DELL Technologies

기술 백서

Dell EMC ECS: 고가용성 설계

백서 소개

이 문서에서는 Dell EMC™ ECS 플랫폼이 엔터프라이즈 가용성을 제공하는 방법에 대한 자세한 아키텍처를 설명합니다.

2021 년 6 월

H16344.6

개정 내역

날짜	설명
2017 년 7 월	최초 릴리스
2017 년 8 월	ECS 버전 3.1 콘텐츠를 포함하도록 업데이트
2019 년 3 월	ECS 버전 3.3 콘텐츠를 포함하도록 업데이트
2020 년 4 월	'운영 중단 중 액세스가 활성화된 TSO 동작' 업데이트
2020 년 12 월	메타데이터 보호 방법 업데이트
2021 년 6 월	ECS 버전 3.6.1 콘텐츠를 포함하도록 업데이트

감사의 말

이 백서는 다음에 의해 작성되었습니다.

작성자: Zhu, Jarvis

본 출판물의 정보는 "있는 그대로" 제공됩니다. Dell Inc.는 본 출판물의 정보와 관련하여 어떠한 종류의 진술이나 보증을 하지 않으며, 특정 목적을 위한 상업성 또는 적합성에 대한 묵시적인 보증을 하지 않습니다. 본 문서에 설명된 소프트웨어를 사용, 복사 및 배포하려면 해당 소프트웨어 라이선스가 필요합니다.

본 문서에는 Dell 의 표현에 대한 현재 지침과 일치하지 않는 특정 단어가 포함되어 있을 수 있습니다. Dell 은 향후 릴리스에 대해 본 문서를 업데이트하고 이에 맞춰 해당 단어를 수정할 계획입니다.

본 문서에는 Dell 의 관리하에 있지 않으며, Dell 자체 콘텐츠에 대한 Dell 의 현재 지침과 일치하지 않는 타사 콘텐츠의 특정 표현이 포함되어 있을 수 있습니다. 이러한 타사 콘텐츠가 해당 업체에 의해 업데이트되면 이 문서도 그에 따라 수정됩니다.

Copyright © 2017–2021 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, Dell EMC 및 기타 상표는 Dell Inc. 또는 해당 자회사의 상표입니다. 기타 모든 상표는 해당 소유주의 상표일 수 있습니다. [10/29/2021] [기술 백서] [H16344.6]

목차

핵심 요약

기업은 가용성을 유지하는 것이 아주 중요한 막대한 양의 데이터를 계속해서 저장하고 있습니다. 시스템이나 사이트에 장애가 발생했을 때 막대한 양의 데이터를 복원하는 일은 복잡하고 비용이 많이 들어 IT 조직에 부담이 될 수 있습니다.

Dell EMC™ ECS™ 플랫폼은 오늘날 기업의 용량, 가용성 요구 사항을 모두 충족하도록 설계되었습니다. ECS 는 전 세계에 분산된 오브젝트 인프라스트럭처를 지원하는 엑사바이트 확장성을 제공합니다. ECS 는 자동 장애 탐지 및 자체 복구 옵션을 사용하여 엔터프라이즈 가용성을 확보하도록 설계되었습니다.

이 문서에서는 ECS 가 엔터프라이즈 가용성을 제공하는 방법에 대한 자세한 아키텍처를 설명합니다. 여기에는 다음과 같은 세부 사항이 포함됩니다.

- 분산된 인프라스트럭처가 시스템 가용성을 높이는 방법
- 데이터 내구성을 제공하는 고급 데이터 보호 방법
- 최적의 가용성을 위한 데이터 배포 방법
- 자동 장애 탐지
- 내장된 자가 복구 방법
- 디스크, 노드, 네트워크 장애 해결 세부 정보
- 재해 복구:
  - ECS 가 사이트 규모의 장애로부터 보호하는 방법
  - 액티브-액티브 다중 사이트 구성에서 일관성이 유지되는 방법
  - 사이트 규모의 장애 탐지 방법
  - 사이트 운영 중단 중 액세스 옵션
  - 사이트 규모의 영구적인 장애가 발생한 후 데이터 내구성을 재수립하는 방법

용어

VDC(Virtual Data Center): 이 백서에서 VDC(Virtual Data Center)라는 용어는 사이트 또는 영역과 동의어로 사용됩니다. 한 VDC 의 ECS 리소스는 동일한 내부 관리 네트워크에 속해야 합니다.

지리적 페더레이션: 여러 데이터 센터에 ECS 소프트웨어를 배치하여 지리적 페더레이션을 생성할 수 있습니다. 지리적 페더레이션에서 ECS 는 느슨하게 결합된 자율 VDC 페더레이션으로 동작합니다. 사이트 페더레이션에는 사이트 간 통신을 위한 복제와 관리 엔드포인트를 제공하는 것이 포함됩니다. 사이트가 페더레이션되면 페더레이션 내의 모든 노드에서 하나의 인프라스트럭처로 관리할 수 있습니다.

복제 그룹: 복제 그룹은 데이터가 보호되는 위치를 정의합니다. 로컬 복제 그룹은 하나의 VDC 를 포함하며 디스크 또는 노드 장애로부터 동일한 VDC 내의 데이터를 보호합니다. 글로벌 복제 그룹에는 둘 이상의 VDC 가 포함되어 있으며 디스크, 노드, 사이트 장애로부터 데이터를 보호합니다. 복제 그룹은 버킷 수준에서 할당됩니다.

1 고가용성 설계 개요

고가용성은 시스템 가용성과 데이터 내구성이라는 두 가지 주요 영역으로 설명할 수 있습니다. 시스템은 클라이언트 요청에 응답할 수 있을 때 사용할 수 있습니다. 데이터 내구성은 시스템 가용성과 무관하게 제공되며 데이터가 손실이나 손상 없이 시스템에 저장될 수 있도록 보장합니다. 즉 네트워크 운영 중단과 같이 ECS 시스템이 다운되어도 데이터가 계속 보호됩니다.

ECS 아키텍처의 분산 특성은 VDC(Virtual Data Center)/사이트의 모든 노드가 클라이언트 요청에 응답할 수 있도록 함으로써 시스템 가용성을 제공합니다. 노드가 다운되면 클라이언트를 수동으로 또는 자동으로(예: DNS 또는 로드 밸런서 사용) 요청을 처리할 수 있는 다른 노드로 리디렉션할 수 있습니다.

ECS 는 3 중 미러링과 삭제 코딩의 조합을 사용하여 디스크, 노드 장애에 대한 회복탄력성을 높이기 위해 분산 방식으로 데이터를 씁니다. ECS 는 사이트 간 복제로 사이트 전체를 장애로부터 보호함으로써 가용성과 회복탄력성을 높입니다. 또한 ECS 는 자가 복구 기능을 가지고 있으며 정기적, 체계적인 데이터 무결성 검사를 합니다.

고가용성을 살펴볼 때는 먼저 아키텍처와 ECS 내에서 최적의 가용성과 성능을 위해 데이터가 배포되는 방식을 이해하는 것이 중요합니다.

1.1 청크

청크는 오브젝트 데이터, 맞춤형 클라이언트에서 제공하는 메타데이터, ECS 시스템 메타데이터를 포함한 모든 유형의 데이터를 저장하는 데 사용되는 논리 컨테이너입니다. 청크에는 그림 1 에 표시된 것처럼 하나의 버킷에서 하나 이상의 오브젝트로 구성된 128MB 의 데이터가 포함되어 있습니다.

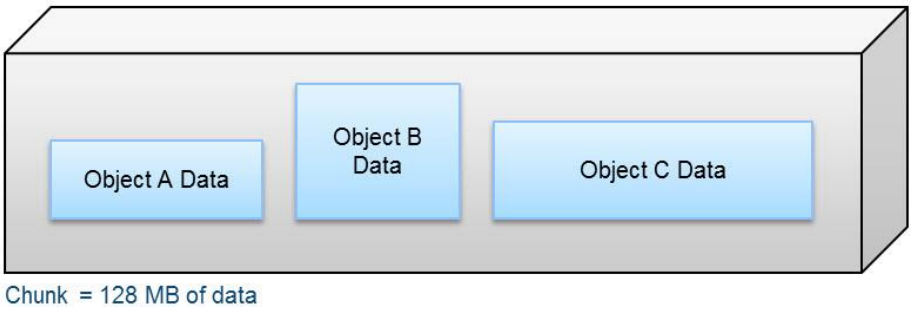

그림 1 논리적 청크

ECS 는 인덱싱을 사용하여 청크 내의 모든 데이터를 추적합니다. 자세한 내용은 1.2 를 참조하십시오.

1.2 ECS 메타데이터

ECS 는 트랜잭션 내역뿐만 아니라 데이터가 존재하는 위치를 추적하는 자체 메타데이터를 관리합니다. 이 메타데이터는 논리 테이블과 저널에서 관리됩니다.

테이블에는 오브젝트와 관련된 정보를 저장하는 키-값 쌍이 있습니다. 해시 함수는 키와 관련된 값을 빠르게 조회하는 데 사용됩니다. 이러한 키-값 쌍은 데이터 위치를 빠르게 인덱싱할 수 있도록 B+ 트리에 저장됩니다. 키-값 쌍을 B+ 트리와 같이 균형 잡힌 검색된 트리 형식에 저장하면 데이터 및 메타데이터의 위치에 빠르게 액세스할 수 있습니다. 또한 이러한 논리 테이블의 쿼리 성능을 더욱 향상시키기 위해 ECS 는 2 단계 LSM(Log-Structured Merge) 트리를 구현합니다. 따라서 더 작은 트리가 메모리(메모리 테이블)에 있고 메인 B+ 트리가 디스크에 있는 두 개의 트리 같은 구조가 있습니다. 따라서 키-값 쌍을 조회하면 먼저 메모리 테이블을 쿼리하고 값이 메모리에 없으면 디스크의 메인 B+ 트리를 검색합니다.

트랜잭션 기록은 저널 로그에 기록되고 이러한 로그는 디스크에 기록됩니다. 저널은 아직 B+ 트리에 커밋되지 않은 인덱스 트랜잭션을 추적합니다. 트랜잭션이 저널에 로깅되면 메모리 테이블이 업데이트됩니다. 메모리 테이블이 가득 차거나 설정된 시간이 지나면 테이블이 정렬되거나 디스크의 B+ 트리에 덤프되고 체크포인트가 저널에 기록됩니다. 이 프로세스는 그림 2 에 설명되어 있습니다.

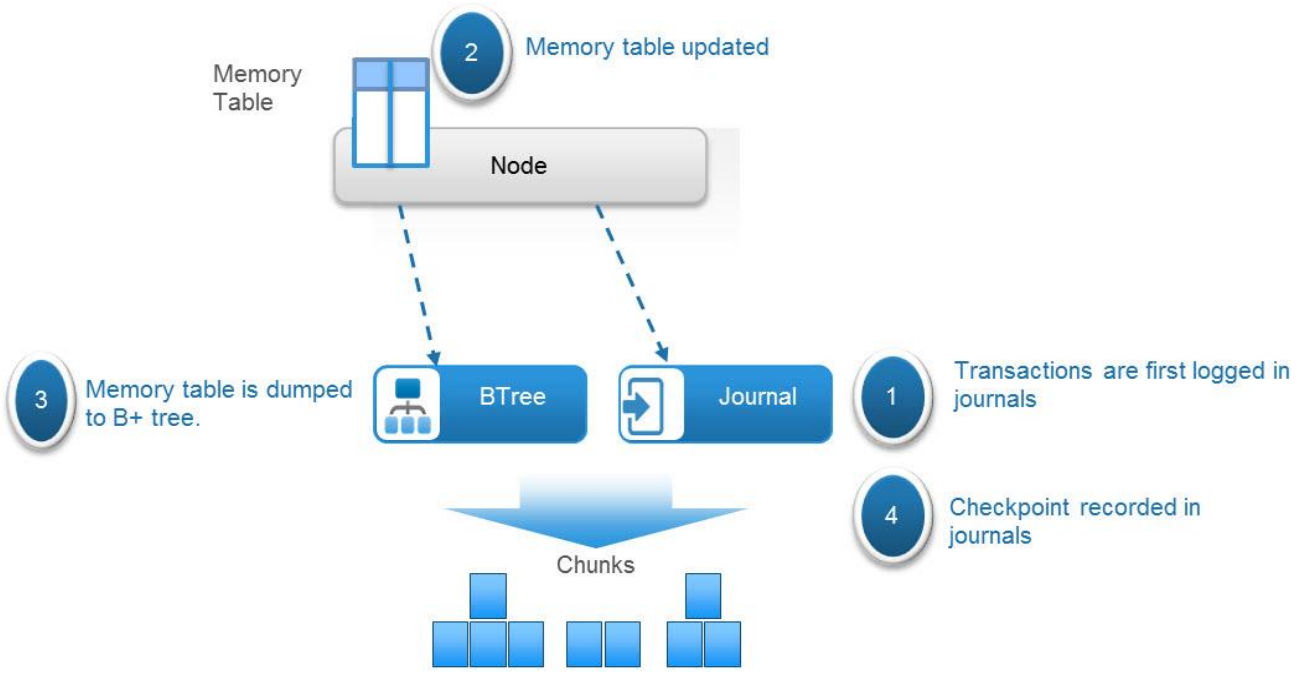

그림 2 ECS 테이블에 대한 트랜잭션 업데이트 워크플로

저널과 B+ 트리 모두 청크에 기록됩니다.

ECS 가 사용하는 테이블에는 여러 가지가 있으며, 각각의 테이블은 아주 커질 수 있습니다. 테이블 조회의 성능을 최적화하기 위해 각 테이블은 VDC/사이트의 노드에 분산된 파티션으로 분할됩니다. 그러면 파티션이 쓰여진 노드가 그 파티션 또는 테이블 섹션의 소유자/주체자가 됩니다.

이러한 테이블 중 하나는 청크 테이블로, 청크 조각과 디스크에 있는 복제본의 물리적 위치를 추적합니다. 표 1 에는 노드 내 디스크, 디스크 내 파일, 그 파일 내 오프셋, 데이터의 길이를 나열하여 각 청크에 대한 물리적 위치를 식별하는 청크 테이블의 파티션 샘플이 나와 있습니다. 여기서 볼 수 있는 청크 ID C1 은 삭제되고 청크 ID C2 는 3 중 미러링됩니다. 3 중 미러링과 삭제 코딩에 대한 자세한 내용은 이 문서의 1.4 를 참조하십시오.

표 1 샘플 청크 테이블 파티션

청크 ID	청크 위치
C1	노드 1:디스크 1:파일 1:오프셋 1:길이 노드 2:디스크 1:파일 1:오프셋 1:길이 노드 3:디스크 1:파일 1:오프셋 1:길이 노드 4:디스크 1:파일 1:오프셋 1:길이 노드 5:디스크 1:파일 1:오프셋 1:길이 노드 6:디스크 1:파일 1:오프셋 1:길이 노드 7:디스크 1:파일 1:오프셋 1:길이 노드 8:디스크 1:파일 1:오프셋 1:길이 노드 1:디스크 2:파일 1:오프셋 1:길이 노드 2:디스크 2:파일 1:오프셋 1:길이 노드 3:디스크 2:파일 1:오프셋 1:길이 노드 4:디스크 2:파일 1:오프셋 1:길이 노드 5:디스크 2:파일 1:오프셋 1:길이 노드 6:디스크 2:파일 1:오프셋 1:길이 노드 7:디스크 2:파일 1:오프셋 1:길이 노드 8:디스크 2:파일 1:오프셋 1:길이
C2	노드 1:디스크 3:파일 1:오프셋 1:길이 노드 2:디스크 3:파일 1:오프셋 1:길이 노드 3:디스크 3:파일 1:오프셋 1:길이

다른 예로는 오브젝트 이름에서 청크로의 매핑에 사용되는 오브젝트 테이블이 있습니다. 표 2 에서는 오브젝트가 있는 청크와 청크 내의 위치를 자세히 설명하는 오브젝트 테이블 파티션의 예를 보여 줍니다.

표 2 샘플 오브젝트 테이블

오브젝트 이름	청크 ID
ImgA	C1:오프셋:길이
FileA	C4:오프셋:길이 C6:오프셋:길이

테이블 파티션 소유자의 매핑은 모든 노드에서 실행되는 vnest 라는 서비스에 의해 유지됩니다. 표 3 에는 vnest 매핑 테이블 부분의 예가 나와 있습니다.

표 3 샘플 vnest 매핑 테이블

테이블 ID	테이블 파티션 소유자
테이블 0 P1	노드 1
테이블 0 P2	노드 2

1.3 장애 도메인

일반적으로 장애 도메인은 장애 가능성이 있는 솔루션 내의 구성 요소를 고려하는 엔지니어링 설계의 개념과 관련이 있습니다. ECS 소프트웨어는 어떤 디스크가 동일한 노드에 있는지, 어떤 노드가 동일한 랙에 있는지 자동으로 인식합니다. 대부분의 장애 시나리오로부터 보호하기 위해 ECS 소프트웨어는 데이터를 쓸 때 이 정보를 활용하도록 설계되었습니다. ECS 가 활용하는 장애 도메인의 기본 지침은 다음을 포함합니다.

- ECS 는 노드의 동일한 디스크에 동일한 청크의 조각을 쓰지 않습니다.
- ECS 는 노드 간에 청크의 조각을 균등하게 분배합니다.
- ECS 는 랙을 인식합니다. 충분한 공간이 있다는 가정 하에 VDC/사이트에 두 개 이상의 랙이 있으면 ECS 는 이러한 랙에 청크의 조각을 균등하게 분배하기 위해 최선의 노력을 기울입니다.

1.4 고급 데이터 보호 방법

ECS 내에서 오브젝트가 생성되면 여기에는 쓰기 데이터, 맞춤형 메타데이터, ECS 메타데이터가 포함됩니다. ECS 메타데이터에는 저널 청크와 btree 청크가 포함됩니다. 각각은 하나 이상의 오브젝트의 최대 128MB 의 데이터를 포함하는 서로 다른 논리 청크에 기록됩니다. ECS 는 VDC(Virtual Data Center)/사이트 내의 데이터를 보호하기 위해 3 중 미러링과 삭제 코딩의 조합을 사용합니다.

- 3 중 미러링은 3 개의 데이터 복사본이 기록되도록 보장하므로 두 개의 노드 장애로부터 데이터를 보호합니다.
- 삭제 코딩은 디스크와 노드 장애로부터 향상된 데이터 보호 기능을 제공합니다. 이는 Reed Solomon 삭제 코딩 체계를 활용하며, 이 방식은 청크를 데이터와 코딩 조각으로 나누고 VDC/사이트 내의 노드 전체에 균등하게 배포합니다.

데이터 크기 및 유형에 따라 표 4 에 나와 있는 데이터 보호 방법 중 하나를 사용하여 쓰게 됩니다.

표 4 서로 다른 유형의 데이터에 사용할 데이터 보호 수준 결정

데이터 유형	사용되는 데이터 보호 방법
저널 청크	3 중 미러링
Btree 청크/맞춤형 메타데이터	중복 데이터 세그먼트를 사용한 삭제 코딩
오브젝트 데이터 <128MB	트리플 미러링 + 삭제 코딩
오브젝트 데이터 >128MB	인라인 삭제 코딩

참고: EXF900 과 같은 올플래시 아키텍처에서 btree 청크 보호는 3 중 미러링입니다.

1.4.1 3 중 미러

3 중 미러 쓰기 방식은 ECS 저널 청크에 적용 가능하며, 여기서 ECS 는 3 개의 복제 복사본을 만듭니다. 각 복제 복사본은 장애 도메인에 걸쳐 서로 다른 노드의 한 디스크에 기록됩니다. 이 방법은 2 노드 또는 2 디스크 장애로부터 청크 데이터를 보호합니다.

그림 3 은 128MB 메타데이터를 포함한 논리 청크에 세 개의 복제 복사본이 있고 각각 다른 노드에 기록되는 3 중 미러링의 예를 보여 줍니다.

