Runyour Cloud SERVICE INFORMATION

복잡한 AI 인프라 운영,
Runyour Cloud에서 심플하게

Pain Point

GPU는 늘어나는데, 왜 운영은 더 어려워질까요?

여러 대의 GPU 서버를 운영중이지만, 누가 얼마나 쓰는지 실시간으로 파악하기 어렵습니다.

사용 현황 불투명

어떤 구성원은 GPU를 독점하고, 다른 팀은 추가 구매를 요청합니다. 이를 판단할 근거가 없습니다.

불공평한 자원 분배

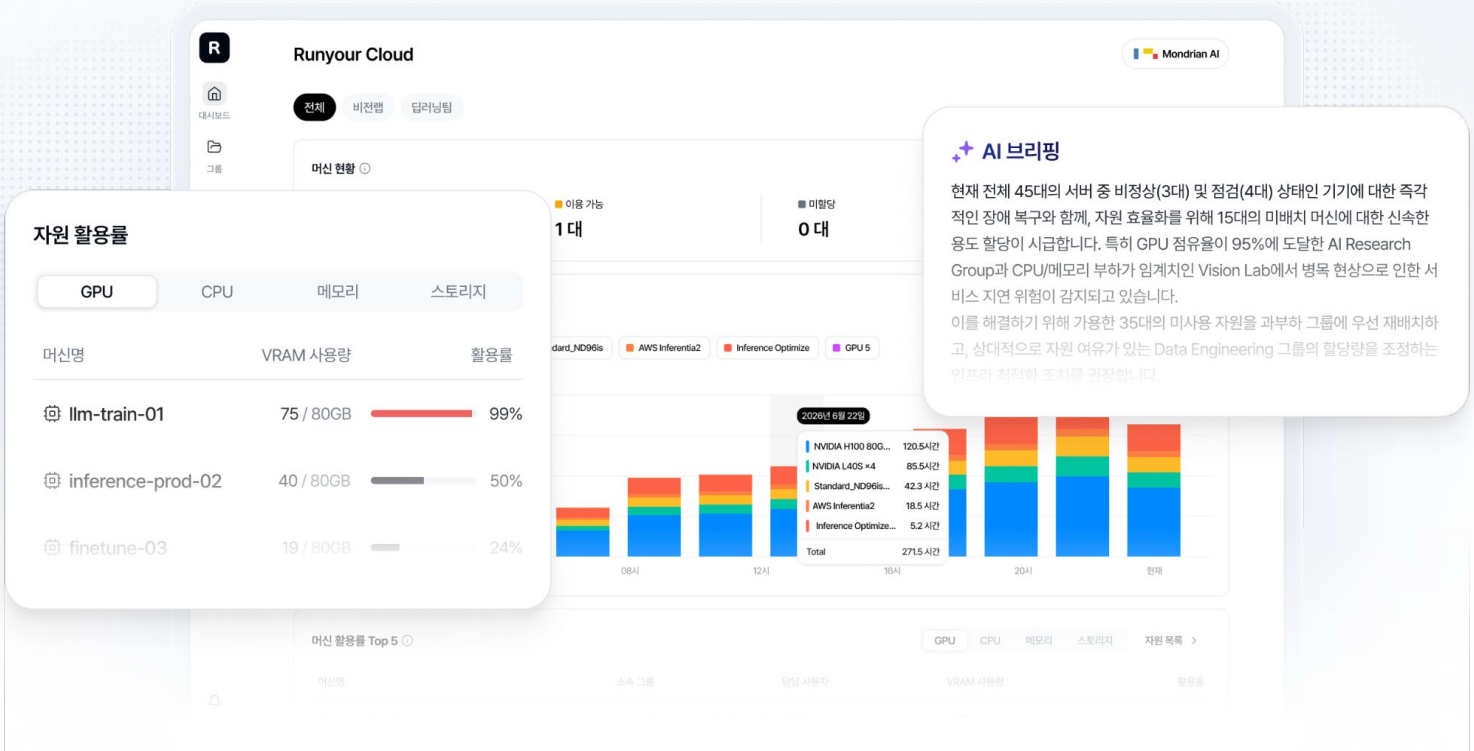
서버가 늘수록 관리는 복잡해지는데, 전담할 전문가가 없어 담당자 혼자 모든 것을 감당해야 합니다.

