HS 효성인포메이션시스템 | Hitachi Vantara

[AI WAVE 2025]

AX와 지속 가능성이 가져올 미래 혁신

권동수 전문위원
HS효성인포메이션시스템
2025년 5월 28일

Agenda

1. AI 트렌드
2. 생성형 AI를 위한 고려사항
3. AI 구축 방안 및 사례

1. AI 트렌드

1. DX에서 AX로의 확산
2. ANI(약인공지능), AGI(일반인공지능) and ASI(초인공지능)
3. 지속가능한 미래를 위한 AI

DX에서 AX로의 확산

AX (AI Transformation, 인공지능 전환)는 새로운 개념이 아닌 DX (Digital Transformation, 디지털 전환)의 진화
데이터 디지털화의 전반 (DX) → DX에 AI를 활용하여 효율성 증대(AX)

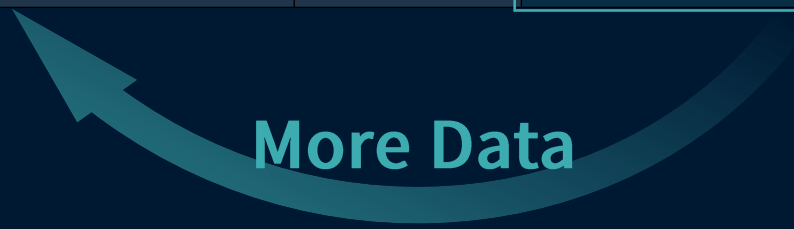
DX

- 기술 도입
- 비즈니스 모델 혁신
- 운영 효율화
- 고객 경험 개선

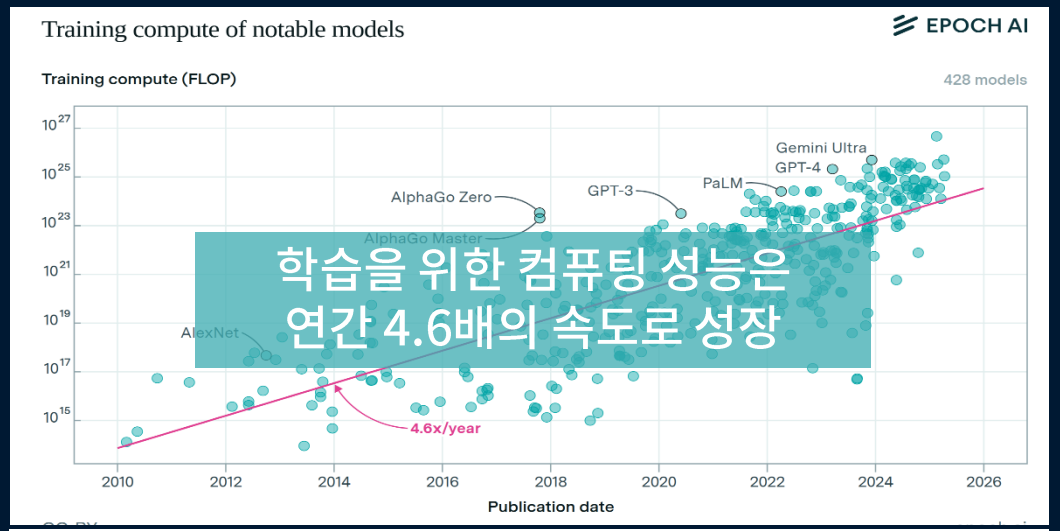
Digital Transformation	항목	AI Transformation
디지털 기술 활용을 통한 전반적 혁신	주요 초점	AI 기술 활용을 통한 심화된 최적화와 자동화
프로세스, 운영, 고객 경험의 디지털화	적용 범위	데이터 중심의 AI 활용과 지능화된 시스템 구축
클라우드, IoT, 빅데이터 등	핵심 기술	머신러닝, 딥러닝, RPA , LLM등

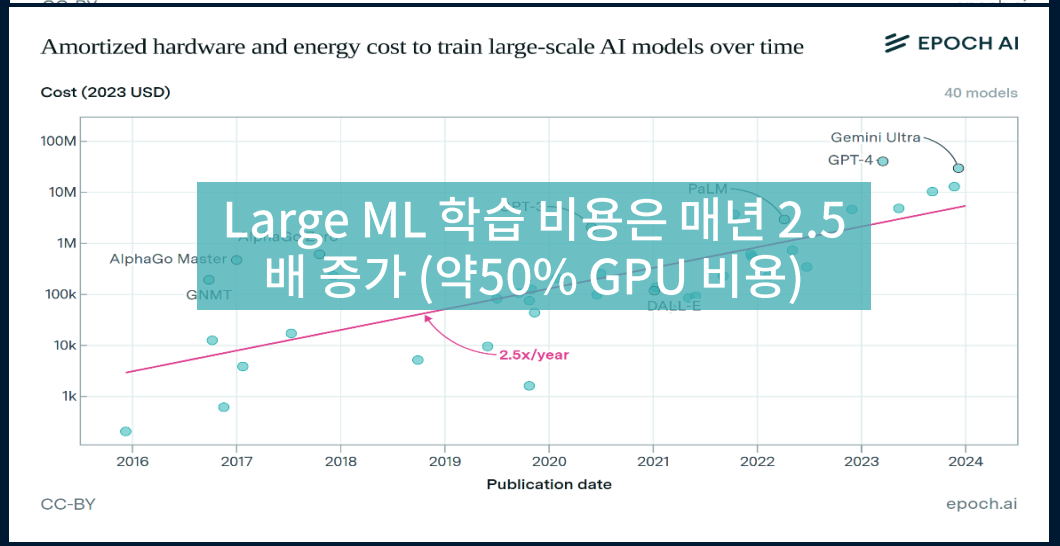
AX

- AI 기반 의사결정
- 자동화 및 최적화
- 개인화 경험
- 예측 분석



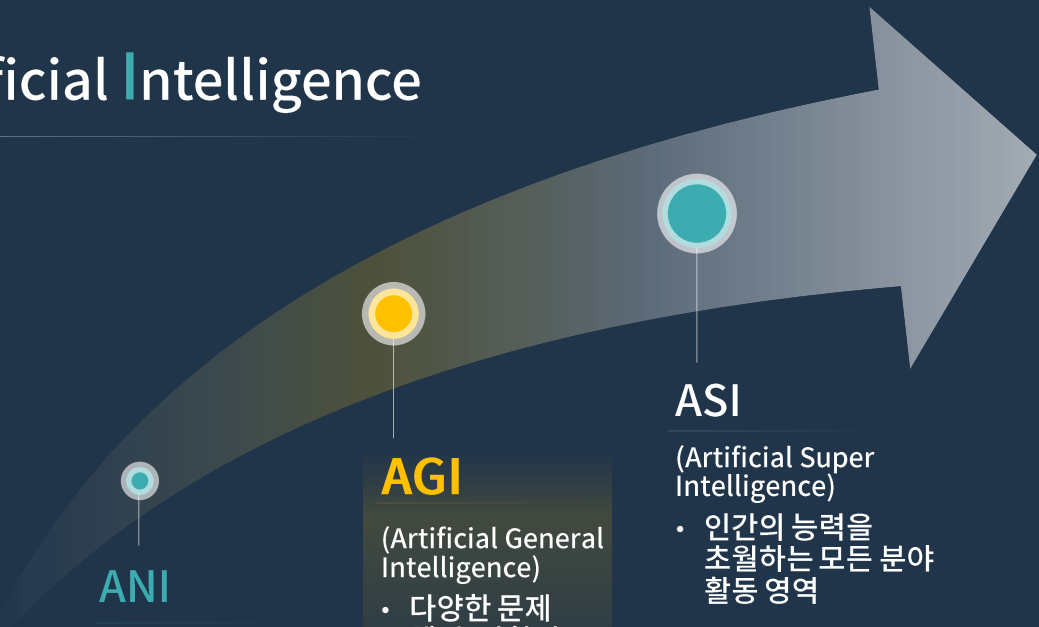
ANI(약인공지능), AGI(일반인공지능) and ASI(초인공지능)





