생성적 AI와 NetApp 가치

NetApp artificial intelligence solutions

NetApp
December 04, 2025

목차

생성적 **AI**와 **NetApp** 가치

생성적 인공지능(AI)에 대한 수요는 산업 전반에 걸쳐 혁신을 촉진하고, 비즈니스 창의성과 제품 혁신을 강화하고 있습니다.

개요

많은 조직에서 생성적 AI를 사용하여 새로운 제품 기능을 구축하고, 엔지니어링 생산성을 개선하고, 더 나은 결과와 고객 경험을 제공하는 AI 기반 애플리케이션의 프로토타입을 개발하고 있습니다. GPT(Generative Pre-trained Transformers)와 같은 생성적 AI는 신경망을 사용하여 텍스트, 오디오, 비디오 등 다양한 새로운 콘텐츠를 생성합니다. 대규모 언어 모델(LLM)과 관련된 극단적인 규모와 엄청난 양의 데이터 세트를 고려할 때, 기업이 AI 솔루션을 설계하기 전에 온프레미스, 하이브리드 및 멀티클라우드 배포 옵션의 매력적인 데이터 저장 기능을 활용하고 데이터 이동성, 데이터 보호 및 거버넌스와 관련된 위험을 줄이는 강력한 AI 인프라를 구축하는 것이 중요합니다. 이 논문에서는 이러한 고려 사항과 학습, 재학습, 미세 조정 및 생성 AI 모델 추론을 위한 AI 데이터 파이프라인 전반의 원활한 데이터 관리 및 데이터 이동을 가능하게 하는 해당 NetApp AI 기능에 대해 설명합니다.

요약

가장 최근에는 2022년 11월 GPT-3의 분사 기업인 ChatGPT가 출시된 이후, 사용자 프롬프트에 따라 텍스트, 코드, 이미지, 심지어 치료용 단백질까지 생성하는 데 사용되는 새로운 AI 도구가 상당한 명성을 얻었습니다. 이는 사용자가 자연어를 사용하여 요청을 하면 AI가 사용자 요청을 반영하는 뉴스 기사나 제품 설명과 같은 텍스트를 해석하고 생성하거나, 이미 존재하는 데이터로 훈련된 알고리즘을 사용하여 코드, 음악, 음성, 시각 효과, 3D 자산을 생성한다는 것을 의미합니다. 그 결과, 안정적 확산, 환각, 신속한 엔지니어링, 가치 정렬과 같은 문구가 AI 시스템 설계에 빠르게 등장하고 있습니다. 이러한 자기 지도 또는 반지도 머신 러닝(ML) 모델은 클라우드 서비스 공급업체와 기타 AI 회사 공급업체를 통해 사전 학습된 기반 모델(FM)로 널리 제공되고 있으며, 다양한 산업 전반의 기업에서 광범위한 다운스트림 NLP(자연어 처리) 작업에 채택되고 있습니다. 맥킨지와 같은 연구 분석 회사가 주장한 것처럼 "생성적 AI가 생산성에 미치는 영향은 글로벌 경제에 수조 달러의 가치를 더할 수 있습니다." 기업들이 AI를 인간의 사고 파트너로 재구성하고, FM이 기업과 기관이 생성적 AI를 통해 할 수 있는 일을 확대하는 동시에, 엄청난 양의 데이터를 관리할 수 있는 기회도 계속해서 늘어날 것입니다. 이 문서에서는 온프레미스와 하이브리드 또는 멀티클라우드 환경 모두에서 NetApp 고객에게 가치를 제공하는 NetApp 기능과 관련된 생성적 AI 및 설계 개념에 대한 소개 정보를 제공합니다.

그렇다면 고객이 **AI** 환경에서 **NetApp** 사용하면 어떤 이점이 있을까요**?**

NetApp 조직이 급속한 데이터 및 클라우드 성장, 멀티 클라우드 관리, AI와 같은 차세대 기술 도입으로 인해 발생하는 복잡성을 해결할 수 있도록 지원합니다. NetApp 다양한 기능을 지능형 데이터 관리 소프트웨어와 스토리지 인프라로 결합했으며, 이는 AI 워크로드에 최적화된 고성능과 균형을 이루고 있습니다. LLM과 같은 생성적 AI 솔루션은 지능을 강화하기 위해 저장소에서 소스 데이터 세트를 여러 번 읽어 메모리로 처리해야 합니다.

