DELL Technologies

White paper técnico

Visão geral e arquitetura do ECS

Resumo geral

Este documento apresenta uma visão geral técnica e o projeto da plataforma de armazenamento em object definida por software e com escala de nuvem Dell EMC™ ECS™.

February 2021

Revisões

Data	Descrição
Dezembro de 2015	Versão inicial
Maio de 2016	Atualização para 2.2.1
Setembro de 2016	Atualização para 3.0
Agosto de 2017	Atualização para 3.1
Março de 2018	Atualização para 3.2
de setembro de 2018	Atualização para o hardware de 3ª geração
Fevereiro de 2019	Atualização para 3.3
Setembro de 2019	Atualização para 3.4
Fevereiro de 2020	Atualização das alterações do ECSDOC-628
Maio de 2020	Atualização para 3.5
Novembro de 2020	Atualização para 3.6
fevereiro de 2021	Atualizado para 3.6.1

Agradecimentos

Este artigo foi produzido por:

Autor: Zhu, Jarvis

Índice

Resumo executivo

As organizações precisam de opções para consumir serviços em nuvem pública com a confiabilidade e o controle de uma infraestrutura em nuvem privada. Dell EMC ECS é uma plataforma de armazenamento em objeto definida por software, com suporte para IPv6 e com escala de nuvem que oferece serviços de armazenamento S3, Atmos, CAS, Swift, NFSv3 e HDFS em uma só plataforma moderna.

Com o ECS, os administradores podem gerenciar facilmente a infraestrutura de armazenamento globalmente distribuída em um só namespace global, com acesso ao conteúdo em qualquer lugar. Os principais componentes do ECS são dispostos em níveis para oferecer flexibilidade e resiliência. Cada nível é abstraído e escalável de modo independente, com alta disponibilidade.

O acesso à API RESTful simples para serviços de armazenamento está sendo adotado pelos desenvolvedores. O uso de semântica HTTP, como GET e PUT, simplifica a lógica de aplicativo necessária em comparação com operações de arquivo tradicionais, mas conhecidas e baseadas em caminhos. Além disso, o sistema de armazenamento subjacente do ECS é altamente consistente, o que significa que ele pode garantir uma resposta autorizada. Os aplicativos que são necessários para garantir a entrega autorizada de dados podem fazer isso sem uma lógica complexa de código usando o ECS.

1 Introdução

Este documento apresenta uma visão geral da plataforma de armazenamento em object Dell EMC ECS. Ele detalha a arquitetura de projeto do ECS e os componentes principais, como os serviços de armazenamento e os mecanismos de proteção de dados.

1.1 Público

Este documento é destinado a qualquer pessoa interessada em entender o valor e a arquitetura do ECS. Ele tem como objetivo apresentar contexto, com links para informações adicionais.

1.2 Escopo

Este documento se concentra principalmente na arquitetura do ECS. Ele não aborda os procedimentos de instalação, administração e upgrade do software ou hardware do ECS. Além disso, não aborda informações específicas sobre o uso e a criação de aplicativos com as APIs do ECS.

As atualizações deste documento são feitas periodicamente e, geralmente, coincidem com as principais versões ou os novos recursos.

2 Valor do ECS

O ECS oferece valor significativo para empresas e prestadores de serviços que buscam uma plataforma projetada para dar suporte ao rápido crescimento de dados. As principais vantagens e recursos do ECS que permitem que as empresas gerenciem e armazenem globalmente o conteúdo distribuído em escala são:

- **Escala de nuvem** — o ECS é uma plataforma de armazenamento em object para cargas de trabalho tradicionais e de última geração. A arquitetura em níveis definida por software do ECS promove uma escalabilidade ilimitada. Os destaques de recursos são:
  - Infraestrutura em object distribuída globalmente
  - Escala superior a exabytes, sem limites sobre a capacidade do pool de armazenamento, do cluster ou do ambiente federado
  - Não existem limites para o número de objetos em um sistema, namespace ou bucket
  - Eficiente em cargas de trabalho de arquivo pequenas e grandes, sem limites sobre o tamanho dos objetos
- **Implementação flexível** — o ECS tem flexibilidade inigualável, com recursos como:
  - Implementação de equipamentos
  - Implementação somente de software com suporte a hardware padrão do setor certificado ou personalizado
  - Suporte multiprotocolo: Object (S3, Swift, Atmos, CAS) e File (HDFS, NFSv3)
  - Várias cargas de trabalho: aplicativos modernos e arquivamento de longo prazo
  - Armazenamento secundário para Data Domain Cloud Tier e Isilon, usando CloudPools
  - Caminhos de upgrade não disruptivos para os modelos do ECS da geração atual
- **Nível empresarial** — o ECS oferece aos clientes mais controle de seus ativos de dados com armazenamento de classe empresarial em um sistema seguro e compatível, como:
  - Dados em repouso (D@RE) com rodízio de chaves e gerenciamento de chaves externas.
  - Comunicação criptografada entre locais
  - Desativa as portas 9101/9206 por padrão para capacitar as organizações a atender às políticas de conformidade
  - Geração de relatórios, retenção de registros baseada em políticas e eventos e fortalecimento de plataforma para conformidade com a regra 17a-4(f) da SEC, inclusive gerenciamento avançado de retenção, como retenção legal e governança mín./máx.
  - Conformidade com as diretrizes de reforço Security Technical Implementation Guide (STIG) da Defense Information Systems Agency (DISA).
  - Autenticação, autorização e controles de acesso com Active Directory e LDAP
  - Integração à infraestrutura de monitoramento e emissão de alertas (traps SNMP e SYSLOG)
  - Recursos empresariais aprimorados (multi-tenancy, monitoramento de capacidade e alertas)
- **Redução do TCO** — o ECS pode reduzir drasticamente o custo total de propriedade (TCO) em relação ao armazenamento tradicional e ao armazenamento em nuvem pública. Ele também oferece um TCO mais baixo que a fita para retenção a longo prazo. Entre os recursos, estão:
  - Namespace global
  - Desempenho de arquivos pequenos e grandes
  - Migração perfeita do Centera
  - Totalmente compatível com Atmos REST
  - Baixa sobrecarga de gerenciamento
  - Pequeno espaço ocupado pelo data center
  - Alta utilização do armazenamento

