Documentación técnica

Dell EMC ECS: diseño de alta disponibilidad

Resumen

En este informe, se describen los detalles arquitectónicos sobre cómo la plataforma Dell EMC™ ECS proporciona disponibilidad empresarial.

Junio de 2021

H16344.6

Revisiones

Fecha	Descripción
Julio de 2017	Versión inicial
Agosto de 2017	Se actualizó para incluir contenido de la versión 3.1 de ECS
Marzo de 2019	Se actualizó para incluir contenido de la versión 3.3 de ECS
Abril de 2020	Se actualizó “Comportamiento de TSO con acceso durante una interrupción habilitada”
Diciembre del 2020	Se actualizó el método de protección de metadatos
Junio de 2021	Se actualizó para incluir contenido de la versión 3.6.1 de ECS

Agradecimientos

La producción de esta documentación estuvo a cargo de las siguientes personas:

Autor: Zhu, Jarvis

Tabla de contenido

Resumen ejecutivo

Las empresas almacenan cantidades cada vez mayores de datos que son fundamentales para mantenerse disponibles. Los costos y las complejidades de tener que restaurar grandes cantidades de datos en caso de fallas del sistema o del sitio pueden ser abrumadores para una organización de TI.

La plataforma Dell EMC™ ECS™ se diseñó para satisfacer las necesidades de capacidad y disponibilidad de las empresas actuales. ECS proporciona escalabilidad de exabytes con soporte para una infraestructura de objetos distribuida globalmente. Está diseñado para proporcionar disponibilidad empresarial con opciones automáticas de detección de fallas y recuperación automática incorporadas.

En este informe, se describen los detalles arquitectónicos sobre cómo ECS proporciona disponibilidad empresarial. Incluye detalles como los siguientes:

- Cómo la infraestructura distribuida proporciona una mayor disponibilidad del sistema
- Métodos avanzados de protección de datos que proporcionan durabilidad de datos
- Cómo se distribuyen los datos para obtener una disponibilidad óptima
- Detección automática de fallas
- Métodos de autorreparación incorporados
- Detalles de la resolución de fallas de disco, nodo y red
- Recuperación ante desastres:
  - Cómo ECS brinda protección contra fallas en todo el sitio
  - Cómo se mantiene la coherencia en una configuración activa-activa de múltiples sitios
  - Cómo se detectan las fallas en todo el sitio
  - Opciones de acceso durante una interrupción del sitio
  - Cómo se restablece la durabilidad de los datos después de una falla permanente en todo el sitio

Terminología

Centro de datos virtual (VDC): en este informe, el término centro de datos virtual (VDC) se utiliza sinónimo de sitio o zona. Los recursos de ECS en un solo VDC deben formar parte de la misma red de administración interna.

Geofederación: puede implementar software de ECS en múltiples centros de datos para crear una geofederación. En una geofederación, ECS se comporta como una federación ligeramente acoplada de VDC autónomos. La federación de sitios implica proporcionar terminales de replicación y administración para la comunicación entre sitios. Una vez federados los sitios, se pueden administrar como una sola infraestructura desde cualquier nodo de la federación.

Grupos de replicación: los grupos de replicación definen dónde están protegidos los datos. Un grupo de replicación local contiene un solo VDC y protege los datos dentro del mismo VDC contra fallas de discos o nodos. Los grupos de replicación global contienen más de un VDC y protegen los datos contra fallas de disco, nodo y sitio. Los grupos de replicación se asignan en el nivel del depósito.

1 Descripción general del diseño de alta disponibilidad

La alta disponibilidad se puede describir en dos áreas principales: la disponibilidad del sistema y la durabilidad de los datos. Un sistema está disponible cuando puede responder a una solicitud de cliente. La durabilidad de los datos se proporciona de forma independiente de la disponibilidad de un sistema de almacenamiento y proporciona garantías para que los datos se almacenen en el sistema sin daños ni pérdidas. Esto significa que, incluso si el sistema ECS está inactivo (por ejemplo, una interrupción de la red), los datos aún están protegidos.

La naturaleza distribuida de la arquitectura de ECS proporciona disponibilidad del sistema, ya que permite que cualquier nodo de un centro de datos virtual (VDC)/sitio responda a las solicitudes del cliente. Si un nodo deja de funcionar, el cliente se puede redirigir de forma manual o automática (por ejemplo, mediante DNS o un balanceador de carga) a otro nodo que pueda atender la solicitud.

ECS utiliza una combinación de espejeado triple y codificación de eliminación para escribir datos de manera distribuida a fin de ser resistentes contra fallas de discos y nodos. ECS es compatible con la replicación entre sitios para aumentar la disponibilidad y la resiliencia mediante la protección contra fallas en todo el sitio. ECS también incluye comprobaciones sistemáticas regulares de integridad de datos con funcionalidad de autorreparación.

Cuando se observa la alta disponibilidad, primero es importante comprender la arquitectura y cómo se distribuyen los datos para obtener disponibilidad y rendimiento óptimos dentro de ECS.

1.1 Fragmentos

Un fragmento es un contenedor lógico que ECS utiliza para almacenar todos los tipos de datos, incluidos los datos de objetos, los metadatos personalizados proporcionados por el cliente y los metadatos del sistema de ECS. Los fragmentos contienen 128 MB de datos que constan de uno o más objetos de un solo depósito, como se muestra en Figura 1.

Figura 1 Fragmento lógico

ECS utiliza la indexación para realizar un seguimiento de todos los datos dentro de un fragmento; consulte 1.2 para obtener más detalles.

1.2 Metadatos de ECS

ECS mantiene sus propios metadatos que rastrean dónde existen los datos, así como el historial de transacciones. Estos metadatos se mantienen en tablas lógicas y registros.

Las tablas contienen pares de valores clave para almacenar información relacionada con los objetos. Una función hash se utiliza para realizar búsquedas rápidas de valores asociados con una clave. Los pares de clave-valor se almacenan en un árbol B+ para la indexación rápida de ubicaciones de datos. Mediante el almacenamiento de clave-valor en un formato de árbol buscado y balanceado, como un árbol B+, se puede acceder rápidamente a la ubicación de los datos y los metadatos. Además, para mejorar aún más el rendimiento de las consultas de estas tablas lógicas, ECS implementa un árbol de combinación estructurada de registros (LSM) de dos niveles. Por lo tanto, hay dos estructuras similares a un árbol, donde un árbol más pequeño está en la memoria (tabla de memoria) y el árbol B+ principal reside en el disco. La búsqueda de pares de clave-valor primero consultará la tabla de memoria y, si el valor no es memoria, se buscará en el árbol B+ principal en el disco.

El historial de transacciones se registra en los registros de registro y estos registros se escriben en los discos. Los registros se rastrean las transacciones de índice que aún no se confirman en el árbol B+. Después de que la transacción se ingresa en un registro, se actualiza la tabla en la memoria. Una vez que la tabla en la memoria se llena o después de un período establecido, la tabla se combina ordenada o volcada al árbol B+ en el disco y se registra un punto de control en el registro. Este proceso se ilustra en Figura 2.

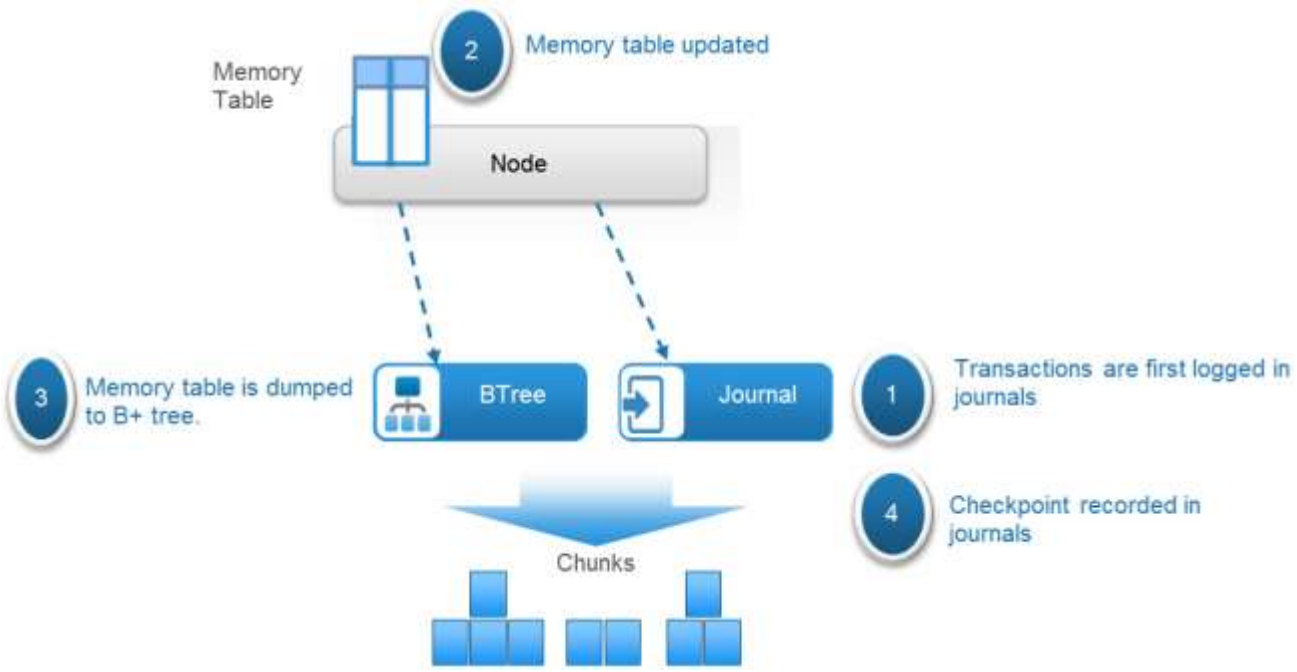

Figura 2 Flujo de trabajo de actualizaciones de transacciones en las tablas de ECS

Los registros y los árboles B+ se escriben en fragmentos.

Hay varias tablas diferentes que utiliza ECS, cada una de las cuales puede ser muy grande. A fin de optimizar el rendimiento de las búsquedas de tabla, cada tabla se divide en particiones que se distribuyen entre los nodos en un VDC/sitio. El nodo donde se escribe la partición se convierte en el propietario/autoridad de esa partición o sección de la tabla.

Una de estas tablas es una tabla de fragmentos, que rastrea la ubicación física de los fragmentos de fragmentos y las copias de réplica en el disco. Tabla 1 muestra una muestra de una partición de la tabla de fragmentos que, para cada fragmento, identifica su ubicación física enumerando el disco dentro del nodo, el archivo dentro del disco, la compensación dentro de ese archivo y la longitud de los datos. Aquí podemos ver que el ID de fragmento C1 está codificado para eliminación y el ID de fragmento C2 tiene un espejeado triple. Puede encontrar más detalles sobre el espejeado triple y la codificación de eliminación en la sección 1.4 de este documento.

Tabla 1 Ejemplo de partición de tabla de fragmentos

ID del fragmento	Ubicación del fragmento
C1	Node1:Disk1:file1:offset1:length Node2:Disk1:File1:offset1:length Node3:Disk1:File1:offset1:length Node4:Disk1:File1:offset1:length Node5:Disk1:File1:offset1:length Node6:Disk1:File1:offset1:length Node7:Disk1:File1:offset1:length Node8:Disk1:File1:offset1:length Node1:Disk2:File1:offset1:length Node2:Disk2:File1:offset1:length Node3:Disk2:File1:offset1:length Node4:Disk2:File1:offset1:length Node5:Disk2:File1:offset1:length Node6:Disk2:File1:offset1:length Node7:Disk2:File1:offset1:length Node8:Disk2:File1:offset1:length
C2	Node1:Disk3:File1:offset1:length Node2:Disk3:File1:offset1:length Node3:Disk3:File1:offset1:length

Otro ejemplo es una tabla de objetos que se utiliza para el mapeo de nombre de objeto a fragmento. Tabla 2 muestra un ejemplo de una partición de una tabla de objetos que detalla qué fragmentos y dónde dentro del fragmento reside un objeto.

Tabla 2 Ejemplo de tabla de objetos

Nombre de objeto	ID del fragmento
ImgA	C1:offset:length
FileA	C4:offset:length C6:offset:length

El mapeo de los propietarios de particiones de tabla se mantiene mediante un servicio, denominado vnest, que se ejecuta en todos los nodos. Tabla 3 muestra un ejemplo de una parte de una tabla de mapeo de vnest.

Tabla 3 Ejemplo de tabla de mapeo de vnest

ID de tabla	Propietario de partición de tabla
Tabla 0 P1	Nodo 1
Tabla 0 P2	Nodo 2

1.3 Dominios de falla

En general, los dominios de fallas se relacionan con un concepto de diseño de ingeniería que tiene en cuenta los componentes dentro de una solución que tienen un potencial de falla. El software ECS detecta automáticamente qué discos están en el mismo nodo y qué nodos se encuentran en el mismo rack. A fin de protegerse contra la mayoría de los escenarios de falla, el software ECS está diseñado para utilizar esta información cuando se escriben datos. Las reglas básicas de los dominios de fallas que utiliza ECS incluyen lo siguiente:

- ECS nunca escribe fragmentos del mismo fragmento en el mismo disco en un nodo
- ECS distribuye fragmentos de un fragmento por igual entre los nodos
- ECS reconoce el rack, si un VDC/sitio contiene más de un rack, suponiendo que hay suficiente espacio, ECS utiliza los mejores esfuerzos para distribuir por igual fragmentos de un fragmento entre estos racks

1.4 Métodos avanzados de protección de datos

Dentro de ECS, cuando se crea un objeto, incluye la escritura de datos, metadatos personalizados y metadatos de ECS. Los metadatos de ECS incluyen fragmentos de registro y fragmentos de btree. Cada uno se escribe en un fragmento lógico diferente que contendrá aproximadamente 128 MB de datos de uno o más objetos. ECS utiliza una combinación de espejeado triple y codificación de eliminación para proteger los datos dentro de un centro de datos virtual (VDC)/sitio.

- El espejeado triple garantiza la escritura de tres copias de datos, lo que brinda protección contra fallas de dos nodos.
- La codificación de eliminación proporciona protección de datos mejorada contra fallas de discos y nodos. Utiliza el esquema de codificación de eliminación de Reed Solomon que divide fragmentos en datos y fragmentos de codificación que se distribuyen equitativamente entre los nodos dentro de un VDC/sitio.

Según el tamaño y el tipo de datos, se escribirán mediante uno de los métodos de protección de datos que se muestran en Tabla 4.

Tabla 4 Determinación del nivel de protección de datos que se utilizará para diferentes tipos de datos

Tipo de datos	Métodos de protección de datos utilizados
Fragmentos de registro	Espejeado triple
Fragmentos de BTree/metadatos personalizados	Codificación de eliminación con segmentos de datos redundantes
<128 MB de datos de objetos	Espejeado triple más codificación de eliminación en el lugar
>128 MB de datos de objetos	Codificación de eliminación en línea

Nota: en la arquitectura todo flash, como EXF900, la protección de fragmentos de btree es el espejeado triple

1.4.1 Espejeado triple

El método de escritura de espejeado triple se aplica a los fragmentos de registro de ECS, de los cuales ECS crea tres copias de réplica. Cada copia de réplica se escribe en un solo disco en diferentes nodos entre dominios de fallas. Este método protege los datos de los fragmentos contra fallas de dos nodos o dos discos.

Figura 3 muestra un ejemplo de espejeado triple en el que un fragmento lógico, que contiene metadatos de 128 MB, tiene tres copias de réplica, cada una escrita en un nodo diferente.

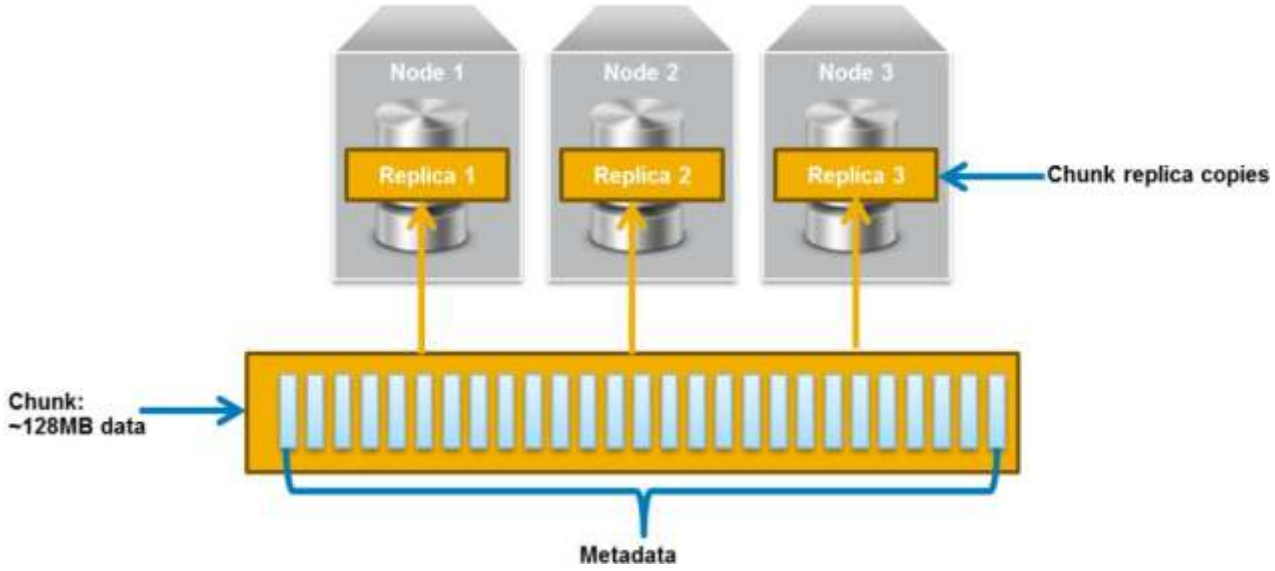

Figura 3 Espejeado triple

1.4.2 Codificación de eliminación con segmentos de datos redundantes

El método de escritura de codificación de eliminación con segmentos de datos redundantes se aplica a los fragmentos de btree de ECS y a los metadatos de objetos personalizados. Incluye 12 segmentos de datos, 12 segmentos de datos replicados y 4 segmentos de paridad. El nuevo esquema de EC de datos redundantes de fragmentos de btree ahorra sobrecarga de protección de metadatos.

1.4.3 Espejeado triple más codificación de eliminación en el lugar

Este método de escritura se aplica a los datos de cualquier objeto que tenga un tamaño inferior a 128 MB.

