Building the Archive of the Future

미래를 위한 아카이브 구축하기

유재근
퀀텀코리아 이사

지난 30년 동안 디지털 데이터가 인간 생활의 모든 측면에서 중요해짐에 따라 끊임없는 데이터 증가는 불가피하며, 향후 30년 동안 훨씬 더 큰 역할을 할 것으로 예상된다. 이러한 데이터의 대부분은 영구적으로 저장될 것이므로 더욱 지능적이고 보안성이 뛰어난 장기 스토리지 인프라의 출현이 요구된다. 보존 요구 사항은 데이터 유형에 따라 크게 다르지만 모든 비즈니스에서 아카이브 데이터가 빠르게 쌓이고 있다. 이러한 잠재적 가치를 고려할 때, 최신 데이터 아카이빙은 백업을 넘어 기업의 핵심 전략이자 하이퍼스케일 데이터센터의 필수 분야가 되었다.

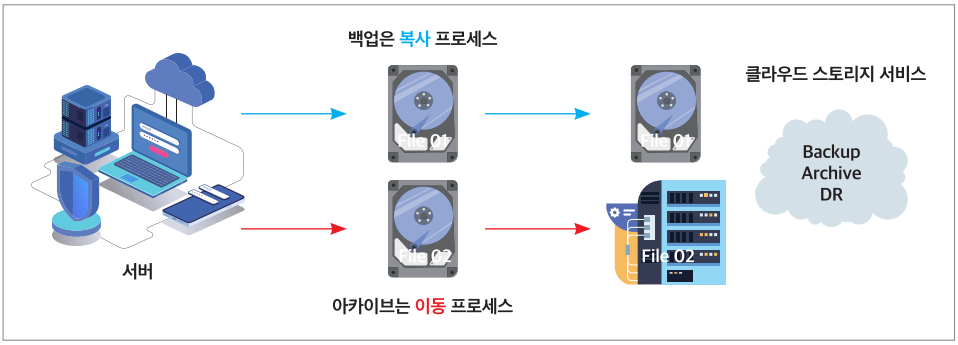
많은 데이터 유형이 언젠가 그 가치가 실현될 것으로 기대하며 무기한 저장되고 있다.
업계 설문조사에 따르면 거의 60%의 기업이 50년에서 100년 이상 동안 점점 더 많은 양의 데이터를 디지털 형식으로 보관할 계획이며, 이 데이터의 대부분은 수정하거나 삭제하지 않을 것이라고 한다. 많은 조직이 처음으로 테라바이트, 페타바이트, 엑사바이트급 아카이브 데이터에 직면하게 되면 전체 스토리지 전략과 인프라를 재설계해야 할 수 있다. 다행히도 오늘날의 아카이브 과제를 관리하는 데 필요한 기술과 아키텍처를 사용할 수 있다.

아카이브 데이터가 중요해진 이유

대부분의 아카이브 데이터는 수익화되지 않아 그 진정한 가치를 알 수 없었지만, 기업들은 이제 막 미개발 디지털 아카이브의 잠재적 가치가 크다는 사실을 깨닫고 있다. 데이터의 가치를 실현하려면 데이터를 활용해야 한다. 지금부터 2025년 사이에 경쟁력을 확보하고자 하는 기업은 아카이브 데이터가 조직의 성공에 얼마나 중요한 역할을 하는지, 그리고 그 기간 동안 데이터 아카이빙 전략이 어떻게 발전할 것인지 이해해야 한다. 아카이빙 팬데믹은 전 세계 데이터 영역으로 빠르게 확산되고 있다. 이러한 추세를 고려할 때 아카이브 스토리지 패러다임은 새롭게 재창조되어야 할 것이다.

백업은 '복사', 아카이브는 '이동' 프로세스

많은 기업이 백업과 아카이브 프로세스를 혼동하여 같은 프로세스라고 생각하는 경우가 많다. 디스크 데이터 백업은 50년 이상 테이프의 초기 및 주요 사용 사례였지만, 이러한 사고방식은 데이터 아카이빙으로 전환되었다. 백업은 데이터 사본을 만드는 프로세스로서, 원본 사본이 손상되거나 손상된 경우 또는 데이터 손실 이벤트가 발생한 후 원본 사본을 복원하는 데 사용할 수 있다. 데이터가 중요할수록 복구 속도는 더욱 중요해진다.



