet d'IA générative appelé Project Helix, une initiative conjointe entre Dell Technologies et NVIDIA, visant à apporter l'IA générative aux centres de données d'entreprise du monde entier. Project Helix est une solution complète qui permet aux entreprises de créer et d'exécuter des modèles d'IA personnalisés, construits avec la connaissance de leur entreprise. Nous avons conçu un une infrastructure évolutive, modulaire et haute performance qui permet aux entreprises du monde entier de créer une vague de solutions d'IA générative qui réinventeront leurs secteurs et leur donner un avantage concurrentiel.

L'IA générative est aujourd'hui l'un des domaines les plus passionnants et en évolution rapide de l'IA. Il s'agit d'une technologie transformatrice et de la combinaison d'une infrastructure et de logiciels puissants de Dell Technologies, associé aux accélérateurs, aux logiciels d'IA et à l'expertise en IA de NVIDIA est sans égal.

À propos de ce document

Dans ce livre blanc, les lecteurs peuvent obtenir un aperçu complet de l'IA générative, y compris ses principes sous-jacents, ses avantages, ses architectures et ses techniques. Ils peuvent également en apprendre davantage sur les différents types de modèles d'IA générative et sur la manière dont ils sont utilisés dans des applications réelles.

Ce livre blanc explore également les défis et les limites de l'IA générative, tels que la difficulté de former des modèles à grande échelle, le potentiel de biais et de préoccupations éthiques, ainsi que le compromis entre la génération de résultats réalistes et le maintien de la confidentialité des données.

Ce livre blanc fournit également des conseils sur la manière de développer et de déployer efficacement des modèles d'IA génératifs. Il comprend des considérations sur l'infrastructure matérielle et logicielle de Dell Technologies et NVIDIA, des mesures de gestion des données et d'évaluation, le tout conduisant à une architecture de production évolutive et hautes performances pour l'IA générative dans l'entreprise.

Public

Ce livre blanc est destiné aux chefs d'entreprise, aux directeurs technologiques (CTO), aux directeurs de l'information (CIO), aux gestionnaires d'infrastructure informatique et aux architectes système qui sont intéressés, impliqués ou envisagent la mise en œuvre de l'IA générative.

Contexte et concepts de l'IA générative

Arrière-plan

L'IA a connu plusieurs phases de développement depuis sa création au milieu du XXe siècle. Les principales phases du développement de l'IA, ainsi que les délais approximatifs, sont :

1. Systèmes basés sur des règles (années 1950-1960) : la première phase du développement de l'IA portait sur la création de systèmes basés sur des règles, dans lesquels les experts codaient leurs connaissances dans un ensemble de règles que l'ordinateur devait suivre. Ces systèmes étaient limités dans leur capacité à apprendre de nouvelles données ou à s'adapter à de nouvelles situations.

2. Apprentissage automatique (années 1960-1990) – La phase suivante du développement de l'IA a porté sur l'utilisation d'algorithmes d'apprentissage automatique pour entraîner les ordinateurs à reconnaître des modèles dans les données et à faire des prédictions ou des décisions basées sur ces modèles. Cette phase a vu le développement d'algorithmes tels que les arbres de décision, la régression logistique et les réseaux de neurones.

3. Apprentissage profond (années 2010 à aujourd'hui) – La phase suivante de l'IA a porté sur l'apprentissage profond. L'apprentissage profond est un sous-ensemble de l'apprentissage automatique qui utilise des réseaux de neurones à plusieurs couches pour reconnaître des modèles complexes dans les données. Cette phase a été efficace pour traiter des images, des vidéos et des données en langage naturel.

4. IA générative (présente) – La phase actuelle porte sur l'IA générative. L'IA générative utilise des algorithmes d'apprentissage en profondeur pour générer du contenu tel que des images, des vidéos, de la musique et même du texte qui ressemble étroitement aux modèles des données d'origine. Cette phase présente un énorme potentiel pour créer de nouveaux types de contenu et générer de nouvelles informations et prédictions basées sur de grandes quantités de données.

Bien que ces phases ne soient pas strictement définies ni mutuellement exclusives, elles représentent des étapes majeures dans le développement de l'IA et démontrent la complexité et la sophistication croissantes des algorithmes et des applications d'IA au fil du temps.

Définition et aperçu

L'IA générative est une branche de l'intelligence artificielle qui construit des modèles capables de générer du contenu (tel que des images, du texte ou de l'audio) qui n'est pas explicitement programmé par les humains et dont le style et la structure sont similaires aux exemples existants. Les techniques d'IA générative utilisent des algorithmes d'apprentissage en profondeur pour apprendre à partir de grands ensembles de données d'exemples, apprendre des modèles et générer un nouveau contenu similaire aux données d'origine.

L'un des aspects importants de l'IA générative est sa capacité à créer du contenu impossible à distinguer du contenu créé par des humains, qui a de nombreuses applications dans des secteurs tels que le divertissement, le design et le marketing. Par exemple, l'IA générative peut créer des images réalistes de produits qui n'existent pas encore, générer de la musique qui imite le style d'un artiste particulier, ou même générer un texte impossible à distinguer du contenu.
écrit par des humains.

Un domaine important de l'IA générative est la génération de langage naturel (NLG), qui est un sous-ensemble du traitement du langage naturel (NLP) et implique la génération d'un texte en langage naturel cohérent, fluide et de style similaire à celui d'un texte existant ou produit par l'homme. NLG a été utilisé pour diverses applications, notamment les chatbots, la traduction linguistique et la génération de contenu.

Dans l'ensemble, l'IA générative a le potentiel de transformer la façon dont nous créons et consommons du contenu. Elle a le potentiel de générer de nouvelles connaissances et perspectives dans divers domaines, ce qui en fait un domaine de développement passionnant pour l'IA.

Évolution

Les progrès des algorithmes d'apprentissage profond et la disponibilité de grands ensembles de données de textes en langage naturel ont conduit à l'évolution de la NLG vers l'IA générative. Les premiers systèmes NLG reposaient sur des approches basées sur des règles ou des modèles, dont la capacité à générer un contenu diversifié et créatif était limitée. Cependant, avec l'essor des techniques d'apprentissage en profondeur telles que les réseaux neuronaux récurrents (RNN) et les transformateurs, il est devenu possible de créer des modèles capables d'apprendre à partir de grands ensembles de données de texte en langage naturel et de générer un nouveau texte plus diversifié et créatif.

Une étape importante dans l'évolution de l'IA générative a été le développement de la série de modèles Generative Pretrained Transformer (GPT) par OpenAI. Le modèle GPT original, publié en 2018, était un modèle basé sur un transformateur formé sur un vaste corpus de données textuelles. Le modèle a pu générer un texte cohérent et fluide, dont le style était similaire à celui de

les données originales. Les versions ultérieures du modèle, notamment GPT-2 et GPT-3, ont repoussé les limites de ce qui est possible avec NLG, générant un texte de plus en plus diversifié, créatif et même humain dans certains cas.

Aujourd'hui, les techniques d'IA générative sont utilisées dans un large éventail d'applications, notamment la génération de contenu, les chatbots, la traduction linguistique, etc. À mesure que le domaine continue d'évoluer, nous pouvons nous attendre à voir apparaître des modèles d'IA générative plus sophistiqués, capables de générer un contenu encore plus créatif et diversifié.

Modèles de transformateurs

Les modèles de transformateur sont un type de modèle d'apprentissage profond couramment utilisé en PNL et dans d'autres applications d'IA générative. Les transformateurs ont été introduits dans un article fondateur de Vaswani et d'autres en 2017. Ils sont depuis devenus un élément clé de nombreux modèles PNL de pointe.

À un niveau élevé, les modèles de transformateur sont conçus pour apprendre les relations contextuelles entre les mots d'une phrase ou d'une séquence de texte. Ils réalisent cet apprentissage en utilisant un mécanisme appelé auto-attention, qui permet au modèle de peser l'importance des différents mots dans une séquence en fonction de leur contexte. Cette méthode contraste avec les modèles de réseaux neuronaux récurrents (RNN) traditionnels, qui traitent les séquences d'entrée de manière séquentielle et n'ont pas une vue globale de la séquence.

L'un des principaux avantages des modèles de transformateur est leur capacité à traiter les séquences d'entrée en parallèle, ce qui les rend plus rapides que les RNN pour de nombreuses tâches NLP. Ils se sont également révélés très efficaces pour une gamme de tâches de PNL, notamment la modélisation linguistique, la classification de textes, la réponse aux questions et la traduction automatique.

Le succès des modèles de transformateurs a conduit au développement de modèles pré-entraînés à grande échelle. modèles de langage, appelés transformateurs de pré-entraînement génératifs (GPT), tels que la série GPT d'OpenAI et le modèle Bidirectionnel Encoder Representations from Transformers (BERT) de Google. Ces modèles pré-entraînés peuvent être ajustés pour des tâches PNL spécifiques avec relativement peu de données de formation supplémentaires, ce qui les rend très efficaces pour un large éventail d'applications PNL.

Dans l'ensemble, les modèles de transformateur ont révolutionné le domaine de la PNL et sont devenus un élément clé de nombreux modèles d'IA générative de pointe. Leur capacité à apprendre les relations contextuelles entre les mots dans une séquence de texte a offert de nouvelles possibilités pour la génération de langage, la compréhension de texte et d'autres tâches de PNL.

Caractéristiques de la charge de travail

Les charges de travail d'IA générative peuvent être globalement classées en deux types : la formation et l'inférence. La formation utilise un vaste ensemble de données d'exemples pour former un modèle d'IA génératif, tandis que l'inférence utilise un modèle entraîné pour générer un nouveau contenu basé sur une entrée. La préparation des données avant la formation peut également constituer une tâche importante lors de la création de modèles personnalisés. Tous ces charges de travail présentent des caractéristiques qui doivent être prises en compte dans la conception des solutions et de leur infrastructure.

Les caractéristiques d'une charge de travail d'IA générative peuvent varier en fonction de l'application spécifique et du type de modèle utilisé. Cependant, certaines caractéristiques communes incluent :

- Intensité de calcul : les charges de travail d'IA générative peuvent être informatiques intensif, nécessitant des quantités importantes de puissance de traitement pour former ou générer

Nouveau contenu. Ce scénario s'applique particulièrement aux modèles à grande échelle tels que GPT-3, qui peuvent nécessiter du matériel spécialisé tel que des GPU pour s'entraîner efficacement.

