2025년 4월 | eBook

Nutanix 기반 엔터프라이즈 AI

가장 단순한 방법이 가장 빠른 방법입니다

NUTANIX

목차

엔터프라이즈 AI 도입 가속화

GenAI의 등장으로 기업들은 AI 전략을 재정비하고 도입 일정을 앞당기고 있습니다

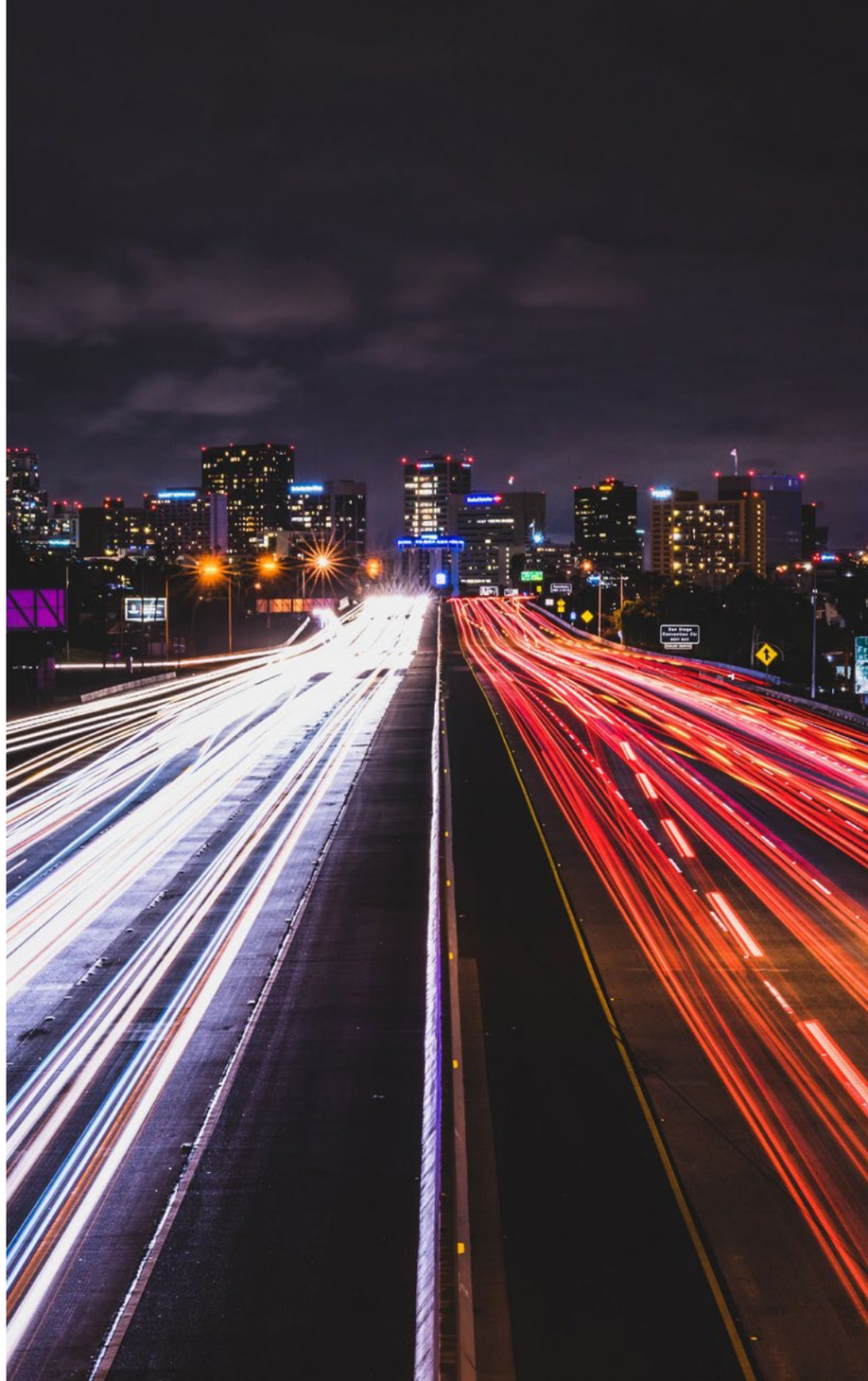
그러나 최근 IT 및 DevOps 플랫폼 엔지니어링 전문가 650명을 대상으로 한 설문조사에서, 다수의 기업이 GenAI 도입의 출발점조차 명확하지 않다고 응답했습니다. 주요 우려 사항은 다음과 같습니다.

