고성능 AI 분석, 기대를 현실로 차세대 데이터 레이크 아키텍처

강준범 / 효성인포메이션시스템 데이터사업팀 컨설턴트

'Dall-E'와 '챗GPT(ChatGPT)'은 아마도 올해 가장 핫한 키워드일 것이다. 두 가지 서비스 영역은 그림과 텍스트로 각각 다르지만, 공통점이 하나 있다. 고성능 분석 환경에 기반한 AI 모델이라는 점이다. 멀게만 느껴지던 AI 분석은 어느새 우리 일상 깊숙이 들어와 버렸다.

데이터가 기하급수적으로 증가하는 요즘, 이러한 고성능 분석 환경은 기업에 더욱 필요해졌다. 그리고 고성능 분석 환경의 핵심에는 AI 분석을 위한 필수 요건인 고성능 데이터 레이크가 자리하고 있다.

데이터 웨어하우스부터 고성능 데이터 레이크까지

데이터 분석에 기반한 인사이트를 도출하는 일은 기업들이 이미 오래전부터 해오던 일이다. 초창기 IT 시장에 등장한 RDBMS 기반 데이터 웨어하우스가 그 시작이다. 당시에는 형태가 정해져 있는 정형 데이터를 기반으로 인사이트를 도출하고자 정보계 시스템을 구축했는데, 정보계 시스템을 기반으로 한 정형 데이터 기반의 데이터 웨어하우스 환경이 바로 그것이다.

그러나 IT가 발전하면서 비정형/반정형 데이터, 실시간성 데이터, IoT 센서 데이터 등으로 데이터 형태가 다양해졌다. 기존의 정형 데이터 기반 정보계 시스템은 구조상 정형 데이터가 아닌 다양한 형태의 데이터를 저장하고 분석하기에는 적합하지 않을 뿐만 아니라, 데이터 양도 방대하게 급증해 기존의 정보계 시스템만으로는 감당할 수 없었다. 더 정확한 데이터 분석을 위해, 다양한 형태의 데이터를 수집하고 분석할 수 있는 환경이 필요하게 되었고, 데이터 레이크가 그 갈증을 해소해 주었다. 정형 데이터 분석을 위한 정보계 시스템과 비정형/반정형 데이터 분석을 위한 하둡 기반의 데이터 레이크 아키텍처는 이렇게 시작되었다.

다만, 하둡 기반 데이터 레이크 아키텍처도 문제가 있었다. 정형 데이터 저장을 위한 기존의 정보계 시스템, 비정형/반정형 데이터를 저장하기 위한 하둡 저장소와 NoSQL과 같이 다양한 형태의 데이터를 수집하기 위해 여러 형태의 저장소로 이뤄졌기 때문이다. 이는 데이터 중복과 데이터 사일로, 데이터 간 연관관계 도출의 어려움 등의 문제가 발생했으며, 하둡 분석 환경인 에코시스템에 대한 Dependency와 하둡 노드 증설에 대한 비용 문제 역시 나타났다.

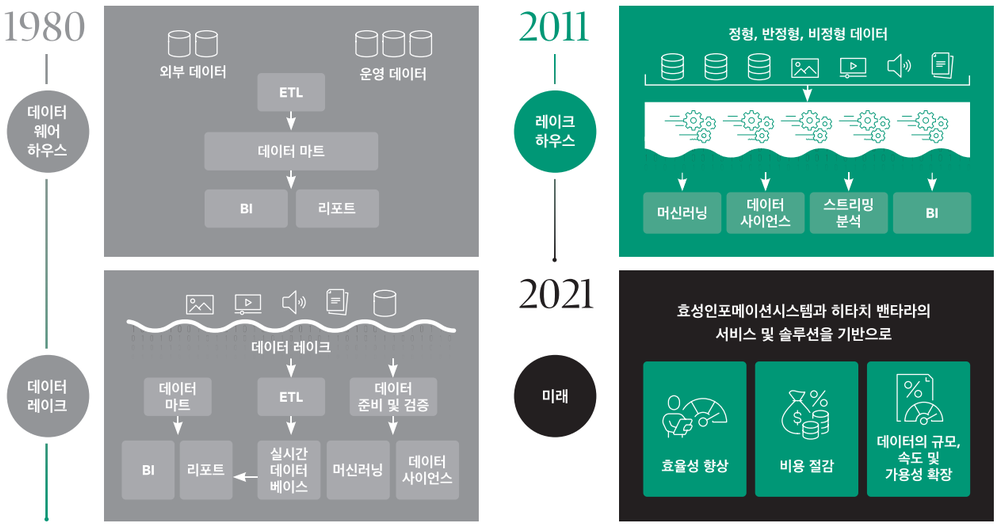
이를 해결하기 위해 등장한 아키텍처가 오브젝트 스토리지 기반의 단일 저장소인 '차세대 데이터 레이크 아키텍처'다.