백업 (데이터 복사)	아카이브 (데이터 이동)	활성 아카이브 (아카이브에 빠르게 액세스)
보호 및 복구를 위해 데이터를 복사하고, 원본 데이터는 그대로 유지	자주 사용하지 않는 데이터를 보다 비용 효율적인 스토리지로 이동하여 소스 디바이스의 공간 확보	지능형 소프트웨어, SSD, HDD, 테이프 시스템의 결합 솔루션
데이터 손실 시 원하는 시점으로 파일 복원, 속도 중요	향후 참조 및 분석을 위해 파일을 검색. 검색 속도는 중요하지 않음	대용량 테이프 라이브러리를 위한 캐시 프론트엔드로 HDD 또는 SDD 사용
주기적 프로세스, 보존 기간 종료 시 자체 덮어쓰기	영구적으로 증가하며 일반적으로 변경되지 않고 덮어쓰지 않음	파일 및 개체 수준에서 항상 보관 날짜에 대한 액세스 제공
단기간 1일 ~ 120일	영구적이고 장기적인 데이터를 변경으로부터 보호	아카이브 데이터에 더 쉽게 액세스하여 워크플로우 개선

아카이빙은 더 이상 활발하게 사용되지는 않지만 장기 보관을 위해 다른 물리적 위치에 안전하게 보관해야 하는 데이터를 이동하여 원본 위치에서 더 많은 비용이 드는 공간을 확보하는 프로세스다. 오늘날 대부분의 애플리케이션은 아카이브 데이터를 읽기 전용으로 처리하여 수정되지 않도록 보호한다.

아카이빙으로 백업 기간에 대한 부담 감소

연구에 따르면 조직 데이터의 85%는 역사적으로 가치가 있지만, 액세스하는 경우는 드물고 삭제되는 경우는 거의 없는 것으로 조사됐다. 이러한 데이터의 60%는 일반적으로 운영 비용이 많이 드는 디스크 드라이브에 저장된다. 아카이빙을 사용하면 백업 세트에서 활동량이 적고 변경되지 않는 데이터를 대부분 제거하여 백업(및 복원) 프로세스의 속도를 높이고 이 과정에서 값비싼 디스크 스토리지 용량을 확보할 수 있다. 중복 제거와 같은 디스크 백업 프로세스가 도움이 될 수 있지만, 엔터프라이즈 데이터 증가율이 매년 25~30%에 달하면서 백업 기간의 증가는 여전히 주요 문제로 남아 있으며 지속적인 압박을 받고 있다. 액티브 아카이브 구현은 HDD(하드 디스크 드라이브) 또는 SSD(솔리드 스테이트 디스크 드라이브)를 로봇 테이프 라이브러리의 캐시 프런트 엔드로 사용하여 아카이브 데이터에 더 빠르게 액세스할 수 있도록 한다. 아카이브 규모가 커질수록 액티브 아카이브는 더 많은 이점을 제공한다.

