DELL Technologies

Beknopt overzicht van de oplossing: Validated Design voor Generative AI Inferencing met NVIDIA

Ontgrendel kwalitatieve inzichten nu sneller met GenAI

Implementeer snel een 'full-stack' oplossing voor GenAI (Generative Artificial Intelligence) inferencing voor grote taalmodellen

Hogere productiviteit en diepere inzichten

Deze gezamenlijke architectuur levert een modulair en flexibel ontwerp voor een groot aantal gebruiksscenario's en computationele vereisten. De onderdelen kunnen worden gecombineerd en onafhankelijk worden geschaald op basis van uw gebruiksbehoeftes.

Enkele opmerkelijke voorbeelden van ondersteunde gebruiksscenario's voor inferencing:

Generatie van natuurlijke taal: Generatiemodellen kunnen worden gebruikt voor tekstgeneratietaken, zoals het schrijven van documenten, het genereren van dialogen of het maken van samenvattingen of van content.

Chatbots en virtuele assistenten: GenAI maakt gebruik van gespreksagents, chatbots en virtuele assistenten door antwoorden in natuurlijk taalgebruik te genereren op basis van gebruikersvragen of -instructies.

Ontwikkeling van code: Krijg hulp bij softwareontwikkeling met functies als codeaanvulling, de mogelijkheid om tests voor units te genereren of een chatfunctie om code uit te leggen.

Genereer betere, snellere time-to-value-voorspellingen en -uitvoer, terwijl u de besluitvorming versnelt met een krachtige GenAI-oplossing van Dell Technologies en NVIDIA. Deze gezamenlijk ontworpen oplossing pakt uitdagingen op het gebied van inferencing aan, zoals latentie, reactietijd en computationele vereisten, en helpt bedrijfsgegevens om te zetten in hoogwaardige, slimmere resultaten.

Met innovatieve technologieën, uitgebreide professionele services en een breed partner-ecosysteem kan uw organisatie GenAI op ondernemingsniveau versnellen. Voortaan kunnen IT-organisaties, datawetenschappers en AI DevOps eenvoudig een modulair en schaalbaar platform leveren voor GenAI- en LLM-inferencing.

- Creëer nieuwe waarde met een veilige infrastructuur voor uw bedrijfskritieke bewerkingen
- Mobiliseer en schaal Gen AI-voorspellingen en -inzichten, core-to-edge
- Verbeter de waarde van IT met een strategische begeleiding
- Schaal uw infrastructuur op de juiste grootte en consolideer al uw behoeften voor AI inferencing

Snellere resultaten met een beproefde oplossing

Bouw snel een on-premises infrastructuur voor uw toepassingsbehoeften met een gevalideerde design- en referentiearchitectuur die speciaal is gemaakt voor een eenvoudige ingebruikname. Door de complexiteit bij elke stap te verminderen, genereert u nu meer inzichten en een snellere besluitvorming. Tegelijkertijd verhoogt u de productiviteit.

NVIDIA

Meer informatie

- Zie de ontwerphandleiding
- AI InfoHub
- delltechnologies.com/ai
- Dell Technologies en NVIDIA

Wat is inferencing?

Inferencing in AI verwijst naar het proces waarbij een getraind model wordt gebruikt om voorspellingen te genereren, beslissingen te nemen of uitvoer te genereren op basis van invoerdata. Hierbij gaat het onder meer om het toepassen van geleerde kennis en patronen die zijn opgedaan tijdens de trainingsfase van het model op nieuwe gegevens die nooit eerder door het model zijn ingezien.

Tijdens inferencing neemt het getrainde model invoergegevens op en verwerkt het deze via computationele algoritmen of een neurale netwerkarchitectuur, waarna een uitvoer of voorspelling wordt geproduceerd. Het model past de geleerde parameters, wegingen of regels toe om de invoergegevens om te zetten in zinvolle informatie of acties.

Inferencing is een cruciale fase in de levenscyclus van een AI-systeem. Nadat een model is getraind op gelabelde of niet-gelabelde data om patronen en verbanden te leren, staat inferencing het model toe de kennis te generaliseren en voorspellingen te doen of antwoorden te genereren op basis van daadwerkelijke of onbekende data.

Behaal sneller resultaten met onze hulp

Dell Services experts helpen u de waarde van GenAI voor uw gegevens sneller te realiseren met een serviceportfolio voor elke fase op uw GenAI-traject:

- **Strategie** - Ontwikkel uw roadmap om de innovatieobjecten van uw IT- en zakelijke belanghebbenden te realiseren
- **Implementatie** - Ontwikkel uw platform en maak gebruik van Dell Validated Design om GenAI-inferencinghardware en -software te implementeren
- **Adoptie** - Versnel de waarde van uw GenAI-gebruiksscenario's door een vooraf getraind inferencing-model te implementeren
- **Scale-up** - Beheer uw GenAI-innovatieportfolio met interne technische experts en trainingsaanbiedingen om de vaardigheden van uw team te ontwikkelen

Technische specificaties

De Validated Design-configuraties zijn gebaseerd op de nieuwste, met AI-versnelling geoptimaliseerde Dell PowerEdge XE- en rack servers, met gebruik van de nieuwste NVIDIA GPU's en NVIDIA AI Enterprise met Triton Inference Server en het NeMo-framework. De snelle en ruime data lake-storage voor Generative AI en grote taalmodellen wordt geleverd door Dell PowerScale all-flash-arrays of hybride storage-arrays.

Computergebruik	Accelerators	Networking	Software	Storage
Dell PowerEdge R760xa servers	NVIDIA A100 of H100 GPU's	NVIDIA Networking, Dell PowerSwitch S5232F-ON of S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ. NVIDIA AI Enterprise met Nemo Framework voor LLM's en Triton Inference Server; NVIDIA Base Command Manager Essentials	Ondersteund door Dell PowerScale, ECS en ObjectScale

Dell Technologies en NVIDIA

Dell Technologies en NVIDIA werken samen om generatieve AI-workloads mogelijk te maken en te versnellen, door technici goedgekeurde hardware en software te bieden om AI te versnellen en ML- en DL-workloads te bieden om aan de behoeften van de klant te voldoen binnen alle bedrijven en verticals. Met Validated Design voor LLM-inferencing kunt u AI-oplossingen implementeren om via realtimegegevens uw digitale transformatie te versnellen en belangrijke besluitvorming te verbeteren met oplossingen die zijn geoptimaliseerd voor snellere time-to-value van uw AI-initiatieven.

Meer informatie over Dell oplossingen

Neem contact op met een Dell Technologies expert

Bekijk meer informatiebronnen

Neem deel aan het gesprek via #HashTag

DELL Technologies

NVIDIA