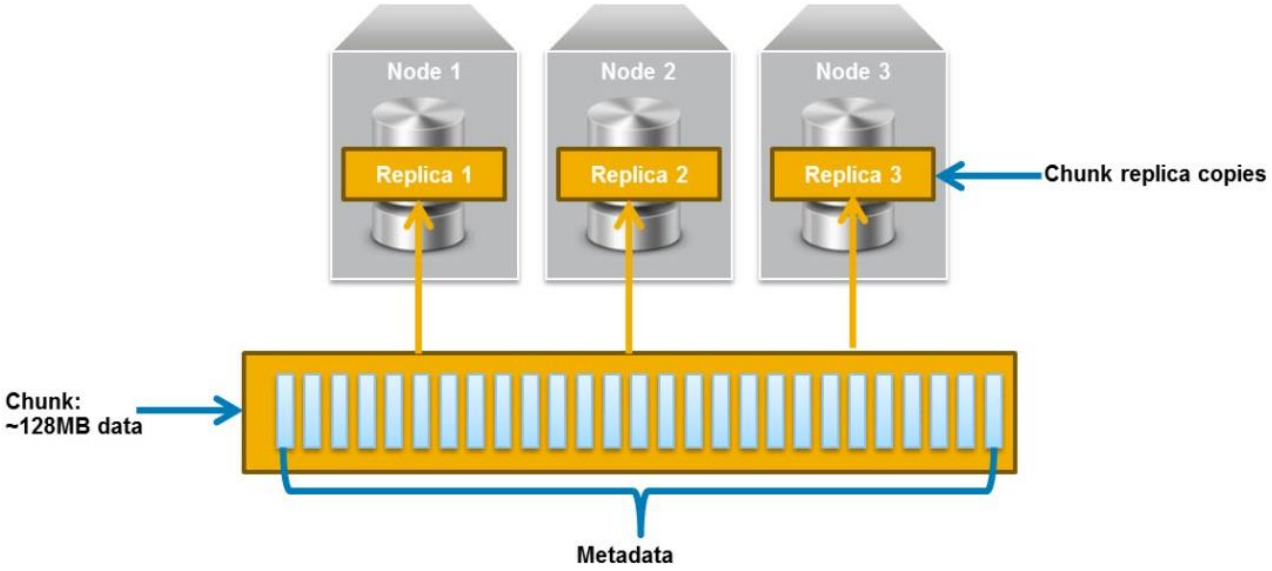

그림 3 3 중 미러링

1.4.2 중복 데이터 세그먼트를 사용한 삭제 코딩

중복 데이터 세그먼트 쓰기 방법을 사용한 삭제 코딩은 ECS btree 청크, 맞춤형 오브젝트 메타데이터에 적용할 수 있습니다. 여기에는 12 개의 데이터 세그먼트, 12 개의 복제된 데이터 세그먼트, 4 개의 패리티 세그먼트가 포함됩니다. 새로운 btree 청크 중복 데이터 EC 체계는 메타데이터 보호 오버헤드를 줄입니다.

1.4.3 트리플 미러 + 원 위치 삭제 코딩

이 쓰기 방법은 크기가 128MB 미만인 오브젝트의 데이터에 적용할 수 있습니다.

오브젝트가 생성되면 청크에 기록되며, ECS 는 다음과 같이 3 개의 복제 복사본을 생성합니다.

- 하나의 복제본은 여러 노드와 디스크에 분산되어 있는 조각으로 작성됩니다. 배포 과정에서 조각은 가능한 한 많은 장애 도메인으로 분산됩니다. 각 디스크에 기록되는 크기는 사용 중인 삭제 코딩 방식에 따라 달라집니다.
  - 삭제 코딩 체계가 기본값(12+4)이면 각 디스크는 최대 10.67MB 의 용량을 받습니다.
  - 삭제 코딩 방식이 콜드 스토리지(10+2)이면 각 디스크는 최대 12.8MB 의 용량을 받습니다.
- 청크의 두 번째 복제본은 노드의 한 디스크에 기록됩니다.
- 청크의 세 번째 복제본은 다른 노드의 한 디스크에 기록됩니다.

이 방법이 3 중 미러링을 하고 2 노드 또는 2 디스크 장애로부터 청크 데이터를 보호합니다.

추가 오브젝트는 최대 128MB 의 데이터를 포함할 때까지의 시간 또는 미리 정의된 시간 중 더 짧은 시간 이후에 동일한 청크에 기록됩니다. 이때 Reed Solomon 삭제 코딩 체계는 청크에 대한 코딩(패리티) 조각을 계산하여 서로 다른 디스크에 씁니다. 이렇게 하면 코딩 조각을 포함하여 청크 내의 모든 조각을 서로 다른 디스크에 기록하고 장애 도메인에 배포하게 됩니다.

코딩 조각이 디스크에 기록되면 두 번째, 세 번째 복제 복사본이 디스크에서 삭제됩니다. 이 작업이 완료된 후 청크는 3 중 미러링보다 높은 수준의 가용성을 제공하는 삭제 코딩에 의해 보호됩니다.

1.4.4 인라인 삭제 코딩

이 쓰기 방법은 128MB 이상의 오브젝트의 데이터에 적용할 수 있습니다. 오브젝트는 128MB 청크로 나뉩니다. Reed Solomon 삭제 코딩 체계는 각 청크에 대한 코딩(패리티) 조각을 계산합니다. 각 조각은 서로 다른 디스크에 기록되고 장애 도메인에 배포됩니다. 각 디스크에 기록되는 크기는 사용 중인 삭제 코딩 방식에 따라 달라집니다.

- 삭제 코딩 체계가 기본값(12+4)이면 조각은 16 개의 디스크에 분산되며 각 조각은 최대 10.67MB 입니다.
- 삭제 코딩 체계가 콜드 스토리지(10+2)이면 조각은 12 개의 디스크에 분산되며 각 조각은 최대 12.8MB 입니다.

128MB 미만 오브젝트의 나머지 부분은 앞에서 설명한 3 중 미러링 + 원 위치 삭제 코딩 체계를 사용하여 기록됩니다. 예를 들어 오브젝트가 150MB 라면 128MB 는 인라인 삭제 코딩을 사용하여 기록되고, 나머지 22MB 는 3 중 미러 + 원 위치 삭제 코딩을 사용하여 기록됩니다.

그림 4 에서는 청크가 장애 도메인에 분산되는 방식을 보여 주는 예입니다. 이 예에는 두 개의 랙에 걸쳐 있는 하나의 VDC/사이트가 있으며 각 랙에는 4 개의 노드가 있습니다.

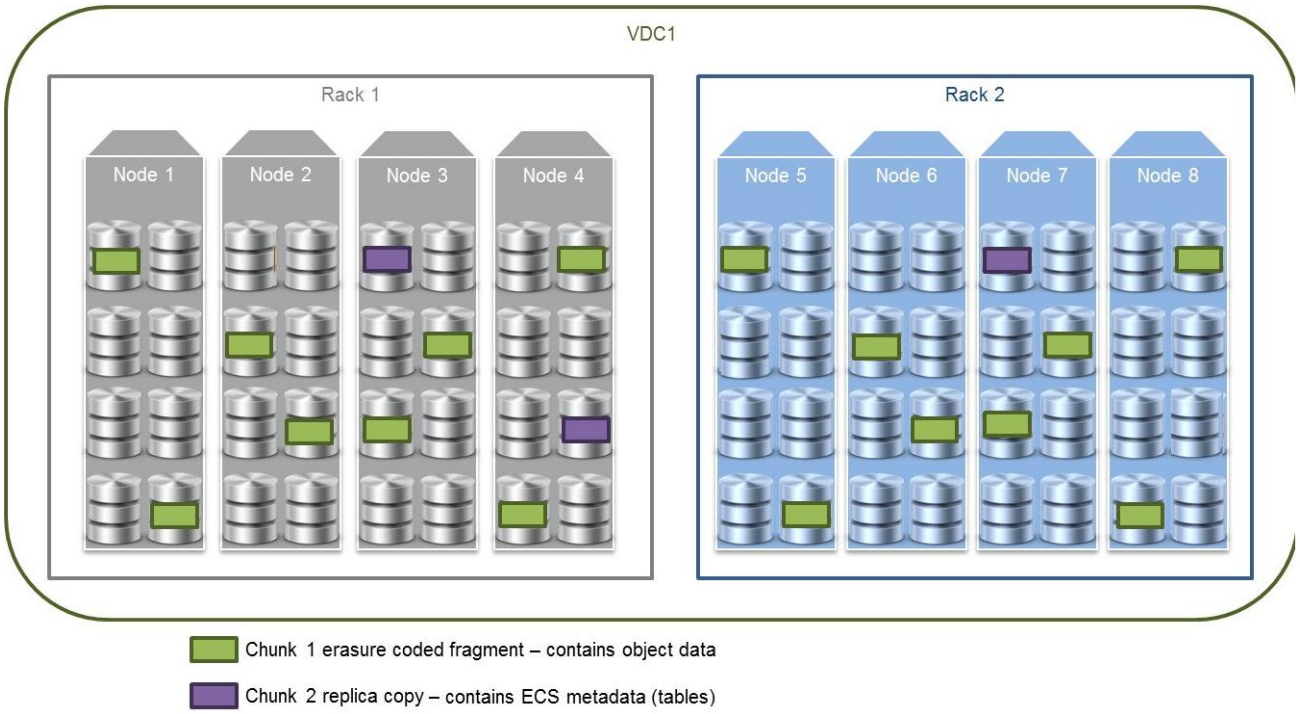

그림 4 장애 도메인에 청크가 분산되는 방식

- 청크 1 은 12+4 의 삭제 코딩을 사용하여 삭제 코딩된 오브젝트 데이터를 포함합니다. 조각은 랙당 4 개씩 8 개 노드 전체에 균등하게 배포됩니다. 각 노드에는 2 개의 조각이 포함되어 있으며, 이를 두 개의 서로 다른 디스크에 씁니다.
- 청크 2 는 ECS 메타데이터(테이블)를 포함하므로 3 중 미러됩니다. 각 복제 복사본은 각각 한 디스크에 있는 서로 다른 노드에 기록됩니다. 복제본은 랙에 걸쳐 있어 최고의 가용성을 가집니다.

1.5 삭제 코딩 보호 수준

스토리지 풀을 생성하는 동안 선택한 삭제 코딩 방식에 따라 삭제 코딩된 데이터는 다음에 발생하는 장애로부터 보호됩니다.

1.5.1 기본 삭제 코딩 체계(12+4):

ECS 가 기본 삭제 코딩 체계를 사용하여 삭제 코딩을 하려면 최소 4 개의 노드가 있어야 합니다. 스토리지 풀에 노드가 4 개 미만이면 삭제 코딩이 중지되므로 보호 수준이 3 중 미러링이 됩니다. 이 시간 동안 세 개의 복제본은 유지되며 어떤 청크에서도 패리티가 계산되지 않습니다. 스토리지 풀에 노드를 추가하고 지원되는 최소 노드 수를 충족하면 여기서뿐만 아니라 새 청크에서도 삭제 코딩이 계속됩니다.

각각의 128 MB 청크에 대해 기본 삭제 코딩 체계는 12 개의 데이터 조각과 4 개의 코딩 조각을 기록하며, 각각의 크기는 최대 10.67MB 입니다. 이는 최대 4 개의 청크 조각 손실로부터 청크 데이터를 보호하며 여기에는 표 5 에 나와 있는 장애 시나리오도 포함할 수 있습니다.

표 5 기본 삭제 코딩 보호

VDC 의 노드 수	노드당 청크 조각 수	코딩된 데이터가 보호되는 장애 상황
5 노드	4	• 최대 4 개 디스크 손실 • 1 개 노드 손실
6 또는 7 노드	3	• 최대 4 개 디스크 손실 • 두 번째 노드에서 1 개 노드 및 1 개 디스크 손실
8 개 이상의 노드	2	• 최대 4 개 디스크 손실 • 2 개 노드 손실 • 2 개 노드 및 2 개 디스크 손실
16 개 이상의 노드	1	• 4 개 노드 손실 • 1 개의 추가 노드에서 3 개 노드 및 디스크 손실 • 최대 2 개의 서로 다른 노드에서 2 개 노드 및 디스크 손실 • 최대 3 개의 서로 다른 노드에서 1 개 노드 및 디스크 손실 • 4 개의 서로 다른 노드에서 4 개 디스크 손실

참고: 표 5 는 청크 조각의 전체 배포로 가능한 보호 수준을 반영한 것입니다. 노드에 사용 가능한 공간이 부족한 경우와 같이 노드에 더 많은 조각이 있는 상황이 있을 수 있습니다. 이 경우 보호 수준이 달라질 수 있습니다.

1.5.2 콜드 스토리지 삭제 코딩 체계(10+2):

ECS 가 콜드 스토리지 삭제 코딩 체계를 사용하여 삭제 코딩을 할 수 있으려면 최소 6 개의 노드가 필요합니다. 스토리지 풀에 노드가 6 개 미만이면 삭제 코딩이 중지됩니다. 즉, 3 개의 복제 복사본이 남고 청크에서 패리티가 계산되지 않습니다. 추가 노드가 스토리지 풀에 추가되면 여기서뿐만 아니라 새 청크에서도 삭제 코딩이 계속됩니다.

각각의 128 MB 청크에서 콜드 스토리지 삭제 코딩 체계는 10 개의 데이터 조각과 2 개의 코딩 조각을 생성하며, 각각의 크기는 최대 12.8MB 입니다. 이는 최대 2 개의 청크 조각 손실로부터 청크 데이터를 보호하며 여기에는 표 6 에 나와 있는 장애 시나리오도 포함할 수 있습니다.

표 6 콜드 스토리지 삭제 코딩 보호

VDC 의 노드 수	노드당 청크 조각 수	코딩된 데이터가 보호되는 장애 상황
11 개 이하의 노드	2	• 최대 2 개 디스크 손실 • 1 개 노드 손실
12 개 이상의 노드	1	• 2 개의 서로 다른 노드에서 디스크 손실 • 2 개 노드 손실

참고: 이 표는 청크 조각의 전체 배포로 가능한 보호 수준을 반영한 것입니다. 노드에 사용 가능한 공간이 부족한 경우와 같이 노드에 더 많은 조각이 있는 상황이 있을 수 있습니다. 이 경우 보호 수준이 달라질 수 있습니다.

1.6 체크섬

ECS 가 데이터 무결성을 보장하기 위해 사용하는 또 다른 메커니즘은 기록된 데이터의 체크섬을 저장하는 것입니다. 체크섬은 쓰기 단위당 최대 2MB 까지 수행됩니다. 따라서 큰 오브젝트 쓰기의 경우 하나의 오브젝트 조각에 대해 체크섬이 발생하거나 2MB 미만의 작은 오브젝트 쓰기의 경우 오브젝트별로 발생할 수 있습니다. 쓰기 작업 중에 체크섬이 메모리에서 계산된 다음 디스크에 기록됩니다. 읽기 시 데이터를 체크섬과 함께 읽은 다음 읽은 데이터로 메모리에서 체크섬을 계산하고 디스크에 저장된 체크섬과 비교하여 데이터 무결성을 확인합니다. 또한 스토리지 엔진은 백그라운드에서 일관성 검사기를 주기적으로 실행하고 전체 데이터 세트에 대해 체크섬 확인을 합니다.

1.7 오브젝트 쓰기

ECS 에서 쓰기 작업이 발생하면 클라이언트가 노드에 요청을 보내는 것으로 시작합니다. ECS 는 분산 아키텍처로 설계되어 VDC/사이트의 모든 노드가 읽기 또는 쓰기 요청에 응답할 수 있도록 합니다. 쓰기 요청에는 오브젝트 데이터, 맞춤형 오브젝트 메타데이터, 트랜잭션을 저널 로그에 기록하는 작업이 포함됩니다. 이 작업이 완료되면 클라이언트에게 확인 메시지가 전송됩니다.

그림 5 와 앞에서 설명한 단계에서 쓰기 워크플로의 간략한 개요를 볼 수 있습니다.

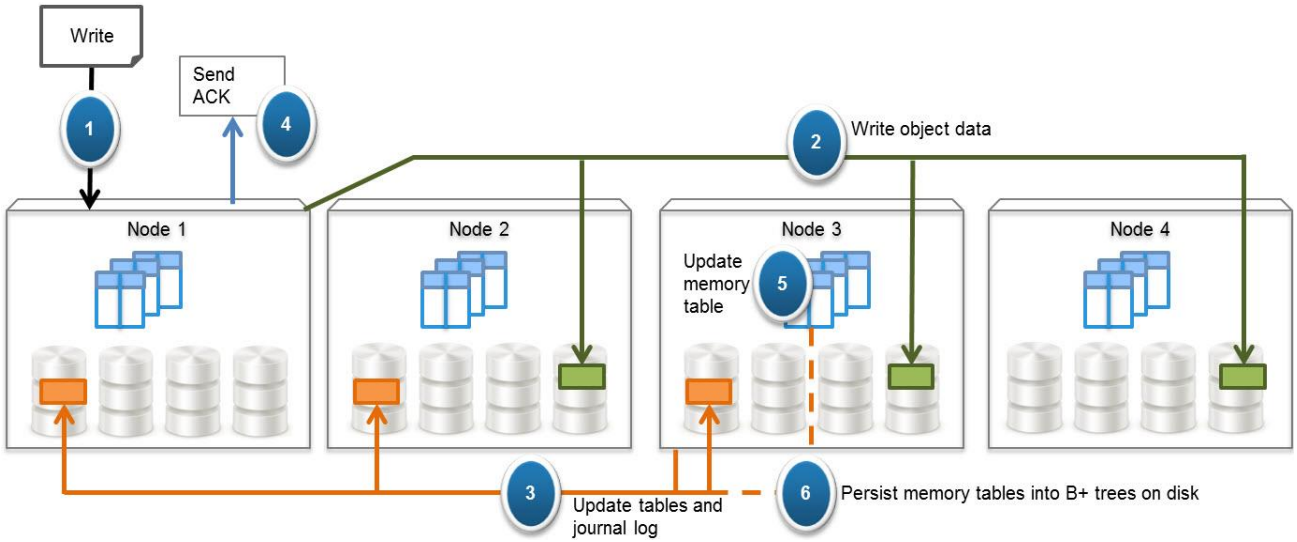

그림 5 오브젝트 쓰기 워크플로

1. 오브젝트 쓰기 요청을 받습니다. 모든 노드가 이 요청에 응답할 수 있지만 이 예에서는 노드 1 이 요청을 처리합니다.
2. 오브젝트의 크기에 따라 데이터가 하나 이상의 청크에 기록됩니다. 각 청크는 3 중 미러링, 삭제 코딩과 같은 고급 데이터 보호 체계를 사용하여 보호됩니다. 데이터를 디스크에 쓰기 전에 ECS 는 체크섬 기능을 실행하고 결과를 저장합니다.

   데이터가 청크에 추가됩니다. 이 오브젝트의 크기는 10MB 에 불과하기 때문에 3 중 미러링과 삭제 코딩 체계를 사용합니다. 따라서 서로 다른 세 노드의 디스크 3 개, 이 예에서는 노드 2, 노드 3, 노드 4 에 쓰게 됩니다. 이 세 노드가 노드 1 에 다시 확인을 보냅니다.

3. 오브젝트 데이터를 성공적으로 쓰면 오브젝트의 메타데이터가 저장됩니다. 이 예에서는 노드 3 이 이 오브젝트가 속한 오브젝트 테이블의 파티션을 소유합니다. 노드 3 은 소유자로서 오브젝트 이름과 청크 ID 를 오브젝트 테이블의 저널 로그의 이 파티션에 기록합니다. 저널 로그는 3 중 미러링되므로 노드 3 은 복제 복사본을 3 개의 서로 다른 노드에 병렬로 보냅니다. 이 예에서는 노드 1, 노드 2, 노드 3 입니다.
4. 클라이언트에게 확인이 전송됩니다.
5. 백그라운드 프로세스로 메모리 테이블이 업데이트됩니다.
6. 메모리 안의 테이블이 가득 차거나 정해진 시간이 지나면 테이블은 B+ 트리에서 청크로 병합, 정렬 또는 덤프되고 체크포인트가 저널에 기록됩니다.

1.8 오브젝트 읽기

ECS 는 분산 아키텍처로 설계되어 VDC/사이트의 모든 노드가 읽기 또는 쓰기 요청에 응답할 수 있도록 합니다. 읽기 요청에는 파티션 레코드 소유자로부터의 테이블 조회를 하여 데이터의 물리적 위치를 찾는 작업뿐 아니라 바이트 오프셋 읽기, 체크섬 검증, 요청한 클라이언트로 데이터를 반환하는 작업이 포함됩니다.

그림 6 과 앞에서 설명한 단계에서 읽기 워크플로의 간략한 개요를 볼 수 있습니다.