- Exigences en matière de mémoire : les modèles d'IA génératifs nécessitent des quantités importantes de mémoire pour stocker les paramètres du modèle et les représentations intermédiaires. Ce scénario s'applique particulièrement aux modèles basés sur des transformateurs tels que GPT-3, qui comportent de nombreuses couches et peuvent nécessiter des centaines de millions, voire des milliards de paramètres. Il est donc essentiel de disposer d'une capacité de mémoire GPU suffisante.
- Dépendances des données : les modèles d'IA générative dépendent fortement des la qualité et la quantité des données d'entraînement, ce qui peut grandement affecter les performances du modèle. La préparation et le nettoyage des données sont des éléments importants d'une solution car L'exploitation de grands ensembles de données de haute qualité est essentielle pour créer des modèles personnalisés.
- Exigences de latence : les charges de travail d'inférence peuvent avoir une latence stricte. exigences, notamment dans les applications temps réel telles que les chatbots ou les assistants vocaux. Les modèles doivent être optimisés pour la vitesse d'inférence, ce qui peut impliquer des techniques telles que la quantification ou l'élagage du modèle. Les considérations de latence favorisent également les solutions sur site ou hybrides, par opposition aux solutions purement basées sur le cloud, pour former et déduire à partir de modèles les plus proches de la source des données.
- Précision du modèle : la précision et la qualité du contenu généré sont un résultat critique pour de nombreuses applications d'IA générative, et est généralement évalué à l'aide de mesures telles que la perplexité, le score de l'étude d'évaluation bilingue (BLEU) ou l'évaluation humaine.

Dans l'ensemble, les charges de travail d'IA générative peuvent être très complexes et difficiles, nécessitant du matériel, des logiciels et une expertise spécialisés pour obtenir des résultats optimaux. Cependant, avec les bons outils et techniques, ils peuvent permettre un large éventail d'applications passionnantes et innovantes dans des domaines tels que la PNL, la vision par ordinateur et les arts créatifs.

Types de charges de travail

Il existe plusieurs types spécifiques de charges de travail d'IA générative ; chacun a des exigences différentes. Les configurations système décrites plus loin dans ce livre blanc reflètent ces exigences.

Inférence

L'inférence est le processus d'utilisation d'un modèle d'IA génératif pour générer un nouveau contenu prédictif basé sur les entrées. Un modèle pré-entraîné est formé sur un vaste ensemble de données et lorsque de nouvelles données sont introduites dans le modèle, il effectue des prédictions basées sur ce qu'il a appris au cours de la formation. Cette formation consiste à introduire une séquence ou une image d'entrée dans le modèle et à recevoir une séquence ou une image de sortie comme résultat. L'inférence est généralement plus rapide et nécessite moins de calculs que la formation, car elle n'implique pas la mise à jour des paramètres du modèle.

Personnalisation du modèle

La personnalisation du modèle pré-entraîné est le processus de recyclage d'un modèle d'IA génératif existant pour des cas d'utilisation spécifiques à une tâche ou à un domaine. Pour les modèles volumineux, il est plus efficace de personnaliser que d'entraîner le modèle sur un nouvel ensemble de données. Les techniques de personnalisation utilisées aujourd'hui comprennent le réglage fin, le réglage des instructions, l'apprentissage rapide (y compris le réglage rapide et P-tuning), apprentissage par renforcement avec feedback humain, apprentissage par transfert et utilisation d'adaptateurs (ou transformateurs adaptables).

Les types de personnalisation les plus utiles sont le réglage fin, l'apprentissage rapide et l'apprentissage par transfert.

Réglage fin Le

réglage fin recycle un modèle pré-entraîné sur une tâche ou un ensemble de données spécifique, en adaptant ses paramètres pour améliorer les performances et le rendre plus spécialisé. Cette méthode traditionnelle de personnalisation soit gèle toutes les couches sauf une et ajuste les pondérations et les biais sur un nouvel ensemble de données, soit ajoute une autre couche au réseau neuronal et recalcule les pondérations et les biais sur un nouvel ensemble de données.

Apprentissage rapide

L'apprentissage rapide est une stratégie qui permet de réutiliser des modèles de langage pré-entraînés pour différentes tâches sans ajouter de nouveaux paramètres ni affiner les données étiquetées. Ces techniques peuvent également être utilisées sur de grands modèles d'images génératives d'IA.

L'apprentissage rapide peut être classé en deux techniques plus larges : le réglage rapide et le réglage P.

- Le réglage rapide est le processus de recyclage d'un modèle d'IA génératif pré-entraîné pour des cas d'utilisation spécifiques à une tâche ou à un domaine. Il utilise des ensembles de données personnalisés pour améliorer ses performances sur un domaine, un cas d'utilisation ou une tâche spécifique, ou pour incorporer des connaissances supplémentaires dans le modèle. Ce processus permet au modèle de s'adapter aux caractéristiques spécifiques du nouvel ensemble de données et peut améliorer sa précision et ses performances sur la tâche.
- Le réglage P, ou réglage des paramètres, se concentre sur l'ajustement des invites ou des instructions pendant l'inférence pour façonner la sortie du modèle sans modifier ses poids sous-jacents. Les deux techniques jouent un rôle dans la personnalisation et l'optimisation de grands modèles de langage pour des cas d'utilisation spécifiques.

Transférer l'apprentissage

L'apprentissage par transfert est une technique traditionnelle permettant d'utiliser des modèles d'IA génératifs pré-entraînés pour accélérer la formation sur de nouveaux ensembles de données. Cette technique commence par un modèle pré-entraîné qui a déjà appris des fonctionnalités utiles à partir d'un grand ensemble de données, puis l'adapte à un nouvel ensemble de données avec une plus petite quantité de données d'entraînement. Cela peut être beaucoup plus rapide et efficace que de former initialement un modèle sur le nouvel ensemble de données, car le modèle pré-entraîné comprend déjà les caractéristiques sous-jacentes des données. L'apprentissage par transfert est utile lorsque les données de formation disponibles pour une nouvelle tâche ou un nouveau domaine sont limitées. L'apprentissage par transfert n'est généralement pas utilisé pour les LLM d'IA génératifs, mais est efficace avec les modèles d'IA généraux.

Dans cette conception de solution, les configurations liées à la personnalisation sont optimisées pour le réglage fin et le réglage P. Cependant, les considérations d'évolutivité et de conception globale de l'architecture s'appliquent toujours à d'autres techniques de personnalisation et aux ensembles de données autres que le texte.

Entraînement

La formation est le processus d'utilisation d'un ensemble de données pour former initialement un modèle d'IA générative. La formation alimente les exemples de modèle à partir de l'ensemble de données et ajuste les paramètres du modèle pour améliorer ses performances sur la tâche. La formation peut être un processus gourmand en calcul, en particulier pour les modèles à grande échelle comme GPT-3.

Dans un workflow de bout en bout pour l'IA générative, la séquence exacte de ces étapes dépend sur l'application et les exigences spécifiques. Par exemple, un flux de travail commun pour les LLM pourrait impliquer:

- Prétraitement et nettoyage des données d'entraînement
- Formation d'un modèle d'IA génératif sur les données
- Évaluation des performances du modèle formé
- Affiner le modèle sur une tâche ou un ensemble de données spécifique
- Évaluation des performances du modèle affiné
- Déploiement du modèle d'inférence dans un environnement de production

L'apprentissage par transfert peut également être utilisé à différents moments de ce flux de travail pour accélérer le processus de formation ou améliorer les performances du modèle. Dans l'ensemble, la clé est de sélectionner les techniques et les outils appropriés pour chaque étape du flux de travail et d'optimiser le processus en fonction des exigences et contraintes spécifiques de l'application.

Types de sorties

Le type de données utilisé et le résultat de l'IA générative varient en fonction du type de données analysées. Bien que ce projet se concentre sur les LLM, d'autres types de modèles d'IA génératifs peuvent produire d'autres types de résultats.

- Texte : les LLM peuvent être utilisés pour générer un nouveau texte basé sur une invite spécifique ou pour compiler de longues sections de texte en résumés plus courts. Par exemple, ChatGPT peut générer des articles d'actualité ou des descriptions de produits à partir de quelques détails clés.
- Image : des modèles d'IA génératifs pour les images peuvent être utilisés pour créer des images réalistes. des images de personnes, d'objets ou d'environnements qui n'existent pas. Par exemple, StyleGAN2 peut générer des portraits réalistes de personnes inexistantes.
- Audio : les modèles d'IA génératifs pour l'audio peuvent être utilisés pour générer de nouveaux sons ou de la musique basés sur des échantillons audio existants ou pour créer des simulations vocales réalistes. Par exemple, Tacotron 2 peut générer des paroles qui ressemblent à celles d'une personne spécifique, même si cette personne n'a jamais prononcé ces mots.
- Vidéo : les modèles d'IA générative pour la vidéo peuvent être utilisés pour créer des vidéos basées sur des séquences existantes ou pour générer des animations réalistes de personnes ou d'objets. Par exemple, DALL-E peut générer des images d'objets qui n'existent pas, et ces images peuvent être combinées pour créer des vidéos animées.

Dans chaque cas, le modèle d'IA générative doit être entraîné sur de grands ensembles de données du type de données approprié. Le processus de formation est adapté aux exigences du type de données et au type de données spécifique car différents formats d'entrée et de sortie sont requis pour chaque type de données. Les progrès récents sont désormais capables d'intégrer différents types de données, par exemple en utilisant une entrée de texte pour générer une image.

Défis commerciaux et techniques

Il existe des défis à la fois commerciaux et techniques à prendre en compte lors de l'utilisation de modèles d'IA génératifs, en particulier les modèles du domaine public qui n'ont pas été développés et contrôlés au sein de l'entreprise.

Les exemples suivants montrent les défis auxquels les entreprises peuvent être confrontées lors de la mise en œuvre de modèles d'IA générative, ainsi que les solutions potentielles pour relever ces défis. Il est important d'aborder chaque défi au cas par cas et de travailler avec des experts du domaine pour développer les meilleures solutions possibles.

Propriété de contenu

Il existe des préoccupations légitimes au sein de l'entreprise concernant la propriété des résultats et la propriété intellectuelle lors de l'utilisation de certains modèles d'IA générative. Ces préoccupations incluent des questions d'exactitude, de véracité et d'attribution de la source. Les données utilisées pour former des modèles publics, bien que nombreuses, peuvent être basées sur des connaissances incomplètes ou obsolètes ou conduire à l'impossibilité de vérifier les faits ou d'accéder à des informations en temps réel.

Qualité des données

L'un des plus grands défis de tout modèle d'apprentissage automatique est de garantir que les données d'entraînement sont de haute qualité. Ce besoin est particulièrement important pour les modèles d'IA génératifs, qui peuvent nécessiter de grandes quantités de données d'entraînement pour générer des résultats précis. Pour relever ce défi, les entreprises doivent s'assurer que leurs données sont propres, bien étiquetées et représentatives du problème qu'elles tentent de résoudre.