O projeto do ECS é otimizado para os seguintes casos de uso principais:

- **Aplicativos modernos** — o ECS foi projetado para o desenvolvimento moderno, como os aplicativos em nuvem, aplicativos móveis e da Web de última geração. O desenvolvimento de aplicativos é simplificado com um armazenamento altamente consistente. Juntamente com o acesso de leitura/gravação simultâneo em vários locais e para vários usuários, à medida que a capacidade do ECS muda e aumenta, os desenvolvedores nunca precisam recodificar seus aplicativos.
- **Armazenamento secundário** — o ECS é usado como armazenamento secundário para liberar o armazenamento primário dos dados acessados com pouca frequência e, ao mesmo tempo, mantê-lo razoavelmente acessível. Os exemplos são produtos de armazenamento em camadas baseados em políticas, como Data Domain Cloud Tier e Isilon CloudPools. O GeoDrive, um aplicativo baseado em Windows, dá aos sistemas Windows acesso direto ao ECS para armazenar dados.
- **Arquivamento com proteção regional**— o ECS serve como uma nuvem segura e acessível no local para fins de arquivamento e retenção em longo prazo. Usar o ECS como um nível de arquivamento pode reduzir significativamente as capacidades de armazenamento primário. Para permitir uma melhor eficiência de armazenamento para os casos de uso de arquivamento estático, um esquema de codificação de eliminação (EC) 10+2 está disponível, além do padrão de 12+4.
- **Repositório de conteúdo global** — muitas vezes, repositórios de conteúdo não estruturados que contêm dados, como imagens e vídeos, são armazenados em sistemas de armazenamento de alto custo, o que torna impossível para as empresas gerenciar o enorme crescimento de dados de modo econômico. O ECS permite a consolidação de vários sistemas de armazenamento em um só repositório de conteúdo globalmente acessível e eficiente.
- **Armazenamento para Internet das Coisas** — a Internet das Coisas (IoT) oferece uma nova oportunidade de receita para empresas que podem extrair valor dos dados do cliente. O ECS oferece uma arquitetura eficiente de IoT para coleta de dados não estruturados em grande escala. Sem limites sobre o número de objetos, o tamanho dos objetos ou metadados personalizados, o ECS é a plataforma ideal para armazenar dados de IoT. O ECS também pode simplificar alguns fluxos de trabalho analíticos, permitindo que os dados sejam analisados diretamente na plataforma ECS, sem a necessidade de processos demorados de extração, transformação e carregamento (ETL). Os clusters do Hadoop podem executar consultas usando dados armazenados no ECS por outra API de protocolo, como S3 ou NFS.
- **Repositório de provas da videovigilância** — ao contrário dos dados da IoT, os dados de videovigilância têm um número muito menor de armazenamento em objeto, mas uma capacidade muito maior por arquivo. Embora a autenticidade dos dados seja importante, a retenção de dados não é tão essencial. O ECS pode ser uma área inicial de baixo custo ou um local de armazenamento secundário para esses dados. O software de gerenciamento de vídeo pode aproveitar os avançados recursos de metadados personalizados para a marcação de arquivos com detalhes importantes como o local da câmera, o requisito de retenção e o requisito de proteção de dados. Além disso, os metadados podem ser usados para configurar o arquivo no status somente leitura para garantir uma cadeia de custódia do arquivo.
- **Data lakes e lógica analítica** — os dados e a lógica analítica se tornaram um diferencial competitivo e uma fonte principal de geração de valor para as organizações. No entanto, transformar dados em um ativo corporativo valioso é um tópico complexo que pode facilmente envolver o uso de dezenas de tecnologias, ferramentas e ambientes. O ECS oferece um conjunto de serviços para ajudar o cliente a coletar, armazenar, controlar e analisar os dados em qualquer escala.