NetApp 엣지에서 코어, 클라우드에 이르는 생태계 전반에 걸쳐 데이터 이동성, 데이터 거버넌스 및 데이터 보안 기술을 선도해 왔으며, 기업 고객에게 대규모 AI 솔루션을 구축할 수 있는 서비스를 제공합니다. 강력한 파트너 네트워크를 갖춘 NetApp 은 최고 데이터 책임자, AI 엔지니어, 엔터프라이즈 아키텍트, 데이터 과학자가 AI 모델 학습 및 추론에 대한 데이터 준비, 데이터 보호, 전략적 데이터 관리 책임을 위한 자유로운 데이터 파이프라인을 설계하고 AI/ML 수명 주기의 성능과 확장성을 최적화하도록 지원해 왔습니다. NetApp ONTAP AI (딥 러닝 데이터 파이프라인), NetApp SnapMirror(스토리지 엔드포인트 간에 데이터를 원활하고 효율적으로 전송), NetApp FlexCache(배치에서 실시간으로 데이터 흐름이 전환되고 데이터 엔지니어링이 신속하게 수행될 때 실시간 렌더링)와 같은 NetApp 데이터 기술과 기능은 실시간 생성 AI 모델 배포에 가치를 더해줍니다. 모든 유형의 기업이 새로운 AI 도구를 도입함에 따라 엣지에서 데이터 센터, 클라우드에 이르기까지 확장 가능하고 책임감 있으며 설명 가능한 AI 솔루션을 요구하는 데이터 과제에 직면하게 됩니다.

하이브리드 및 멀티 클라우드 분야의 데이터 기관으로서 NetApp 생성적 AI 모델 학습(사전 학습), 미세 조정, 컨텍스트

기반 추론 및 LLM의 모델 붕괴 모니터링을 위한 데이터 파이프라인 및 데이터 레이크 구축의 모든 측면을 지원할 수 있는 파트너 및 공동 솔루션 네트워크를 구축하는 데 전념하고 있습니다.

생성적 **AI**란 무엇인가**?**

생성적 AI는 콘텐츠를 만드는 방식, 새로운 디자인 컨셉을 창출하는 방식, 참신한 구성을 탐구하는 방식을 바꾸고 있습니다. 여기에는 텍스트, 코드, 이미지, 오디오, 비디오 및 합성 데이터와 같은 새로운 콘텐츠를 생성할 수 있는 GAN(Generative Adversarial Network), VAE(Variational Autoencoders), GPT(Generative Pre-Trained Transformers)와 같은 신경망 프레임워크가 설명되어 있습니다. OpenAI의 Chat-GPT, Google의 Bard, Hugging Face의 BLOOM, Meta의 LLaMA와 같은 트랜스포머 기반 모델은 대규모 언어 모델의 많은 발전을 뒷받침하는 기반 기술로 떠올랐습니다. 마찬가지로 OpenAI의 Dall-E, Meta의 CM3leon, Google의 Imagen은 텍스트-이미지 확산 모델의 예입니다. 이러한 모델은 고객에게 데이터 세트 증강 및 텍스트-이미지 합성을 사용하여 텍스트와 시각적 의미를 연결하여 고품질의 상황 인식 이미지를 생성하거나 기존 이미지를 편집하여 새롭고 복잡한 이미지를 처음부터 만들 수 있는 전례 없는 수준의 포토리얼리즘을 제공합니다. 디지털 아티스트는 NeRF(Neural Radiance Field)와 같은 렌더링 기술과 생성 AI를 결합하여 정적인 2D 이미지를 몰입형 3D 장면으로 변환하기 시작했습니다. 일반적으로 LLM은 크게 4가지 매개변수로 특징지어집니다. (1) 모델 크기(일반적으로 수십억 개의 매개변수), (2) 학습 데이터 세트 크기, (3) 학습 비용, (4) 학습 후 모델 성능. LLM도 주로 3가지 변압기 아키텍처로 분류됩니다. (i) 인코더 전용 모델. 예: BERT(Google, 2018); (ii) 인코더-디코더 예: BART(Meta, 2020) 및 (iii) 디코더 전용 모델. 예를 들어 LLaMA(Meta, 2023), PaLM-E(Google, 2023). 비즈니스 요구 사항에 따라, 회사가 어떤 아키텍처를 선택하든 일반적으로 학습 데이터 세트의 모델 매개변수 수(N)와 토큰 수(D)에 따라 LLM 학습(사전 학습) 또는 미세 조정의 기준 비용이 결정됩니다.

엔터프라이즈 사용 사례 및 다운스트림 **NLP** 작업

다양한 산업의 기업들은 AI가 기존 데이터에서 비즈니스 운영, 판매, 마케팅, 법률 서비스를 위한 새로운 형태의 가치를 추출하고 생산할 수 있는 잠재력을 점점 더 많이 발견하고 있습니다. IDC(International Data Corporation)의 글로벌 생성적 AI 활용 사례 및 투자에 대한 시장 정보에 따르면, 소프트웨어 개발 및 제품 설계 분야의 지식 관리가 가장 큰 영향을 받을 것으로 예상되며, 그 다음으로 마케팅을 위한 스토리라인 생성과 개발자를 위한 코드 생성이 영향을 받을 것으로 예상됩니다. 의료 분야에서 임상 연구 기관은 의학의 새로운 영역을 개척하고 있습니다. ProteinBERT와 같은 사전 학습된 모델은 유전자 온톨로지(GO) 주석을 통합하여 의약품의 단백질 구조를 빠르게 설계하며, 이는 약물 발견, 생물정보학, 분자생물학 분야에서 중요한 이정표를 나타냅니다. 바이오 기술 회사들은 폐 조직에 돌이킬 수 없는 상처를 남기는 폐 질환인 폐 섬유증(IPF)과 같은 질병을 치료하는 것을 목표로 하는 생성적 AI 기반 의학에 대한 인체 실험을 시작했습니다.