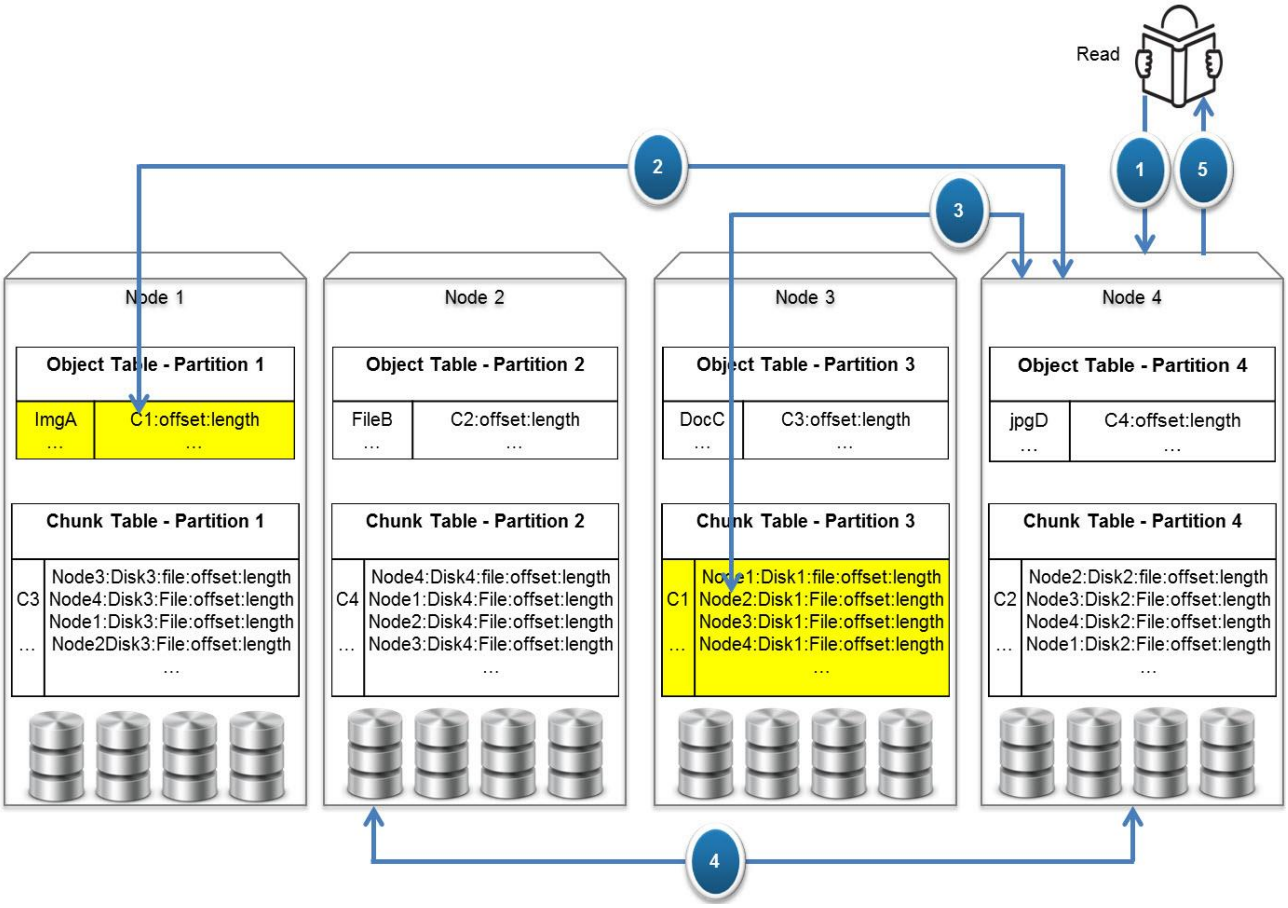

그림 6 오브젝트 읽기 워크플로

1. ImgA 에 대한 읽기 요청이 수신됩니다. 모든 노드가 이 요청에 응답할 수 있지만 이 예에서는 노드 4 가 요청을 처리합니다.
2. 노드 4 는 노드 1(ImgA 의 오브젝트 테이블 파티션 소유자)에 청크 정보를 요청합니다.
3. ImgA 가 특정 오프셋과 길이로 C1 에 있다는 것을 알고 있는 노드 4 는 노드 3(C1 에 대한 청크 테이블 파티션 소유자)에 청크의 물리적 위치를 요청합니다.
4. 이제 노드 4 가 ImgA 의 물리적 위치를 알게 되었으므로 그 파일의 데이터 조각을 포함하는 노드(이 예에서는 노드 2 디스크 1)에게 데이터를 요청합니다. 노드가 바이트 오프셋 읽기를 하고 데이터를 노드 4 로 반환합니다.
5. 노드 4 가 체크섬을 확인한 다음 데이터를 요청 클라이언트로 반환합니다.

참고: 4 단계에서, 각 노드가 자신의 데이터 저장소 자체만 읽을 수 있는 EX300, EX500, EX3000 과 같은 하드 디스크 드라이브 아키텍처가 아닌 EXF900 과 같은 올플래시 아키텍처는 다른 노드의 데이터도 직접 읽을 수 있습니다.

2 로컬 사이트 가용성

ECS 아키텍처의 분산 특성은 다양한 장애에 대비한 시스템 가용성과 데이터 내구성 형태 모두로 고가용성을 유지합니다. 이 섹션에서는 로컬 사이트 장애 시 가용성을 유지하는 방법을 중점으로 설명합니다.

2.1 디스크 장애

아키텍처 섹션에서 ECS 가 다양한 장애 시나리오에서 회복탄력성을 가지고 데이터를 쓰기 위해 3 중 미러링과 삭제 코딩의 조합을 사용하는 방법을 설명했습니다.

데이터의 일관성을 위해 읽기 시, 그리고 일관성 검사기가 체크섬의 일관성을 검증합니다. 일관성 검사기는 전체 데이터 세트에 대해 주기적으로 체크섬 검증을 하는 백그라운드 프로세스입니다. 읽기 요청도 체크섬 검증을 실행합니다.

드라이브가 응답하지 않거나 체크섬 검증에 실패하여 읽기 요청에 조각이 없으면 청크 관리자에게 알림이 전송됩니다. 청크 관리자는 나머지 삭제 코딩된 데이터와 패리티 조각 또는 복제 복사본을 사용하여 누락된 조각의 재구성을 시작하고 그 후에 청크 정보를 업데이트합니다. 조각이 다시 생성되면 미결 또는 새로운 읽기 요청이 업데이트된 청크 정보를 사용하여 데이터 조각을 요청하고 읽기 요청을 처리합니다.

ECS 노드는 직접 연결된 디스크에 대해 지속적으로 상태 점검을 실행합니다. 디스크가 응답하지 않게 되면 ECS 노드가 청크 관리자에게 새 쓰기 작업에 이를 포함시키는 것을 중지하도록 알립니다. 일정 시간(기본값: 60 분)이 지난 후에도 응답이 없으면 청크 관리자에게 장애가 발생한 드라이브에서 데이터를 다시 생성하라는 알림이 전송됩니다. ECS 노드는 장애가 발생한 드라이브에 블록이 있어 복구해야 하는 청크를 파악합니다. 그리고 장애가 발생한 드라이브에 저장된 모든 청크 조각의 병렬 복구를 시작할 청크 관리자에게 이 정보를 보냅니다. 청크 조각은 나머지 삭제 코딩된 조각 또는 복제 복사본을 사용하여 다른 디스크로 복구됩니다. 새 조각이 기록되면 관련 청크 테이블이 이 정보로 업데이트됩니다. 가능하면 청크 관리자가 장애가 발생한 드라이브의 조각도 삭제합니다. 디스크가 나중에 다시 온라인 상태가 되면 다음과 같이 됩니다.

- 설정된 기간(기본값 90 분) 내에 응답하지 않으면 나머지 복구 작업이 취소됩니다.
- 설정된 시간(기본값 90 분) 동안 응답하지 않거나 하드웨어 관리자가 실패한 것으로 보고하면 ECS 가 드라이브를 제거합니다. 드라이브가 제거되면 나머지 복구 작업이 완료될 때까지 계속됩니다. 복구가 완료되면 청크 관리자가 장애가 발생한 이 디스크에 대한 모든 참조를 청크 테이블에서 제거합니다.

  이 드라이브가 제거된 후 온라인 상태가 되면 새 드라이브로 추가되고 청크 관리자가 새 쓰기 작업에 이 드라이브를 포함합니다.

2.2 ECS 노드 장애

ECS 는 노드에서 지속적으로 상태 점검을 합니다. 시스템 가용성을 유지하기 위해, ECS 분산 아키텍처는 어떤 노드라도 클라이언트 요청을 수락할 수 있도록 합니다. 노드 하나가 다운되면 클라이언트를 수동으로 또는 자동으로 요청을 처리할 수 있는 노드로 리디렉션(예: DNS 또는 로드 밸런서 사용)할 수 있습니다.

잘못된 이벤트로 인해 재구성 작업을 트리거하지 않기 위해 노드가 몇 개의 순차적 상태 점검을 통과하지 못하지 않는 이상 전체 재구성 작업은 트리거되지 않습니다. 기본값은 60 분입니다. 응답하지 않지만 전체 재구성이 트리거되기 전에 노드에 IO 요청이 들어오는 경우:

- 응답하지 않는 노드에서 호스팅되는 파티션 테이블에 대한 요청이 들어오면 요청된 파티션 테이블의 소유권을 사이트의 나머지 노드에 재분배하게 됩니다. 이 작업이 완료되면 요청이 성공적으로 완료됩니다.
- 응답하지 않는 노드의 디스크에 있는 데이터에 대한 IO 요청은 나머지 삭제 코딩된 데이터, 패리티 조각 또는 복제 복사본을 사용하여 재구성되며, 그 후에 청크 정보가 업데이트됩니다. 조각이 다시 생성되면 미결 또는 새로운 읽기 요청이 업데이트된 청크 정보를 사용하여 데이터 조각을 요청하고 읽기 요청을 처리합니다.

설정된 수의 순차적 상태 점검(기본값 60 분)에 실패한 노드는 다운된 것으로 간주됩니다. 이렇게 되면 실패한 노드가 소유한 디스크의 파티션 테이블과 청크 조각을 자동으로 다시 생성합니다.

재생성 작업의 알림이 청크 관리자에게 전송되어 실패한 노드의 디스크에 저장된 모든 청크 조각의 병렬 복구를 시작하게 됩니다. 여기에는 오브젝트 데이터, 맞춤형 클라이언트가 제공한 메타데이터, ECS 메타데이터를 포함하는 청크가 포함될 수 있습니다. 장애가 발생한 노드가 다시 온라인 상태가 되면 업데이트된 상태가 청크 관리자로 전송되고 완료되지 않은 복구 작업이 취소됩니다. 청크 조각 복구에 대한 자세한 내용은 위의 디스크 장애 섹션에 나와 있습니다.

하드웨어 모니터링 외에도 ECS 는 각 노드의 모든 서비스와 데이터 테이블을 모니터링합니다.

- 테이블 장애가 발생했지만 노드가 여전히 작동 중이면 동일한 노드에서 테이블을 자동으로 다시 초기화하려고 시도합니다.
- 서비스 장애가 탐지되면 먼저 서비스 재시작을 시도합니다.

이 작업이 실패하면 다운된 노드 또는 서비스가 소유한 테이블의 소유권을 VDC/사이트의 나머지 모든 노드에 재분배합니다. 소유권 변경에는 vnest 정보를 업데이트하고 장애가 발생한 노드에서 소유한 메모리 테이블을 다시 생성하는 작업이 포함됩니다. 새 파티션 테이블 소유자 정보를 사용하여 나머지 노드에서 vnest 정보가 업데이트됩니다.

실패한 노드의 메모리 테이블은 마지막으로 성공한 저널 체크포인트 이후에 작성된 저널 항목을 재생하여 다시 생성됩니다.

2.2.1 다중 노드 장애

사이트 내에서 여러 노드에 장애가 발생할 수 있는 경우가 있습니다. 노드에 하나씩 장애가 발생할 수도 있고 동시에 장애가 발생할 수도 있습니다.

- **하나씩 장애 발생**: 노드에 하나씩 장애가 발생한다는 것은 한 노드에서 장애가 발생하고 모든 복구 작업이 완료된 다음 두 번째 노드에서 장애가 발생함을 의미합니다. 이 문제는 여러 번 발생할 수 있으며 VDC 가 4 개 사이트 → 3 개 사이트 → 2 개 사이트 → 1 개 사이트가 되는 것과 비슷합니다. 이렇게 하려면 나머지 노드에 복구 작업을 완료하기에 충분한 공간이 필요합니다.
- **동시에 장애 발생**: 노드에 동시에 장애가 발생한다는 것은 노드 장애가 거의 동시에 발생하거나 장애가 발생한 이전 노드에서 복구가 완료되기 전에 다른 노드에 장애가 발생하는 것을 말합니다.

장애의 영향은 어떤 노드가 다운되는가에 따라 달라집니다. 표 7 및 표 8 에 삭제 코딩 체계와 VDC 의 노드 수에 따라 한 사이트의 장애 허용 능력에 대한 최상의 시나리오가 설명되어 있습니다.

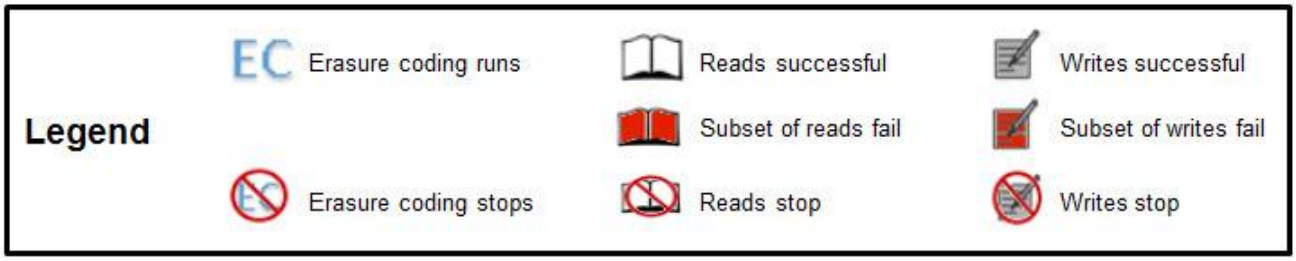

표 7 기본 12+4 삭제 코딩 체계를 기반으로 한 사이트 내의 다중 노드 장애에 대한 최상의 시나리오

생성 시 VDC 의 노드 수	VDC 생성 이후 장애가 발생한 총 노드 수	동시 장애 후 상태	최근 하나씩 장애 발생 후 상태	이전에 하나씩 장애가 발생한 이후 현재 VDC 상태
5 노드	1	EC	EC	1 개 노드에 장애가 발생한 5 노드 VDC
	2	EC	EC	VDC 가 전에 5 → 4 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
	3~4	EC	EC	VDC 가 전에 5 → 4 → 3 또는 5 → 4 → 3 → 2 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
6 노드	1	EC	EC	1 개 노드에 장애가 발생한 6 노드 VDC

생성 시 VDC 의 노드 수	VDC 생성 이후 장애가 발생한 총 노드 수	동시 장애 후 상태	최근 하나씩 장애 발생 후 상태	이전에 하나씩 장애가 발생한 이후 현재 VDC 상태
	2	EC	EC	VDC 가 전에 6 → 5 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
	3			VDC 가 전에 6 → 5 → 4 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
	4~5			VDC 가 전에 6 → 5 → 4 → 3 또는 6 → 5 → 4 → 3 → 2 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
8 노드	1~2	EC	EC	8 노드 VDC 또는 VDC 가 전에 8 → 7 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생
	3~4	EC	EC	VDC 가 전에 8 → 7 → 6 또는 8 → 7 → 6 → 5 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생
	5			VDC 가 전에 8 → 7 → 6 → 5 → 4 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생
	6~7			VDC 가 전에 8 → 7 → 6 → 5 → 4 → 3 개 노드 또는 8 → 7 → 6 → 5 → 4 → 3 → 2 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생

표 8 콜드 스토리지 10+2 삭제 코딩 체계를 기반으로 한 사이트 내의 다중 노드 장애에 대한 최상의 시나리오

생성 시 VDC 의 노드 수	VDC 생성 이후 장애가 발생한 총 노드 수	동시 장애 후 상태	최근 하나씩 장애 발생 후 상태	전에 하나씩 장애가 발생한 이후 현재 VDC 상태
6 노드	1			1 개 노드에 장애가 발생한 6 노드 VDC
	2			VDC 가 전에 6 → 5 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
	3			VDC 가 전에 6 → 5 → 4 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
	4~5			VDC 가 전에 6 → 5 → 4 → 3 또는 6 → 5 → 4 → 3 → 2 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
8 노드	1			1 개 노드에 장애가 발생한 8 노드 VDC
	2			VDC 가 전에 8 → 7 개 노드가 되었고, 이제 추가로 1 개의 노드에서 장애 발생
	3~5			VDC 가 전에 8 → 7 → 6 또는 8 → 7 → 6 → 5 또는 8 → 7 → 6 → 5 → 4 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생
	6~7			VDC 가 전에 8 → 7 → 6 → 5 → 4 → 3 개 노드 또는 8 → 7 → 6 → 5 → 4 → 3 → 2 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생
12 노드	1~2			12 노드 VDC 또는 12 노드 VDC 가 전에 12 → 11 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생

생성 시 VDC 의 노드 수	VDC 생성 이후 장애가 발생한 총 노드 수	동시 장애 후 상태	최근 하나씩 장애 발생 후 상태	전에 하나씩 장애가 발생한 이후 현재 VDC 상태
	3~6			VDC 가 전에 12→ 11 → 10 또는 12→ 11 → 10 → 9 또는 12→ 11 → 10 → 9 → 8 또는 12→ 11 → 10 → 9 → 8 → 7 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생
	7~9			VDC 가 전에 12→ 11 → 10 → 9 → 8 → 7 → 6 또는 12→ 11→ 10→ 9→ 8 → 7→ 6→ 5 또는 12→ 11→ 10→ 9→ 8→ 7→ 6→ 5→ 4 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생
	10~11			VDC 가 전에 12→11→10→9→8→7→6→5→4→3 또는 12→11→10→9→8→7→6→5→4→3→ 2 개 노드가 되었고 이제 추가로 1 개의 노드에서 장애 발생

하나의 사이트에서 여러 노드 장애가 발생했을 때 어떤 작업이 실패하는지 파악하는 기본 규칙은 다음과 같습니다.

- 노드 장애가 3 개 이상 동시에 발생하는 경우 연결된 3 중 미러링된 메타데이터 청크의 복제본 3 개가 모두 손실될 수 있기 때문에 일부 읽기와 쓰기가 실패합니다.
- 쓰기는 최소 3 개의 노드가 필요합니다.
- 노드 수가 각 삭제 코딩 체계에 필요한 최소 수보다 작으면 삭제 코딩이 중지되고 삭제 코딩된 청크가 3 중 미러 보호로 변환됩니다. 기본 삭제 코딩 체계인 12+4 는 4 개의 노드가 필요하므로 노드가 4 개 미만일 경우 삭제 코딩이 중단됩니다. 콜드 스토리지 삭제 코딩인 10+2 의 경우 노드가 6 개 미만일 경우 삭제 코딩이 중단됩니다.

- 노드 수가 삭제 코딩 체계에 필요한 최소 수 이하로 내려가면 삭제 코딩 청크가 3 중 미러 보호로 변환됩니다. 예를 들어, 기본 삭제 코딩을 하고 4 개 노드가 있는 VDC 에서 노드 장애가 발생하면 다음과 같은 상황이 발생합니다.

  - 노드 장애로 인해 4 개 조각이 손실됩니다.
  - 누락된 조각이 재구축됩니다.
  - 청크는 각 노드에 하나씩 3 개의 복제 복사본을 생성합니다.
  - EC 복제본이 삭제됩니다.

- 하나씩 장애가 발생할 경우 하나의 노드 장애와 같은 방식으로 처리됩니다. 예를 들어 두 개의 노드를 하나씩 잃을 경우 각 장애에 대한 대처는 하나의 노드 운영 중단 시 데이터를 복구하는 작업으로만 구성됩니다.