- **IT 인프라 부족:** 응답자의 91%가 AI 워크로드를 지원하기 위해 IT 인프라 개선이 필요하다고 답했습니다.
- **데이터 보안 및 복원력 확보:** 데이터 보안, 복원력, 확장성은 엔터프라이즈 AI의 핵심 과제입니다.
- **기술 인력 부족:** 모든 조직이 향후 12개월 동안 AI 관련 이니셔티브를 지원하기 위해 추가 기술 인력이 필요하다고 답했습니다.

AI 도입 전 아래와 같은 세 가지 일반적인 접근 방식을 고려합니다.

- **클라우드에만 의존하기:** 클라우드를 활용해 AI 실험, 개발, 앱 프로덕션을 진행할 수 있지만 비용이 많이 들 수 있습니다. 또한 다른 모든 사용자와 클라우드 자원을 두고 경쟁해야 하고, 데이터가 위험에 노출될 수 있습니다.
- **데이터센터 및 엣지에서 맞춤형 AI 스택 구축:** 이 접근 방식은 상당한 전문성과 시간이 필요합니다. 사내 전문 인력이 있거나 헌신적인 파트너가 있는 경우에만 고려해야 합니다. 그렇지 않으면 AI 도입 속도가 늦어지고 시행착오가 반복될 수 있습니다.
- **턴키 솔루션:** 턴키 솔루션은 경험이 부족한 팀도 빠르게 시작할 수 있게 해주며, 최고 수준의 하드웨어와 AI 소프트웨어를 통해 유연성을 높일 수 있습니다.

이 eBook은 AI를 계획하는 팀들이 직면하는 어려움을 살펴보고, Nutanix 클라우드 플랫폼과 Nutanix GPT-in-a-Box가 온프레미스, 클라우드, 엣지 환경에서 AI를 빠르고 쉽게 프로덕션 환경으로 전환할 수 있도록 돕는 방법을 설명합니다. 이를 통해 더 짧은 시간 안에 더 나은 결과를 얻을 수 있습니다.



AI 주요 과제

조직의 85%는 기존 AI 모델을 구매하거나 오픈 소스 AI 모델을 활용해 AI 애플리케이션을 구축할 계획이며, 단 10%만이 자체 모델을 구축할 계획입니다.

많은 기업에서 AI 모델을 처음부터 설계하고 학습시킬 필요가 없습니다. 사전에 학습된 파운데이션 모델을 라이선스 받아 프라이빗 데이터를 활용해 확장하는 것이 훨씬 효율적입니다. 대표적인 접근 방식은 다음과 같습니다.

- **파인튜닝:** 프라이빗 데이터를 활용해 모델을 파인튜닝하여 요구사항에 맞게 최적화.
- **검색 증강 생성(RAG):** 모델 외부 데이터를 활용해 기업 맞춤형 결과를 제공.

이러한 옵션을 활용해 실험 단계에서 AI를 실제 프로덕션 배포로 전환하는 과정에서 다음과 같은 중요한 도전 과제에 직면하게 됩니다.

복잡성

GenAI를 활용하려면 컨테이너화된 LLM을 배포하고, 파인튜닝 또는 RAG를 구현하며, 이를 MLOps를 통해 반복적으로 프로덕션 환경에 배포할 수 있어야 합니다.