그림 1: 생성적 AI를 구동하는 사용 사례

[그림 1: 생성적 AI를 구동하는 사용 사례]

생성적 AI로 인한 자동화 도입 증가는 많은 직업에 대한 업무 활동의 수요와 공급에도 변화를 가져오고 있습니다. 맥킨지에 따르면 미국 노동 시장(아래 다이어그램)은 빠른 변화를 겪었으며, AI의 영향을 고려하면 이러한 변화는 계속될 가능성이 있습니다.

출처: 맥킨지앤컴퍼니

[그림 2: 출처: McKinsey Company]

생성적 **AI**에서 스토리지의 역할

LLM은 딥러닝, GPU, 컴퓨팅에 크게 의존합니다. 하지만 GPU 버퍼가 가득 차면 데이터를 저장소에 빠르게 기록해야 합니다. 일부 AI 모델은 메모리에서 실행할 만큼 작은 반면, LLM은 대용량 데이터 세트에 빠르게 액세스하려면 높은 IOPS와 높은 처리량의 스토리지가 필요합니다. 특히 수십억 개의 토큰이나 수백만 개의 이미지가 포함된 경우 더욱 그렇습니다. LLM의 일반적인 GPU 메모리 요구 사항의 경우, 10억 개의 매개변수로 모델을 학습하는 데 필요한

메모리는 32비트 전체 정밀도에서 최대 80GB까지 늘어날 수 있습니다. 이 경우, 70억에서 700억 개의 매개변수 규모를 가진 LLM 제품군인 Meta의 LLaMA 2에는 약 70x80, 5600GB 또는 5.6TB의 GPU RAM이 필요할 수 있습니다. 게다가 필요한 메모리 양은 생성하려는 토큰의 최대 수에 직접 비례합니다. 예를 들어 최대 512개 토큰(약 380개 단어)의 출력을 생성하려면 다음이 필요합니다."512MB" . 별로 중요하지 않은 것처럼 보일 수도 있지만, 대량으로 처리하려고 하면 비용이 늘어나기 시작합니다. 따라서 조직이 메모리에서 LLM을 훈련하거나 미세 조정하는 데 드는 비용이 매우 많이 들고, 따라서 저장은 생성적 AI의 초석이 됩니다.

LLM에 대한 세 가지 주요 접근 방식

대부분의 기업의 경우, 현재 추세를 바탕으로 LLM을 배포하는 접근 방식은 3가지 기본 시나리오로 요약될 수 있습니다. 최근에 설명된 바와 같이"하버드 비즈니스 리뷰" 기사: (1) LLM을 처음부터 학습(사전 학습) - 비용이 많이 들고 전문적인 AI/ML 기술이 필요함; (2) 기업 데이터를 사용하여 기초 모델 미세 조정 - 복잡하지만 실행 가능; (3) 검색 증강 생성(RAG)을 사용하여 회사 데이터가 포함된 문서 저장소, API 및 벡터 데이터베이스를 쿼리함. 이들 각각은 구현 과정에서 노력, 반복 속도, 비용 효율성, 모델 정확도 간에 상충 관계가 있으며, 이는 다양한 유형의 문제를 해결하는 데 사용됩니다(아래 다이어그램).

그림 3: 문제 유형

[그림 3: 문제 유형]

기초 모델

기초 모델(FM)은 기본 모델이라고도 하며, 대규모의 레이블이 지정되지 않은 데이터로 학습된 대규모 AI 모델(LLM)로, 대규모 자체 감독을 사용하여 일반적으로 광범위한 다운스트림 NLP 작업에 맞게 조정됩니다. 훈련 데이터에 사람이 레이블을 지정하지 않으므로 모델이 명시적으로 인코딩된 것이 아니라 자연스럽게 나타납니다. 즉, 모델은 명시적으로 프로그래밍되지 않고도 자체적으로 스토리나 내러티브를 생성할 수 있습니다. 따라서 FM의 중요한 특징은 동질화인데, 이는 많은 영역에서 동일한 방법이 사용된다는 것을 의미합니다. 하지만 개인화와 미세 조정 기술을 통해 요즘 출시되는 제품에 통합된 FM은 텍스트 생성, 텍스트-이미지 변환, 텍스트-코드 변환에만 유용한 것이 아니라 도메인별 작업을 설명하거나 코드를 디버깅하는 데도 유용합니다. 예를 들어, OpenAI의 Codex나 Meta의 Code Llama와 같은 FM은 프로그래밍 작업에 대한 자연어 설명을 기반으로 여러 프로그래밍 언어로 코드를 생성할 수 있습니다. 이러한 모델은 Python, C#, JavaScript, Perl, Ruby, SQL을 포함한 12개 이상의 프로그래밍 언어에 능숙합니다. 그들은 사용자의 의도를 이해하고 소프트웨어 개발, 코드 최적화, 프로그래밍 작업 자동화에 유용한 원하는 작업을 수행하는 특정 코드를 생성합니다.