AI

Artificial Intelligence



ANI

(Artificial Narrow Intelligence)

- 음성 인식, 이미지 처리, 자연어 처리, 추천 시스템, 자율주행 등

AGI

(Artificial General Intelligence)

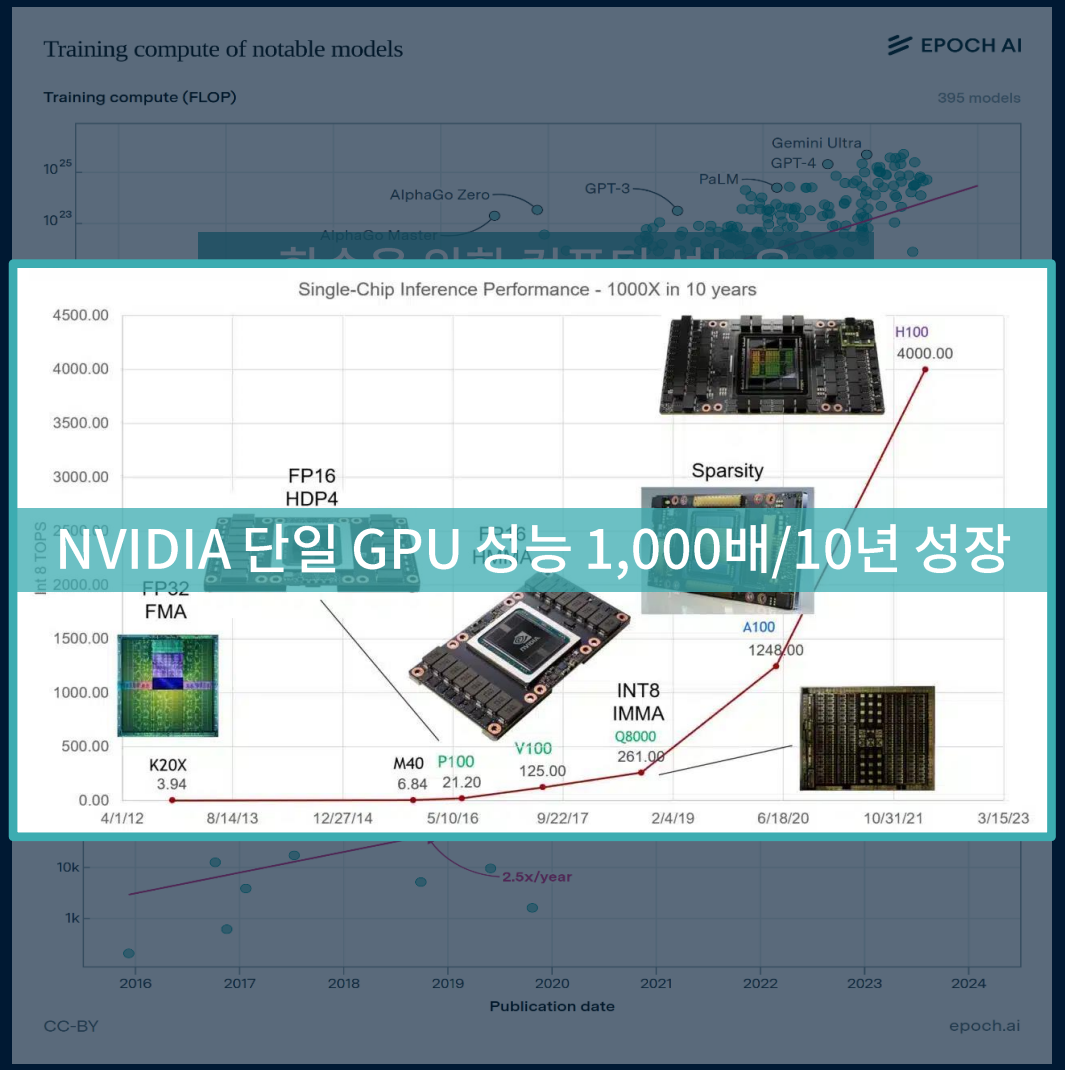
- 다양한 문제 해결, 과학적 발견, 창의적 작업 등

ASI

(Artificial Super Intelligence)

- 인간의 능력을 초월하는 모든 분야 활동 영역

ANI(약인공지능), AGI(일반인공지능) and ASI(초인공지능)



NVIDIA 단일 GPU 성능 1,000배/10년 성장

AI

Artificial Intelligence

ANI

(Artificial Narrow Intelligence)

- 음성 인식, 이미지 처리, 자연어 처리, 추천 시스템, 자율주행 등

AGI

(Artificial General Intelligence)

- 다양한 문제 해결, 과학적 발견, 창의적 작업 등

ASI

(Artificial Super Intelligence)

- 인간의 능력을 초월하는 모든 분야 활동 영역

지속가능한 미래를 위한 AI

멀티모달 AI

텍스트, 이미지, 음성 등
다양한 데이터를
동시에 처리

AI Agent

개인화된 AI Agent를
통한 단순 작업 자동화에서
복잡한 업무까지 수행

물리 AI

물리적 법칙 + 데이터 기반
학습을 통해 실제 현상을
보다 정확히 예측

2. 생성형 AI 를 위한 고려사항

1. 생성형 AI를 위한 고려사항
2. 생성형 AI 도입 전략
3. AX를 위한 HS효성 포트폴리오

생성형 AI를 위한 고려사항

전력이슈	고성능 연산 자원	데이터 활용	데이터과학자/ AI전문가
고성능 인프라 사용에 따른 대량 전력 소비	수요 대비 공급 부족 가격부담	출처/보안/품질등 고객사 고려사항증가	고객사 Domain 전문 인력 확보
ARM / Liquid Cooling 등 전력 감소를 위한 솔루션	유연성과 확장성을 위한 Hybrid GPU 구성	기업 내 데이터 활용 및 외부 데이터 연계 관리 필요	기업에서 요구되는 needs와 의미 있는 insight 도출 전문가

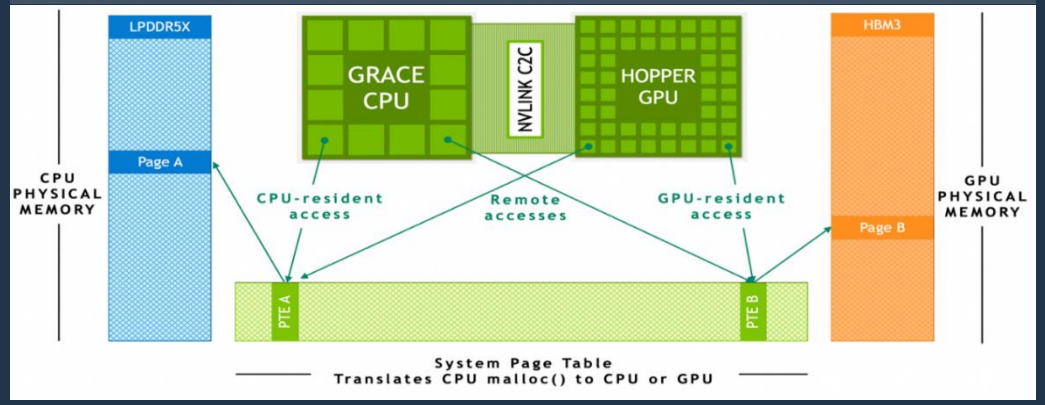
전력 이슈 해결 방안

생성형 AI = GPU = 탄소 배출

학습 및 추론을 위해 많은 전력 사용!