전담 관리 인력의 부재

보이지 않으면 효율적으로 관리 운영할 수 없습니다

Solution

Runyour Cloud는 인프라 관리부터 운영까지 모든 과정을 지원합니다.



How it works

자원 등록부터 조직 관리, 실시간 모니터링, 개발환경까지 하나의 플랫폼에서

사용 시나리오

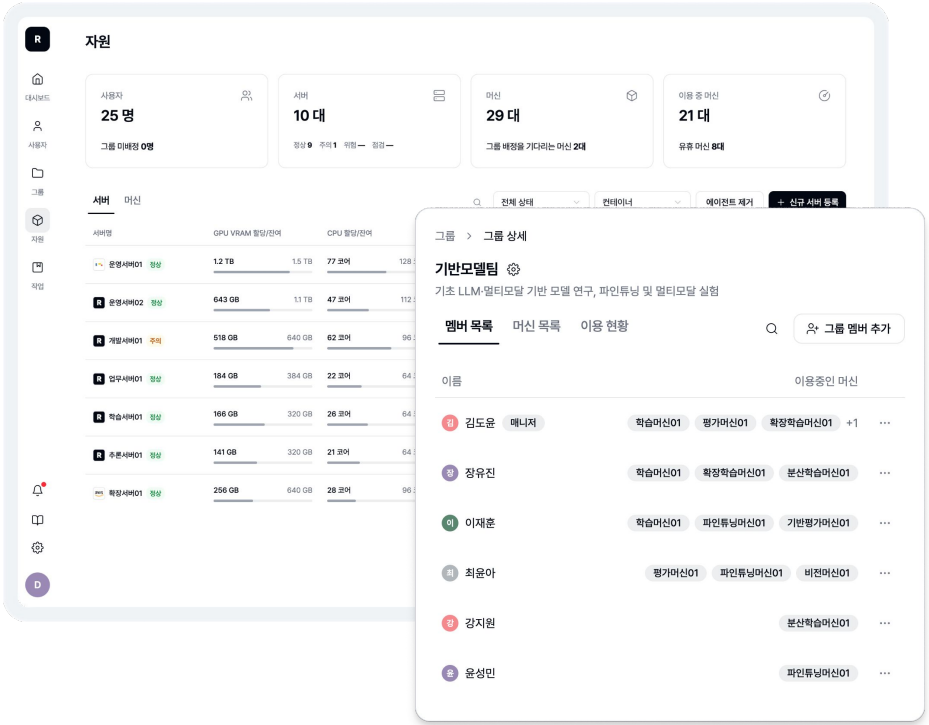
자원 등록부터 회수까지 자동화된 운영 흐름

통합 자원 등록	물리 및 가상 자원을 기관 시스템에 등록하여 통합 관리 환경을 구성	↓
그룹, 자원 배치	등록 또는 확보한 자원을 목적에 맞게 그룹 및 구성원에 배치	↓
머신 분할 실행	자원 유형에 따라 컨테이너(1:N) 또는 베어메탈(1:1)로 머신 생성 및 실행	↓
실시간 미터링	그룹 및 구성원 단위로 사용량 실시간 집계	↓
자동 회수 정책	이용 종료 시 자원 자동 회수 정책으로 유휴 자원 제로화	↓

핵심 기능 1

조직 · 자원 관리부터 실시간 모니터링까지

구성원과 그룹, 자원 배분을 한 콘솔에서 관리하고, 자원 상태를 실시간으로 모니터링하세요.



