GenAI로 높은 가치의 통찰력 활용

GenAI(Generative Artificial Intelligence) 대규모 언어 모델 추론을 위한 전체 스택 솔루션을 신속하게 구축

생산성 및 통찰력 향상

이 공동 아키텍처는 여러 활용 사례와 컴퓨팅 요구 사항을 지원하는 유연한 모듈형 설계를 제공합니다. 애플리케이션 요구 사항에 따라 구성 요소를 다양하게 조합할 수도 있고 개별적으로 확장할 수도 있습니다.

지원되는 추론 활용 사례의 몇 가지 주된 예는 다음과 같습니다.

자연어 생성: 문서 작성, 대화 생성, 내용 요약, 콘텐츠 제작과 같은 텍스트 생성 작업에 생성 모델 사용

챗봇 및 가상 어시스턴트: 사용자의 질문이나 지시에 따라 자연어 응답을 생성하는 GenAI 기반의 대화형 상담원, 챗봇, 가상 어시스턴트 구현

코드 개발: 코드 완성, 단위 테스트 생성 기능, 코드 설명용 채팅 기능과 같은 여러 가지 기능으로 소프트웨어 개발 지원

Dell Technologies와 NVIDIA의 강력한 GenAI 솔루션으로 예측과 결과물의 품질을 높이고 가치 실현 시간을 단축하는 동시에 의사 결정 속도를 높일 수 있습니다. 공동 엔지니어링된 이 솔루션은 레이턴시, 응답성, 컴퓨팅 요구 사항 등의 추론 문제를 해결하여 기업 데이터를 더욱 스마트한 고가치 결과물로 전환하도록 지원합니다.

조직은 혁신 기술, 포괄적 전문 서비스, 광범위한 파트너 생태계를 통해 전사적인 차원에서 GenAI를 가속할 수 있습니다. 이제 IT 조직, 데이터 과학자와 AI DevOps는 GenAI 및 LLM 추론을 위해 확장 가능한 모듈형 플랫폼을 쉽게 제공할 수 있습니다.

검증된 솔루션으로 결과 실현 시간 단축

도입 간소화를 지원하는 것으로 검증된 설계와 레퍼런스 아키텍처로 애플리케이션 요구 사항에 맞는 온프레미스 인프라스트럭처를 신속하게 구축해 보십시오. 도입 과정에 수반되는 모든 단계의 복잡성을 줄임으로써 이제 더 많은 통찰력을 확보하고 더 빠르게 결정을 내림과 동시에 생산성을 크게 높일 수 있습니다.

자세한 정보:

- 설계 가이드 참조
- AI InfoHub
- delltechnologies.com/ai
- Dell Technologies와 NVIDIA

추론이란?

AI에서 추론이란 학습된 모델을 사용하여 입력 데이터를 기반으로 예측을 생성하고 의사 결정을 내리거나 결과물을 생성하는 프로세스를 의미합니다. 이는 모델의 훈련 단계에서 학습한 지식과 습득한 패턴을 처음 보는 새로운 데이터에 응용하는 작업을 말합니다.

학습된 모델은 추론 시 입력 데이터를 가져온 후 컴퓨팅 알고리듬이나 중립 네트워크 아키텍처를 통해 처리하여 결과나 예측을 생성합니다. 이 모델은 학습된 매개변수, 가중치 또는 규칙을 적용하여 입력 데이터를 의미 있는 정보나 작업으로 바꿉니다.

추론은 AI 시스템의 수명주기에서 매우 중요한 단계입니다. 레이블이 지정되거나 미지정된 데이터로 패턴과 상관관계를 학습하도록 모델을 훈련시키면, 해당 모델은 추론을 통해 정보를 일반화하고 실제 데이터 또는 처음 보는 데이터에 대해 예측하거나 응답을 생성할 수 있습니다.

Dell의 지원을 활용하여 더 빠르게 성과 실현

Dell Services 전문가는 GenAI 여정의 모든 단계를 지원하는 서비스 포트폴리오를 통해 데이터와 관련한 GenAI의 가치를 더욱 신속하게 실현하도록 도와줍니다.

- **전략 수립** - IT 및 비즈니스 이해 관계자의 혁신 목표를 달성하기 위한 로드맵 구축
- **구현** - Dell Validated Designs를 활용하여 GenAI 추론 하드웨어 및 소프트웨어를 구현하는 플랫폼 구축
- **채택** - 사전 학습된 추론 모델을 구현하여 GenAI 활용 사례의 가치 실현 시간 단축
- **확장** - 상주 기술 전문가 및 교육 오퍼링으로 GenAI 혁신 포트폴리오를 관리하여 팀의 능력 증진

기술 사양

Validated Design 구성은 AI 가속에 최적화된 최신 Dell PowerEdge XE 및 랙 서버에 기반하며, 최신 NVIDIA GPU 및 NVIDIA AI Enterprise를 활용하고, Triton Inference Server 및 NeMo 프레임워크를 사용합니다. Generative AI와 대규모 언어 모델을 위한 빠르고 풍부한 데이터 레이크 스토리지는 Dell PowerScale 올플래시 또는 하이브리드 스토리지 어레이를 통해 제공됩니다.

컴퓨팅	가속기	네트워킹	소프트웨어	스토리지
Dell PowerEdge R760xa 서버	NVIDIA A100 또는 H100 GPU	NVIDIA Networking, Dell PowerSwitch S5232F-ON 또는 S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ. LLM용 Nemo Framework 및 Triton Inference Server 기반의 NVIDIA AI Enterprise, NVIDIA Base Command Manager Essentials	Dell PowerScale, ECS 및 ObjectScale 기반

Dell Technologies와 NVIDIA

Dell Technologies와 NVIDIA는 협력을 통해 Generative AI 워크로드를 지원하고 가속하며 엔지니어링 검증을 거친 하드웨어와 소프트웨어를 제공하여 AI, ML, DL 워크로드를 가속함으로써 모든 비즈니스와 업종에서 고객의 요구 사항에 부응하고 있습니다. LLM 추론을 위한 이 Validated Design를 활용하면 AI 이니셔티브의 가치 실현 시간을 크게 단축하는 최적화된 솔루션을 통해 규모에 맞춰 주요 의사 결정을 향상시키는 실시간 데이터로 디지털 혁신을 가속할 수 있습니다.

Dell 솔루션에 대한 자세한 정보

Dell Technologies 전문가에게 문의

추가 리소스 보기

대화 참여: #HashTag

DELL Technologies

NVIDIA