Complexité du modèle

Les modèles d'IA générative peuvent être complexes et nécessiter d'importantes ressources informatiques pour s'entraîner et courir. Cette exigence peut constituer un défi pour les entreprises qui n'ont pas accès à du matériel puissant ou qui travaillent avec de grands ensembles de données.

Considérations éthiques

Les modèles d'IA générative peuvent avoir des implications éthiques, surtout s'ils sont utilisés pour créer du contenu ou prendre des décisions qui affectent la vie des gens. Pour relever ce défi, les entreprises doivent examiner attentivement les implications éthiques potentielles de leurs modèles d'IA générative et veiller à ce qu'ils ne causent aucun préjudice.

Durabilité

Les modèles d'IA générative à grande échelle nécessitent des ressources informatiques et une puissance considérables pour fonctionner. Les processus de formation et d'inférence pour de tels modèles peuvent consommer des quantités importantes d'énergie, contribuant ainsi à l'augmentation des émissions de carbone, des demandes de refroidissement et de l'impact environnemental.

Conformité réglementaire

Selon le secteur et l'application, les entreprises doivent respecter des exigences réglementaires lors de la mise en œuvre de modèles d'IA génératifs. Par exemple, dans le domaine de la santé, il peut exister des réglementations concernant la confidentialité des patients et la sécurité des données. Pour relever ce défi, les entreprises doivent travailler en étroite collaboration avec les équipes juridiques et de conformité pour garantir que leurs modèles d'IA générative répondent à toutes les exigences réglementaires.

Avantages

Avantages de l'IA générative

L'IA générative peut offrir de nombreux avantages à une organisation dans plusieurs dimensions. Ces avantages comprennent :

- Productivité améliorée : pour automatiser les tâches répétitives et chronophages, permettant aux employés de se concentrer sur des tâches de plus haut niveau et augmentant la productivité globale.
- Expérience client améliorée : développer des interfaces conversationnelles et des chatbots capables d'améliorer l'engagement et la satisfaction des clients en fournissant des réponses personnalisées et rapides.
- Meilleure prise de décision : pour générer des informations et des recommandations à partir de données qui peuvent aider à éclairer les décisions commerciales et à améliorer les performances globales de l'entreprise.
- Économies de coûts : pour aider à réduire les coûts opérationnels en automatisant les tâches et en améliorant l'efficacité des processus, ce qui se traduit finalement par des économies de coûts.
- Innovation accrue : pour générer de nouvelles idées et solutions qui peuvent aider stimuler l'innovation et créer de nouvelles sources de revenus
- Avantage concurrentiel : aider les entreprises à garder une longueur d'avance sur la concurrence en permettant des processus plus rapides et plus efficaces, un meilleur engagement client et une meilleure prise de décision.

Avantages Dell et NVIDIA

Les avantages offerts par Dell Technologies et NVIDIA sont importants, car ensemble nous:

- Fournir des solutions d'IA générative complètes construites sur le meilleur de l'infrastructure et des logiciels Dell, avec les derniers accélérateurs NVIDIA, les logiciels NVIDIA AI et l'expertise en IA.
- Fournir des conceptions validées qui réduisent le temps et les efforts de conception et de spécification. Solutions d'IA, accélérant le délai de valorisation
- Fournir des conseils en matière de dimensionnement et de mise à l'échelle afin que votre infrastructure soit efficacement adaptée à vos besoins, mais puisse également croître à mesure que ces besoins augmentent.
- Permettre aux entreprises de créer, de personnaliser et d'exécuter sur site une IA générative spécialement conçue pour résoudre des défis commerciaux spécifiques, et d'utiliser la même plate-forme informatique accélérée pour créer des modèles de pointe.
- Aider les entreprises tout au long du cycle de vie de l'IA générative, depuis l'infrastructure provisionnement, formation de grands modèles, réglage fin du modèle pré-entraîné, déploiement de modèles multisites et inférence de grands modèles
- Activer des modèles d'IA génératifs personnalisés qui se concentrent sur le fonctionnement souhaité domaine, avoir une connaissance à jour de votre entreprise, avoir les compétences nécessaires et pouvoir améliorer continuellement votre production
- Inclure des modèles de base pré-entraînés de pointe pour accélérer rapidement le création de modèles d'IA génératifs personnalisés

- Garantir la sécurité et la confidentialité des données sensibles et exclusives de l'entreprise.
  comme le respect des réglementations gouvernementales
- Conceptions de matériel de stockage et de serveur puissantes mais aux performances optimisées
  couplé à une accélération GPU, ainsi qu'à un logiciel de gestion système qui comprend une gestion
  avancée de l'alimentation, une optimisation thermique et une surveillance globale de l'utilisation de l'énergie
- Inclure la capacité de développer une IA plus sûre et plus fiable avec des modèles et des ensembles de données
  connus : une exigence fondamentale des entreprises d'aujourd'hui.

Cas d'utilisation

Les modèles d'IA générative ont le potentiel de répondre à un large éventail de cas d'utilisation et de résoudre de nombreux défis commerciaux dans différents secteurs. Les modèles d'IA générative peuvent être utilisés pour :

- Service client : pour améliorer l'identification des intentions du chatbot, résumez
  conversations, répondre aux questions des clients et diriger les clients vers les
  ressources.
- Création de contenu : pour créer du contenu tel que des descriptions de produits, des réseaux sociaux.
  des publications dans les médias, des articles de presse et même des livres. Cette capacité peut aider les
  entreprises à économiser du temps et de l'argent en automatisant le processus de création de contenu.
- Ventes et marketing : pour créer des expériences personnalisées pour les clients, telles que des recommandations
  de produits personnalisées ou des messages marketing personnalisés.
- Conception de produits : pour concevoir de nouveaux produits ou améliorer des produits existants. Par exemple, un
  modèle d'IA générative peut être formé sur des images de produits existants pour générer de nouvelles conceptions
  répondant à des critères spécifiques.
- Éducation : pour créer des expériences d'apprentissage personnelles, similaires à celles des tuteurs, et générer
  des plans d'apprentissage et du matériel d'apprentissage personnalisé.
- Détection de fraude : pour détecter et prévenir la fraude dans les transactions financières ou dans d'autres contextes. Par
  exemple, un modèle d'IA générative peut être entraîné pour reconnaître les modèles de comportement
  frauduleux et signaler les transactions suspectes.
- Soins de santé : pour analyser des images médicales ou des données de patients afin de faciliter le diagnostic ou le
  traitement. Par exemple, un modèle d'IA générative peut être formé pour analyser des images médicales
  afin d'identifier des cellules cancéreuses ou analyser des structures protéiques pour la découverte de nouveaux
  médicaments.
- Jeux : pour créer des expériences de jeu plus réalistes et plus engageantes. Par exemple, un modèle d'IA
  générative peut être entraîné pour créer des animations plus réalistes ou générer de nouveaux niveaux
  de jeu.
- Développement de logiciels : pour écrire du code à partir du langage humain, convertissez le code
  d'un langage de programmation à un autre, corriger un code erroné ou expliquer un code.

Ces exemples montrent les nombreux défis commerciaux que les modèles d'IA générative peuvent aider à résoudre. La clé est d'identifier les défis spécifiques les plus urgents pour un domaine spécifique.

entreprise ou industrie, puis pour déterminer comment les modèles d'IA génératifs peuvent être utilisés pour relever ces défis.

Architecture des solutions Dell et NVIDIA

Haut niveau architecture

Dell Technologies et NVIDIA ouvrent la voie depuis des années en proposant des innovations conjointes pour l'IA et le calcul haute performance. Avec ce projet, nous avons conçu conjointement une solution complète centrée sur les flux de travail qui permet aux entreprises de créer et d'exécuter des modèles d'IA génératifs à n'importe quelle échelle, de l'expérimentation de l'IA à la production de l'IA.

L'architecture est modulaire, évolutive et équilibre performances et efficacité. La modularité permet à l'architecture de prendre en charge de nombreux flux de travail d'IA différents, comme expliqué dans les sections suivantes.

Esprit de modularité

La pierre angulaire de cette architecture commune est la modularité, offrant une conception flexible qui répond à une multitude de cas d'utilisation, de secteurs et d'exigences informatiques. Une infrastructure d'IA véritablement modulaire est conçue pour être adaptable et évolutive, avec des composants qui peuvent être mélangés et assortis en fonction des exigences spécifiques du projet. La solution Dell-NVIDIA utilise cette approche, permettant aux entreprises de se concentrer sur certains aspects des charges de travail d'IA générative lors de la construction de leur infrastructure. Cette approche modulaire est réalisée grâce à des conceptions de cas d'utilisation spécifiques pour la formation, le réglage du modèle et l'inférence qui utilisent efficacement chaque type de calcul. Chaque conception commence par l'unité minimale pour chaque cas d'utilisation, avec des options d'extension.

Une pile logicielle modulaire est également essentielle pour permettre aux chercheurs en IA, aux data scientists, aux ingénieurs de données et aux autres utilisateurs de concevoir rapidement leur infrastructure et d'obtenir une rentabilisation rapide. La solution Dell-NVIDIA utilise le meilleur des logiciels d'IA NVIDIA, avec des solutions partenaires pour créer une plate-forme d'IA adaptable et prise en charge à chaque couche : du système d'exploitation au planificateur en passant par plusieurs opérations d'IA (AIOps) et d'apprentissage automatique (MLOps).) solutions.

