ORACLE

AI 워크로드의 퍼포먼스를 극대화하는 OCI Supercluster

어디서든 구축 가능한 고성능 확장형 Oracle AI 인프라

AI 인프라 혁신의 역사

확장 가능하고 비용 효율적인 AI

생성형 AI를 사용하기 위해서는 모델 훈련 및 배포를 확장하고 가속화하기 위한 비용 효율적 고성능 인프라가 필요합니다.

Oracle은 현재 AI 워크로드의 필수 요소로 자리잡은 고성능 인프라 분야를 오랫동안 선도해 왔습니다. 2008년 출시된 Oracle Exadata는 독특하고 혁신적인 Remote Direct Memory Access(RDMA) 기술을 선보였습니다. 2016년, Oracle Cloud Infrastructure(OCI)는 고객을 위한 최대의 성능 및 제어 능력을 갖춘 베어메탈 컴퓨트 인스턴스를 최초로 제공했습니다.

2023년 출시된 OCI Supercluster는 전 세계에서 가장 뛰어난 성능을 저렴한 비용으로 이용할 수 있는 업계 최고의 GPU 클러스터 중 하나입니다. 논블로킹 네트워크의 하드웨어 지원 RDMA, 상당한 규모의 로컬 NVMe 스토리지, 베어메탈 컴퓨팅은 AI 훈련을 위한 이상적인 환경을 제공합니다.

Oracle은 최근 대규모 언어 모델(LLM)의 훈련을 지원하기 위해 페타바이트 규모의 관리형 고성능 마운트 타겟 파일 시스템을 출시했습니다. 최근에는 최고 성능의 스토리지를 필요로 하는 조 단위 매개변수 모델을 위한 관리형 Lustre 파일 시스템 서비스를 추가로 출시했습니다.

OCI Supercluster로 혁신하기

OCI Supercluster는 NVIDIA Blackwell GPU를 131,072개까지 확장할 수 있는 업계 최고 성능의 AI 구축 및 배포용 클라우드 환경을 제공합니다.* 동시에 최대로 확장시 이전 세대와 동일한 성능을 제공하면서도 25배 낮은 에너지 소비량과 25배 향상된 TCO를 보장합니다.

NVIDIA GB200(Grace Blackwell) 및 B200(Blackwell)이 탑재된 OCI Supercluster를 대규모 AI 훈련 및 추론에 지금 바로 사용할 수 있습니다. OCI Superclusters는 NVIDIA H100 Tensor Core GPU가 탑재된 전 세대 OCI Supercluster 대비 AI 추론 성능은 최대 30배, AI 훈련 성능은 4배 향상되었습니다.

또한 NVIDIA H200 Tensor Core GPU가 탑재된 OCI Supercluster는 전 세대 NVIDIA H100 GPU 인스턴스 대비 76% 높은 대역폭의 GPU 메모리 용량 및 40% 많은 GPU 메모리 대역폭을 갖춘 인스턴스를 제공함으로써 LLM 추론 성능을 최대 1.9배까지 향상시킬 수 있습니다. 데이터 수집 및 검색을 위한 프론트엔드 네트워크 처리량이 인스턴스마다 초당 200Gb로 2배 향상되어 클러스터 간 데이터 전송을 획기적으로 개선함으로써 AI 모델 훈련 속도 역시 더욱 빨라졌습니다.