A medida que se crea un objeto, se escribe en un fragmento, el cual ECS crea tres copias de réplica de la siguiente manera:

- Una copia se escribe en fragmentos que se distribuyen entre diferentes nodos y discos.
  La distribución distribuye los fragmentos en tantos dominios de fallas como sea posible. El tamaño que se escribe en cada disco depende del esquema de codificación de eliminación que se utiliza.
  - Si el esquema de codificación de eliminación es el predeterminado (12+4), cada disco obtendría un máximo de ~10,67 MB
  - Si el esquema de codificación de eliminación es almacenamiento inactivo (10+2), cada disco obtendría un máximo de aproximadamente 12,8 MB
- Se escribe una segunda copia de réplica del fragmento en un solo disco de un nodo.
- Una tercera copia de réplica del fragmento se escribe en un solo disco en un nodo diferente.

Este método proporciona espejeado triple y protege los datos de fragmentos contra fallas de dos nodos o dos discos.

Se escribirán objetos adicionales en el mismo fragmento hasta que contenga aproximadamente 128 MB de datos o después de un tiempo predefinido, lo que sea menor. En este momento, el esquema de codificación de eliminación de Reed Solomon calcula los fragmentos de codificación (paridad) para el fragmento y los escribe en diferentes discos. Esto garantiza que todos los fragmentos dentro de un fragmento, incluidos los fragmentos de codificación, se escriban en diferentes discos y se distribuyan entre dominios de fallas.

Una vez que los fragmentos de codificación se hayan escrito en el disco, las copias de segunda y tercera réplica se eliminarán del disco. Una vez que finaliza esto, el fragmento está protegido por la codificación de eliminación, lo que proporciona niveles más altos de disponibilidad que el espejeado triple.

1.4.4 Codificación de eliminación en línea

Este método de escritura se aplica a los datos de cualquier objeto de 128 MB o más. Los objetos se dividen en fragmentos de 128 MB. El esquema de codificación de eliminación de Reed Solomon calcula los fragmentos de codificación (paridad) para cada fragmento. Cada fragmento se escribirá en diferentes discos y se distribuirá entre dominios de fallas. El tamaño que se escribe en cada disco depende del esquema de codificación de eliminación que se utiliza.

- Si el esquema de codificación de eliminación es el predeterminado (12+4), los fragmentos se distribuirán entre 16 discos, cada fragmento será de aproximadamente 10,67 MB
- Si el esquema de codificación de eliminación es almacenamiento inactivo (10+2), los fragmentos se distribuirán en 12 discos, cada fragmento será de aproximadamente 12,8 MB

Cualquier parte restante de un objeto que sea inferior a 128 MB se escribirá mediante el espejeado triple más el esquema de codificación de eliminación en el lugar mencionado anteriormente. Por ejemplo, si un objeto es de 150 MB, 128 MB se escribirán mediante la codificación de eliminación en línea, los 22 MB restantes se escribirán con espejeado triple y la codificación de eliminación en el lugar.

Figura 4 muestra un ejemplo de cómo los fragmentos se distribuyen entre dominios de fallas. Este ejemplo tiene un solo VDC/sitio que abarca dos racks, cada uno de los que contiene cuatro nodos.

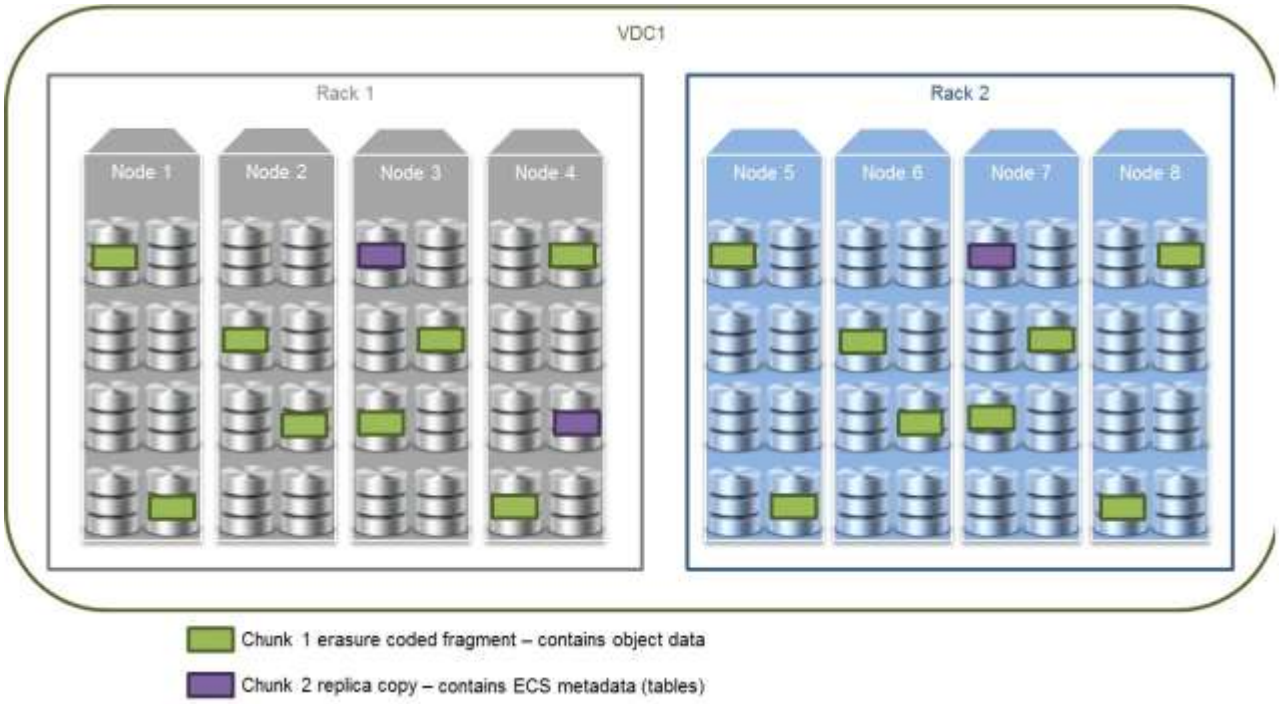

Figura 4 Cómo se distribuyen los fragmentos entre dominios de fallas

- El fragmento 1 contiene datos de objetos que se codificaron con una codificación de eliminación de 12+4. Los fragmentos se distribuyen por igual en los 8 nodos, cuatro por rack. Cada nodo contiene 2 fragmentos que escribe en dos discos diferentes.
- El fragmento 2 contiene metadatos de ECS (tablas) y, por lo tanto, es espejeado triple. Cada copia de réplica se escribe en un nodo diferente, cada una en un solo disco. Las copias abarcan racks para proporcionar la mayor disponibilidad.

1.5 Niveles de protección de codificación de eliminación

Según el esquema de codificación de eliminación que se seleccionó durante la creación del pool de almacenamiento, los datos codificados de eliminación están protegidos contra las siguientes fallas.

1.5.1 Esquema de código de eliminación predeterminado (12+4):

ECS requiere un mínimo de cuatro nodos para poder realizar la codificación de eliminación mediante el esquema de codificación de eliminación predeterminado. La codificación de eliminación se detendrá si un pool de almacenamiento contiene menos de cuatro nodos, lo que significa que el nivel de protección será el espejeado triple. Durante este tiempo, las tres copias de réplica permanecerán y la paridad no se calculará en ningún fragmento. Una vez que se agregan nodos adicionales al pool de almacenamiento y se cumple con la cantidad mínima admitida de nodos, la codificación de eliminación continuará en estos, así como en los fragmentos nuevos.

Para cada fragmento de 128 MB, el esquema de codificación de eliminación predeterminado escribe doce fragmentos de datos y cuatro fragmentos de codificación, cada uno de aproximadamente 10,67 MB de tamaño. Protege los datos de los fragmentos contra la pérdida de hasta cuatro fragmentos de un fragmento, lo que puede incluir los siguientes escenarios de falla que se muestran en Tabla 5.

Tabla 5 Protección de codificación de eliminación predeterminada

N.º de nodos en VDC	Cantidad de fragmentos por nodo	Datos codificados de eliminación protegidos contra
5 nodos	4	• Pérdida de hasta cuatro discos o • Pérdida de un nodo
6 ó 7 nodos	3	• Pérdida de hasta cuatro discos o • Pérdida de un nodo y de un disco de un segundo nodo
8 o más nodos	2	• Pérdida de hasta cuatro discos o • Pérdida de dos nodos o • Pérdida de un nodo y de dos discos
16 o más nodos	1	• Pérdida de cuatro nodos o • Pérdida de tres nodos y discos de un nodo adicional o • Pérdida de dos nodos y discos de hasta dos nodos diferentes o • Pérdida de un nodo y de discos de hasta tres nodos diferentes o • Pérdida de cuatro discos de cuatro nodos diferentes

Nota: Tabla 5 refleja los niveles de protección posibles con la distribución completa de fragmentos. Puede haber escenarios en los que existan más fragmentos en un nodo, como si un nodo no tiene suficiente espacio disponible. En este caso, los niveles de protección pueden variar.

1.5.2 Esquema de codificación de eliminación del almacenamiento inactivo (10+2):

ECS requiere un mínimo de seis nodos para poder realizar la codificación de eliminación mediante el esquema de codificación de eliminación del almacenamiento inactivo. La codificación de eliminación se detendrá si un pool de almacenamiento contiene menos de seis nodos, lo que significa que las tres copias de réplica permanecerán y la paridad no se calculará en un fragmento. Una vez que se agregan nodos adicionales al pool de almacenamiento, la codificación de eliminación continuará en estos, así como en los nuevos fragmentos.

Para cada fragmento de 128 MB, el esquema de codificación de eliminación del almacenamiento inactivo escribe diez fragmentos de datos y dos fragmentos de codificación, cada uno de aproximadamente 12,8 MB de tamaño. Protege los datos del fragmento contra la pérdida de hasta dos fragmentos de un fragmento, lo que puede incluir los siguientes escenarios de falla que se muestran en Tabla 6.

Tabla 6 Protección de codificación de eliminación de almacenamiento inactivo

N.º de nodos en VDC	Cantidad de fragmentos por nodo	Datos codificados de eliminación protegidos contra
11 o menos nodos	2	• Pérdida de hasta dos discos o • Pérdida de un nodo
12 o más nodos	1	• Pérdida de cualquier cantidad de discos de dos nodos diferentes o • Pérdida de dos nodos

Nota: esta tabla refleja los niveles de protección posibles con la distribución completa de fragmentos. Puede haber escenarios en los que existan más fragmentos en un nodo, como si un nodo no tiene suficiente espacio disponible. En este caso, los niveles de protección pueden variar.

1.6 Sumas de comprobación

Otro mecanismo que ECS utiliza para garantizar la integridad de los datos es almacenar la suma de comprobación de los datos escritos. Las sumas de comprobación se realizan por unidad de escritura, hasta 2 MB. Por lo tanto, se pueden producir sumas de comprobación para un fragmento de objeto para escrituras de objetos grandes o por objeto para escrituras de objetos pequeños de menos de 2 MB. Durante las operaciones de escritura, la suma de comprobación se calcula en la memoria y, a continuación, se escribe en el disco. En las lecturas, los datos se leen junto con la suma de comprobación y, a continuación, la suma de comprobación se calcula en la memoria a partir de la lectura de datos y se compara con la suma de comprobación almacenada en el disco para determinar la integridad de los datos. Además, el motor de almacenamiento ejecuta un comprobador de coherencia periódicamente en segundo plano y verifica la suma de comprobación en todo el conjunto de datos.

1.7 Escritura de objetos

Cuando se produce una operación de escritura en ECS, comienza con un cliente que envía una solicitud a un nodo. ECS se diseñó como una arquitectura distribuida, lo que permite que cualquier nodo en un VDC/sitio responda a una solicitud de lectura o escritura. Una solicitud de escritura implica escribir los datos del objeto, los metadatos de objetos personalizados y registrar la transacción en un registro de registro. Una vez que se completa, se envía un acuso de recibo al cliente.

Figura 5 y los pasos mencionados anteriormente recorren una descripción general de alto nivel de un flujo de trabajo de escritura.

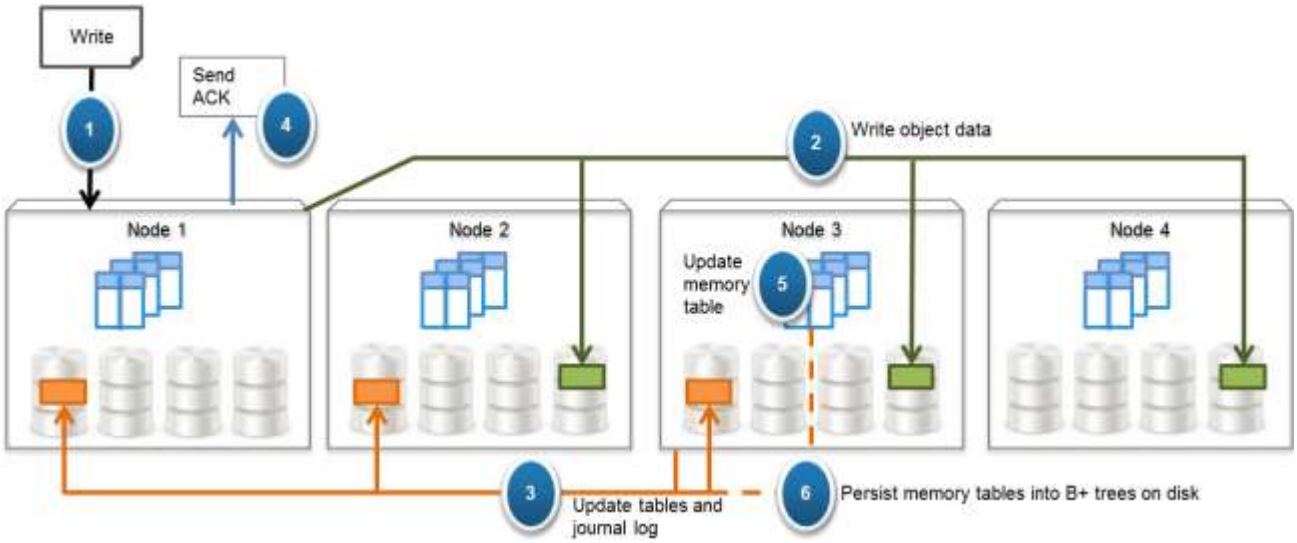

Figura 5 Flujo de trabajo de escritura de objetos

1. Se recibe una solicitud de objeto de escritura. Cualquier nodo puede responder a esta solicitud, pero, en este ejemplo, el nodo 1 procesa la solicitud.
2. Según el tamaño del objeto, los datos se escriben en uno o más fragmentos. Cada fragmento está protegido mediante esquemas de protección de datos avanzados, como el espejeado triple y la codificación de eliminación. Antes de escribir los datos en el disco, ECS ejecuta una función de suma de comprobación y almacena el resultado.

   Los datos se agregan a un fragmento; dado que este objeto tiene solo 10 MB de tamaño, utiliza el espejeado triple más el esquema de codificación de eliminación en el lugar. Esto da como resultado escrituras en tres discos en tres nodos diferentes, en este ejemplo, el nodo 2, el nodo 3 y el nodo 4. Estos tres nodos envían reconocimientos al nodo 1.
3. Después de que los datos del objeto se escriben correctamente, se almacenan los metadatos del objeto. En este ejemplo, el nodo 3 es propietario de la partición de la tabla de objetos a la que pertenece este objeto. Como propietario, el nodo 3 escribe el nombre del objeto y el ID del fragmento en esta partición de los registros de registro de la tabla de objetos. Los registros de registro son espejeados triples, por lo que el nodo 3 envía copias de réplica a tres nodos diferentes en paralelo, en este ejemplo, el nodo 1, el nodo 2 y el nodo 3.
4. El reconocimiento se envía al cliente.
5. En un proceso en segundo plano, se actualiza la tabla de memoria.
6. Una vez que la tabla en la memoria se llena o después de un período establecido, la tabla se combina, se ordena o se vuelca en árboles B+ como fragmentos y se registra un punto de control en el registro.

1.8 Lectura de objetos

ECS se diseñó como una arquitectura distribuida, lo que permite que cualquier nodo en un VDC/sitio responda a una solicitud de lectura o escritura. Una solicitud de lectura implica encontrar la ubicación física de los datos mediante búsquedas de tabla desde el propietario del registro de partición, así como lecturas de compensación de bytes, validación de suma de comprobación y devolución de los datos al cliente solicitante.

Figura 6 y los pasos mencionados anteriormente recorren una descripción general del flujo de trabajo de lectura.