3 Arquitetura

O ECS foi projetado com alguns princípios fundamentais de projeto, como namespace global com forte consistência; recurso de scale-out, multi-tenancy seguro; e desempenho superior para objetos pequenos e grandes. Ele foi desenvolvido como um sistema completamente distribuído, seguindo o princípio de aplicativos em nuvem, em que todas as funções do sistema são criadas como uma camada independente. Com esse projeto, cada camada é horizontalmente escalável entre todos os nós do sistema. Os recursos são distribuídos entre todos os nós para aumentar a disponibilidade e compartilhar a carga.

Esta seção analisará em detalhes a arquitetura do ECS e o projeto de software e hardware.

3.1 Visão geral

O ECS é implementado em um conjunto de hardware qualificado padrão do setor ou como um equipamento de armazenamento turnkey. Os principais componentes do ECS são:

- **Portal e serviços de provisionamento do ECS** — IU na Web e CLI com base em API para autoatendimento, automação, geração de relatórios e gerenciamento dos nós do ECS. Essa camada também lida com licenciamento, autenticação, multi-tenancy e serviços de provisionamento, como criação de namespace.
- **Serviços de dados** — serviços, ferramentas e APIs para dar suporte ao acesso a arquivos e a objetos do sistema.
- **Mecanismo de armazenamento** — principal serviço responsável por armazenar e recuperar dados, gerenciar transações e proteger e replicar dados localmente e entre locais.
- **Fabric** — serviço de organização em clusters para alertas e gerenciamento de upgrade, integridade e configuração.
- **Infraestrutura** — SUSE Linux Enterprise Server 12 para o sistema operacional base no equipamento turnkey ou sistemas operacionais Linux qualificados para a configuração de hardware padrão do setor.
- **Hardware** — um equipamento turnkey ou hardware qualificado padrão do setor.

A Figura 1 mostra uma visualização gráfica dessas camadas, que são descritas em detalhes nas seções a seguir.

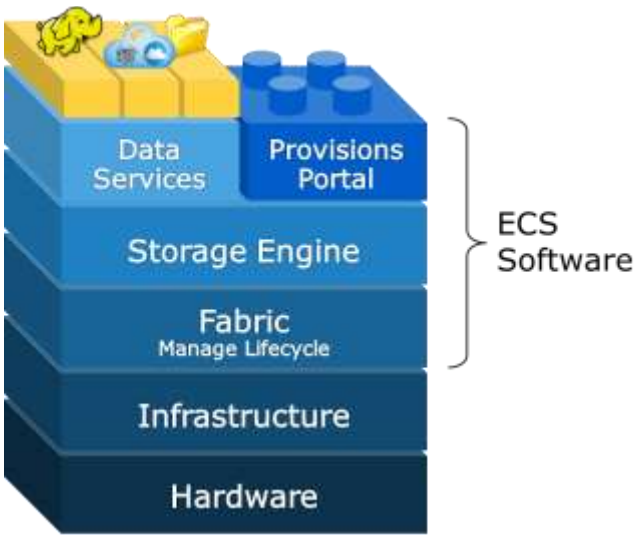

Figura 1 Camadas da arquitetura do ECS

3.2 Portal e serviços de provisionamento do ECS

Os administradores de armazenamento gerenciam o ECS usando o portal e os serviços de provisionamento do ECS. O ECS oferece uma GUI baseada na Web (IU na Web) para gerenciar, licenciar e provisionar nós do ECS. O portal apresenta recursos abrangentes de geração de relatórios que incluem:

- Utilização da capacidade por local, pool de armazenamento, nó e disco.
- Monitoramento de desempenho em termos de latência, throughput e andamento da replicação.
- Informações de diagnóstico, como status de recuperação de nós e discos.

O painel de indicadores do ECS apresenta informações gerais sobre a integridade e o desempenho em nível do sistema. Essa visualização unificada aprimora a visibilidade geral do sistema. Os alertas notificam os usuários sobre eventos críticos, como limites de capacidade, limites de cota, falhas de discos ou nós, ou falhas de software. O ECS também oferece uma interface de linha de comando para instalar, fazer upgrade e monitorar o ECS. O acesso aos nós para uso da linha de comando é feito por meio de SSH. Uma captura de tela do painel de indicadores do ECS é exibida na Figura 2 abaixo.