OCI를 활용 중인 혁신 기업

CSM

Common Sense Machines는 OCI의 전문가 조언 서비스를 활용해 생성형 AI 스타트업을 지원합니다.
더 알아보기

Twelve Labs

Twelve Labs는 모델 훈련 효율을 5배에서 10배까지 향상시킴으로써 시장 출시 기간을 예상보다 훨씬 앞당겼습니다.
더 알아보기

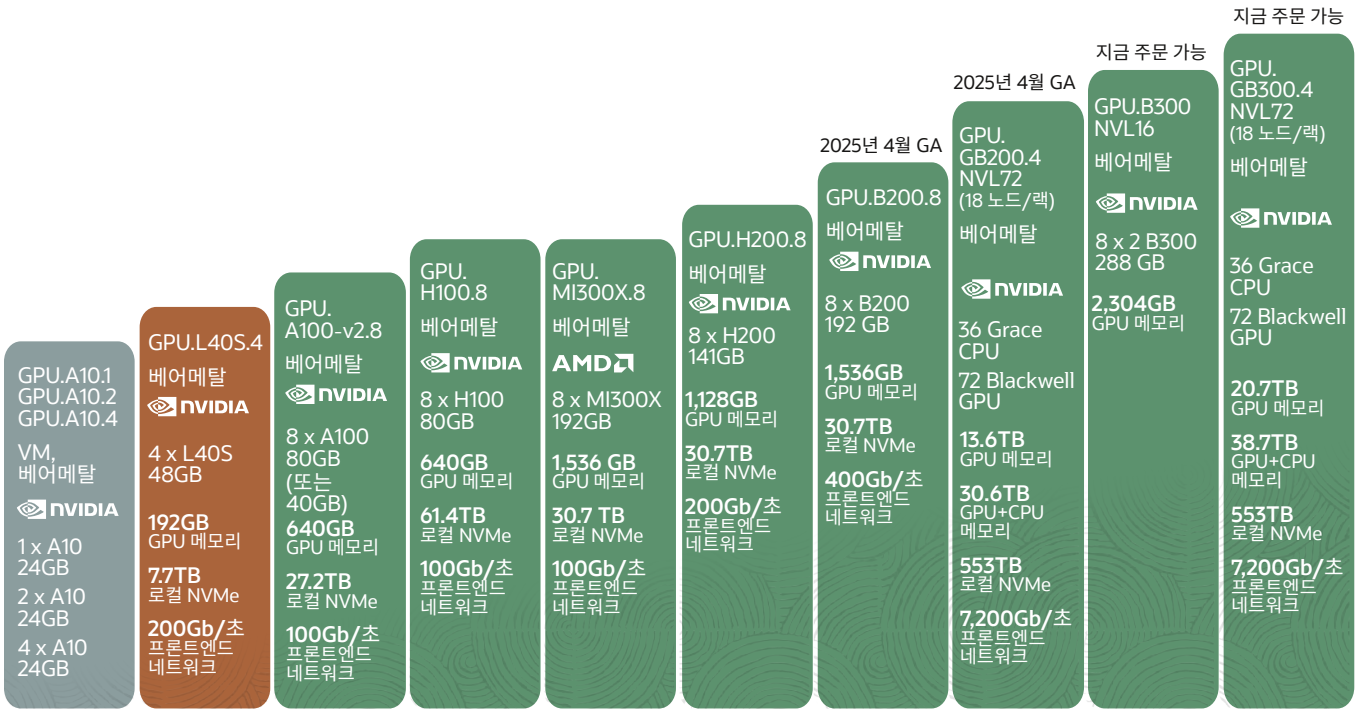
SUNO

Suno AI는 확장 가능한 고성능 OCI Supercluster에서 훈련된 AI 기반 모델을 통해 고품질의 음악 및 음향을 생성합니다.
더 알아보기

*CSP 1의 확장성: NVIDIA H200 GPU 20,000개, CSP 2 및 CSP 3의 확장성: 공개되지 않음

모든 워크로드에 최적화되어있는 AI 인프라

OCI는 모든 규모의 AI 워크로드를 지원합니다. 아래의 차트를 통해 OCI Supercluster의 확장성과 더불어 현재 사용 가능하거나 곧 출시될 인스턴스의 범위를 확인할 수 있습니다. OCI는 OCI Dedicated Region, Oracle Alloy 등 업계에서 가장 광범위한 옵션을 통해 어디든 배포할 수 있습니다.



OCI는 Compute Cloud at Customer용 NVIDIA L40S GPU 및 엣지 배포용 NVIDIA L4 GPU를 제공합니다.

OCI는 OCI Dedicated Region 및 Oracle Alloy용 NVIDIA H100 및 H200 Tensor Core GPU를 제공합니다. NVIDIA B200 Tensor Core GPU는 현재 주문 가능 단계입니다.

Oracle AI 인프라만의 장점

탁월한 확장성

OCI는 2025년 기준 최대 131,072개 NVIDIA GPU를 지원하는 슈퍼클러스터 확장성을 바탕으로 세계에서 가장 큰 규모의 생성형 AI 배포 관련 요구 사항을 충족할 수 있습니다.