이를 위해서는 파인튜닝용 인프라와 추론용 인프라를 구분해 갖춰야 합니다. 추론은 지연 시간을 줄이고 응답성을 높이기 위해 서비스 소비자와 가까운 엣지에서 수행되는 경우가 많습니다.

- 엣지에서 추론을 수행하면 원격 인프라 관리에 어려움이 따릅니다.
- 데이터센터이든 엣지든, 추론과 학습 모두 요구 변화에 따라 빠르게 확장하면서도 과도한 리소스 투입을 피할 수 있는 적절한 연산 능력, GPU, 스토리지 자원의 조합이 필요합니다.

AI 운영을 안정적으로 구축하려면, 발생하는 문제의 근본 원인이 모델, 인프라, 또는 MLOps 프로세스의 약점 중 어디에 있는지를 파악할 수 있도록 도와주는 지능형 도구가 필요합니다.

규정 준수

AI 팀은 종종 IT 부서와 별도로 구성되며, 핵심 업무의 복원력, 데이터 관리, 개인정보 보호(PII) 같은 요소를 충분히 고려하지 않고 솔루션 개발에 집중하는 경우가 많습니다.

이로 인해 GenAI 또는 기타 AI 기반 애플리케이션을 배포·운영할 때 IT 정책(보안, 데이터 보호, 복원력, 운영 규정)을 훼손하지 않고 준수하는 것이 쉽지 않습니다.

모든 학습 데이터는 반드시 정제되어야 하며, 이름, 주소, 전화번호, 주민등록번호, 금융·신용카드 정보 등 PII가 포함되지 않도록 해야 합니다.

비용

클라우드는 AI 운영에 장점이 있지만, 클라우드에서 AI 모델을 실행하는 비용은 온프레미스 대비 훨씬 높을 수 있습니다.

반대로, AI 인프라는 본질적으로 많은 전력을 소비합니다. 따라서 전력, 냉각, 물리적 공간, 운영 관리 등 2차적 요소까지 고려해야 데이터센터나 엣지에서 AI 워크로드를 운영할 수 있습니다.

거버넌스

AI에는 여러 거버넌스 과제가 뒤따릅니다.

- **데이터 관리 및 데이터 주권:** 엣지, 코어, 클라우드 전반에서 정책 기반으로 일관된 데이터 보호 및 보안을 유지하는 것은 어려운 일입니다. 엣지에서 생성된 데이터는 추가 학습에 필요할 수 있으므로, 위치 간 데이터 이동과 관리에는 단순·자동화된 방식이 필수적입니다.

  경우에 따라, 데이터 주권 규제를 위반하지 않기 위해 데이터를 해당 위치에 안전하게 보관해야 하며, 학습이 필요한 경우 모델을 데이터가 있는 곳으로 옮겨 학습을 수행해야 합니다.

  이러한 위치가 여러 곳일 경우, 모델을 각 위치에서 순차적으로 학습시키는 연합학습(Federated Learning) 방식이 필요할 수 있습니다.

- **MLOps:** AI 팀은 모델의 각 버전에 대해 학습에 사용된 정확한 데이터세트, 학습 시점, 배포 장소 등을 완벽히 추적·관리할 수 있어야 합니다.

  이는 전통적인 소프트웨어 개발에서 코드 버전을 관리하는 것보다 훨씬 복잡할 수 있습니다. 학습 반복 주기와 데이터셋의 양·복잡성에 따라 난이도가 크게 증가합니다.

보안 및 데이터 개인정보 보호

AI와 관련해 여러 추가적인 보안 우려가 발생합니다.

