Dell AI Factory

Dell Generative AI-oplossing met AMD

Versnel innovatie, verlaag kosten en bescherm data met een schaalbare en modulaire architectuur voor complexe GenAI.

75% rendabelere inferencing in vergelijking met public cloud IaaS[1]

86% vermindering van de time-to-value[2]

Voor belangrijke gebruiksscenario's zijn kracht, flexibiliteit en schaalbaarheid vereist

- Assistenten, chatbots en contentcreatie
- Accelerator as a service
- Multimodale Retrieval Augmented Generation (RAG)

Vereenvoudigd

Stroomlijn GenAI-implementaties met bewezen gevalideerde oplossingen, ondersteund door meer dan 340.000 uur aan engineering.

Optimaliseer de prestaties
Krachtige accelerator, open architectuur en voor AI geoptimaliseerde fabrics

Overal AI
Overal data beschikbaar met flexibiliteit van multicloudstorage

Meerdere kant-en-klare knooppunten
Bewezen full-stack AI-fundamenten voor snellere resultaten

Maatwerk

AMD ROCm™ open-sourcesoftware en open ecosystemen stimuleren de ontwikkeling en activiteiten van AI.

Sneller innoveren
Gebruik open-sourcesoftware en ecosystemen om unieke applicaties te ontwikkelen.

Ontwikkeling versnellen
Maak gebruik van frameworks die voldoen aan de industriestandaard met flexibele technologiestacks.

Uw data activeren
Voer op efficiënte wijze meerdere AI-gebruiksscenario's tegelijkertijd uit.

Vertrouwd

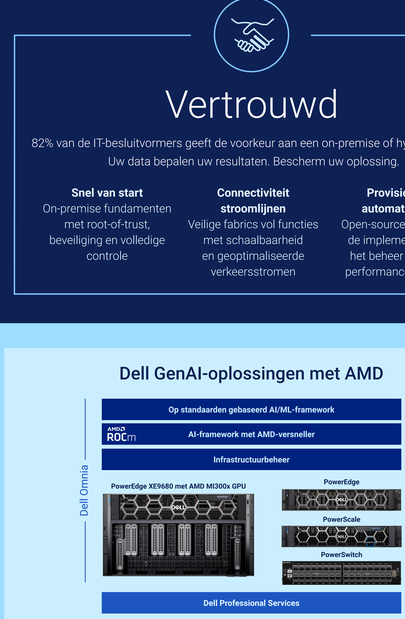
82% van de IT-besluitvormers geeft de voorkeur aan een on-premise of hybride model.[3] Uw data bepalen uw resultaten. Bescherm uw oplossing.

Snel van start
On-premise fundamenten met root-of-trust, beveiliging en volledige controle

Connectiviteit stroomlijnen
Veilige fabrics vol functies met schaalbaarheid en geoptimaliseerde verkeersstromen

Provisioning automatiseren
Open-sourcebasis voor de implementatie en het beheer van high performance clusters

Dell GenAI-oplossingen met AMD



Conclusie

Voer een 70B-parametermodel uit op een enkele AMD Instinct™ MI300X-versneller.[4]

Aanpassen

Implementeer acht gelijktijdige 70B-modellen en stel deze af op één Dell PowerEdge XE9680.[4]

Uitbreiden

Neem uw data op in het generatieve proces.

Onderscheid uzelf van de concurrentie met een beproefde, open oplossing die veilige, on-premise AI-applicaties op schaal levert.

Meer informatie

[1] Enterprise Strategy Group, AI-rendement maximaliseren: Inferentie on-premise met Dell Technologies kan 75% rendabeler zijn dan public clouds, april 2024.
[2] Schatting gebaseerd op een analyse van Dell in mei 2024, waarbij de tijd voor het opzetten van een Kubernetes-cluster met 2 knooppunten voor een LLM voor algemene doeleinden met behulp van geautomatiseerde scripts werd vergeleken met het handmatig implementeren van een veelgebruikt ontwerp. De installatietijd omvat alleen de basisinstallatie. De werkelijke installatietijd is afhankelijk van de configuratie van de oplossing.
[3] Dell Technologies, Generative AI Pulse Survey, augustus en september 2023.
[4] Dell Technologies blog, Silicon Diversity: GenAI implementeren op de PowerEdge XE9680 met AMD Instinct MI300X versnellers, mei 2024.