베어메탈 인스턴스

Oracle AI Infrastructure는 이더넷 및 NVIDIA Quantum-2 InfiniBand 기반의 전용 네트워크 및 가상화 계층의 오버헤드를 제거해 성능을 향상시킬 수 있는 베어메탈 컴퓨팅 인스턴스를 함께 제공합니다.

업계를 선도하는 분산 클라우드

OCI는 퍼블릭 클라우드, 소버린 및 정부 클라우드, 온프레미스 데이터 센터, 파트너사 데이터 센터를 모두 아우르는 선도적인 분산 클라우드 인프라 공급업체입니다.

맞춤형 지원

OCI는 연중무휴 24시간 운영 지원 및 AI에 대한 전문성을 갖추고 AI 인프라 배포, 문제 해결, 관리 전반을 안내해 주는 전담 클라우드 엔지니어를 제공합니다.

가격 정책

세계 어디서든 AI 인프라에서 실행되는 서비스를 비롯한 모든 서비스에 균일한 가격 정책을 적용하는 OCI의 GPU 인스턴스는 타사 CSP 대비 비용이 월등하게 저렴합니다.

Uber가 Oracle을 선택한 이유

Uber

Uber, OCI를 사용해 시간당 백만 건 이상의 운행 정보 처리

Uber는 자사의 애플리케이션 계층 및 AI 인프라를 최신화하고 운영 빅데이터 및 스트리밍 스택의 대부분을 OCI로 이전했습니다. 이로써 수익성을 늘리면서도 사업을 확장했고, 새로운 제품과 서비스를 더욱 빨리 출시하며 혁신적인 변화를 이끌고 있습니다.

1천 4백만

초당 예측 건수

1백만

시간당 운행 횟수

Zoom이 Oracle을 선택한 이유

zoom

기업의 업무 방식을 혁신하는 Zoom AI Companion

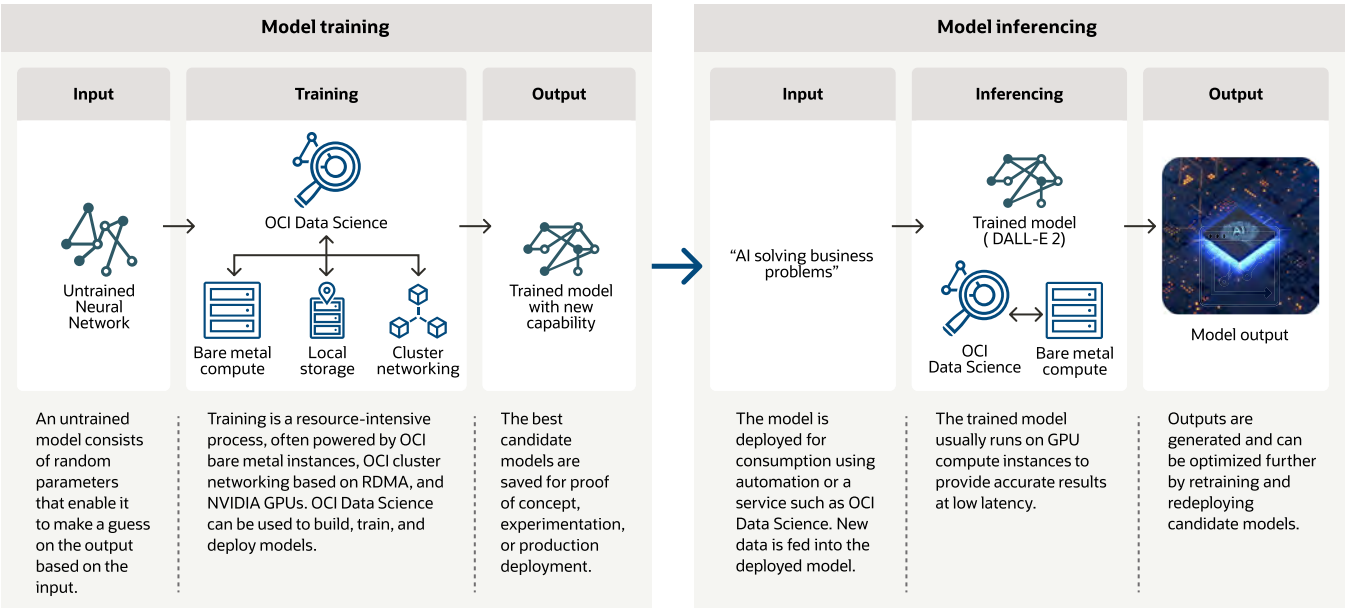
사람과 사람을 연결하는 AI 기반 업무 플랫폼인 Zoom은 사용자 수를 늘리는 과정에서도 성능 개선과 인프라 비용 절감을 동시에 성공해냈습니다. Zoom의 개인 AI 어시스턴트인 Zoom AI Companion은 사용자가 정보 검색, 이메일 및 채팅 메시지 초안 작성, 미팅 내용 요약 및 실행 가능성 향상, 브레인스토밍 개선, 콘텐츠 제작 등의 작업을 Zoom Workplace 앱에서 바로 수행할 수 있도록 도움을 줍니다.