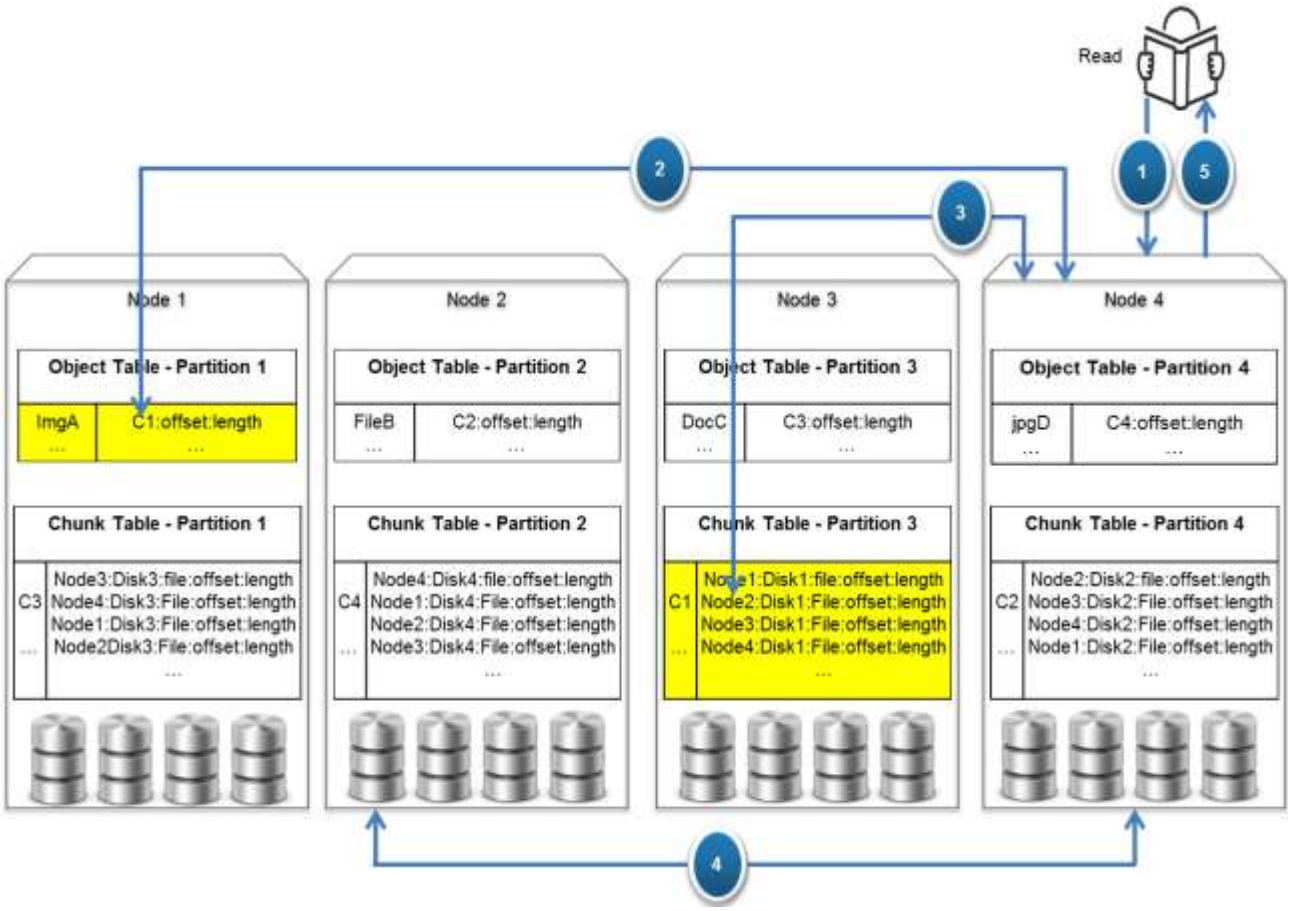

Figura 6 Flujo de trabajo de lectura de objetos

1. Se recibe una solicitud de lectura para ImgA. Cualquier nodo puede responder a esta solicitud, pero, en este ejemplo, el nodo 4 procesa la solicitud.
2. El nodo 4 solicita la información del fragmento del nodo 1 (propietario de la partición de tabla de objetos para ImgA).
3. Sabiendo que ImgA está en C1 en una compensación y longitud específicas, el nodo 4 solicita la ubicación física del fragmento desde el nodo 3 (propietario de la partición de tabla de fragmentos para C1).
4. Ahora que el nodo 4 conoce la ubicación física de ImgA, solicitará los datos de los nodos que contienen los fragmentos de datos de ese archivo, en este ejemplo, el nodo 2, disco 1. Los nodos realizarán una lectura de compensación de bytes y devolverán los datos al nodo 4.
5. El nodo 4 valida la suma de comprobación y, a continuación, devuelve los datos al cliente solicitante.

Nota: en el paso 4, para la arquitectura todo flash como EXF900, cada nodo puede leer datos de otro nodo directamente, a excepción de la arquitectura de la unidad de disco duro, como EX300, EX500 y EX3000, y cada nodo solo puede leer el almacén de datos en sí mismo.

2 Disponibilidad del sitio local

La naturaleza distribuida de la arquitectura de ECS proporciona alta disponibilidad en forma de disponibilidad del sistema y durabilidad de los datos frente a una variedad de fallas. Esta sección se centra en cómo se mantiene la disponibilidad durante fallas del sitio local.

2.1 Disk failure

En la sección de arquitectura, se describió cómo ECS utiliza una combinación de espejeado triple y codificación de eliminación para escribir datos de manera distribuida con el fin de ser resiliente frente a diversos escenarios de falla.

Para garantizar que la coherencia de las sumas de comprobación de datos se valide en lecturas y mediante un comprobador de coherencia. El comprobador de coherencia es un proceso en segundo plano que realiza periódicamente la verificación de la suma de comprobación en todo el conjunto de datos. Las solicitudes de lectura también ejecutan la verificación de suma de comprobación.

Si falta un fragmento en una solicitud de lectura, debido a que una unidad no responde o a una verificación de suma de comprobación fallida, se enviará una notificación al administrador de fragmentos. El administrador de fragmentos iniciará una reconstrucción de los fragmentos faltantes mediante el uso de los datos y fragmentos de paridad de eliminación restantes o las copias de réplica, después de lo cual actualizará la información del fragmento. Una vez que se vuelvan a crear los fragmentos, cualquier solicitud de lectura pendiente o nueva utilizará la información actualizada del fragmento para solicitar los fragmentos de datos y reparar la solicitud de lectura.

Los nodos de ECS realizan evaluaciones del estado constantemente en los discos conectados directamente a ellos. Si un disco deja de responder, el nodo de ECS notificará al administrador de fragmentos para que deje de incluirlo en las nuevas operaciones de escritura. Si deja de responder después de un período (el valor predeterminado es 60 minutos), se enviará una notificación al administrador de fragmentos para volver a crear los datos desde la unidad fallida. El nodo de ECS identificará qué fragmentos tienen bloques en la unidad fallida y, por lo tanto, se deben recuperar. Enviará esta información al administrador de fragmentos, que iniciará la recuperación paralela de todos los fragmentos almacenados en la unidad fallida. Los fragmentos de fragmentos se recuperan en otros discos mediante las copias de réplica o los fragmentos codificados de eliminación restantes. A medida que se escriben nuevos fragmentos, las tablas de fragmentos asociadas se actualizan con esta información. Si es posible, el administrador de fragmentos también eliminará los fragmentos de la unidad fallida. Si el disco vuelve a estar en línea y:

- No ha respondido durante menos de un período establecido (el valor predeterminado es 90 minutos), se cancelarán las operaciones de recuperación restantes.
- No ha respondido durante un período establecido (el valor predeterminado es 90 minutos) o se informa como fallido por el administrador de hardware, ECS eliminará la unidad. Una vez que se quita una unidad, las operaciones de recuperación restantes continuarán hasta que se completen. Cuando se completa la recuperación, el administrador de fragmentos elimina todas las referencias a este disco fallido en la tabla de fragmentos.

  Si esta unidad se encuentra en línea después de quitarla, se agregará como una nueva unidad y el administrador de fragmentos la incluirá en las nuevas operaciones de escritura.

2.2 Falla de un nodo ECS

ECS realiza constantemente evaluaciones del estado en los nodos. A fin de mantener la disponibilidad del sistema, la arquitectura distribuida de ECS permite que cualquier nodo acepte las solicitudes del cliente. Si un nodo está inactivo, un cliente se puede redirigir de forma manual o automática (p. ej., mediante DNS o un balanceador de carga) a otro nodo que puede brindar servicio a la solicitud.

Para no desencadenar operaciones de reconstrucción para eventos falsos, no se activa una operación de reconstrucción completa, a menos que un nodo falle una cantidad establecida de evaluaciones del estado secuenciales, el valor predeterminado es 60 minutos. Si aparece una solicitud de I/O para un nodo que no responde, pero antes de que se active una reconstrucción completa:

- Cualquier solicitud de una tabla de particiones alojada en un nodo que no responde desencadenará que la propiedad de las tablas de partición solicitadas se redistribuya entre los nodos restantes en el sitio. Una vez que esto finalice, la solicitud se completará correctamente.
- Las solicitudes de I/O de datos que existen en los discos del nodo que no responde se reconstruirán con los datos codificados de eliminación restantes y los fragmentos de paridad o las copias de réplica, después de lo cual se actualizará la información del fragmento. Una vez que se vuelvan a crear los fragmentos, cualquier solicitud de lectura pendiente o nueva utilizará la información actualizada del fragmento para solicitar los fragmentos de datos y reparar la solicitud de lectura.

Después de que un nodo falla una cantidad establecida de evaluaciones del estado secuenciales, el valor predeterminado es 60 minutos, se considera que un nodo está inactivo. Esto activa automáticamente una operación de recreación de las tablas de partición y los fragmentos de fragmentos en los discos que pertenecen al nodo fallido.

Como parte de la operación de recreación, se enviará una notificación al administrador de fragmentos que inicia una recuperación paralela de todos los fragmentos almacenados en los discos de los nodos fallidos. Esto puede incluir fragmentos que contienen datos de objetos, metadatos proporcionados por el cliente personalizado y metadatos de ECS. Si el nodo fallido vuelve a estar en línea, se enviará un estado actualizado al administrador de fragmentos y se cancelarán todas las operaciones de recuperación que no se completaron. Encontrará más detalles sobre la recuperación de fragmentos en la sección de falla de disco anterior.

Además del monitoreo de hardware, ECS también monitorea todos los servicios y las tablas de datos en cada nodo.

- Si hay una falla de tabla, pero el nodo aún está activo, intentará reinicializar automáticamente la tabla en el mismo nodo.
- Si detecta una falla de servicio, primero intentará reiniciar el servicio.

Si eso falla, redistribuirá la propiedad de las tablas que pertenecen al nodo inactivo o al servicio en todos los nodos restantes en el VDC/sitio. Los cambios de propiedad implican actualizar la información de vnest y volver a crear las tablas de memoria que pertenecen al nodo fallido. La información de vnest se actualizará en los nodos restantes con la nueva información del propietario de la tabla de particiones.

Las tablas de memoria del nodo fallido se volverán a crear mediante la reproducción de entradas de registro escritas después del punto de control de registro correcto más reciente.

2.2.1 Fallas de múltiples nodos

Hay escenarios en los que varios nodos pueden fallar dentro de un sitio. Varios nodos pueden fallar uno por uno o simultáneamente.

- **Falla de uno a uno:** cuando los nodos fallan uno por uno, significa que falla un nodo, se completan todas las operaciones de recuperación y, a continuación, se produce un error en un segundo nodo. Esto puede ocurrir varias veces y es análogo a un VDC que va desde algo así como 4 sitios → 3 sitios → 2 sitios → 1 sitio. Esto requiere que los nodos restantes tengan suficiente espacio para completar las operaciones de recuperación.
- **Falla simultánea:** cuando los nodos fallan simultáneamente, significa que los nodos fallan casi al mismo tiempo o un nodo falla antes de que se complete la recuperación de un nodo fallido anterior.

El impacto de la falla depende de los nodos que se desactivan. Tabla 7 y Tabla 8 describen el escenario del mejor de los casos para las tolerancias a fallas de un solo sitio en función del esquema de codificación de eliminación y la cantidad de nodos en un VDC.

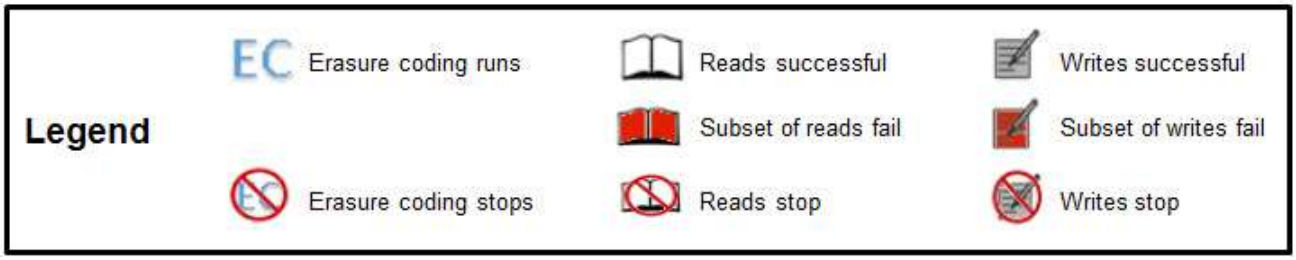

Tabla 7 Escenario de mejor de los casos de fallas de múltiples nodos dentro de un sitio según el esquema de codificación de eliminación 12+4 predeterminado

Cantidad de nodos en VDC durante la creación	Cantidad total de nodos fallidos desde la creación del VDC	Estado después de fallas simultáneas	Estado después de la falla más reciente, una por una	Estado actual del VDC después de fallas de una por una anteriores
5 nodos	1	EC	EC	VDC de 5 nodos con falla de 1 nodo
	2			El VDC anteriormente pasó de 5 → 4 nodos, ahora 1 nodo adicional presenta fallas
	3 a 4			Anteriormente, el VDC pasó de 5 → 4 → 3 o 5 → 4 → 3 → 2 nodos, ahora 1 nodo adicional presenta fallas
6 nodos	1	EC	EC	VDC de 6 nodos con falla de 1 nodo
	2	EC	EC	El VDC anteriormente pasó de 6 → 5 nodos, ahora 1 nodo adicional presenta fallas

Cantidad de nodos en VDC durante la creación	Cantidad total de nodos fallidos desde la creación del VDC	Estado después de fallas simultáneas	Estado después de la falla más reciente, una por una	Estado actual del VDC después de fallas de una por una anteriores
	3			El VDC anteriormente pasó de 6 → 5 → 4 nodos, ahora 1 nodo adicional presenta fallas
	4 a 5			El VDC anteriormente pasó de 6 → 5 → 4 → 3 o 6 → 5 → 4 → 3 → 2 nodos, ahora 1 nodo adicional presenta fallas
8 nodos	1 a 2			VDC de 8 nodos o VDC que pasó de 8 → 7 nodos, ahora falla 1 nodo
	3 a 4			El VDC anteriormente pasó de 8 → 7 → 6 u 8 → 7 → 6 → 5 nodos, ahora 1 nodo adicional presenta fallas
	5			El VDC anteriormente pasó de 8 → 7 → 6 → 5 → 4 nodos, ahora 1 nodo adicional presenta fallas
	6 a 7			El VDC anteriormente pasó de 8 → 7 → 6 → 5 → 4 → 3 nodos o 8 → 7 → 6 → 5 → 4 → 3 → 2 nodos, ahora falla 1 nodo adicional

Tabla 8 Escenario de mejor de los casos de fallas de múltiples nodos dentro de un sitio según el esquema de codificación de eliminación de almacenamiento inactivo 10+2

Cantidad de nodos en VDC durante la creación	Cantidad total de nodos fallidos desde la creación del VDC	Estado después de fallas simultáneas	Estado después de la falla más reciente, una por una	Estado actual del VDC después de fallas anteriores una por una
6 nodos	1			VDC de 6 nodos con falla de 1 nodo
	2			El VDC anteriormente pasó de 6 → 5 nodos, ahora 1 nodo adicional presenta fallas
	3			El VDC anteriormente pasó de 6 → 5 → 4 nodos, ahora 1 nodo adicional presenta fallas

Cantidad de nodos en VDC durante la creación	Cantidad total de nodos fallidos desde la creación del VDC	Estado después de fallas simultáneas	Estado después de la falla más reciente, una por una	Estado actual del VDC después de fallas anteriores una por una
	4 a 5			El VDC anteriormente pasó de 6 → 5 → 4 → 3 o 6 → 5 → 4 → 3 → 2 nodos, ahora 1 nodo adicional presenta fallas
8 nodos	1			VDC de 8 nodos con falla de 1 nodo
	2			El VDC anteriormente pasó de 8 → 7 nodos, ahora 1 nodo adicional presenta fallas
	3 a 5			El VDC anteriormente pasó de 8 → 7 → 6 o 8 → 7 → 6 → 5 o 8 → 7 → 6 → 5 → 4 nodos, ahora 1 nodo adicional presenta fallas
	6 a 7			El VDC anteriormente pasó de 8 → 7 → 6 → 5 → 4 → 3 nodos o 8 → 7 → 6 → 5 → 4 → 3 → 2 nodos, ahora falla 1 nodo adicional
12 nodos	1 a 2			VDC de 12 nodos o VDC de 12 nodos que pasaron de 12 → 11 nodos, ahora 1 nodo adicional presenta fallas
	3 a 6			El VDC anteriormente pasó de 12→ 11 → 10 o 12→ 11 → 10 → 9 o 12→ 11 → 10 → 9 → 8 o 12→ 11 → 10 → 9 → 8 → 7 nodos, ahora 1 nodo adicional presenta fallas
	7 a 9			El VDC anteriormente pasó de 12→ 11 → 10 → 9 → 8 → 7 → 6 o 12→ 11→ 10→ 9→ 8 → 7→ 6→ 5 o 12→ 11→ 10→ 9→ 8→ 7→ 6→ 5→ 4 nodos, ahora 1 nodo adicional presenta fallas
	10 a 11			El VDC anteriormente pasó de 12→11→10→9→8→7→6→5→4→3 o 12→11→10→9→8→7→6→5→4→3→ 2 nodos, ahora 1 nodo adicional presenta fallas

Las reglas básicas para determinar qué operaciones fallan en un solo sitio con fallas de múltiples nodos incluyen:

- Si tiene tres o más fallas simultáneas de nodos, algunas lecturas y escrituras fallarán debido a la posible pérdida de las tres copias de réplica de los fragmentos de metadatos espejeados triples asociados.
- Las escrituras requieren tres nodos como mínimo.
- La codificación de eliminación se detendrá y los fragmentos codificados de eliminación se convertirán en protección de espejeado triple si la cantidad de nodos es menor que el mínimo necesario para cada esquema de codificación de eliminación. Dado que el esquema de codificación de eliminación predeterminado, 12+4 requiere 4 nodos, la codificación de eliminación se detendrá si hay menos de 4 nodos. Para la codificación de eliminación del almacenamiento inactivo, 10+2, la codificación de eliminación se detendrá si hay menos de 6 nodos.
- Si el conteo de nodos es inferior al mínimo necesario para el esquema de codificación de eliminación, los fragmentos codificados de eliminación se convertirán en protección de espejeado triple. Por ejemplo, en un VDC con codificación de eliminación predeterminada y 4 nodos, después de una falla de nodo, sucedería lo siguiente:

  - La falla del nodo hace que se pierdan 4 fragmentos.
  - Los fragmentos faltantes se reconstruyen.
  - El fragmento crea 3 copias de réplica, una en cada nodo.
  - Se elimina la copia de EC.

- Las fallas una por una actúan como fallas de nodo único. Por ejemplo, si pierde dos nodos uno por uno, cada falla consiste únicamente en recuperar datos de una interrupción de nodo único.