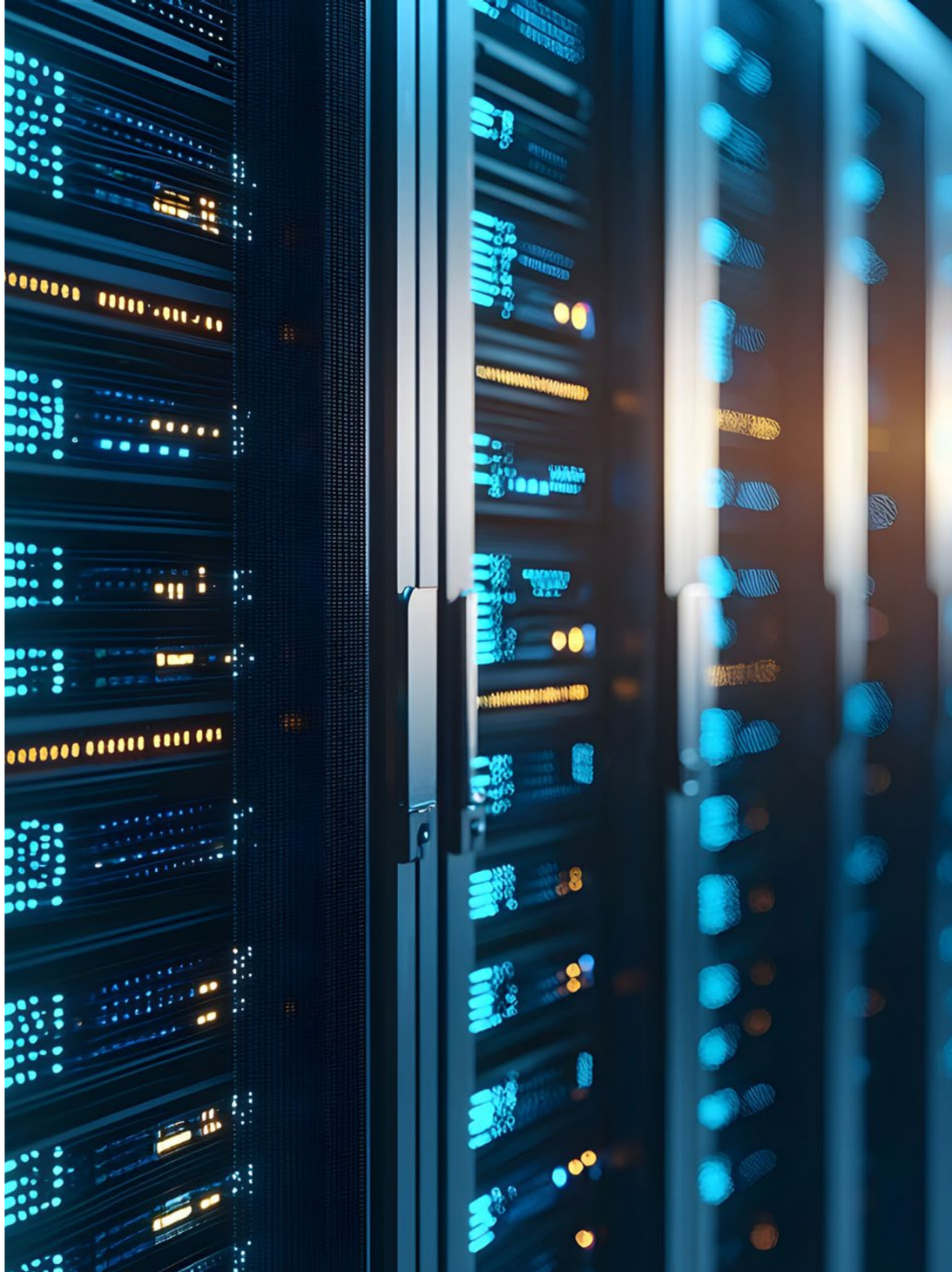
- **데이터 유출:** 클라우드에서 실행되는 비주권 AI 서비스를 활용할 때, 기밀 또는 독점 데이터를 포함한 회사 정보를 입력하고 싶은 유혹이 큽니다.

  예를 들어, 출판되지 않은 보고서 내용을 Copilot, Gemini, ChatGPT 등에 입력해 요약을 요청하거나, OpenAI의 Whisper 모델을 사용해 내부 회의의 녹취록을 생성하는 행위는 겉보기에는 무해해 보일 수 있습니다.