미세 조정, 도메인 특이성 및 재교육

데이터 준비 및 데이터 전처리 후 LLM 배포와 관련된 일반적인 관행 중 하나는 대규모의 다양한 데이터 세트에서 학습된 사전 학습된 모델을 선택하는 것입니다. 미세 조정의 맥락에서 이는 다음과 같은 오픈 소스 대규모 언어 모델이 될 수 있습니다."메타의 라마 2" 700억 개의 매개변수와 2조 개의 토큰을 통해 학습되었습니다. 사전 학습된 모델을 선택하면 다음 단계는 도메인별 데이터에 맞춰 세부적으로 조정하는 것입니다. 여기에는 모델의 매개변수를 조정하고 새로운 데이터로 모델을 훈련하여 특정 도메인과 작업에 적응시키는 작업이 포함됩니다. 예를 들어, BloombergGPT는 금융 산업에 서비스를 제공하는 광범위한 금융 데이터에 대한 교육을 받은 독점 LLM입니다.

특정 작업을 위해 설계되고 훈련된 도메인별 모델은 일반적으로 해당 범위 내에서 정확도와 성능이 더 높지만, 다른 작업이나 도메인으로의 이전 가능성이 낮습니다. 비즈니스 환경과 데이터가 일정 기간 동안 변경되면 FM의 예측 정확도는 테스트 시의 성능에 비해 떨어지기 시작할 수 있습니다. 모델을 재교육하거나 미세 조정하는 것이 중요해지는 시점입니다.

기존 AI/ML에서 모델 재교육은 배포된 ML 모델을 새로운 데이터로 업데이트하는 것을 말하며, 일반적으로 발생하는 두 가지 유형의 드리프트를 제거하기 위해 수행됩니다. (1) 개념 드리프트 - 입력 변수와 대상 변수 간의 연결이 시간이 지남에 따라 변경되면 예측하려는 내용에 대한 설명이 변경되므로 모델이 부정확한 예측을 생성할 수 있습니다. (2) 데이터 드리프트 - 입력 데이터의 특성이 변경될 때 발생합니다. 예를 들어 시간이 지남에 따라 고객 습관이나 행동이 변경되어 모델이 이러한 변경에 대응할 수 없게 되는 경우입니다.

비슷한 방식으로 재교육은 FM/LLM에도 적용되지만 비용이 훨씬 많이 들 수 있으므로(수백만 달러) 대부분의 조직에서는 고려하지 않는 사항입니다. 이는 활발하게 연구 중이며, LLMOps 분야에서는 아직 새로운 분야로 떠오르고 있습니다. 따라서 미세 조정된 FM에서 모델 쇠퇴가 발생하면 기업은 재교육을 하는 대신 새로운 데이터 세트를 사용하여 다시 미세 조정을 (훨씬 저렴하게) 선택할 수 있습니다. 비용 관점에서 볼 때, 아래는 Azure-OpenAI 서비스의 모델 가격표 예입니다. 각 작업 범주에 대해 고객은 특정 데이터 세트에 대한 모델을 미세 조정하고 평가할 수 있습니다.

출처: Microsoft Azure

[출처: Microsoft Azure]

신속한 엔지니어링 및 추론

신속한 엔지니어링은 모델 가중치를 업데이트하지 않고도 원하는 작업을 수행하기 위해 LLM과 통신하는 효과적인 방법을 말합니다. NLP 애플리케이션에 있어 AI 모델 학습과 미세 조정이 중요한 만큼, 학습된 모델이 사용자 프롬프트에 응답하는 추론도 마찬가지로 중요합니다. 추론에 필요한 시스템 요구 사항은 일반적으로 LLM에서 GPU로 데이터를 공급하는 AI 저장 시스템의 읽기 성능에 훨씬 더 중점을 둡니다. 최상의 응답을 생성하기 위해 수십억 개의 저장된 모델 매개변수를 적용할 수 있어야 하기 때문입니다.

LLMOps, 모델 모니터링 및 벡터스토어

기존의 머신 러닝 운영(MLOps)과 마찬가지로 대규모 언어 모델 운영(LLMOps)에도 프로덕션 환경에서 LLM을 관리하기 위한 도구와 모범 사례를 갖춘 데이터 과학자와 DevOps 엔지니어의 협업이 필요합니다. 그러나 LLM의 워크플로와 기술 스택은 여러 면에서 다를 수 있습니다. 예를 들어, LangChain과 같은 프레임워크를 사용하여 구축된 LLM 파이프라인은 벡터스토어나 벡터 데이터베이스와 같은 외부 임베딩 엔드포인트에 대한 여러 LLM API 호출을 연결합니다. 다운스트림 커넥터(예: 벡터 데이터베이스)에 임베딩 엔드포인트와 벡터스토어를 사용하는 것은 데이터가 저장되고 액세스되는 방식에 있어서 중요한 발전을 나타냅니다. 처음부터 개발되는 기존 ML 모델과 달리 LLM은 종종 전이 학습에 의존합니다. 이는 이러한 모델이 새로운 데이터로 미세 조정된 FM으로 시작하여 보다 구체적인 도메인에서 성능을 개선하기 때문입니다. 따라서 LLMOps가 위험 관리 및 모델 붕괴 모니터링 기능을 제공하는 것이 중요합니다.