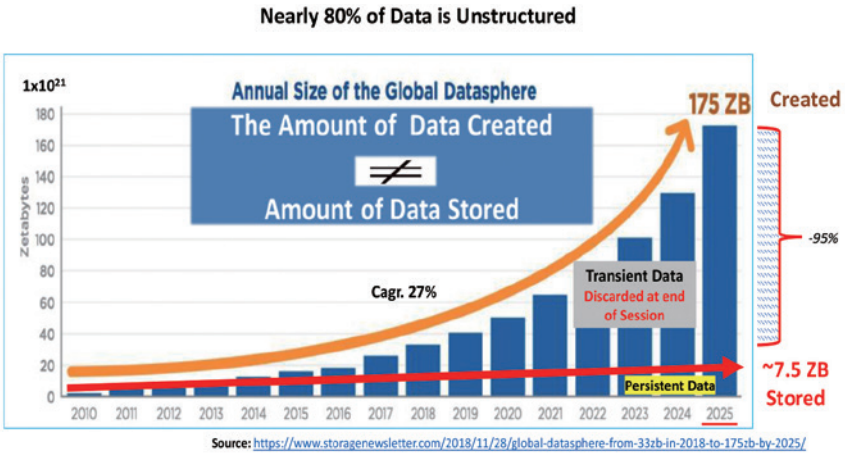
아카이빙 되고 있는 데이터량

불과 20년 전 1년 동안 생성되는 데이터보다 현재 시간당 생성되는 데이터가 더 많다. 2025년까지 생성되는 데이터의 총합은 175ZB에 달할 것으로 예상되지만, 생성되는 대부분의 데이터의 수명이 짧고 일시적이기 때문에 실제로 저장되는 데이터는 약 7.5ZB에 불과할 것이다. 전체

데이터의 60~80%는 아카이브 데이터이며, 이 중 상당수가 잘못된 장소, 즉 HDD에 저장되어 2025년까지 총 4.5~6ZB의 아카이브 데이터가 저장되어 아카이브가 가장 큰 분류 카테고리가 될 것으로 예상된다. 아카이브 데이터를 비용 효율성이 높은 테이프 스토리지 시스템에 올바르게 할당하면 상당한 비용 절감 기회를 얻을 수 있다.

여기서 주목해야 할 점은 현대의 아카이브는 대부분 비정형 데이터라는 점이다

최신 아카이브 데이터는 주로 검색이 쉽지 않은 비정형 데이터로, 사무용 문서, 비디오, 오디오, 이미지, 과거 기록 등 기본적으로 데이터베이스에 없는 모든 것을 포함한다. 오브젝트 스토리지는 비정형 데이터 저장, 대규모 데이터 세트 저장, 사용자 지정 데이터 보존, 삭제, 보존 정책으로 데이터를 저장하는 데 더 적합하기 때문에 선호되는 아카이브 형식이 되고 있다. 빅데이터는 대부분 비정형 아카이브 데이터로 구성되어 있어 기존 방법으로는 검색 및 분석이 어렵다. 다행히도 빅데이터를 분석하는 많은 도구가 메타데이터, 태그 및 카탈로그를 활용하여 아카이브 데이터를 더 쉽고 빠르게 검색하고 액세스할 수 있도록 지원하기 시작했다.

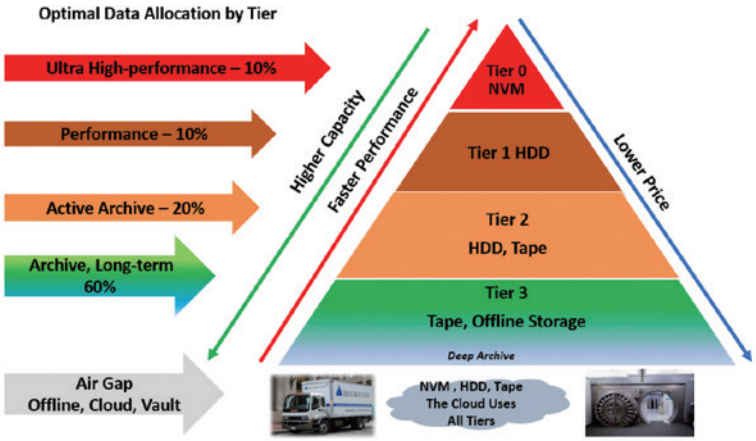


글로벌 데이터스피어 확장은 끝이 없다. (데이터의 약 80%가 비정형 데이터)
글로벌 데이터스피어의 연간 규모 : 생성되는 데이터의 양 ≠ 저장된 데이터의 양

데이터 분류 가이드라인

데이터 분류 프로세스는 데이터를 효과적으로 관리하기 위해 매우 중요하며, 스토리지 풀이 커질수록 그 중요성은 더욱 커진다. 원하는 만큼 스토리지 계층을 정의할 수 있지만, 데이터를 분류하는 사실상의 표준 계층은 일반적으로 네 가지가 사용된다: 초고성능 데이터(OLTP), 성능 데이터(미션 크리티컬), 활성 아카이브(낮은 활동 데이터), 그리고 가장 크고 빠르게 성장하는 스토리지 세그먼트인 장기 아카이브다. 데이터 분류를 통해 데이터 특성을 최적의 스토리지 기술 계층에 맞게 조정하고 매핑할 수 있다. 가능한 한 많은 데이터를 가장 비용이 적게 드는 스토리지 계층으로 옮기는 것이 최신 아카이빙 전략의 핵심 요소이며 가장 큰 경제적 가치를 제공한다.