  Por ejemplo, con 6 nodos y codificación de eliminación predeterminada:

  - Primera falla: cada uno de los seis nodos tiene hasta tres fragmentos (16 fragmentos/6 nodos). Los tres fragmentos faltantes se vuelven a crear en los nodos restantes. Una vez finalizada la recuperación, el VDC termina con 5 nodos.
  - Segunda falla: cada uno de los cinco nodos tiene hasta 4 fragmentos (16 fragmentos/5 nodos). Los cuatro fragmentos faltantes se vuelven a crear en los nodos restantes. Una vez finalizada la recuperación, el VDC termina con 4 nodos.
  - Tercera falla: cada uno de los cuatro nodos tiene hasta 4 fragmentos (16 fragmentos/4 nodos). Los cuatro fragmentos faltantes se vuelven a crear en los nodos restantes. Una vez que se completa la recuperación, el VDC termina con 3 nodos y, dado que este es inferior al mínimo para la codificación de eliminación, el fragmento codificado de eliminación se reemplaza por tres copias de réplica distribuidas en los nodos restantes.
  - Falla sucesiva: cada uno de los tres nodos tiene una copia de réplica. La copia de réplica faltante se vuelve a crear en uno de los nodos restantes. Una vez finalizada la recuperación, el VDC termina con 2 nodos.
  - Quinta falla: hay tres copias de réplica, dos en un nodo y una en el otro nodo. Las copias de réplica faltantes se vuelven a crear en el nodo restante. Una vez finalizada la recuperación, el VDC termina con 1 nodo.

3 Descripción general del diseño de múltiples sitios

Más allá de la disponibilidad del sistema y la durabilidad de los datos diseñadas dentro de un solo sitio, ECS también incluye protección contra una falla completa en todo el sitio. Esto se logra en una implementación de múltiples sitios mediante la federación de varios VDC/sitios y la configuración de la replicación geográfica.

La federación de sitios implica proporcionar terminales de replicación y administración para la comunicación entre sitios. Una vez federados los sitios, se pueden administrar como una sola infraestructura.

Las políticas de grupos de replicación determinan cómo se protegen los datos y desde dónde se puede acceder a estos. ECS es compatible con la replicación geoactiva y la replicación geopasiva. La replicación geoactiva proporciona acceso activo-activo a los datos, lo que permite que se lean y escriban desde cualquier sitio dentro de su grupo de replicación definido.

Cuando se replican series todo flash como EXF900 entre sitios, se deben considerar los posibles impactos en el rendimiento a través de la WAN. Una gran ingesta puede poner una carga alta en el vínculo, lo que causa saturación o un RPO diferido, además de que un usuario/aplicación puede experimentar tiempos de latencia más altos en lecturas y escrituras remotas en comparación con las solicitudes locales. El otro debe considerar la recolección parcial de elementos no utilizados fallida, una gran ingesta de sitios locales y replicados puede causar que el sistema alcance el 90 % pronto, lo que hará que el sistema deje de escribir datos y recuperar datos.

Nota: recolección de elementos no utilizados parcial: cuando los 2/3 de un fragmento son elementos no utilizados, recupera el fragmento combinando las partes válidas de este con otros fragmentos parcialmente llenos en un nuevo fragmento y luego recupera el espacio.

La replicación también se puede configurar como geopasivo, lo que designa de dos a cuatro sitios de origen y uno o dos sitios que se usarán solo como destino de replicación. Los destinos de replicación se utilizan solo con fines de recuperación. Los destinos de replicación bloquean el acceso directo del cliente para las operaciones de creación/actualización/eliminación.

Los beneficios de la replicación geopasiva incluyen:

- Puede optimizar la eficiencia del almacenamiento aumentando las posibilidades de que las operaciones XOR se ejecuten, ya que garantiza que las escrituras de ambos sitios de origen vayan al mismo destino de replicación.
- Permite que el administrador controle dónde existe la copia de replicación de los datos, como en un escenario de respaldo en la nube.

ECS proporciona opciones de configuración de replicación geográfica en el nivel del depósito, lo que permite que el administrador configure diferentes niveles de replicación para diferentes depósitos.

Figura 7 muestra un ejemplo de cómo un administrador podría configurar la replicación de tres depósitos:

- Depósito A: datos de prueba/desarrollo de ingeniería; no replicar, mantener solo localmente
- Depósito B: datos de ventas europeos: replicar entre sitios dentro de Europa únicamente
- Depósito C: datos de capacitación en toda la empresa: replicar a todos los sitios dentro de la empresa

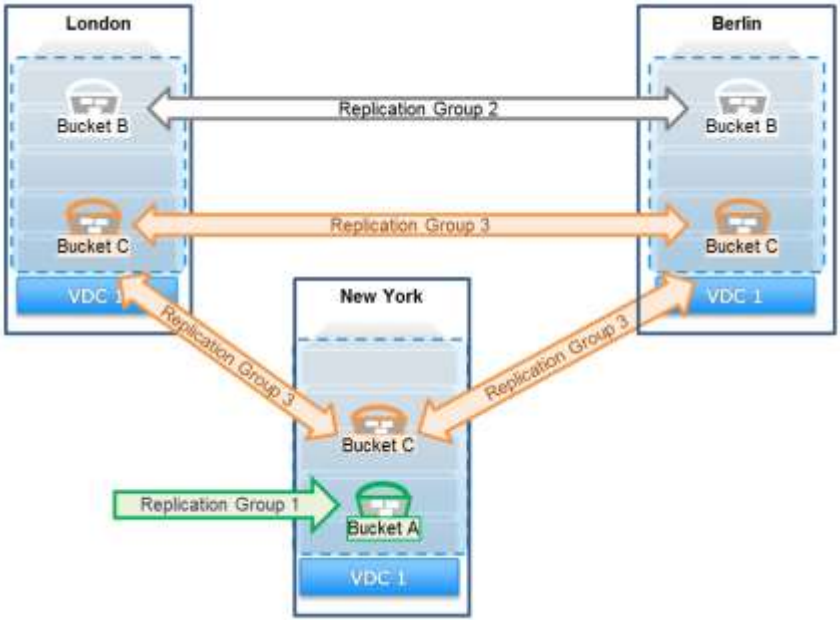

Figura 7 Ejemplo que muestra cómo diferentes depósitos pueden usar diferentes políticas de replicación

Como práctica recomendada, se recomienda configurar grupos de replicación para rutas de replicación específicas. Por ejemplo, en Figura 7 hay un grupo de replicación que replica los datos entre Londres y Berlín. Esto se debe utilizar para todos los depósitos que necesitan la replicación entre Londres y Berlín.

Los datos replicados geográficamente están protegidos mediante el almacenamiento de una copia principal de los datos en el sitio local y una copia secundaria de los datos en uno o más sitios remotos. La cantidad de copias y la cantidad de espacio que ocupa la copia secundaria se determinan en función de la cantidad de sitios configurados en el grupo de replicación, la forma en que se escriben los datos en los sitios, así como si **replicate to all sites** está habilitado o no.

Cada sitio es responsable de la protección de datos local, lo que significa que las copias locales y secundarias protegerán individualmente los datos mediante la codificación de eliminación o el espejeado triple. Los esquemas de codificación de eliminación en cada sitio no tienen que ser los mismos, lo que significa que un sitio puede usar el esquema de codificación de eliminación predeterminado de 12+4 y el otro sitio puede utilizar el esquema de codificación de eliminación de almacenamiento inactivo de 10+2.

Los datos replicados se cifran (AES256) y se comprimen antes de enviarse al otro sitio a través de HTTP.

Para mantener la coherencia entre los sitios, debe haber una autoridad responsable de mantener la versión más reciente de los metadatos. La autoridad se define en el nivel del sitio y determina la propiedad de espacios de nombres, depósitos y objetos. La información de propiedad se almacena primero en el sitio propietario, pero también se replica a los otros sitios como parte de los metadatos de ECS.

- Versión autorizada: la versión autorizada siempre es el propietario y se utiliza para proporcionar una coherencia sólida.
- Versión replicada: es posible que las versiones replicadas no sean las más recientes, pero se utilizan durante las operaciones de falla, incluidas las siguientes:
  - Si **access during outage** está habilitado (coherencia eventual).
  - Y durante las operaciones permanentes de conmutación por error del sitio.

Hay versiones autorizadas de los propietarios de objetos y depósitos.

Propietario del espacio de nombres:

- El sitio que crea el espacio de nombres es el propietario del espacio de nombres.
- Es responsable de mantener la versión autorizada de la lista de depósitos.

Propietario del depósito:

- El sitio que crea el depósito es el propietario del depósito.
- Es responsable de mantener la versión autorizada de:
  - La lista de depósitos que incluye la versión más actualizada de los objetos que se encuentran en un depósito.
  - La lista de propiedad de objetos para los objetos dentro de su depósito

Propietario del objeto:

- Inicialmente, el sitio que creó por primera vez el objeto es el propietario del objeto. Esto puede cambiar; consulte la sección “Acceso durante una interrupción” para obtener más información.
- Es responsable de mantener la versión autorizada de los metadatos del objeto.

3.1 Tablas del administrador de fragmentos

Las ubicaciones de los fragmentos se conservan en la tabla del administrador de fragmentos que se almacena de manera independiente en todos los sitios. La ubicación en la que se creó originalmente un fragmento se conoce como el sitio principal; la ubicación a la que se replica se conoce como el sitio secundario.

Cuando se crea un fragmento, se determinan los sitios principal y secundario, y el sitio principal transmite la información del sitio de los fragmentos a los otros nodos del grupo de replicación.

Además, dado que cada sitio mantiene su propia tabla de administrador de fragmentos, también incluirá información sobre el tipo de propiedad para cada fragmento. Los tipos de propiedad incluyen:

- **Local:** en el sitio en el que se creó el fragmento
- **Copy:** en el sitio al que se replicó el fragmento
- **Remote:** en los sitios en los que ni el fragmento ni su réplica se almacenan localmente
- **Parity:** en fragmentos que contienen el resultado de una operación XOR de otros fragmentos (consulte la sección XOR a continuación para obtener más detalles)
- **Encoded:** en fragmentos cuyos datos se reemplazaron localmente por datos XOR (consulte la sección XOR a continuación para obtener más detalles).

Tabla 9 a Tabla 11 muestran ejemplos de partes de listas de tablas del administrador de fragmentos de cada uno de los tres sitios.

Tabla 9 Ejemplo de tabla de administrador de fragmentos del sitio 1

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 2	Local
C2	Sitio 2	Sitio 3	Remoto
C3	Sitio 1	Sitio 3	Local
C4	Sitio 2	Sitio 1	Copiar

Tabla 10 Ejemplo de tabla de administrador de fragmentos del sitio 2

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 2	Copiar
C2	Sitio 2	Sitio 3	Local
C3	Sitio 1	Sitio 3	Remoto
C4	Sitio 2	Sitio 1	Local

Tabla 11 Ejemplo de tabla de administrador de fragmentos del sitio 3

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 2	Remoto
C2	Sitio 2	Sitio 3	Copiar
C3	Sitio 1	Sitio 3	Copiar
C4	Sitio 2	Sitio 1	Remoto

3.2 Codificación XOR

A fin de maximizar la eficiencia del almacenamiento de los datos configurados con un grupo de replicación que contiene tres o más sitios, ECS utiliza la codificación XOR. A medida que aumenta la cantidad de sitios en un grupo de replicación, el algoritmo de ECS es más eficiente en la reducción de la sobrecarga.

La codificación XOR se realiza en cada sitio. Escanea su tabla de administrador de fragmentos y, cuando encuentra fragmentos de tipo COPY que provienen de cada uno de los otros sitios de su grupo de replicación, puede realizar la codificación XOR en estos fragmentos. Por ejemplo, en la Tabla 12, se muestra el sitio 3 de una configuración de tres sitios que tiene fragmentos **C2** y **C3** que son de tipo COPY, cada uno con un sitio principal diferente. Esto permite que el sitio 3 los XOR se reúnan y almacenen el resultado. El resultado será un nuevo fragmento, **C5**, que es un XOR de **C2** y **C3** (matemáticamente C2 ⊕ C3) y tendrá un tipo enumerado como **Parity**, sin un sitio secundario. Los ID de fragmentos de los fragmentos de paridad no se transmiten a otros sitios.

Tabla 12muestra un ejemplo de una tabla de administrador de fragmentos en el sitio 3 mientras ejecuta XOR de fragmentos **C2** y **C3** juntos en el fragmento **C5**.

Tabla 12 Tabla del administrador de fragmentos del sitio 3 durante la operación XOR

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 2	Remoto
C2	Sitio 2	Sitio 3	Copiar
C3	Sitio 1	Sitio 3	Copiar
C4	Sitio 2	Sitio 1	Remoto
C5	Sitio 3		Paridad (C2 y C3)

Una vez que se completa XOR, la copia de datos para **C2** y **C3** se elimina, lo que libera espacio en el disco y el tipo de tabla del administrador de fragmentos para estos fragmentos cambia al tipo Codificado. La operación XOR es únicamente una operación de sitio secundario; el sitio principal no reconoce que sus fragmentos se codificaron. Después de que se complete la codificación XOR y se elimine la copia de datos para **C2** y **C3**, la tabla del administrador de fragmentos del sitio 3 se enumeraría como se muestra en Tabla 13.

Tabla 13 Tabla del administrador de fragmentos del sitio 3 después de completar la codificación XOR de C2 y C3

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 2	Remoto
C2	Sitio 2	Sitio 3	Codificado
C3	Sitio 1	Sitio 3	Codificado
C4	Sitio 2	Sitio 1	Remoto
C5	Sitio 3		Paridad (C2 y C3)

El sitio que contiene la copia principal atenderá las solicitudes de datos en un fragmento codificado. Para obtener más detalles sobre la eficiencia del almacenamiento, consulte la Documentación técnica Descripción general y arquitectura de ECS.

3.3 Replicar a todos los sitios

Replicar a todos los sitios es una opción de grupo de replicación que se utiliza cuando tiene tres o más sitios y desea que todos los fragmentos se repliquen en todos los VDC/sitios configurados en el grupo de replicación. Esto también impide que se ejecuten las operaciones de XOR. Cuando la opción **replicate to all sites** está:

- **Habilitada:** la cantidad de copias de datos escritos es igual a la cantidad de sitios en el grupo de replicación. Por ejemplo, si tiene cuatro sitios en el grupo de replicación, tendrá una copia principal más tres copias secundarias; uno en cada sitio.
- **Deshabilitada:** la cantidad de copias de datos escritos es dos. Tiene la copia principal más una copia replicada en un sitio remoto, independientemente de la cantidad total de sitios.

Nota: replicate to all sites no se puede habilitar en el grupo de replicación configurado como geopasivo.

Esta configuración no tiene ningún impacto en los grupos de replicación que contienen solo dos sitios, ya que, en esencia, ya replica todos los datos a ambos sitios. El administrador puede elegir qué depósitos utilizan este grupo de replicación.

La habilitación de replicate to all sites tiene el siguiente impacto:

- Puede mejorar el rendimiento de lectura porque, después de que se completa la replicación, las lecturas posteriores se realizan localmente.
- Elimina el impacto en el rendimiento causado por la decodificación de XOR
- Aumenta la durabilidad de los datos.
- Disminuye la eficiencia de la utilización del almacenamiento.
- Aumenta la utilización de WAN para la replicación geográfica; el aumento es proporcional a la cantidad de VDC/sitios en el grupo de replicación.
- Disminuye la utilización de WAN para las lecturas de datos replicados.

Por estos motivos, la activación de esta opción solo se recomienda en depósitos específicos en entornos que cumplan con los siguientes criterios:

- Cuya carga de trabajo es de lectura intensiva para los mismos datos de sitios geográficamente dispersos.
- Cuya infraestructura tiene suficiente ancho de banda WAN entre sitios en el grupo de replicación para permitir un mayor tráfico de replicación geográfica.
- Quién valora el rendimiento de lectura sobre el costo de la eficiencia de la utilización del almacenamiento.

3.4 Flujo de datos de escritura en un entorno replicado geográficamente

Los fragmentos contienen 128 MB de datos que constan de uno o más objetos de depósitos que comparten la misma configuración de grupo de replicación. El propietario de la partición del fragmento realiza la replicación de manera asíncrona e inicia en el nivel del fragmento. Si el fragmento está configurado con datos de replicación geográfica se agregará a una línea de espera de replicación a medida que se escribe en el fragmento del sitio principal, no espera que el fragmento se sella. Hay subprocesos de I/O de trabajador que procesan continuamente la línea de espera.

La operación de escritura se produce primero localmente, incluida la adición de protección de datos y, a continuación, se replica y protege en el sitio remoto. Figura 8 muestra un ejemplo del proceso de escritura para un objeto de 128 MB en un depósito replicado geográficamente.

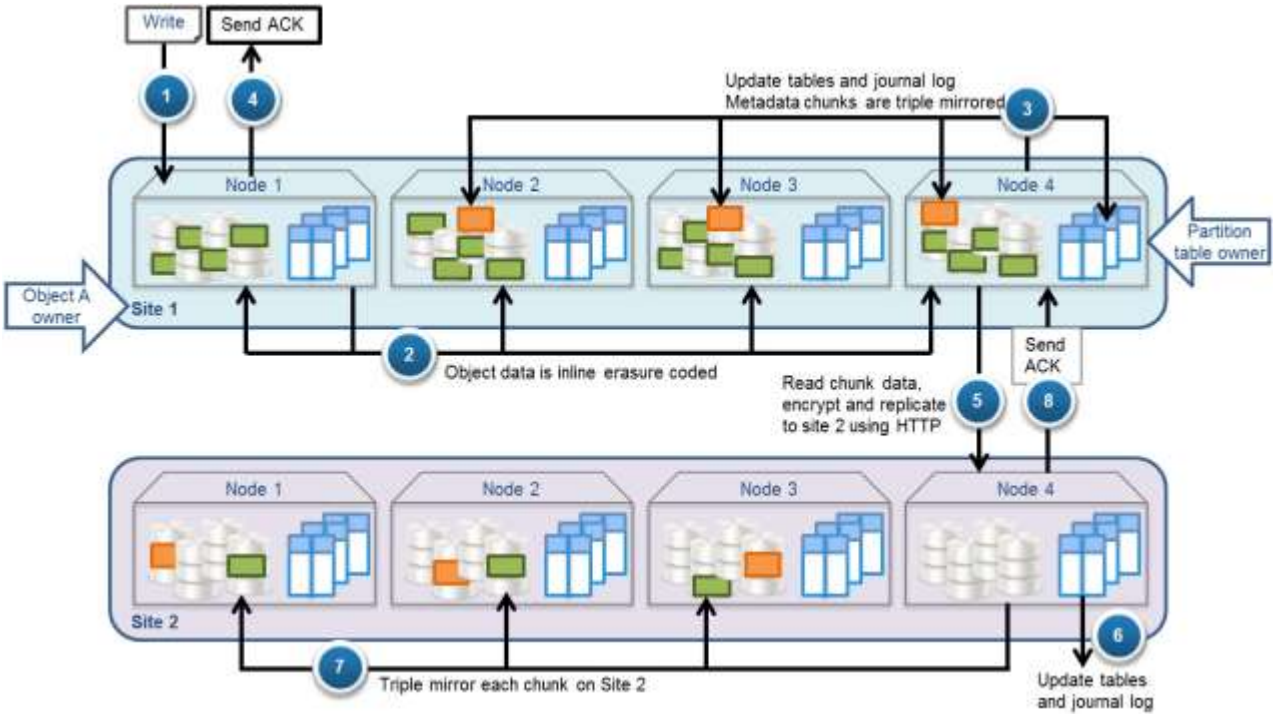

Figura 8 Flujo de trabajo de escritura de datos para un objeto de 128 MB en un depósito replicado geográficamente

1. La solicitud de escritura para el objeto A se envía a un nodo, en este ejemplo, el nodo 1 del sitio 1. El sitio 1 se convierte en el propietario del objeto A.
2. Los datos están codificados para eliminación en línea y escritos en un fragmento del sitio 1.
3. Los propietarios de particiones de tabla, en este ejemplo nodo 4, actualizan las tablas correspondientes (p. ej., tablas de lista de fragmentos, objetos y depósitos) y escriba las transacciones en sus registros de registro. Estos metadatos se escriben en un fragmento de metadatos que se duplica triplemente en el sitio 1.