생성적 **AI** 시대의 위험과 윤리

"ChatGPT – 매끄럽지만 여전히 말도 안 되는 소리를 낸다."– MIT Tech Review. 쓰레기를 넣으면 쓰레기가 나오는 것은 컴퓨팅에서 항상 어려운 문제였습니다. 생성적 AI와의 유일한 차이점은 쓰레기를 매우 신뢰할 만하게 만들어서 부정확한 결과를 낳는 데 탁월하다는 것입니다. LLM은 자신이 만들고 있는 이야기에 맞게 사실을 만들어내는 경향이 있습니다. 따라서 생성적 AI를 AI 대응 제품의 비용을 절감할 수 있는 좋은 기회로 보는 기업은 딥페이크를 효율적으로 감지하고, 편견을 줄이고, 위험을 낮춰 시스템의 정직성과 윤리성을 유지해야 합니다. 종단 간 암호화와 AI 가드레일을 통해 데이터 이동성, 데이터 품질, 데이터 거버넌스 및 데이터 보호를 지원하는 강력한 AI 인프라를 갖춘 자유롭게 흐르는 데이터 파이프라인은 책임감 있고 설명 가능한 생성적 AI 모델을 설계하는 데 매우 중요합니다.

고객 시나리오 및 **NetApp**

그림 3: 머신 러닝/대규모 언어 모델 워크플로

[그림 3: 머신 러닝/대규모 언어 모델 워크플로]

우리는 훈련을 하고 있는가**,** 아니면 미세조정을 하고 있는가**?** (a) LLM 모델을 처음부터 학습시킬지, 미리 학습된 FM을 미세 조정할지, RAG를 사용하여 기초 모델 외부의 문서 저장소에서 데이터를 검색하고 프롬프트를 증강할지, (b) 오픈 소스 LLM(예: Llama 2)이나 독점 FM(예: ChatGPT, Bard, AWS Bedrock)을 활용할지 여부는 조직이 내려야 할 전략적 결정입니다. 각 접근 방식은 비용 효율성, 데이터 중력, 운영, 모델 정확도 및 LLM 관리 간에 상충 관계가 있습니다.

NetApp 회사 내부적으로 업무 문화와 제품 설계 및 엔지니어링 노력에 대한 접근 방식에 AI를 도입했습니다. 예를 들어, NetApp의 자율형 랜섬웨어 보호 기능은 AI와 머신 러닝을 사용하여 구축되었습니다. 이 기능은 파일 시스템 이상을 조기에 감지하여 운영에 영향을 미치기 전에 위협을 식별하는 데 도움이 됩니다. 둘째, NetApp 판매 및 재고 예측과 같은 비즈니스 운영에 예측 AI를 활용하고, 챗봇을 통해 콜센터 제품 지원 서비스, 기술 사양, 보증, 서비스 매뉴얼 등에서 고객을 지원합니다. 셋째, NetApp NetApp ONTAP AI, NetApp SnapMirror, NetApp FlexCache와 같은 NetApp 제품 및 기능을 통해 수요 예측, 의료 영상, 감정 분석, 제조 분야의 산업 이미지 이상 탐지 및 은행 및 금융 서비스의 자금 세탁 방지 및 사기 탐지와 같은 생성 AI 솔루션과 같은 예측 AI 솔루션을 구축하는 고객에게 서비스 NetApp 제품과 솔루션을 통해 AI ONTAP NetApp SnapMirror / NetApp FlexCache 에 고객 가치를 제공합니다.

NetApp 기능

챗봇, 코드 생성, 이미지 생성 또는 게놈 모델 표현과 같은 생성적 AI 애플리케이션에서 데이터의 이동과 관리가 엣지, 프라이빗 데이터 센터 및 하이브리드 멀티클라우드 생태계 전반에 걸쳐 이루어질 수 있습니다. 예를 들어, ChatGPT와 같은 사전 훈련된 모델의 API를 통해 노출된 최종 사용자 앱에서 승객의 항공권을 비즈니스 클래스로 업그레이드하도록 돕는 실시간 AI 봇은 승객 정보가 인터넷에 공개적으로 제공되지 않기 때문에 그 자체로 해당 작업을 달성할 수 없습니다. API는 하이브리드 또는 멀티클라우드 생태계에 존재할 수 있는 항공사의 승객 개인 정보 및 항공권 정보에 대한 액세스가 필요합니다. 유사한 시나리오는 LLM을 사용하여 일대다 생물의학 연구 기관이 참여하는 약물 발견 전반에 걸쳐 임상 시험을 수행하는 최종 사용자 애플리케이션을 통해 약물 분자와 환자 데이터를 공유하는 과학자에게도 적용될 수 있습니다. FM이나 LLM에 전달되는 민감한 데이터에는 PII, 재무 정보, 건강 정보, 생체 인식 데이터, 위치 데이터, 통신 데이터, 온라인 행동 및 법률 정보가 포함될 수 있습니다. 실시간 렌더링, 즉각적인 실행 및 에지 추론이 발생하는 경우 오픈 소스 또는 독점 LLM 모델을 통해 최종 사용자 앱에서 스토리지 엔드포인트로 데이터가 이동하고, 이를 온프레미스 또는 퍼블릭 클라우드 플랫폼의 데이터 센터로 전송합니다. 이러한 모든 시나리오에서 데이터 이동성과 데이터 보호는 대규모 학습 데이터 세트와 이러한 데이터의 이동에 의존하는 LLM과 관련된 AI 작업에 매우 중요합니다.