이 과정에서 가장 큰 역할을 한 건 아마존의 AWS다. 아마존이 S3와 함께 다양한 분석 도구를 출시한 덕분에 S3 API를 통해 S3 오브젝트 스토리지 기반 퍼블릭 클라우드 상의 데이터 레이크 아키텍처를 수립할 수 있었다. 그리고 스토리지 벤더들도 여기에 맞춰 A3 API를 지원하는 오브젝트 스토리지를 출시하면서 S3 호환 오브젝트 스토리지가 차세대 데이터 레이크의 1차 저장소로 자리잡게 되었다.

데이터 레이크에 저장되는 데이터의 양이 기하급수적으로 증가하다 보니, 많은 기업에서는 좀 더 정확하고 빠른 분석을 위해 GPU 기반 데이터베이스, GPU 딥러닝 등 고성능 분석 환경을 도입했다. 그리고 이러한 고성능 분석 환경이 요구하는 저장소의 성능요건을 맞추기 위해 AI 분석을 위한 고성능 데이터 레이크를 위한 고성능 병렬 분산 파일 시스템 저장소가 필요하게 되었다. 스토리지 Layer의 병목 현상을 해결하면서 무제한 급의 확장이 가능한, 지금과 같은 병렬 분산 파일 시스템이 차세대 고성능 데이터 레이크 아키텍처로 자리잡게 된 것이다.

↓ 데이터 레이크의 발전

현대적인 데이터 아키텍처



(출처: 히타치 밴타라)

AI 분석을 위한 차세대 고성능 데이터 레이크의 조건

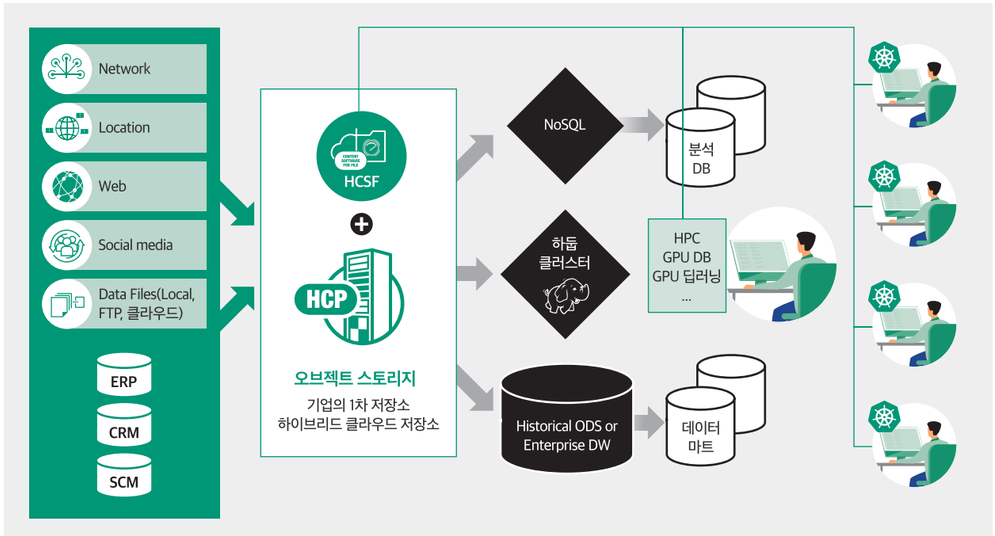
그렇다면 AI 분석을 위한 차세대 고성능 데이터 레이크는 어떤 조건들을 갖춰야 할까?