  예를 들어, 6 개의 노드가 있고 기본 삭제 코딩을 할 경우:

  - 첫 번째 장애: 각 6 개 노드에는 최대 3 개의 조각(16 개 조각/6 개 노드)이 있습니다. 누락된 3 조각은 나머지 노드에 다시 생성됩니다. 복구가 완료된 후 VDC 는 5 개의 노드가 됩니다.
  - 두 번째 장애: 각 5 개 노드에는 최대 4 개의 조각(16 개 조각/5 개 노드)이 있습니다. 누락된 4 조각은 나머지 노드에 다시 생성됩니다. 복구가 완료된 후 VDC 는 4 개의 노드가 됩니다.
  - 세 번째 장애: 각 4 개 노드에는 최대 4 개의 조각(16 개 조각/4 개 노드)이 있습니다. 누락된 4 조각은 나머지 노드에 다시 생성됩니다. 복구가 완료된 후 VDC 는 3 개의 노드가 되고 삭제 코딩을 위한 최소 수보다 적기 때문에 삭제 코딩 청크는 나머지 노드에 분산된 3 개의 복제 복사본으로 대체됩니다.
  - 네 번째 장애: 세 개의 노드 각각에 하나의 복제 복사본이 있습니다. 누락된 복제 복사본은 나머지 노드 중 하나에 다시 생성됩니다. 복구가 완료된 후 VDC 는 2 개의 노드가 됩니다.
  - 다섯 번째 장애: 3 개의 복제 복사본이 있으며, 한 노드에 2 개, 다른 노드에 1 개 있습니다. 누락된 복제 복사본은 나머지 노드에 다시 생성됩니다. 복구가 완료된 후 VDC 는 1 개의 노드가 됩니다.

3 다중 사이트 설계 개요

하나의 사이트 내에 설계된 시스템 가용성과 데이터 내구성 외에, ECS 는 전체 사이트 장애에 대한 보호 기능도 제공합니다. 이는 다중 사이트 배포에서 여러 VDC/사이트를 함께 페더레이션하고 지리적 복제를 구성하여 수행됩니다.

사이트 페더레이션에는 사이트 간 통신을 위한 복제와 관리 엔드포인트를 제공하는 것이 포함됩니다. 사이트가 페더레이션되면 하나의 인프라스트럭처로 관리할 수 있습니다.

복제 그룹 정책이 데이터가 보호되는 방식과 액세스 가능한 위치를 결정합니다. ECS 는 지리적 액티브 복제와 지리적 패시브 복제를 모두 지원합니다. 지리적 액티브 복제는 정의된 복제 그룹 내의 모든 사이트에서 데이터를 읽고 쓸 수 있도록 데이터에 대한 액티브-액티브 액세스를 제공합니다.

EXF900 과 같은 올플래시 시리즈를 사이트 사이에서 복제할 때는 WAN 을 사용함으로 인해 발생할 수 있는 성능 영향을 고려해야 합니다. 수집량이 많을 경우 링크에 로드가 많이 걸려 포화 상태 또는 RPO 지연이 발생할 수 있으며, 사용자/애플리케이션에서 로컬 요청에 비해 원격 읽기와 쓰기 레이턴시가 길어질 수 있습니다. 또 하나 고려할 것은 부분적인 가비지 컬렉션이 실패한다는 것입니다. 로컬 사이트와 복제 사이트 모두에서 대량의 수집이 이루어지면 시스템 로드가 곧 90%에 도달하여 데이터 쓰기와 데이터 회수가 중지될 수 있습니다.

참고: 부분 가비지 컬렉션 - 청크가 2/3 가비지인 경우, 유효한 부분을 다른 부분적으로 채워진 청크와 함께 새로운 청크로 병합하여 청크를 재확보한 다음 공간을 재확보합니다.

복제를 지리적 패시브로 구성할 수도 있습니다. 이는 2~4 개의 소스 사이트와 1~2 개의 사이트를 복제 타겟으로만 사용하도록 지정합니다. 복제 타겟은 복구 목적으로만 사용됩니다. 복제 타겟은 생성/업데이트/삭제 작업을 위해 직접 클라이언트 액세스를 차단합니다.

지리적 패시브 복제의 이점은 다음과 같습니다.

- 두 소스 사이트의 쓰기가 동일한 복제 타겟에 전송되도록 함으로써 XOR 작업의 실행 가능성을 높여 스토리지 효율성을 최적화할 수 있습니다.
- 덕분에 관리자는 클라우드 백업과 같은 상황에서 데이터 복제 복사본이 있을 위치를 제어할 수 있습니다.

ECS 는 버킷 레벨에서 지리적 복제 구성 옵션을 제공하므로 관리자가 버킷마다 서로 다른 수준의 복제를 구성할 수 있습니다.

그림 7 은 관리자가 세 버킷의 복제를 구성하는 방법의 예를 보여 줍니다.

- 버킷 A: 엔지니어링 테스트/개발 데이터 - 복제하지 않고 로컬에서만 보관
- 버킷 B: 유럽 영업 데이터 - 유럽 내 사이트 간에만 복제
- 버킷 C: 전사적 교육 데이터 - 회사 내 모든 사이트에 복제

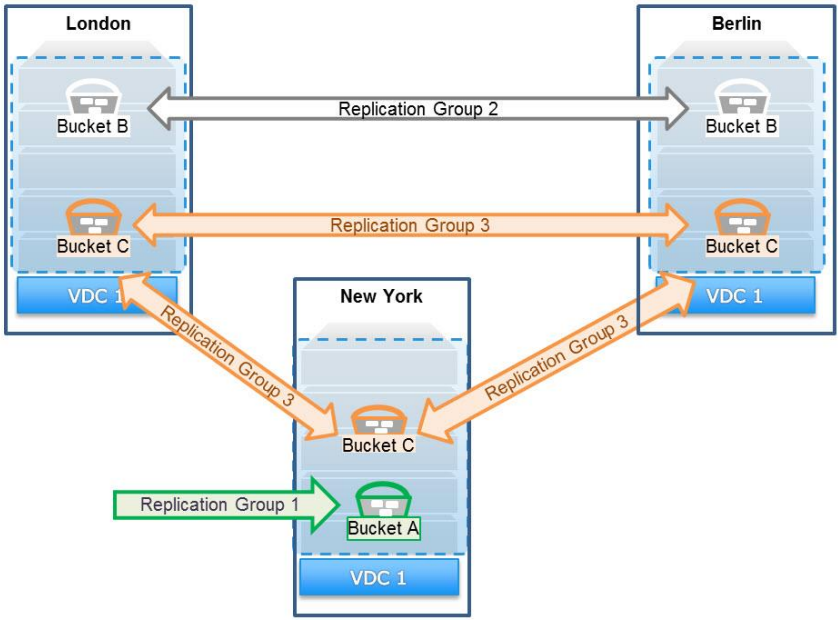

그림 7 서로 다른 버킷에서 서로 다른 복제 정책을 사용하는 방법을 보여 주는 샘플

모범 사례로, 특정 복제 경로에 대해 복제 그룹을 구성하는 것이 좋습니다. 예를 들어 그림 7 을 보면 런던과 베를린 사이에 데이터를 복제하는 복제 그룹이 있습니다. 이는 런던과 베를린 사이만의 복제가 필요한 모든 버킷에 사용해야 합니다.

지리적 복제 데이터는 로컬 사이트에 데이터의 주 복제본을 저장하고 하나 이상의 원격 사이트에 데이터의 보조 복사본을 저장하여 보호합니다. 복제 그룹에 구성된 사이트 수, 사이트 간 데이터 기록 방법, **모든 사이트로의 복제**가 활성화되었는지 여부에 따라 복제본 수와 보조 복제본에 사용되는 공간의 양이 결정됩니다.

각 사이트는 로컬 데이터 보호를 담당합니다. 즉 로컬, 보조 복제본 모두 삭제 코딩 및/또는 3 중 미러링을 사용하여 데이터를 각각 보호합니다. 각 사이트의 삭제 코딩 체계는 동일할 필요가 없습니다. 즉 한 사이트는 12+4 의 기본 삭제 코딩 체계를 사용하고 다른 사이트는 10+2 의 콜드 스토리지 삭제 코딩 체계를 사용할 수 있음을 의미합니다.

복제된 데이터는 HTTP 를 사용하여 다른 사이트로 전송되기 전에 암호화되고(AES256) 압축됩니다.

사이트 간의 일관성을 유지하기 위해서는 최신 버전의 메타데이터를 관리하는 일을 담당하는 주체가 있어야 합니다. 이 주체는 사이트 수준에서 정의되며 네임스페이스, 버킷, 오브젝트의 소유권을 결정합니다. 소유권 정보는 처음에 소유자 사이트에 저장되지만 ECS 메타데이터의 일부로 다른 사이트에도 복제됩니다.

- 권한 있는 버전: 권한 있는 버전은 항상 소유자이며 강력한 일관성을 위해 사용됩니다.
- 복제된 버전: 복제된 버전은 최신 버전이 아닐 수 있지만 다음을 포함한 장애 처리 작업 중에 사용됩니다.
  - **운영 중단 중 액세스**가 활성화된 경우(일시적 일관성).
  - 그리고 영구 사이트 페일오버 작업 중.

버킷 및 오브젝트 소유자의 권한 있는 버전이 있습니다.

네임스페이스 소유자:

- 네임스페이스를 생성하는 사이트는 네임스페이스 소유자입니다.
- 이 소유자가 버킷 목록의 권한 있는 버전을 유지 관리합니다.

버킷 소유자:

- 버킷을 생성하는 사이트는 버킷 소유자입니다.
- 이 소유자가 다음의 권한 있는 버전을 유지 관리할 책임이 있습니다.
  - 버킷에 있는 오브젝트의 최신 버전이 포함된 버킷 목록.
  - 버킷 내의 오브젝트에 대한 오브젝트 소유권 목록.

오브젝트 소유자:

- 처음에 오브젝트를 처음 생성한 사이트가 오브젝트 소유자입니다. 이는 변경될 수 있습니다. 자세한 내용은 "운영 중단 중 액세스" 섹션을 참조하십시오.
- 이 소유자가 오브젝트 메타데이터의 권한 있는 버전을 유지 관리하는 역할을 합니다.

3.1 청크 관리자 테이블

청크 위치는 모든 사이트에 독립적으로 저장되는 청크 관리자 테이블에 유지됩니다. 청크가 생성되는 위치를 원래 주 사이트라고 하고, 청크가 복제되는 위치를 보조 사이트라고 합니다.

청크가 생성되면 주 사이트와 보조 사이트가 결정되고 주 사이트는 복제 그룹의 다른 노드에 청크의 사이트 정보를 브로드캐스트합니다.

또한 각 사이트는 자체 청크 관리자 테이블을 유지 관리하기 때문에 각 청크의 속성 유형에 대한 정보도 가지고 있습니다. 속성 유형에는 다음이 포함됩니다.

- **로컬**: 청크가 생성된 사이트
- **복사**: 청크가 복제된 사이트
- **원격**: 청크와 그 복제본이 로컬로 저장되지 않은 사이트
- **패리티**: 다른 청크의 XOR 작업 결과가 포함된 청크(자세한 내용은 아래 XOR 섹션 참조)
- **인코딩됨**: 데이터가 XOR 데이터로 로컬로 대체된 청크(자세한 내용은 아래 XOR 섹션 참조)

표 9~표 11 은 세 사이트 각각에서 청크 관리자 테이블 목록의 샘플 부분을 보여 줍니다.

표 9 사이트 1 의 샘플 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 2	로컬
C2	사이트 2	사이트 3	원격
C3	사이트 1	사이트 3	로컬
C4	사이트 2	사이트 1	복사

표 10 사이트 2 의 샘플 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 2	복사
C2	사이트 2	사이트 3	로컬
C3	사이트 1	사이트 3	원격
C4	사이트 2	사이트 1	로컬

표 11 사이트 3 의 샘플 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 2	원격
C2	사이트 2	사이트 3	복사
C3	사이트 1	사이트 3	복사
C4	사이트 2	사이트 1	원격

3.2 XOR 인코딩

3 개 이상의 사이트를 포함하는 복제 그룹으로 구성된 데이터의 스토리지 효율성을 극대화하기 위해 ECS 는 XOR 인코딩을 활용합니다. 복제 그룹의 사이트 수가 증가할수록 ECS 알고리듬은 오버헤드를 줄이는 데 더 효율적입니다.

XOR 인코딩은 각 사이트에서 수행됩니다. 이는 청크 관리자 테이블을 스캔하고 복제 그룹의 다른 각 사이트에서 가져온 COPY 유형 청크를 찾으면 그 청크에 대해 XOR 인코딩을 할 수 있습니다. 예를 들어, 표 12 는 서로 다른 주 사이트를 가진 COPY 유형인 청크 **C2** 와 **C3** 을 모두 포함하는 3 개 사이트 구성의 사이트 3 을 보여줍니다. 이를 사용하여 사이트 3 이 청크를 함께 XOR 하고 결과를 저장할 수 있습니다. 결과는 새 청크인 **C5** 로, **C2** 와 **C3** 의 XOR(수학적으로 C2 ⊕ C3) 보조 사이트 없이 **패리티**로 나열되는 형식을 가집니다. 패리티 청크의 청크 ID 는 다른 사이트로 브로드캐스트되지 않습니다.

표 12 는 사이트 3 이 청크 **C2** 와 **C3** 을 함께 청크 **C5** 로 XOR 하는 동안 사이트 3 의 청크 관리자 테이블의 예를 보여 줍니다.

표 12 XOR 작업 중 사이트 3 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 2	원격
C2	사이트 2	사이트 3	복사
C3	사이트 1	사이트 3	복사
C4	사이트 2	사이트 1	원격
C5	사이트 3		패리티(C2 및 C3)

XOR 이 완료되면 **C2** 와 **C3** 의 데이터 복제본이 삭제되어 디스크 공간이 확보되며 이러한 청크의 청크 관리자 테이블 유형이 인코딩됨 유형으로 변경됩니다. XOR 작업은 순전히 보조 사이트 작업이며 주 사이트는 청크가 인코딩되었음을 인식하지 못합니다. XOR 인코딩이 완료되고 **C2** 와 **C3** 에 대한 데이터 복제본이 삭제되면 사이트 3 의 청크 관리자 테이블은 표 13 처럼 나열됩니다.

표 13 C2 와 C3 의 XOR 인코딩을 완료한 후 사이트 3 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 2	원격
C2	사이트 2	사이트 3	인코딩됨
C3	사이트 1	사이트 3	인코딩됨
C4	사이트 2	사이트 1	원격
C5	사이트 3		패리티(C2 및 C3)

인코딩된 청크의 데이터 요청은 주 복제본을 포함하는 사이트에서 서비스합니다. 스토리지 효율성에 대한 자세한 내용은 ECS 개요 및 아키텍처 백서를 참조하십시오.

3.3 모든 사이트에 복제

모든 사이트에 복제는 세 개 이상의 사이트가 있고 모든 청크를 복제 그룹에 구성된 모든 VDC/사이트에 복제하려는 경우에 사용되는 복제 그룹 옵션입니다. 이는 XOR 작업의 실행도 막습니다. **모든 사이트에 복제** 옵션의 각 값과 효과:

- **활성화됨:** 쓴 데이터 복제본의 수가 복제 그룹에 있는 사이트 수와 같습니다. 예를 들어, 복제 그룹에 사이트가 4 개 있으면 주 복제본과 각 사이트에 하나씩 있는 세 개의 보조 복제본을 갖게 됩니다.
- **비활성화됨:** 쓴 데이터의 복제본 수가 두 개입니다. 총 사이트 수에 관계없이 주 복제본과 원격 사이트에 복제된 복제본 하나가 있습니다.

참고: 지리적 패시브로 구성된 복제 그룹에서는 모든 사이트로 복제를 활성화할 수 없습니다.

이 설정은 기본적으로 모든 데이터를 두 사이트 모두에 이미 복제하므로 두 사이트만 포함하는 복제 그룹에는 영향을 미치지 않습니다. 관리자는 이 복제 그룹을 사용할 버킷을 선택할 수 있습니다.

모든 사이트에 복제를 활성화하면 다음과 같은 영향이 있습니다.

- 복제가 완료된 후 후속 읽기가 로컬에서 처리되므로 읽기 성능을 향상시킬 수 있습니다.
- XOR 디코딩으로 인한 성능 저하가 사라집니다.
- 데이터 내구성이 향상됩니다.
- 스토리지 활용 효율이 낮아집니다.
- 지리적 복제에 대한 WAN 활용도를 높입니다. 높아지는 정도는 복제 그룹의 VDC/사이트 수에 비례합니다.
- 복제된 데이터의 읽기에 대한 WAN 활용도가 낮아집니다.

이러한 이유로 다음 조건을 충족하는 환경의 특정 버킷에서만 이 옵션을 활성화하는 것이 좋습니다.

- 지리적으로 분산된 사이트의 동일한 데이터에 대한 읽기 작업이 많은 워크로드.
- 복제 그룹의 사이트 간 지리적 복제 트래픽을 늘리기에 충분한 WAN 대역폭이 있는 인프라스트럭처.
- 스토리지 활용 효율로 인한 비용보다 읽기 성능이 중요한 워크로드.

3.4 지리적 복제 환경에서 데이터 쓰기 흐름

청크에는 동일한 복제 그룹 설정을 공유하는 버킷의 하나 이상의 오브젝트로 구성된 128MB 의 데이터가 포함됩니다. 복제는 비동기적으로 수행되며 청크 파티션 소유자에 의해 청크 수준에서 시작됩니다. 청크가 지리적 복제 데이터로 구성되면 주 사이트의 청크에 기록될 때 복제 대기열에 추가됩니다. 청크가 밀폐되기를 기다리지 않습니다. 대기열을 계속 처리하는 작업자 I/O 스레드가 있습니다.

쓰기 작업은 데이터 보호 추가를 포함하여 로컬에서 먼저 실행된 다음 원격 사이트에서 복제되고 보호됩니다. 그림 8 은 128MB 오브젝트를 지리적 복제된 버킷에 쓰는 프로세스의 예를 보여 줍니다.

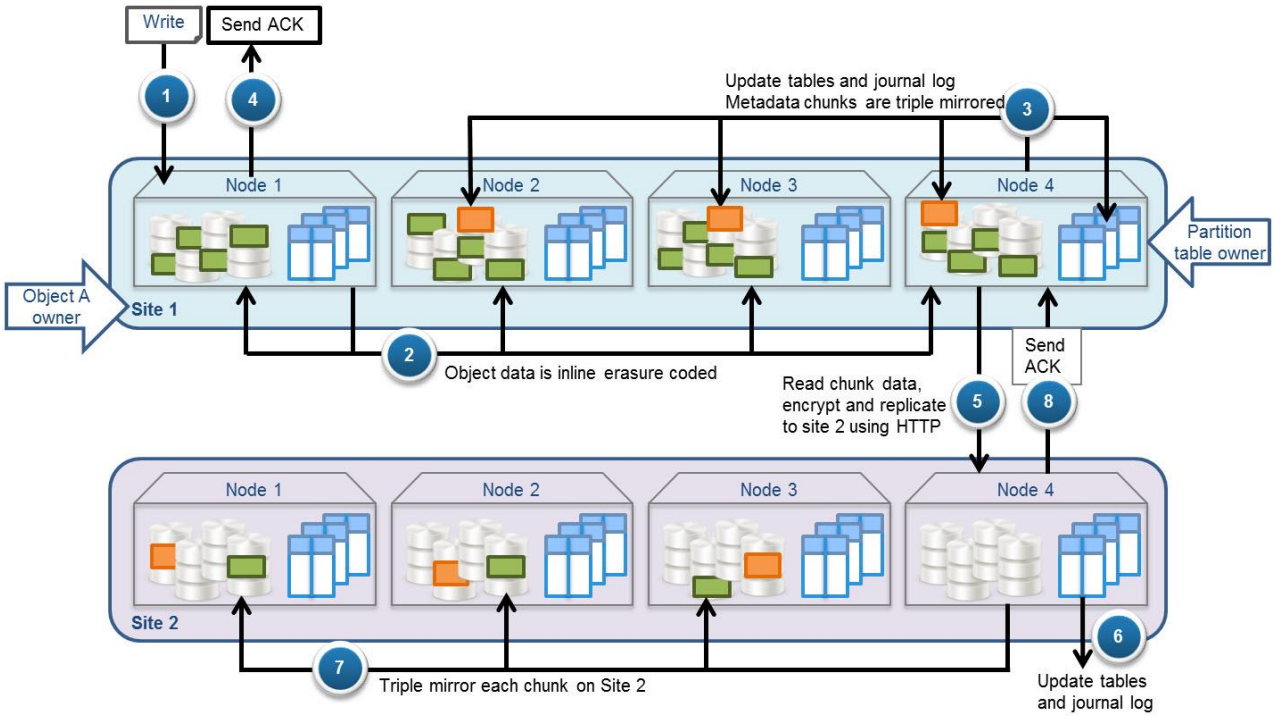

그림 8 128MB 오브젝트를 지리적 복제된 버킷에 쓰는 데이터 쓰기 워크플로

1. 오브젝트 A 에 대한 쓰기 요청이 노드(이 예에서는 사이트 1 노드 1)로 전송됩니다. 사이트 1 이 오브젝트 A 의 소유자가 됩니다.
2. 데이터는 인라인 삭제 코딩되어 사이트 1 의 청크에 기록됩니다.

