Dells AI-fabrik

Dells generativa AI-lösning med AMD

Snabba upp innovationen, minska kostnaderna och skydda data med en skalbar och modulär arkitektur för komplex GenAI.

75 % mer kostnadseffektiv inferens jämfört med IaaS i offentligt moln[1]

86 % kortare tid till värde[2]

Viktiga användningsområden kräver kraft, flexibilitet och skalning

- Assistenter, chatbots och innehållsproduktion
- Accelerator som tjänst
- Multimodal RAG (Retrieval Augmented Generation)

Förenklad

Effektivisera GenAI-driftsättningar med beprövade och validerade lösningar, uppbackade av mer än 340 000 tekniktimmar.

Optimera prestandan
Högpresterande accelerator, öppen arkitektur och AI-optimerade strukturer

AI var som helst
På data överallt med flexibilitet för flermolnsbaserad lagring

Nyckelfärdig multi-nod
Beprövade fullständiga AI-grunder som levererar snabbare resultat

Anpassad

AMD ROCm™ -mjukvara med öppen källkod och öppna ekosystem ökar AI-utvecklingen och AI-driften.

Öka innovationstakten
Använd mjukvara och ekosystem med öppen källkod för att utveckla unika program.

Accelerera utvecklingen
Utnyttja ramverk av branschstandard med flexibla teknikstackar.

Aktivera dina data
Kör effektivt flera AI-användningsfall samtidigt.

Betrodd

82 % av IT-beslutsfattare föredrar en lokal modell eller en hybridmodell.[3]
Dina data avgör dina resultat. Skydda dem.

Starta snabbt
Lokala stiftelser med säkerhet som bygger på förtroenderot och fullständig kontroll

Effektivisera anslutningsmöjligheterna
Säkra, funktionsrika strukturer med skalbarhet och optimerade trafikflöden

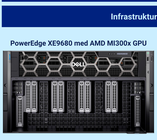
Automatisera tillhandahållandet av tjänster
Lokala stiftelser för distribution och hantering av högpresterande kluster

Dell GenAI-lösningar med AMD

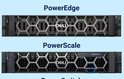
Dell Omnia

- Standardbaserat AI/ML-ramverk
- AMD ROCm – AI-ramverk för AMD-accelerator
- Infrastrukturhantering

PowerEdge XE9680 med AMD MI300x GPU



PowerEdge



PowerScale

PowerSwitch

- Dell Professional Services

Inferens

Kör en 70 MD-parametermodell på en enda AMD Instinct™ MI300X-accelerator.[4]

Anpassa

Installera och finjustera åtta samtidiga 70 MD-modeller på en enda Dell PowerEdge XE9680.[4]

Förstärk

Införliva dina data i den generativa processen.

Uppnå konkurrenskraftig differentiering med en beprövad, öppen lösning som levererar säkra AI-program på plats i stor skala.

Mer information

[1] Enterprise Strategy Group, maximera AI-avkastning på investering: Lokal inferens med Dell Technologies kan vara 75 % mer kostnadseffektivt än offentliga moln, April 2024.
[2] Uppskattning baserad på Dells analys i maj 2024 där man jämför tiden för att konfigurera ett Kubernetes-kluster med två noder för en allmän språkmodell med hjälp av automatiserade skript jämfört med att driftsätta en vanlig design manuellt. Installationstiden innefattar endast en basinstallation. Den faktiska installationstiden varierar beroende på lösningens konfiguration.
[3] Dell Technologies Generative AI Pulse Survey, augusti och september 2023.
[4] Dell Technologies-blogget Silicon Diversity: Driftsätt GenAI på PowerEdge XE9680 med AMD Instinct MI300X-accelerator, maj 2024.

Upphovsrätt © 2024 Dell Inc. eller dess dotterbolag. Med ensamrätt. Dell och andra varumärken är varumärken som tillhör Dell Inc. eller dess dotterbolag. AMD, AMD Arrow-logotypen, ROCm™, Instinct™ och kombinationer av dessa är varumärken som tillhör Advanced Micro Devices, Inc. PyTorch® är ett varumärke eller registrerat varumärke som tillhör PyTorch eller PyTorcha licensgivare. TensorFlow™ och alla relaterade märken är varumärken som tillhör Google Inc. Andra varumärken kan tillhöra sina respektive ägare. Publicerat i USA 10/24 infografik

Dell Technologies anser att informationen i detta dokument är korrekt på publiceringsdagen. Informationen kan komma att ändras utan föregående meddelande.