4. El reconocimiento de la escritura correcta se envía al cliente.
5. Para cada fragmento, el propietario de la tabla de partición del fragmento, en este ejemplo, nodo 4:
   a. agrega los datos dentro del fragmento a la línea de espera de replicación después de que se escriben localmente, no espera a que el fragmento se sella.
   b. lee los fragmentos de datos del fragmento (los fragmentos de paridad solo se leen si es necesario para volver a crear un fragmento de datos faltante).
   c. cifra y replica los datos en el sitio 2 a través de HTTP.
6. Los propietarios de particiones de tablas para los fragmentos replicados, en este sitio de ejemplo, nodo 2, nodo 4, actualizan las tablas correspondientes y escriben las transacciones en sus registros de registro que son espejeados triples.
7. Cada fragmento replicado se escribe inicialmente en el segundo sitio mediante el espejeado triple.
8. El reconocimiento se envía de vuelta al propietario de la tabla de partición de fragmentos del sitio principal.

Nota: los datos escritos en el sitio replicado se codificarán después de un retraso, lo que permitirá que otros procesos, como las operaciones XOR, se completen primero.

3.5 Flujo de datos de lectura en un entorno replicado geográficamente

Dado que ECS replica datos de manera asíncrona a varios VDC dentro de un grupo de replicación, requiere un método para garantizar la coherencia de los datos entre sitios/VDC. ECS garantiza una coherencia sólida mediante la obtención de la copia más reciente de los metadatos desde el sitio que es el propietario del objeto. Si el sitio solicitante contiene una copia (tipo de fragmento = local o copia) del objeto que utilizará para atender la solicitud de lectura, de lo contrario, buscará los datos del propietario del objeto. Un ejemplo que muestra el flujo de datos de lectura se ve en Figura 9, que representa una solicitud de lectura para el objeto A desde un sitio que no es el sitio propietario del objeto.

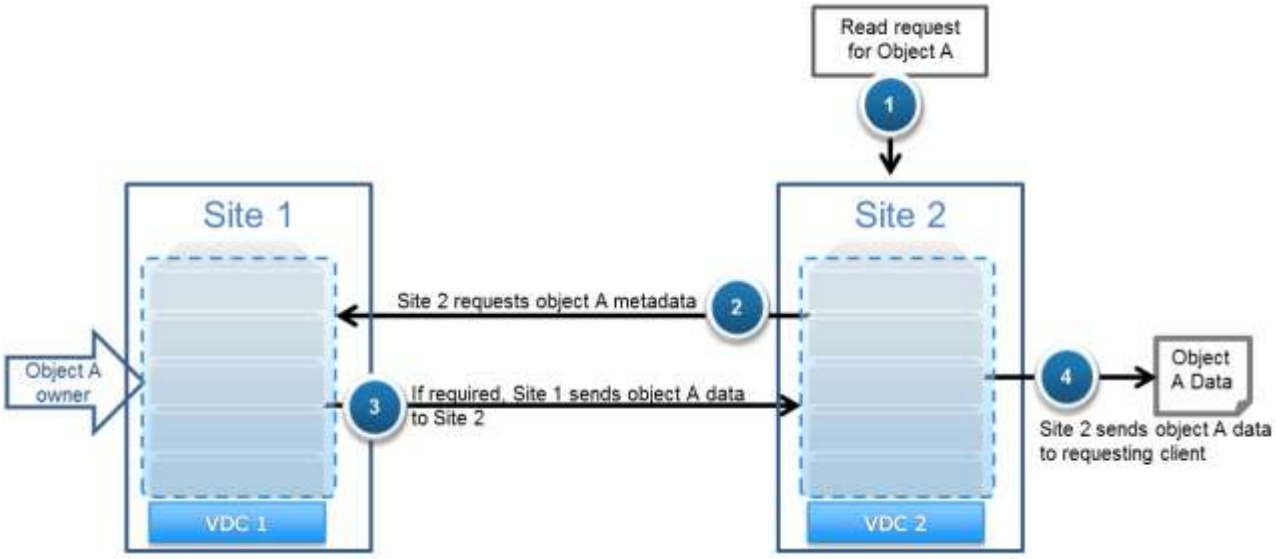

Figura 9 Flujo de trabajo de lectura de datos para un objeto replicado geográficamente que pertenece a otro sitio

En este ejemplo, el flujo de la operación de lectura se muestra de la siguiente manera:

1. El sitio 2 recibe una solicitud de lectura para el objeto A que es propiedad del sitio 1.
2. El sitio 2 se comunica con el depósito y el propietario del objeto, en este ejemplo, el sitio 1, para obtener la versión más reciente de los metadatos.

Propiedad del objeto:

- Si **access during outage** está deshabilitado, comprobará su información local para determinar si es el propietario del objeto. Si no es así, se pondrá en contacto con el propietario del depósito para determinar quién es el propietario del objeto.
- Si **access during outage** está habilitado en el depósito, el sitio solicitante comprobará con el propietario del depósito para ver quién es el propietario actual del objeto.

3. Si el sitio 2 no contiene una copia del objeto (tipo de fragmento = local o copia), el sitio 1 enviará los datos del objeto A al sitio 2.
4. El sitio 2 envía los datos del objeto A al cliente solicitante.

3.6 Actualizar el flujo de datos en un entorno replicado geográficamente

ECS está diseñado para permitir la actualización activa-activa de datos desde los nodos dentro del grupo de replicación de depósitos asociado. Puede lograr esto al requerir que los sitios que no son propietarios de objetos envíen de manera síncrona información sobre una actualización de objetos al sitio del propietario principal y esperen a que se confirme antes de que puedan enviar la confirmación de vuelta al cliente. Los datos relacionados con el objeto actualizado se replican como parte de las actividades normales de replicación asíncrona de fragmentos. Si los datos aún no se replican en el sitio propietario y recibe una solicitud de lectura para los datos, los solicitará desde el sitio remoto. Figura 10 muestra un ejemplo que representa una solicitud de actualización para el objeto A desde un sitio que no es el sitio propietario del objeto.

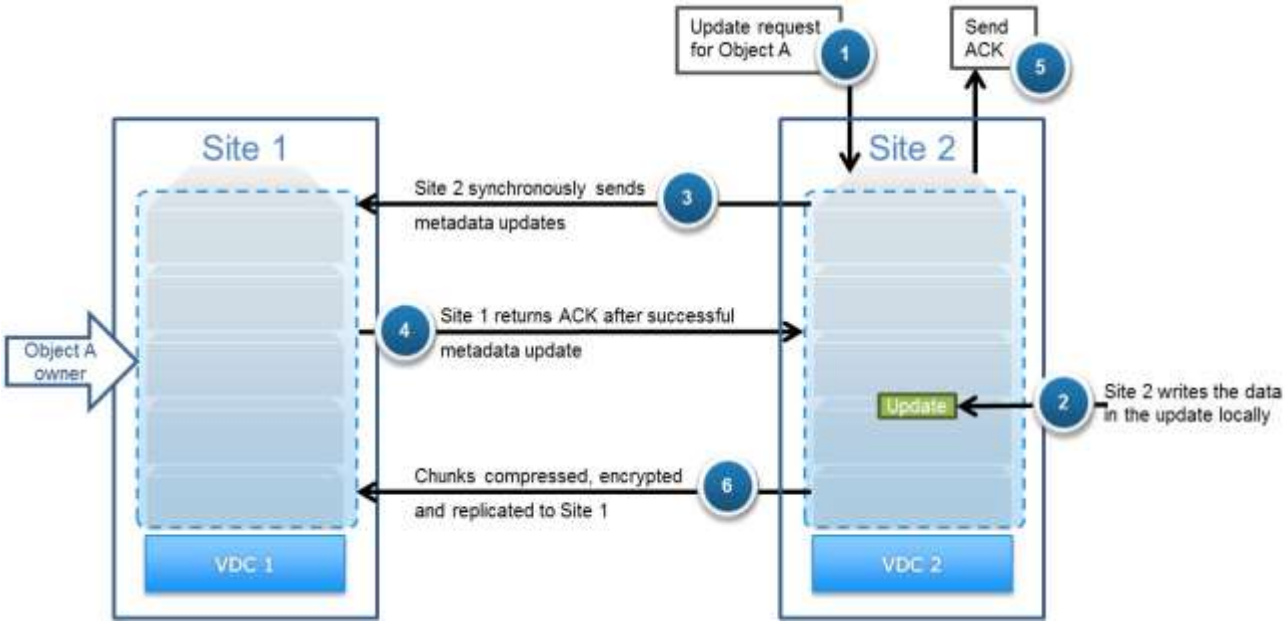

Figura 10 Actualizar el flujo de trabajo de objetos para un objeto replicado geográficamente que pertenece a otro sitio

En este ejemplo, el flujo de la operación de actualización se muestra de la siguiente manera:

1. El sitio 2 recibe una solicitud de actualización para el objeto A que es propiedad del sitio 1.
2. El sitio 2 escribe los datos en la actualización localmente.
3. El sitio 2 envía actualizaciones de metadatos de manera síncrona al propietario del objeto, en este ejemplo, el sitio 1.

Nota: si **access during outage** está habilitado en el depósito o el sitio 2 no es el propietario del objeto, primero se comunica con el propietario del depósito para determinar quién es el propietario actual del objeto.

4. El sitio 1 envía confirmación al sitio 2 de que la actualización de metadatos se realizó correctamente.
5. El sitio 2 envía la confirmación al cliente para solicitar que la actualización se haya realizado correctamente.
6. Los fragmentos se agregan a la línea de espera de replicación, se cifran y se replican de manera asíncrona al sitio 1 como de costumbre.

En condiciones normales, el propietario del objeto no cambia después de realizar una actualización, independientemente del sitio desde el que se originó la actualización. En este ejemplo, el propietario del objeto permanece en el sitio 1, incluso después de la actualización correcta que se originó desde el sitio 2. La única excepción es durante una interrupción temporal del sitio si **access during outage** está habilitado; consulte la sección 4.1.2 para obtener más detalles.

4 Disponibilidad de múltiples sitios

ECS proporciona una coherencia sólida, lo que requiere que las solicitudes de I/O se verifiquen con el propietario antes de responder. Por este motivo, si no se puede acceder a un sitio a algunos depósitos y objetos, es posible que se interrumpa temporalmente.

Las interrupciones del sitio pueden ocurrir por diferentes duraciones y por muchas razones diferentes, como las siguientes:

- Temporalmente, por ejemplo, por la pérdida de conectividad de red entre sitios federados o por la falla de un sitio completo, como una interrupción de la alimentación del edificio.
- Permanentemente, como por un desastre natural.

Para detectar interrupciones temporales del sitio, los sitios federados establecen un latido entre los sitios. Si el latido se pierde entre sitios durante un período de tiempo sostenido, el valor predeterminado es 15 minutos:

- En una configuración de dos sitios, cada uno marcará el otro como fallido.
- En una configuración de tres o más sitios, un sitio solo se marcará como fallido si ambos:
  - La mayoría de los sitios pierden latidos durante el período sostenido al mismo sitio de ECS
  - Y si todos los sitios restantes están marcados actualmente como en línea

Por ejemplo, en una configuración de tres sitios, si los sitios 2 y 3 pierden conectividad de red con el sitio 1 durante un período sostenido, ECS marcará el sitio 1 como fallido temporalmente.

Cuando falla un sitio federado, la disponibilidad del sistema se puede mantener mediante la dirección del acceso a otros sistemas federados. Durante la falla en todo el sitio, los datos replicados geográficamente que pertenecen al sitio no disponible dejarán de estar disponibles temporalmente. La duración durante la cual los datos permanecen no disponibles está determinada por lo siguiente:

- Ya sea que **access during outage** esté habilitado o no
- Cuánto tiempo permanece la interrupción temporal del sitio
- El tiempo que tarda una conmutación por error permanente en el sitio para completar las operaciones de recuperación

La falla del sitio puede ser temporal o permanente. Una interrupción temporal del sitio significa que el sitio puede volver a estar en línea y, por lo general, se debe a interrupciones en la alimentación o a la pérdida de redes entre sitios. Una interrupción permanente del sitio es cuando todo el sistema es irrecuperable, por ejemplo, desde un incendio de laboratorio. Solo un administrador puede determinar si una interrupción del sitio es permanente y requiere operaciones de recuperación.

4.1 Interrupción temporal del sitio (TSO)

Las interrupciones temporales del sitio se producen cuando no se puede acceder temporalmente a un sitio a otros sitios de un grupo de replicación. ECS permite a los administradores dos opciones de configuración que afectan la manera en que se puede acceder a los objetos durante una interrupción temporal del sitio.

- Deshabilite la opción **access during outage** (ADO) que conserva una coherencia sólida mediante lo siguiente:
  - Continuar permitiendo el acceso a los datos que pertenecen a sitios a los que se puede acceder.
  - Impide el acceso a los datos que pertenecen a un sitio inaccesible.

- Habilite la opción **access during outage** que permite el acceso de lectura y, opcionalmente, de escritura a todos los datos replicados geográficamente, incluido el que pertenece al sitio marcado como fallido. Durante una TSO con la opción **access during outage** habilitada, los datos en el depósito cambian temporalmente a coherencia eventual; una vez que todos los sitios estén en línea nuevamente, volverá a tener una coherencia sólida.

El valor predeterminado es que **access during outage** esté deshabilitado.

La opción **access during outage** puede configurarse en el nivel de depósito; lo que significa que puede habilitar esta opción para algunos depósitos y no para otros. Esta opción de depósito se puede cambiar en cualquier momento, siempre y cuando todos los sitios estén en línea, no se puede cambiar durante una falla del sitio.

Durante una interrupción temporal del sitio:

- Los depósitos, espacios de nombres, usuarios de objetos, proveedores de autenticación, grupos de replicación y mapeos de usuarios y grupos de NFS no se pueden crear, eliminar ni actualizar desde ningún sitio (los grupos de replicación se pueden eliminar de un VDC durante una conmutación por error permanente del sitio).
- No puede enumerar depósitos para un espacio de nombres cuando el sitio propietario del espacio de nombres no está accesible.
- Los sistemas de archivos en depósitos HDFS/NFS que son propiedad del sitio no disponible son de solo lectura.
- Cuando copia un objeto de un depósito que pertenece al sitio no disponible, la copia es una copia completa del objeto de origen. Esto significa que los datos del mismo objeto se almacenan más de una vez. En circunstancias normales que no sean de TSO, la copia del objeto consta de los índices de datos del objeto, no es un duplicado completo de los datos del objeto.
- Los usuarios de OpenStack Swift no pueden iniciar sesión en OpenStack durante una TSO debido a que ECS no puede autenticar usuarios de Swift durante la TSO. Después de la TSO, los usuarios Swift deben volver a autenticarse.

4.1.1 Comportamiento predeterminado de TSO

Dado que ECS proporciona una coherencia sólida, las solicitudes de I/O requieren una comprobación con el propietario antes de responder. Si no se puede acceder a un sitio desde otros sitios dentro de un grupo de replicación, es posible que se interrumpa el acceso a depósitos y objetos.

Tabla 14 muestra qué acceso se requiere para que una operación se realice correctamente.

Tabla 14 Requisitos de acceso

Operación	Requisitos para el éxito
Crear objeto	Requiere que se pueda acceder al propietario del depósito
Listado de objetos	Requiere que el propietario del depósito y el nodo solicitante puedan acceder a todos los objetos del depósito
Leer objeto Actualización de un objeto	Requiere que el solicitante sea: • El propietario del objeto y el propietario del depósito (la propiedad del depósito solo es necesaria si **access during outage** está habilitado en el depósito que contiene el objeto) • O que tanto el propietario del objeto como el propietario del depósito sean accesibles para el nodo solicitante

- Una operación de creación de objetos incluye la actualización de la lista de depósitos con el nuevo nombre de objeto. Esto requiere acceso al propietario del depósito y, por lo tanto, fallará si el sitio solicitante no tiene acceso al propietario del depósito.
- La enumeración de objetos en un depósito requiere la enumeración de información del propietario del depósito y la información del cabezal para cada objeto en el depósito. Por lo tanto, las siguientes solicitudes de lista de depósitos fallarán si **access during outage** está deshabilitado:
  - Solicitudes para enumerar depósitos que pertenecen a un sitio que no es accesible para el solicitante.
  - Depósito que contiene objetos que pertenecen a un sitio al que el solicitante no puede acceder.
- El objeto de lectura requiere la lectura inicial de los metadatos del objeto desde el propietario del objeto.
  - Si el sitio solicitante es el propietario del objeto y **access during outage** está deshabilitado, la solicitud se realizará correctamente.
  - Si el sitio solicitante es el objeto y el propietario del depósito, la solicitud se realizará correctamente.
  - Si el propietario del objeto no es local, el sitio debe consultar con el propietario del depósito para encontrar al propietario del objeto. Si el propietario del objeto o los sitios propietarios del depósito no están disponibles para el solicitante, la operación de lectura fallará.
- Las actualizaciones de objetos requieren la actualización correcta de los metadatos del objeto en el propietario del objeto.
  - Si el sitio solicitante es el propietario del objeto y **access during outage** está deshabilitado, la solicitud se realizará correctamente.
  - Si el sitio solicitante es el objeto y el propietario del depósito, la solicitud se realizará correctamente.
  - Si el propietario del objeto no es local, el sitio debe consultar con el propietario del depósito para encontrar al propietario del objeto. Si el propietario del objeto o los sitios propietarios del depósito no están disponibles para el solicitante, la operación de lectura fallará.