ARM 기반 서버

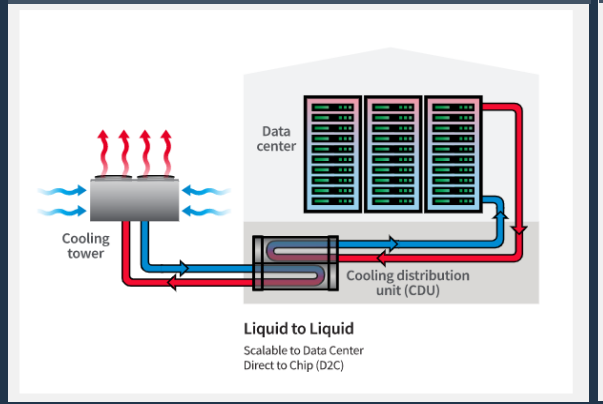
NVIDIA GraceHopper SuperChip



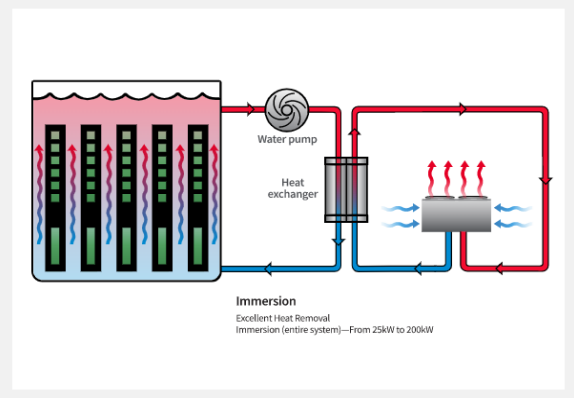
- NVLink-C2C 하드웨어 일관성을 통해 GPU와 CPU가 마이그레이션 없이 서로의 메모리에 엑세스 가능
- ARM Neoverse 아키텍처를 통해 고성능 제공 동시에 높은 에너지 효율성 제공, 지속 가능한 컴퓨팅 솔루션 구현

액체 냉각(Liquid Cooling)

Direct Liquid Cooling 시스템



Immersion Cooling 시스템



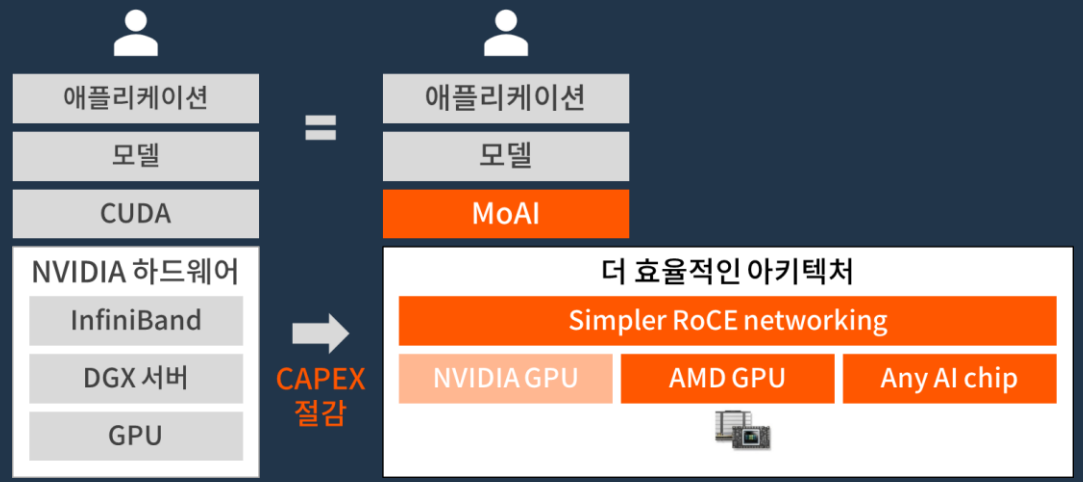
- 기존 Air Cooling 방식 외 Liquid Cooling 방식의 needs 증가 (비용 절감의 효과 및 친환경)
- SMC의 자체 생산 시스템과 기술력, 원스탑 Total IT 솔루션 기반 최적의 DLC 환경 구축 제공

고성능 연산 자원 구성 및 활용 방안

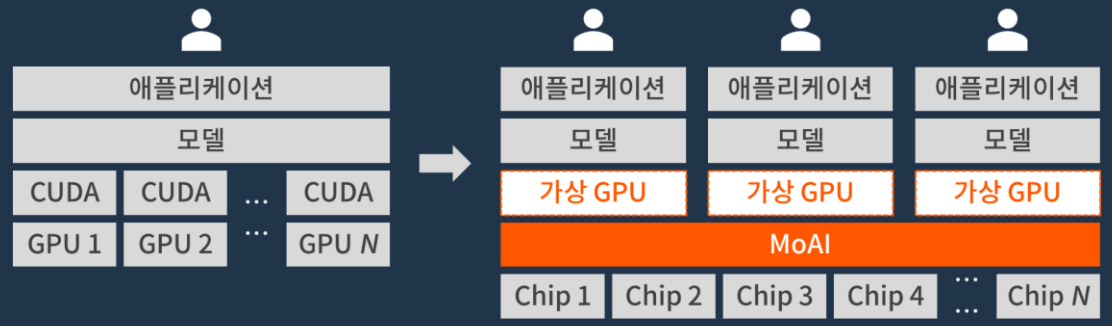
GPU 가상화 : **모레 MoAI** + GPU

비용/성능 효율적인 구축!