이상적인 아카이브 솔루션은 수집 시 데이터를 분류하고 메타데이터를 생성한다. 아카이브 데이터의 가치를 실현하기 위한 4가지 주요 과제는 다음과 같다: 1) 수집 시 아카이브 데이터(분류, 색인, 카탈로그, 메타데이터)에 대한 접근성 확보, 2) 장기 아카이브 스토리지 인프라 관리,



데이터 분류 및 저장 계층
계층별 최적의 데이터 할당 (초고성능 : 10% / 성능 : 10% / 액티브 아카이브 : 20% / 아카이브, 장기간 : 60%)

3) 필요한 아카이브 데이터만 실제로 저장, 4) 아카이브 데이터의 보안 및 가용성 보장이다.

테이프 르네상스에서 현대 테이프 시대로의 전환

1950년대 초 최초의 테이프 드라이브가 등장한 이래 테이프는 주로 디스크 데이터의 백업 장치로 사용되어 왔다. 가장자리 손상, 늘어남, 찢어짐, 로딩 문제, DAT, DDS, DLT, 8밀리미터 테이프와 같은 구형(현재는 사용되지 않는) 테이프 포맷의 미디어 정렬 문제 등 과거의 테이프 문제들은 성공적으로 해결되었다. 2000년에 이르러 레거시 테이프 시대가 막을 내리고 테이프 업계가 새롭게 등장하는 많은 데이터 집약적 애플리케이션을 처리하기 위한 새로운 기반을 구축하는 테이프 르네상스가 진행되면서 테이프의 초기 사용 사례(백업)가 아카이빙에 자리를 내어주기 시작했다.

과거의 모습 (레거시 테이프 시대)	테이프 르네상스의 시작	현재의 모습 (최신 테이프 시대)
• 다양한 포맷 및 파일 시스템 (Travan, 8mm, DLT, DDS, DAT...) • 엣지, 늘어남, 찢김, 압착 • 미디어 수명 *4~10년으로 • HDD보다 낮은 신뢰성 • '테이프는 죽었다' • 성공적인 로봇 라이브러리 등장 • 주요 애플리케이션 - 백업	• LTO와 LTFS(표준 테이프 및 파일 시스템 - 상호 교환) • 30+년 미디어 수명을 위해 MP에서 Bafe(산화) 미디어로 전환 • 최고의 신뢰성을 위해 HDD에서 PRML ECC 차용 • HDDS에서 GMR 헤드 차용 • 서보 트랙을 테이프 가장자리에서 중간 대역으로 이동 • 견고한 카트리지	• HDD보다 높은 신뢰성(BER) • HDD보다 2배 이상 빠른 데이터 속도 • 30년 이상의 미디어 수명 • 엑사바이트+(1x10¹⁸) 라이브러리 - 에어 갭 • RAIT, RAO 및 TAOS 어라이브 (성능) • 최저 에너지 소비 및 TCO • 지능형 로봇 공학 • 주요 애플리케이션 - 아카이브

테이프 르네상스, 과거와는 전혀 다른 오늘날의 테이프

인터넷, 하이퍼스케일 데이터 센터, 클라우드, 빅 데이터, 원격 의료, 규정 준수, 분석 및 IoT의 물결로 인해 전례 없는 데이터 증가가 예상되는 지금이 바로 고급 테이프 기능의 적기라고 할 수 있다. 다음과 같은 기능을 제공하는 최신 테이프 시대가 도래했다.

