DELL Technologies

AMD

Dell AI Factory

AMD 기반의 Dell Generative AI 솔루션

복잡한 GenAI를 위한 확장 가능한 모듈식 아키텍처로 혁신을 가속하고 비용을 절감하며 데이터를 보호하십시오.

75%
퍼블릭 클라우드 IaaS 대비 더 높은 비용 효율적 추론[1]

86%
가치 실현 시간 단축 비율[2]

강력한 성능, 유연성 및 확장성을 요구하는 주요 활용 사례에 대응

어시스턴트, 챗봇 및 콘텐츠 제작

as-a-Service로 제공되는 가속기

멀티모달 RAG(Retrieval Augmented Generation)

간소화

340,000시간 이상의 엔지니어링 시간을 지원하는 검증된 솔루션으로 GenAI 구축을 간소화하십시오.

성능 최적화
고성능 가속기, 개방형 아키텍처 및 AI 최적화 패브릭

어디서나 지원되는 AI
멀티클라우드 스토리지의 유연성 덕분에 어디서나 데이터 전송 가능

턴키 멀티 노드
더욱 빠른 결과를 제공하는 검증된 풀 스택 AI 기반

맞춤형

AMD ROCm™ 오픈 소스 소프트웨어와 개방형 생태계가 AI 개발과 운영을 촉진합니다.

더 빠르게 혁신 추진
오픈 소스 소프트웨어와 생태계를 사용하여 고유한 애플리케이션을 개발합니다.

개발 속도 향상
유연한 기술 스택으로 업계 표준 프레임워크를 활용합니다.

데이터 활성화
여러 AI 활용 사례를 동시에 효율적으로 실행합니다.

신뢰성

82%의 ITDM이 온프레미스 또는 하이브리드 모델을 선호합니다.[3] 결과를 좌우하는 데이터를 보호하십시오.

빠른 시작
RoT(Root of Trust) 보안과 완벽한 제어 기능을 갖춘 온프레미스 기반

연결 간소화
확장성과 최적화된 트래픽 흐름을 갖춘 안전하고 풍부한 기능의 패브릭

프로비저닝 자동화
고성능 클러스터 배포 및 관리를 위한 오픈 소스 기반

AMD 기반의 Dell GenAI 솔루션

Dell Omnia

표준 기반 AI/ML 프레임워크

AMD ROCm AMD 가속기 AI 프레임워크

인프라스트럭처 관리

AMD MI300x GPU가 장착된 PowerEdge XE9680

PowerEdge

PowerScale

PowerSwitch

Dell Professional Services

추론
단일 AMD Instinct™ MI300X 가속기에서 70B 매개변수 모델을 실행합니다.[4]

맞춤 구성
단일 Dell PowerEdge XE9680에서 8개의 동시 70B 모델을 배포하고 미세 조정합니다.[4]

보강
데이터를 생성 프로세스에 통합합니다.

안전한 온프레미스 AI 애플리케이션을 규모에 따라 제공하는 검증된 개방형 솔루션으로 차별화된 경쟁력을 확보하십시오.

자세한 정보

[1] Enterprise Strategy Group, Maximizing AI ROI: Inferencing On-premises With Dell Technologies Can Be 75% More Cost-effective Than Public Cloud, 2024년 4월.
[2] 범용 LLM(Large Language Model)을 지원하는 2노드 Kubernetes 클러스터의 설치 소요 시간을 자동화된 스크립트를 사용하는 경우와 공용 설계를 수동으로 배포하는 경우에 대해 비교한 2024년 5월 Dell 분석 결과를 기반으로 추정했습니다. 설치 시간은 기본 설치만 포함합니다. 실제 설치 시간은 솔루션 구성에 따라 다릅니다.
[3] Dell Technologies, Generative AI Pulse Survey, 2023년 8월 및 9월.
[4] Dell Technologies 블로그 Silicon Diversity: Deploy GenAI on the PowerEdge XE9680 with AMD Instinct MI300X Accelerators, 2024년 5월

Dell Technologies는 발행일을 기준으로 본 문서의 정보가 정확한 것으로 간주합니다. 이 정보는 예고 없이 변경될 수 있습니다.