PyTorch와 TensorFlow를 다양한 GPU에서 코드 변경 없이 실행



인프라의 복잡성을 숨기고 가상화하여 제공

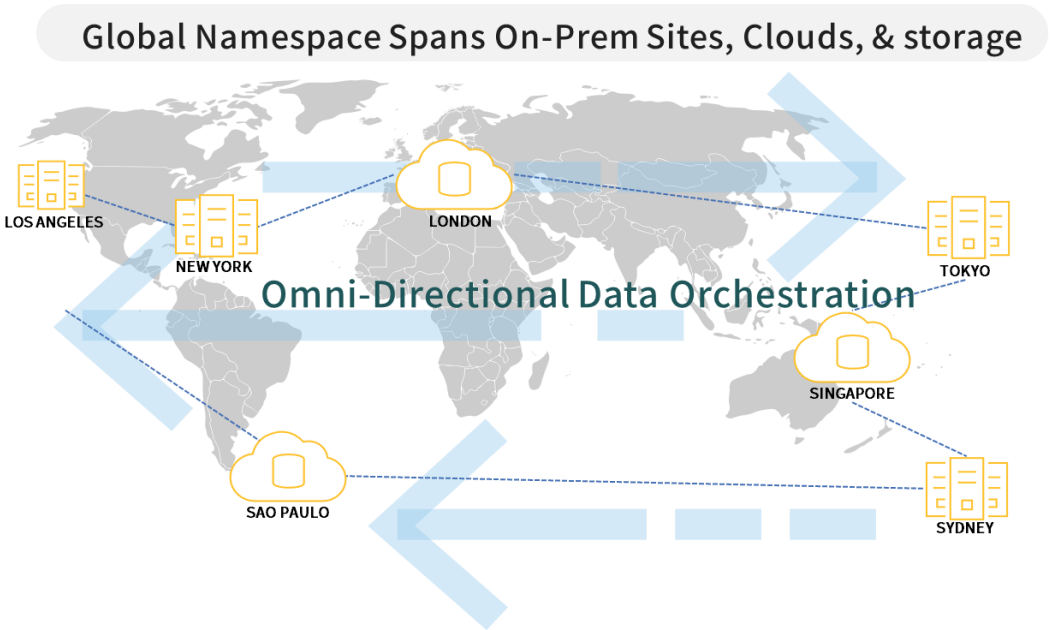


데이터 오케스트레이션, AI 데이터 통합

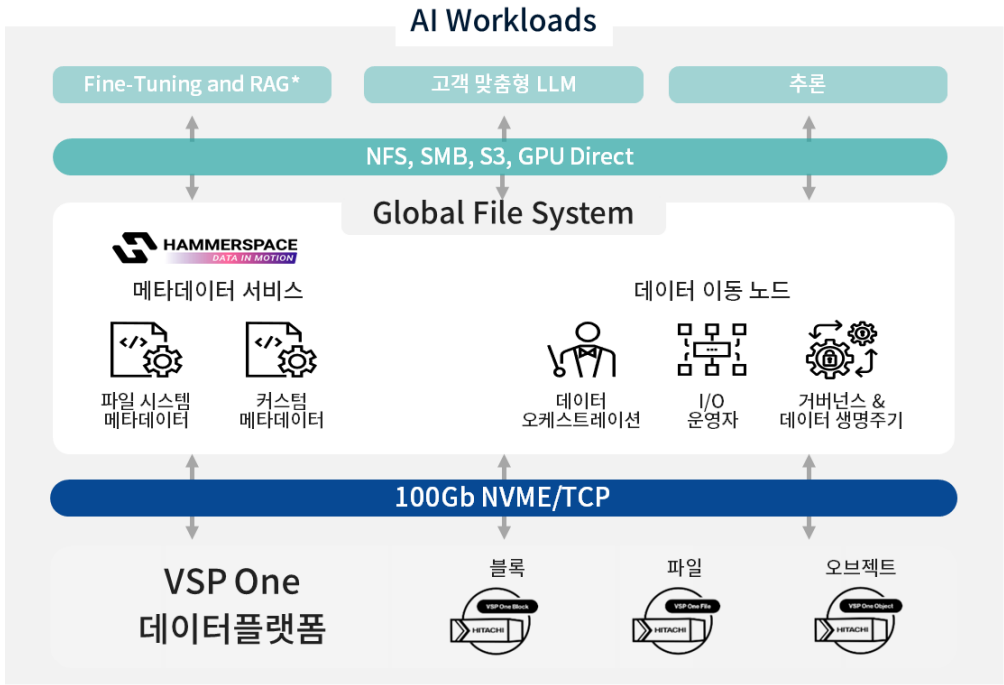
AI 발전으로 데이터 중요성 증가!!!

