금융권 인공지능 시대 이끄는 단계별 AI 솔루션을 주목하라

김형섭 / 효성인포메이션시스템 컨설턴트

AI는 산업 분야와 상관없이 이미 우리 생활 속에 깊숙이 들어와 있다. 휴머노이드 로봇부터 완전자율주행 자동차까지, 자동차 산업에 AI가 일상화된 지 오래다. 금융 서비스 영역에서도 많은 기업이 AI, 분석 등 신기술 활용에 적극적이다.

금융산업에 AI를 적용하기 위한 요소

금융권은 다른 산업에 비해 AI를 적용하기에 훨씬 더 좋은 인프라 조건을 갖추고 있다. 이는 두 가지 이유 때문이다. 다른 산업과는 비교할 수 없을 정도로 많은 데이터, 그리고 데이터의 정확도가 높다는 것이다. 방대한 양의 데이터는 정확한 예측과 의사결정에 중요한 밑거름이 된다.

실제로 금융권에서는 금융 시장 분석 및 전망, 금융상품 추천, 담보물의 가격 결정, 리스크 측정 및 관리, 업무 자동화 등 사실상 전 분야에서 이미 AI를 활용하고 있거나 시도하고 있다.

그렇다면 금융권을 포함한 전 산업 분야에서 AI를 적용하는 데 필요한 요소는 무엇일까?

크게 데이터, AI 모델, 그리고 연산 자원 세 가지로 정리할 수 있다. 가장 먼저 해야 할 일은 데이터 준비다. 산재해 있는 데이터를 AI 모델에 맞게 준비해야 그에 적합한 알고리즘을 선택해 AI 모델로 개발할 수 있다. 그리고 나서 학습과 평가 과정을 거치고 마지막으로 활용 단계인 추론으로 이어진다.

효성인포메이션시스템의 AI 플랫폼은 이러한 프로세스의 각 단계에 적합한 AI 솔루션을 제공한다. 1단계 데이터 운영은 초고성능 스토리지인 HCSF가, 2단계 AI 모델 서비스는 래블업(Lablup)의 Backend.AI가, 그리고 3단계 연산자원 성능은 GPU 시스템인 NVIDIA DGX와 HGX가 각각 지원한다.

효성인포메이션시스템의 AI 플랫폼이 단계별 AI 프로세스를 어떻게 지원하는지 하나씩 살펴보자.

HCSF, 완벽한 데이터 운영 솔루션

HCSF(Hitachi Content Software for File)는 NVMe 기반의 초고성능 병렬 파일 시스템과 대용량 오브젝트 스토리지가 하나로 통합된 솔루션이다. 필요할 때만 고속으로 데이터를 가져와 분석에 활용할 수 있기 때문에 진정한 데이터 레이크 스토리지 환경을 구축할 수 있다. 다음은 HCSF의 특장점 다섯 가지를 꼽아봤다.

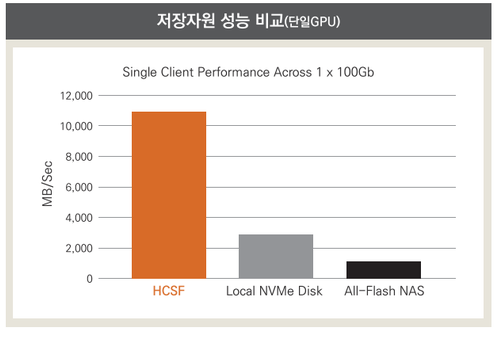
01 | 알고리즘 트레이딩으로 리스크는 낮추고 분석은 향상

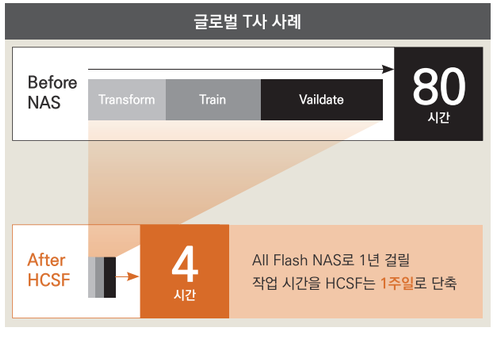
초고성능 병렬 파일 스토리지인 HCSF는 스토리지 병목 현상이 발생하지 않기 때문에 방대한 규모의 데이터 레이크 스토리지 환경을 구축해 시장 트렌드 분석에 활용할 수 있다. NVMe의 고성능과 짧은 대기시간의 이점을 활용하면, 이전에는 불가능했던 알고리즘 트레이딩과 백테스팅도 가능하다. 또 NVIDIA GPUDirect가 업무 워크로드를 가속하므로 분석 통찰력도 실시간으로 확보할 수 있다.

02 | 데이터 병목 현상 제거, 대량 거래 가능

대량 거래가 많았던 한 금융 서비스 기업의 이야기다. 이 회사는 자동화된 전용 트레이딩 솔루션이 공유 병렬 파일 시스템에 배치되어 있어 병목 현상이 자주 발생했고, 이에 따라 성능이 75%나 저하됐다. 그러나 HCSF를 도입한 이후, 전반적인 성능이 3배 정도 향상되는 놀라운 성과를 얻고 있다. 로컬 NVMe 드라이브를 능가하는 성능 덕분에 데이터를 애플리케이션 서버에 복제하지 않게 되어 스토리지 비용도 65% 절감할 수 있었다.