구성원 · 그룹 관리

구성원 현황과 권한을 조회하고 팀 · 학과 · 부서 단위로 그룹 구성

간편한 자원 배치 및 회수

사용 현황에 따라 자원을 원클릭 배분하고 만료 · 유휴 자원은 회수

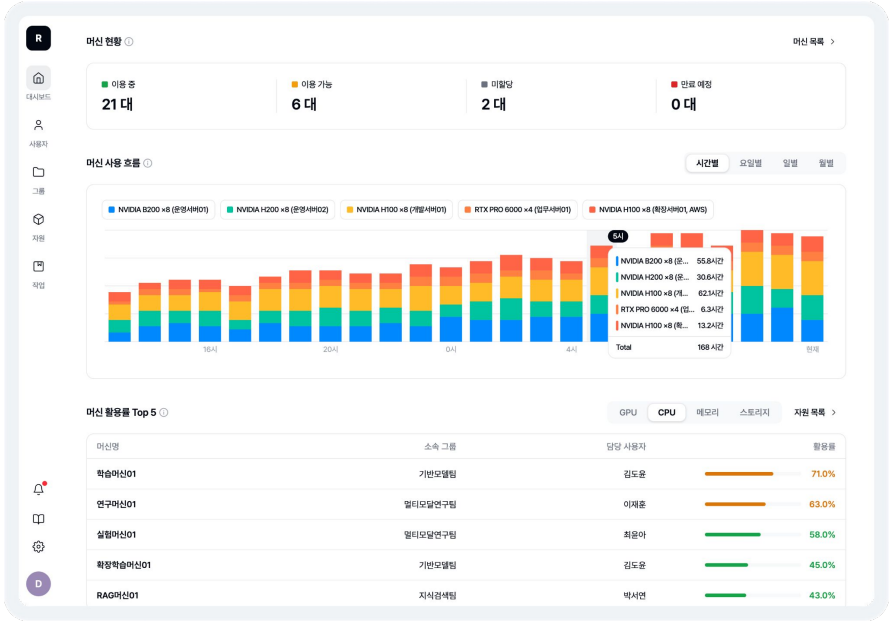
전체 자원 모니터링

GPU · CPU · 메모리 사용률과 프로세스를 실시간 추적하고 이상 징후 파악 및 알림

핵심 기능 2

실시간 미터링으로,
사용량과 비용을 투명하게

그룹별, 구성원별 자원 사용량을 실시간으로 수집하고 자원 배분과 비용 정산의 데이터 기반을 마련하세요.



그룹 · 개인별 사용량 집계

GPU · CPU · 메모리 사용량을 그룹 · 구성원 단위로 누적 집계

기간별 사용량과 추이 시각화

누적 사용량, 활용률 추이 등을 쉽게 조회 및 관리

빌링 · 차지백 연동

집계된 사용량을 비용 산정 · 차지백의 근거 데이터로 연결

핵심 기능 3

잡 스케줄링 기능을 통해 유휴 자원 제로화

작업을 우선순서에 따라 자원에 배치하고, 유휴 자원 없이 24시간동안 가동하세요. 매번 수동으로 자원을 가동할 필요가 없습니다.