La figure suivante présente une vue générale de l'architecture de la solution, en mettant l'accent sur la pile logicielle, depuis la couche d'infrastructure jusqu'au logiciel d'application d'IA :

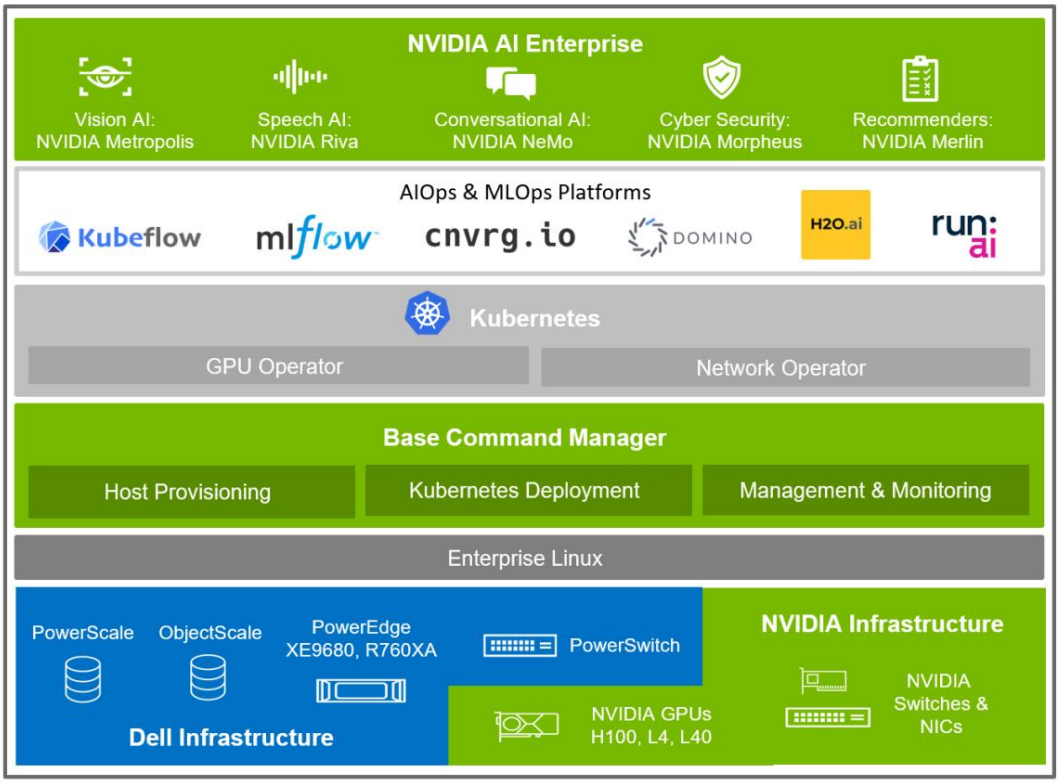


Figure 1. Architecture de la solution et pile logicielle

À un niveau élevé, l'architecture de la solution commence par les composants matériels de base de Dell Technologies et NVIDIA, qui sont combinés dans des permutations axées sur des charges de travail d'IA spécifiques, telles que la formation, le réglage fin et l'inférence. Ce livre blanc décrit les composants matériels individuels dans une section ultérieure.

Chaque plan de contrôle ou élément de calcul prend en charge Red Hat Enterprise Linux ou Ubuntu comme système d'exploitation, qui est préchargé avec des pilotes GPU NVIDIA et Compute Unified Device Architecture (CUDA) pour une utilisation sans système d'exploitation.

NVIDIA Base Command Manager (BCM) sert de gestionnaire de cluster en installant des logiciels sur les systèmes hôtes du cluster, en déployant Kubernetes et en surveillant l'état du cluster. Le provisionnement des hôtes est essentiel au bon fonctionnement d'un cluster, avec la possibilité de charger le système d'exploitation, le pilote, le micrologiciel et d'autres logiciels critiques sur chaque système hôte. Le déploiement de Kubernetes comprend l'installation d'un opérateur GPU et d'un opérateur réseau, un élément essentiel de l'activation du GPU et de la structure réseau. NVIDIA BCM prend en charge la gestion des hôtes avec et sans état, en suivant chaque système, son état de santé et en collectant des métriques que les administrateurs peuvent visualiser en temps réel ou regrouper dans des rapports.

Au niveau supérieur de la solution se trouve le logiciel NVIDIA AI Enterprise qui accélère le pipeline de science des données et rationalise le développement et le déploiement de l'IA de production, notamment l'IA générative, la vision par ordinateur, l'IA vocale, etc. Qu'il s'agisse de développer initialement un nouveau modèle d'IA ou d'utiliser l'un des workflows d'IA de référence comme modèle pour commencer,

NVIDIA AI Enterprise propose des logiciels de bout en bout sécurisés, stables, en croissance rapide et entièrement pris en charge par NVIDIA.

Avec Kubernetes déployé dans la solution, plusieurs solutions MLOps différentes peuvent être installées, qu'il s'agisse de solutions open source comme Kubeflow et MLFlow, ou de solutions prises en charge telles que cnvrg.io, Domino, H2O.ai, Run:ai, etc. . Chacune de ces solutions peut être déployée pour fonctionner dans un scénario multicluster et cloud hybride.

Modules architecturaux

L'architecture de la solution d'IA générative répond à trois flux de travail principaux :

- Inférence de grands modèles
- Personnalisation de grands modèles (réglage fin et réglage P)
- Formation sur grands modèles

Chacun de ces flux de travail a des exigences distinctes en matière de calcul, de stockage, de réseau et de logiciels. La conception de la solution est modulaire et chacun des composants peut être mis à l'échelle indépendamment en fonction du flux de travail du client et des exigences de l'application. En outre, certains modules sont facultatifs ou échangeables avec des solutions existantes équivalentes dans l'infrastructure d'IA d'une organisation, telles que leur module MLOps et Data Prep préféré ou leur module de données préféré. Le tableau suivant montre les fonctionnalités modules dans l'architecture de la solution :

Tableau 1. Modules d'architecture fonctionnelle pour la solution d'IA générative

Module	Description
Entraînement	Module pour serveurs optimisés par l'IA pour la formation, alimenté par Serveurs PowerEdge XE9680 et XE8640 avec GPU NVIDIA H100
Inférence	Module pour serveurs optimisés par l'IA pour l'inférence, alimentés par Serveurs PowerEdge XE9680 avec serveurs NVIDIA H100 ou R760xa avec GPU NVIDIA L40 ou L4
Gestion	Module de gestion du système et du cluster, comprenant un nœud principal pour NVIDIA BCM, alimenté par des serveurs PowerEdge R660.
MLOps et préparation des données	Module pour les opérations d'apprentissage automatique et la préparation des données pour l'exécution du logiciel MLOps, de la base de données et d'autres tâches basées sur le processeur pour la préparation des données, optimisé par des serveurs PowerEdge R660
Données	Module pour le stockage en réseau évolutif et à haut débit (NAS) optimisé par Dell PowerScale, ainsi qu'un stockage d'objets évolutif à haut débit optimisé par Dell ECS et ObjectScale
InfiniBande	Module pour une communication GPU à GPU à très faible latence et à large bande passante, alimenté par les commutateurs NVIDIA QM9700 InfiniBand
Ethernet	Module pour une communication à haut débit et à large bande passante entre d'autres modules de la solution optimisée par Dell PowerSwitch Z9432F-ON

Évolutivité

Dans l'architecture de la solution, les modules fonctionnels peuvent être évolutifs en fonction des cas d'utilisation et des exigences de capacité. Par exemple, l'unité minimale du module de formation pour
La formation sur les grands modèles comprend huit serveurs PowerEdge XE9680 avec 64 GPU NVIDIA H100.

À titre d'exemple théorique, le module de formation avec un module InfiniBand pourrait former un modèle de paramètres 175B en 112 jours. Pour illustrer l'évolutivité, six exemplaires de ceux-ci
les modules pourraient entraîner le même modèle en 19 jours. Comme autre exemple, si vous entraînez un modèle de paramètres 40B, deux copies du module de formation suffisent pour entraîner le modèle en 14 jours.

Il existe un concept d'évolutivité similaire pour le module InfiniBand. Par exemple, un module doté de deux commutateurs QM9700 peut prendre en charge jusqu'à 24 serveurs PowerEdge XE9680. Si vous doublez le module InfiniBand, dans une architecture fat-tree, vous pouvez évoluer jusqu'à 48
Serveurs PowerEdge XE9680. Le module Ethernet et les modules d'inférence fonctionnent de la même manière.

Le module Data est alimenté par des solutions de stockage à architecture de stockage évolutive, qui peut évoluer de manière linéaire pour répondre aux exigences de performances et de capacité, à mesure que vous augmentez le nombre de serveurs et de GPU dans vos modules de formation et d'inférence.

L'évolutivité et la modularité sont intrinsèques à la conception Dell et NVIDIA pour l'IA générative à tous les niveaux.

Sécurité

L'approche Dell en matière de sécurité est intrinsèque par nature : elle est intégrée, et non ajoutée ultérieurement, et elle est intégrée à chaque étape du cycle de vie de développement sécurisé Dell. Nous nous efforçons de faire évoluer continuellement nos contrôles, fonctionnalités et solutions de sécurité PowerEdge pour répondre au paysage de menaces toujours croissant, et nous continuons d'ancrer la sécurité avec une racine de confiance en silicium.

Les fonctionnalités de sécurité sont intégrées à la plateforme PowerEdge Cyber Resilient, activées par le contrôleur d'accès à distance Dell intégré (iDRAC). De nombreuses fonctionnalités ont été ajoutées au système, allant du contrôle d'accès au cryptage des données en passant par l'assurance de la chaîne d'approvisionnement.
Ces fonctionnalités incluent l'analyse du BIOS en direct, la personnalisation du démarrage sécurisé UEFI, la MFA RSA Secure ID, la gestion sécurisée des clés d'entreprise (SEKM), la vérification des composants sécurisés (SCV), l'effacement amélioré du système, l'inscription et le renouvellement automatiques des certificats, la sélection de chiffrement et la prise en charge CNSA. Toutes les fonctionnalités font largement appel à l'intelligence et à l'automatisation pour vous aider à garder une longueur d'avance sur les menaces et pour permettre l'évolutivité exigée par des modèles d'utilisation en constante expansion.

À mesure que les entreprises se tournent vers l'IA de production, il peut s'avérer difficile de maintenir une plateforme d'IA sécurisée et stable. Ce défi est particulièrement vrai pour les entreprises qui ont construit leur propre plate-forme d'IA à l'aide de bibliothèques et de frameworks d'IA open source et non pris en charge. Pour répondre à ce problème et minimiser la charge liée à la maintenance d'une plate-forme d'IA, l'abonnement au logiciel NVIDIA AI Enterprise comprend une surveillance continue des vulnérabilités de sécurité, des corrections continues et des correctifs de sécurité, ainsi que des notifications prioritaires des vulnérabilités critiques.
Cette surveillance permet aux développeurs d'entreprise de se concentrer sur la création d'applications d'IA innovantes au lieu de maintenir leur plate-forme de développement d'IA. De plus, maintenir la stabilité de l'API peut s'avérer difficile en raison des nombreuses dépendances open source. Avec NVIDIA AI Enterprise, les entreprises peuvent compter sur la stabilité des API en utilisant une branche de production qui
Les experts NVIDIA AI soutiennent. L'accès aux experts du support NVIDIA signifie que les projets d'IA restent sur la bonne voie.

Considérations sur les composants d'infrastructure pour l'IA

Il existe de nombreuses considérations importantes concernant les différents composants de l'infrastructure matérielle d'un système d'IA générative, notamment le calcul haute performance, la mise en réseau à haut débit et le stockage évolutif, de grande capacité et à faible latence, pour n'en nommer que quelques-uns.

Calculer

Les modèles d'IA générative nécessitent une puissance de calcul importante, en particulier pendant la phase de formation, car ils impliquent généralement une multiplication matricielle à grande échelle et d'autres opérations gourmandes en calcul. Pour la formation, il est courant d'utiliser de nombreux GPU puissants pour accélérer le processus. Pour l'inférence, un matériel moins puissant peut être utilisé, mais une puissance de calcul importante est nécessaire pour fournir des réponses rapides.