그림 4: 생성 AI - LLM 데이터 파이프라인

[그림 4: 생성 AI-LLM 데이터 파이프라인]

NetApp의 스토리지 인프라, 데이터 및 클라우드 서비스 포트폴리오는 지능형 데이터 관리 소프트웨어로 구동됩니다.

데이터 준비: LLM 기술 스택의 첫 번째 기둥은 기존의 전통적인 ML 스택과 크게 다르지 않습니다. AI 파이프라인에서 데이터 전처리는 학습이나 미세 조정에 앞서 데이터를 정규화하고 정리하는 데 필요합니다. 이 단계에는 Amazon S3 계층 형태나 파일 저장소 또는 NetApp StorageGRID 와 같은 온프레미스 스토리지 시스템 형태에 있는 데이터를 수집하기 위한 커넥터가 포함됩니다.

- NetApp ONTAP*은 데이터 센터와 클라우드에서 NetApp의 핵심 스토리지 솔루션을 뒷받침하는 기반 기술입니다. ONTAP 에는 사이버 공격에 대한 자동 랜섬웨어 보호, 내장형 데이터 전송 기능, 온프레미스, 하이브리드, 멀티클라우드, NAS, SAN, 객체 및 소프트웨어 정의 스토리지(SDS) 상황의 LLM 배포 등 다양한 아키텍처에 대한 스토리지 효율성 기능을 포함하여 다양한 데이터 관리 및 보호 기능이 포함되어 있습니다.
- 딥 러닝 모델 학습을 위한 NetApp ONTAP AI* NetApp ONTAP ONTAP 스토리지 클러스터와 NVIDIA DGX 컴퓨트 노드를 사용하는 NetApp 고객을 위해 RDMA를 통한 NFS를 사용하여 NVIDIA GPU 직접 스토리지를 지원합니다. 저장소에서 소스 데이터 세트를 여러 번 읽어 메모리로 처리하는 비용 효율적인 성능을 제공하여 지능성을 강화하고, 조직이 LLM에 대한 교육, 미세 조정 및 확장을 수행할 수 있도록 지원합니다.
- NetApp FlexCache*는 파일 배포를 간소화하고 적극적으로 읽은 데이터만 캐싱하는 원격 캐싱 기능입니다. 이는 LLM 교육, 재교육 및 미세 조정에 유용할 수 있으며, 실시간 렌더링 및 LLM 추론과 같은 비즈니스 요구 사항을 가진 고객에게 가치를 제공합니다.
- NetApp SnapMirror *는 두 개의 ONTAP 시스템 간에 볼륨 스냅샷을 복제하는 ONTAP 기능입니다. 이 기능은 엣지에서 온프레미스 데이터 센터나 클라우드로 데이터를 최적으로 전송합니다. 고객이 기업 데이터가 포함된 RAG를 사용하여 클라우드에서 생성적 AI를 개발하려는 경우, SnapMirror 사용하면 온프레미스와 하이퍼스케일러 클라우드 간에 데이터를 안전하고 효율적으로 이동할 수 있습니다. 변경 사항만 효율적으로 전송하여 대역폭을 절약하고 복제 속도를 높여 FM 또는 LLM의 교육, 재교육 및 미세 조정 작업 중에 필수적인 데이터 이동성 기능을

제공합니다.

