

AI 전성시대, 기업의 생존 전략

챗GPT로 촉발된 국내 AI 비즈니스 열기가 나날이 뜨거워지고 있다.

생성형 AI를 도입하려는 기업들의 움직임이 활발해지고 있으며,
전에 없던 형태의 비즈니스 수요가 폭발하고 있다. 말 그대로 'AI 전성시대'다.

AI 전성시대 한 가운데서 자사에 적합한 솔루션을 찾고 있다면, Cover Story를 주목해 보자.

국내 대표 AI 서비스 기업 업스테이지가 효성인포메이션시스템의 솔루션으로

성과와 관리, 운영의 편의성을 한 단계 업그레이드한 사례와

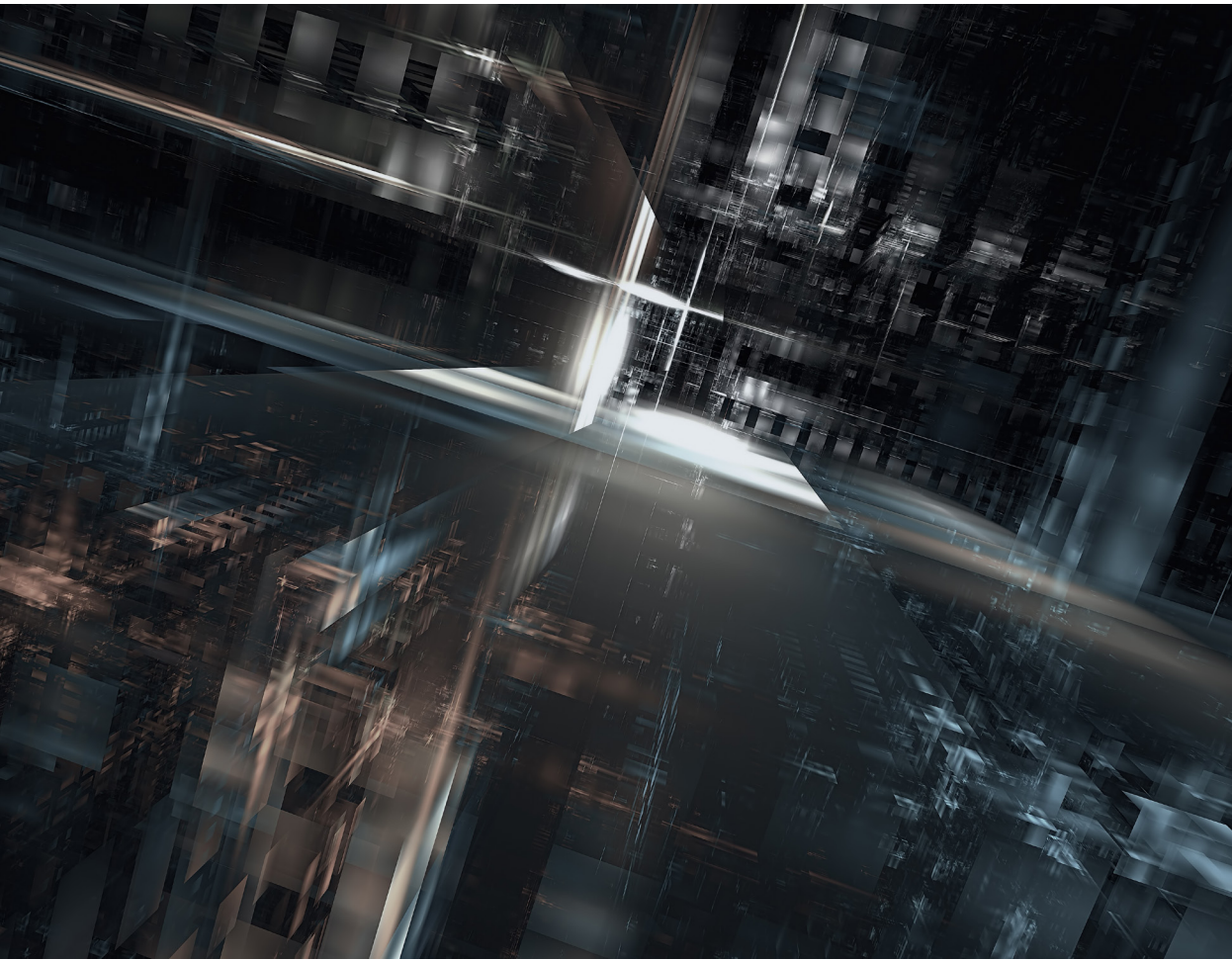
생성형 AI를 위한 차세대 고성능 데이터 저장소인 HCSF를 소개한다.