데이터 통합 관리를 위한 글로벌파일시스템 **해머스페이스**

데이터 오케스트레이션:위치 제약 없는 데이터 운영



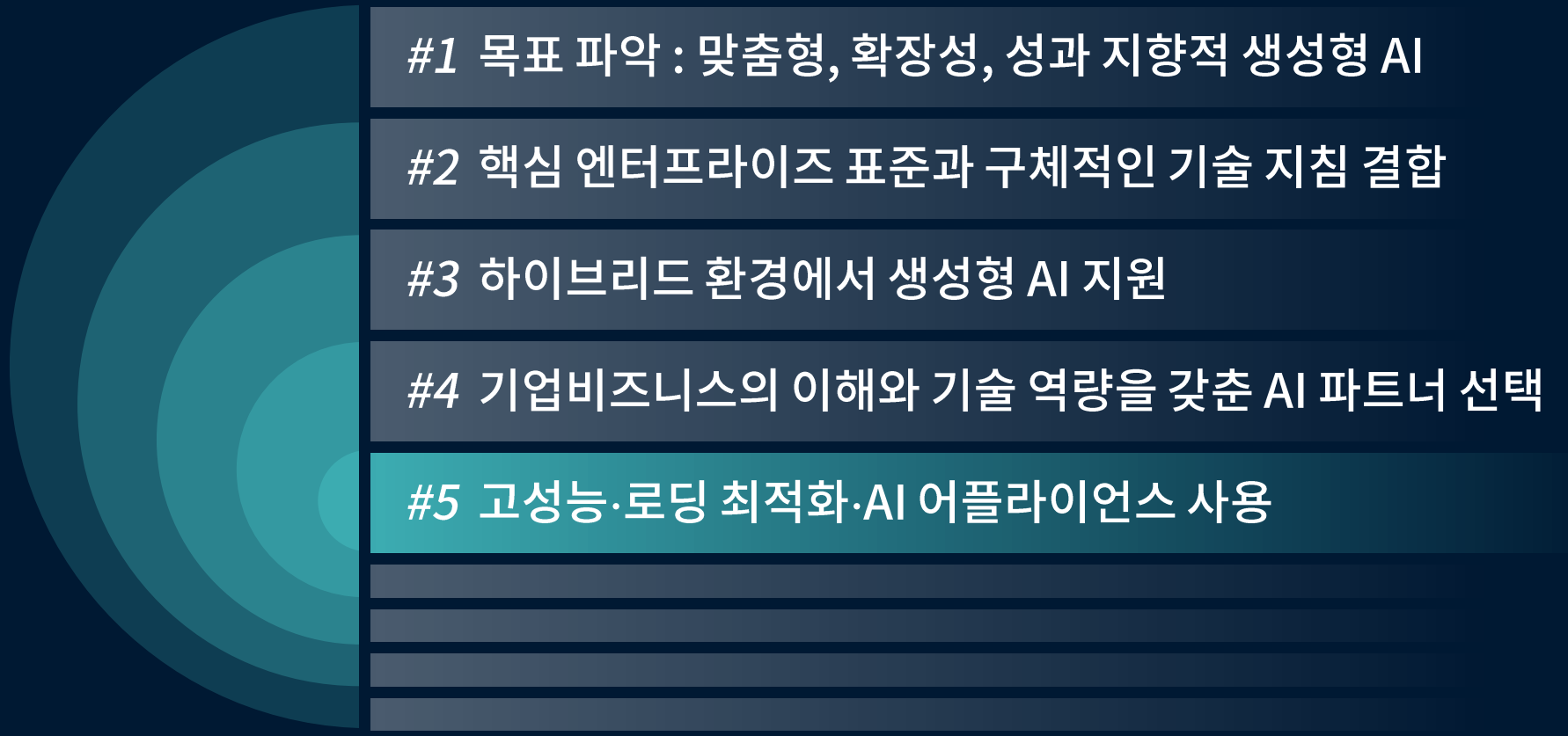
글로벌파일시스템:데이터 통합 관리



※ RAG (Retrieval Augmented Generation) : 검색증강생성

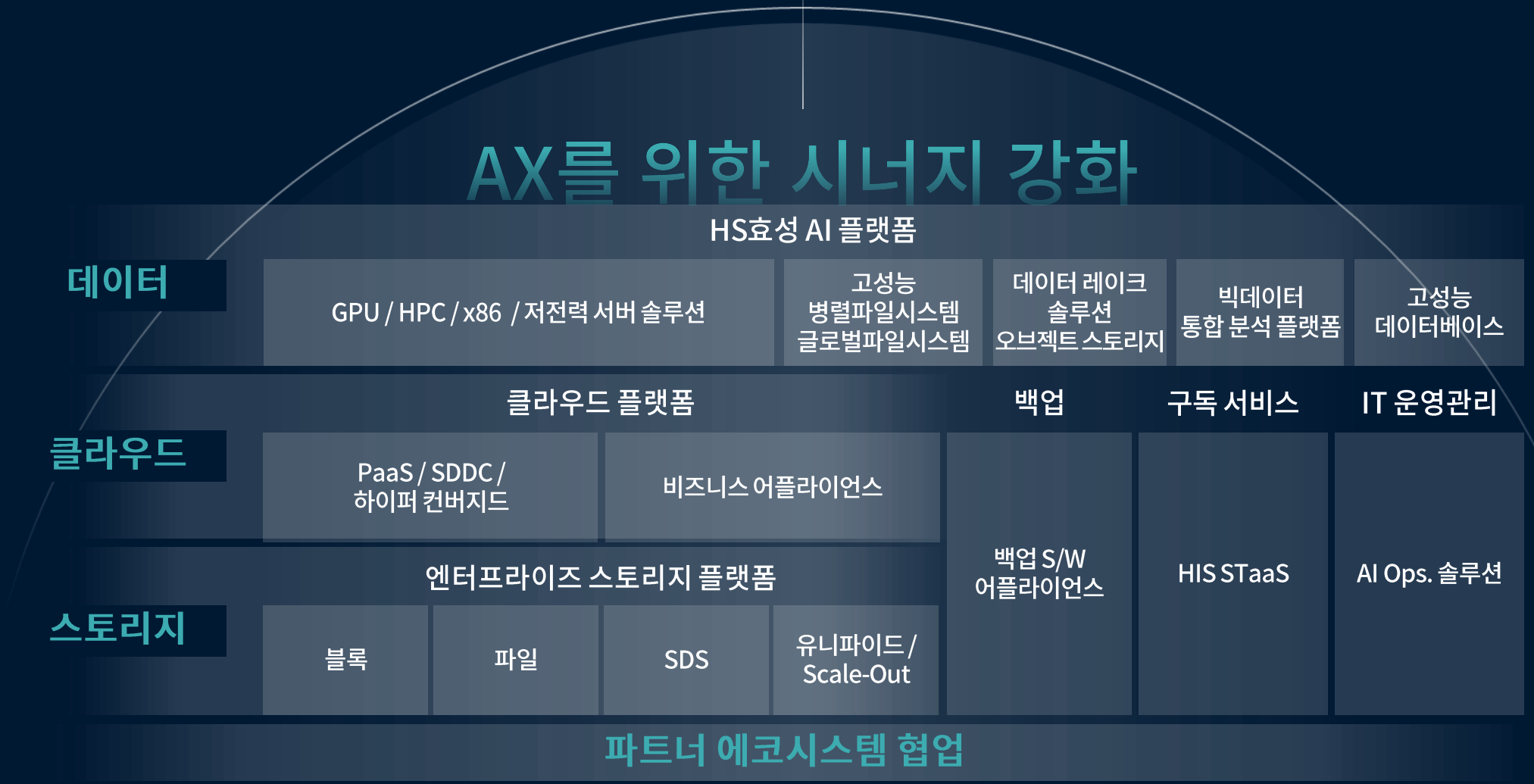
생성형 AI 도입 전략

비즈니스에 성공적인 생성형 AI를 도입하고 궁극적으로 지속가능 성장을 위해 기업이 고려할 5가지 중요한 사항



[출처] Five actions enterprises can take to accelerate GenAI adoption and improve business outcomes (Hitachi Vantara)

