lablup

lablup VULTR

NVIDIA HGX B200과 함께 도메인 특화 LLM에 날개를 달아보세요

글로벌 무역 기업, Vultr GPU Cloud의 NVIDIA HGX B200 위에서 Backend.AI로 워크로드를 오케스트레이션하다.

Use Case

NVIDIA HGX B200과 함께 도메인 특화 LLM에 날개를 달아보세요

글로벌 무역 기업, Vultr GPU Cloud의 NVIDIA HGX B200 위에서 Backend.AI로 워크로드를 오케스트레이션하다.

"Vultr는 고성능과 신뢰성을 갖춘 클라우드 인프라를 바탕으로 최신의 고사양 NVIDIA와 AMD GPU를 자유롭게 활용할 수 있는 환경을 제공합니다. 전 세계 32개 클라우드 데이터 센터 리전을 통해 기업이 어디서든 AI 애플리케이션을 손쉽게 구축하고 배포하며 확장할 수 있도록 지원하고, 데이터 주권과 컴플라이언스, 유연한 컴퓨트 옵션, 투명한 가격 정책까지 모두 갖추고 있습니다. 저희가 래블업을 선택한 이유는 사용량에 기반한 혁신적인 워크로드 관리 기술과 GPU 활용을 극대화하는 GPU 분할가상화를 구현해낸 뛰어난 기술력을 높이 평가했기 때문입니다."

—Kevin Cochrane, CMO at Vultr

무역 분야의 AI 대전환에 앞서 인프라 준비가 필요합니다.

글로벌 무역 환경은 다양한 국가, 다양한 언어, 복잡한 규제, 비정형 데이터, 방대한 문서 작업 등으로 인해 매우 높은 복잡도를 지니고 있습니다. 이러한 복잡성은 업무 부담을 가중시키고, 업무 부담은 오류로 이어질 가능성이 높으며, 반복적이고 잦은 오류는 전반적인 업무 시스템의 가시성을 저하시키는 원인이 됩니다.

세계무역기구(WTO)는 인공지능이 무역 분야에 도입될 경우, 언어 장벽을 해소하고 규제 및 데이터 검색에 드는 비용을 크게 절감할 수 있다고 밝히고 있습니다. 특히 중소기업에게 AI 솔루션은 국제 무역 진출의 문턱을 낮추는 핵심 도구가 될 수 있습니다. (*1)

각 조직의 환경과 데이터에 특화된 무역 전용 LLM을 도입하면 기업의 경쟁력을 한층 강화할 수 있습니다. AI가 글로벌 무역의 판도를 빠르게 바꾸고 있는 지금, 인프라에 투자하는 기업과 그렇지 않은 기업 간의 격차는 더욱 커질 것입니다.

무역 분야에서 AI를 도입하면 얻을 수 있는 주요 이점:

- 무역 프로세스의 속도 향상과 운영 비용 절감
- 시장 변화와 새로운 리스크에 대한 신속한 대응
- 점점 복잡해지는 규제 환경에서의 컴플라이언스 강화
- 다양한 비정형 문서 양식의 표준화
- 국가별 규제 사항을 한 번에 확인할 수 있는 원스톱 검증

(*1) Source: Trading with intelligence: How AI shapes and is shaped by international trade

기술 조직이 당면한 복잡한 과제들이 있습니다.

많은 무역 기업들이 디지털 인프라와 소프트웨어의 한계로 인해 AI의 잠재력을 충분히 실현하지 못하고 있습니다. 현재의 무역 업무에는 레거시 시스템, 분산된 데이터 사일로, 수작업 중심의 비효율적인 프로세스가 존재합니다. 세계무역기구(WTO) 역시 상호운용 가능한 인프라와 표준화된 프로세스의 부재를 AI 기반 무역 혁신의 주요 장애 요소로 지적합니다.

온프레미스 AI 리소스를 보유하더라도, 고성능 GPU 서버의 부족이나 팀 간 워크로드 공유의 복잡성 등 다양한 기술적 난관이 존재합니다. 글로벌 환경에서 AI를 구축·운영하려면 상당한 시간과 전문성이 필요하며, 자체 GPU 서버를 마련하는 데는 높은 비용이 들고, DevOps 엔지니어 채용도 쉽지 않습니다. 최근에는 국가별 개인정보 보호법 준수까지 요구되면서 기업의 부담이 더욱 높아지고 있습니다. 빠르고, 유연하며, 컴플라이언스까지 충족하는 인프라와 관리 소프트웨어 없이는 진보된 생성형 AI 기술을 성공적으로 도입하기 어렵습니다.