  하지만 모델이 데이터를 처리한 후 그 데이터가 어떻게 되는지는 알 수 없습니다. Microsoft, Google, AWS, OpenAI 같은 대기업은 신뢰할 만할지 몰라도, GenAI 앱과 서비스의 생태계가 급격히 확장되면서 위험이 증가하고 있습니다.

- **데이터 오염:** 승인되지 않은 악의적인 데이터가 LLM에 주입되면, 해당 모델이 여러분의 통제 하에 있더라도, 잘못된 생성 결과나 편향된 결과를 초래할 수 있습니다.
- **프롬프트 인젝션:** 악의적인 프롬프트가 삽입되면 AI의 안전 장치를 우회하고 LLM의 응답을 조작할 수 있습니다.
- **보안 도구:** 현재 사용 중인 보안 도구가 MLOps와 호환되지 않을 수 있으며, 이 경우 업그레이드나 완전히 새로운 도구가 필요합니다.

Nutanix 클라우드 플랫폼은 혁신적인 소프트웨어와 독자적인 하이퍼컨버지드 인프라(HCI)를 활용하여 이러한 과제를 온프레미스, 엣지, 퍼블릭 클라우드 환경에서 해결합니다.



왜 Nutanix 클라우드 플랫폼이 AI에 적합한가?

Nutanix는 인프라 복잡성을 줄이고 하이브리드 멀티클라우드 운영을 가능하게 하는 데 집중하고 있습니다. 검증된 하이퍼컨버지드 인프라(HCI)를 기반으로 구축된 Nutanix 클라우드 플랫폼(NCP)은 엣지부터 코어 데이터센터, 퍼블릭 클라우드에 이르기까지 AI 요구사항을 충족하는 민첩하고 탄력적인 인프라를 제공합니다.

AI 운영 간소화

운영 복잡성은 AI를 프로덕션 환경에 배포하는 과정에서 또 다른 큰 도전 과제입니다. 자원 제약으로 인해 많은 조직이 AI 실험을 클라우드에서 수행합니다.

하지만 클라우드에서 진행한 작업을 데이터센터로 옮기거나, 엣지에서 실제 운영으로 전환하려면 어떻게 해야 할까요?

NCP는 인프라의 복잡성과 한계를 제거하여 AI 및 머신러닝 이니셔티브를 빠르게 시작할 수 있도록 합니다. Nutanix를 활용하면 엣지, 데이터센터, 클라우드 환경 전반에서 애플리케이션, 워크로드, 데이터의 원활한 이동성을 확보할 수 있습니다.

또한 NCP는 어디서나 실행 가능하기 때문에, 팀은 모든 환경에서 효율적으로 작업할 수 있습니다.

NCP at the Edge

엣지는 원격으로 복잡한 인프라를 배포·관리해야 한다는 점에서 특히 어려운 지점이 될 수 있습니다. 그러나 NCP는 검증된 HCI 설계(컴퓨트, 네트워킹, 스토리지 통합)를 기반으로, 엣지 배포의 요구사항을 고유하게 충족합니다.

- 컴팩트한 설치 공간
- 원격 관리 및 원격 앱 배포의 용이성
- 고급 데이터 보호 및 보안 기능
- 제약 없는 운영
- 완전한 데이터 서비스 패키지 제공

Nutanix GPT-in-a-Box

많은 기업들이 데이터 주권, 거버넌스, 개인정보 보호 문제로 인해 퍼블릭 클라우드에서 실행할 수 없는 사용 사례에서 GenAI를 활용하는 데 어려움을 겪고 있습니다.

Nutanix GPT-in-a-Box는 완전한 통제를 유지하면서 LLM 및 기타 AI 모델을 파인튜닝 및 실행할 수 있는 즉시 사용 가능한 턴키형 프라이빗 AI 솔루션을 제공합니다.