3. 이 예에서 노드 4 인 테이블 파티션 소유자가 적절한 테이블(예: 청크, 오브젝트, 버킷 목록 테이블)을 업데이트하고 그 트랜잭션을 저널 로그에 기록합니다. 이 메타데이터는 사이트 1 에서 3 중 미러링된 메타데이터 청크에 기록됩니다.
4. 쓰기 성공 확인이 클라이언트에 전송됩니다.
5. 각 청크에 대해 청크 파티션 테이블 소유자, 이 예에서는 노드 4 가 다음을 합니다.

   a. 청크 내부의 데이터를 로컬로 쓴 후 복제 대기열에 추가합니다. 청크가 밀폐될 때까지 기다리지 않습니다.
   b. 청크의 데이터 조각을 읽습니다. 패리티 조각은 누락된 데이터 조각을 다시 만드는 데 필요한 경우에만 읽습니다.
   c. 데이터를 암호화하고 HTTP 를 사용하여 사이트 2 로 복제합니다.

6. 복제된 청크의 테이블 파티션 소유자, 이 예에서는 사이트 2 노드 4 가 적절한 테이블을 업데이트하고 트랜잭션을 3 중 미러링된 저널 로그에 씁니다.
7. 복제된 각 청크는 처음에 3 중 미러링을 사용하여 두 번째 사이트에 기록됩니다.
8. 주 사이트 청크 파티션 테이블 소유자에게 확인이 다시 전송됩니다.

참고: 복제된 사이트에 쓴 데이터는 지연 후 삭제 코딩되므로 XOR 작업과 같은 다른 프로세스가 먼저 완료될 수 있습니다.

3.5 지리적 복제 환경에서 데이터 읽기 흐름

ECS 는 복제 그룹 내의 여러 VDC 에 비동기식으로 데이터를 복제하므로 사이트/VDC 간에 데이터의 일관성을 보장하는 방법이 필요합니다. ECS 는 오브젝트 소유자인 사이트에서 메타데이터의 최신 복제본을 가져와 강력한 일관성을 보장합니다. 요청 사이트에 오브젝트의 복사본(청크 유형 = 로컬 또는 복제본)이 들어 있으면 읽기 요청을 처리하는 데 사용되며, 그렇지 않으면 오브젝트 소유자에서 데이터를 가져옵니다. 데이터 읽기 흐름을 보여주는 예를 그림 9 에서 볼 수 있습니다. 이는 오브젝트 소유자 사이트가 아닌 다른 사이트에서 오브젝트 A 에 대한 읽기 요청을 하는 경우를 보여줍니다.

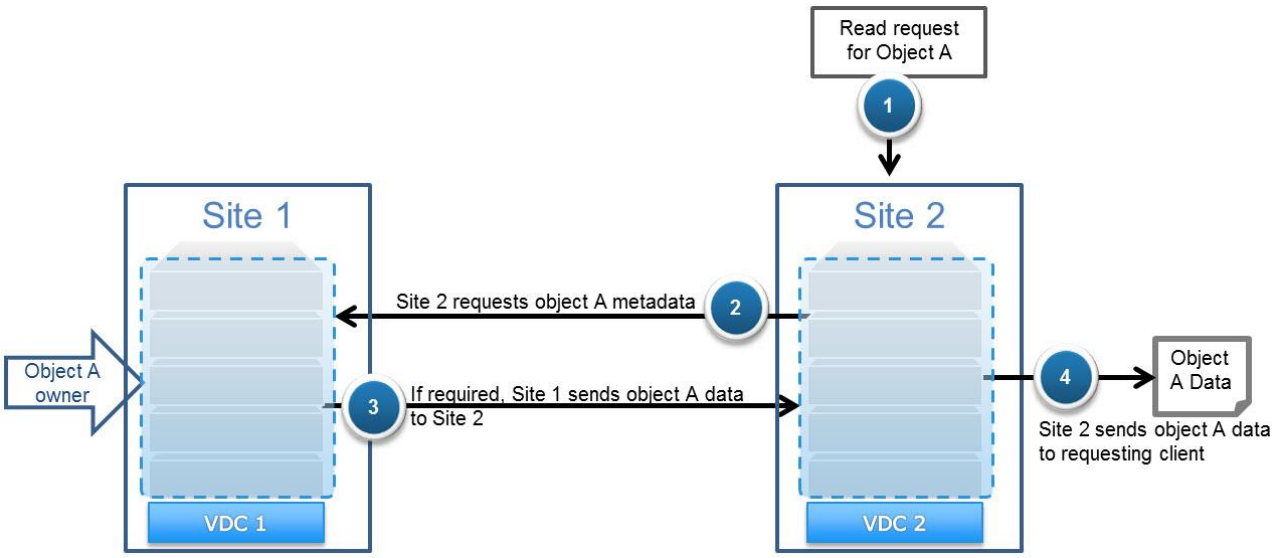

그림 9 다른 사이트가 소유한 지리적 복제 오브젝트에 대한 데이터 읽기 워크플로

이 예에서 읽기 작업의 흐름은 다음과 같습니다.

1. 사이트 2 가 사이트 1 이 소유한 오브젝트 A 에 대한 읽기 요청을 받습니다.
2. 사이트 2 는 최신 버전의 메타데이터를 가져오기 위해 버킷 및 오브젝트 소유자, 이 예에서는 사이트 1 에 연락합니다.

오브젝트 소유권:

- **운영 중단 중 액세스**가 비활성화되어 있으면 로컬 정보를 확인하여 오브젝트 소유자인지 확인합니다. 그렇지 않으면 버킷 소유자에게 연락하여 오브젝트 소유자를 확인합니다.
- **운영 중단 중 액세스**가 버킷에서 활성화되어 있으면 요청 사이트에서 버킷 소유자에게 현재 오브젝트 소유자를 확인합니다.

3. 사이트 2 에 오브젝트의 복제본(일반 유형 = 로컬 또는 복제본)이 없으면 사이트 1 이 오브젝트 A 데이터를 사이트 2 로 보냅니다.
4. 사이트 2 가 오브젝트 A 데이터를 요청 클라이언트로 보냅니다.

3.6 지리적 복제된 환경의 데이터 업데이트 흐름

ECS 는 연결된 버킷 복제 그룹 내에 있는 노드의 데이터를 액티브-액티브로 업데이트할 수 있도록 설계되었습니다. 이렇게 하려면 오브젝트 소유자가 아닌 사이트가 오브젝트 업데이트에 대한 정보를 주 소유자 사이트에 동기식으로 보내고 확인을 기다린 후 클라이언트에게 확인을 다시 보내야 합니다. 업데이트된 오브젝트와 관련된 데이터는 일반적인 비동기 청크 복제 작업의 일부로 복제됩니다. 데이터가 소유 사이트로 아직 복제되지 않은 상태에서 데이터에 대한 읽기 요청을 받으면 원격 사이트에서 데이터를 요청합니다. 그림 10 은 오브젝트 소유자 사이트가 아닌 다른 사이트에서 오브젝트 A 에 대한 업데이트를 요청하는 과정의 예를 보여줍니다.

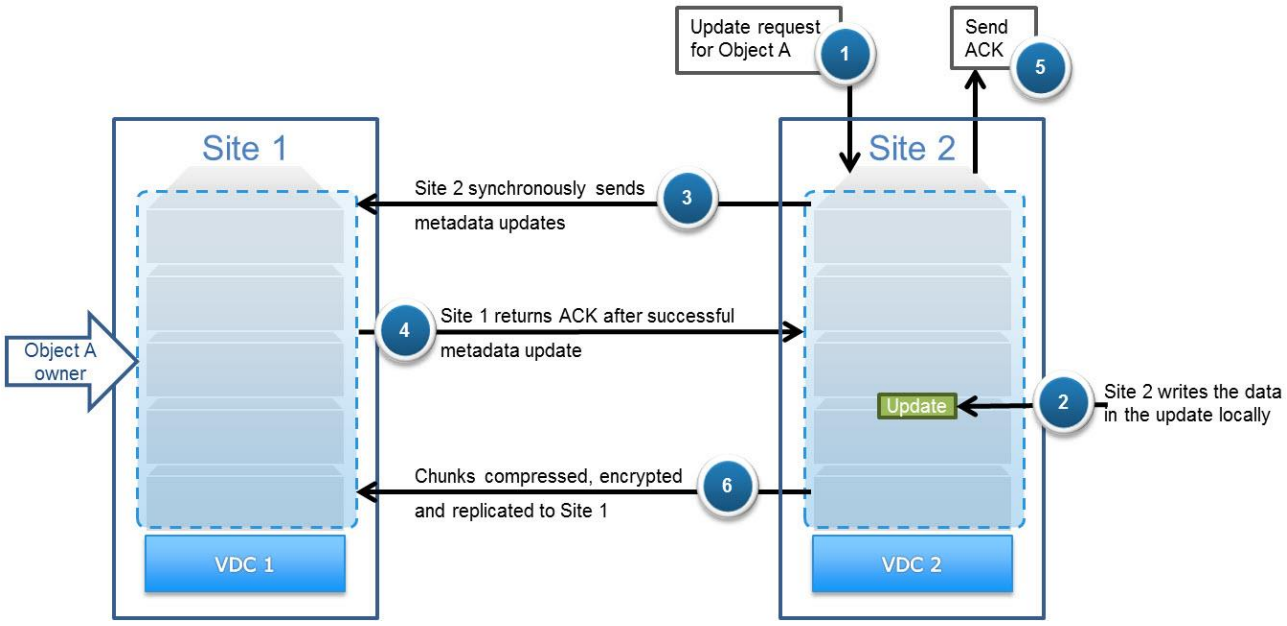

그림 10 다른 사이트가 소유한 지리적 복제 오브젝트에 대한 오브젝트 업데이트 워크플로

이 예에서 업데이트 작업의 흐름은 다음과 같습니다.

1. 사이트 2 가 사이트 1 이 소유한 오브젝트 A 에 대한 업데이트 요청을 받습니다.
2. 사이트 2 가 업데이트의 데이터를 로컬로 씁니다.
3. 사이트 2 가 메타데이터 업데이트를 오브젝트 소유자, 이 예에서 사이트 1 에게 동기식으로 보냅니다.

참고: 운영 중단 중 액세스가 버킷에 활성화되어 있거나 사이트 2 가 오브젝트 소유자가 아닌 경우 먼저 버킷 소유자에게 연락하여 현재 오브젝트 소유자를 확인합니다.

---

4. 사이트 1 이 메타데이터 업데이트에 성공했다는 확인을 사이트 2 로 보냅니다.
5. 사이트 2 가 요청한 클라이언트에게 업데이트에 성공했다는 확인을 보냅니다.
6. 청크가 복제 대기열에 추가되고 평소처럼 암호화되고 비동기식으로 사이트 1 에 복제됩니다.

정상 조건에서는 업데이트가 시작된 사이트에 관계없이 업데이트를 한 후에도 오브젝트 소유자가 변경되지 않습니다. 이 예에서 오브젝트 소유자는 사이트 2 에서 시작된 업데이트 성공 후에도 사이트 1 로 유지됩니다. 유일한 예외는 **운영 중단 중 액세스**가 활성화된 상태에서 임시 사이트 운영 중단이 발생하는 것입니다. 자세한 내용은 섹션 4.1.2 를 참조하십시오.

4 다중 사이트 가용성

ECS 는 강력한 일관성을 제공하므로 응답하기 전에 소유자와 I/O 요청을 확인해야 합니다. 이로 인해 사이트에 액세스할 수 없으면 일부 버킷에 대한 액세스가 일시적으로 중단될 수 있습니다.

사이트 운영 중단은 다음과 같은 다양한 이유로 인해 발생할 수 있습니다.

- 페더레이션 사이트 간의 네트워크 연결이 끊기거나 건물 정전과 같은 전체 사이트의 장애와 같은 일시적 문제.
- 자연 재해 등으로 인한 영구적 문제.

일시적인 사이트 운영 중단을 탐지하기 위해 페더레이션 사이트는 사이트 간 하트비트를 설정합니다. 사이트 간에 하트비트가 지속해서 손실되는 경우(기본값은 15 분) 다음과 같이 됩니다.

- 2 사이트 구성에서는 각각이 상대 사이트를 장애가 발생한 것으로 표시합니다.
- 3 사이트 이상의 사이트 구성에서는 다음이 모두 해당할 경우만 사이트를 장애가 발생한 것으로 표시합니다.
  - 대부분의 사이트에서 동일한 ECS 사이트에 대한 하트비트가 지속적으로 손실됩니다.
  - 그리고 나머지 모든 사이트가 현재 온라인 상태로 표시되어 있습니다.

예를 들어 3 사이트 구성에서 사이트 2 와 3 이 모두 지속적으로 사이트 1 로의 네트워크 연결이 끊어지면 ECS 는 사이트 1 을 일시적으로 장애가 발생한 것으로 표시합니다.

페더레이션 사이트에 장애가 발생하면 다른 페더레이션 시스템으로의 액세스를 유도하여 시스템 가용성을 유지할 수 있습니다. 사이트 전체에 장애가 발생하면 사용할 수 없는 사이트가 소유한 지리적 복제 데이터를 일시적으로 사용할 수 없게 됩니다. 데이터를 사용할 수 없는 기간은 다음에 의해 결정됩니다.

- **운영 중단 중 액세스** 활성화 여부
- 임시 사이트 운영 중단이 남아 있는 기간
- 영구 사이트 페일오버가 복구 작업을 완료하는 데 걸리는 시간

사이트 장애는 일시적일 수도 있고 영구적일 수도 있습니다. 일시적인 사이트 운영 중단은 사이트를 다시 온라인 상태로 돌릴 수 있음을 의미하며 일반적으로 정전이나 사이트 간의 네트워킹 손실로 인해 발생합니다. 영구적인 사이트 운영 중단은 연구실 화재와 같이 전체 시스템을 복구할 수 없는 경우입니다. 관리자만 사이트 운영 중단이 영구적인지 확인하고 복구 작업을 시작할 수 있습니다.

4.1 TSO(Temporary Site Outage)

복제 그룹의 다른 사이트에 일시적으로 액세스할 수 없는 사이트가 있을 때 사이트 운영 중단이 발생합니다. 관리자는 ECS 를 사용하여 임시 사이트 운영 중단 중 오브젝트에 액세스할 수 있는 방법에 영향을 미치는 두 가지 구성 옵션을 설정할 수 있습니다.

- 다음 방법으로 강력한 일관성을 유지하는 ADO(**Access During Outage**) 옵션을 비활성화합니다.
  - 액세스할 수 있는 사이트가 소유한 데이터에 대한 액세스를 계속 허용합니다.
  - 액세스할 수 없는 사이트에서 소유한 데이터에 대한 액세스를 차단합니다.
- **운영 중단 중 액세스** 옵션을 활성화합니다. 이렇게 하면 장애가 발생한 것으로 표시된 사이트를 포함하여 모든 지리적 복제 데이터에 대한 읽기와 쓰기 액세스를 선택적으로 허용합니다. **운영 중단 중 액세스**가 활성화된 TSO 의 경우 버킷의 데이터가 일시적 일관성으로 전환됩니다. 모든 사이트가 다시 온라인 상태가 되면 다시 강력한 일관성을 띄게 됩니다.

기본값은 **운영 중단 중 액세스**를 비활성화하는 것입니다.

운영 중단 중 액세스 옵션은 버킷 수준에서 설정할 수 있습니다. 즉, 일부 버킷에는 이 옵션을 활성화하고 다른 버킷에는 비활성화할 수 있습니다. 이 버킷 옵션은 모든 사이트가 온라인 상태인 한 언제든지 변경할 수 있으며 사이트 장애 중에는 변경할 수 없습니다.

일시적 사이트 운영 중단 동안:

- 버킷, 네임스페이스, 오브젝트 사용자, 인증 공급자, 복제 그룹, NFS 사용자, 그룹 매핑을 모든 사이트에서 생성, 삭제 또는 업데이트할 수 없습니다(영구 사이트 페일오버 중에 복제 그룹을 VDC 에서 제거할 수 있음).
- 네임스페이스 소유자 사이트에 연결할 수 없는 경우 네임스페이스의 버킷을 나열할 수 없습니다.
- 사용할 수 없는 사이트에서 소유하는 HDFS/NFS 버킷 내의 파일 시스템은 읽기 전용입니다.
- 사용할 수 없는 사이트에서 소유한 버킷의 오브젝트를 복제하는 경우 복제본이 소스 오브젝트의 전체 복제본입니다. 즉, 동일한 오브젝트의 데이터가 두 번 이상 저장됩니다. TSO 가 아닌 정상 상황에서는 오브젝트 복제본이 오브젝트 데이터의 전체 복제본이 아닌 오브젝트의 데이터 인덱스로 구성됩니다.
- TSO 시 OpenStack Swift 사용자가 OpenStack 에 로그인 할 수 없습니다. TSO 중에는 ECS 가 Swift 사용자를 인증할 수 없기 때문입니다. TSO 후 Swift 사용자를 다시 인증해야 합니다.

4.1.1 기본 TSO 동작

ECS 는 강력한 일관성을 제공하므로 IO 요청에 응답하기 전에 소유자와 확인해야 합니다. 복제 그룹 내의 다른 사이트에 액세스할 수 없으면 버킷, 오브젝트에 대한 일부 액세스가 중단될 수 있습니다.

표 14 는 작업이 성공하는 데 필요한 액세스 권한을 보여줍니다.

표 14 액세스 요구 사항

작업	성공을 위한 요구 사항
오브젝트 생성	버킷 소유자에 액세스할 수 있어야 합니다.
오브젝트 나열	요청 노드에서 버킷 소유자와 버킷의 모든 오브젝트에 액세스할 수 있어야 합니다.
오브젝트 읽기 오브젝트 업데이트	요청자가 다음이어야 합니다. • 오브젝트 소유자 및 버킷 소유자(버킷 소유권은 오브젝트가 포함된 버킷에서 **운영 중단 중 액세스**가 활성화되었을 때만 필요) • 또는 요청 노드에서 오브젝트 소유자와 버킷 소유자를 모두 액세스할 수 있음.

- 오브젝트 생성 작업에는 새 오브젝트 이름으로 버킷 목록을 업데이트하는 작업이 포함됩니다. 이렇게 하려면 버킷 소유자로의 액세스 권한이 필요하며, 요청 사이트가 버킷 소유자에 대한 액세스 권한이 없으면 실패합니다.
- 버킷에 오브젝트를 나열하려면 버킷 소유자의 목록 정보와 버킷의 각 오브젝트에 대한 헤드 정보가 모두 필요합니다. 따라서 **운영 중단 중 액세스**가 비활성화되면 다음 버킷 나열 요청이 실패합니다.
  - 요청자가 액세스할 수 없는 사이트에서 소유한 버킷을 나열하도록 요청.
  - 요청자가 액세스할 수 없는 사이트에서 소유한 오브젝트를 포함한 버킷.

- 오브젝트를 읽으려면 먼저 오브젝트 소유자로부터 오브젝트 메타데이터를 읽어야 합니다.
  - 요청 사이트가 오브젝트 소유자이고 **운영 중단 중 액세스**가 비활성화되면 요청이 성공합니다.
  - 요청 사이트가 오브젝트와 버킷 소유자이면 요청이 성공합니다.
  - 오브젝트 소유자가 로컬이 아닌 경우 사이트에서 버킷 소유자에게 확인하여 오브젝트 소유자를 찾아야 합니다. 요청자가 오브젝트 소유자 또는 버킷 소유자 사이트를 사용할 수 없으면 읽기 작업이 실패합니다.

- 오브젝트를 업데이트하려면 오브젝트 소유자의 오브젝트 메타데이터를 업데이트해야 합니다.
  - 요청 사이트가 오브젝트 소유자이고 **운영 중단 중 액세스**가 비활성화되면 요청이 성공합니다.
  - 요청 사이트가 오브젝트와 버킷 소유자이면 요청이 성공합니다.
  - 오브젝트 소유자가 로컬이 아닌 경우 사이트에서 버킷 소유자에게 확인하여 오브젝트 소유자를 찾아야 합니다. 요청자가 오브젝트 소유자 또는 버킷 소유자 사이트를 사용할 수 없으면 읽기 작업이 실패합니다.

그림 11 에 3 사이트의 예, 버킷과 오브젝트 레이아웃이 있습니다.