↓ 저장자원 성능 비교





03 | 비용은 줄이고, 확장성은 증가

주식 거래 플랫폼을 제공하는 뉴욕증권거래소의 한 서비스 기관은 인프라 관련 비용이 감당할 수 없는 수준까지 증가했다. 하지만 인프라 확장은 쉽지 않았고, 지나치게 높은 비용 부담으로 인해 강력한 DR 전략도 수립하기 힘겨웠다. 그러나 HCSF를 도입하자 성능이 7배까지 향상되었으며, 컴퓨팅 리소스 비용은 1/7로 줄었다. 또한 대규모 재해가 발생하더라도 복구가 간편하고, 컴퓨팅과 스토리지 비용도 대폭 절감할 수 있게 되었다.

04 | 보안, 데이터 보호, 랜섬웨어 보호, 재해 복구

데이터는 기업의 생명선이다. 그러나 안타깝게도 보안 침해와 랜섬웨어 공격은 이미 일상이 되어버렸다. 갈수록 고도화되는 위협과 엄격해지는 규제 속에서 기존의 스토리지로는 컴플라이언스를 제대로 준수하기 어렵다. HCP(Hitachi Contents Platform)는 민감한 데이터 스토리지의 최신 요구사항을 처리할 수 있도록 설계된 선도적인 컴플라이언스 전략 기술 솔루션이다.

"기업이 컴플라이언스를 준수하기 위해서는
최신 데이터 스토리지 아키텍처가 필요하다. 정확히 히타치 밴타라가 제공하는
솔루션이 바로 이것이다. 현대 디지털 시대에 중요한 데이터를 안정적으로
관리하고자 한다면, 반드시 히타치 밴타라를 고려해야 한다."

-ESG(Enterprise Strategy Group)

05 | 제로카피/제로튜닝 아키텍처로 고성능 애플리케이션 지원

HCSF를 이용하면 동일한 스토리지 백엔드에서 전체 파이프라인을 운영할 수 있어, 복제 관련 비용과 지연을 줄일 수 있다. HCSF는 엄청나게 빠른 파일시스템이기 때문에, 로우 레이턴시 환경에서 높은 I/O, 작은 파일, 혼합 워크로드 및 데이터 이동성을 지원한다. 온프레미스와 클라우드에서 모두 동작하고, 플랫폼 간 확장도 용이하다. 멀티 카피 병목현상이 제거되어 '로컬 스토리지보다 빠른 속도'를 보장하며, 대기 시간도 줄어 대규모 데이터 파이프라인을 효과적으로 가속할 수 있다. 또 AI/ML 애플리케이션에 대한 추론 시간도 단축할 수 있다. HCSF는 NVIDIA GPUDirect 스토리지뿐 아니라 POSIX, NFS, SMB, S3 등 다양한 멀티 프로토콜을 지원하는 솔루션이다.

Backend.AI, AI 모델 서비스 지원

변화가 거의 없는 기존 애플리케이션과 달리 최근에 등장하는 애플리케이션은 유동적이고 변화도 잦다. AI 플랫폼은 컨테이너 운영 환경이 필수적이며, 데이터 과학자들이 AI 플랫폼을 선호하는 것도 이 때문이다.

효성인포메이션시스템이 제공하는 래블업사의 'Backend.AI'는 아태지역 최초로 NVIDIA에서 성능과 신뢰성을 검증한 DGX-Ready 소프트웨어로, 다음과 같은 장점을 가지고 있다.

GPU 활용 극대화 컨테이너 환경에서 GPU 자원을 쪼개 분할 할당할 수 있으므로 다른 자원을 추가로 요구하지 않는다.	**직관적 관리와 사용자 경험** 웹 UI와 데스크톱 앱을 지원하므로 브라우저만 지원된다면 포털 환경에서도 로그인만으로 관리자와 사용자가 모두 시스템을 관리할 수 있다.
사전 정의된 AI 개발환경 제공 데이터 과학자들은 대부분 선호하는 개발 환경이 있다. 'Backend.AI'에서는 Tensorflow, Pytorch 등 다양한 개발 환경을 사전 정의된 이미지로 제공하므로, 원하는 환경을 클릭하기만 하면 곧바로 이용할 수 있다.	**AI 및 HPC 성능 최적화** HCSF와 스토리지 프록시 기능을 통해 성능 인티그레이션이 되어있어 독자적 엔진으로 최적의 GPU 연산자원을 배치하고 구현할 수 있다.

'Backend.AI'의 가장 큰 장점은 컨테이너 기반으로 GPU를 스케일링할 수 있다는 점이다. 자체 개발한 쿠다(CUDA) 기반 가상화 기술 덕분이다. 이를 통해 교육 및 추론 워크로드를 위한 단일 GPU를 공유할 수 있고, 모델 학습 등 대규모 워크로드를 위한 다중 GPU도 할당할 수 있다.