Accélérateurs

Comme mentionné précédemment, les accélérateurs tels que les GPU sont souvent utilisés pour accélérer le processus de formation. Ces accélérateurs sont spécialement conçus pour le traitement parallèle de grandes quantités de données, ce qui les rend bien adaptés à la multiplication matricielle et à d'autres opérations requises par les modèles d'IA génératifs. Outre le matériel spécialisé, il existe également des techniques d'accélération logicielles telles que la formation de précision mixte, qui peuvent accélérer le processus de formation en réduisant la précision de certains calculs.

Stockage

Les modèles d'IA générative peuvent être importants, avec de nombreux paramètres et résultats intermédiaires. Ce volume signifie que les modèles nécessitent des quantités de stockage importantes pour contenir toutes les données. Il est courant d'utiliser des systèmes de stockage distribués tels que Hadoop ou Spark pour stocker les données de formation et les sorties intermédiaires pendant la formation. À des fins d'inférence, il peut être possible de stocker le modèle sur un disque local, mais pour les modèles plus grands, il peut être nécessaire d'utiliser un stockage en réseau ou des solutions de stockage basées sur le cloud. Des composants de stockage évolutifs, de grande capacité et à faible latence pour les objets fichiers et les magasins de fichiers sont essentiels dans les systèmes d'IA.

La mise en réseau

La mise en réseau est une considération importante pour l'IA générative, en particulier dans les scénarios de formation distribuée. Pendant la formation, les données sont généralement distribuées sur plusieurs nœuds, chacun avec son propre accélérateur et son propre stockage. Ces nœuds doivent communiquer fréquemment entre eux pour échanger des données et mettre à jour le modèle. Les solutions de mise en réseau à haut débit telles qu'InfiniBand ou RDMA sont souvent utilisées pour minimiser la latence de ces communications.
et améliorer considérablement les performances du processus de formation.

Résumé

L'IA générative nécessite des quantités importantes de puissance de calcul et de stockage, et implique souvent l'utilisation d'accélérateurs spécialisés tels que les GPU. En outre, les solutions de mise en réseau à haut débit sont importantes pour minimiser la latence lors de la formation distribuée. En examinant attentivement ces exigences, les entreprises peuvent créer et déployer des modèles d'IA génératifs rapides, efficaces et précis.

Infrastructure Dell et composants logiciels

Cette section décrit les principaux composants matériels et logiciels Dell utilisés dans l'architecture de la solution d'IA générative.

Dell PowerEdge les serveurs

Dell Technologies propose une gamme de serveurs optimisés pour l'accélération et une vaste gamme d'accélérations avec des GPU NVIDIA. Deux serveurs Dell sont présentés dans la solution d'IA générative.

L'approche de calcul adaptatif PowerEdge permet aux serveurs conçus pour optimiser les dernières avancées technologiques pour des résultats rentables prévisibles. Les améliorations apportées au portefeuille PowerEdge incluent :

- Concentration sur l'accélération : prise en charge de la gamme la plus complète de GPU, offrant des performances maximales pour l'IA, l'apprentissage automatique et l'apprentissage profond formation et inférence, modélisation et simulation du calcul haute performance (HPC), analyses avancées et suites d'applications et charges de travail de collaboration riche
- Conception thermique réfléchie : nouvelles solutions et conceptions thermiques pour répondre des composants denses produisant de la chaleur et, dans certains cas, des conceptions refroidies par air d'avant en arrière
- Refroidissement multivecteur Dell : conception thermique rationalisée et avancée pour la circulation de l'air chemins au sein du serveur

Serveur PowerEdge XE9680

Le serveur PowerEdge XE9680 est un serveur d'applications hautes performances conçu pour les charges de travail exigeantes d'IA, d'apprentissage automatique et d'apprentissage profond qui vous permettent de développer, former et déployer rapidement de grands modèles d'apprentissage automatique.

Le serveur PowerEdge XE9680 est le premier serveur du secteur à être livré avec huit GPU NVIDIA H100 et le logiciel NVIDIA AI. Il est conçu pour maximiser le débit de l'IA, en fournissant aux entreprises une plate-forme hautement raffinée, systématisée et évolutive pour les aider à réaliser des percées en matière de PNL, de systèmes de recommandation, d'analyse de données, etc.

Son châssis 6U refroidi par air prend en charge la nouvelle génération de puissance la plus élevée technologies jusqu'à 35°C ambiant. Il offre neuf fois plus de performances et deux fois plus une mise en réseau plus rapide avec les cartes d'interface réseau intelligentes NVIDIA ConnectX-7 (SmartNIC) et une évolutivité à grande vitesse pour NVIDIA SuperPOD.

Serveur PowerEdge XE8640

Le serveur PowerEdge XE8640 est un serveur 4U refroidi par air aux performances optimisées, doté de quatre GPU NVIDIA H100 Tensor Core et de la technologie NVIDIA NVLink, ainsi que de deux prochains processeurs Intel Xeon Scalable de 4e génération. Il est conçu pour aider les entreprises à se développer, à se former et à déployez des modèles d'apprentissage automatique pour accélérer et automatiser l'analyse.

Serveur PowerEdge R760xa

Optimisé pour les GPU PCIe, le serveur PowerEdge R760xa 2U à double socket permet aux entreprises d'accélérer une grande variété d'applications, notamment la formation et l'inférence de l'IA, l'analyse, la virtualisation et les applications de rendu des performances, le tout dans une conception refroidie par air.
Le serveur PowerEdge R760xa offre des performances exceptionnelles grâce aux processeurs Intel et prend en charge un ensemble diversifié d'accélérateurs GPU d'AMD, Intel et NVIDIA pour répondre à une gamme large et puissante de besoins de traitement exigeants.
Déployez et activez des applications graphiques exigeantes et des applications d'inférence d'IA denses à l'échelle de l'entreprise avec des fonctionnalités et des capacités puissantes, en utilisant les dernières technologies.

Stockage de fichiers Dell

Dell PowerScale prend en charge les charges de travail d'IA les plus exigeantes avec des solutions de stockage de fichiers NVMe 100 % Flash qui offrent des performances et une efficacité exceptionnelles dans un format compact.

Il existe plusieurs modèles utilisés dans l'architecture de la solution d'IA générative, tous alimentés par le système d'exploitation PowerScale OneFS et prenant en charge la compression et la déduplication des données en ligne. Le nombre minimum de nœuds PowerScale par cluster est de trois nœuds et la taille maximale du cluster est de 252 nœuds.

PowerScale F900

PowerScale F900 offre les performances maximales des disques 100 % NVMe dans une configuration rentable pour répondre aux besoins de stockage des charges de travail d'IA exigeantes.
Chaque nœud mesure 2U de hauteur et héberge 24 SSD NVMe.
PowerScale F900 prend en charge les disques TLC ou QLC pour des performances maximales. Il vous permet de faire évoluer le stockage brut de 46 To à 736 To par nœud et jusqu'à 186 Po de capacité brute par cluster.

PowerScale F600

PowerScale F600 comprend des disques NVMe pour offrir une plus grande capacité avec des performances massives dans un format 1U compact et économique pour alimenter les charges de travail exigeantes. Le PowerScale F600 prend en charge les disques TLC ou QLC pour des performances maximales. Chaque nœud vous permet de faire évoluer la capacité de stockage brute de 15,36 To à 245 To et jusqu'à 60 Po de capacité brute par cluster.

Stockage d'objets Dell

Dell Technologies propose un choix de produits de stockage basés sur les objets, tous évolutifs et économiques pour les volumes élevés de données non structurées destinés aux charges de travail d'IA.

Dell ECS

Le stockage objet d'entreprise ECS combine la simplicité de S3 avec des performances extrêmes à grande échelle pour les charges de travail modernes telles que les applications d'IA, d'apprentissage automatique et d'analyse en temps réel. L'ECS EXF900 offre des performances NVMe 100 % Flash avec une capacité évolutive jusqu'à 5 898 Po par rack, ainsi que des performances 21 fois plus rapides* que la génération précédente. L'utilisation d'ECS pour alimenter les serveurs GPU avec un stockage à débit optimisé expose rapidement les algorithmes et les applications de formation à davantage de fonctionnalités.
données que jamais auparavant.

*Basé sur une analyse interne de Dell Technologies comparant la bande passante maximale de l'ECS EXF900 (511 Mo/s) à la bande passante maximale de l'ECS EX300 (24 Mo/s) pour 10 Ko d'écritures, novembre 2020. Les performances réelles varient.

Dell ObjectScale

ObjectScale est un stockage d'objets défini par logiciel qui offre des performances à grande échelle pour prendre en charge l'IA.
charges de travail. Il fournit des ensembles de données à des taux de transfert élevés aux serveurs CPU et GPU les plus exigeants, exposant les algorithmes d'entraînement de l'IA à davantage de données sans introduire la complexité du stockage HPC. Ce stockage inclut une prise en charge rapide et stable pour des objets allant jusqu'à 30 To. Les clusters peuvent être facilement étendus pour améliorer les performances et la capacité de manière linéaire. Avec la possibilité de déployer sur des disques 100 % Flash basés sur NVMe, les performances de stockage ne constituent plus un goulot d'étranglement. De plus, le marquage d'objets fournit aux modèles d'inférence des ensembles de données plus riches à partir desquels effectuer des prédictions plus intelligentes.

Dell
Interrupteur
la mise en réseau

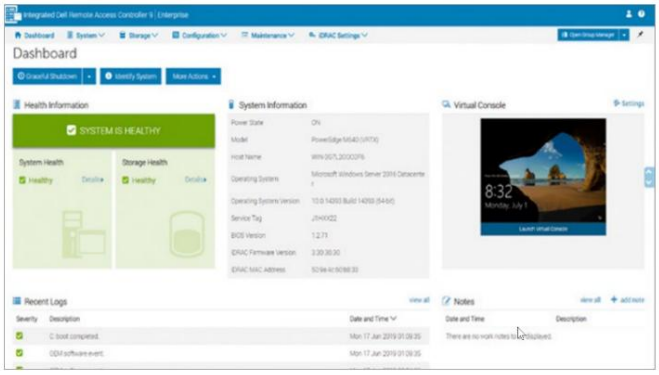
La technologie réseau prête pour l'avenir vous aide à améliorer les performances du réseau, ce qui réduit les coûts globaux et la complexité de la gestion du réseau, et offre la flexibilité nécessaire pour adopter de nouvelles innovations.

Le commutateur fixe Dell PowerSwitch Z9432F-ON 100/400GbE est composé du dernier des solutions de mise en réseau matérielles et logicielles désagrégées pour les centres de données, fournissant des ports 100/400 GbE haute densité de pointe et une large gamme de fonctionnalités pour répondre aux demandes croissantes de l'environnement des centres de données d'aujourd'hui. Ce commutateur d'agrégation haute densité de réseau ouvert innovant de nouvelle génération offre une flexibilité et une rentabilité optimales pour les fournisseurs de services Web 2.0, d'entreprise, de taille intermédiaire et cloud avec
environnements de trafic de calcul et de stockage exigeants.