- NetApp SnapLock*은 ONTAP 기반 스토리지 시스템에 변경 불가능한 디스크 기능을 제공하여 데이터 세트 버전을 관리합니다. 마이크로코어 아키텍처는 FPolicy Zero Trust 엔진을 통해 고객 데이터를 보호하도록 설계되었습니다. NetApp 공격자가 특히 리소스를 많이 소모하는 방식으로 LLM과 상호 작용할 때 서비스 거부(DoS) 공격을 저항하여 고객 데이터의 가용성을 보장합니다.
- NetApp Cloud Data Sense*는 기업 데이터 세트에 존재하는 개인 정보를 식별, 매핑 및 분류하고, 정책을 시행하고, 온프레미스 또는 클라우드에서 개인 정보 보호 요구 사항을 충족하고, 보안 태세를 개선하고 규정을 준수하는 데 도움이 됩니다.
- NetApp BlueXP* 분류, Cloud Data Sense 기반. 고객은 데이터 자산 전반에 걸쳐 데이터를 자동으로 스캔, 분석, 분류하고 조치를 취할 수 있으며, 보안 위험을 탐지하고, 스토리지를 최적화하고, 클라우드 배포를 가속화할 수 있습니다. 통합 제어 평면을 통해 스토리지와 데이터 서비스를 결합합니다. 고객은 컴퓨팅에 GPU 인스턴스를 사용하고, 콜드 스토리지 계층화와 보관 및 백업에 하이브리드 멀티클라우드 환경을 사용할 수 있습니다.
- NetApp 파일-객체 이중성*. NetApp ONTAP NFS 및 S3에 대한 이중 프로토콜 액세스를 지원합니다. 이 솔루션을 사용하면 고객은 NetApp Cloud Volumes ONTAP 의 S3 버킷을 통해 Amazon AWS SageMaker 노트북에서 NFS 데이터에 액세스할 수 있습니다. 이 기능은 NFS와 S3 모두에서 데이터를 공유할 수 있는 기능을 통해 이기종 데이터 소스에 쉽게 액세스해야 하는 고객에게 유연성을 제공합니다. 예를 들어, SageMaker에서 파일 객체 버킷에 대한 액세스를 통해 Meta의 Llama 2 텍스트 생성 모델과 같은 FM을 미세 조정합니다.
- NetApp Cloud Sync* 서비스는 클라우드나 온프레미스의 모든 대상으로 데이터를 마이그레이션하는 간단하고 안전한 방법을 제공합니다. Cloud Sync 온프레미스 또는 클라우드 스토리지, NAS, 개체 저장소 간에 데이터를 원활하게 전송하고 동기화합니다.
- NetApp XCP*는 빠르고 안정적인 any-to- NetApp 및 NetApp-to- NetApp 데이터 마이그레이션을 가능하게 하는 클라이언트 소프트웨어입니다. XCP는 Hadoop HDFS 파일 시스템에서 ONTAP NFS, S3 또는 StorageGRID 로 대량 데이터를 효율적으로 이동하는 기능을 제공하며, XCP 파일 분석은 파일 시스템에 대한 가시성을 제공합니다.
- NetApp DataOps Toolkit*은 데이터 과학자, DevOps 및 데이터 엔지니어가 다양한 데이터 관리 작업을 간편하게 수행할 수 있도록 해주는 Python 라이브러리로, 고성능 확장형 NetApp 스토리지로 지원되는 데이터 볼륨이나 JupyterLab 작업 공간을 거의 즉각적으로 프로비저닝, 복제 또는 스냅샷하는 등의 작업이 가능합니다.

NetApp의 제품 보안. LLM은 실수로 답변에서 기밀 데이터를 공개할 수 있으므로 LLM을 활용하는 AI 애플리케이션과 관련된 취약성을 연구하는 CISO에게는 우려스러운 일입니다. OWASP(Open Worldwide Application Security Project)에서 설명한 대로, 데이터 오염, 데이터 유출, 서비스 거부, LLM 내의 즉각적인 주입과 같은 보안 문제는 데이터 노출부터 공격자의 무단 접근까지 기업에 영향을 미칠 수 있습니다. 데이터 저장 요구 사항에는 구조화된 데이터, 반구조화된 데이터, 구조화되지 않은 데이터에 대한 무결성 검사와 변경 불가능한 스냅샷이 포함되어야 합니다. NetApp Snapshots와 SnapLock 데이터 세트 버전 관리에 사용됩니다. 엄격한 역할 기반 액세스 제어(RBAC)와 보안 프로토콜, 업계 표준 암호화를 제공하여 저장 중인 데이터와 전송 중인 데이터를 모두 보호합니다. Cloud Insights 와 Cloud Data Sense는 함께 위협의 근원을 법의학적으로 식별하고 복원할 데이터의 우선순위를 지정하는 데 도움이 되는 기능을 제공합니다.

* DGX BasePOD를 탑재한 ONTAP AI*

NVIDIA DGX BasePOD 탑재한 NetApp ONTAP AI 참조 아키텍처는 머신 러닝(ML) 및 인공 지능(AI) 워크로드를 위한 확장 가능한 아키텍처입니다. LLM의 중요한 교육 단계에서는 일반적으로 데이터가 정기적으로 데이터 저장소에서 교육 클러스터로 복사됩니다. 이 단계에서 사용되는 서버는 GPU를 사용하여 계산을 병렬화하여 엄청난 데이터 수요를 생성합니다. GPU 활용도를 높게 유지하려면 원시 I/O 대역폭 요구 사항을 충족하는 것이 중요합니다.

* NVIDIA AI Enterprise를 탑재한 ONTAP AI*

NVIDIA AI Enterprise는 NVIDIA 인증 시스템을 통해 VMware vSphere에서 실행되도록 NVIDIA 에서 최적화, 인증 및 지원하는 엔드 투 엔드 클라우드 기반 AI 및 데이터 분석 소프트웨어 제품군입니다. 이 소프트웨어는 최신 하이브리드 클라우드 환경에서 AI 워크로드를 간단하고 빠르게 배포, 관리, 확장할 수 있도록 지원합니다. NetApp 과 VMware

기반의 NVIDIA AI Enterprise는 간소화되고 친숙한 패키지로 엔터프라이즈급 AI 워크로드와 데이터 관리를 제공합니다.