자원 유휴 시간 제로화

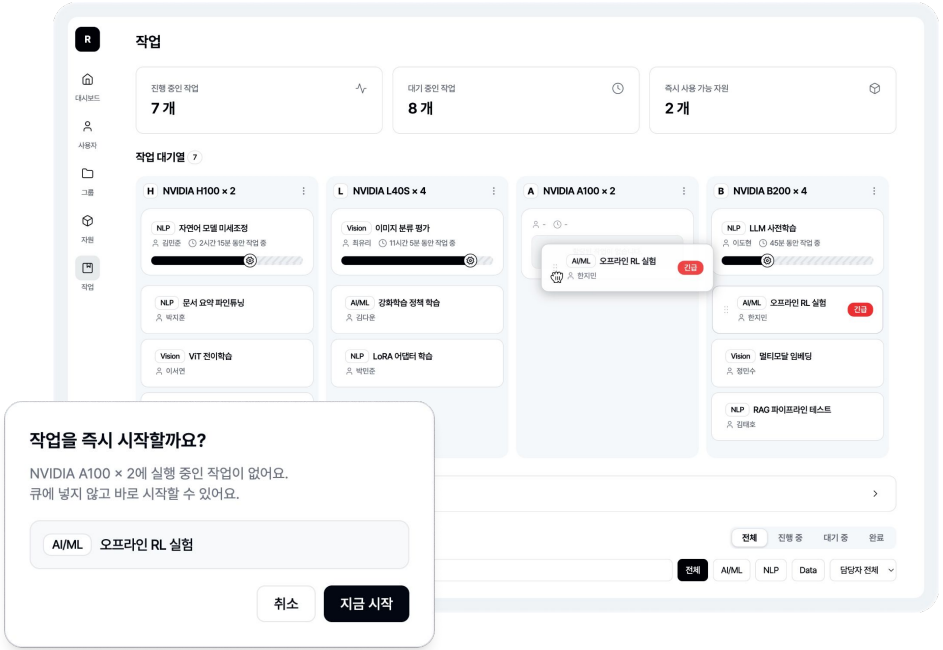
작업을 대기열에 배치해 값비싼 컴퓨팅 자원의 투자 효율을 극대화

공정하고 예측 가능한 배분

우선순위 대기열로 자원 독점을 막고 대기 순번과 현황을 투명하게 공유

운영 부담 제거

수동 배정 없이 자동 스케줄링되어 더 많은 작업을 더 빠르게 처리



핵심 기능 4

환경 설정 없이,
즉시 시작하는 개발환경

복잡한 환경 설정 없이, 템플릿으로 즉시 개발하며 필요한 개발 환경은 직접 구성하여 AI 워크로드를 바로 시작하세요.