Part I

국가대표 AI 기업 ‘업스테이지’
HCSF 도입으로 HPC 환경 마련,
성능 개선·운영 편의성 UP!

기업들이 규모를 막론하고 앞다퉈 AI 비즈니스에 뛰어들고 있는 가운데, AI 서비스 전문기업 ‘업스테이지’의 기세가 두드러진다. 글로벌 빅테크 기업과 견주어도 손색없는 기술력으로 두각을 나타내고 있는 업스테이지. 지난해 더욱 발전된 비즈니스를 위해 효성인포메이션시스템의 초고성능 병렬 파일 스토리지 ‘HCSF’를 도입해 성능 개선뿐만 아니라 관리와 운영의 편의성을 한층 높였다.



업스테이지, 국내 AI 비즈니스 분야에서 고공행진 중
‘Making AI Beneficial’

업스테이지는 인공지능 거대 언어 모델(LLM)과 문서 자동화(DocAI) 분야에서 세계 최고 수준의 기술을 보유한 국내 AI 스타트업으로, 기업들의 디지털 혁신을 위한 B2B 솔루션을 제공하고 AI 교육 프로그램을 운영하고 있다.

올해 3월, 업스테이지는 모바일 메신저 카카오톡에 업스테이지의 광학문자인식(OCR) 기술과 챗GPT¹⁾를 더한 AskUp(애스크업)을 런칭했다. AskUp은 사용자가 문서의 사진을 찍거나 전송하면 그 내용을 읽어 이해하고 답변할 수 있는 ‘눈 달린 챗GPT’다. 출시한 지 5개월여 만에 AskUp은 130만 명이 활용하는 국내 대표 AI 서비스로 등극했다.

최근에는 업스테이지가 개발한 생성 AI 모델이 세계 최대 머신러닝 플랫폼 허깅페이스에서 운영하는 ‘오픈 거대 언어 모델(LLM) 리더보드’ 평가에서 72.3점을 획득하며, 챗GPT의 기반인 GPT-3.5보다 성능이 뛰어나다는 평가를 받았다.

성능 개선과 운영의 편의성을 위한 선택, HCSF

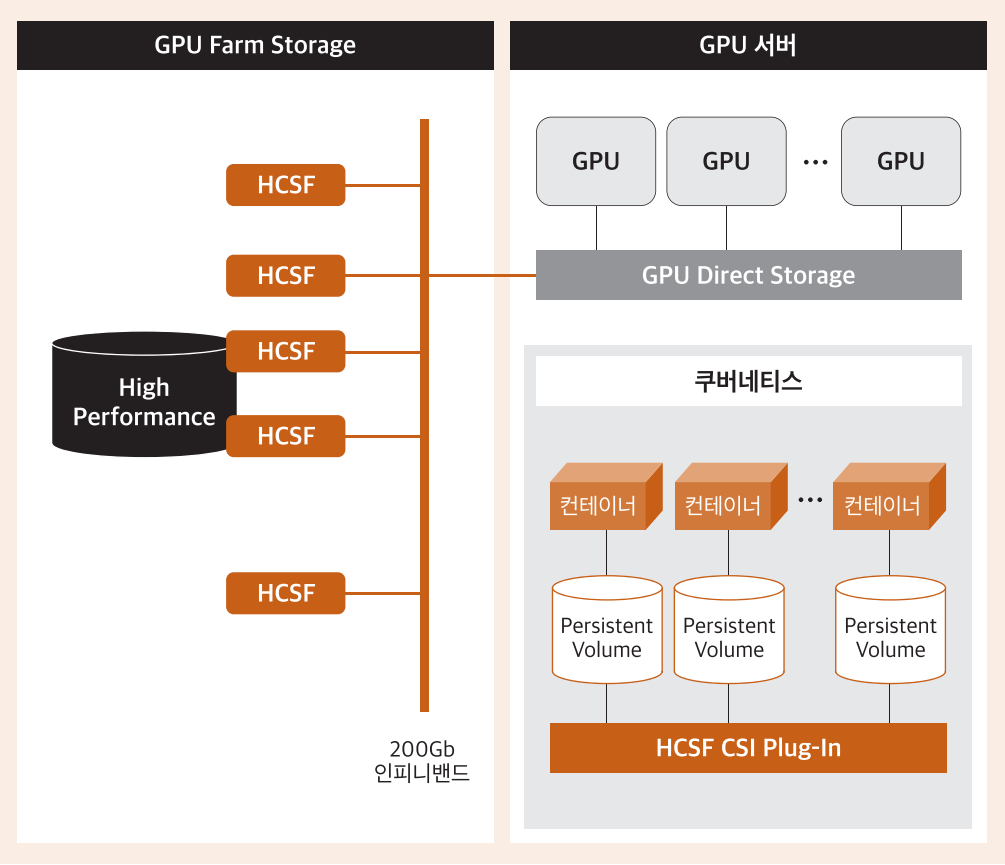
국내 AI 시장의 강자로 부상한 업스테이지는 한 단계 발전된 AI 비즈니스를 선보이기 위해 효성인포메이션시스템을 선택했다. AI 교육과 AI 학습 모델 테스트 환경을 위한 프라이빗 클라우드 기반의 HPC(고성능 컴퓨팅) 환경을 구축하면서, 초고성능 병렬 파일 스토리지인 ‘HCSF(Hitachi Content Software for File)’를 도입한 것이다.