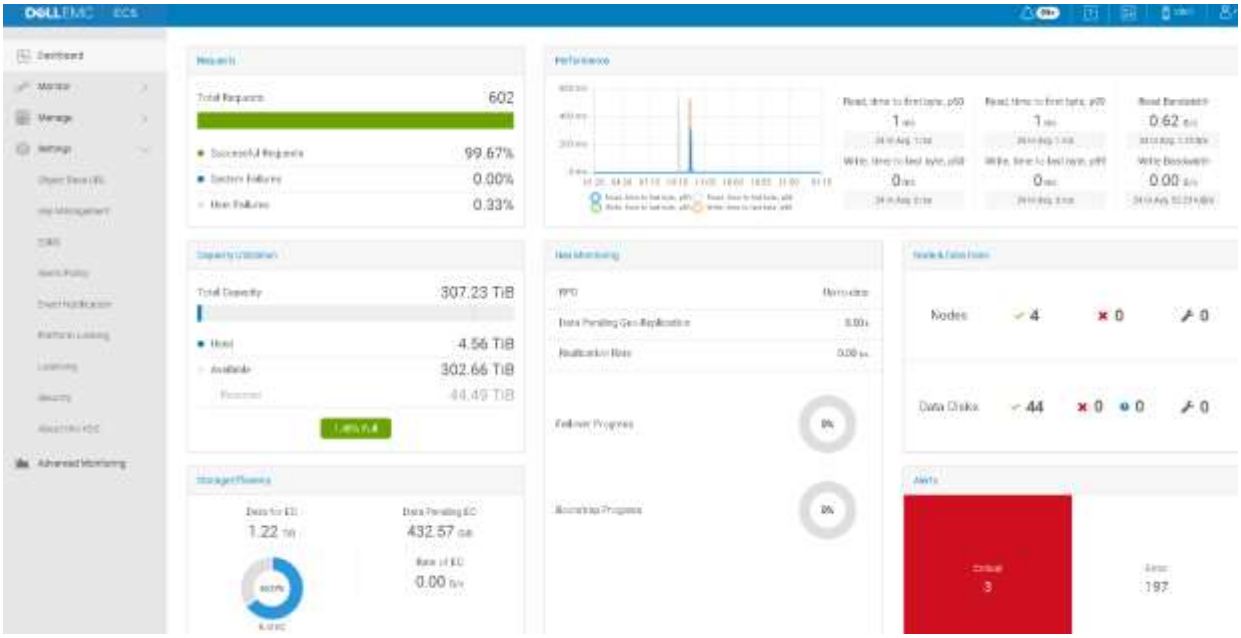

Figura 2 Painel de indicadores do ECS na IU da Web

Os relatórios detalhados de desempenho estão disponíveis na UI, na pasta Advanced Monitoring. Os relatórios são exibidos em um painel de indicadores do Grafana. Há filtros disponíveis para detalhar namespaces, protocolos ou nós especificados. Um exemplo de um relatório de desempenho do protocolo S3 é mostrado abaixo em Figura 3.