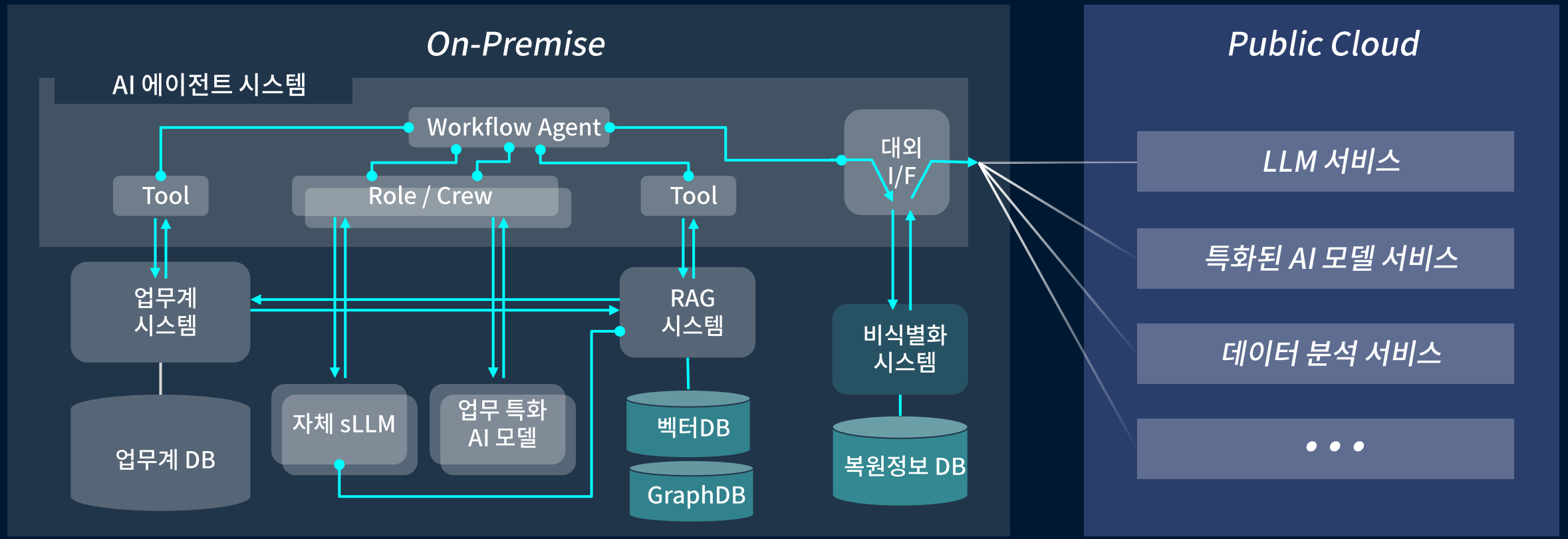
AX를 위한 HS효성 포트폴리오



3. AI 구축 방안 및 사례

1. AI 확장형 하이브리드 구성
2. HS효성 통합 AI 플랫폼
3. 확장형 구축 사례
4. AI 구축을 위한 체크포인트

AI 확장형 하이브리드 구성



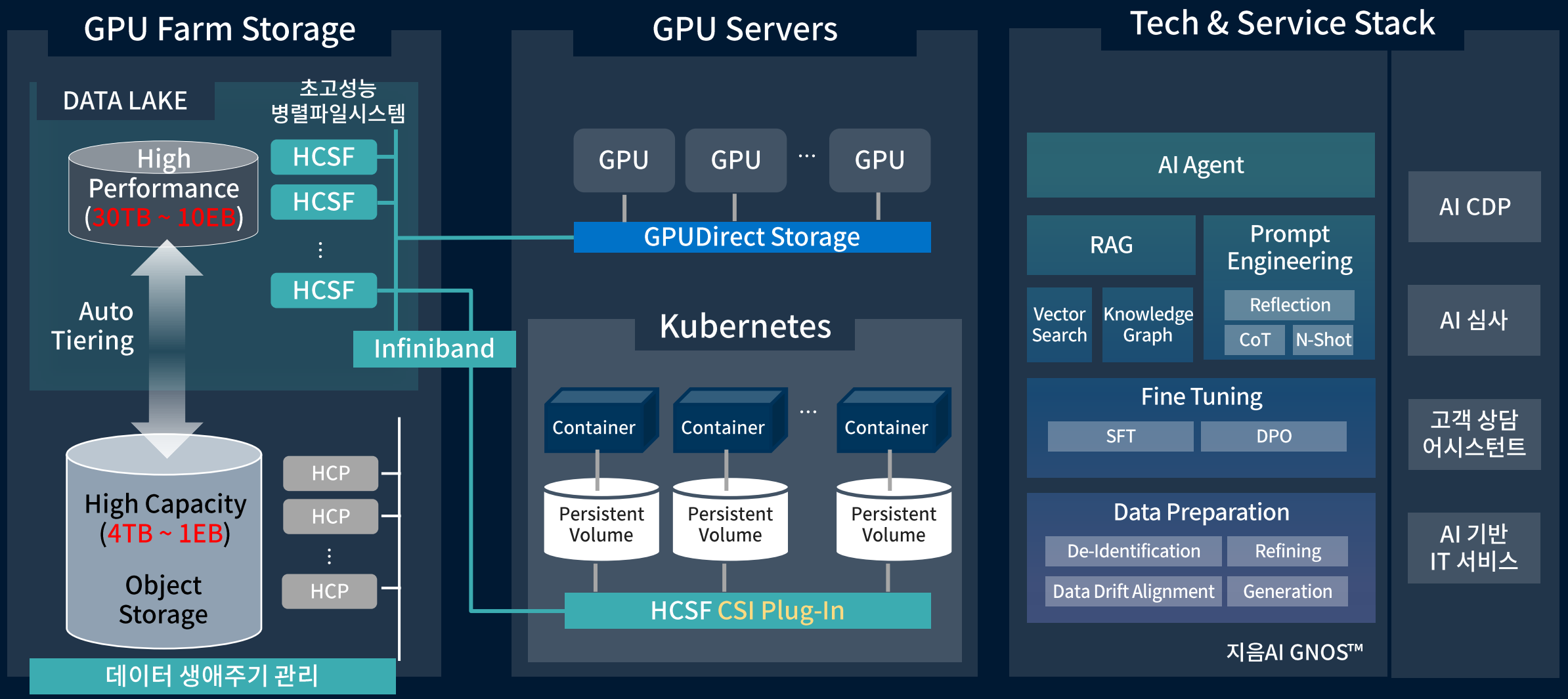
On-Premise Usage

- *개인정보 직접활용 업무 (초개인화)*
- *업무 노하우 기반 처리 업무*
- *대외비 지식 정보 활용 RAG 처리 업무*

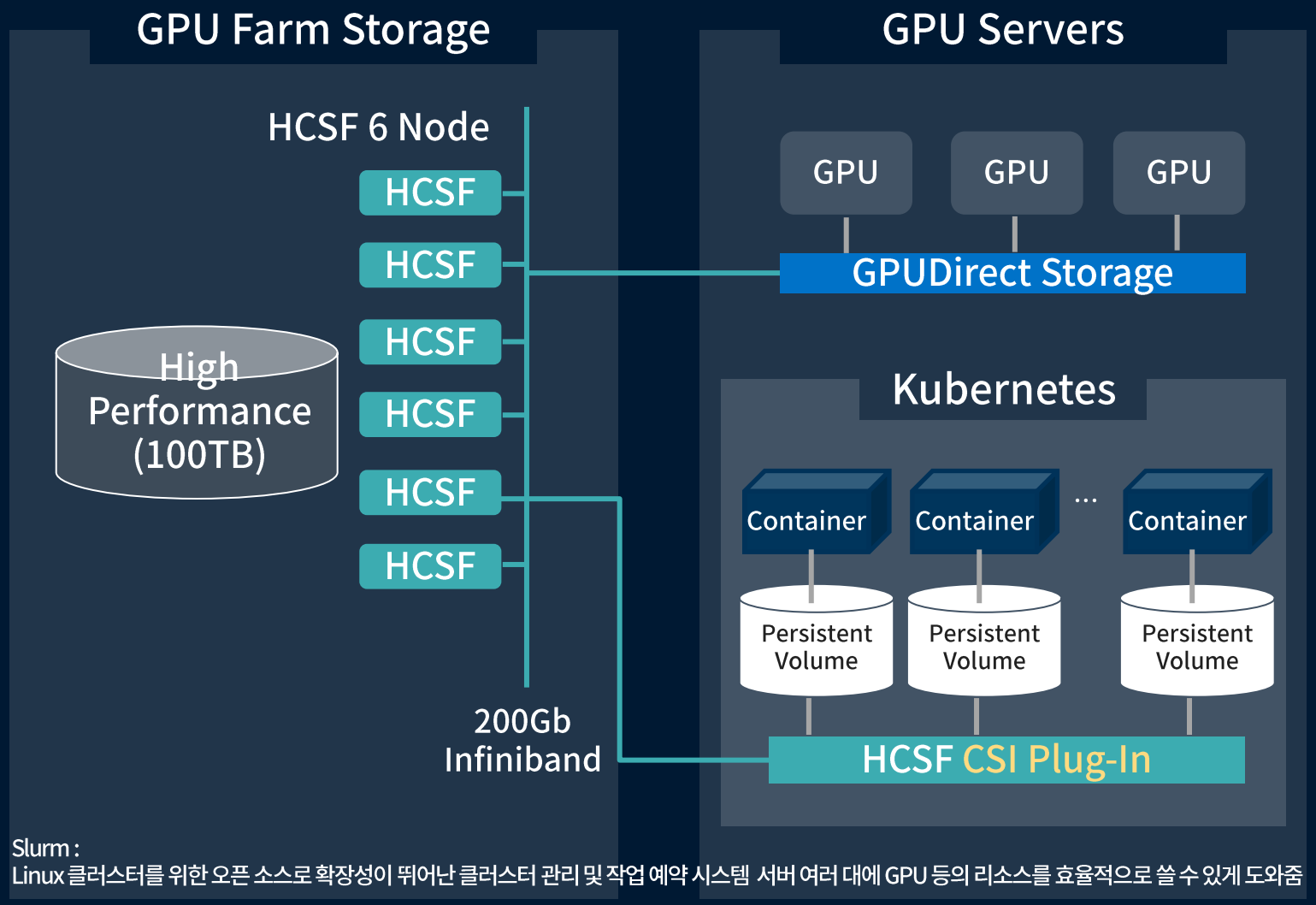
Public Cloud Usage

- *알려진 분석 기법 서비스*
- *공개된 LLM 기반 서비스*
- *가명정보 활용 서비스*

HS효성 통합 AI 플랫폼



사례 1. 국내 그룹 SI 기업 (100TB)



Slurm :
Linux 클러스터를 위한 오픈 소스로 확장성이 뛰어난 클러스터 관리 및 작업 예약 시스템 서버 여러 대에 GPU 등의 리소스를 효율적으로 쓸 수 있게 도와줌