> “Zoom은 OCI의 AI 추론 기능을 바탕으로 더 빠르고 정확한 결과를 제공함으로써 사용자들이 원활히 협업하고, 간단히 커뮤니케이션하고, 전례없는 생산성, 효율성, 잠재력 향상을 달성할 수 있도록 지원하고 있습니다.
>
> **Bo Yan**
> Zoom, Head of AI

AI 인프라의 주요 사용 사례 살펴보기

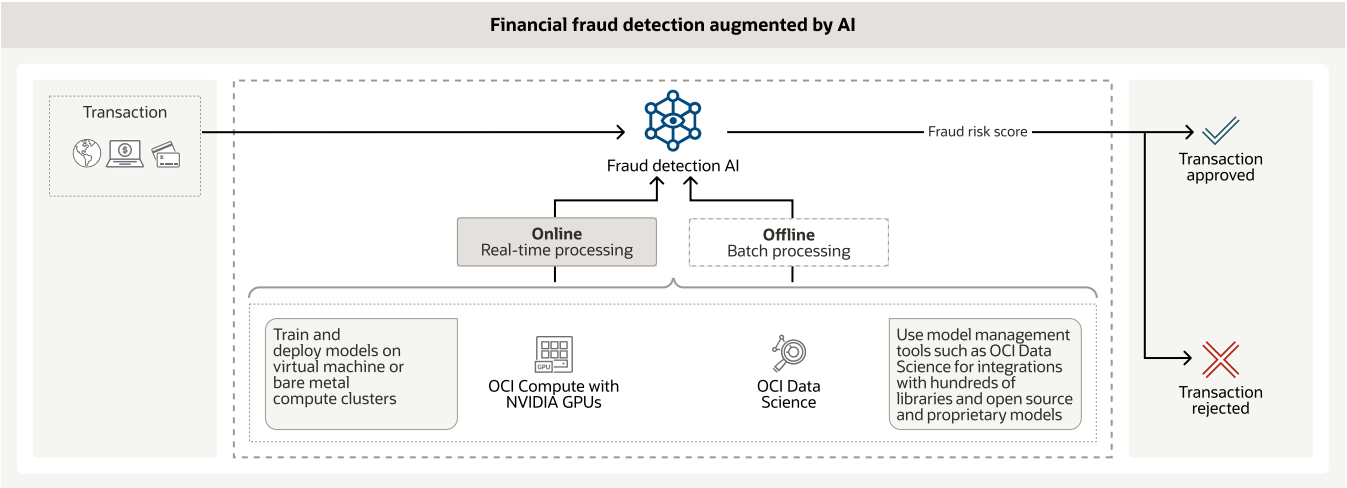
AI 훈련 및 추론

AI 스타트업 및 기업은 베어메탈 GPU 인스턴스 및 초고속 클러스터 네트워킹을 사용해 AI 모델을 학습시킬 수 있습니다. Oracle 관리형 및 자체 관리형 Kubernetes 통합관리를 위한 개발자 서비스를 활용해 보세요. GPU, 사전 구축된 AI 서비스용 API, PyTorch, TensorFlow, Kueue와 같은 타사 도구 등을 지원합니다.