첫째는 자동 티어링(Auto-tiering)이다. AI 분석에서는 대부분의 데이터를 핫-웜-콜드(Hot-Warm-Cold) 데이터로 분류하고, 과거 데이터는 비용 효율을 높이기 위해 별도의 저장소로 아카이빙한다. 이때 콜드 데이터(cold data)로 아카이빙된 데이터를 분석에 활용하려면 고성능 스토리지로 다시 저장하는 과정을 거쳐야 한다. 반면, 자동 티어링 기능이 있으면 이 과정이 생략된다.

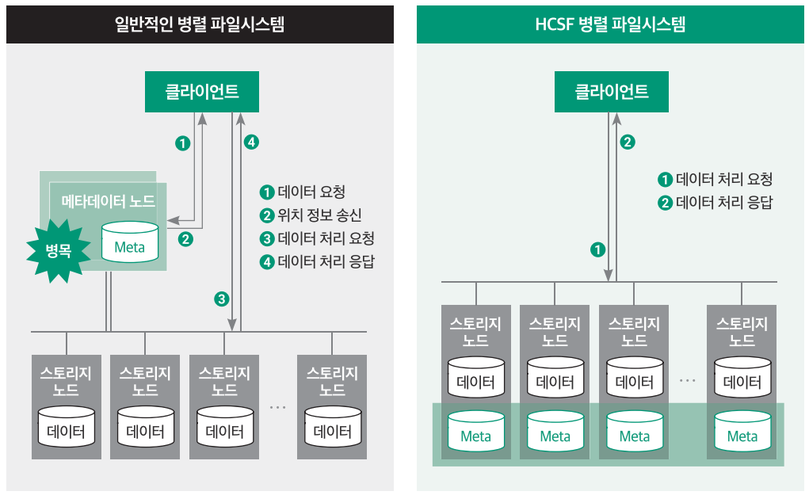
두 번째 요건은 멀티 프로토콜 지원이다. 자동 티어링 구조를 갖춘 스토리지라 하더라도, 분석에 쓰일 분석 도구를 미리 한정해선 안 된다. 끊임없이 새로운 분석 방법과 분석 도구가 출시되기 때문에 멀티 프로토콜 지원은 필수 요소다.

세 번째, 높은 IOPS와 Throughput의 보장이다. 딥러닝IO는 하나의 큰 데이터 파일이 아니라 데이터셋을 잘게 쪼개 사용하므로 Small IO 요청이 빈번하게 일어난다. 따라서 높은 IOPS가 필수적으로 요구되고, 저장소가 메타데이터 오버헤드를 최소화할 수 있는 구조를 갖춰야 한다.

↓ 차세대 고성능 데이터 레이크 아키텍처



↓ HCSF 병렬 파일 시스템의 메타데이터 구조



최신 엔터프라이즈 워크로드에 적합한 초고성능 솔루션 'HCSF'

HCSF(Hitachi Content Software for File)는 가장 빠르게 성장하고 있는 초고성능 병렬 파일 스토리지로 AI/ML, 고성능 데이터 분석 등 최신 엔터프라이즈 워크로드에 적합하다. 최첨단 클라우드, 컴퓨팅, 스토리지, 고속 네트워킹 기술이 적용되어, 최대 EB(엑사바이트) 규모의 데이터라도 필요할 때마다 신속하게 액세스가 가능해 데이터의 가치를 충분히 실현할 수 있다. 이를 통해 기업은 데이터를 활용해 혁신을 추진할 수 있고, 새로운 시장 기회를 모색하며, 신제품 출시 주기를 단축할 수 있다.

HCSF의 특장점은 다음 몇 가지로 요약할 수 있다.

01 | 단일 네임스페이스로 자동 티어링

HCSF는 핫티어(hot-tier)와 콜드티어(cold-tier) 간 자동 티어링이 가능해 데이터가 어디에 있든 분석가가 호출하면 즉각 로드된다. 자동 티어링 기능이 없는 일반적인 병렬 분산 스토리지는 티어 간 데이터를 이동하는 과정을 거쳐야 하므로 분석가가 데이터를 손에 넣기까지 적지 않은 시간이 소요된다. 자동 티어링은 이 문제를 간단히 해결해 주고, 인프라 담당자 역시 별도의 아카이빙 서버를 도입해 관리하지 않아도 된다.