그간 업스테이지는 AI 교육을 위해 내부 GPU 분석 서버 인프라와 올플래시 NAS를 운영해 왔다. 그러던 중, 비즈니스 활용도 및 편의성을 높이기 위해 스토리지 성능을 개선하고 운영 부담을 덜어줄 수 있는 솔루션에 대한 필요성이 점점 높아졌다. 업스테이지는 다방면의 검토 끝에 NVIDIA GPU 분석 환경의 성능을 완벽하게 활용할 수 있는 효성인포메이션시스템의 HCSF를 데이터 저장소로 선정하고, 인피니밴드²⁾ 네트워크 환경 구성을 통해 빠르고 효율적인 AI 교육 서비스를 제공할 수 있는 최상의 AI 분석 환경을 구현했다.

1) 챗GPT: GPT-3.5와 GPT-4를 기반으로 하는 대화형 인공지능 서비스로, 채팅을 하듯이 챗봇에 질문을 입력하면 AI가 빅데이터 분석을 바탕으로 대화하듯이 답변을 해주는 시스템이다.
2) 인피니밴드(InfiniBand): 고성능 컴퓨팅과 기업 데이터센터에서 사용되는 스위치 방식의 통신 연결 방식으로, 높은 스루풋과 낮은 레이턴시, 높은 안정성과 확장성이 특징이다.

특히 HCSF를 NVMe 고성능 Tier1로 구성해 성능을 극대화하고, GPU 서버와 스토리지 간 직접 연결을 위한 GPU 다이렉트 기술(GDS: GPUDirect Storage)로 GPU 환경의 스토리지 응답 성능을 극대화했다. 또한 CSI(Container Storage Interface Plug-in)를 통한 컨테이너 기반 관리로 스토리지 볼륨 운영을 최적화하고 효율성을 높일 수 있었다.

