Dell AI Factory

Solution d'IA générative Dell avec AMD

Accélérez l'innovation, réduisez les coûts et protégez les données avec une architecture évolutive et modulaire pour l'IA générative complexe.

75 %
d'inférence plus rentable par rapport à l'IaaS de type Cloud public[1]

86 %
de réduction du délai de rentabilisation[2]

Les principaux cas d'utilisation exigent puissance, flexibilité et évolutivité

Assistants, chatbots et création de contenu

Accelerator as-a-service

Retrieval Augmented Generation (RAG) multimodale

Simplifiée

Rationalisez les déploiements d'IA générative avec des solutions éprouvées et validées, étayées par plus de 340 000 heures d'ingénierie.

Optimisation des performances
Accélérateur hautes performances, architecture ouverte et fabrics optimisés par l'IA

L'IA dans tous ses états
Des données partout grâce à la flexibilité du stockage multicloud

Plusieurs nœuds clé en main
Les bases éprouvées de l'IA pour des résultats plus rapides

Adaptée

Le logiciel Open Source AMD ROCm™ et les écosystèmes ouverts stimulent le développement et les opérations d'IA.

Innovez plus vite
Utilisez des logiciels et des écosystèmes Open Source pour développer des applications uniques.

Accélérez le développement
Tirez parti des cadres standard du secteur avec des piles de technologies flexibles.

Activez vos données
Exécutez efficacement et simultanément plusieurs cas d'utilisation de l'IA.

Fiable

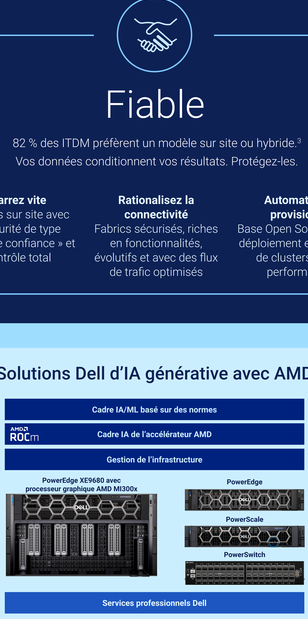
82 % des ITDM préfèrent un modèle sur site ou hybride.[3]
Vos données conditionnent vos résultats. Protégez-les.

Démarrez vite
Des bases sur site avec une sécurité de type « Racine de confiance » et un contrôle total

Rationalisez la connectivité
Fabrics sécurisés, riches en fonctionnalités, évolutifs et avec des flux de trafic optimisés

Automatiser le provisioning
Base Open Source pour le déploiement et la gestion de clusters hautes performances

Solutions Dell d'IA générative avec AMD



Inférer

Exécutez un modèle de 70 milliards de paramètres sur un seul accélérateur AMD Instinct™ MI300X[4].

Personnaliser

Déployez et affinez huit modèles simultanés de 70 milliards sur un seul serveur Dell PowerEdge XE9680[4].

Augmenter

Intégrez vos données dans le processus génératif.

Démarquez-vous de la concurrence grâce à une solution ouverte et éprouvée qui fournit des applications d'IA sécurisées sur site à grande échelle.

En savoir plus

[1] Enterprise Strategy Group, « Maximizing AI ROI: Inferencing On-premises With Dell Technologies Can Be 75% More Cost-effective Than Public Cloud », avril 2024.
[2] Estimation basée sur une analyse réalisée par Dell en mai 2024 comparant le temps de configuration d'un cluster Kubernetes à 2 nœuds pour un LLM général utilisant des scripts automatisés au délai de déploiement manuel d'une conception commune. Le temps de configuration inclut uniquement l'installation de la base. Le temps de configuration réel varie en fonction de la configuration de la solution.
[3] Dell Technologies, « Generative AI Pulse Survey », août et septembre 2023.
[4] Blog Dell Technologies, « Silicon Diversity: Deploy GenAI on the PowerEdge XE9680 with AMD Instinct MI300X Accelerators, mai 2024.