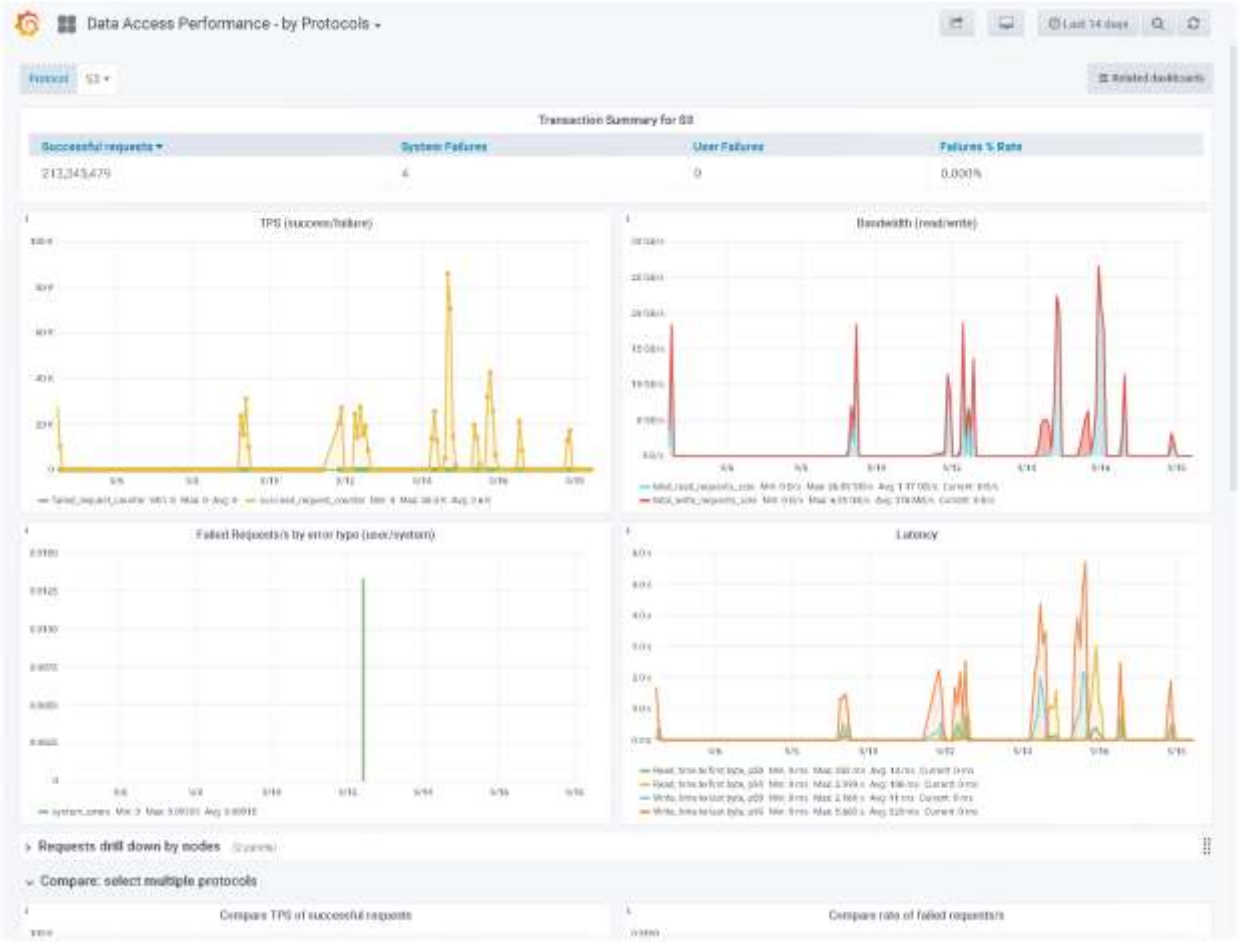

Figura 3 Visualização do monitoramento avançado usando o Grafana

O ECS também pode ser gerenciado usando APIs RESTful. A API de gerenciamento permite que os usuários administrem o ECS em suas próprias ferramentas, scripts e aplicativos novos ou existentes. A IU do ECS na Web e as ferramentas de linha de comando foram criadas usando APIs REST de gerenciamento do ECS.

O ECS dá suporte aos seguintes servidores de notificação de eventos, que podem ser configurados usando a IU na Web, a API ou a CLI:

- Servidores SNMP (Simple Network Management Protocol, protocolo simples de gerenciamento de rede)
- Servidores Syslog

O *Guia do Administrador do ECS* apresenta mais informações e detalhes sobre como configurar serviços de notificação.

3.3 Serviços de dados

Os métodos padrão de object e file são usados para acessar os serviços de armazenamento do ECS. Para S3, Atmos e Swift, APIs RESTful via HTTP são usadas para acesso. Para armazenamento associativo (CAS), é usado um método/SDK exclusivo de acesso. O ECS oferece suporte nativamente a todos os procedimentos de NFSv3, exceto LINK. Os buckets do ECS agora podem ser acessados pelo S3a.

O ECS oferece acesso por vários protocolos: os dados que são incluídos por meio de um protocolo podem ser acessados por meio de outro. Isso significa que os dados podem ser incluídos por meio do S3 e modificados por meio de NFSv3 ou Swift, ou vice-versa. O acesso por vários protocolos tem algumas exceções, devido à semântica dos protocolos e às representações do projeto dos protocolos. A Tabela 1 destaca os métodos de acesso e quais protocolos são interoperáveis.

Tabela 1 Interoperabilidade de protocolos e serviços de dados compatíveis com o ECS

Protocolos		**Compatível**	**Interoperabilidade**
Objeto	S3	Recursos adicionais, como Atualizações do Intervalo de Bytes e ACLS Avançado	HDFS, NFS, Swift
	Atmos	Versão 2.0	NFS (somente objetos com base em caminho, não os com base no estilo de ID de objeto)
	Swift	Autenticação Swift e Keystone v3 e APIs V2	HDFS, NFS, S3
	CAS	SDK v3.1.544 ou versões posteriores	N/A
File	HDFS	Compatibilidade com Hadoop 2.7	S3, NFS, Swift
	NFS	NFSv3	S3, Swift, HDFS, Atmos (somente objetos com base em caminho, não os com base no estilo de ID de objeto)

Os serviços de dados, que também são chamados de serviços principais, são responsáveis por atender a solicitações do client, extrair as informações necessárias e passá-las ao mecanismo de armazenamento para processamento adicional. Todos os serviços principais são combinados em um só processo, *dataheadsvc,* que é executado na camada de infraestrutura. Esse processo é encapsulado, ainda mais em um contêiner do Docker chamado *object-main*, que é executado em todos os nós do ECS. A seção *Infraestrutura* deste documento abrange mais detalhes sobre o Docker. Os requisitos de porta de serviço de protocolo do ECS, como a porta 9020 para comunicação de S3, estão disponíveis no mais recente *Guia de Configuração de Segurança do ECS*.

3.3.1 Objeto

O ECS dá suporte a APIs CAS, S3, Atmos e Swift para acesso a objetos. Com exceção do CAS, os objetos ou os dados são gravados, recuperados, atualizados e excluídos por meio de chamadas HTTP ou HTTPS de GET, POST, PUT, DELETE e HEAD. Para CAS, a comunicação TCP padrão e os métodos e chamadas de acesso específicos são usados.