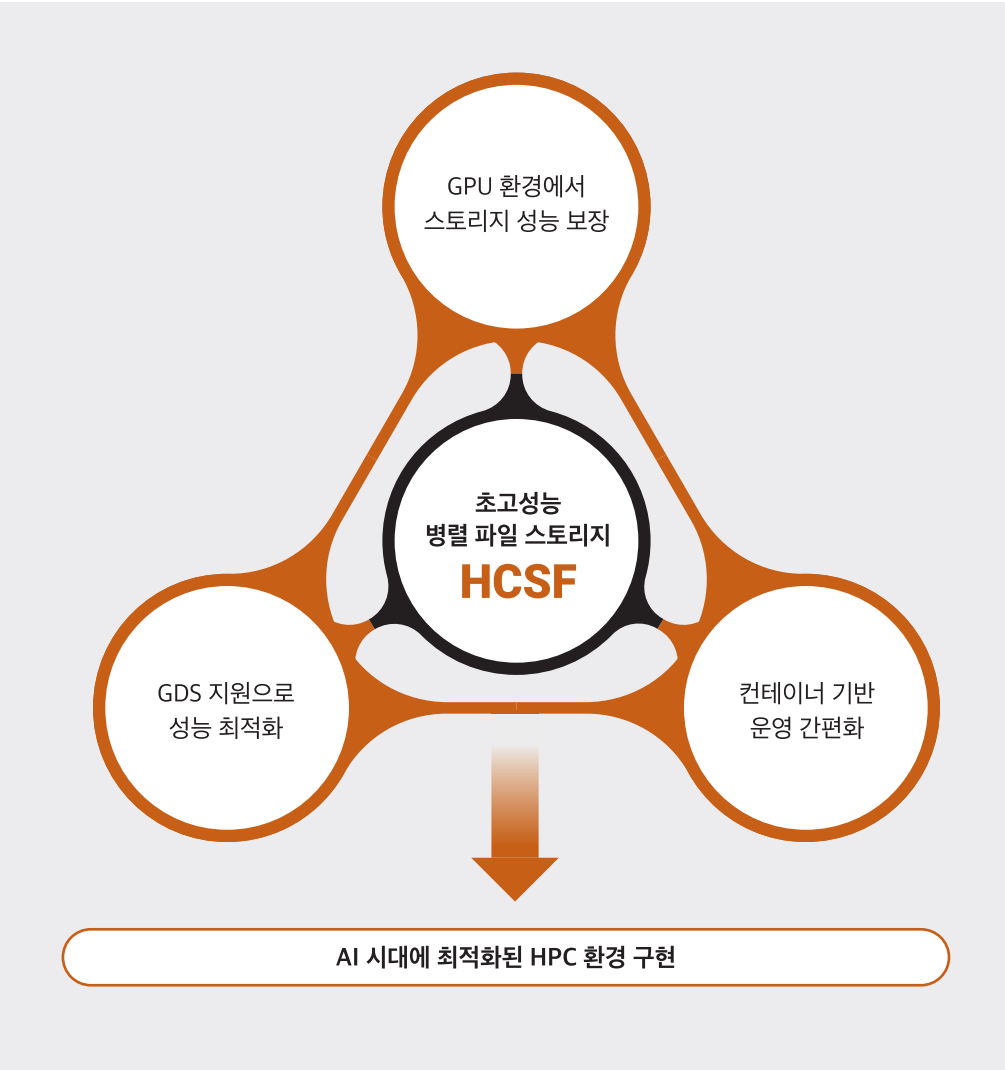
↓ 업스테이지의 HCSF 구축 모형



관리 및 운영 편의성, ‘따라올 자가 없다’

현재 업스테이지는 HCSF를 활용해 HPC 환경을 구현하고 안정적으로 서비스를 운영 중이다. 업스테이지 담당자는 “스타트업의 특성상 IT 운영 인력이 충분하지 않습니다. 스토리지의 성능 향상뿐만 아니라 운영의 편의성이 높은 솔루션이어야 한다는 점이 중요했지요. HCSF는 이러한 요구사항을 완벽히 충족시켜 줬고, 구축 이후 현재까지 매우 만족하고 있습니다.”라고 전했다.

↓ 업스테이지가 선택한 HCSF



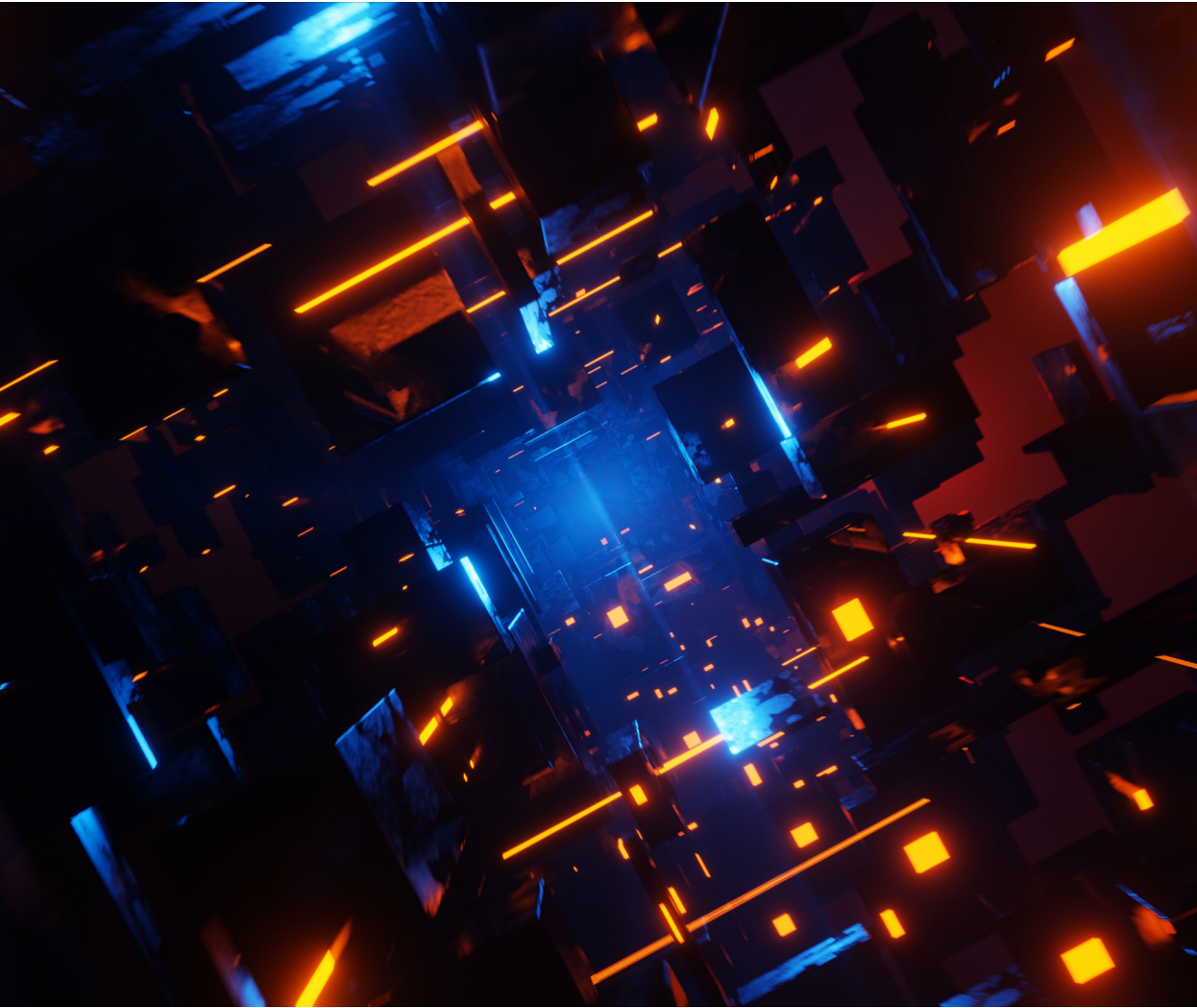
업스테이지는 HCSF를 도입함으로써 다수의 개발자, 분석가, 교육생 공동 운영을 위한 GPU 전용 스토리지 인프라를 확보했다. 또한 향후 데이터 및 사용자 증가 등 갑작스러운 요구사항의 변화에도 유연하게 대응할 수 있게 됐다.