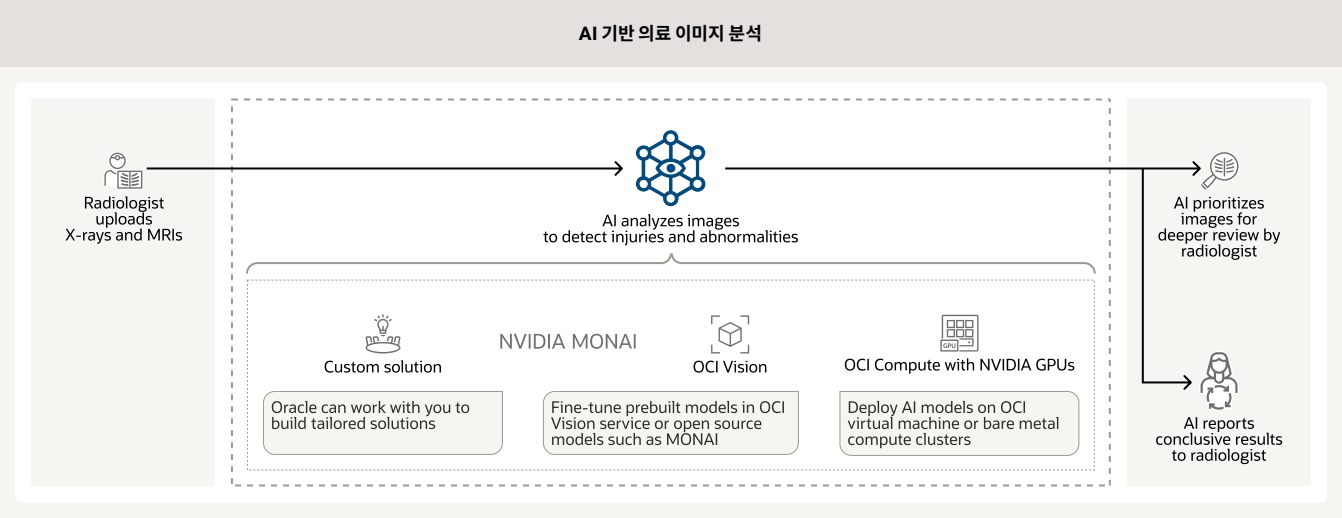
AI로 강화된 사기 감지

매일 수십억 건에 달하는 금융 거래를 보호하기 위해서는 방대한 과거 거래 데이터를 학습한 향상된 AI 도구를 사용해야 합니다. OCI에서 실행되고 NVIDIA 및 AMD GPU로 가속화된 AI 모델은 금융 기관의 사기 방지에 도움이 됩니다.



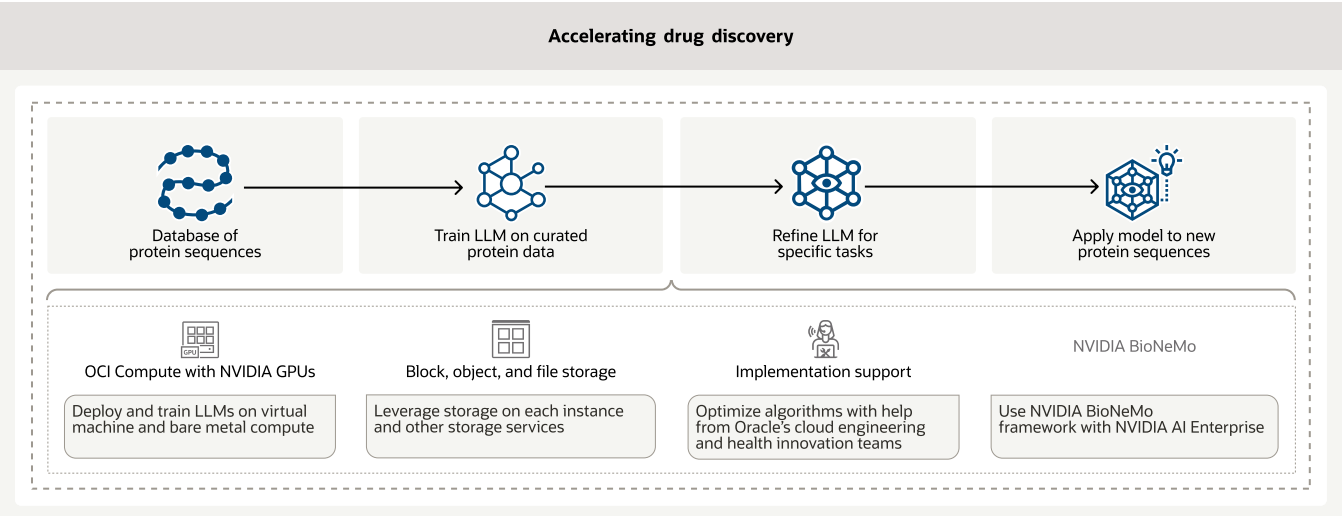
AI 기반 의료 이미지 분석

훈련된 모델은 엑스레이, CT 스캔, MRI를 분석해 방사선 전문의가 즉각적으로 검토해야 하는 이미지들을 우선순위로 선정하고, 다른 이미지들을 분석해 최종 결과를 보고합니다.