사업 목적

- sLLM 모델 개발 업무를 위한 AI/HPC 인프라 구축

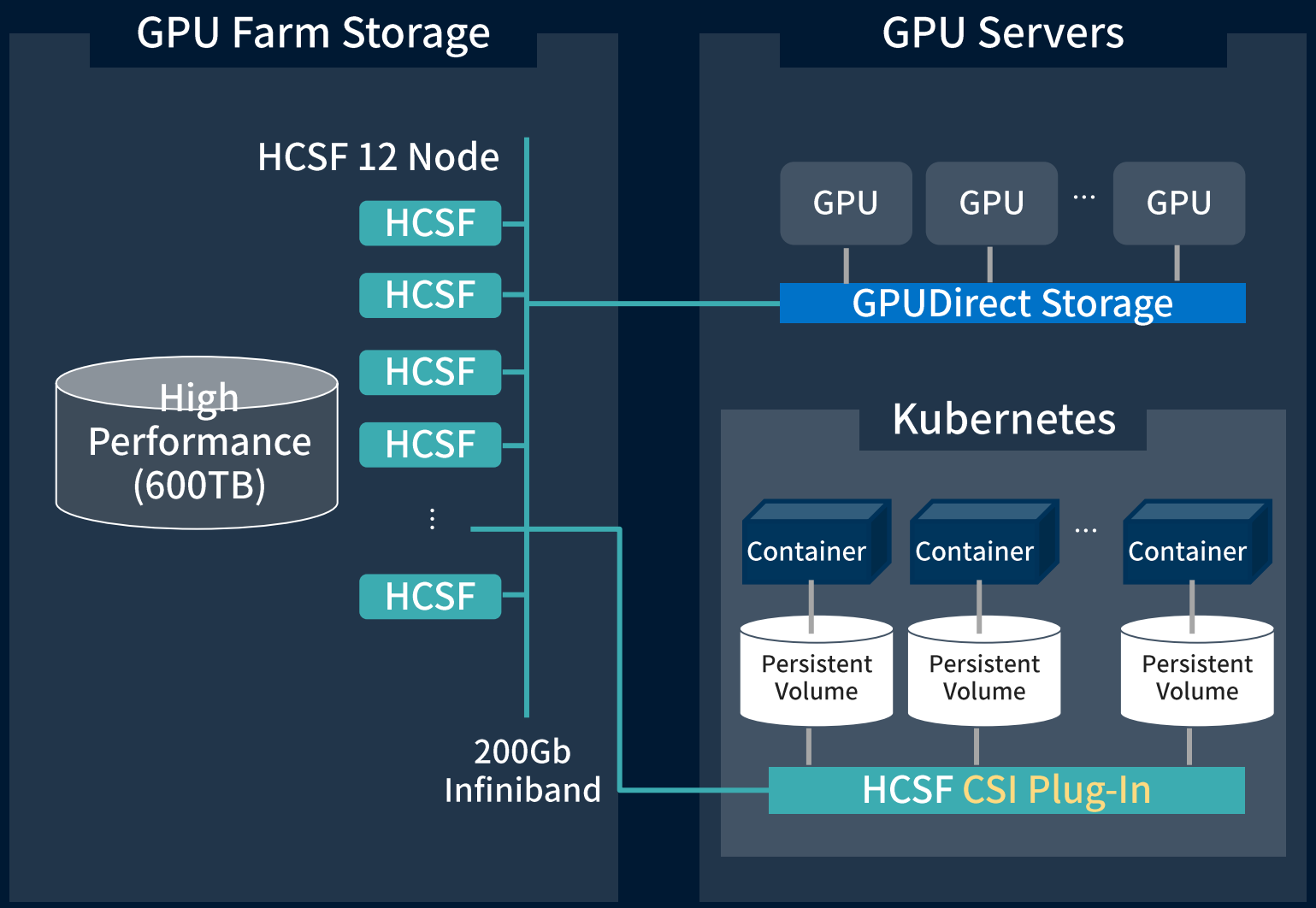
구축 내용

- 최신 Supermicro GPU 서버, 고성능 스토리지 HCSF, 고속 네트워크, Slurm AI HPC 인프라 구축
- Kubernetes 연계(CSI) 기능을 제공하여 볼륨 운영 최적화 / 효율화 지원

도입 효과

- MLPerf 테스트 검증된 최고성능 GPU서버와 GPUDirect Storage 적용으로 GPU 연산성능 최적화
- 확장에 따른 성능저하 이슈 없는 무제한급 증설지원 저장자원 제안
- AI플랫폼 구축 경험 있는 전문가로 구성된 전담 조직으로 AI플랫폼 연계 기술지원 보장

사례 2. 국내 AI 서비스 기업 (600TB)