AI로 이로운 세상을 만들겠다는 포부를 지닌 업스테이지. AI 대전환 시대에 최고의 AI 기술력을 구현하기 위해 인프라 혁신에 투자를 아끼지 않는 업스테이지가 국내를 넘어 세계 최고의 자리에 우뚝 서는 건 시간문제다.

Part II

생성형 AI 열풍, 고성능 스토리지 시대를 열다

최근 AI의 급발전으로 데이터가 폭증하면서, 분석해야 할 데이터의 양 역시 폭발적으로 늘고 있다. 다양한 종류의 데이터를 비용 효율적으로 저장하고 활용하는 데이터 운영의 혁신이 필요한 때다. AI/ML, 고성능 데이터 분석 등 최신 엔터프라이즈 워크로드를 위해서는 그에 걸맞은 최고의 스토리지가 필요하다.



데이터 레이크, 고성능을 요구하다

데이터 유형에 상관없이 모든 데이터를 한 저장소에 모아 놓은 ‘데이터 레이크’. 그동안 데이터 레이크는 인프라 성능에 큰 부담을 미치지 않는 요소였다. 문제는 생성형 AI¹⁾ 열풍이 불면서부터다. 기업들이 수많은 콘텐츠에서 생성하는 대규모 언어 모델(LLM)²⁾에 대한 관심이 높아지면서, AI 모델 학습을 위한 고성능의 데이터 레이크 인프라에 대한 관심이 부상하고 있다.

전통적인 SAN이나 NAS의 성능과 확장성으로는 데이터 레이크를 처리하는 데 한계가 있다. 데이터 노드를 늘린다 해도 메타 정보를 찾는 속도가 서버 속도를 따라가지 못한다.

GPT는 데이터 수집, 전처리, 모델 생성, 평가, 아카이빙 등의 여러 단계를 거친다. 이 과정에서 전수 데이터를 읽으며 상당한 디스크 IO를 일으킨다. NVMe 플래시가 개별 속도는 빠를지 모르지만, 데이터를 찾아서 GPU로 이동시키는 게 느리다면 전반적인 시스템 속도가 현저히 떨어질 수밖에 없다.

즉, AI 시스템에 고속으로 정확하게 데이터를 공급하고 분석하려면, 서버 외에 스토리지나 네트워크도 고성능의 요건을 반드시 충족해야 한다. 특히 스토리지의 경우에는 고속 처리를 위해 하드디스크보다 NVMe 플래시를, 방대한 양의 데이터를 더욱 고성능으로 저장하기 위해 병렬 분산 파일 시스템을 권장하는 추세다.