Dell OpenManage Entreprise

La gestion informatique est la base du succès opérationnel et l'exécution d'un grand système multi-nœuds nécessaire aux charges de travail d'IA générative peut s'avérer particulièrement complexe.

OpenManage Enterprise réduit le temps et les efforts nécessaires à la gestion des implémentations informatiques. Il permet des capacités de gestion du cycle de vie des serveurs qui génèrent de la valeur grâce à des efficacités en temps réel et à des économies de coûts.



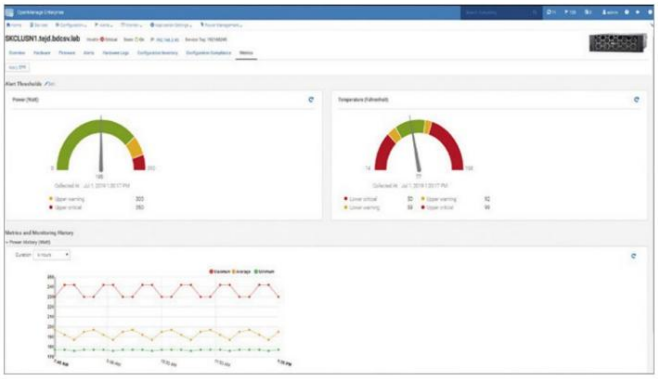
OpenManage Enterprise stimule l'innovation grâce à une analyse prédictive, des informations supplémentaires et un contrôle étendu qui améliorent la sécurité, améliorent l'efficacité et accélèrent le délai de rentabilisation. Il prend en charge jusqu'à 8 000 appareils, gère les serveurs Dell et surveille l'infrastructure de réseau et de stockage Dell ainsi que les produits tiers.

Grâce à l'automatisation intelligente, OpenManage Enterprise offre une gestion de la configuration du cycle de vie complet avec des modèles modifiables et une gestion de la configuration avec détection des dérives du micrologiciel. Il dispose également d'une architecture de plug-in extensible et de capacités de gestion à distance rationalisées.

Dell OpenManage Puissance d'entreprise Directeur

Puissance d'entreprise OpenManage Manager vous permet d'optimiser la disponibilité du centre de données, de contrôler la consommation d'énergie, de surveiller et de budgétiser la puissance du serveur en fonction des besoins de consommation et de charge de travail, ainsi que de surveiller la température.

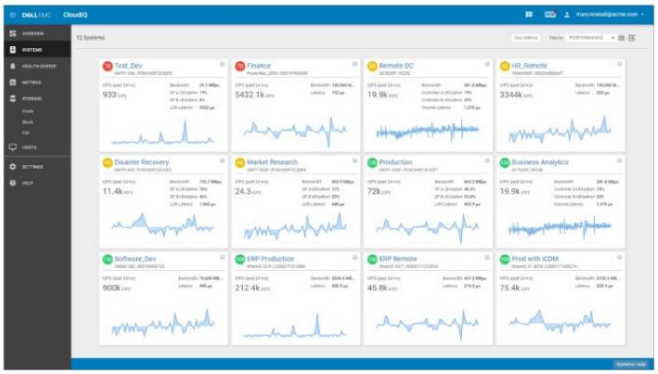
conditions. Étant donné que l'IA générative implique des charges de travail exigeantes et gourmandes en ressources, il est essentiel de gérer efficacement la consommation d'énergie.



Grâce à l'intégration avec le contrôleur d'accès à distance Dell (iDRAC) intégré à tous les serveurs PowerEdge, vous pouvez définir des contrôles basés sur des politiques pour maximiser l'utilisation des ressources et réduire la puissance lorsque la demande de performances diminue. En utilisant des politiques d'alimentation prédéfinies, OpenManage Enterprise Power Manager peut contribuer à atténuer les risques opérationnels et garantir que vos serveurs et leurs charges de travail clés continuent de fonctionner.

Dell CloudIQ

CloudIQ est une application cloud de surveillance proactive et d'analyse prédictive pour le portefeuille d'infrastructures Dell. Il combine l'intelligence humaine de l'ingénierie experte et l'intelligence artificielle de l'IA et de l'apprentissage automatique pour vous fournir les informations nécessaires pour gérer votre infrastructure informatique de manière efficace et proactive afin de répondre à la demande de votre entreprise.



CloudIQ intègre les données de toutes vos consoles OpenManage Enterprise Power Manager pour surveiller l'état, la capacité, les performances et la cybersécurité des composants Dell.
sur tous vos sites.

Le portail CloudIQ affiche vos systèmes d'infrastructure Dell dans une seule vue pour simplifier la surveillance de vos centres de données principaux et secondaires et de vos emplacements périphériques, ainsi que la protection des données dans les cloud publics. Avec CloudIQ, vous pouvez facilement garantir que les charges de travail critiques de l'entreprise obtiennent la capacité et les performances dont elles ont besoin, passer moins de temps à surveiller et à dépanner l'infrastructure, et consacrer plus de temps à innover et à vous concentrer sur des projets qui ajoutent de la valeur à votre entreprise.

Services Dell

Dell Technologies fournit plusieurs services, reliant les personnes, les processus et la technologie pour accélérer l'innovation et permettre des résultats commerciaux optimaux pour les solutions d'IA et tous les besoins de votre centre de données.

Des services de consultation

Les services de conseil vous aident à créer un avantage concurrentiel pour votre entreprise. Nos consultants experts travaillent avec des entreprises à toutes les étapes de l'analyse des données pour vous aider à planifier, mettre en œuvre et optimiser des solutions qui vous permettent de libérer votre capital de données et de prendre en charge des techniques avancées, telles que l'IA, l'apprentissage automatique et l'apprentissage profond.

Services de déploiement

Les services de déploiement vous aident à rationaliser la complexité et à mettre en ligne de nouveaux investissements informatiques le plus rapidement possible. Utilisez nos plus de 30 années d'expérience pour un déploiement de solutions efficace et fiable afin d'accélérer l'adoption et le retour sur investissement (ROI) tout en libérant le personnel informatique pour un travail plus stratégique.

Services de soutien

Les services d'assistance basés sur l'IA et l'apprentissage profond changeront votre façon de concevoir l'assistance grâce à une technologie intelligente et révolutionnaire soutenue par des experts pour vous aider à maximiser la productivité, la disponibilité et la commodité. Découvrez bien plus qu'une résolution rapide des problèmes –
notre moteur d'IA détecte et prévient de manière proactive les problèmes avant qu'ils n'aient un impact sur les performances.

Services gérés

Les services gérés peuvent vous aider à réduire les coûts, la complexité et les risques liés à la gestion informatique afin que vous puissiez concentrer vos ressources sur l'innovation et la transformation numériques tandis que nos experts vous aident à optimiser vos opérations et investissements informatiques.

Services de résidence

Les services de résidence fournissent l'expertise nécessaire pour conduire une transformation informatique efficace et maintenir l'infrastructure informatique à son apogée. Les experts résidents travaillent sans relâche pour relever les défis et répondre aux exigences, avec la capacité de s'adapter à mesure que les priorités évoluent.

Infrastructure NVIDIA et composants logiciels

Cette section décrit les principaux composants logiciels d'accélération matérielle et d'IA NVIDIA utilisés dans l'architecture de la solution d'IA générative.

Accélérateurs NVIDIA

Les GPU NVIDIA suivants font partie des composants d'accélération NVIDIA utilisés dans ce document. architecture de solution d'IA générative.

GPU NVIDIA H100 Tensor Core

Le GPU NVIDIA H100 Tensor Core offre des performances, une évolutivité et une sécurité sans précédent pour chaque charge de travail. Avec le système de commutation NVLInk NVIDIA de quatrième génération, le GPU NVIDIA H100 accélère les charges de travail d'IA avec un

Transformer Engine pour des modèles de langage de paramètres de milliers de milliards. Le GPU NVIDIA H100 utilise des innovations révolutionnaires dans l'architecture NVIDIA Hopper pour fournir une IA conversationnelle de pointe, accélérant les grands modèles de langage 30 fois par rapport à la génération précédente.

Pour les petits travaux, le GPU NVIDIA H100 peut être partitionné en partitions GPU multi-instances (MIG) de bonne taille. Avec Hopper Confidential Computing, cette puissance de calcul évolutive peut sécuriser les applications sensibles sur l'infrastructure de centre de données partagée. L'inclusion de la suite logicielle NVIDIA AI Enterprise réduit le temps de développement et simplifie le déploiement des charges de travail d'IA et fait du GPU NVIDIA H100 la plate-forme de centre de données IA et HPC de bout en bout la plus puissante.

GPU NVIDIA L40

L'accélérateur GPU NVIDIA L40 est une solution graphique pleine hauteur et pleine longueur (FHFL) à double emplacement PCI Express Gen4 de 10,5 pouces basée sur la dernière architecture NVIDIA Ada Lovelace. La carte est refroidie passivement et est capable d'une puissance maximale de 300 W.

Le GPU NVIDIA L40 prend en charge le dernier traçage de rayons accéléré par le matériel, les fonctionnalités d'IA révolutionnaires, l'ombrage avancé et de puissantes capacités de simulation pour un large éventail de cas d'utilisation graphiques et informatiques dans les déploiements de centres de données et de serveurs périphériques. Cette prise en charge inclut NVIDIA Omniverse, les jeux dans le cloud, le rendu par lots, les postes de travail virtuels et la formation en apprentissage profond ainsi que les charges de travail d'inférence.

Faisant partie de la plate-forme serveur NVIDIA OVX, le GPU NVIDIA L40 offre le plus haut niveau de performances graphiques, de traçage de rayons et de simulation pour NVIDIA Omniverse. Avec 48 Go de mémoire GDDR6, même les applications graphiques les plus intenses fonctionnent avec le plus haut niveau de performances.

GPU NVIDIA L4 Tensor Core

Le GPU NVIDIA Ada Lovelace L4 Tensor Core offre une accélération et efficacité énergétique pour la vidéo, l'IA, les postes de travail virtualisés et les applications graphiques dans l'entreprise, dans le cloud et à la périphérie. Grâce à la plateforme IA et à l'approche full-stack de NVIDIA, le GPU NVIDIA L4 est optimisé pour l'inférence à grande échelle pour une large gamme d'applications IA, notamment les recommandations, les assistants d'avatar IA vocaux, l'IA générative, la recherche visuelle et l'automatisation des centres de contact pour offrir les meilleures expériences personnalisées.