기본 템플릿 제공

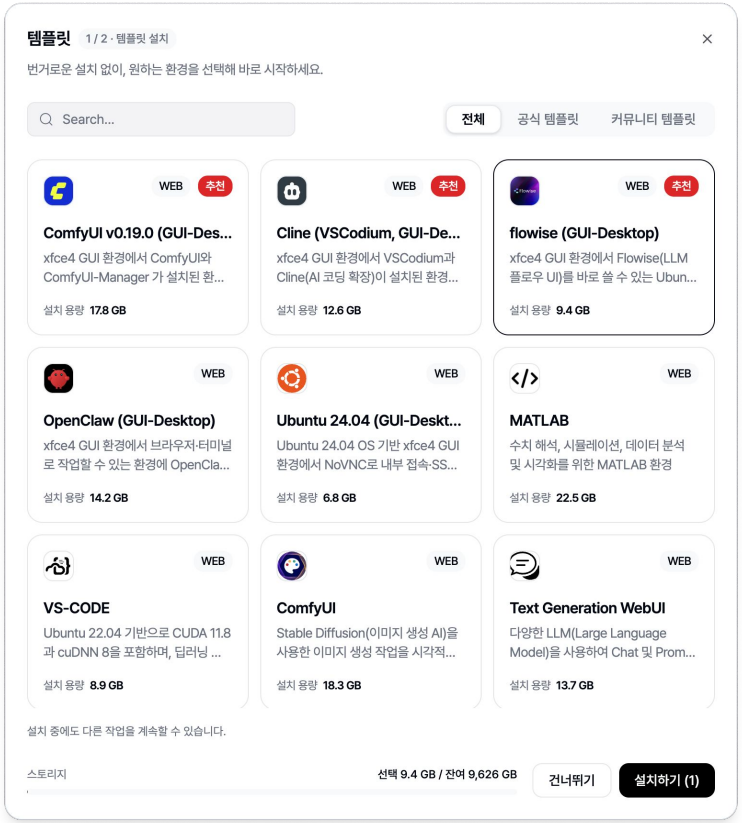
PyTorch, Jupyter 등 표준 AI 개발환경 즉시 실행

맞춤형 환경 구축

필요한 패키지, 설정으로 커스텀 환경 직접 구성 가능

WEB, SSH 즉시 접속

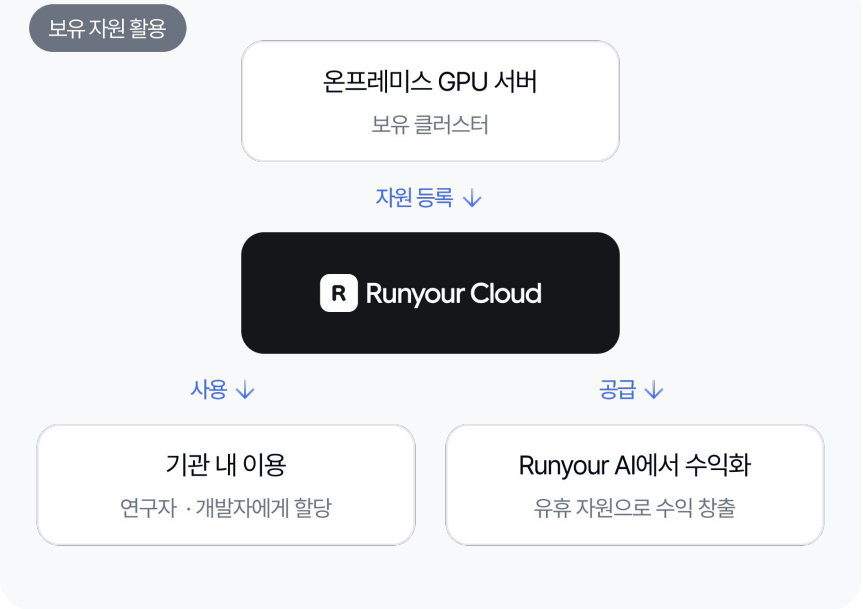
브라우저 Web IDE 또는 SSH로 바로 접속

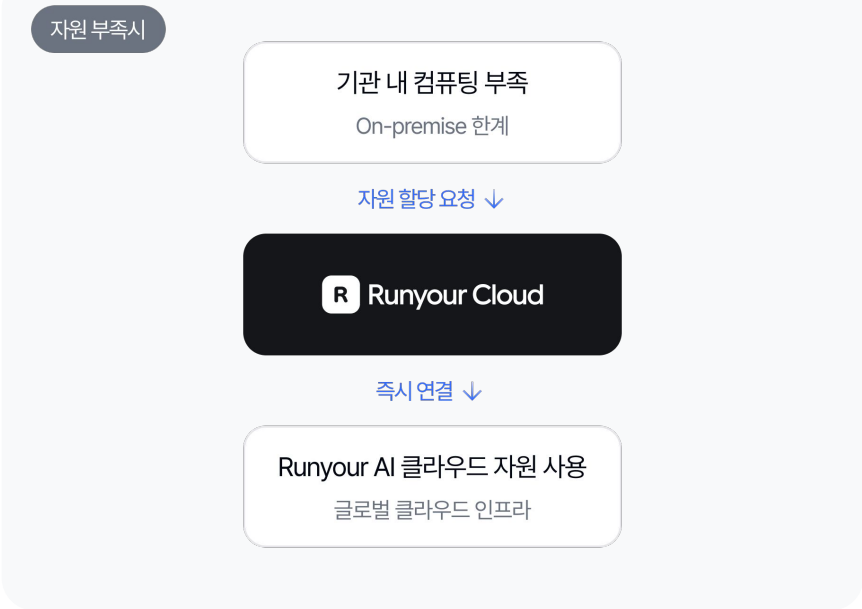


핵심 기능 5

온프레미스와 클라우드를 하나로

등록한 GPU 서버를 기관 내에서 이용하거나 Runyour AI에 공급해
수익화하고, 부족한 자원은 클라우드로 즉시 확장하세요.





핵심 사용자

우리 회사에 꼭 필요한 Private Cloud
지금 바로 시작하세요

민간기업

부서별 개발팀에 연구 자원을 유연하게 할당하고,
프로젝트 단위로 비용과 사용량을 정밀하게
정산하고자 하는 기업

공공기관

보안이 강화된 폐쇄망 내 GPU 인프라를 구축하고
여러 협력 기관에 자원을 안정적으로 분배하고자
하는 공공기관(바이오, 교통 등 데이터 활용 허브)