Si se observa un ejemplo de tres sitios, el diseño de objetos y depósitos se muestra en la Figura 11.

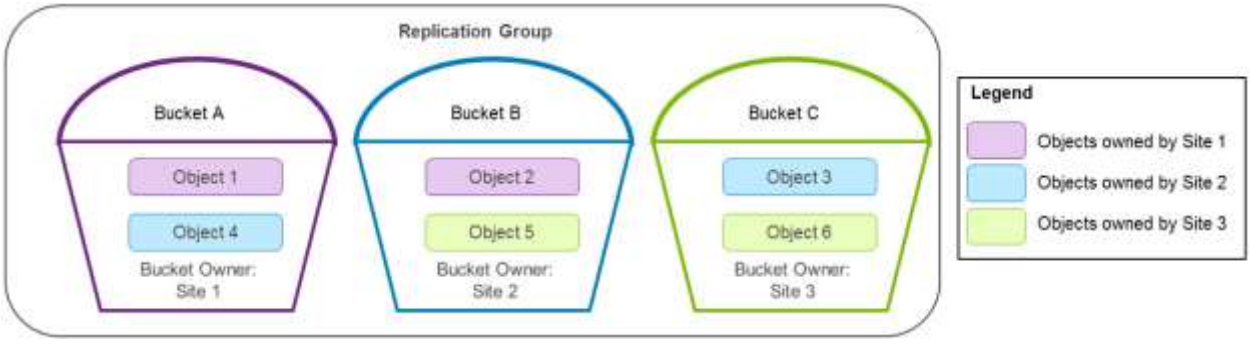

Figura 11 Ejemplo de propiedad de objetos y depósitos

Tabla 15 enumera las operaciones que se realizarán correctamente o fallarán en el ejemplo de configuración de tres sitios que se muestra en Figura 11 si el sitio 1 no está accesible para los otros sitios del grupo de replicación. Para simplificar la interpretación de la tabla, el sitio inaccesible se muestra como fallido y los otros dos se enumeran como en línea.

Tabla 15 Operaciones que se realizan correctamente o fallan si no se puede acceder al sitio 1 para otros sitios

Operación	Depósito/objeto	Solicitud enviada a		
		Sitio 1 (fallido)	**Sitio 2 (en línea)**	**Sitio 3 (en línea)**
Crear objetos en	Depósito A	Satisfactorio Depósito de propiedad local	No aprobada No se puede acceder al propietario del depósito	No aprobada No se puede acceder al propietario del depósito
	Depósito B	No aprobada No se puede acceder al propietario del depósito	Satisfactorio Depósito de propiedad local	Satisfactorio Depósito que pertenece al sitio en línea
	Depósito C	No aprobada No se puede acceder al propietario del depósito	Satisfactorio Depósito que pertenece al sitio en línea	Satisfactorio Depósito de propiedad local
Listar objetos en	Depósito A	No aprobada A pesar de que el depósito es propiedad local, contiene un objeto que pertenece a un sitio al que no puede acceder	No aprobada No se puede acceder al propietario del depósito	No aprobada No se puede acceder al propietario del depósito
	Depósito B	No aprobada No se puede acceder al propietario del depósito	No aprobada A pesar de que el depósito es propiedad local, contiene un objeto que es propiedad del sitio fallido	No aprobada A pesar de que el propietario del depósito está en línea, el depósito contiene un objeto que es propiedad del sitio fallido
	Depósito C	No aprobada No se puede acceder al propietario del depósito	Satisfactorio El propietario del depósito es un sitio en línea y todos los objetos son de sitios en línea	Satisfactorio El depósito es propiedad local y todos los objetos son de sitios en línea
Leer o actualizar objeto	Objeto 1	Satisfactorio Objeto de propiedad local	No aprobada No se puede acceder al propietario del objeto	No aprobada No se puede acceder al propietario del objeto
	Objeto 2	Satisfactorio Objeto de propiedad local	No aprobada No se puede acceder al propietario del objeto	No aprobada No se puede acceder al propietario del objeto
	Objeto 3	No aprobada No se puede acceder al propietario del objeto	Satisfactorio Objeto de propiedad local	Satisfactorio El objeto no es de propiedad local, por lo que obtiene el propietario del objeto del propietario del depósito que está en línea

Operación	Depósito/objeto	Solicitud enviada a		
		Sitio 1 (fallido)	**Sitio 2 (en línea)**	**Sitio 3 (en línea)**
	Objeto 4	No aprobada No se puede acceder al propietario del objeto	Satisfactorio Objeto de propiedad local	No aprobada El objeto no es de propiedad local, por lo que requiere acceder al propietario del depósito, que es el sitio fallido
	Objeto 5	No aprobada No se puede acceder al propietario del objeto	Satisfactorio El objeto no es de propiedad local, por lo que obtiene el propietario del objeto del propietario del depósito, que es	Satisfactorio Objeto de propiedad local
	Objeto 6	No aprobada No se puede acceder al propietario del objeto	Satisfactorio El objeto no es de propiedad local, por lo que obtiene el propietario del objeto del propietario del depósito que está en línea	Satisfactorio Objeto de propiedad local

4.1.2 Comportamiento de TSO con acceso durante una interrupción habilitado

Cuando un sitio es inaccesible por primera vez para otros sitios dentro de un grupo de replicación, el comportamiento es el que se detalla en la sección de comportamiento de TSO predeterminada. Después de que se pierde el latido entre los sitios durante un período de tiempo sostenido, el valor predeterminado es 15 minutos en que ECS marca un sitio como fallido. Habilitar **access during outage** (ADO) en un depósito cambia el comportamiento de la TSO después de que un sitio se marca como fallido, lo que permite que los objetos en ese depósito utilicen la coherencia eventual. Esto significa que, después de que un sitio se marca como fallido temporalmente, cualquier depósito que tenga habilitada la opción **access during outage** admitirá lecturas y, opcionalmente, escrituras desde un sitio no propietario. Para ello, permite el uso de los metadatos replicados cuando la copia autorizada en el sitio propietario no está disponible. Puede cambiar la opción de depósito **access during outage** en cualquier momento, excepto durante una falla del sitio.

El beneficio de habilitar **access during outage** es permitir el acceso a los datos después de que un sitio se marca como fallido; la desventaja es que los datos devueltos pueden estar desactualizados.

Desde la versión 3.1, se agregó una opción de depósito adicional para **read-only access during outage**, lo que garantiza que la propiedad de los objetos nunca cambie y elimina la posibilidad de que se produzcan conflictos causados por las actualizaciones de objetos en los sitios fallidos y en línea durante un TSO. La desventaja de **read-only access during outage** es que, después de que un sitio se marca como fallido, no se pueden crear objetos nuevos ni se pueden actualizar objetos existentes en el depósito hasta que todos los sitios estén nuevamente en línea. La opción **read-only access during outage** solo está disponible durante la creación de depósitos; no se puede modificar posteriormente.

Como se mencionó anteriormente, un sitio se marca como fallido cuando el latido se pierde entre los sitios durante un período prolongado, el valor predeterminado es de 15 minutos. Por lo tanto, si el latido se pierde durante un período prolongado:

- En una configuración de dos sitios, cada uno se considerará en línea y marcará el otro como fallido.
- En una configuración con tres o más sitios, un sitio solo se marcará como fallido si ambos:
  - La mayoría de los sitios pierden latidos durante el período sostenido al mismo sitio de ECS
  - Y si todos los sitios restantes están marcados actualmente como en línea

Es posible que los clientes y las aplicaciones aún puedan acceder a un sitio fallido, como cuando la red interna de una empresa pierde conectividad con un solo sitio, pero las redes de extranet siguen funcionando. Por ejemplo, en una configuración de cinco sitios si los sitios del 2 al 5 pierden la conectividad de red con el sitio 1 durante un período sostenido, ECS marcará el sitio 1 como fallido temporalmente. Si el sitio 1 sigue siendo accesible para los clientes y las aplicaciones, podrá realizar solicitudes de servicio para depósitos y objetos de propiedad local, ya que no se requieren búsquedas en otros sitios. Sin embargo, las solicitudes al sitio 1 para depósitos y objetos que no son de propiedad fallarán. Tabla 16 muestra qué acceso se requiere después de que un sitio se marca como fallido para que una operación se realice correctamente si **access during outage** está configurado en habilitado.

Tabla 16 Operaciones correctas después de que un sitio se marca como fallido con access during outage habilitado

Operación	**Solicitud enviada al sitio fallido (en una federación que contiene tres o más sitios)**	**Solicitud enviada a un sitio en línea, que incluye:** • **Cualquier sitio en línea de una federación que contenga tres o más sitios** • **O cualquiera de los sitios de una federación que contiene solo dos sitios**
Crear objeto	Correcto para los depósitos de propiedad local, a menos que **read-only access during outage** esté habilitado en el depósito. Falla para depósitos de propiedad remota	Correcto, a menos que **read-only access during outage** esté habilitado en el depósito.
Listado de objetos	Solo enumera los objetos de sus depósitos de propiedad local si todos los objetos también son de propiedad local	Successful No incluirá objetos que pertenecen a un sitio fallido que no haya terminado de replicarse
Leer objeto	Operación correcta para objetos de propiedad local en depósitos de propiedad local (es posible que no sea la versión más reciente) Falla para objetos de propiedad remota	Successful Si el objeto es propiedad del sitio fallido, requiere que el objeto original haya terminado la replicación antes de que se produjera la falla
Actualización de un objeto	Correcto para objetos de propiedad local en depósitos de propiedad local, a menos que **read-only access during outage** esté habilitado en el depósito. Falla para objetos de propiedad remota	Correcto, a menos que **read-only access during outage** esté habilitado en el depósito. Adquiere propiedad del objeto

- Crear objeto

  Una vez que un sitio se marca como fallido, la creación de objetos no se realizará correctamente si **read-only access during outage** está habilitado en el depósito. Si está deshabilitado:

  - En una federación que contiene tres o más sitios, el sitio marcado como fallido, si los clientes o las aplicaciones pueden acceder a él, puede crear objetos solo en sus depósitos de propiedad local. Solo se puede acceder a estos nuevos objetos desde este sitio; otros sitios no conocerán estos objetos hasta que el sitio fallido vuelva a estar en línea y obtengan acceso al propietario del depósito.

  - Los sitios en línea pueden crear objetos en cualquier depósito, incluidos los depósitos que pertenecen al sitio marcados como fallidos. La creación de un objeto requiere la actualización de la lista de depósitos con el nuevo nombre de objeto. Si el propietario del depósito está inactivo, creará un historial de objetos que insertará el objeto en la tabla de lista de depósitos durante las operaciones de recuperación o reincorporación del propietario del depósito.

- Listar objetos

  - En una federación que contiene tres o más sitios, el sitio marcado como fallido requiere la propiedad local del depósito y de todos los objetos dentro del depósito para enumerar correctamente los objetos en un depósito. La lista del sitio fallido no incluirá objetos creados de manera remota mientras se marca como fallido temporalmente.
  - Los sitios en línea pueden enumerar objetos en cualquier depósito, incluido un depósito que pertenece al sitio marcado como fallido. Enumerará la versión más reciente de la lista de depósitos que tiene, puede estar ligeramente desactualizada.

- Leer objeto

  - En una federación que contiene tres o más sitios, el sitio fallido, si es accesible para clientes o aplicaciones, puede leer objetos de propiedad local en depósitos de propiedad local.
  - La solicitud de lectura a un sitio fallido primero necesita acceso al propietario del depósito para validar al propietario del objeto actual. Si puede acceder al propietario del depósito y el propietario actual del objeto es local, la solicitud de lectura se realizará correctamente. Si no se puede acceder al propietario del depósito o al propietario actual del objeto, la solicitud de lectura fallará.
  - Los sitios en línea pueden leer cualquier objeto, incluidos los que pertenecen al sitio marcado como fallido, siempre y cuando el objeto original haya completado la replicación. Comprobará el historial de objetos y responderá con la versión más reciente del objeto que está disponible. Si un objeto se actualizó posteriormente en el sitio marcado como fallido y la replicación geográfica de la versión actualizada no se completó, la versión anterior se utilizará para atender la solicitud de lectura.

Nota: las solicitudes de lectura enviadas a sitios en línea donde el propietario del depósito es el sitio fallido utilizarán su información de lista de depósitos locales y el historial de objetos para determinar el propietario del objeto.

- Actualización de un objeto

  - Una vez que un sitio se marca como fallido, la actualización de objetos no se realizará correctamente si **read-only access during outage** está habilitado en el depósito. Si está deshabilitado, en una federación que contiene tres o más sitios, el sitio fallido, si los clientes o las aplicaciones pueden acceder a él, puede actualizar solo objetos de propiedad local en depósitos de propiedad local.
  - La solicitud de actualización primero necesita acceso al propietario del depósito para validar la propiedad del objeto actual. Si puede acceder al propietario del depósito y el propietario actual del objeto es local, la solicitud de actualización se realizará correctamente. Si no se puede acceder al propietario del depósito o al propietario actual del objeto, la solicitud de actualización fallará.
  - Después de completar las operaciones de reincorporación, esta actualización no se incluirá en las operaciones de lectura si el sitio remoto también actualizó el mismo objeto durante la misma TSO.

Un sitio en línea puede actualizar los objetos que pertenecen a sitios en línea y sitios fallidos. Si se envía una solicitud de actualización de objeto a un sitio en línea para un objeto que pertenece al sitio marcado como fallido, actualizará la versión más reciente del objeto disponible en un sistema marcado como en línea.

El sitio que realiza la actualización se convertirá en el nuevo propietario del objeto y actualizará el historial de objetos con la información del nuevo propietario y el número de secuencia. Esto se utilizará en las operaciones de recuperación o reincorporación del propietario del objeto original para actualizar su historial de objetos con el nuevo propietario.

Nota: las solicitudes de actualización enviadas a sitios en línea donde el propietario del depósito es el sitio fallido utilizarán su depósito local que enumerará la información y el historial de objetos para determinar el propietario del objeto.

En este ejemplo se muestra lo que sucedería con el depósito y el diseño de objetos para el espacio de nombres 1 en una configuración de tres sitios, como se muestra en Figura 12.

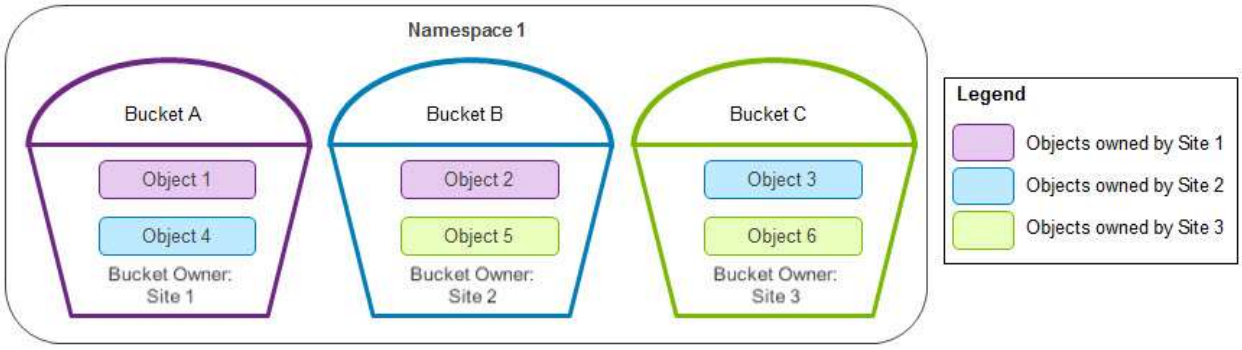

Figura 12 Propiedad de objetos y depósitos para el espacio de nombres 1

Tabla 17 muestra un ejemplo en esta configuración de tres sitios si las tres son verdaderas:

- **Access during outage** está habilitado
- Y **read-only access during outage** está deshabilitado
- Y el sitio 1 está marcado como fallido

Tabla 17 Ejemplo de operaciones que se realizan correctamente o fallan con **access during outage** y **read-only access during outage** deshabilitados con el sitio 1 marcado como fallido temporalmente en una configuración de tres sitios

Operación	Depósito/objeto	Solicitud enviada a	
		Sitio 1 (marcado como fallido)	**Sitio 2 o sitio 3 (en línea)**
Crear objetos en	Depósito A	Satisfactorio	Satisfactorio
	Depósito B	No aprobada El sitio fallido solo puede crear objetos en depósitos de propiedad local	Satisfactorio
	Depósito C	No aprobada El sitio fallido solo puede crear objetos en depósitos de propiedad local	Satisfactorio
Listar objetos en	Depósito A	No aprobada A pesar de que el depósito es de propiedad local, contiene objetos de propiedad remota	Satisfactorio No incluirá objetos que pertenecen a un sitio fallido que no se haya replicado
	Depósito B	No aprobada El sitio fallido solo puede enumerar objetos en depósitos de propiedad local	Satisfactorio
	Depósito C	No aprobada El sitio fallido solo puede enumerar objetos en depósitos de propiedad local	Satisfactorio
Leer o actualizar objeto	Objeto 1	Correcto, tanto el objeto como el depósito son propiedad local	Satisfactorio Requiere que el objeto haya completado la replicación antes de la TSO La actualización adquiere propiedad de objetos
	Objeto 2	Falla, el depósito no es de propiedad local	
	Objeto 3 Objeto 4 Objeto 5 Objeto 6	No aprobada El sitio fallido solo puede leer y actualizar objetos de propiedad local en depósitos de propiedad local	Satisfactorio

Una vez que se restablece el latido entre los sitios, el sistema marca el sitio como en línea y el acceso a estos datos continúa como lo hacía antes de la falla. La operación rejoin:

- Actualiza las tablas de lista de depósitos
- Actualiza las propiedades de objetos cuando sea necesario
- Reanuda el procesamiento de la línea de espera de replicación de los sitios fallidos anteriormente

Nota: ECS solo admite el acceso durante la falla temporal de un solo sitio.

En otros dos ejemplos de sitios, como se muestra en Figura 13. Ambos sitios se pensarán que están en línea y marcarán al otro sitio como fallido cuando ocurrió una TSO, todas las operaciones create, list, read y update serán exitosas.