O ECS oferece uma instalação para pesquisa de metadados de objeto usando uma linguagem de consulta avançada. Esse é um recurso avançado do ECS que permite que os clients de objeto do S3 pesquisem objetos em buckets usando metadados personalizados e do sistema. Embora seja possível pesquisar usando quaisquer metadados, ao pesquisar em metadados que foram configurados especificamente para ser indexados em um bucket, o ECS pode retornar as consultas de maneira mais rápida, especialmente para buckets com bilhões de objetos.

Até 30 campos de metadados definidos pelo usuário podem ser indexados por bucket. Os metadados são especificados no momento da criação do bucket. O recurso de pesquisa de metadados pode ser ativado nos buckets com criptografia ativada no servidor; no entanto, qualquer atributo indexado de metadados do usuário utilizado como uma chave de pesquisa não será criptografado.

Nota: o desempenho é afetado ao gravar dados em buckets configurados para indexação de metadados. O impacto sobre as operações aumenta à medida que aumenta o número de campos indexados. O impacto sobre o desempenho precisa de uma consideração cuidadosa sobre a escolha quanto à indexação de metadados em um bucket ou não e, em caso afirmativo, quantos índices serão mantidos.

Para objetos do CAS, a API de consulta do CAS oferece uma capacidade semelhante de pesquisar objetos com base nos metadados que são mantidos para objetos do CAS que não precisam ser ativados explicitamente.

Para obter mais informações sobre APIs do ECS e APIs para pesquisa de metadados, consulte o mais recente *Guia de Acesso a Dados do ECS*. Para os SDKs do Atmos e do S3, consulte o site do GitHub SDK de Serviços de Dados da Dell EMC ou Dell EMC ECS. Para CAS, consulte o site Centera Community. O acesso a vários exemplos, recursos e assistência para desenvolvedores pode ser encontrado na ECS Community.

Os aplicativos client, como S3 Browser e Cyberduck, oferecem uma maneira de testar ou acessar rapidamente os dados armazenados no ECS. O ECS Test Drive é oferecido gratuitamente pela Dell EMC, o que permite o acesso a um sistema ECS voltado para o público para fins de teste e desenvolvimento. Depois de inscrever-se no ECS Test Drive, os endpoints REST recebem credenciais do usuário para cada um dos protocolos de object. Qualquer pessoa pode usar o ECS Test Drive para testar o aplicativo de API do S3.

Nota: apenas o número de metadados que podem ser indexados por bucket é limitado a 30 no ECS. Não há limite sobre o número total de metadados personalizados armazenados por objeto, apenas o número indexado para pesquisa rápida.

3.3.2 HDFS

O ECS pode armazenar dados do file system do Hadoop. Como um file system compatível com Hadoop, as organizações podem criar repositórios de Big Data no ECS que a lógica analítica do Hadoop pode consumir e processar. O serviço de dados do HDFS é compatível com o Apache Hadoop 2.7, com suporte a ACLs refinadas e atributo estendido do file system.

O ECS foi validado e testado com Hortonworks (HDP 2.7). O ECS também oferece suporte a serviços como YARN, MapReduce, Pig, Hive/Hiveserver2, HBase, Zookeeper, Flume, Spark e Sqoop.

3.3.2.1 Suporte a Hadoop S3A

O ECS é oferece suporte ao client Hadoop S3A para armazenar dados do Hadoop. O S3A é um conector de código aberto para Hadoop, baseado no SDK oficial da Amazon Web Services (AWS). Ele foi criado para resolver problemas de dimensionamento de armazenamento e custo que muitos administradores do Hadoop tinham com o HDFS. O Hadoop S3A conecta clusters do Hadoop a qualquer armazenamento de objetos compatível com S3 que esteja na nuvem pública, na nuvem híbrida ou no local.

Nota: o suporte a S3A está disponível no Hadoop 2.7 ou em versões posteriores

Figura 4 Arquitetura do Hadoop e do ECS

Conforme mostrado na Figura 4, quando o cliente configura o cluster do Hadoop no HDFS tradicional, sua configuração do S3A aponta para os dados do objeto ECS para fazer todas as atividades do HDFS. Em cada nó do Hadoop HDFS, qualquer componente tradicional do Hadoop usaria o client S3A do Hadoop para executar a atividade do HDFS.

Análise de configuração do Hadoop usando o console de serviços do ECS

O Console de Serviços (SC) do ECS pode ler e interpretar os parâmetros de configuração do Hadoop relacionados às conexões com ECS para S3A. Além disso, o SC fornece uma função, *Get_Hadoop_Config*, que lê a configuração em cluster do Hadoop e verifica as configurações do S3A em busca de erros e valores. Entre em contato com a equipe de suporte do ECS para obter assistência com a instalação do ECS SC.

Implementação do Privacera com o Hadoop S3A

O Privacera é um fornecedor terceirizado que implementou um agente do Hadoop no client e realizou uma integração com a segurança granular do Ambari for S3 (AWS e ECS). Embora o Privacera seja compatível com o Cloudera Distribution of Hadoop (CDH), o Cloudera (outro fornecedor terceirizado) não é compatível com o Privacera no CDH.

Nota: os usuários do CDH devem usar os serviços de segurança IAM do ECS. Se você quiser acesso seguro ao S3A sem usar IAM do ECS, entre em contato com o a equipe de suporte.