대학 연구기관

학부생, 대학원생, 연구실 별로 연구 자원을
공정하게 배분하고 관리 효율을 극대화하고자 하는
교육 기관

Use Case. 1

완벽한 통제력을 갖춘 프라이빗 클라우드

LG AI Research

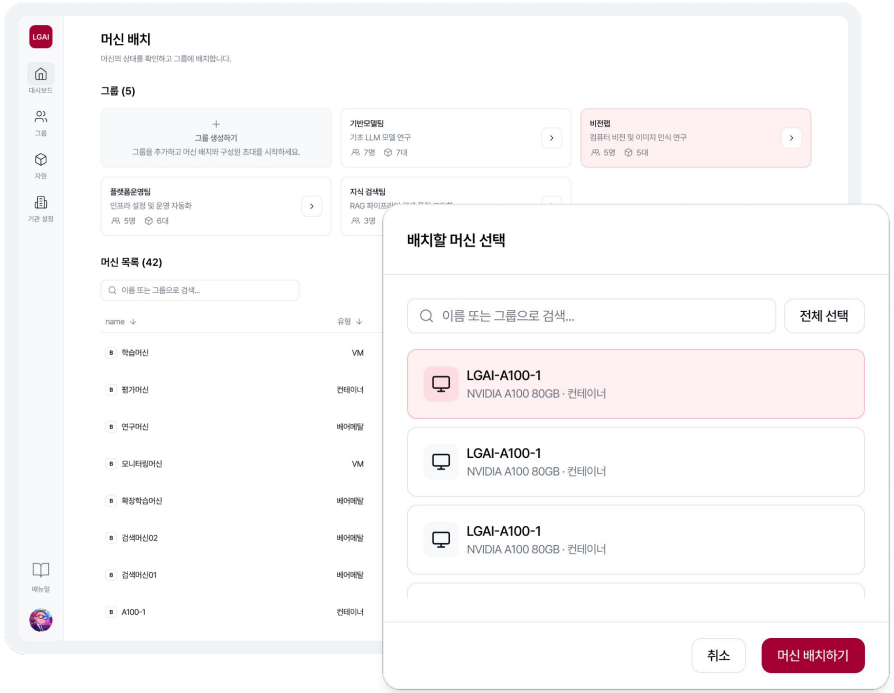
도입 배경

- 부서별 개발팀의 GPU 자원 파편화
- 프로젝트별 정밀한 비용 정산 불가, 벤더 종속적인 인프라 관리

솔루션

SaaS 형

- 고객사 전용 화이트 라벨링: LG AI 연구원 로고와 도메인이 완벽히 적용된 로그인 페이지 및 대시보드 구축
- 독립적 거버넌스 확보: 내부 관리자(Admin)가 직접 유저별 자원 할당량 및 스토리지를 실시간으로 통제하고 회수



결과

운영 주도권 확보를 위한 비즈니스 연속성 제공 및 획기적인 비용 절감

Use Case. 2

보안망 기반의 AI 인프라 운영 관리 플랫폼

금융감독원

도입 배경

- 가상 자산 시장 조작 및 금융 사기 탐지 AI 모델 개발을 위한 AI 인프라 운영 플랫폼 필요
- 외부망과 철저히 분리된 폐쇄망 내 인프라 구축 필수

솔루션

설치형 License

- 국내 최고 수준의 보안이 적용된 폐쇄망 내 인프라 통합 관리 체계 구축
- 실시간 대규모 금융 데이터 수집 · 정제 · 분석을 위한 컨테이너 기반 고성능 서버 인프라 운영 플랫폼 제공

결과

머신러닝 기반 사기 탐지 모델의 효율성 극대화 및 금융시장 신뢰도와 안정성 제고

Use Case. 3

캠퍼스 통합 GPU 팜

고려대학교
KOREA UNIVERSITY

도입 배경

- 교내 수십개의 연구실 및 외부 파트너 대상으로 중앙 집중식 GPU 서버 관리 필요성 대두
- 더불어, 연구실은 각기 다른 연구 환경을 구축하는데 많은 시간이 소요되어 연구 효율 저하

솔루션

설치형 License

- H100 94GB 및 NVL 자원의 중앙 관리
- SSO 인증 연동 및 로그 정밀 제어
- 교내 보안 및 사용량 기반 빌링 시스템 도입

결과

'향후 10년 AI 주도 대학' 성장을 위한 최적화된 교내 AI 자산 인프라 완성

인프라 확장성

필요한 GPU 자원,
Runyour Cloud에서 바로 이용하세요

글로벌 네트워크 Global Region 95개국 Data Center Tier 3

NAVER Cloud aws NHN CLOUD Azure CORE Google Cloud

최신 GPU 서버

- RTX Series 엔트리
- A100 고성능
- H100 AI 학습
- H200 최신 세대
- B200 차세대
- B300 최상위
- L40s 추론 특화
- A5000 워크스테이션

GPU 인프라, 이제 Runyour Cloud로 운영 방식을 바꾸세요

인천 본사
인천광역시 연수구 송도과학로 16번길 13-46, 701호, 702호

서울 지사
서울특별시 서초구 서초대로 398, 그레이츠강남 4층

Web. www.cloud.runyour.ai
Email. runyour@mondrian.ai
Tel. 032-459-2239
Fax. 032-232-1104