여기서 가장 중요한 문제는 비용효율적인 저장소의 아키텍처이다. 오브젝트 스토리지와의 단일화된 아키텍처(Unified Namespace) 구성이야 말로 차세대 고성능 데이터 레이크 저장소로 적합하다. 또한 딥러닝은 하나의 데이터셋을 잘게 쪼개서 사용하기 때문에 작은 IO 요청이 빈번하게 일어난다. 이 때문에 높은 IOPS(초당 입출력 횟수)가 요구되고, 저장소 역시 메타정보 오버헤드를 최소화할 수 있는 구조여야 한다.

효율적인 데이터 관리를 위한 고성능 솔루션, HCSF

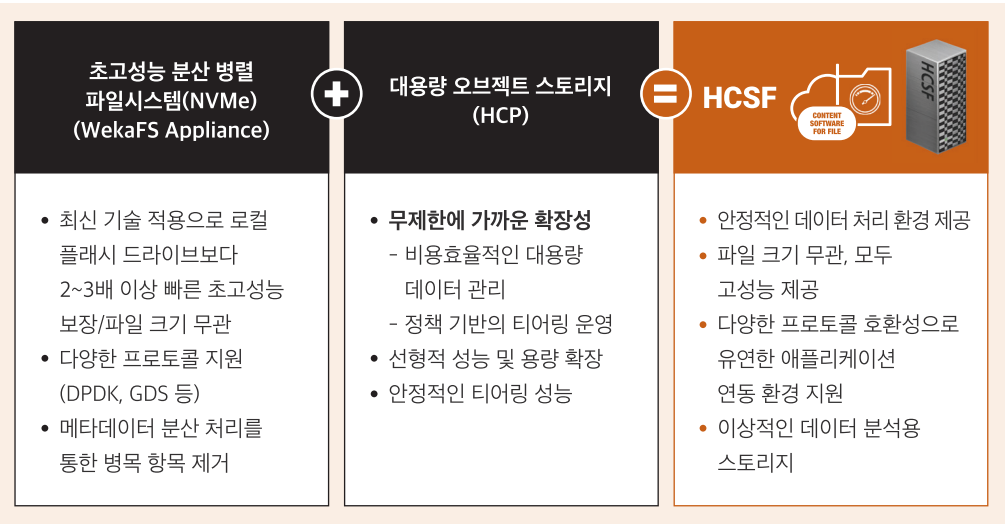
효성인포메이션시스템이 제공하는 HCSF(Hitachi Content Software for File)는 병렬 파일시스템과 오브젝트 스토리지를 통합한 파일 스토리지로, 이러한 문제를 완벽하게 해결한다.

1) 생성형 AI(Generative AI): 텍스트, 오디오, 이미지 등 기존 콘텐츠의 패턴을 학습해 추론한 후 사용자의 특정한 요구에 대한 결과를 생성해 내는 인공지능(AI) 기술. 새로운 콘텐츠 생산에 그치지 않고 생성자와 만들어진 콘텐츠를 평가하는 판별자가 끊임없이 서로 대립하고 경쟁하며 또 새로운 콘텐츠를 생성해 낸다.

2) LLM(Large Language Model, 대규모 언어 모델): 자연어 처리와 이해를 수행하는 러닝 모델. 방대한 양의 자연언어를 처리해서 사람이 생성한 것인지 구별할 수 없을 정도의 응답을 할 수 있는 인공지능을 말한다.

HCSF는 오브젝트 스토리지의 가용량과 클라우드 기능을 갖추면서도 분산 파일 시스템 속도를 제공하며, 멀티 프로토콜을 지원해 데이터 수집이 편리하다. 특히 정책 기반 티어링 파일 시스템 운영으로 비용 효율적으로 데이터를 관리할 수 있다.

↓ AI 워크로드에 적합한 고성능 스토리지 'HCSF'



대부분의 데이터가 초기에는 많은 IO를 요구하지만, 시간이 지날수록 액세스 빈도가 급격히 낮아진다. 따라서 병렬 처리 시스템의 확장에 따른 메타데이터 서버도 함께 확장될 수 있는 아키텍처가 필요하며, 액세스 빈도가 낮아지는 데이터를 병렬 처리 시스템 스토리지 레이어에서 자동으로 티어링을 통해 저렴한 스토리지로 이동시킨다면, 비용을 상당히 절감할 수 있다.