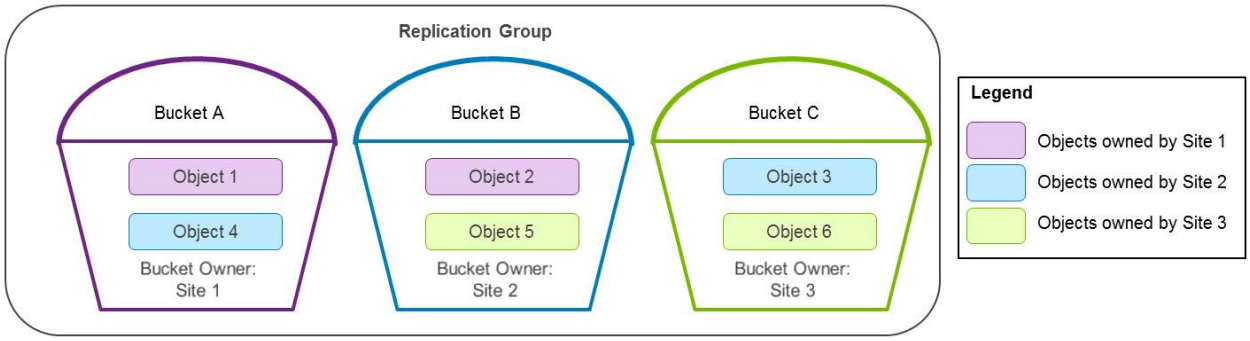

그림 11 버킷 및 오브젝트 소유권의 예

표 15 는 3 사이트 구성에서 어떤 작업이 성공하고 실패하는지 보여주며 사이트 1 이 복제 그룹의 다른 사이트에 액세스할 수 없는 경우의 예를 그림 11 에서 볼 수 있습니다. 표를 간단하게 해석할 수 있도록 액세스할 수 없는 사이트는 실패로 나열하고 나머지 두 사이트는 온라인으로 나열했습니다.

표 15 다른 사이트가 사이트 1 에 액세스할 수 없을 때 성공 또는 실패하는 작업

작업	버킷/오브젝트	요청 전송 대상		
		사이트 1(실패)	사이트 2(온라인)	사이트 3(온라인)
오브젝트 생성 위치	버킷 A	성공 로컬에서 소유한 버킷	실패 버킷 소유자에 액세스할 수 없음	실패 버킷 소유자에 액세스할 수 없음
	버킷 B	실패 버킷 소유자에 액세스할 수 없음	성공 로컬에서 소유한 버킷	성공 온라인 사이트에서 소유한 버킷
	버킷 C	실패 버킷 소유자에 액세스할 수 없음	성공 온라인 사이트에서 소유한 버킷	성공 로컬에서 소유한 버킷

작업	버킷/오브젝트	요청 전송 대상		
		사이트 1(실패)	사이트 2(온라인)	사이트 3(온라인)
오브젝트 나열 위치	버킷 A	실패 로컬에서 소유한 버킷이지만 액세스할 수 없는 사이트에서 소유한 오브젝트가 포함됨	실패 버킷 소유자에 액세스할 수 없음	실패 버킷 소유자에 액세스할 수 없음
	버킷 B	실패 버킷 소유자에 액세스할 수 없음	실패 로컬에서 소유한 버킷이지만 실패한 사이트에서 소유한 오브젝트가 포함됨	실패 버킷 소유자가 온라인 상태이지만 실패한 사이트에서 소유한 오브젝트가 버킷에 포함되어 있음
	버킷 C	실패 버킷 소유자에 액세스할 수 없음	성공 버킷 소유자가 온라인 사이트이며 모든 오브젝트가 온라인 사이트의 오브젝트임	성공 로컬에서 소유한 버킷이며 모든 오브젝트가 온라인 사이트의 오브젝트임
오브젝트 읽기 또는 업데이트	오브젝트 1	성공 로컬에서 소유한 오브젝트	실패 오브젝트 소유자에 액세스할 수 없음	실패 오브젝트 소유자에 액세스할 수 없음
	오브젝트 2	성공 로컬에서 소유한 오브젝트	실패 오브젝트 소유자에 액세스할 수 없음	실패 오브젝트 소유자에 액세스할 수 없음
	오브젝트 3	실패 오브젝트 소유자에 액세스할 수 없음	성공 로컬에서 소유한 오브젝트	성공 로컬에서 소유한 오브젝트가 아니므로 온라인 상태인 버킷 소유자로부터 오브젝트 소유자를 가져옴
	오브젝트 4	실패 오브젝트 소유자에 액세스할 수 없음	성공 로컬에서 소유한 오브젝트	실패 로컬에서 소유한 오브젝트가 아니므로 실패한 사이트인 버킷 소유자에 액세스해야 함

작업	버킷/오브젝트	요청 전송 대상		
		사이트 1(실패)	사이트 2(온라인)	사이트 3(온라인)
	오브젝트 5	실패 오브젝트 소유자에 액세스할 수 없음	성공 로컬에서 소유한 오브젝트가 아니므로 온라인 상태인 버킷 소유자로부터 오브젝트 소유자를 가져옴	성공 로컬에서 소유한 오브젝트
	오브젝트 6	실패 오브젝트 소유자에 액세스할 수 없음	성공 로컬에서 소유한 오브젝트가 아니므로 온라인 상태인 버킷 소유자로부터 오브젝트 소유자를 가져옴	성공 로컬에서 소유한 오브젝트

4.1.2 운영 중단 중 액세스가 활성화되었을 때 TSO 동작

복제 그룹 내의 다른 사이트에 처음 액세스할 수 없을 때 사이트의 동작은 기본 TSO 동작 섹션에 자세히 설명되어 있습니다. 사이트 간에 하트비트가 지속해서 손실되면(기본값은 15 분) ECS 는 사이트를 장애가 발생한 것으로 표시합니다. 버킷에서 ADO(**Access During Outage**)를 활성화하면 사이트가 장애가 발생한 표시된 후 TSO 동작이 바뀌어 이 버킷의 오브젝트가 일시적 일관성을 활용할 수 있습니다. 즉 사이트가 일시적으로 장애가 발생한 것으로 표시된 후 **운영 중단 중 액세스** 옵션을 활성화된 모든 버킷은 소유자가 아닌 사이트의 읽기를 지원하고 쓰기는 선택적으로 지원합니다. 소유자 사이트에서 권한 있는 복사본을 사용할 수 없을 때 복제된 메타데이터를 사용할 수 있게 함으로써 이렇게 할 수 있습니다. 사이트 장애 중을 제외하고 언제든지 **운영 중단 중 액세스** 버킷 옵션을 변경할 수 있습니다.

운영 중단 중 액세스를 활성화하면 사이트에 장애가 발생한 후 데이터에 액세스할 수 있다는 이점이 있지만 반환된 데이터가 오래되었을 수 있다는 단점이 있습니다.

버전 3.1 부터는 **운영 중단 증 읽기 전용 액세스**를 위한 추가 버킷 옵션이 추가되어 오브젝트 소유권이 절대 변경되지 않도록 하고, TSO 동안 장애가 발생한 사이트와 온라인 사이트 모두에서 오브젝트를 업데이트하여 충돌이 발생할 가능성을 없앴습니다. **운영 중단 증 읽기 전용 액세스**의 단점은 사이트가 장애가 발생한 것으로 표시된 후 새 오브젝트를 생성할 수 없으며 모든 사이트가 다시 온라인 상태가 될 때까지 버킷의 기존 오브젝트를 업데이트할 수 없다는 것입니다. **운영 중단 증 읽기 전용 액세스** 옵션은 버킷을 생성할 때만 사용할 수 있으며 이후에는 수정할 수 없습니다.

앞에서 설명한 것처럼 사이트 간에 하트비트가 계속 손실되면 사이트에 장애가 발생한 것으로 표시되며 기본값은 15 분입니다. 따라서 하트비트가 지속적으로 손실되면 다음과 같이 됩니다.

- 2 사이트 구성에서는 각각이 자신을 온라인 상태로 간주하고 다른 하나는 장애가 발생한 것으로 표시합니다.
- 3 사이트 이상의 사이트 구성에서는 다음이 모두 해당할 경우만 사이트를 장애가 발생한 것으로 표시합니다.
  - 대부분의 사이트에서 동일한 ECS 사이트에 대한 하트비트가 지속적으로 손실됩니다.
  - 그리고 나머지 모든 사이트가 현재 온라인 상태로 표시되어 있습니다.

기업의 내부 네트워크에서 한 사이트에 연결할 수 없지만 엑스트라넷 네트워킹은 계속 작동하는 경우와 같이 장애가 발생한 사이트를 클라이언트와 애플리케이션이 여전히 액세스할 수 있습니다. 예를 들어 5 사이트 구성에서 사이트 2~5 가 모두 지속적으로 사이트 1 로의 네트워크 연결이 끊어지면 ECS 는 사이트 1 을 일시적으로 장애가 발생한 것으로 표시합니다. 클라이언트와 애플리케이션에서 사이트 1 에 계속 액세스할 수 있으면 다른 사이트에 대한 조회가 필요하지 않기 때문에 로컬에서 소유한 버킷과 오브젝트에 대한 서비스 요청을 처리할 수 있습니다. 그러나 소유하지 않은 버킷 및 오브젝트에 대한 사이트 1 로의 요청은 실패합니다. 표 16 은 **운영 중단 중 액세스**가 활성화로 설정되었을 때 사이트가 장애가 발생한 것으로 표시된 후 작업이 성공하기 위해 필요한 액세스 권한을 보여 줍니다.

표 16 운영 중단 중 액세스가 활성화된 상태에서 사이트가 장애가 발생한 것으로 표시된 후 성공하는 작업

작업	**장애가 발생한 사이트에 요청을 전송 (3 개 이상의 사이트를 포함하는 페더레이션)**	**다음 중 하나를 포함한 온라인 사이트에 요청을 전송:** • **3 개 이상의 사이트를 포함하는 페더레이션의 온라인 사이트** • **또는 두 개의 사이트만 포함된 페더레이션의 한 사이트**
오브젝트 생성	버킷에 **운영 중단 중 읽기 전용 액세스**를 활성화하지 않았다면 로컬 소유 버킷에서 성공. 원격 소유한 버킷에서는 실패	버킷에 **운영 중단 중 읽기 전용 액세스**를 활성화하지 않았다면 성공.
오브젝트 나열	모든 오브젝트가 로컬 소유이면 로컬 소유 버킷의 오브젝트 나열	성공 복제가 완료되지 않은 장애가 발생한 사이트에서 소유한 오브젝트는 포함되지 않음

작업	장애가 발생한 사이트에 요청을 전송 (3 개 이상의 사이트를 포함하는 페더레이션)	다음 중 하나를 포함한 온라인 사이트에 요청을 전송: • 3 개 이상의 사이트를 포함하는 페더레이션의 온라인 사이트 • 또는 두 개의 사이트만 포함된 페더레이션의 한 사이트
오브젝트 읽기	로컬 소유 버킷의 로컬 소유 오브젝트에서 성공 (최신 버전이 아닐 수 있음) 원격 소유한 오브젝트에서는 실패	성공 오브젝트를 장애가 발생한 사이트에서 소유한 경우 장애가 발생하기 전에 원래 오브젝트가 복제를 완료해야 함
오브젝트 업데이트	버킷에 **운영 중단 중 읽기 전용 액세스**를 활성화하지 않았다면 로컬 소유 버킷의 로컬 소유 오브젝트에서 성공. 원격 소유한 오브젝트에서는 실패	버킷에 **운영 중단 중 읽기 전용 액세스**를 활성화하지 않았다면 성공. 오브젝트 소유권 획득

- 오브젝트 생성

  버킷에 **운영 중단 중 읽기 전용 액세스**를 활성화했다면 사이트에 장애가 발생한 것으로 표시된 후 오브젝트 생성이 성공하지 않습니다. 비활성화된 경우:

  - 세 개 이상의 사이트가 포함된 페더레이션에서 사이트가 장애가 발생한 것으로 표시되고 클라이언트나 애플리케이션이 여기에 액세스할 수 있으면 로컬에서 소유한 버킷에만 오브젝트를 만들 수 있습니다. 이러한 새 오브젝트는 이 사이트에서만 액세스할 수 있으며, 장애가 발생한 사이트가 다시 온라인 상태가 되고 버킷 소유자에 액세스할 때까지 다른 사이트에서는 이러한 오브젝트를 인식하지 못합니다.
  - 온라인 사이트는 장애가 발생한 것으로 표시된 사이트에서 소유한 버킷을 포함하여 모든 버킷에 오브젝트를 만들 수 있습니다. 오브젝트를 생성하려면 버킷 목록을 새 오브젝트 이름으로 업데이트해야 합니다. 버킷 소유자가 다운되면 버킷 소유자의 복구 또는 재가입 작업 중에 버킷 목록 테이블에 오브젝트를 삽입하는 오브젝트 내역이 생성됩니다.

- 오브젝트 나열
  - 세 개 이상의 사이트가 포함된 페더레이션에서 장애가 발생한 것으로 표시된 사이트는 버킷과 버킷 내의 모든 오브젝트에 대한 로컬 소유권이 모두 필요합니다. 장애가 발생한 사이트의 목록에는 일시적으로 실패한 것으로 표시된 동안 원격으로 생성된 오브젝트는 포함되지 않습니다.
  - 온라인 사이트는 장애가 발생한 것으로 표시된 사이트에서 소유한 버킷을 포함하여 모든 버킷의 오브젝트를 나열할 수 있습니다. 가지고 있는 최신 버전의 버킷 목록이 나열되며 약간 오래되었을 수 있습니다.
- 오브젝트 읽기
  - 세 개 이상의 사이트가 포함된 페더레이션에서 클라이언트나 애플리케이션이 장애가 발생한 사이트에 액세스할 수 있으면 로컬 소유 버킷의 로컬 소유 오브젝트만 읽을 수 있습니다.
  - 장애가 발생한 사이트에 대한 읽기 요청을 받으면 현재 오브젝트 소유자를 검증하기 위해 먼저 버킷 소유자에 액세스해야 합니다. 버킷 소유자에 액세스할 수 있고 현재 오브젝트 소유자가 로컬이면 읽기 요청이 성공합니다. 버킷 소유자 또는 현재 오브젝트 소유자에 액세스할 수 없으면 읽기 요청이 실패합니다.
  - 원본 오브젝트가 복제를 완료했다면 온라인 사이트는 장애가 발생한 것으로 표시된 사이트가 소유한 오브젝트를 포함하여 모든 오브젝트를 읽을 수 있습니다. 온라인 사이트는 오브젝트 내역을 확인하고 사용 가능한 최신 버전의 오브젝트로 응답합니다. 나중에 장애가 발생한 것으로 표시된 사이트에서 오브젝트가 업데이트되고 업데이트된 버전의 지리적 복제가 완료되지 않으면 읽기 요청을 처리하는 데 이전 버전이 사용됩니다.

참고: 읽기 요청은 온라인 사이트로 전송되고 장애가 발생한 사이트인 버킷 소유자는 로컬 버킷 목록 정보와 오브젝트 내역을 사용하여 오브젝트 소유자를 확인합니다.

- 오브젝트 업데이트
  - 버킷에 **운영 중단 중 읽기 전용 액세스**를 활성화했다면 사이트에 장애가 발생한 것으로 표시된 후 오브젝트 업데이트가 성공하지 않습니다. 비활성화된 경우, 세 개 이상의 사이트가 포함된 페더레이션에서 클라이언트나 애플리케이션이 장애가 발생한 사이트에 액세스할 수 있으면 로컬 소유 버킷의 로컬 소유 오브젝트만 업데이트할 수 있습니다.
  - 업데이트 요청은 현재 오브젝트 소유권을 확인하기 위해 먼저 버킷 소유자에 액세스해야 합니다. 버킷 소유자에 액세스할 수 있고 현재 오브젝트 소유자가 로컬이면 업데이트 요청이 성공합니다. 버킷 소유자 또는 현재 오브젝트 소유자에 액세스할 수 없으면 업데이트 요청이 실패합니다.

- 재가입 작업이 완료된 후 원격 사이트에서 동일한 TSO 중에 동일한 오브젝트를 업데이트한다면 이 업데이트는 읽기 작업에 포함되지 않습니다.

온라인 사이트는 온라인 사이트와 장애가 발생한 사이트가 소유한 오브젝트를 모두 업데이트할 수 있습니다. 장애가 발생한 것으로 표시된 사이트가 소유한 오브젝트에 대한 오브젝트 업데이트 요청이 온라인 사이트로 전송되면 온라인 상태로 표시된 시스템에서 사용할 수 있는 오브젝트의 최신 버전이 업데이트됩니다.

업데이트를 하는 사이트가 새 오브젝트 소유자가 되고 새 소유자 정보와 시퀀스 번호로 오브젝트 내역을 업데이트합니다. 이 작업은 원래 오브젝트 소유자의 복구 또는 재가입 작업에 사용되어 새 소유자로 오브젝트 내역을 업데이트합니다.

참고: 업데이트 요청은 온라인 사이트로 전송되고 장애가 발생한 사이트인 버킷 소유자는 로컬 버킷 목록 정보와 오브젝트 내역을 사용하여 오브젝트 소유자를 확인합니다.

이 예에서는 그림 12 에 표시된 것처럼 3 사이트 구성에서 네임스페이스 1 의 버킷과 오브젝트 레이아웃에서 발생하는 일을 보여 줍니다.

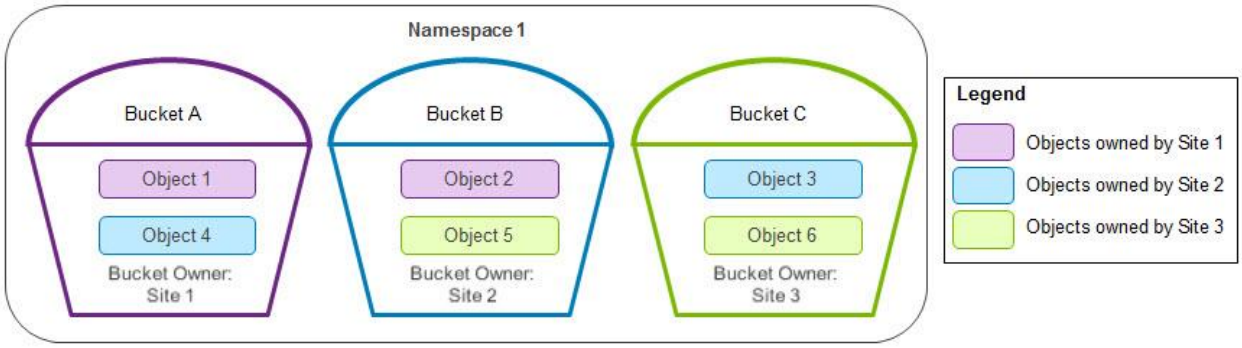

그림 12 네임스페이스 1 에 대한 버킷 및 오브젝트 소유권

표 17 은 다음 세 가지에 모두 해당하는 경우 이 3 사이트 구성의 예입니다.

- **운영 중단 중 액세스**가 활성화됨
- **운영 중단 중 읽기 전용 액세스**가 비활성화됨
- 사이트 1 이 장애가 발생한 것으로 표시되었습니다.

표 17 3 사이트 구성에서 **운영 중단 중 액세스**와 **운영 중단 중 읽기 전용 액세스**가 비활성화되어 있고 사이트 1 이 일시적 장애가 발생한 것으로 표시된 상태에서 성공 또는 실패하는 작업의 예

작업	버킷/오브젝트	요청 전송 대상	
		사이트 1(장애가 발생한 것으로 표시됨)	사이트 2 또는 사이트 3(온라인)
오브젝트 생성 위치	버킷 A	성공	성공
	버킷 B	실패 장애가 발생한 사이트는 로컬 소유 버킷에만 오브젝트를 생성할 수 있음	성공
	버킷 C	실패 장애가 발생한 사이트는 로컬 소유 버킷에만 오브젝트를 생성할 수 있음	성공
오브젝트 나열 위치	버킷 A	실패 버킷이 로컬 소유이지만 원격으로 소유한 오브젝트가 포함되어 있음	성공 복제가 완료된 장애가 발생한 사이트에서 소유한 오브젝트는 포함되지 않음
	버킷 B	실패 장애가 발생한 사이트는 로컬 소유 버킷의 오브젝트만 나열할 수 있음	성공
	버킷 C	실패 장애가 발생한 사이트는 로컬 소유 버킷의 오브젝트만 나열할 수 있음	성공
오브젝트 읽기 또는 업데이트	오브젝트 1	성공, 오브젝트와 버킷을 모두 로컬로 소유	성공 오브젝트가 TSO 이전에 복제를 완료해야 함 업데이트에서 오브젝트 소유권 획득
	오브젝트 2	실패. 버킷을 로컬로 소유하지 않음	

작업	버킷/오브젝트	요청 전송 대상	
		사이트 1(장애가 발생한 것으로 표시됨)	사이트 2 또는 사이트 3(온라인)
	오브젝트 3 오브젝트 4 오브젝트 5 오브젝트 6	실패 장애가 발생한 사이트는 로컬 소유 버킷의 로컬 소유 오브젝트만 읽고 업데이트할 수 있음	성공

사이트 간에 하트비트가 다시 설정되면 시스템이 사이트를 온라인 상태로 표시하고 장애 이전과 마찬가지로 이 데이터에 대한 액세스가 계속됩니다. 재가입 작업이 다음을 처리합니다.