02 | 빈번한 데이터 복제로 인한 병목현상 제거

일반적인 병렬 파일 시스템은 대부분 NAS 스토리지 기반의 변형된 솔루션이다. 그러나 특허 받은 기술에 기반한 HCSF는 설계될 때부터 AI 분석에 특화된 솔루션이다. 스토리지 노드가 증설되면 메타데이터 서버 역할을 하는 프로세스도 같이 증설되는 MSA(Micro Service Architecture) 구조로, HCSF에서는 노드 증설로 인해 성능 저하가 발생하지 않는다. 데이터 레이크를 도입하는 기업들이 처음부터 수백TB, 수백PB 급의 대규모로 시작하는 것은 아니다. 기본 100TB로 시작해 점진적으로 데이터 레이크를 증설하기 때문에 노드 증설에 따른 선형적 성능 향상이 무엇보다 중요하다.

03 | IOPS와 Throughput 동시에 극대화

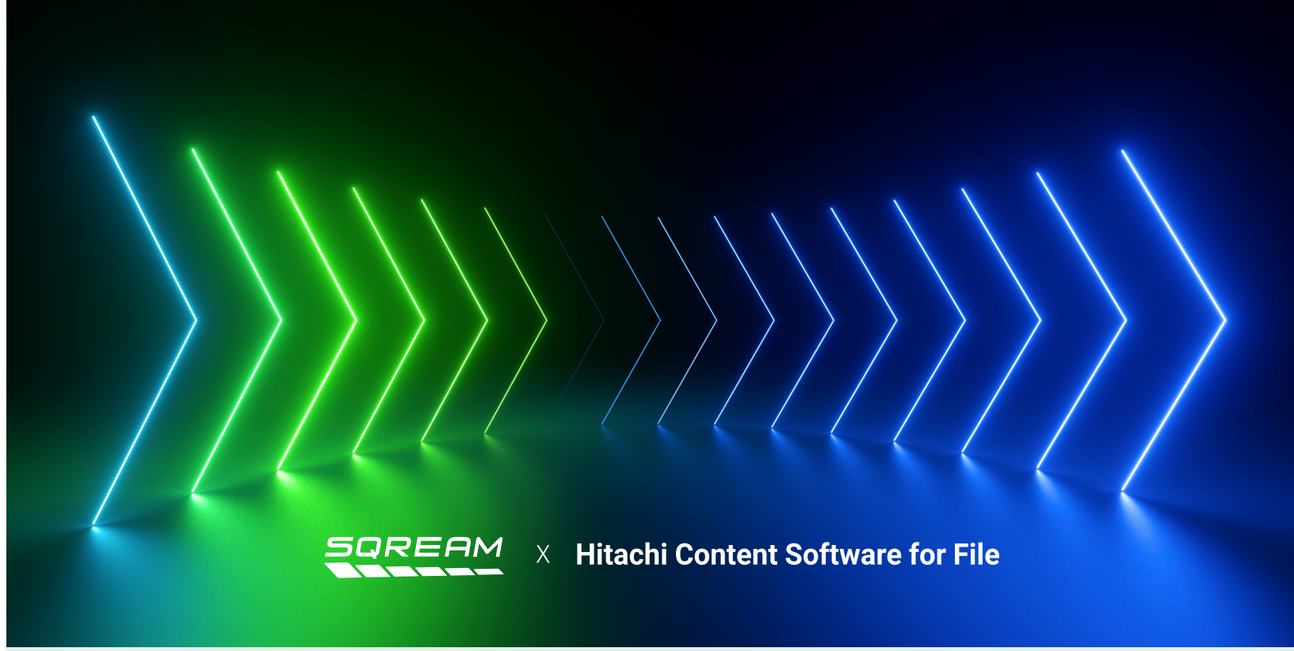
HCSF는 분산해서 저장되는 블록 사이즈가 4K로 AI/ML과 같은 IO 집약적인 혼합 워크로드에 적합한 솔루션이다. 딥러닝에서 가장 중요한 요건인 스몰 파일(small file)에 대해 높은 IOPS, 높은 Throughput, 짧고 일정한 지연시간을 동시에 만족시킨다.