Consulte o *Guia de Acesso a Dados do ECS* mais recente para obter mais informações sobre o suporte ao S3A

Segurança do Hadoop S3A

O IAM de ECS permite que o administrador do Hadoop configure políticas de acesso para controlar o acesso aos dados do Hadoop do S3A. Depois que as políticas de acesso são definidas, há duas opções de acesso do usuário que os administradores do Hadoop podem configurar:

- Usuários/grupos do IAM
  - Criar grupos do IAM que se conectam às políticas
  - Criar usuários do IAM que são membros de um grupo do IAM

- Afirmações SAML (usuários federados)
  - Criar funções do IAM que se conectam às políticas
  - Configurar CrossTrustRelationship entre o provedor de identidade (AD FS) e o ECS que associam grupos do AD às funções do IAM

O administrador do ECS e o administrador do Hadoop precisam trabalhar juntos para definir previamente as políticas apropriadas. Os exemplos fictícios que seguem descrevem três tipos de usuários do Hadoop para os quais criaremos políticas. São eles:

- **Administrador do Hadoop** — realiza todas as operações, exceto criar bucket e excluir bucket
- **Usuário power do Hadoop** — realiza todas as operações, exceto criar bucket, excluir bucket e excluir objetos
- Usuário somente leitura do Hadoop — apenas lista e lê objetos

Para obter mais informações sobre IAM do ECS, consulte AM do ECS na página 38.

3.3.2.2 Suporte de client do ECS HDFS

O ECS é integrado ao Ambari, o que permite que você implemente facilmente o arquivo jar do client ECS HDFS e especifique o ECS HDFS como o file system padrão em um cluster do Hadoop. O arquivo jar é instalado em cada nó de um cluster participante do Hadoop. O ECS oferece funcionalidades de file system e armazenamento equivalentes às que os nós de nome e dados fazem em uma implementação Hadoop. Ele simplifica o fluxo de trabalho do Hadoop, eliminando a necessidade de migração de dados para um DAS local do Hadoop e/ou de criação de um mínimo de três cópias. A Figura 5 abaixo mostra o arquivo jar do client ECS HDFS instalado em cada nó de computação do Hadoop e o fluxo geral de comunicação.

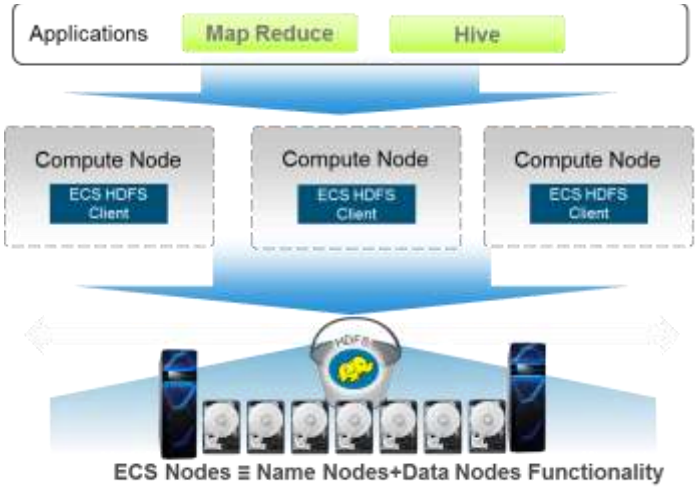

Figura 5 ECS atuando como nós de nome e dados para um cluster do Hadoop

Outros aprimoramentos adicionados ao ECS para HDFS são:

- **Autenticação de usuário proxy** — personificação de Hive, HBase e Oozie.
- **Segurança** — imposição de ACLs no servidor e adição de superusuários e grupos de superusuários do Hadoop, bem como grupos padrão nos buckets.

3.3.3 NFS

O ECS inclui suporte nativo a arquivos com o NFSv3. Os principais recursos do serviço de dados de arquivo do NFSv3 incluem:

- **Namespace global** — acesso ao arquivo de qualquer nó, em qualquer local.
- **Bloqueio global** — no NFSv3, o bloqueio é **somente informativo**. O ECS oferece suporte às implementações de client compatíveis que permitem bloqueios compartilhados e exclusivos, obrigatórios e baseados em intervalo.
- **Acesso por vários protocolos** — acesso aos dados usando diferentes métodos de protocolo.

As exportações NFS, as permissões e os mapeamentos de grupos de usuários são criados usando a IU na Web ou a API. Os clients compatíveis com o NFSv3 montam as exportações usando nomes de namespace e bucket. Este é um exemplo de comando para montar um bucket:

```
mount –t nfs –o vers=3 s3.dell.com:/namespace/bucket
```

Para obter transparência dos clients durante a falha de um nó, um balanceador de carga é recomendado para esse fluxo de trabalho.