AI로 가속화되는 신약 개발

이제 연구자들은 AI 인프라 및 데이터 분석을 활용해 과거에는 몇 년씩 걸리곤 했던 신약 개발을 가속화할 수 있게 되었습니다. 또한 NVIDIA BioNeMo™와 같은 AI 워크플로 관리 도구는 연구자들이 데이터를 선별하고 사전 처리하는 데 도움을 줄 수 있습니다.



분산 클라우드 및 소버린 AI

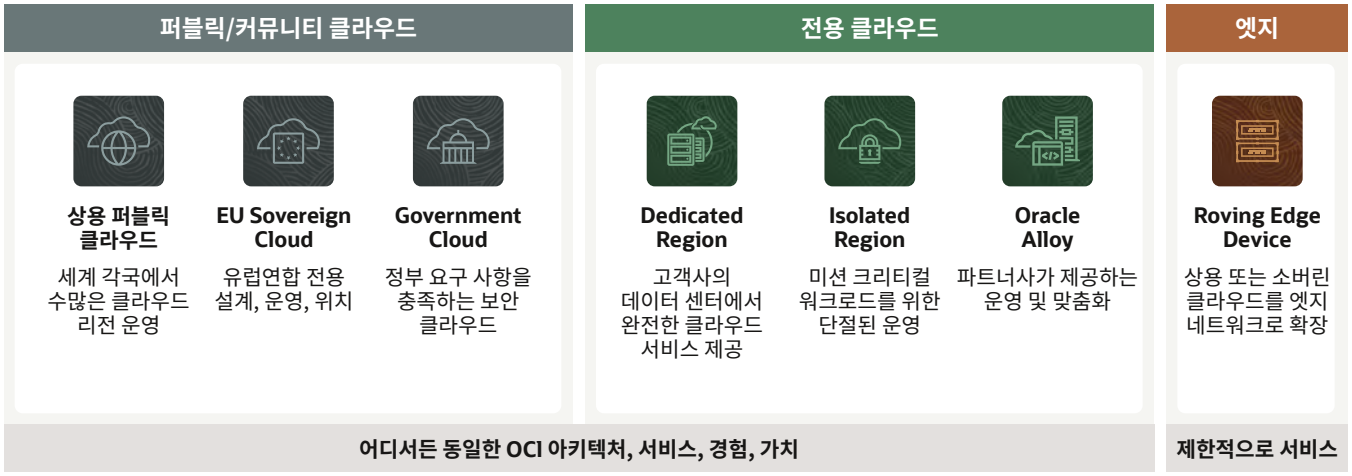
OCI는 Gartner Magic Quadrant for Distributed Hybrid Infrastructure 부문의 리더로 선정되었으며 해당 보고서의 '주권' 및 '분산 클라우드 아키텍처' 관련 핵심 기능 부문에서 높은 점수를 획득했습니다.

OCI는 160개 이상의 현재 사용 가능하거나 향후 개소 예정인 리전들을 통해 다른 어떤 하이퍼스케일러보다도 많은 지역에서의 워크로드 실행을 지원합니다. 또한 OCI는 고객이 보유 중인 가장 유용한 데이터를 원하는 하이퍼스케일 클라우드로 가져올 수 있는 Oracle Database@Azure, Oracle Database@Google Cloud, Oracle Database@AWS 등의 독보적인 멀티클라우드 기능을 제공합니다.

OCI Dedicated Region 및 Oracle Alloy는 정부 또는 기업이 AI 기술 및 관련 데이터를 직접 관리할 수 있는 소버린 AI를 지원합니다. Oracle 고객은 하드웨어 및 소프트웨어 인프라, AI 기술 운영 및 데이터 보호에 사용되는 정책과 인력 등 AI 기술을 배포하고 운영하는 방법과 위치를 직접 결정할 수 있습니다.