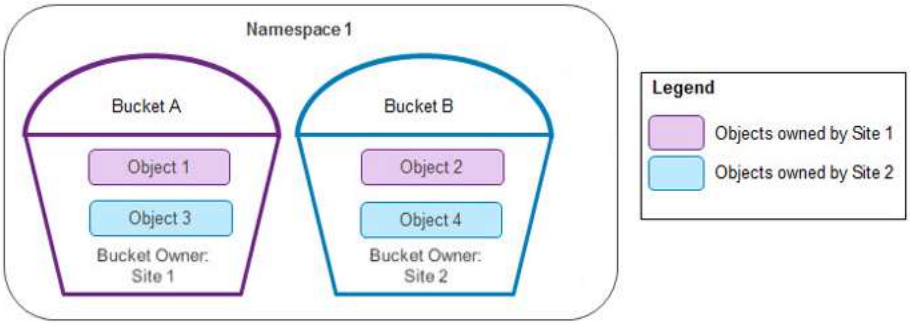

Figura 13 Propiedad de objetos y depósitos en dos sitios

Durante la TSO, todos los objetos se actualizan en cada sitio. Tabla 18 muestra los datos finales en el sitio una vez que se restablece el latido entre los sitios.

Tabla 18 Sitio ganador después de restablecer el sitio

Objetos	Nombre de depósito	Propietario del depósito	Propietario del objeto	Entonces, el sitio “ganador” es...
Objeto 1	Depósito A	Sitio 1	Sitio 1	Sitio 2
Objeto 2	Depósito B	Sitio 2	Sitio 1	El sitio tiene el registro de fecha y hora más reciente
Objeto 3	Depósito A	Sitio 1	Sitio 2	El sitio tiene el registro de fecha y hora más reciente
Objeto 4	Depósito B	Sitio 2	Sitio 2	Sitio 1

Nota: en este ejemplo, el registro de fecha y hora más reciente significa la hora de actualización más reciente del objeto en el sitio.

4.1.2.1 Decodificación XOR con tres o más sitios

Como vimos en la sección codificación XOR, ECS maximiza la eficiencia del almacenamiento de los datos configurados con un grupo de replicación que contiene tres o más sitios. Los datos en copias secundarias de fragmentos se pueden reemplazar por datos en un fragmento de paridad después de una operación XOR. El sitio que contiene la copia principal atenderá las solicitudes de datos en un fragmento codificado. Si este sitio falla, la solicitud se dirigirá al sitio con la copia secundaria del objeto. Sin embargo, dado que esta copia se codificó, el sitio secundario primero debe recuperar la copia de los fragmentos que se utilizaron para la codificación desde los sitios principales en línea. A continuación, realizará una operación XOR para reconstruir el objeto solicitado y responder a la solicitud. Después de que se reconstruyan los fragmentos, también se almacenan para que el sitio pueda responder más rápidamente a las solicitudes subsiguientes.

Tabla 19 muestra un ejemplo de una parte de una tabla de administrador de fragmentos en el sitio 4 en una configuración de cuatro sitios.

Tabla 19 Tabla del administrador de fragmentos del sitio 4 después de completar la codificación XOR

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 4	Codificado
C2	Sitio 2	Sitio 4	Codificado
C3	Sitio 3	Sitio 4	Codificado
C4	Sitio 4		Paridad (C1, C2 y C3)

Figura 14 ilustra las solicitudes involucradas en la recreación de un fragmento para atender una solicitud de lectura durante una TSO.

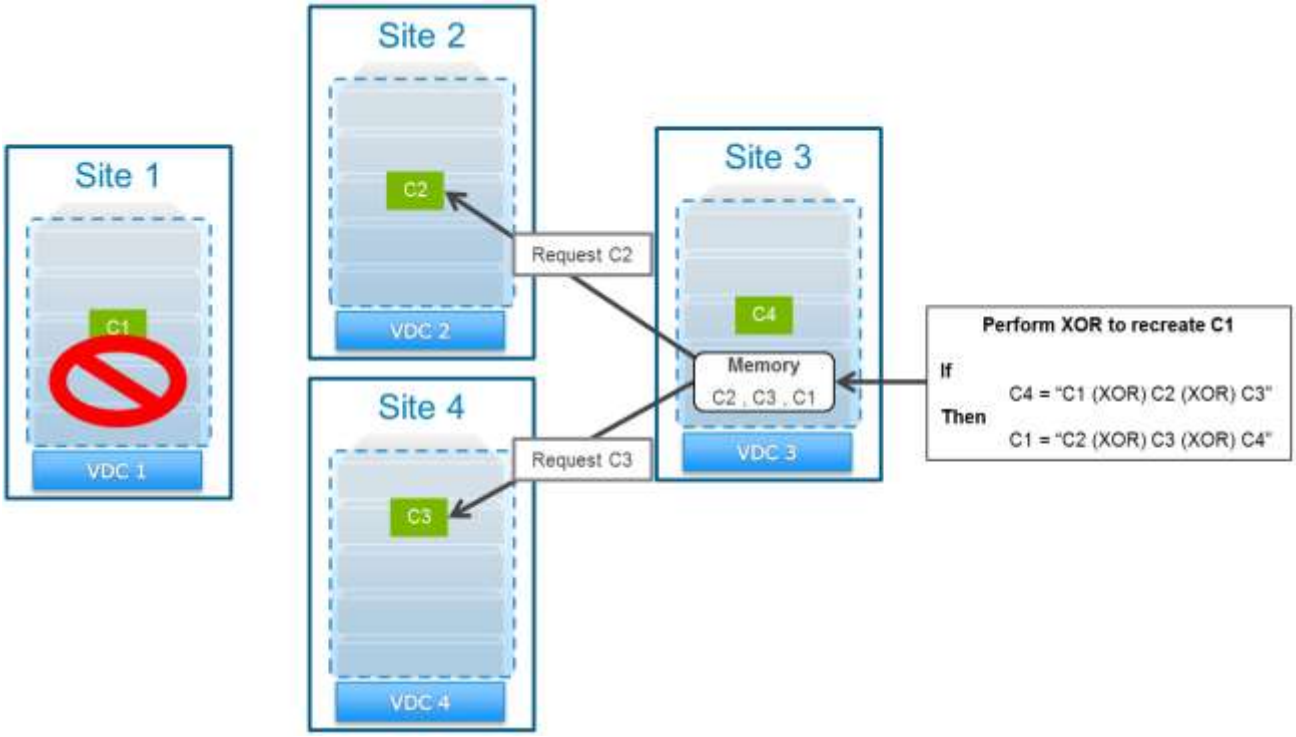

Figura 14 Reparación de una solicitud de lectura mediante la reconstrucción de un fragmento con XOR

En este ejemplo, si se produjo una lectura para un objeto en el fragmento **C1** cuando el **sitio 1** está marcado como fallido, sucedería lo siguiente:

- Dado que el sitio 1 falló, la solicitud se enviaría al sitio secundario del fragmento C1, sitio 4
- El sitio 4 ya realizó XOR en los fragmentos **C1**, **C2** y **C3**, lo que significa que reemplazó su copia local de los datos de estos fragmentos con datos en el fragmento de paridad **C4.**
- El sitio 4 solicita una copia del fragmento **C2** de su sitio principal (sitio 2) y almacena en caché esto localmente.
- El sitio 4 solicita una copia del fragmento **C3** de su sitio principal (sitio 3) y almacena en caché esto localmente.
- A continuación, el sitio 4 realiza una operación XOR entre los fragmentos **C2** y **C3** almacenados en caché con el fragmento de paridad **C4** para volver a crear el fragmento **C1** y almacenarlo localmente en la caché.
- A continuación, el sitio 4 responde a la solicitud de lectura para el objeto en el fragmento **C1**.

Nota: el tiempo que tardan las operaciones de reconstrucción en completarse aumenta linealmente en función de la cantidad de sitios en un grupo de replicación.

4.1.2.2 Con replicación geopasiva

Cualquier dato en un depósito configurado con replicación geopasiva tendrá de dos a cuatro sitios de origen y uno o dos destinos de replicación dedicados. Los datos escritos en los destinos de replicación se pueden reemplazar por datos en un fragmento de paridad después de una operación XOR. El sitio que contiene la copia principal atenderá las solicitudes de datos replicados geográficamente. Si este sitio no es accesible para el sitio solicitante, los datos deberán recuperarse de uno de los sitios de destino de replicación.

Con la replicación geopasiva, los sitios de origen siempre son los propietarios de objetos y depósitos. Dado esto, si un sitio de destino de replicación se marca como fallido temporalmente, todas las operaciones de I/O continuarán como de costumbre. La única excepción es la replicación, que continuará en línea de espera hasta que el sitio de destino de replicación vuelva a unirse a la federación.

Si falla uno de los sitios de origen, las solicitudes al sitio de origen en línea deberán recuperar datos que no son de propiedad local desde uno de los sitios de destino de replicación. Veamos un ejemplo donde el sitio 1 y el sitio 2 son los sitios de origen y el sitio 3 es el sitio de destino de replicación. En este ejemplo, la copia principal de un objeto existe en el fragmento C1 que es propiedad del sitio 1 y el fragmento se replicó en el destino de destino, sitio 3. Si el sitio 1 falla y entra una solicitud en el sitio 2 para leer ese objeto, el sitio 2 tendrá que obtener una copia del sitio 3. Si la copia se codificó, el sitio secundario primero debe recuperar la copia del otro fragmento que se utilizó para la codificación desde el sitio principal en línea. A continuación, realizará una operación XOR para reconstruir el objeto solicitado y responder a la solicitud. Después de que se reconstruyan los fragmentos, también se almacenan para que el sitio pueda responder más rápidamente a las solicitudes subsiguientes.

Tabla 20 muestra un ejemplo de una parte de una tabla de administrador de fragmentos en el destino de replicación geopasivo.

Tabla 20 Tabla de administrador de fragmentos de destino de replicación geopasiva después de completar la codificación XOR

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 3	Codificado
C2	Sitio 2	Sitio 3	Codificado
C3	Sitio 3		Paridad (C1 y C2)

Figura 15 ilustra las solicitudes involucradas en la recreación de un fragmento para atender una solicitud de lectura durante una TSO.

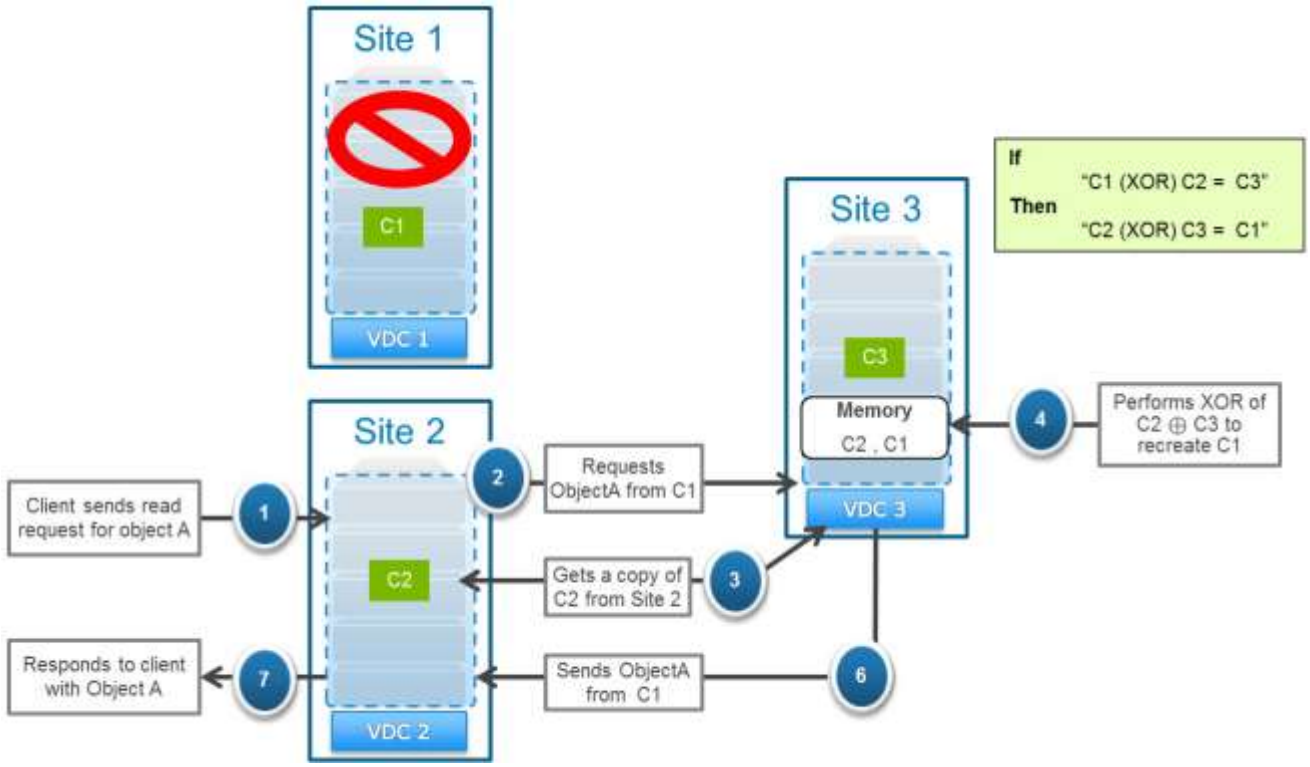

Figura 15 Reparación de una solicitud de lectura mediante la reconstrucción de un fragmento con XOR

En este ejemplo, si se produjo una lectura para un objeto en el fragmento **C1** cuando el **sitio 1** está marcado como fallido temporalmente, sucedería lo siguiente:

- Dado que el sitio 1 falló, la solicitud se enviaría al sitio secundario del fragmento C1, sitio 3.
- El sitio 3 ya realizó XOR en los fragmentos **C1** y **C2**, lo que significa que reemplazó su copia local de los datos de estos fragmentos por datos en el fragmento de paridad **C3**.
- El sitio 3 solicita una copia del fragmento **C2** de su sitio principal (sitio 2) y almacena en caché esto localmente.
- A continuación, el sitio 3 realiza una operación XOR entre el fragmento **C2** almacenado en caché con el fragmento de paridad **C3** para volver a crear el fragmento **C1** y almacenarlo localmente en la caché.
- A continuación, el sitio 3 responde a la solicitud de lectura para el objeto en el fragmento **C1**.

4.1.2.3 Con replicate to all sites habilitado

Los depósitos configurados con las opciones **replicate to all sites** y **access during outage** pueden proporcionar un rendimiento de lectura más rápido. El rendimiento de lectura más rápido se produce en un momento en que todos los sitios están en línea, pero también durante una interrupción temporal del sitio, ya que no se requiere ninguna operación de decodificación XOR y hay más posibilidades de que los datos se lean localmente.

Los datos en depósitos con **replicate to all sites habilitado** se replican en cada sitio. Los objetos de creación y actualización se manejan de la misma manera que si **replicate to all sites** está deshabilitado. Sin embargo, los objetos de lectura y lista se manejan ligeramente diferentes, ya que es posible que algunos datos solo hayan completado la replicación a algunos sitios, pero no a todos antes de que falle el sitio principal.

Durante una operación de lectura, el nodo que atiende la solicitud comprueba primero la versión más reciente de los metadatos del propietario del objeto. Si el nodo solicitante:

- **Es el propietario del objeto**:
  - Si tiene una copia local de los datos que se solicitan, la utilizará para reparar la solicitud.
  - Si otro sitio actualizó el objeto que falló antes de replicar los datos, devolverá la versión que tiene localmente.
- **No es el propietario del objeto**
  - Si el sitio propietario del objeto está en línea y la replicación de los datos del objeto:
    - Se ha completado en este sitio y, a continuación, presta servicios a la solicitud con su copia local de los datos.
    - No se ha completado en este sitio, solicitará una copia del propietario del objeto y la utilizará para reparar la solicitud.
  - Si el propietario del objeto está inactivo
    - Si la replicación del objeto se completó en este sitio, utilizará su copia local de los datos, es posible que esta no sea la versión más reciente.
    - Si la replicación de los datos de objetos al sitio solicitante no completó, el sitio solicitante solicitará una copia del sitio secundario enumerado primero en la tabla del administrador de fragmentos. Si ese sitio es en sí mismo, la operación de lectura fallará.

Durante una operación de enumeración de objetos en un depósito, el nodo requiere información del propietario del depósito y del cabezal para cada objeto del depósito. Si el sitio que es el propietario del objeto o el propietario del depósito está inactivo y **access during outage** también está habilitado, aún puede atender la solicitud si todos los sitios restantes en el grupo de replicación están en línea. Enumerará la versión más reciente de la lista de depósitos que tiene, puede estar ligeramente desactualizada y puede variar entre sitios.

4.1.3 Fallas de múltiples sitios

ECS solo admite el acceso durante una falla temporal de un solo sitio dentro de un grupo de replicación; además, solo un sitio se puede marcar como fallido. Esto significa que si más de un sitio dentro de un grupo de replicación falla simultáneamente, algunas operaciones fallarán. El primer sitio que se determinará como fallido (debido a la pérdida sostenida de latidos) se marcará como fallido. Los sitios restantes que también tengan una pérdida sostenida de latido no se marcarán como fallidos y, por lo tanto, se considerarán en línea.

Por ejemplo, si tenemos cinco sitios en un grupo de replicación y el sitio 1 se identifica como una pérdida sostenida de latido, se marca como fallido. Si el sitio 2 también se identifica como que tiene una pérdida sostenida de latido, permanecerá en línea. Se producirá lo siguiente:

- Si **access during outage** está habilitado y el propietario del depósito es el sitio 2, las lecturas/creaciones/actualizaciones enviadas a otros sitios fallarán independientemente del propietario del objeto. Esto se debe a que primero comprueba con el propietario del depósito para determinar el propietario del objeto. A menos que el propietario del depósito esté marcado como fallido, el solicitante enviará la solicitud al sitio 2, lo cual fallará.
- Las solicitudes de lectura y actualización enviadas al sitio 2 solo se realizarán correctamente si es el propietario del objeto (y el propietario del depósito si **access during outage** está habilitado).