Le GPU NVIDIA L4 est l'accélérateur NVIDIA le plus efficace pour une utilisation grand public. Les serveurs équipé du GPU NVIDIA L4, il offre des performances vidéo IA jusqu'à 120 fois supérieures et des performances IA génératives 2,5 fois supérieures par rapport aux solutions CPU, ainsi que des performances graphiques plus de quatre fois supérieures à celles de la génération GPU précédente. La polyvalence du GPU NVIDIA L4 et son facteur de forme économe en énergie, à emplacement unique et à profil bas le rendent idéal pour les déploiements mondiaux, y compris dans les emplacements périphériques.

NVIDIA NVLink et NVSwitch

NVIDIA NVLink est une interconnexion rapide et évolutive qui permet aux systèmes multi-nœuds et multi-GPU d'assurer une communication transparente et à haut débit entre chaque GPU. La quatrième génération de la technologie NVIDIA NVLink offre une bande passante 1,5 fois supérieure et une évolutivité améliorée pour les configurations système multi-GPU. Un seul GPU NVIDIA H100 Tensor Core prend en charge jusqu'à 18 connexions NVLink pour une bande passante totale de 900 gigaoctets par seconde (Go/s), soit plus de sept fois la bande passante du PCIe Gen5.

Pour une évolutivité encore plus grande, NVIDIA NVSwitch s'appuie sur la capacité de communication avancée de NVIDIA NVLink pour offrir une bande passante plus élevée et une latence réduite pour les charges de travail gourmandes en calcul. Pour permettre des opérations collectives à grande vitesse, chaque NVIDIA NVSwitch dispose de 64 ports NVIDIA NVLink équipés de moteurs pour le protocole SHARP (Scalable Hierarchical Aggregation Reduction Protocol) NVIDIA pour les réductions en réseau et l'accélération de la multidiffusion.

Logiciel NVIDIA IA

Les solutions logicielles d'entreprise NVIDIA sont conçues pour offrir aux administrateurs informatiques, aux data scientists, les architectes et les concepteurs accèdent aux outils dont ils ont besoin pour gérer et optimiser facilement leurs systèmes accélérés.

NVIDIA IA Entreprise

NVIDIA AI Enterprise, la couche logicielle de la plateforme NVIDIA AI, accélère le pipeline de science des données et rationalise le développement et le déploiement de l'IA de production, notamment l'IA générative, la vision par ordinateur, l'IA vocale et bien plus encore. Cette plate-forme cloud native de logiciels d'IA sécurisée, stable et comprend plus de 100 frameworks, modèles pré-entraînés et outils qui accélérez le traitement des données, simplifiez la formation et l'optimisation des modèles et rationalisez le déploiement.

- Préparation des données : améliorez jusqu'à 5 fois le temps de traitement des données réduisant les coûts opérationnels de 4 fois avec l'accélérateur NVIDIA RAPIDS pour Apache Spark.

- Formation IA : créez des modèles personnalisés et précis en quelques heures, au lieu de plusieurs mois. à l'aide de NVIDIA TAO Toolkit et de modèles pré-entraînés.

- Optimisation pour l'inférence : accélère les performances des applications jusqu'à 40 fois sur les plates-formes CPU uniquement lors de l'inférence avec NVIDIA TensorRT.
- Déploiement à grande échelle : simplifiez et optimisez le déploiement de modèles d'IA à grande échelle et en production avec NVIDIA Triton Inference Server.

Disponible dans le cloud, dans le data center et en périphérie, NVIDIA AI Enterprise permet aux organisations de développer une seule fois et de s'exécuter n'importe où. Étant donné que la pile complète est gérée par NVIDIA, les organisations peuvent compter sur des examens de sécurité et des correctifs réguliers, sur la stabilité des API et sur l'accès aux experts en IA et aux équipes d'assistance de NVIDIA pour garantir la continuité des activités et le maintien des projets d'IA sur la bonne voie.

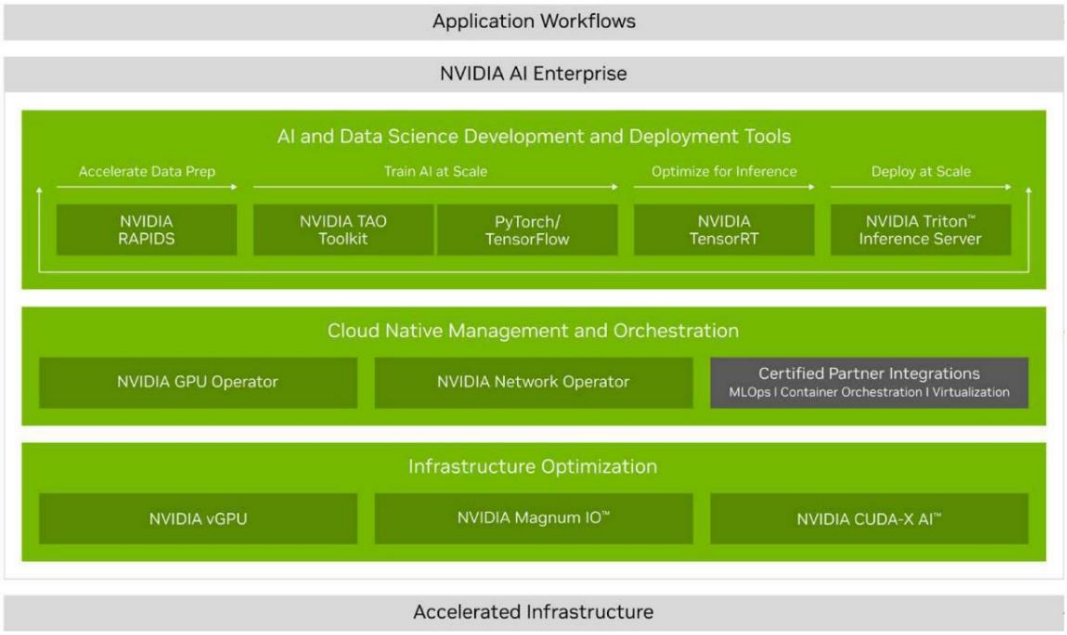


Figure 2. NVIDIA IA Entreprise

NVIDIA AI Enterprise inclut NVIDIA NeMo, un framework permettant de créer, personnaliser et déployer des modèles d'IA génératifs avec des milliards de paramètres. Le framework NVIDIA NeMo fournit un flux de travail accéléré pour la formation aux techniques de parallélisme 3D. Il offre un choix de plusieurs techniques de personnalisation et est optimisé pour l'inférence à grande échelle de modèles à grande échelle pour les applications de langage et d'image, avec des configurations multi-GPU et multi-nœuds. NVIDIA NeMo rend le développement de modèles d'IA génératifs simple, rentable et rapide pour les entreprises.

Gestionnaire de commandes de base NVIDIA

NVIDIA BCM est le gestionnaire de cluster de NVIDIA pour l'infrastructure IA. Il facilite une opérationnalisation transparente du développement de l'IA à grande échelle en fournissant des fonctionnalités telles que le provisionnement du système d'exploitation, les mises à niveau du micrologiciel, la configuration du réseau et du stockage, la planification des tâches multi-GPU et multi-nœuds et la surveillance du système, maximisant ainsi l'utilisation et les performances de l'architecture matérielle sous-jacente.

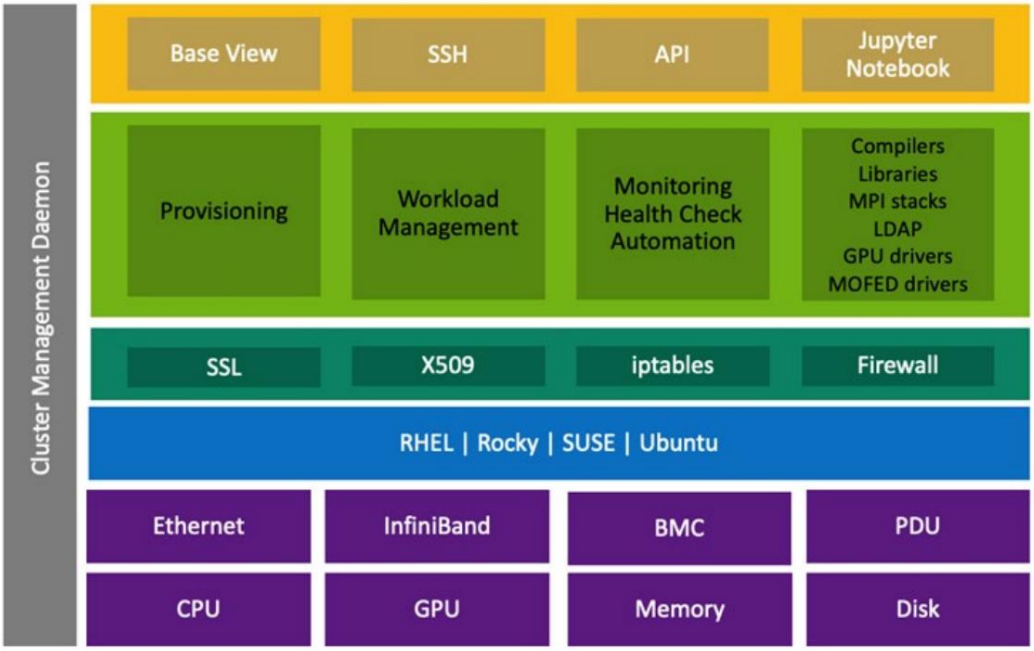


Figure 3. Gestionnaire de commandes de base NVIDIA

NVIDIA BCM prend en charge le provisionnement et la gestion automatiques des modifications apportées aux nœuds tout au long de la durée de vie du cluster.

Avec un cadre extensible et personnalisable, il dispose d'intégrations transparentes avec les multiples gestionnaires de charge de travail HPC, notamment Slurm IBM Spectrum LSF, OpenPBS, Univa Grid Engine et autres. Il offre une prise en charge étendue des technologies de conteneurs, notamment Docker, Harbor, Kubernetes et les opérateurs. Il dispose également d'un solide cadre de gestion de la santé couvrant les mesures, les contrôles de santé et les actions.

Configurations système

Basée sur l'architecture modulaire et évolutive pour l'IA générative décrite précédemment et alimentée par des composants Dell et NVIDIA, cette famille de conceptions propose initialement trois configurations système, chacune axée sur un cas d'utilisation particulier. Les trois configurations système optimisées sont conçues pour les cas d'utilisation d'inférence, de personnalisation et de formation.

Les sections suivantes décrivent les configurations système pour chaque domaine d'intérêt à un niveau élevé. Notez que le plan de contrôle, le stockage des données et la mise en réseau Ethernet pour chaque cas sont similaires. Par conséquent, si vous créez une infrastructure d'IA qui répond à deux cas ou plus, ces ressources de base peuvent être partagées.