기술 조직이 해결해야 할 주요 과제:

- 복잡하고 시간이 많이 소요되는 인프라 관리
- 필요할 때 충분하지 않은 GPU 자원 확보
- 비효율적인 GPU 자원 활용
- 지역별 규제를 준수하는 GPU 서버 인프라 구축의 어려움

한 눈에 살펴보기

무역 도메인에서의 AI 통합은 산업의 미래를 위해 꼭 필요한 전략입니다. 인간 중심의 업무에서 일정 수준의 자동화와 AI 도입을 통해 생산성을 향상시킬 수 있으며, 이는 수준 높은 하드웨어와 함께 하드웨어의 가능성을 모두 활용할 수 있는 인텔리전트한 소프트웨어와 함께 가능합니다. Vultr와 Backend.AI의 통합은 이러한 고객의 요구를 통합하는 훌륭한 제안입니다.

적용 산업군

AI 소프트웨어

무역 AI의 활용처

문서 자동화 및 분류:
다양한 종류의 무역 문서를 분류하고, 핵심 정보를 추출하여 전반적인 운영 효율성을 개선할 수 있습니다.

규정 준수여부 검토:
수출입 법률, 국가별 규정과 같은 정보들을 바탕으로 문서를 자동으로 검토하고, 잠재적인 오류나 위험을 진단하며, 혹시 모를 인적 오류를 줄일 수 있습니다.

다국어 번역:
LLM은 무역 문서 번역, 파트너와의 커뮤니케이션, 현지 법률 및 규정 해석을 용이하게 하여 글로벌 비즈니스 환경의 장벽을 허물어 줍니다.

래블업과 Vultr에 관하여

래블업의 Backend.AI는 지능형 오케스트레이션을 통해 인프라 관리를 자동화하고, GPU 활용을 최적화하는 플랫폼입니다. Backend.AI는 GPU를 컨테이너 단위로 할당 및 사용할 수 있도록 하여, 대규모 분산 모델 학습부터 추론까지의 모든 워크플로우를 효율적으로 지원합니다. 강력한 파이프라인 관리 기능을 바탕으로 멀티 노드, 멀티 테넌트 환경에서의 배포도 원활하게 처리할 수 있으며, 모든 프로세스를 하나의 소프트웨어 플랫폼에서 안전하게, 동적으로, 동시에 실행할 수 있도록 보장합니다.

Vultr는 32개 글로벌 클라우드 데이터 센터 리전에서 NVIDIA HGX B200 GPU를 즉시 사용할 수 있는 환경을 제공, 고성능 AI 인프라를 손쉽게 구축할 수 있도록 지원합니다. Vultr의 예측 가능한 가격 정책과 ISO 27001, SOC 2 Type II, GDPR 등 다양한 인증 및 컴플라이언스 지원을 통해 기업이 엄격한 요구사항을 충족하는 AI 환경을 신속하게 도입할 수 있습니다.

도메인 특화 대규모 언어모델(LLM) 개발의 어려움

어느 무역 기업이 비정형 문서를 표준화하고 국가별 규제를 준수하기 위해 무역 도메인에 특화된 LLM을 개발하고 있습니다. LLM이 다양한 비정형 데이터를 효과적으로 처리하기 위해서는 도메인 특화 데이터를 활용한 파인튜닝이 필수적입니다. 그러나 실제 무역 도메인에서 그러한 데이터를 충분히 수집·정제하는 데에는 많은 시간과 인적 자원이 소요됩니다. 따라서, 모델의 성능을 확보하고 개발 일정을 단축하기 위해 합성 데이터를 활용한 학습이 필수적입니다. 그러나 합성 데이터를 생성하는 데에는 상당한 시간이 소요되고, 사내에 보유하고 있는 GPU 서버는 이미 모두 사용되고 있어 추가적인 작업 확장이 어려운 상황입니다.