고객은 Oracle AI 및 OCI 분산 클라우드 솔루션을 활용해 AI 주권을 확보함으로써 AI 워크로드의 실행 위치, 데이터 및 시스템 관리 방식을 직접 결정할 수 있습니다.

OCI의 분산 클라우드는 모든 지점에서의 AI 애플리케이션 배포를 지원합니다



귀사의 성공을 위한 NVIDIA와의 파트너십

OCI의 AI 클라우드 인프라와 결합된 NVIDIA 소프트웨어는 광범위한 도구 및 서비스 포트폴리오를 편리하게 제공함으로써 귀사의 대규모 AI 훈련 및 추론을 지원합니다.

OCI와 함께 사용 가능한 NVIDIA 소프트웨어는 다음과 같습니다.

- 어디서든 생성형 AI 모델 배포를 가속화시켜주는 NVIDIA NIM™ 마이크로서비스
- NVIDIA DGX™ Cloud - 최적화된 가속 컴퓨팅 클러스터를 제공하는 완전 관리형 고성능 AI 플랫폼
- NVIDIA AI Enterprise - 운영 AI를 위한 엔드투엔드 소프트웨어 플랫폼
- NVIDIA RAPIDS™ - GPU를 사용해 Apache Spark 워크로드를 비롯한 데이터 과학 및 모델 훈련을 가속화할 수 있는 오픈 소스 라이브러리 및 API 모음집
- NVIDIA TensorRT-LLM - LLM 추론 최적화를 위한 라이브러리로서 커스텀 어텐션 커널, 인플라이트 배칭, 페이징된 KV 캐싱, 양자화(FP8, INT4 AWQ, INT8 SmoothQuant, ++) 등의 최첨단 최적화 기술을 제공해 NVIDIA GPU를 활용한 효율적 추론을 지원합니다.
- NVIDIA Triton Inference Server - 모든 워크로드의 AI 모델 배포 및 실행을 표준화할 수 있는 오픈 소스 소프트웨어
- NVIDIA BioNeMo™ - 컴퓨터 기반 신약 개발을 지원하는 프로그래밍 도구, 라이브러리, 모델 모음집

Oracle Cloud 및 NVIDIA 솔루션을 함께 활용해 AI를 가속화하는 방법을 살펴보세요.

> "AI 기반 혁신 기술은 모든 비즈니스의 혁신을 지원하는 무한한 기회를 제공합니다. NVIDIA는 Oracle Cloud Infrastructure(OCI)와의 파트너십을 통해 모든 기업에게 NVIDIA 가속 컴퓨팅 플랫폼의 탁월한 슈퍼컴퓨팅 성능을 제공할 수 있게 되었습니다.
>
> **Justin Boitano**
> NVIDIA, Vice President of Enterprise AI

AI 인프라 시작하기

AI 인프라 추가 리소스
RDMA 클러스터 네트워킹, GPU 인스턴스, 베어메탈 서버 등에 대한 더 많은 정보를 살펴보세요.
AI 인프라 살펴보기

OCI 고객사의 성과 확인하기
확장성, 안전성, 고가용성, 내결함성, 고성능을 제공하는 Oracle의 클라우드 환경에서 애플리케이션을 구축 및 실행 중인 다른 기업들의 사례를 확인해 보세요.
고객 사례 읽어보기

OCI를 통해 얻을 수 있는 비용 절감 효과 확인하기
Oracle Cloud는 간단하고 전 세계에 일관적으로 적용되는 저렴한 가격 정책을 바탕으로 다양한 사용 사례를 지원합니다. 비용 계산기를 사용해 귀사의 요구 사항에 부합하는 서비스를 구성하고 OCI를 통해 얻을 수 있는 비용 절감 효과를 직접 확인해 보세요.
비용 계산하기

AI 인프라 전문가와 상담하기
Oracle 전문가들이 OCI AI 인프라에서의 새로운 AI 솔루션 구축, 워크로드 배포와 같은 다양한 AI 관련 주제에 대한 상담을 제공합니다.
문의하기

문의처

한국오라클 대표번호 02-2194-8000, 또는 oracle.com/kr 웹사이트를 통해 Oracle 담당자에게 연락하실 수 있습니다.

북미 지역 외 국가인 경우 oracle.com/kr/contact 에서 현지 지사를 찾을 수 있습니다.