이 솔루션은 GenAI 애플리케이션 개발 시 직면하는 복잡성, 확장성, 보안 문제를 해결합니다. 소프트웨어 정의형 풀스택 아키텍처로 구축된 AI-ready 플랫폼인 GPT-in-a-Box는 NCP 위에서 실행되며, 배포를 단순화하고 AI 이니셔티브의 속도를 가속화합니다.

GPT-in-a-Box의 주요 기능

- **풀스택 AI 플랫폼:** 원하는 하드웨어(CPU/GPU), VM 또는 컨테이너, 다양한 LLM 및 AI 프레임워크를 선택해 어디서든 GenAI를 배포 가능.
- **AI 전용 데이터 서비스 내장:** 파일, 블록, 오브젝트 데이터를 위한 보안적이고 확장성 있는 데이터 서비스 제공. 스냅샷 및 재해 복구 기능을 통합적으로 제어 가능.
- **확장 가능한 AI 워크로드**: 일관된 플랫폼 서비스를 활용하여 클라우드 운영을 통합하고, 최소한의 노력으로 어디서든 확장 가능한 AI 제공.
- **즉시 사용 가능한 LLM 배포:** 검증된 GenAI 모델에 접근하여 엣지부터 클라우드까지 빠르게 배포하고, Time-to-Value를 가속화.
- **AI 앱을 위한 보안 API 운영:** 역할 기반 보안 API를 손쉽게 생성해 앱과 AI 모델을 안전하게 연결 가능.

Nutanix GPT-in-a-Box를 통해 기업은 AI 학습 데이터와 모델을 내부적으로 유지·관리함으로써 보안, 개인정보 보호, 규정 준수 요구사항을 충족하는 동시에 IT 비용을 최적화할 수 있습니다.

GPT-in-a-Box에는 GenAI 모델 실행을 시작하는 데 필요한 모든 것이 포함되어 있으며, 기업이 준비해야 할 것은 선택한 파운데이션 모델뿐입니다.

유연한 GPU 및 CPU 옵션

AI 성공을 위해서는 적절한 GPU와 CPU를 갖추는 것이 핵심입니다.

GPU

Nutanix는 다양한 비즈니스 요구에 맞춰 NVIDIA GPU 전체를 지원하며, NCP의 GPU 패스스루 기능을 통해 가상화·컨테이너 환경에서도 GPU를 효율적으로 활용할 수 있습니다.

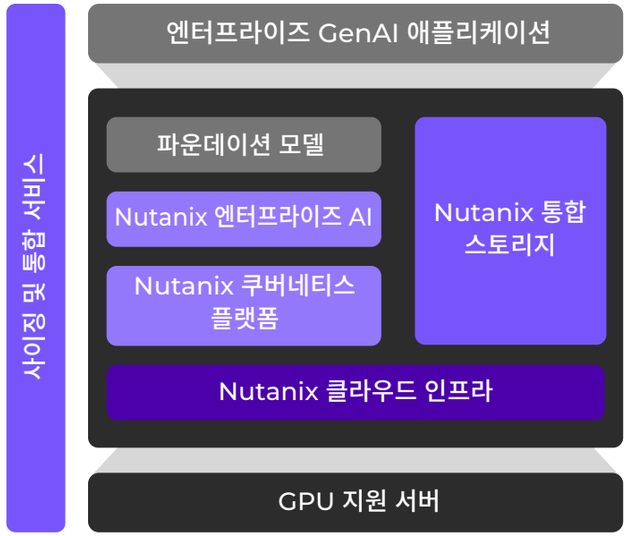
GPU 패스스루를 통해 VM에서 GPU 리소스에 직접 액세스할 수 있으며, 클러스터 전체의 GPU 뷰를 제공해 사용 가능한 GPU를 어떤 VM에도 할당할 수 있습니다. 사용 가능한 GPU를 VM에 할당할 수 있도록 합니다. 하나의 VM에 여러 GPU를 할당할 수 있습니다. 패스스루를 사용하면 한 번에 하나의 VM만 GPU를 사용할 수 있습니다.