연산자원 성능, NVIDIA DGX와 HGX로 최적화

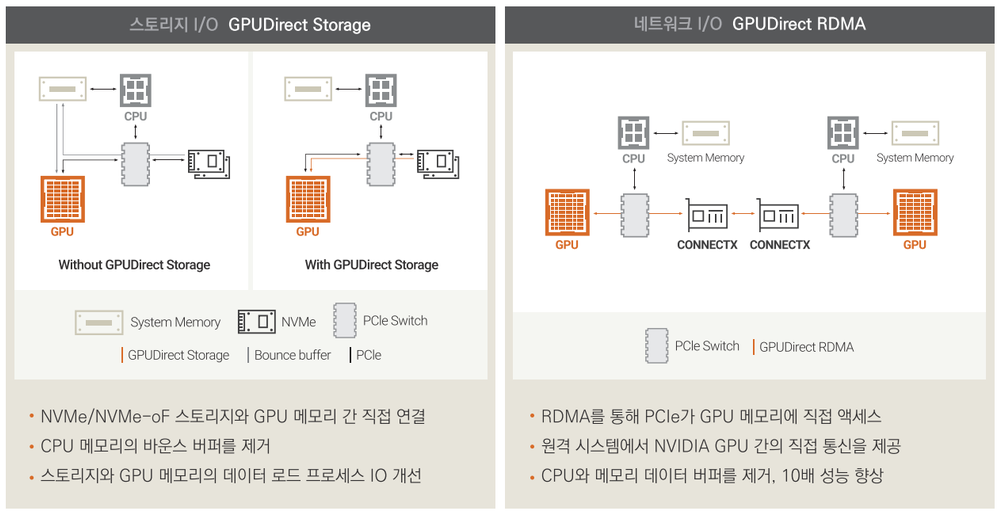
AI 플랫폼을 기획할 때 중요한 부분 중 하나는 연산자원의 성능이다.

최근의 GPU 시스템은 PCIe 방식보다 NVIDIA 보드 일체형을 많이 선호하고 있다. 4세대 NVIDIA NV링크는 900GB/s GPU 대역폭으로, PCI Gen4에 비해 14배 빠른 성능을 자랑하며, NVIDIA A100 Tensor 코어 GPU의 고속 상호 연결을 구현한다. 20대 이상 규모의 대형 GPU 팜을 구축하는 경우라면 이는 반드시 검토해야 할 요소다.

GPU 성능 최적화를 위한 요소 기술로는 스토리지 I/O인 'GPUDirect 스토리지', 네트워크 I/O인 'GPUDirect RDMA', 고속 네트워크인 'Infiniband 구성'을 들 수 있다.

GPUDirect 스토리지는 NVMe/NVMe-oF 스토리지와 GPU 메모리를 직접 연결하므로 CPU 메모리의 바운스 버퍼를 제거한다. 네트워크 I/O인 'GPUDirect RDMA'는 RDMA를 통해 PCIe가 GPU 메모리에 직접 액세스하므로 CPU와 메모리 데이터 버퍼가 제거돼 성능을 10배까지 향상할 수 있다. 또 GPU 연산 성능과 고성능 저장자원의 성능 확보를 원한다면, 장치 간 100G/200G 고속 네트워크 구성이 필수적으로 요구된다.

↓ GPUDirect Storage/RDMA



효성인포메이션시스템의 AI 플랫폼 ↑

금융권의 AI 도입, 기업별 상황에 따른 AI 솔루션 제안

금융권에서 AI를 도입하려는 기업의 상황은 모두 다를 수밖에 없다. GPU 인프라와 AI 솔루션을 이미 도입해 사용하고 있는 기업이 있지만, AI를 처음 도입해 성과가 불확실한 상태에서 사업 방향성을 고민하는 기업도 있다. 또 이미 대형 GPU 팜을 구축해 본격적으로 AI를 적용하는 기업도 적지 않다.

효성인포메이션시스템의 목표는 기업이 현재 처한 상황에 따라 그에 맞는 AI 솔루션을 제공하는 것이다. 따라서 AI 솔루션을 이미 도입해 운영 중인 기업에는 기존 자원을 잘 활용하는 데 집중할 수 있는 방안을, AI에 첫발을 떼는 기업에는 AI를 통해 기업이 원하는 실질적인 ROI를 도출할 수 있는지를 점검할 수 있도록 AI 테스트 베드 환경을 제안하고 있다. 그리고 본격적인 AI 적용을 위해 대형 GPU 팜을 구축하려는 기업에게는 효성인포메이션시스템의 풀 패키지 AI플랫폼 도입을 제안한다.

기업의 현재 상황에 맞는 AI 솔루션을 고민 중이라면, 고성능 기술에 대해 설계부터 구축, 운영, 지원까지 완벽하게 통합 지원하는 HCSF를 비롯해 AI 프로세스의 단계별 솔루션을 모두 갖춘 효성인포메이션시스템이 최고의 파트너라고 자부한다.