Dell AI Factory

Soluzione di AI generativa Dell con AMD

Accelera l'innovazione, riduci i costi e proteggi i dati con un'architettura scalabile e modulare per l'AI generativa complessa.

75% di convenienza in più rispetto all'IaaS su public cloud[1]

86% di riduzione del time-to-value[2]

I casi d'uso chiave richiedono potenza, flessibilità e scalabilità

Assistenti, chatbot e creazione di contenuti

Acceleratore as-a-Service

RAG (Retrieval-Augmented Generation) multimodale

Semplificato

Semplifica l'implementazione dell'AI generativa con soluzioni convalidate comprovate, supportate da oltre 340.000 ore di progettazione.

Ottimizzazione delle prestazioni
Acceleratore a prestazioni elevate, architettura aperta e fabric ottimizzati per l'AI

AI ovunque
Dati ovunque con la flessibilità dello storage multicloud

Multi-nodo pronto all'uso
Basi comprovate di AI full-stack per risultati più rapidi

Su misura

Il software open source AMD ROCm™ e gli ecosistemi aperti potenziano lo sviluppo e le operazioni dell'AI.

Innovate più velocemente
Usa software ed ecosistemi open source per sviluppare applicazioni esclusive.

Accelera lo sviluppo
Sfrutta i framework standard del settore con stack tecnologici flessibili.

Attiva i tuoi dati
Esegui in modo efficiente più casi d'uso di AI contemporaneamente.

Affidabile

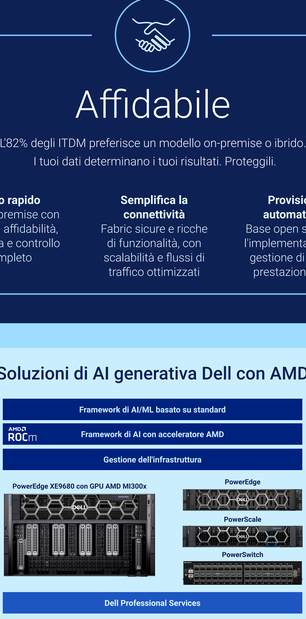
L'82% degli ITDM preferisce un modello on-premise o ibrido.[3]
I tuoi dati determinano i tuoi risultati. Proteggili.

Avvio rapido
Basi on-premise con radice di affidabilità, sicurezza e controllo completo

Semplifica la connettività
Fabric sicure e ricche di funzionalità, con scalabilità e flussi di traffico ottimizzati

Provisioning automatizzato
Base open source per l'implementazione e la gestione di cluster a prestazioni elevate

Soluzioni di AI generativa Dell con AMD



Inferenza

Esegui un modello di parametri 70B su un singolo acceleratore AMD Instinct™ MI300X.[4]

Personalizza

Implementa e ottimizza otto modelli 70B simultanei su un singolo Dell PowerEdge XE9680.[4]

Aumento

Incorpora i dati nel processo generativo.

Ottieni una differenziazione competitiva con una soluzione aperta e comprovata che offre applicazioni AI on-premise sicure su larga scala.

Ulteriori informazioni

[1] Enterprise Strategy Group, Maximizing AI ROI: Inferencing On-premises With Dell Technologies Can Be 75% More Cost-effective Than Public Cloud, aprile 2024.
[2] Stima basata su un'analisi condotta da Dell a maggio 2024, in cui è stato confrontato il tempo necessario per configurare un cluster Kubernetes a 2 nodi per un LLM generico utilizzando script automatizzati e l'implementazione manuale di una progettazione comune. Il tempo di configurazione include solo l'installazione di base. Il tempo di configurazione effettivo varia a seconda della configurazione della soluzione.
[3] Dell Technologies, Generative AI Pulse Survey, agosto e settembre 2023.
[4] Blog di Dell Technologies, Silicon Diversity: Deploy GenAI on the PowerEdge XE9680 with AMD Instinct MI300X Accelerators, maggio 2024.