클라우드 기반 GPU 활용으로 한계 극복

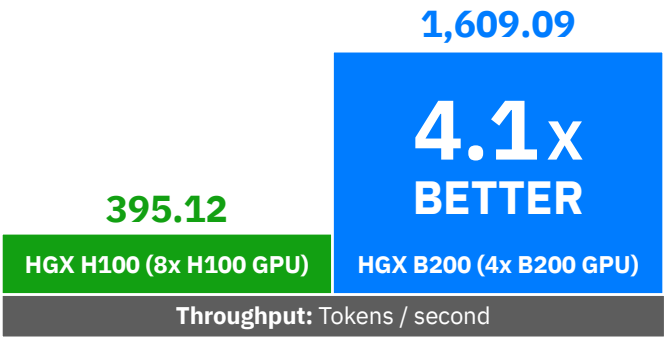
이러한 한계를 해결하기 위해, 기업은 Vultr의 NVIDIA HGX B200 클라우드 GPU를 도입하고, Backend.AI를 활용해 효율적인 GPU 오케스트레이션 환경을 구축했습니다. 또한 Backend.AI FastTrack 2 MLOps를 통해 합성 데이터를 신속하게 생성하고, LLM을 파인튜닝한 뒤 실제 서비스에 배포할 수 있었습니다. 이 과정에서 Backend.AI가 GPU 사용량에 따라 워크로드를 자동으로 분산·관리함으로써 사용자가 자원을 관리할 필요 없이 효율적이고 신뢰성 높은 모델 서비스를 제공합니다.

벤치마크 진행 방법

본 벤치마크는 합성 데이터 구축을 위한 성능 평가를 목적으로 Qwen-3-235B-A22B 모델을 이용하여 진행되었습니다. 팀은 25종의 무역 자격시험(물류관리사, 국제무역사 등)에서 4,000개의 Q&A 세트를 추출하고, 20개의 데이터셋 예시를 기반으로 100개의 질문을 선정했습니다. 벤치마크는 vLLM 0.8.5, CUDA 12.8, NVIDIA 드라이버 570 환경에서 실시되었으며, 비교 대상은 Vultr 클라우드 환경의 NVIDIA HGX B200 GPU 4장과 온프레미스 클러스터의 HGX H100 GPU 8장입니다.

벤치마크: Vultr의 NVIDIA HGX B200과 Backend.AI를 활용한 합성 데이터 생성

래블업과 Vultr는 무역 특화 LLM 구축 과정에서 발생하는 다양한 기술적 과제를 효과적으로 해결할 수 있는 강력한 솔루션을 제공합니다. 두 기업의 역량을 결합함으로써, 조직은 빠르고 안전하며 효율적인 생성형 AI 도입 플랫폼을 구축할 수 있습니다.



Metrics	HGX H100 (8x H100 GPU)	HGX B200 (4x B200 GPU)
처리량 (Throughput) 토큰/초	395.12	1609.09
비용 효율성	단위비용의 1.3배	GPU 개수 ½개
종합 효율성	-	약 6배 이상

주요 결과 및 성과

처리량 대폭 향상:
Vultr Cloud GPU의 NVIDIA HGX B200과 Backend.AI를 활용한 결과, 온프레미스 H100 GPU 대비 4.1배 높은 처리량을 달성했습니다.

데이터의 신속한 생성:
대용량 배치 처리와 빠른 연산이 가능해지면서, LLM 파인튜닝에 필요한 데이터를 신속하게 생성할 수 있었습니다.

인프라 구축 시간의 단축:
클러스터 구축 시간이 베어메탈 환경 대비 1/20 이하로 줄어들었으며, 컨테이너 수준의 GPU 가상화를 통해 GPU 활용도를 극대화하여 시간과 비용을 크게 절감할 수 있었습니다.

적은 자원으로 더 높은 효율 달성:
B200은 절반의 자원만으로도 H100과 동등한 성능을 보여 약 6배의 효율을 실현했습니다. 이 덕분에 불필요한 자원 낭비 없이 GPU를 최대한 활용할 수 있었습니다.

Backend.AI와 Vultr에 대해 자세히 알아보세요.
래블업에 문의하거나 backend.ai를 방문하세요. →

추가 리소스 보기:
Backend.AI - The NVIDIA® DGX™-Ready Software