- 테이프는 디스크보다 구입 비용이 저렴($/TB) - 테이프는 디스크보다 소유 및 운영 비용이 5~8배 낮아져 TCO를 절감할 수 있다
- 테이프는 HDD의 신뢰성을 3배 이상 능가한다

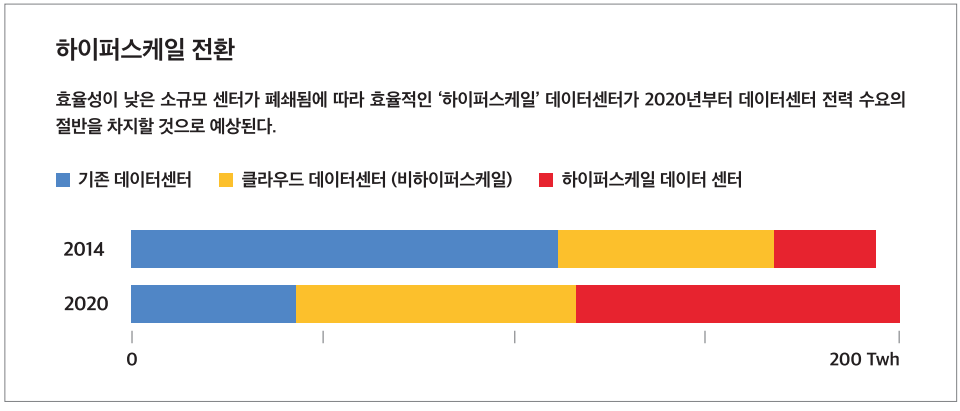
- 최신 테이프의 미디어 수명은 모든 새로운 미디어에 대해 30년 이상이다
- RAIT를 통해 테이프 드라이브 성능(처리량)을 향상시키고, RAIL을 통해 가용성을 높이며, RAO 및 TAOS를 통해 파일 액세스 시간을 단축한다
- 테이프 라이브러리를 위한 지능적이고 빠르며 효율적인 로봇 이동
- 2:5 - 1 이상의 테이프 데이터 압축률
- 암호화, WORM 및 에어 갭 스토리지를 통한 보안 기능
- 보다 빠른 '디스크와 같은' 액세스를 위한 미디어 파티션이 있는 표준 개방형 테이프 파일 시스템인 LTFS
- 테이프 기술에 대한 10년간의 LTO 로드맵은 예측 가능한 한계가 거의 없이 잘 정의되어 있다

향후 10년을 내다보면, 많은 새로운 애플리케이션과 워크로드, 그리고 대부분의 대규모 하이퍼스케일 데이터 센터에서 데이터가 폭발적으로 증가함에 따라 테이프의 모멘텀은 더욱 커질 것이다. 테이프가 HDD나 SSD를 대체하지는 못하겠지만, 가까운 미래에는 사실상 표준 아카이브 아키텍처가 될 것이다. 오늘날 많은 데이터 센터 관리자들이 '데이터를 사용하지 않는다면 에너지를 소비해서는 안 된다'는 공통된 목표를 가지고 있다는 점은 테이프를 가장 친환경적인 스토리지 솔루션으로 자리매김하고 있다. 다양한 테이프 기술 개선이 이루어지면서 테이프는 앞으로 직면하게 될 엄청난 고용량 및 아카이빙 과제를 해결하는 가장 비용 효율적인 스토리지 솔루션으로 계속 자리매김할 수 있는 기반을 마련했다. 최신 테이프는 과거의 테이프와는 전혀 다르다.

클라우드 및 하이퍼스케일 데이터센터(HSDC)의 테이프 아카이브 솔루션 구현

2020년부터 가장 빠르게 성장하는 데이터센터 유형으로 전 세계적으로 570개에 달하는 HSDC가 구축되었다. 데이터 센터와 정보 기술은 현재 전 세계 전력의 약 2%를 소비하고 있으며 2030년에는 8%까지 급증할 것으로 예상된다. HSDC는 최신 아카이빙 전략의 진원지가 될 수 있다.

HSDC는 운영 예산과 에너지원을 소모하고 데이터 센터를 과밀하게 만드는 디스크 팜의 극복할 수 없는 성장에 직면하여 아카이빙 및 활동량이 적은 데이터를 테이프 솔루션으로 마이그레이션해야 하는 상황에 직면해 있다. 하이퍼스케일 환경에서는 '디스크를 추가하는 것은 전



술적이지만 테이프를 추가하는 것은 전략적'이다. 테이프 카트리지는 수명의 대부분을 라이브러리 슬롯이나 선반에 보관하며 테이프 드라이브에 장착되어 있지 않을 때는 에너지를 소비하지 않으므로 테이프는 이상적인 아카이빙 스토리지다. 이레이저(삭제) 코딩, 엑사스케일 용량으로 지리적 확산을 지원하는 동시에 최저 TCO, 최고의 신뢰성, 향상된 사이버 보안 보호를 제공하는 확장성이 뛰어난 고급 에어 갭 테이프 아키텍처는 앞으로 직면하게 될 막대한 HSDC 스토리지 과제를 해결하는 데 점점 더 중요한 역할을 할 것이다.