- 버킷 목록 테이블 업데이트
- 필요하면 오브젝트 소유권 업데이트
- 이전에 장애가 발생한 사이트의 복제 대기열 처리 재개

참고: ECS 는 한 사이트의 일시적 장애 중에만 액세스를 지원합니다.

그림 13 은 다른 두 사이트의 예입니다. 두 사이트 모두 자신을 온라인 상태라고 가정하고, TSO 가 발생했을 때 다른 사이트를 장애가 발생한 것으로 표시합니다. 모든 생성, 나열, 읽기, 업데이트 작업이 성공합니다.

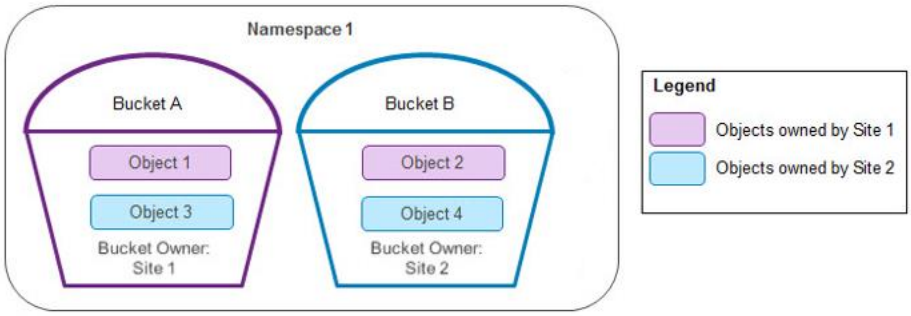

그림 13 두 사이트의 버킷 및 오브젝트 소유권

TSO 동안 각 사이트의 모든 오브젝트가 업데이트됩니다. 표 18 은 사이트 간에 하트비트가 다시 설정되었을 때 사이트의 최종 데이터를 보여 줍니다.

표 18 사이트 다시 설정 후 승리 사이트

오브젝트	버킷 이름	버킷 소유자	오브젝트 소유자	그럼 "승리" 사이트는...
오브젝트 1	버킷 A	사이트 1	사이트 1	사이트 2
오브젝트 2	버킷 B	사이트 2	사이트 1	사이트에 최신 타임스탬프 있음
오브젝트 3	버킷 A	사이트 1	사이트 2	사이트에 최신 타임스탬프 있음
오브젝트 4	버킷 B	사이트 2	사이트 2	사이트 1

참고: 이 예에서 최신 타임스탬프는 오브젝트의 최신 업데이트 시간을 의미합니다.

4.1.2.1 3 개 이상의 사이트를 사용한 XOR 디코딩

XOR 인코딩 섹션에서 살펴본 것처럼 ECS 는 3 개 이상의 사이트가 포함된 복제 그룹으로 구성된 데이터의 스토리지 효율성을 극대화합니다. 청크의 보조 복사본에 있는 데이터는 XOR 작업 후 패리티 청크의 데이터로 대체될 수 있습니다. 인코딩된 청크의 데이터에 대한 요청은 주 복제본을 포함한 사이트에서 처리합니다. 이 사이트에 장애가 발생하면 요청이 오브젝트의 보조 복사본이 있는 사이트로 이동합니다. 그러나 이 복사본이 인코딩되었으므로 보조 사이트는 먼저 온라인 주 사이트에서 인코딩에 사용된 청크의 복사본을 가져와야 합니다. 그런 다음 XOR 작업을 하여 요청된 오브젝트를 재구성하고 요청에 응답합니다. 청크가 재구성된 후 사이트가 후속 요청에 더 신속하게 응답할 수 있도록 청크가 캐싱됩니다.

표 19 는 4 사이트 구성의 사이트 4 에 있는 청크 관리자 테이블 일부를 예로 보여주고 있습니다.

표 19 XOR 인코딩 완료 후 사이트 4 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 4	인코딩됨
C2	사이트 2	사이트 4	인코딩됨
C3	사이트 3	사이트 4	인코딩됨
C4	사이트 4		패리티(C1, C2, C3)

그림 14 - TSO 중에 읽기 요청을 서비스하기 위해 청크를 다시 생성하는 것과 관련된 요청을 보여 줍니다.

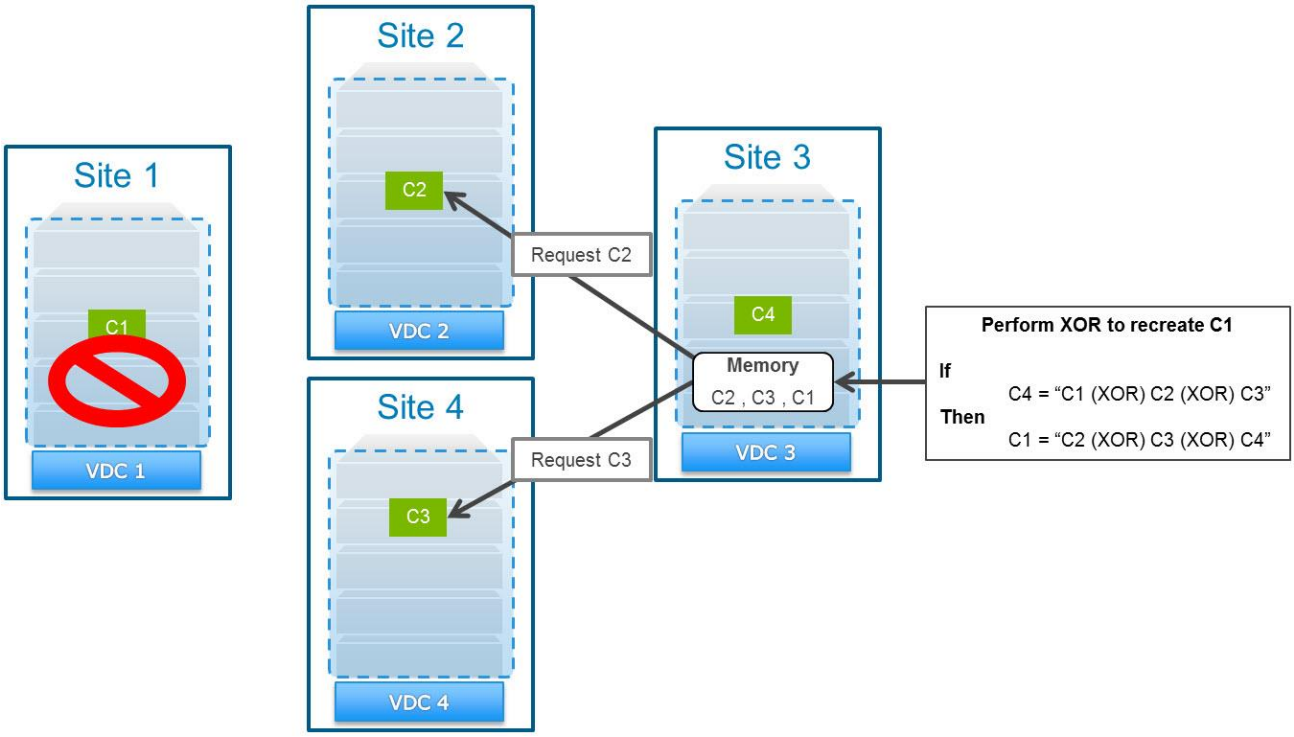

그림 14 XOR 된 청크를 재구성하여 읽기 요청 서비스

이 예에서 **사이트 1** 이 장애가 발생한 것으로 표시될 때 청크 **C1** 의 오브젝트에 대한 읽기가 요청되면 다음과 같이 됩니다.

- 사이트 1 에 장애가 있으므로 요청은 청크 C1 의 보조 사이트인 사이트 4 로 전송됩니다.
- 사이트 4 는 이미 청크 **C1**, **C2**, **C3** 에 XOR 을 했습니다. 이는 사이트 4 가 이러한 청크에서 데이터의 로컬 복제본을 패리티 청크 **C4** 의 데이터로 교체했음을 의미합니다.
- 사이트 4 가 주 사이트(사이트 2)에 청크 **C2** 의 복제본을 요청하고 로컬로 캐시합니다.
- 사이트 4 가 주 사이트(사이트 3)에 청크 **C3** 의 복제본을 요청하고 로컬로 캐시합니다.
- 그런 다음 사이트 4 는 패리티 청크 **C4** 를 사용하여 캐시된 청크 **C2** 와 **C3** 사이의 XOR 작업을 하여 청크 **C1** 을 다시 생성하고 캐시에 로컬로 저장합니다.
- 그런 다음 사이트 4 가 청크 **C1** 의 오브젝트에 대한 읽기 요청에 응답합니다.

참고: 재구성 작업 완료 시간은 복제 그룹의 사이트 수에 비례하여 증가합니다.

4.1.2.2 지리적 패시브 복제 사용

지리적 패시브 복제를 사용하여 구성된 버킷의 모든 데이터는 2~4 개의 소스 사이트와 1 개 또는 2 개의 전용 복제 타겟을 가집니다. 복제 타겟에 쓴 데이터는 XOR 작업 후 패리티 청크의 데이터로 대체될 수 있습니다.

지리적 패시브 복제된 데이터에 대한 요청은 주 복제본을 포함하고 있는 사이트에 의해 처리됩니다. 요청한 사이트가 이 사이트에 액세스할 수 없으면 복제 타겟 사이트 중 하나에서 데이터를 복구해야 합니다.

지리적 패시브 복제에서는 항상 소스 사이트가 오브젝트와 버킷 소유자가 됩니다. 따라서 복제 타겟 사이트가 일시적으로 장애가 발생한 것으로 표시되면 모든 IO 작업이 정상적으로 계속됩니다. 유일한 예외는 복제 타겟 사이트가 페더레이션에 재가입할 때까지 대기열을 계속 유지하는 복제입니다.

소스 사이트 중 하나에 장애가 발생하면 온라인 소스 사이트에 대한 요청이 복제 타겟 사이트 중 하나에서 로컬로 소유하지 않은 데이터를 복구해야 합니다. 사이트 1 과 사이트 2 가 소스 사이트이고 사이트 3 이 복제 타겟 사이트인 예를 살펴보겠습니다. 이 예에서는 오브젝트의 주 복제본이 사이트 1 에서 소유한 청크 C1 에 있으며 청크가 타겟 사이트 3 으로 복제되었습니다. 사이트 1 에 장애가 발생하고 그 오브젝트를 읽어달라는 요청이 사이트 2 에 들어오면 사이트 2 는 사이트 3 에서 복제본을 가져와야 합니다. 복제본이 인코딩되면 보조 사이트는 먼저 온라인 주 사이트에서 인코딩에 사용된 다른 청크의 복사본을 가져와야 합니다. 그런 다음 XOR 작업을 하여 요청된 오브젝트를 재구성하고 요청에 응답합니다. 청크가 재구성된 후 사이트가 후속 요청에 더 신속하게 응답할 수 있도록 청크가 캐싱됩니다.

표 20 은 지리적 패시브 타겟의 청크 관리자 테이블의 일부를 예로 보여주고 있습니다.

표 20 XOR 인코딩 완료 후 지리적 패시브 복제 타겟 청크 관리자

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 3	인코딩됨
C2	사이트 2	사이트 3	인코딩됨
C3	사이트 3		패리티(C1 및 C2)

그림 15 - TSO 중에 읽기 요청을 서비스하기 위해 청크를 다시 생성하는 것과 관련된 요청을 보여 줍니다.

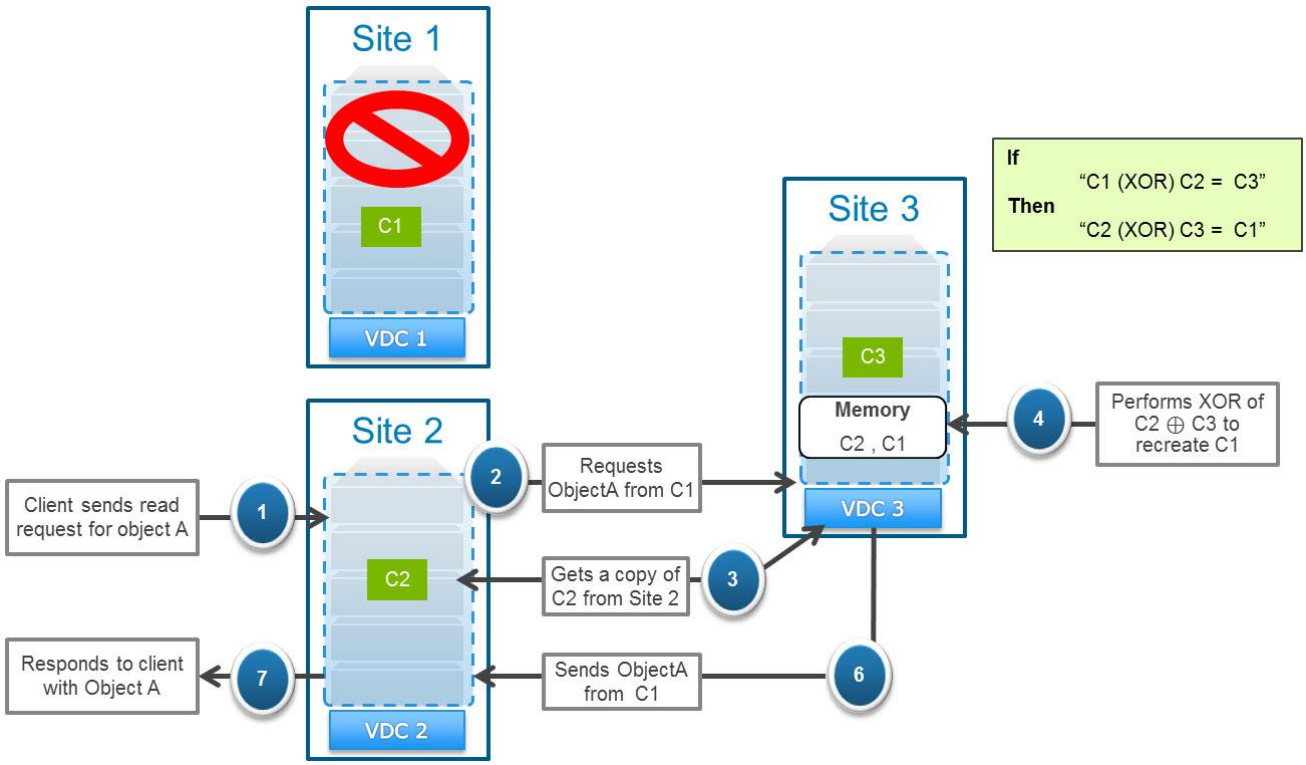

그림 15 XOR 된 청크를 재구성하여 읽기 요청 서비스

이 예에서 **사이트 1** 이 일시적으로 장애가 발생한 것으로 표시될 때 청크 **C1** 의 오브젝트에 대한 읽기가 요청되면 다음과 같이 됩니다.

- 사이트 1 에 장애가 있으므로 요청은 청크 C1 의 보조 사이트인 사이트 3 으로 전송됩니다.
- 사이트 3 은 이미 청크 **C1** 과 **C2** 에 XOR 을 했습니다. 이는 사이트 3 이 이러한 청크에서 데이터의 로컬 복제본을 패리티 청크 **C3** 의 데이터로 교체했음을 의미합니다.
- 사이트 3 이 주 사이트(사이트 2)에 청크 **C2** 의 복제본을 요청하고 로컬로 캐시합니다.
- 그런 다음 사이트 3 은 패리티 청크 **C3** 을 사용하여 캐시된 청크 **C2** 사이의 XOR 작업을 하여 청크 **C1** 을 다시 생성하고 캐시에 로컬로 저장합니다.
- 그런 다음 사이트 3 이 청크 **C1** 의 오브젝트에 대한 읽기 요청에 응답합니다.

4.1.2.3 모든 사이트에 복제 사용

모든 사이트에 복제와 **운영 중단 중 액세스** 옵션으로 구성된 버킷을 사용하면 읽기 성능을 높일 수 있습니다. XOR 디코딩 작업이 필요하지 않고 데이터를 로컬에서 읽을 가능성이 더 높기 때문에 모든 사이트가 온라인 상태일 때뿐만 아니라 일시적인 사이트 운영 중단 중에도 읽기 성능이 빠릅니다.

모든 사이트에 복제를 활성화한 버킷의 데이터는 각 사이트에 복제됩니다. 오브젝트 생성과 업데이트는 **모든 사이트에 복제**를 비활성화한 것과 동일하게 처리됩니다. 그러나 오브젝트 읽기와 나열은 주 사이트에 장애가 발생하기 전에 일부 사이트에 대한 복제만 완료되었을 수 있기 때문에 약간 다르게 처리됩니다.

읽기 작업 중에 요청을 서비스하는 노드는 먼저 오브젝트 소유자의 메타데이터 최신 버전을 확인합니다. 요청 노드에 따라 다음과 같이 달라집니다.

- **오브젝트 소유자인 경우**:
  - 요청 중인 데이터의 로컬 복제본이 있으면 요청을 처리하는 데 사용됩니다.
  - 데이터를 복제하기 전에 장애가 발생한 다른 사이트에 의해 오브젝트가 업데이트되면 로컬로 있는 버전을 반환합니다.
- **오브젝트 소유자가 아닌 경우**
  - 오브젝트 소유자 사이트가 온라인 상태이고 오브젝트 데이터를 복제하는 경우:
    - 이 사이트에 완료한 후 로컬 데이터 복제본으로 요청을 처리합니다.
    - 이 사이트에 완료하지 않고 오브젝트 소유자로부터 복제본을 요청하고 요청을 처리하는 데 사용합니다.

- 오브젝트 소유자가 다운된 경우
  - 이 사이트로의 오브젝트 복제가 완료되면 로컬 데이터 복제본을 사용하게 되며, 최신 버전이 아닐 수 있습니다.
  - 요청 사이트에 대한 오브젝트 데이터 복제가 완료되지 않은 경우 요청 사이트가 청크 관리자 테이블에 먼저 나열된 보조 사이트에서 복제본을 요청합니다. 사이트가 그 자체이면 읽기 작업이 실패합니다.

버킷의 오브젝트 나열 작업 중에 노드에 버킷 소유자와 버킷의 각 오브젝트에 대한 헤드 정보 모두가 필요합니다. 오브젝트 소유자 또는 버킷 소유자인 사이트가 다운되고 **운영 중단 중 액세스**가 활성화되면 복제 그룹의 나머지 사이트가 모두 온라인 상태일 때 요청을 계속 처리할 수 있습니다. 최신 버전의 버킷 목록을 나열합니다. 약간 오래되었을 수 있으며 사이트마다 다를 수 있습니다.

4.1.3 여러 사이트 장애

ECS 는 복제 그룹 내의 하나의 사이트에 일시적인 장애가 발생한 경우에만 액세스를 지원합니다. 또한 한 사이트만 장애로 표시할 수 있습니다. 즉 복제 그룹 내에서 둘 이상의 사이트에 동시에 장애가 발생할 일부 작업이 실패합니다. (지속적인 하트비트 손실로 인해) 장애가 확인된 첫 번째 사이트는 장애가 발생한 것으로 표시됩니다. 하트비트가 지속적으로 손실된 나머지 사이트는 장애가 발생한 것으로 표시되지 않으며 온라인 상태로 간주됩니다.

예를 들어 복제 그룹에 사이트가 5 개 있는데 사이트 1 이 지속적인 하트비트 손실로 파악되면 장애가 발생한 것으로 표시됩니다. 사이트 2 도 지속적인 하트비트 손실로 파악되었지만 온라인 상태로 유지됩니다. 다음과 같은 상황이 발생합니다.