사업 목적

- AI 교육 및 자사 AI 학습 모델 테스트 환경을 위한 프라이빗 클라우드 기반 HPC 인프라 구축

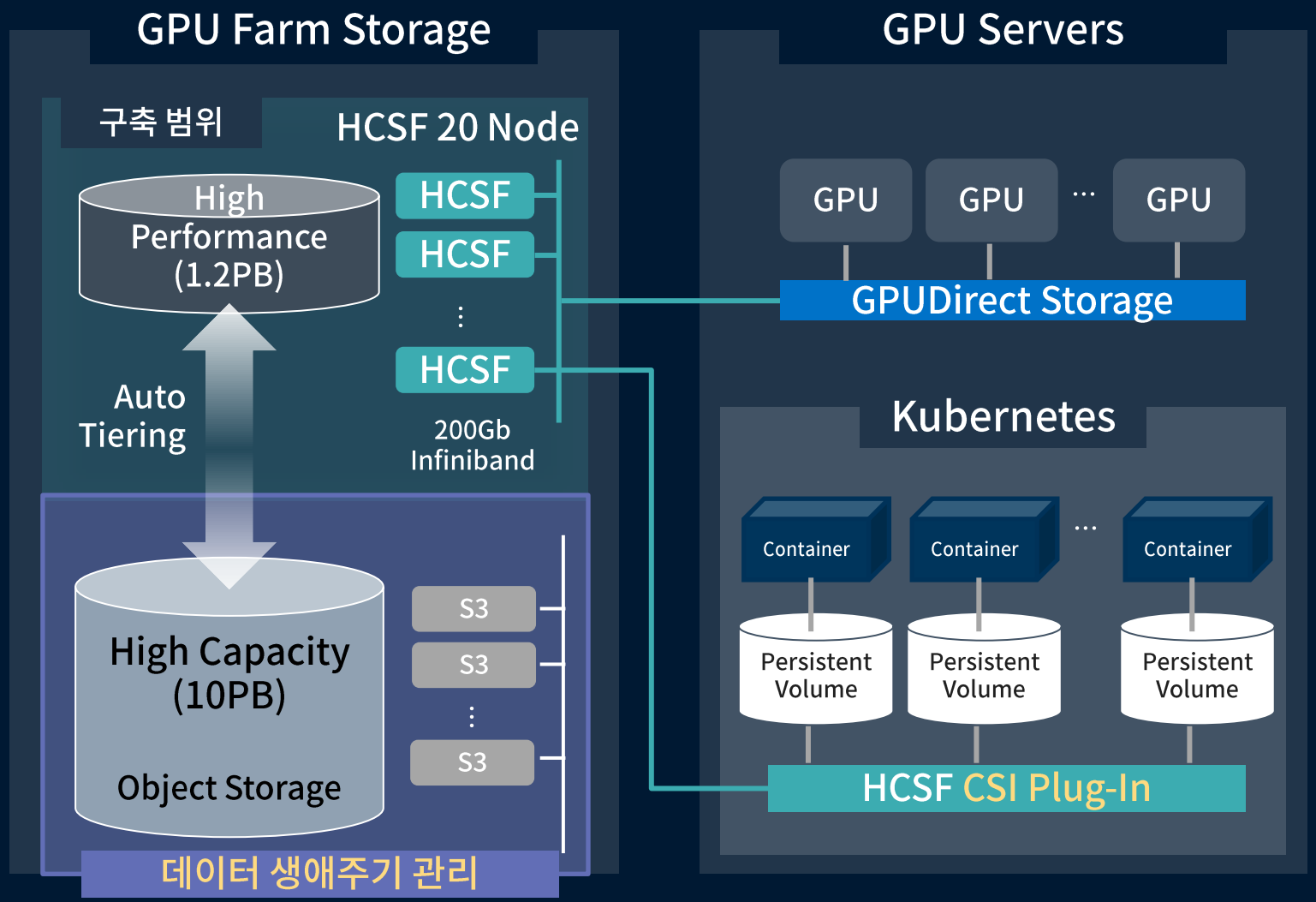
구축 내용

- GPU 환경 구현을 위한 고성능 스토리지 HCSF 구축
- GPU 성능 최적화 GDS 지원 스토리지 구축
- Kubernetes 연계(CSI) 기능을 제공하여 볼륨 운영 최적화 / 효율화 지원

도입 효과

- 다수 개발자/분석가/교육생 공동 운영을 위한 GPU 전용 스토리지 인프라 확보
- 향후 요구 사항 변화 (데이터 증가, 사용자 증가 등)에 유연한 대응 확보
- GPU 서버 인프라 고려한 스토리지 최적화

사례 3. 국내 R&D기업 (1.2PB)



사업 목적

- 자연어 처리 알고리즘 학습 성능 개선을 위한 GPU 서버와 전용 스토리지를 도입

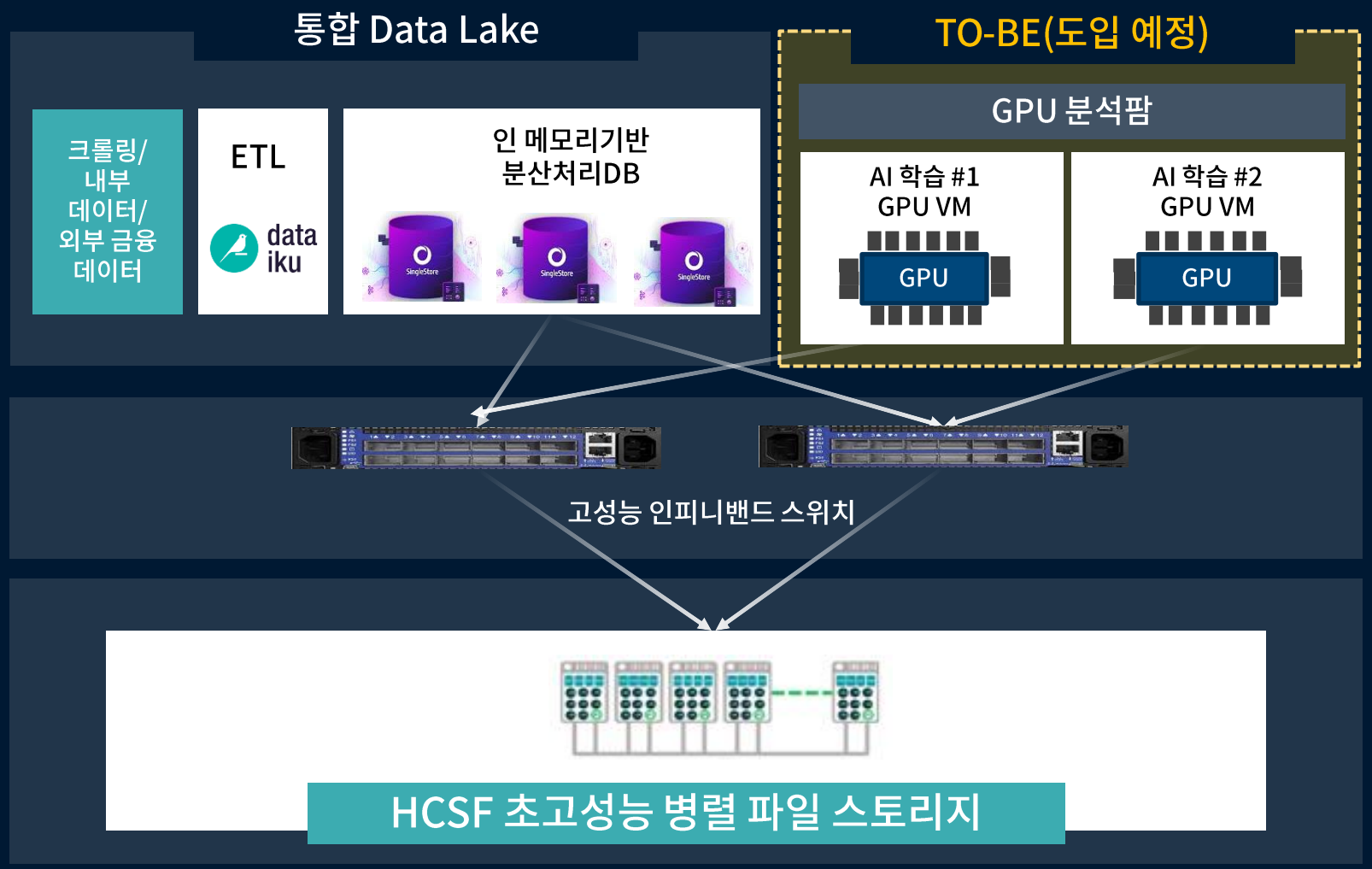
구축 내용

- GPU 환경 구현을 위한 고성능 스토리지 HCSF 구축
- GPU 성능 최적화 GDS 지원 스토리지 구축
- Kubernetes 연계(CSI) 기능을 제공하여 볼륨 운영 최적화 / 효율화 지원
- 고객사 S3 스토리지와 Tiering 구성

도입 효과

- 자연어 처리 기계학습 연산 성능 향상
- 서버 인프라 구성 고려한 스토리지 최적화
- 자체 S3 스토리지 활용성 확보를 통한 비용 효율적인 구성

사례 4. 국내 금융기관



사업 목적

- 외부기관 데이터/국내금융기관 데이터 통합으로 효과적인 데이터 분석 및 분석 체계 확립

구축 내용

- 정형/반정형 데이터 분석을 위한 데이터 레이크 구축
- 차년도 예정 사업인 GPU기반 딥러닝 분석환경으로 확장 가능한 아키텍처 수립

도입 효과

- 통합 플랫폼(CPU → + GPU)을 이용해 관리 포인트 최소화 및 확장 유연성 제공
  → 단 2명의 소수 운영자로 관리가 가능해짐
- 고성능 단일 데이터 레이크 저장소 구축 운영
  → 여러 대의 업무 별 스토리지를 하나로 통합
- 고성능 쿼리 분석 엔진 도입

AI 구축을 위한 체크포인트

AI/HPC 요구 사항 확인 및 PoC/BMT 필수!!!

IOPS, 처리량 요구 성능, 요구 기능, 솔루션 호환성 검증

AI/HPC 시스템 데이터 레이크 구축 경험 필수!!!

대용량의 파일 처리, 자동 티어링, 안정적인 확장

국내외 실 사례를 통한 국내 기술력(인력) 필수!!!

장애 지원, 버티컬 AI 솔루션 & 도메인 전문가 연계 지원

감사합니다.