- Las solicitudes de lectura y actualización enviadas a sitios que no sean el sitio 2 se realizarán correctamente solo si el propietario del objeto (y el propietario del depósito si **access during outage** está habilitado) no es el sitio 2.
- La creación del objeto fallará si el propietario del depósito es el sitio 1 o el sitio 2. Esto se debe a que la creación de un objeto requiere la actualización de la lista de depósitos con el nuevo nombre de objeto. Dado que no puede hacer esto correctamente en todos los sitios marcados como en línea, la operación de creación fallará.
- Las solicitudes para enumerar objetos en un depósito solo se realizarán correctamente si el sitio solicitante puede acceder al propietario del depósito y a todos los objetos.
  - Si la solicitud se envía al sitio 2, solo se realizará correctamente si posee el depósito y todos los objetos en el depósito.
  - Si la solicitud se envía a otro sitio, solo se realizará correctamente si ni el depósito ni los objetos dentro del depósito pertenecen al sitio 2.

4.2 Interrupción permanente del sitio (PSO)

Si se produce un desastre en un sitio y el administrador determina que el sitio es irrecuperable, puede iniciar una conmutación por error permanente del sitio (elimine el VDC de la federación). Cuando se inicia una conmutación por error permanente del sitio, todos los fragmentos del sitio fallido se recuperarán en los sitios restantes para restablecer la durabilidad de los datos.

El proceso de recuperación implica que los sitios restantes analicen su tabla de administrador de fragmentos local en busca de referencias a sitios que incluyan el sitio fallido. Cualquiera que encuentre con un tipo de fragmento de:

- **Codificado**
  a. Para los fragmentos cuyo tipo está codificado y cuyo sitio principal está en línea, volverá a crear los datos localmente mediante los datos del sitio principal. Cuando se completa, marca este fragmento como un tipo de copia.
  b. A continuación, volverá a crear el fragmento codificado cuyo sitio principal es el sitio fallido mediante la ejecución de una operación XOR de los tipos de copia previamente recreados con el fragmento de paridad. Este sitio ahora se convierte en el sitio principal de los fragmentos y tiene un tipo de local.
  c. A continuación, estos fragmentos se agregan a la línea de espera de replicación para replicarse a otros sitios enumerados dentro del grupo de replicación.
- **Copy** y un sitio principal enumerado como el sitio fallido se convierten en el nuevo sitio principal. A continuación, agrega el fragmento a su línea de espera de replicación para que se replique en un nuevo sitio secundario.
- **Local** y su sitio secundario es el sitio fallido; se inserta una tarea para replicar el fragmento en un nuevo sitio secundario.

Una vez que se inicia la conmutación por error permanente del sitio, el acceso a los datos que pertenecen al sitio fallido no estará disponible hasta que se complete el proceso de conmutación por error permanente del sitio. La replicación de datos es independiente de las operaciones de conmutación por error y, por lo tanto, no tiene que completarse para que se restaure el acceso a los datos que pertenecen al sitio fallido.

Si se observa un ejemplo de tres sitios, el sitio 1 falla Tabla 21 y Tabla 22 son las tablas del administrador de fragmentos de los dos sitios restantes.

Tabla 21 Tabla del administrador de fragmentos del sitio 2

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 2	Copiar
C2	Sitio 2	Sitio 3	Local
C3	Sitio 1	Sitio 3	Remoto
C4	Sitio 2	Sitio 1	Local

El sitio 2 realizaría lo siguiente:

- Agregue el fragmento **C1** a la línea de espera de replicación que se replicará. El sitio 2 se convertirá en el nuevo sitio principal y el sitio con el nuevo fragmento se convertirá en el nuevo sitio secundario.
- Agregue el fragmento **C4** a la línea de espera de replicación que se replicará y actualice el sitio secundario en la tabla

Tabla 22 Tabla del administrador de fragmentos del sitio 3

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 2	Remoto
C2	Sitio 2	Sitio 3	Codificado
C3	Sitio 1	Sitio 3	Codificado
C4	Sitio 2	Sitio 1	Remoto
C5	Sitio 3		Paridad (C2 y C3)

El sitio 3 realizaría lo siguiente:

1. Vuelva a crear los datos del fragmento **C2** localmente utilizando los datos del sitio principal (**Sitio 2**). Cambie el tipo de fragmento a Copy.
2. Reconstruya el fragmento **C3** con datos **C2** y los datos de paridad **C5** mediante la operación de XOR C2 ⊕ C5. El sitio 3 se convertirá en el nuevo sitio principal.
3. Elimine el fragmento **C5**.
4. Agregue el fragmento **C3** a la línea de espera de replicación que se volverá a replicar; el sitio con el nuevo fragmento se convertirá en el nuevo sitio secundario.

Una vez que la conmutación por error permanente del sitio completa el administrador de fragmentos, las tablas de los dos sitios restantes serían similares a las que se muestran en Tabla 23 y Tabla 24.

Tabla 23 Tabla del administrador de fragmentos del sitio 2 después de que se completa la PSO

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 2	Sitio 3	Local
C2	Sitio 2	Sitio 3	Local
C3	Sitio 3	Sitio 2	Copiar
C4	Sitio 2	Sitio 3	Local

Tabla 24 Tabla del administrador de fragmentos del sitio 3 después de que se completa la PSO

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 2	Sitio 3	Copiar
C2	Sitio 2	Sitio 3	Copiar
C3	Sitio 3	Sitio 2	Local
C4	Sitio 2	Sitio 3	Copiar

4.2.1 PSO para replicación geopasiva

La interrupción permanente del sitio se maneja de manera levemente diferente para los datos replicados mediante la replicación geopasiva. Los datos replicados geopasivos no restablecerán la durabilidad de los datos durante una PSO; en su lugar, se restablece después de que se agrega un tercer sitio nuevo al grupo de replicación. Las operaciones de PSO difieren en función de si el sitio que ha fallado permanentemente es uno de los sitios de origen o el sitio de destino de replicación.

El proceso de recuperación aún implica que los sitios restantes analicen su tabla de administrador de fragmentos local en busca de referencias a sitios que incluyan el sitio fallido. Cualquiera que encuentre con un tipo de fragmento de:

- **Encoded** (existe en el sitio de destino de replicación)
  - Para los fragmentos cuyo tipo está codificado y cuyo sitio principal está en línea, volverá a crear los datos localmente mediante los datos del sitio principal. Cuando se completa, marca este fragmento como un tipo de copia.
  - A continuación, volverá a crear el fragmento codificado cuyo sitio principal es el sitio de origen fallido mediante la ejecución de una operación XOR de los tipos de copia previamente recreados con el fragmento de paridad. Este sitio ahora se convierte en el sitio principal de los fragmentos y tiene un tipo de local. No se crearán sitios secundarios hasta después de que se agregue un tercer sitio al grupo de replicación.
- **Copy** y un sitio principal enumerado como el sitio fallido se convierten en el nuevo sitio principal y su tipo se cambia a Local. No se crearán sitios secundarios hasta después de que se agregue un tercer sitio al grupo de replicación.
- **Local** y su sitio secundario (el destino de replicación) es el sitio fallido. No se crearán nuevos sitios secundarios hasta después de que se agregue un tercer sitio al grupo de replicación.

Después de la PSO, se puede agregar un tercer sitio para restablecer la durabilidad de los datos y proteger contra fallas en todo el sitio. Después de agregar un tercer sitio al grupo de replicación geopasiva, los dos sitios anteriores analizarán su tabla de administrador de fragmentos local en busca de fragmentos sin que se enumere ningún fragmento secundario. Luego:

- Los fragmentos locales en un sitio de origen sin ningún fragmento secundario enumerado iniciarán la replicación de un fragmento secundario al nuevo sitio de destino de replicación. La tabla del administrador de fragmentos se actualizará para incluir la nueva ubicación del fragmento secundario.
- Los fragmentos locales en el sitio de destino de replicación iniciarán una replicación del fragmento a un nuevo sitio de origen. Una vez que se complete la replicación, el tipo de sitio de destino de replicación cambiará de Local a Copy y el tipo de sitio de origen cambiará de Copy a Local. Las operaciones de XOR continuarán en el destino de manera normal.

Una vez que se inicia la PSO en un sitio de origen, el acceso a los datos que pertenecen al sitio fallido no estará disponible hasta que se complete el proceso de conmutación por error permanente del sitio. Una vez que PSO completa el acceso a los datos, se restaura. Hasta que se agregue un tercer sitio al grupo de replicación, cualquier escritura nueva en el sitio de origen en línea se replicará en el destino de replicación, pero no se producirán operaciones XOR. Esto se debe a que XOR solo se ejecuta en fragmentos de dos sitios de origen diferentes, ya que todos los sitios de origen nuevos son los mismos, XOR no puede ejecutarse.

Una vez que se completa la PSO, se puede agregar un tercer sitio al grupo de replicación para restaurar la durabilidad de los datos y protegerse contra fallas en todo el sitio. Además, el destino de replicación puede reanudar la ejecución de operaciones XOR en dos fragmentos de diferentes sitios de origen marcados como tipo Copy.

Echemos un vistazo a un par de ejemplos donde el sitio 1 y el sitio 2 son los sitios de origen, y el sitio 3 es el sitio de destino. Tabla 25 y Tabla 26 son las tablas del administrador de fragmentos de los dos sitios de origen y Tabla 27 es la tabla del administrador de fragmentos del sitio de destino de replicación.

Tabla 25 Sitio 1, tabla del administrador de fragmentos del sitio de origen

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 3	Local
C2	Sitio 2	Sitio 3	Remoto
C3	Sitio 1	Sitio 3	Local

Tabla 26 Sitio 2, tabla del administrador de fragmentos del sitio de origen

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 3	Remoto
C2	Sitio 2	Sitio 3	Local
C3	Sitio 1	Sitio 3	Remoto

Tabla 27 Sitio 3, tabla del administrador de fragmentos de destino de replicación

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 3	Codificado
C2	Sitio 2	Sitio 3	Codificado
C3	Sitio 1	Sitio 3	Copiar
C4	Sitio 3		Paridad (C1 y C2)

Ejemplo 1: si el sitio 3 se elimina debido a una PSO, todos los sitios secundarios quedarán vacíos, pero los sitios y los tipos principales permanecerán. Hasta que se agregue un nuevo destino de replicación, todas las escrituras nuevas tendrán un sitio principal enumerado, pero ningún sitio secundario.

Ejemplo 2: si el sitio 1 se elimina debido a una PSO, sucedería lo siguiente:

- El sitio 3 se convertirá en el nuevo sitio principal con un tipo de local para el fragmento "**C3**" y ningún sitio se enumerará como el sitio secundario.
- El sitio 3 volverá a crear los datos del fragmento **C2** utilizando los datos del sitio principal (sitio 2) y cambiará su tipo de fragmento a Copy.
- El sitio 3 reconstruirá el fragmento **C1** con datos **C2** y los datos de paridad **C4** mediante la operación de XOR C2 ⊕ C4. El sitio 3 se convertirá en el nuevo sitio principal; no se enumerará ningún sitio secundario.
- El sitio 3 eliminará el fragmento **C4**.

Después de que se completa una conmutación por error permanente del sitio 1, las tablas del administrador de fragmentos de los dos sitios restantes serían similares a las que se muestran en Tabla 28 y Tabla 29.

Tabla 28 Sitio 2, tabla del administrador de fragmentos del sitio de origen después de que se complete la PSO

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 3		Remoto
C2	Sitio 2	Sitio 3	Local
C3	Sitio 3		Remoto

Tabla 29 Sitio 3, tabla del administrador de fragmentos del sitio de destino de replicación después de que se complete la PSO

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 3		Local
C2	Sitio 2	Sitio 3	Copiar
C3	Sitio 3		Local

Hasta que se agregue un nuevo sitio de origen, cualquier escritura nueva tendrá un sitio principal del sitio 2 con un tipo de sitio local y un sitio secundario del sitio 3 con un tipo de copia.

Después de agregar un nuevo sitio de origen al grupo de replicación, la durabilidad de los datos para protegerse contra fallas en todo el sitio se restablecerá mediante la adición de un sitio secundario y la replicación de los datos en él. Las operaciones de XOR también se reanudarán en el destino de replicación. Las nuevas tablas del administrador de fragmentos serían las que se muestran Tabla 30 en Tabla 32.

- Los fragmentos C1 y C3 se replicarán en el nuevo sitio de origen, sitio 1. Una vez finalizada la replicación, el sitio principal se enumerará como sitio 1 y el sitio secundario como sitio 3.
- El sitio 3 ejecutará la codificación XOR en los fragmentos C1 y C2, lo que dará como resultado un nuevo fragmento C4 con un tipo de paridad y los fragmentos de tipo C1 y C2 se cambiarán a codificados.

Tabla 30 Nueva tabla de administración de fragmentos del sitio 1 después de restablecer la durabilidad de los datos

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 3	Local
C2	Sitio 2	Sitio 3	Remoto
C3	Sitio 1	Sitio 3	Local

Tabla 31 Tabla del administrador de fragmentos del sitio 2 después de que se agregó el nuevo sitio 1 y se volvió a establecer la durabilidad de los datos

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 3	Remoto
C2	Sitio 2	Sitio 3	Local
C3	Sitio 1	Sitio 3	Remoto

Tabla 32 Sitio 3, tabla de administración de fragmentos del sitio de destino de replicación después de agregar el nuevo sitio 1 y restablecer la durabilidad de los datos

ID del fragmento	Sitio principal	Sitio secundario	Tipo
C1	Sitio 1	Sitio 3	Codificado
C2	Sitio 2	Sitio 3	Codificado
C3	Sitio 1	Sitio 3	Copiar
C4	Sitio 3		Paridad (C1 y C2)

4.2.2 Capacidad de recuperación de múltiples fallas del sitio

ECS solo admite la recuperación a partir de una falla de un solo sitio a la vez. ECS puede recuperarse de múltiples fallas del sitio si las operaciones de PSO y recuperación de datos se completan entre las interrupciones del sitio. Si el segundo sitio falla antes de que se complete la recuperación:

- Para que se ejecuten las operaciones permanentes de recuperación ante interrupciones del sitio, todos los demás sitios del sistema deben estar en línea. Si hay varias fallas simultáneas del sitio, todas menos una deben recuperarse de la TSO antes de que se pueda ejecutar una PSO en un sitio.
- Si se produce una segunda falla en el sitio después de que se completa la PSO, pero antes de que se complete la recuperación de datos, es posible que se pierdan algunos datos.

Si observamos un escenario de cuatro sitios, nos recuperamos correctamente de la pérdida de todo, excepto un sitio (supone que hay suficiente espacio para almacenar todos los datos en el sitio restante):

- El sitio 4 falla
  - El administrador inicia una operación de PSO para quitar el sitio 4
  - Los datos se recuperan en los sitios restantes para restablecer la durabilidad de los datos

  Ahora nos queda una federación de tres sitios que contiene el sitio 1, el sitio 2 y el sitio 3.
- El segundo sitio falla, sitio 2:

  En algún momento después de que PSO y la recuperación de datos completan otro sitio puede fallar, como el sitio 2.
  - El administrador inicia una operación de PSO para quitar el sitio 2
  - Los datos se recuperan en los sitios restantes para restablecer la durabilidad de los datos

  Ahora nos queda una federación de dos sitios que contiene el sitio 1 y el sitio 3.
- El tercer sitio falla, sitio 1:

  En algún momento después de que PSO y la recuperación de datos completan otro sitio puede fallar, como el sitio 1.
  - El administrador inicia una operación de PSO para quitar el sitio 1

  Ahora nos queda una federación de un solo sitio que contiene el sitio 3.

En este ejemplo, se analizaron varias fallas del sitio. Este no es un escenario normal; por lo general, las fallas permanentes del sitio se deben a escenarios de desastre, como terremotos e incendios, y por lo tanto, no son comunes en varias ubicaciones en breve herencia. Por lo general, después de que un solo sitio falla permanentemente, se agrega un nuevo sitio antes de que falle un sitio posterior.

5 Conclusión

La arquitectura de ECS se diseñó desde cero para proporcionar disponibilidad del sistema y durabilidad de los datos. ECS permite la granularidad del administrador en la forma en que equilibran los requisitos de disponibilidad con el TCO. Las funciones como la detección automática de fallas y las funcionalidades de autorreparación minimizan las cargas de trabajo administrativas de TI en los momentos más críticos, cuando hay un evento no planificado, como una interrupción del sitio.

ECS protege los datos dentro de un sitio/VDC contra fallas de disco mediante una combinación de espejeado triple y codificación de eliminación. ECS ofrece dos niveles de protección de codificación de eliminación, el valor predeterminado para los casos de uso típicos y el almacenamiento inactivo, que es más eficiente para los objetos a los que se accede con poca frecuencia. También distribuye los datos entre dominios de fallas para proporcionar protección contra la mayoría de los escenarios de falla.

ECS garantiza la integridad de los datos mediante el cálculo y la escritura de sumas de comprobación como parte de una operación de escritura y la validación de estas sumas de comprobación durante una operación de lectura. La validación de la suma de comprobación también se realiza proactivamente en una tarea en segundo plano.

ECS está diseñado para continuar proporcionando disponibilidad del sistema. Esto se logra con el diseño de arquitectura distribuida, que permite que cualquier nodo de un sitio/VDC pueda atender las solicitudes del cliente.

El diseño de ECS amplía la disponibilidad del sistema y la protección de la durabilidad de los datos mediante la adición de protección opcional contra una falla completa en todo el sitio. Para ello, federa sitios y permite que el administrador configure una variedad de opciones de políticas de grupo de replicación. Estas opciones se pueden configurar en el nivel del depósito y determinar dónde replicar los datos y cómo almacenarlos en los sitios remotos, así como las opciones de acceso durante la interrupción.

Además, ECS ofrece a los clientes una opción de “access during outage”, que permite leer, listar y, opcionalmente, realizar operaciones de escritura y actualización enviadas a un sitio en línea cuando el depósito o el objeto se marcan como fallidos.

Si un administrador determina que un sitio no se puede recuperar, puede iniciar una interrupción permanente del sitio. Esto eliminará el VDC/sitio del grupo de replicación y volverá a crear los datos según sea necesario para restablecer la durabilidad de los datos.

En conclusión, ECS proporciona una solución de almacenamiento de nube de calidad empresarial con resiliencia incorporada en la que puede confiar.

A Soporte técnico y recursos

Dell.com/support se centra en satisfacer las necesidades de los clientes con servicios y soporte comprobados.

Los videos y documentos técnicos de almacenamiento proporcionan pericia que ayuda a garantizar el éxito de los clientes en las plataformas de almacenamiento Dell EMC.

A.1 Recursos relacionados

- Documentación técnica de la visión general y la arquitectura de ECS
- Comunidad de ECS
- Prueba de producto de ECS
- Documentación de productos de ECS en el sitio de soporte o en el sitio de la comunidad
- SolVe Desktop (generador de procedimientos)