- **운영 중단 중 액세스**가 활성화되어 있고 버킷 소유자가 사이트 2 이면 다른 사이트로 전송된 읽기/생성/업데이트는 오브젝트 소유자와 관계없이 실패합니다. 먼저 버킷 소유자와 확인하여 오브젝트 소유자를 결정하기 때문입니다. 버킷 소유자가 장애가 발생한 것으로 표시되지 않아 요청자가 사이트 2 로 요청을 보내지만 이 요청은 실패합니다.
- 사이트 2 로 전송된 읽기 및 업데이트 요청은 사이트 2 가 오브젝트 소유자(그리고 **운영 중단 중 액세스**가 활성화된 경우 버킷 소유자)일 경우에만 성공합니다.
- 사이트 2 가 아닌 다른 사이트로 전송된 읽기 및 업데이트 요청은 오브젝트 소유자(그리고 **운영 중단 중 액세스**가 활성화된 경우 버킷 소유자)가 사이트 2 가 아닌 경우에만 성공합니다.
- 버킷 소유자가 사이트 1 또는 사이트 2 이면 오브젝트 생성이 실패합니다. 오브젝트를 생성하려면 버킷 목록을 새 오브젝트 이름으로 업데이트해야 하기 때문입니다. 온라인으로 표시된 모든 사이트에서 이 작업을 할 수 없으므로 생성 작업이 실패합니다.

- 버킷의 오브젝트를 나열하는 요청은 요청하는 사이트가 버킷 소유자와 모든 오브젝트에 액세스할 수 있을 경우에만 성공합니다.
  - 사이트 2 로 요청이 전송되면 버킷과 버킷의 모든 오브젝트를 소유한 경우에만 성공합니다.
  - 요청을 다른 사이트로 보내는 경우 버킷이나 버킷 내의 오브젝트를 사이트 2 에서 소유하지 않은 경우에만 성공합니다.

4.2 PSO(Permanent Site Outage)

사이트에서 재해가 발생하고 관리자가 그 사이트를 복구할 수 없는 것으로 판단하면 영구 사이트 페일오버(페더레이션에서 VDC 제거)를 시작할 수 있습니다. 영구 사이트 페일오버가 시작되면 장애가 발생한 사이트의 모든 청크가 나머지 사이트에서 복구되어 데이터 내구성을 재수립합니다.

복구 프로세스에는 나머지 사이트가 로컬 청크 관리자 테이블을 스캔하여 장애가 발생한 사이트를 포함한 사이트로의 참조를 찾는 과정이 포함됩니다. 찾는 청크 유형은 다음과 같습니다.

- **인코딩됨**

  a. 유형이 인코딩되고 주 사이트가 온라인인 청크의 경우 주 사이트의 데이터를 사용하여 로컬로 데이터를 다시 생성합니다. 완료되면 이 청크를 복사 유형으로 표시합니다.
  b. 그런 다음 패리티 청크로 이전에 재생성한 복사 유형의 XOR 작업을 하여 주 사이트가 장애 사이트인 인코딩된 청크를 재생성합니다. 이 사이트는 이제 청크의 주 사이트가 되며 로컬 유형을 가집니다.
  c. 그런 다음 이러한 청크는 복제 대기열에 추가되어 복제 그룹 내에 나열된 다른 사이트에 복제됩니다.

- 장애가 발생한 사이트로 나열된 **복사** 및 주 사이트가 새 주 사이트가 됩니다. 그런 다음 새 보조 사이트에 복제할 복제 대기열에 청크를 추가합니다.
- **로컬** 및 그 보조 사이트가 장애가 발생한 사이트이며, 새 보조 사이트로 청크를 복제하는 작업이 추가됩니다.

영구 사이트 페일오버가 시작되면 영구 사이트 페일오버 프로세스가 완료될 때까지 장애가 발생한 사이트가 소유한 데이터에 액세스할 수 없습니다. 데이터 복제는 페일오버 작업과는 별개이므로 장애가 발생한 사이트가 소유한 데이터에 액세스하기 위해 완료할 필요가 없습니다.

3 사이트 예를 보겠습니다. 사이트 1 에 장애가 발생하고 표 21 및 표 22 가 나머지 두 사이트의 청크 관리자 테이블입니다.

표 21 사이트 2 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 2	복사
C2	사이트 2	사이트 3	로컬
C3	사이트 1	사이트 3	원격
C4	사이트 2	사이트 1	로컬

사이트 2 는 다음을 수행합니다.

- 복제할 복제 대기열에 청크 **C1** 을 추가합니다. 사이트 2 가 새 주 사이트가 되고 새 청크가 있는 사이트가 새 보조 사이트가 됩니다.
- 복제할 복제 대기열에 청크 **C4** 를 추가하고 테이블의 보조 사이트를 업데이트합니다.

표 22 사이트 3 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 2	원격
C2	사이트 2	사이트 3	인코딩됨
C3	사이트 1	사이트 3	인코딩됨
C4	사이트 2	사이트 1	원격
C5	사이트 3		패리티(C2 및 C3)

사이트 3 은 다음을 수행합니다.

1. 주 사이트(**사이트 2**)의 데이터를 사용하여 청크 **C2** 데이터를 로컬로 다시 생성합니다. 청크 유형을 복사로 변경합니다.
2. **C2** 데이터를 사용하여 청크 **C3** 을 재구성하고 XOR 작업 C2 ⊕ C5 를 사용하여 **C5** 패리티 데이터를 재구성합니다. 사이트 3 이 새로운 주 사이트가 됩니다.
3. 청크 **C5** 를 삭제합니다.
4. 복제할 복제 대기열에 청크 **C3** 을 추가하고 새 청크가 있는 사이트가 새 보조 사이트가 됩니다.

영구 사이트 페일오버가 청크 관리자를 완료한 후 나머지 두 사이트의 테이블은 표 23 및 표 24 와 같습니다.

표 23 PSO 가 완료된 후 사이트 2 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 2	사이트 3	로컬
C2	사이트 2	사이트 3	로컬
C3	사이트 3	사이트 2	복사
C4	사이트 2	사이트 3	로컬

표 24 PSO 가 완료된 후 사이트 3 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 2	사이트 3	복사
C2	사이트 2	사이트 3	복사
C3	사이트 3	사이트 2	로컬
C4	사이트 2	사이트 3	복사

4.2.1 지리적 패시브 복제를 사용한 PSO

영구 사이트 운영 중단은 지리적 패시브 복제를 사용하여 복제된 데이터에서 약간 다르게 처리됩니다. PSO 동안 지리적 패시브 복제된 데이터는 데이터 내구성을 재수립하지 않으며, 대신 복제 그룹에 새 세 번째 사이트를 추가한 후 재수립됩니다. PSO 작업은 영구적으로 장애가 발생한 사이트가 소스 사이트인지 복제 타겟 사이트인지에 따라 달라집니다.

복구 프로세스에는 나머지 사이트가 로컬 청크 관리자 테이블을 스캔하여 장애가 발생한 사이트를 포함한 사이트로의 참조를 찾는 과정이 포함됩니다. 찾는 청크 유형은 다음과 같습니다.

- **인코딩됨**(복제 타겟 사이트에 있음)
  - 유형이 인코딩되고 주 사이트가 온라인인 청크의 경우 주 사이트의 데이터를 사용하여 로컬로 데이터를 다시 생성합니다. 완료되면 이 청크를 복사 유형으로 표시합니다.
  - 그런 다음 패리티 청크로 이전에 재생성한 복사 유형의 XOR 작업을 하여 주 사이트가 장애가 발생한 소스 사이트인 인코딩된 청크를 재생성합니다. 이 사이트는 이제 청크의 주 사이트가 되며

로컬 유형을 가집니다. 복제 그룹에 세 번째 사이트를 추가할 때까지 보조 사이트가 생성되지 않습니다.

- 장애가 발생한 사이트로 나열된 **복사** 및 주 사이트가 새 주 사이트가 되고 유형이 로컬로 변경됩니다. 복제 그룹에 세 번째 사이트를 추가할 때까지 보조 사이트가 생성되지 않습니다.
- **로컬** 및 그 보조 사이트(복제 타겟)가 장애가 발생한 사이트입니다. 복제 그룹에 세 번째 사이트를 추가할 때까지 새 보조 사이트가 생성되지 않습니다.

PSO 이후 세 번째 사이트를 추가하여 데이터 내구성을 재수립하고 사이트 전체의 장애로부터 보호할 수 있습니다. 세 번째 사이트가 지리적 패시브 복제 그룹에 추가된 후 이전 두 사이트는 로컬 청크 관리자 테이블을 검색하여 보조 청크가 나열되지 않은 청크를 찾습니다. 그런 다음 다음과 같이 됩니다.

- 보조 청크가 나열되지 않은 소스 사이트의 로컬 청크가 새 복제 타겟 사이트로 보조 청크의 복제를 시작합니다. 새 보조 청크 위치를 포함하도록 청크 관리자 테이블이 업데이트됩니다.
- 복제 타겟 사이트의 로컬 청크가 새 소스 사이트로 청크의 복제를 시작합니다. 복제가 완료되면 복제 타겟 사이트 유형이 로컬에서 복사로 변경되고 소스 사이트 유형이 복사에서 로컬로 변경됩니다. XOR 작업은 대상에서 정상적으로 계속됩니다.

PSO 가 소스 사이트에서 시작되면 영구 사이트 페일오버 프로세스가 완료될 때까지 장애가 발생한 사이트에서 소유한 데이터에 액세스할 수 없습니다. PSO 가 완료되면 데이터에 대한 액세스가 복원됩니다. 복제 그룹에 세 번째 사이트를 추가할 때까지 온라인 소스 사이트에 대한 모든 새 쓰기는 복제 타겟에 복제되지만 XOR 작업은 수행되지 않습니다. 이는 XOR 이 오직 두 다른 소스 사이트의 청크에서만 실행되기 때문이며 모든 새로운 소스 사이트가 동일해서 XOR 을 실행할 수 없기 때문입니다.

PSO 가 완료되면 세 번째 사이트를 복제 그룹에 추가하여 데이터 내구성을 복원하고 사이트 규모의 장애로부터 보호할 수 있습니다. 또한 복제 타겟은 복사 유형으로 표시된 서로 다른 소스 사이트의 두 청크에 대해 XOR 작업 실행을 재개할 수 있습니다.

사이트 1 과 사이트 2 가 소스 사이트이고 사이트 3 이 타겟 사이트인 몇 가지 예를 살펴보겠습니다. 표 25 및 표 26 은 두 소스 사이트의 청크 관리자 테이블이며 표 27 은 복제 타겟 사이트의 청크 관리자 테이블입니다.

표 25 사이트 1, 소스 사이트 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 3	로컬
C2	사이트 2	사이트 3	원격
C3	사이트 1	사이트 3	로컬

표 26 사이트 2, 소스 사이트 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 3	원격
C2	사이트 2	사이트 3	로컬
C3	사이트 1	사이트 3	원격

표 27 사이트 3, 복제 타겟 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 3	인코딩됨
C2	사이트 2	사이트 3	인코딩됨
C3	사이트 1	사이트 3	복사
C4	사이트 3		패리티(C1 및 C2)

예 1: PSO 때문에 사이트 3 이 제거되면 보조 사이트는 모두 비어 있지만 주 사이트와 유형은 남아 있습니다. 새 복제 타겟이 추가될 때까지 새 쓰기에 주 사이트가 나열되지만 보조 사이트는 나열되지 않습니다.

예 2: PSO 로 인해 사이트 1 이 제거되면 다음과 같은 상황이 발생합니다.

- 사이트 3 은 청크 "**C3**"에 대한 로컬 유형으로 새로운 주 사이트가 되며 보조 사이트로 나열되지 않습니다.
- 사이트 3 은 주 사이트(사이트 2)의 데이터를 사용하여 청크 **C2** 데이터를 다시 생성하고 그 청크 유형을 복사로 변경합니다.

- 사이트 3 은 **C2** 데이터를 사용하여 청크 **C1** 을 재구성하고 XOR 연산 C2 ⊕ C4 를 사용하여 **C4** 패리티 데이터를 재구성합니다. 사이트 3 이 새로운 주 사이트가 되며, 보조 사이트가 나열되지 않습니다.
- 사이트 3 에서 청크 **C4** 를 삭제합니다.

사이트 1 의 영구 사이트 페일오버가 완료된 후 나머지 두 사이트의 청크 관리자 테이블은 표 28 및 표 29 와 같습니다.

표 28 사이트 2, PSO 완료 후 소스 사이트 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 3		원격
C2	사이트 2	사이트 3	로컬
C3	사이트 3		원격

표 29 사이트 3, PSO 가 완료된 후 복제 타겟 사이트 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 3		로컬
C2	사이트 2	사이트 3	복사
C3	사이트 3		로컬

새 소스 사이트가 추가될 때까지 새 쓰기는 로컬 유형의 주 사이트 사이트 2 와 복사 유형의 보조 사이트 사이트 3 을 가집니다.

새 소스 사이트를 복제 그룹에 추가한 후에는 보조 사이트를 추가하고 그 사이트에 데이터를 복제하여 사이트 규모의 장애로부터 보호하기 위한 데이터 내구성이 재수립됩니다. XOR 작업도 복제 타겟에서 재개됩니다. 새 청크 관리자 테이블은 표 30~표 32 과 같습니다.

- 청크 C1 및 C3 이 새 소스 사이트인 사이트 1 에 복제됩니다. 복제가 완료되면 주 사이트가 사이트 1 로 나열되고 보조 사이트가 사이트 3 으로 나열됩니다.
- 사이트 3 은 청크 C1 과 C2 에 대해 XOR 인코딩을 하여 패리티 유형을 가진 새로운 C4 청크를 생성하고 청크 C1 과 C2 의 유형이 인코딩됨으로 변경됩니다.

표 30 데이터 내구성이 재수립된 후 새 사이트 1 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 3	로컬
C2	사이트 2	사이트 3	원격
C3	사이트 1	사이트 3	로컬

표 31 새 사이트 1 이 추가되고 데이터 내구성이 재수립된 후 사이트 2 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 3	원격
C2	사이트 2	사이트 3	로컬
C3	사이트 1	사이트 3	원격

표 32 사이트 3, 새 사이트 1 이 추가되고 데이터 내구성이 재수립된 후 복제 타겟 사이트 청크 관리자 테이블

청크 ID	주 사이트	보조 사이트	유형
C1	사이트 1	사이트 3	인코딩됨
C2	사이트 2	사이트 3	인코딩됨
C3	사이트 1	사이트 3	복사
C4	사이트 3		패리티(C1 및 C2)

4.2.2 여러 사이트 장애로부터 복구 가능성

ECS 는 한 번에 한 사이트 장애로부터의 복구만 지원합니다. 사이트 운영 중단 사이에 PSO 와 데이터 복구 작업이 모두 완료되면 ECS 는 여러 사이트 장애로부터 복구할 수 있습니다. 복구를 완료하기 전에 두 번째 사이트에 장애가 발생한 경우:

- 영구 사이트 운영 중단 복구 작업을 실행하려면 시스템의 다른 모든 사이트가 온라인 상태여야 합니다. 동시에 사이트 장애가 여러 번 발생하는 경우 사이트에서 PSO 를 실행하려면 하나를 제외하고 모두 TSO 에서 복구해야 합니다.

- PSO 가 완료된 후 데이터 복구가 완료되기 전에 두 번째 사이트에 장애가 발생하면 일부 데이터가 손실될 수 있습니다.

4 사이트 시나리오를 살펴보면, 한 사이트를 제외한 모든 사이트를 손실로부터 성공적으로 복구합니다(나머지 사이트에 모든 데이터를 저장할 공간이 충분하다고 가정).

- 사이트 4 장애
  - 관리자가 사이트 4 를 제거하는 PSO 작업을 시작합니다.
  - 나머지 사이트에서 데이터를 복구하여 데이터 내구성을 재수립합니다.

  이제 사이트 1, 사이트 2, 사이트 3 을 포함하는 3 사이트 페더레이션이 남습니다.

- 두 번째 사이트, 사이트 2 장애:

  PSO 와 데이터 복구가 완료된 후 사이트 2 와 같은 다른 사이트에 장애가 발생할 수 있습니다.

  - 관리자가 사이트 2 를 제거하는 PSO 작업을 시작합니다.
  - 나머지 사이트에서 데이터를 복구하여 데이터 내구성을 재수립합니다.

  이제 사이트 1 과 사이트 3 을 포함하는 2 사이트 페더레이션이 남습니다.

- 세 번째 사이트, 사이트 1 장애:

  PSO 와 데이터 복구가 완료된 후 사이트 1 과 같은 다른 사이트에 장애가 발생할 수 있습니다.

  - 관리자가 사이트 1 을 제거하는 PSO 작업을 시작합니다.

  우리는 이제 사이트 3 이 포함된 1 사이트 페더레이션이 남습니다.

이 예에서는 여러 사이트 장애 상황을 살펴보았습니다. 이것은 일반적인 상황은 아닙니다. 영구 사이트 장애는 보통 지진이나 화재와 같은 재해에 의해 발생하며, 여러 장소에서 단기간에 발생하는 것은 흔하지 않습니다. 일반적으로 하나의 사이트에 영구적 장애가 발생한 후 후속 사이트 장애가 발생하기 전에 새 사이트가 추가됩니다.

5 결론

ECS 아키텍처는 처음부터 시스템 가용성과 데이터 내구성을 모두 제공하도록 설계되었습니다. 관리자는 ECS 를 사용하여 가용성 요구 사항과 TCO 간의 균형을 조정할 수 있습니다. 자동 장애 탐지, 자가 복구와 같은 기능은 사이트 운영 중단과 같은 계획되지 않은 이벤트가 발생하는 가장 중요한 시기에 IT 관리 워크로드를 최소화합니다.

ECS 는 3 중 미러링과 삭제 코딩의 조합을 사용하여 디스크 장애로부터 사이트/VDC 내의 데이터를 보호합니다. ECS 는 일반적인 사용 사례에 대한 기본값과 자주 액세스하지 않는 오브젝트를 위한 더 효율적인 콜드 스토리지라는 두 가지 수준의 삭제 코딩 보호가 가능합니다. 또한 데이터를 장애 도메인에 분산하여 대부분의 장애 시나리오에서 보호합니다.

ECS 는 쓰기 작업의 일부로 체크섬을 계산하여 쓰고 읽기 작업 중에 체크섬을 검증함으로써 데이터 무결성을 보장합니다. 체크섬 검증은 백그라운드 작업에서도 능동적으로 수행됩니다.

ECS 는 시스템 가용성을 계속 제공하도록 설계되었습니다. 이는 사이트/VDC 의 모든 노드로 클라이언트 요청을 서비스하는 분산 아키텍처 설계를 사용하기에 가능합니다.

ECS 설계는 전체 사이트 장애에 대비한 선택적 보호 기능을 추가하여 시스템 가용성과 데이터 내구성 보호 기능을 확장합니다. 이를 위해 사이트를 페더레이션하고 관리자가 다양한 복제 그룹 정책 옵션을 구성할 수 있습니다. 이러한 옵션은 버킷 수준에서 설정할 수 있으며, 운영 중단 중 액세스뿐만 아니라 원격 사이트의 데이터 복제 위치와 저장 방법도 결정할 수 있습니다.

또한 ECS 는 버킷 및/또는 오브젝트가 장애가 발생한 것으로 표시될 때 온라인 사이트에 읽기, 나열, 업데이트 작업을 선택적으로 할 수 있는 "운영 중단 중 액세스" 옵션을 고객에게 제공합니다.

관리자가 복구할 수 없는 사이트라고 판단하면 영구 사이트 운영 중단을 시작할 수 있습니다. 이렇게 하면 복제 그룹에서 VDC/사이트가 제거되고 필요에 따라 데이터를 다시 생성해 데이터 내구성을 재수립합니다.

결론적으로 ECS 는 신뢰할 수 있는 회복탄력성을 갖춘 엔터프라이즈급 클라우드 스토리지 솔루션을 제공합니다.

A 기술 지원 및 리소스

Dell.com/support 는 검증된 서비스와 지원으로 고객의 요구 사항에 부응하기 위해 최선을 다하고 있습니다.

스토리지 기술 문서 및 비디오에서는 고객이 Dell EMC 스토리지 플랫폼을 성공적으로 활용하는 데 필요한 전문 지식을 제공합니다.

A.1 관련 리소스

- ECS 개요 및 아키텍처 백서
- ECS 커뮤니티
- ECS 테스트 드라이브
- 지원 사이트 또는 커뮤니티 사이트의 ECS 제품 설명서
- SolVe Desktop(프로시저 생성기)