04 | 다양한 멀티 프로토콜 지원

고성능 데이터 레이크 저장소의 요건은 기업이 사용하는 분석 및 수집 도구에 대해 개방적이어야 한다. 즉 멀티 프로토콜을 지원해야 한다는 의미다. 저장소에 사용되는 분석 도구가 어떤 것이든 유기적으로 연동되어야 하기 때문이다. 현재 가장 많이 사용하는 프로토콜은 POSIX, NFS/SMB/S3 프로토콜, GPU 분석 환경을 위한 GPU다이렉트 스토리지(GPUDirect storage), 쿠버네티스 CSI 등이다.

05 | 효율적 자원 운용 가능

분석가가 직접 스토리지까지 컨트롤할 수 있는 수준의 손쉬운 구성, 장애 처리, 모니터링이 가능하다는 점은 HCSF의 최대 장점 중 하나다.



세계에서 가장 빠른 파일 시스템과 SQL 플랫폼의 만남

서로 다른 시스템과 스토리지를 도입하지 않고도 TB에서 PB까지 선형으로 가용 용량 및 성능의 확장 가능한 솔루션을 위해 히타치와 SQream이 힘을 합했다. 세계에서 가장 빠른 파일 시스템에 가장 빠른 GPU 기반 고성능 SQL 분석 엔진을 장착한 분석 및 데이터 플랫폼인 'SQreamDB'가 그 주인공이다.

SQreamDB는 방대한 양의 데이터에서 신속하게 비즈니스 인사이트를 도출할 수 있도록 설계된 GPU 기반 SQL 분석 플랫폼이다. 데이터 취합부터 쿼리에 이르기까지 PB 규모의 데이터까지도 복잡한 분석을 빠르게 수행하며, 온프레미스 또는 프라이빗 클라우드에서 선형으로 확장되어 인사이트 도출 시간이 최소화되고, TCO도 절감할 수 있다.

SQreamDB의 GPU 고성능 연산이 가능한 것은 최신 네트워크 패킷 처리 기술이 적용된 초고성능 NVMe 병렬파일 시스템인 HCSF의 성능과의 시너지라 말할 수 있다.

HCSF는 속도가 느린 커널 기반 디바이스 드라이버를 거치지 않고, 네트워킹과 스토리지 디바이스를 직접 관리하기 때문에 CPU와 GPU의 IO 대기시간이 대폭 줄어든다. 즉, 더 짧은 시간에 더 적은 리소스로 더 많은 작업을 수행할 수 있다.

세계에서 가장 빠른 GPU 가속화 SQL 플랫폼으로 인정받고 있는 SQreamDB와 초고성능 병렬 파일 시스템과 오브젝트 스토리지가 하나로 통합된 HCSF와 함께 제공됨에 따라 기존의 빅데이터 플랫폼에 비해 훨씬 저렴한 비용으로 최고 수준의 확장성 확보, 엔터프라이즈급 데이터 보호, 데이터 라이프사이클 관리가 가능해졌다.

*출처: GPU Accelerated Insights with Hitachi Content Software for File and SQream, www.hitachivantara.com, 2023년 1월

데이터 레이크 구축으로 비즈니스 경쟁력 Up!

다양한 산업 분야에서 데이터 레이크를 활용해 비즈니스 환경을 업그레이드한 사례를 소개한다.

01 | 글로벌 바이오 연구소

A 연구소는 입자가 손상되지 않는 최단 시간 동안 이미지 쓰기와 3D 시각화를 위한 대량의 읽기 IOPS가 필요했다. 연구소에서 사용하는 방식은 저온 전자현미경법으로, 생체 입자의 수용액을 섭씨 약 -190도의 초저온으로 냉각시킨 후 극히 짧은 시간 내에 전자파를 쏘면서 수천, 수만 장의 2D 이미지를 생성하고, 이 이미지로 3D 모델링을 수행한다.

지극히 짧은 시간에 디스크에 쓰기를 수행하고, 3D 모델링에서 데이터를 다시 읽어야 하기 때문에 정교한 모델링과 병목 현상을 유발하지 않는 IOPS가 무엇보다 중요했다.

이 연구소에서 HCSF가 빛을 발한 건 다양한 이미지의 사이즈를 신경 쓸 필요 없는 제로 튜닝 기법 덕분이다. 방대한 양의 이미지가 저장되지만, 제로 튜닝 덕분에 이미지 저장 방식 등은 신경 쓸 필요가 없게 됐다.