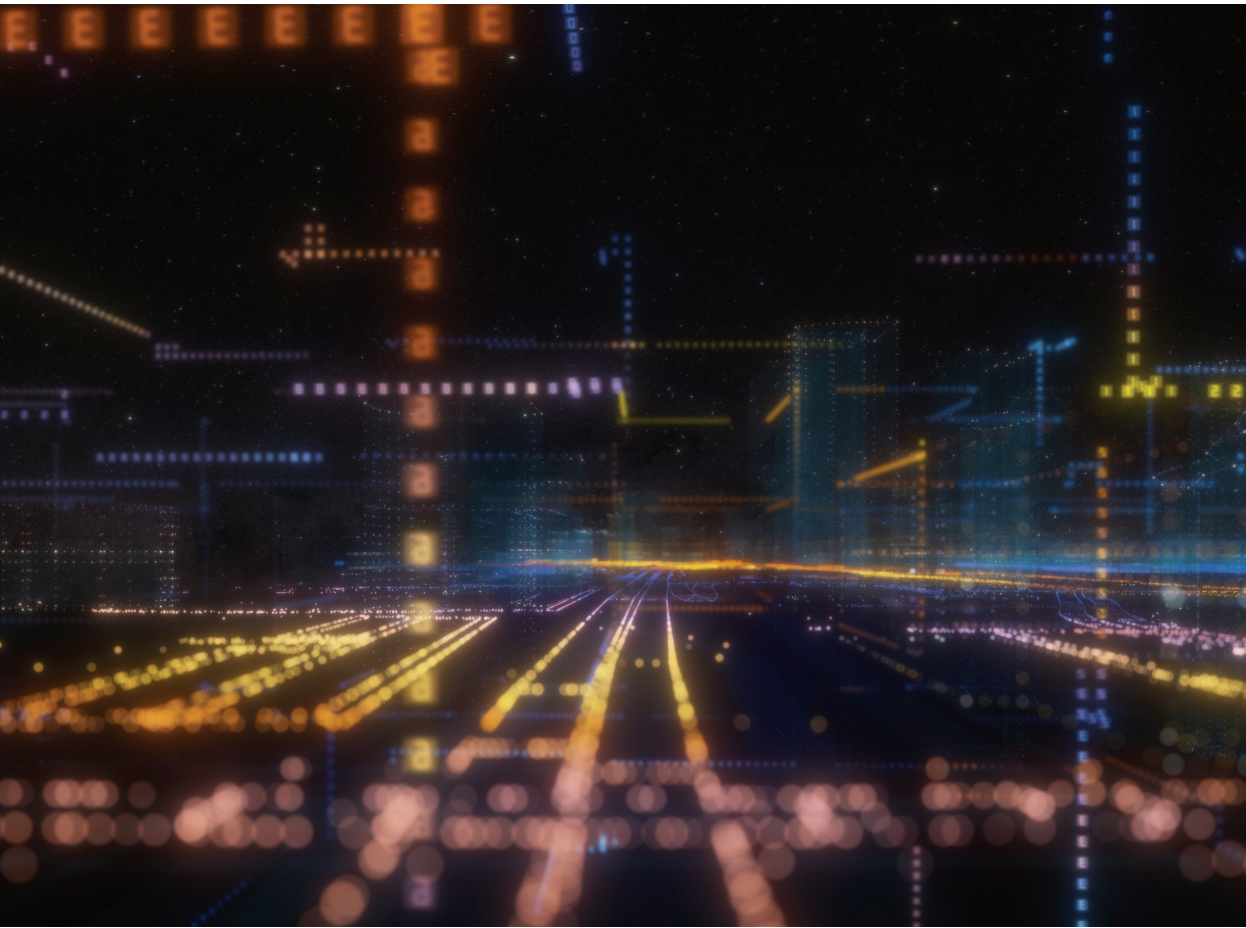
효성인포메이션시스템의 HCSF는 AI/ML, 고성능 데이터 분석 등 최신 엔터프라이즈 워크로드에 적합한 데이터 처리환경을 제공한다. 실시간 데이터 처리를 위한 고속 네트워킹 기술이 적용되고, NVMe Tier1 스토리지와 S3 Compatible Tier2 오브젝트 스토리지 간 오토 티어링이 적용된다. 최대 EB(엑사바이트) 규모의 데이터까지 확장 가능한 Tier1과, 무제한 확장 가능한 Tier2 데이터에 신속하게 액세스할 수 있어 데이터 가치와 활용을 극대화할 수 있다.

HCSF는 GPU와 스토리지 간 직접 연결을 위한 GPU 다이렉트 기술(GDS)로 응답 성능을 최적화하며, CSI(Container Storage Interface Plug-in)를 통한 컨테이너 스토리지 볼륨 운영을 최적화하고 효율성을 높여준다.

생성형 AI 열풍으로 촉발된 HCSF 수요

효성인포메이션시스템이 HCSF 비즈니스를 시작한 지는 약 2년 전으로 거슬러 올라간다. 그런데 최근 몇 달 사이에 수요가 눈에 띄게 증가했다. 이는 AI의 유행에 따라 GPU에 대한 관심과 활용도가 높아졌고, 그에 적합한 스토리지로 HCSF가 주목받게 된 것이다.