CPU

GPU가 AI에 필수라는 인식이 있지만, 최신 세대의 CPU는 학습과 추론을 가속화하는 기능을 갖추고 있어 일부 AI 작업은 GPU 없이도 충분히 지원 가능합니다.

적절한 CPU는 병렬 처리 능력, 메모리 용량, 대역폭을 제공해 GPU 활용 효율을 극대화합니다.

제공 가능한 서비스에는 계획 워크샵, 스택 설계 워크샵 및 스택 배포 서비스가 있습니다.



AI-Ready 번들 서비스는 주요 오픈소스 AI 및 MLOps 프레임워크를 활용하여 선별된 LLM 세트를 배포할 수 있도록, 적절한 인프라 규모 산정과 구성을 지원합니다.

업계 파트너십

Nutanix는 업계를 선도하는 AI 기업들과 협력하여 풀스택 솔루션을 제공합니다.

특히 NVIDIA와의 파트너십을 통해, Nutanix AHV 하이퍼바이저는 NVIDIA AI Enterprise 실행 인증을 획득했습니다.
NVIDIA AI Enterprise는 클라우드 네이티브 소프트웨어 플랫폼으로, 프로덕션 환경 수준의 AI 솔루션 개발과 배포를 간소화합니다.

여기에는 AI 에이전트, 컴퓨터 비전, 음성 AI 등을 포함한 다양한 기능이 지원되며, NVIDIA AI Enterprise 인증을 통해 이러한 기능이 NCP에서 안정적으로 활용 가능함을 보장합니다.

NCP는 NVIDIA AI Enterprise와 최적의 조합을 이루며, 민첩성, 효율성, 확장성을 극대화하는 환경을 제공합니다. 또한 NVIDIA GPU 연산 가속기와의 통합을 통해, AI 워크로드가 GPU 리소스에 직접 접근할 수 있어 지연 시간을 줄이고 성능을 가속화합니다.

Nutanix는 NVIDIA뿐만 아니라 Intel 및 AMD와도 긴밀한 파트너십을 맺고 있습니다. 이를 통해 해당 기술 리더들의 AI 혁신을 신속히 추적하고, 자사 플랫폼에 빠르게 지원 기능을 추가할 수 있습니다.

AI 사용 사례

많은 기업들에게 지능형 챗봇, 지원 코파일럿, 스마트 문서 처리와 같은 GenAI 활용 사례는 최우선 과제입니다. 그러나 Nutanix는 이보다 더 넓은 범위의 AI 활용 사례를 지원할 수 있도록 다양한 옵션을 제공합니다.

NVIDIA AI Enterprise 및 NVIDIA NIM

Nutanix는 NVIDIA AI Enterprise 인증을 받았기 때문에, 컴퓨터 비전, 음성, 언어 이해, 분자 생성을 포함한 다양한 활용 사례와 도메인을 아우르는 폭넓은 모델을 지원합니다.

NVIDIA AI Enterprise 고객은 GPT-in-a-Box를 통해 GenAI를 위한 최적화된 클라우드 네이티브 마이크로서비스 세트인 NVIDIA NIM을 손쉽게 배포할 수 있습니다.