02 | 국내 대형 제조기업

대형 제조기업 B사는 데이터 웨어하우스/하둡 데이터 분석 시스템을 이용하고 있으나, 확장성과 성능 저하가 가장 큰 문제였다. 이를 해결하기 위해 전사 통합 저장소를 구축해 차세대 전사 데이터 분석 체계로 전환했다. B사의 데이터 분석 목적은 제조기업에 가장 중요한 이슈라고 할 수 있는 수율 관리다. 수율은 불량품 관리 측면도 있지만 고객 신뢰도 측면에서도 매우 중요한 문제이기 때문이다.

현장에서 발생하는 데이터에 대해 고성능 데이터 분석을 기반으로 대용량 쿼리가 가능하도록 했으며, 향후 AI/ML을 위한 전사 분석 체계도 마련했다. 공정 데이터, 환경 데이터, 생산관리 데이터, 이미지 데이터에 대해 실시간 또는 준실시간으로 쿼리를 수집하고, 거기에 맞춰 티어1, 2로 나누어 티어1으로 수집할 수 있도록 했다.

현재는 고성능 데이터 분석 기반과 각기 다른 데이터 인터페이스 클라이언트 환경을 지원하는 통합 스토리지인 오브젝트 스토리지가 NVMe 티어링 용도와 비정형 데이터 서비스 용도 두 가지로 나누어 동시에 운영되고 있다. 2021년에 1차 구축 완료 후 지난해 추가 증설까지 진행되었으며, 앞으로도 계속 증설 요건이 발생할 것으로 예상된다. 구체적인 성과로는 쓰기 100GB 이상, 읽기 470GB 이상의 성능을 나타냈으며, 특히 IOPS가 읽기의 경우 2천만 건 이상을 지원한다는 점을 들 수 있다.

03 | 자율 주행 전기차 업체

C사는 클러스터와 인프라 규모가 상상할 수 없을 정도로 큰 자율 주행 전기차 업체로, 10개의 GPU 클러스터에서 ML(머신러닝) 모델을 개발했다. 그러나 스토리지에서 GPU로 데이터를 전송할 때 심각한 병목 현상이 발생해 속도가 느려지고 있었다. 딥러닝을 할 때는 수PB의 데이터로 동시 학습을 해야 하는데, 성능이 미치지 못해 ML 모델링에 어려움을 겪고 있는 것이었다.

원인 파악이 제대로 되지 않았던 초기에는 GPU 서버 증설, NVMe 연계 등 여러 가지 방안을 시도했으나 충분한 효과를 보지 못해 마지막으로 병렬 분산 파일 시스템을 선택했다. C사는 2018년에 6PB 단위의 학습 클러스터 3개로 시스템을 런칭했는데, 병렬 분산 파일 시스템을 도입한 이후 기존의 2주 주기가 4시간 주기로 단축되었으며, IOPS도 증가했다.

04 | 방송/엔터테인먼트 기업

D사는 애니메이션 스튜디오를 운영하면서 동시에 글로벌 OTT도 서비스하는 기업이다. 방송/엔터테인먼트 업계는 최근 들어 가장 많은 변화를 겪고 있는 산업으로, 몇 년 전만 해도 시간당 용량이 수십 GB 정도에 불과했지만 지금은 수십 TB까지 증가했다.

소비자의 눈높이에 맞춰 색, 명암 대비 등을 현실에 가깝게 표현하려면 아티스트들이 데이터를 끌어와 컬러를 수정하고 렌더링을 해야 한다. 하지만 기존 시스템의 성능이 저하돼 작업 효율성이 떨어지고 있다는 점이 문제였다. 그래픽 편집 작업을 위한 이미지 읽기 성능을 강화하는 것이 가장 중요한 목표가 됐다. IOPS를 제대로 제공하는 스토리지가 절실한 상황이었다.

이를 해결하기 위해 1단계로 500TB NVMe를 구축해 최적의 짧은 지연시간을 확보하고, 업무 효율성을 향상했다. 대부분의 애플리케이션이 맥 또는 윈도우에서 동작하기 때문에 주요 프로토콜은 SMB-W를 사용하고 있다. 현재는 600PB 규모의 용량을 처리하고 있으며, 3조 개(평균 2MB 파일)의 오브젝트에 대해 안정적인 서비스를 운영하고 있다.