흔히 HPC 환경에서는 GPU만 도입하면 완벽하다고 생각하기 쉽다. 그러나 이를 잘 활용하기 위해서는 네트워크 환경 또한 중요하다. AI 모델을 학습할 때 방대한 양의 데이터 IO가 필요하고, 재학습하면서 또 무수한 데이터를 받아들이는 과정을 반복하게 되는데, 이는 결국 스토리지의 병목 현상을 유발한다. 즉 스토리지의 성능이 탄탄하게 받쳐 줘야 완벽한 HPC 환경을 구현했다고 말할 수 있다.



초기에는 기업들이 GPU나 네트워크 부분에 주로 투자했지만, 이제는 모든 인프라의 원활한 흐름을 위해 이에 적합한 고성능 스토리지를 도입하려는 수요가 많아지고 있다. HCSF 도입 사례를 보면 이런 트렌드를 알 수 있다.

대형 제조기업 A사는 데이터 웨어하우스/하둡 데이터 분석 시스템을 이용하고 있었으나, 확장성과 성능 저하가 가장 큰 문제가 되었다. A사는 이를 해결하기 위해 전사 통합 리포지터리를 구축해 차세대 전사 데이터 분석 체계로 전환했다. 현장에서 발생하는 데이터는 고성능 데이터 분석을 기반으로 대용량 쿼리가 가능하도록 했으며, 향후 AI/ML을 위한 전사 분석 체계도 마련했다.

앞서 소개한 업스테이지 사례도 같은 맥락이다. 업스테이지는 자사가 제공하는 AI 비즈니스의 성능 개선과 운영 부담을 덜어줄 수 있는 솔루션으로 HCSF를 선택했고, 관리와 운영 면에서도 크게 만족하고 있다.

최근에는 내부적으로 LLM(대규모 언어 모델)을 구축하려는 기업이나 기관들이 특히 HCSF 및 HPC 인프라에 관심을 보인다. 기업 내부의 데이터를 오픈할 수 없기 때문에 경량화된 LLM 모델을 자체적으로 구축해서 사내 서비스를 제공하기 위해서다. 특히 이전까지는 챗봇 서비스를 많이 활용했던 금융사들도 보안과 전문성을 결합한 형태의 서비스를 선보이기 위해 관심을 나타내고 있다.

로드맵 수립부터 원스톱 지원까지, 전천후 전문가

AI 비즈니스를 준비하는 기업 중 LLM 모델 및 AI 모델을 한 가지 관점에서만 고민하는 경우가 많다. 그런데, 이러한 모델은 끊임 없이 진화하고 있어 지속적으로 운영하기가 쉽지 않다. 탄탄한 인프라를 위한 컴퓨팅 파워, 네트워크 속도, 스토리지 구성 등을 초기부터 종합적으로 살펴보고 기업 환경에 맞는 전략을 수립해야 향후에 기회비용을 최소화할 수 있다. 이에 대해 조직 내에서 자체적으로 판단하기 어렵다면, 전문 벤더의 컨설팅을 통해 환경에 걸맞은 로드맵을 수립한 후 최적의 인프라를 체계적으로 도입하는 것이 무엇보다 중요하다.

효성인포메이션시스템은 컴퓨팅부터 네트워크, 스토리지까지 HPC 구현에 필요한 모든 인프라 기술과 컨설팅을 종합적으로 제공한다. 수년간 다양한 산업 분야에서 쌓은 노하우와 전문화된 인력, 레퍼런스를 보유하고 있으며, 이를 바탕으로 안정적인 원스톱 지원이 가능하다.

기업마다 원하는 바와 현재 상황은 다를 수밖에 없다. 효성인포메이션시스템은 기업의 상황과 목표하는 바를 종합적으로 고려하여 차별화된 로드맵과 인프라 계획을 처음부터 끝까지 함께 수립하는 유일한 벤더다.

↓ HPC를 위한 초고성능 데이터 레이크하우스 구축 시 체크포인트



“최근 생성형 AI를 도입하려는 기업 대부분이 컴퓨팅과 네트워크뿐만 아니라 데이터 병목 현상을 줄이기 위해 고성능 스토리지 도입의 필요성을 인식하기 시작했다. 효성인포메이션시스템은 AI 전문 기업뿐만 아니라 대기업, 중견/중소기업, 금융 등 기업의 규모나 업종에 관계 없이 안정적으로 AI를 구현하는 최적의 환경을 제안하고 지원할 것이다.”