1P 클라우드 플랫폼

완전 관리형 클라우드 스토리지 제품은 Microsoft Azure에서 Azure NetApp Files (ANF), AWS에서 Amazon FSx for NetApp ONTAP (FSx ONTAP) 및 Google에서 Google Cloud NetApp Volumes (GNCV)로 기본적으로 제공됩니다. 1P는 고객이 퍼블릭 클라우드에서 향상된 데이터 보안으로 고가용성 AI 워크로드를 실행하고 AWS SageMaker, Azure-OpenAI 서비스, Google의 Vertex AI와 같은 클라우드 기반 ML 플랫폼을 사용하여 LLM/FM을 미세 조정할 수 있도록 하는 관리형 고성능 파일 시스템입니다.

NetApp 파트너 솔루션 제품군

NetApp 핵심 데이터 제품, 기술 및 기능 외에도 강력한 AI 파트너 네트워크와 긴밀히 협력하여 고객에게 부가가치를 제공합니다.

- AI 시스템의 NVIDIA 가드레일*은 AI 기술의 윤리적이고 책임감 있는 사용을 보장하는 안전 장치 역할을 합니다. AI 개발자는 LLM 기반 애플리케이션의 동작을 특정 주제에 대해 정의하고 원치 않는 주제에 대한 토론에 참여하지 못하도록 방지할 수 있습니다. 오픈 소스 툴킷인 가드레일은 LLM을 다른 서비스에 원활하고 안전하게 연결하여 신뢰할 수 있고 안전하며 보안성이 높은 LLM 대화형 시스템을 구축할 수 있는 기능을 제공합니다.

*Domino Data Lab*은 빠르고 안전하며 경제적인 Generative AI를 구축하고 제품화하기 위한 다목적 엔터프라이즈급 도구를 제공합니다. AI 여정의 어느 단계에 있든 상관없습니다. Domino’s Enterprise MLOps 플랫폼을 사용하면 데이터 과학자는 선호하는 도구와 모든 데이터를 사용하고, 어디서나 쉽게 모델을 훈련하고 배포하고, 위험을 관리하고 비용 효율적으로 관리할 수 있습니다. 이 모든 것이 하나의 제어 센터에서 가능합니다.

Edge AI를 위한 **Modzy**. NetApp 과 Modzy는 이미지, 오디오, 텍스트, 표를 포함한 모든 유형의 데이터에 대규모 AI를 제공하기 위해 협력했습니다. Modzy는 AI 모델을 배포, 통합, 실행하기 위한 MLOps 플랫폼으로, 데이터 과학자에게 모델 모니터링, 드리프트 감지, 설명 가능성 기능을 제공하며, 원활한 LLM 추론을 위한 통합 솔루션을 제공합니다.

*Run:AI*와 NetApp Run:AI 클러스터 관리 플랫폼을 통해 AI 워크로드 오케스트레이션을 간소화하는 NetApp ONTAP AI 솔루션의 고유한 기능을 보여주기 위해 파트너십을 맺었습니다. 이 솔루션은 Spark, Ray, Dask, Rapids를 위한 기본 통합 프레임워크를 통해 수백 대의 머신으로 데이터 처리 파이프라인을 확장하도록 설계되어 GPU 리소스를 자동으로 분할하고 결합합니다.

결론

생성적 AI는 모델이 대량의 고품질 데이터를 통해 훈련될 때에만 효과적인 결과를 낼 수 있습니다. LLM은 놀라운 이정표를 달성했지만 데이터 이동성과 데이터 품질과 관련된 한계, 설계상의 과제 및 위험을 인식하는 것이 중요합니다. LLM은 다양한 데이터 소스의 방대하고 다양한 교육 데이터 세트에 의존합니다. 모델에서 생성된 부정확한 결과나 편향된 결과는 기업과 소비자 모두를 위험에 빠뜨릴 수 있습니다. 이러한 위험은 데이터 품질, 데이터 보안, 데이터 이동성과 관련된 데이터 관리 과제로 인해 잠재적으로 LLM에 발생하는 제약에 해당할 수 있습니다. NetApp 조직이 급속한 데이터 증가, 데이터 이동성, 멀티 클라우드 관리 및 AI 도입으로 인해 발생하는 복잡성을 해결할 수 있도록 지원합니다. 대규모 AI 인프라와 효율적인 데이터 관리는 생성 AI와 같은 AI 애플리케이션의 성공을 정의하는 데 매우 중요합니다. 고객이 비용 효율성, 데이터 거버넌스, 윤리적인 AI 관행을 통제하는 동시에 기업의 필요에 따라 확장할 수 있는 능력을 저하시키지 않고 모든 배포 시나리오를 포괄하는 것이 중요합니다. NetApp 고객이 AI 배포를 간소화하고 가속화할 수 있도록 끊임없이 노력하고 있습니다.