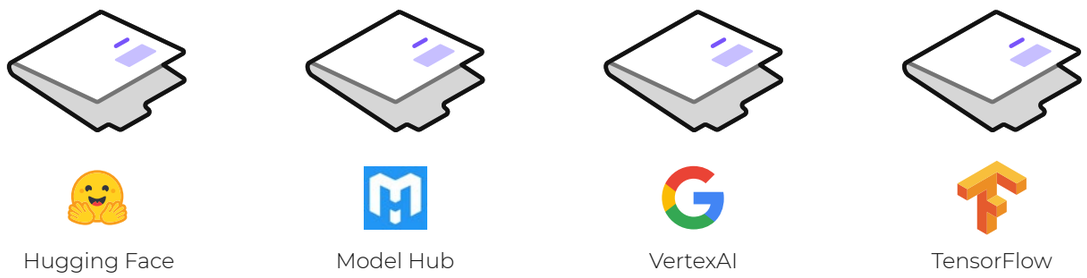
Hugging Face 및 기타 모델 허브

Nutanix는 Hugging Face와 파트너십을 맺어 모델 배포를 신속하게 지원합니다. 이를 통해 Hugging Face의 LLM을 GPT-in-a-Box와 손쉽게 통합할 수 있습니다. 검증된 AI LLM을 검색, 다운로드, 배포하는 과정을 전폭적으로 지원하여 매끄러운 워크플로우를 제공합니다.

또한 GPT-in-a-Box는 사용자가 직접 선택한 비검증·비지원 모델도 업로드 및 배포할 수 있도록 지원합니다.

모델 허브

자유롭게 활용 가능한 오픈소스 AI 모델 저장소



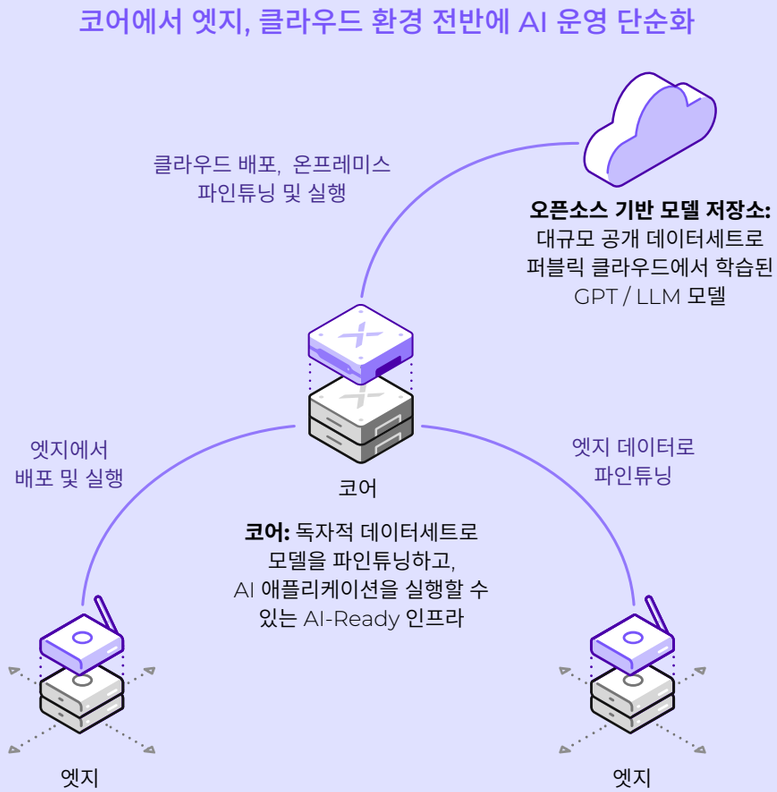
Nutanix 시작하기

Nutanix는 복잡성과 장애 요인을 제거하고, 기업이 GenAI 및 기타 AI 애플리케이션을 성공적으로 프로덕션 환경에 배포할 수 있도록 지원하는 데 집중하고 있습니다.

NCP는 엣지에서 코어 데이터센터, 그리고 클라우드까지 AI 운영을 단순화하며, 데이터 서비스를 단순화하고, 앞으로 나아가는 데 필요한 데이터 보호와 보안을 제공합니다. 여기에 턴키형 GPT-in-a-Box 솔루션까지 더해져, 기업은 GenAI를 더욱 손쉽게 시작할 수 있습니다.

Nutanix 엔터프라이즈 AI 솔루션에 대해 자세히 알아보려면 AI 솔루션 페이지를 클릭하고 AI 테스트 드라이브를 직접 체험해 보시기 바랍니다.

테스트 드라이브 체험하기



NUTANIX
info@nutanix.com | www.nutanix.com/kr | @nutanix