미래를 위한 아카이브 구축

미래의 성공적인 아카이브 아키텍처는 여러 위치에 분산된 엑사바이트급 용량을 저장, 보호, 보존, 검색하고 쉽게 확장할 수 있는 기능을 갖추게 될 것이다. 대부분의 기업과 서비스 제공업체는 자연재해, 사이버 공격, EMP 또는 기타 재난으로 인해 전체 데이터센터가 오프라인 상태가 되거나 액세스가 차단될 수 있기 때문에 단일 위치만으로는 고가용성 데이터 보호 전략을 제공하기에 충분하지 않을 수 있다.

미래 아카이브의 주요 구성 요소에는 스마트 데이터 무버, 데이터 분류 및 메타데이터 기능을 갖춘 액티브 아카이브 소프트웨어, 확장성이 뛰어난 테이프 라이브러리 기술, RAIL(지능형 라이브러리의 중복 어레이) 아키텍처, 이레이저(삭제) 코딩, 내결함성, 중복성 및 가용성을 높이기 위해 여러 위치의 구역에 데이터를 지리적으로 분산하는 기능이 포함된다.

구성 요소	기능
액티브 아카이브 소프트웨어	비정형 및 오브젝트 데이터를 확장하고 지리적으로 분산하여 아카이브 스토리지 요구 사항을 관리하고 보호하는 오브젝트 스토리지 소프트웨어다.
데이터 무버	스토리지 장치에서 데이터를 검색하여 네트워크 클라이언트에서 사용할 수 있도록 한다.
분류 - 카탈로그, 메타데이터 생성	조직의 모든 데이터 자산에 대한 상세한 인벤토리를 수집 시 생성하여 분석 또는 비즈니스 목적에 가장 적합한 데이터를 신속하게 배치, 보호 및 찾을 수 있다.
최적의 아카이브 스토리지	최적의 TCO, 안정성, 미디어 수명, 확장성, 최소한의 리마스터링, 최저 비용을 위한 최신 테이프다. 고성능 액티브 아카이브를 위한 대용량 HDD
RAIL	중복성 및 가용성 향상을 위해 여러 테이프 라이브러리에 걸쳐 데이터를 스트라이핑한다.
이레이저 코딩	데이터를 샤드로 분할하고 인코딩한 후 여러 위치(노드)에 저장하여 고가용성을 보장하는 패리티 기반 보호 체계다.
지리적 분산	여러 곳에 위치한 여러 물리적 스토리지 시스템에 데이터를 분산한다.

현대 아카이브 분석

전 세계의 기업, 정부, 사회, 개인이 데이터에 대한 의존도가 높아지면서 데이터 보존과 아카이빙이 중요한 IT 업무로 빠르게 자리 잡고 있다. 디지털 아카이브의 보존 규모는 현재 페타스케일(1×10^{15}), 엑사스케일(1×10^{18}) 수준에 도달했으며 가까운 미래에 제타스케일(1×10^{21}) 용량에 근접할 것으로 예상된다. 활동량은 적지만 잠재적으로 가치 있는 아카이브 데이터를 최적의 스토리지 계층으로 이동하는 전략은 데이터를 스토리지 계층에 맞게 조정하면서 스토리지 비용을 가장 크게 절감할 수 있는 방법이다. 새로운 아카이빙 기술이 등장하지 않는 한, 테이프의 수많은 개선 사항으로 인해 테이프는 가까운 미래에 가장 확실한 최적의 데이터 아카이빙 스토리지가 될 것이다. 이제 비용 효율적인 고가용성 아카이브 인프라를 구현하기 위한 하드웨어, 소프트웨어 및 관리 구성 요소가 갖추어져 있다.