Inférence de grand modèle

De nombreuses entreprises choisissent de commencer avec un modèle pré-entraîné et de l'utiliser sans modification ou de procéder rapidement à une ingénierie ou à un réglage P pour mieux utiliser le modèle pour une fonction spécifique. Il est essentiel de commencer par le déploiement en production dans le cas des LLM, car il existe une forte demande en puissance de calcul. Selon la taille du modèle, de nombreux modèles plus grands nécessitent plusieurs systèmes GPU 8x pour atteindre un débit de niveau deuxième ou inférieur à la seconde. La configuration minimale pour déduire des modèles pré-entraînés commence avec un seul serveur PowerEdge R760XA avec jusqu'à quatre GPU NVIDIA H100 ou un serveur PowerEdge XE9680 avec huit GPU NVIDIA H100 en fonction de la taille du modèle et du nombre d'instances. Le nombre de nœuds peut ensuite évoluer selon les besoins en termes de performances ou de capacité, bien que deux nœuds soient recommandés pour des raisons de fiabilité.

Les considérations de conception pour l'inférence de grands modèles incluent :

- Les grands modèles ont tendance à avoir une empreinte mémoire importante. Bien qu'il n'y ait pas de limite claire définissant un grand modèle, par souci de simplicité, tout ce qui dépasse les paramètres 10B peut être considéré comme un grand modèle.
- Lorsque le modèle est divisé entre GPU, la communication entre GPU joue un rôle crucial dans l'obtention de performances optimales. Par conséquent, le logiciel NVIDIA Triton Inference Server avec déploiement multi-GPU utilisant la technologie de transformateur rapide pourrait être utilisé.
- Pour les grands modèles au-dessus des paramètres 40B, nous recommandons le PowerEdge XE9680 serveur. Pour les tailles de modèle inférieures à 40 B de paramètres, le serveur PowerEdge R760xa offre d'excellentes performances.
- Le PowerSwitch Z9432F prend en charge 32 ports de 400 (émetteurs-récepteurs optiques QSFP56-DD) ou jusqu'à 128 ports de 100 GbE. L'inférence n'a pas le module InfiniBand ni l'exigence d'un débit élevé ; par conséquent, il évolue linéairement pour répondre aux besoins de concurrence jusqu'à 32 nœuds.
- Les exigences de débit (inférence par seconde) nécessitent que plusieurs GPU soient déployé en fonction des besoins de la charge de travail.

Personnalisation grand modèle

De nombreuses entreprises renoncent à la formation initiale et choisissent d'utiliser et de personnaliser un modèle pré-entraîné comme base de leur solution. En utilisant le réglage fin et le réglage P, il est possible d'appliquer données spécifiques à l'entreprise pour recycler une partie d'un modèle existant ou créer une meilleure interface d'invite. Cette méthode nécessite beaucoup moins de puissance de calcul que la formation initiale d'un modèle, avec la possibilité de démarrer avec une configuration similaire à la configuration d'inférence uniquement. La principale différence réside dans l'ajout de la mise en réseau InfiniBand entre les systèmes informatiques.

Considérations de conception pour la personnalisation de grands modèles avec réglage fin ou formation P à l'aide
Les grands modèles pré-entraînés comprennent les éléments suivants :

Même si cette tâche est relativement moins gourmande en calcul que celle d'un grand modèle
Lors de la formation, il existe un besoin énorme d'échange d'informations (par exemple, des pondérations) entre les GPU de différents nœuds. Par conséquent, InfiniBand est requis pour optimiser les performances et le débit avec un GPU à huit voies et une connexion NVLInk tout-à-tout. Dans certains cas, lorsque la taille du modèle est inférieure à 40 paramètres B et en fonction des exigences de latence de l'application, le
Le module InfiniBand peut être facultatif.

P-tuning utilise un petit modèle entraînable avant d'utiliser le LLM. Le petit modèle est utilisé pour coder l'invite de texte et générer des jetons virtuels spécifiques à la tâche. Le réglage des invites et des préfixes, qui ajustent uniquement les invites continues avec un modèle de langage figé, réduit considérablement le stockage par tâche et l'utilisation de la mémoire lors de la formation.

Pour les modèles comportant moins de 40 B de paramètres, vous pourrez peut-être utiliser un serveur PowerEdge XE8640. Pour les modèles plus grands, nous recommandons le PowerEdgeXE9680 serveur.

Le module Données est facultatif car il n'y a aucune exigence d'instantané.
Certaines techniques d'ingénierie rapide peuvent nécessiter un grand ensemble de données et nécessiter un module de données performant.

Formation grand modèle

Parmi les trois cas d'utilisation, la formation de grands modèles est la charge de travail la plus exigeante en calcul, les modèles les plus volumineux nécessitant des centres de données dotés d'un grand nombre de GPU pour entraîner un modèle en quelques mois. La configuration minimale pour la formation nécessite huit PowerEdge Serveurs XE9680 avec huit GPU NVIDIA H100 chacun. La plus grande formation de modèle nécessite extension à des tailles de cluster plus grandes de 16 fois, 32 fois ou même des configurations plus grandes.

Les considérations de conception pour la formation sur de grands modèles comprennent :

Les grands modèles d'IA générative ont des exigences de calcul importantes pour la formation.
Selon OpenAI, pour Chat GPT-3 avec 175 B de paramètres, la taille du modèle est d'environ 350 Go, et il faudrait 355 ans pour entraîner GPT-3 sur un seul GPU NVIDIA Tesla V100. Alternativement, il faudrait 34 jours pour s'entraîner avec 1 024
GPU NVIDIA A100.

Le modèle de formation a une empreinte mémoire considérable qui ne rentre pas dans un un seul GPU ; par conséquent, vous devez diviser le modèle sur plusieurs GPU (N-GPU).

La combinaison de la taille du modèle, des techniques de parallélisme pour les performances et de la taille de l'ensemble de données de travail nécessite un débit de communication élevé entre

GPU, bénéficiant ainsi de serveurs PowerEdge XE9680 avec huit GPU NVIDIA entièrement connectés entre eux par NVIDIA NVLink et NVIDIA NVSwitch.

Durant la phase de formation, il y a également une quantité importante d'informations échange (par exemple, poids) entre GPU de différents nœuds ; InfiniBand est requis pour optimiser les performances et le débit.

Le commutateur QM9700 InfiniBand dispose de 64 ports de détection et de réponse réseau (NDR). Par conséquent, 24 nœuds des serveurs PowerEdge XE9680 de ce cluster remplissent les ports du QM9700 dans le module InfiniBand. Ajoutez des modules InfiniBand supplémentaires dans une topologie de réseau Fat Tree.

À mesure que vous ajoutez des nœuds de serveur PowerEdgeXE9680 supplémentaires à votre cluster, développez les commutateurs PowerScale de manière appropriée pour répondre aux exigences de performances d'entrée/sortie.

Le point de contrôle est une technique standard utilisée dans la formation de grands modèles. La taille des points de contrôle dépend de la taille et des dimensions du modèle et du parallélisme du pipeline utilisé dans la formation.

Quatre plates-formes de stockage Dell PowerScale F600 Prime offrent 8 Go en écriture et 40 Performances de débit de lecture GBS avec mise à l'échelle linéaire.

Résumé

Les informations contenues dans cette section constituent un aperçu de haut niveau des caractéristiques et des principales considérations de conception des configurations suggérées pour l'inférence, la personnalisation et la formation de modèles d'IA génératifs en langage étendu. Comme mentionné précédemment, de plus amples détails sur chaque cas d'utilisation suivront ce livre blanc dans une série de guides de conception pour ces conceptions validées par Dell pour l'IA.

Conclusion

Avantage de l'IA générative

Ce document a exploré les concepts, les avantages, les cas d'utilisation et les défis de l'IA générative, et a présenté une architecture de solution évolutive et modulaire conçue par Dell Technologies et NVIDIA.

Le projet Helix est une collaboration unique entre Dell Technologies et NVIDIA qui concrétise la promesse de l'IA générative pour l'entreprise. Ensemble, nous proposons une solution complète, construite sur l'infrastructure et les logiciels Dell, et utilisant la pile logicielle et la technologie d'accélérateur primées de NVIDIA. Réunir les connaissances approfondies et la créativité de NVIDIA avec la connaissance client mondiale et l'expertise technologique de Dell
Technologies, Projet Helix :

- Fournit des solutions d'IA générative complètes construites sur le meilleur de l'infrastructure et des logiciels Dell, en combinaison avec les derniers accélérateurs NVIDIA, les logiciels d'IA et l'expertise en IA.
- Permet aux entreprises d'utiliser sur site une IA générative spécialement conçue pour résoudre des défis commerciaux spécifiques.
- Aide les entreprises tout au long du cycle de vie de l'IA générative, depuis la fourniture de l'infrastructure, le développement et la formation de grands modèles, le réglage fin des modèles pré-entraînés, le déploiement de modèles multisites et l'inférence de grands modèles.
- Garantit la confiance, la sécurité et la confidentialité des données sensibles et propriétaires de l'entreprise, ainsi que le respect des réglementations gouvernementales.

Avec Project Helix, Dell Technologies et NVIDIA permettent aux organisations d'automatiser des processus complexes, d'améliorer les interactions avec les clients et d'ouvrir de nouvelles possibilités grâce à une meilleure intelligence machine. Ensemble, nous ouvrons la voie à la prochaine vague d'innovation dans le paysage de l'IA d'entreprise.

Nous apprécions vos commentaires

Dell Technologies et les auteurs de ce document apprécient vos commentaires sur ce document. Contactez l'équipe Dell Technologies Solutions par e-mail.

Pour plus d'informations sur cette solution, vous pouvez consulter un expert en envoyant un e-mail à AI.Assist@dell.com.

Les références

Ces documents peuvent fournir des informations supplémentaires sur les solutions et les composants présentés ici, ainsi que sur les offres associées.

Dell Les technologies Documentation

La documentation et les ressources Dell Technologies suivantes fournissent des informations supplémentaires et pertinentes par rapport à celles contenues dans ce livre blanc.

- Solutions d'IA Dell Technologies
- Dell Technologies Info Hub pour les solutions d'intelligence artificielle
- Serveurs Dell PowerEdge XE
- Serveurs et accélérateurs (GPU) accélérés Dell PowerEdge
- Stockage Dell PowerScale
- Stockage d'objets d'entreprise Dell ECS
- Stockage Dell ObjectScale
- Commutateurs Dell PowerSwitch série Z
- Gestion des systèmes Dell OpenManage

Nvidia Documentation

La documentation et les ressources NVIDIA suivantes fournissent également des informations supplémentaires et pertinentes :

- NVIDIA AI Entreprise NVIDIA NeMo
- GPU NVIDIA pour centres de données
- Réseau NVIDIA