O ECS integrou rigidamente as outras implementações do servidor NFS, como *lockmgr*, *statd*, *nfsd* e *mountd*; assim, esses serviços não dependem da camada de infraestrutura (sistema operacional do host) para gerenciar. O suporte ao NFSv3 tem os seguintes recursos:

- Não há limites de projeto sobre o número de arquivos ou diretórios.
- O tamanho da gravação do arquivo pode ser de até 16 TB.
- Capacidade de dimensionar em até 8 locais com um só namespace/exportação global.
- Suporte a autenticação Kerberos e AUTH_SYS.

Os serviços de arquivo do NFS processam as solicitações do NFS provenientes dos clientes; no entanto, os dados são armazenados como objetos no ECS. Um identificador de arquivo do NFS é associado a um ID de objeto. Como o arquivo é basicamente mapeado a um objeto, o NFS tem recursos como o serviço de dados de objeto, que inclui:

- Gerenciamento de cotas em nível de bucket.
- Criptografia em nível de objeto.
- Write-Once-Read-Many (WORM) ao nível de bucket.
  - O WORM é implementado usando o período de confirmação automática durante a criação de novos buckets.
  - O WORM só é aplicável a buckets incompatíveis.

3.3.4 Conectores e gateways

Vários produtos de software de terceiros têm a capacidade de acessar o armazenamento em object ECS. Fornecedores de software independentes (ISVs), como Panzura, Ctera e Syncplicity, criam uma camada de serviços que oferece acesso do client ao armazenamento em objeto ECS por meio de protocolos tradicionais, como SMB/CIFS, NFS e iSCSI. As organizações também podem acessar ou carregar dados no armazenamento ECS com os seguintes produtos da Dell EMC:

- **Isilon CloudPools** — armazenamento em camadas dos dados e baseado em políticas do Isilon ao ECS.
- **Data Domain Cloud Tier** — armazenamento em camadas nativo e automatizado de dados desduplicados do Data Domain ao ECS para retenção em longo prazo. O Data Domain Cloud Tier oferece uma solução segura e econômica para criptografar dados na nuvem com um espaço ocupado reduzido do armazenamento e menor largura de banda da rede.

- **GeoDrive** — serviço de armazenamento ECS baseado em stub para desktops e servidores do Microsoft® Windows®.

3.4 Mecanismo de armazenamento

O mecanismo de armazenamento é o núcleo do ECS. A camada do mecanismo de armazenamento contém os principais componentes responsáveis por processar as solicitações e por armazenar, recuperar, proteger e replicar dados.

Esta seção descreve os princípios de projeto e como os dados são representados e tratados internamente.

3.4.1 Serviços de armazenamento

O mecanismo de armazenamento do ECS inclui os seguintes serviços, como exibido na Figura 6.

Figura 6 Serviços do mecanismo de armazenamento

Os serviços do mecanismo de armazenamento são encapsulados em um contêiner Docker que é executado em todos os nós do ECS para oferecer um serviço distribuído e compartilhado.

3.4.2 Dados

Os principais tipos de dados armazenados no ECS podem ser resumidos da seguinte maneira:

- **Dados** — conteúdo armazenado em nível de usuário ou aplicativo, como uma imagem. Os dados são usados sinônima com objetos, arquivos ou conteúdo. Os aplicativos podem armazenar uma quantidade ilimitada de metadados personalizados com cada objeto. O mecanismo de armazenamento grava os dados e os metadados personalizados associados, oferecidos pelo aplicativo, juntos em um repositório lógico. Metadados personalizados são um sólido recurso dos sistemas modernos de armazenamento, que apresentam mais informações ou categorização sobre os dados que estão sendo armazenados. Os metadados personalizados são formatados como pares de chave-valor e recebem solicitações de gravação.
- **Metadados do sistema** — informações e atributos do sistema relacionados aos dados do usuário e recursos do sistema. Os metadados do sistema podem ser amplamente categorizados da seguinte maneira:

- **Identificadores e descritores** — um conjunto de atributos, usados internamente para identificar objetos e suas versões. Identificadores são IDs numéricos ou valores de hash que não são usados fora do contexto do software ECS. Os descritores definem informações como o tipo de codificação.
- **Chaves de criptografia em formato criptografado** — as chaves de criptografia de dados são consideradas como metadados do sistema. Elas são armazenadas em formato criptografado na estrutura de tabela do diretório principal.
- **Indicadores internos** — um conjunto de indicadores usados para monitorar se as atualizações do intervalo de bytes ou a criptografia estão ativadas, além de coordenar o armazenamento em cache e a exclusão.
- **Informações sobre localização** — conjunto de atributos com localização de índices e de dados como offsets de bytes.
- **Registros de data e hora** — conjunto de atributos que monitora o tempo, como o de criação ou de atualização de objetos.
- **Informações de configuração/tenancy** — controle de acesso a namespaces e objetos.

Os metadados dos dados e do sistema são gravados em *fragmentos* no ECS. Um fragmento do ECS é um contêiner lógico de 128 MB de espaço contíguo. Cada fragmento pode ter dados de diferentes objetos, como exibido abaixo na Figura 7. O ECS usa a indexação para monitorar todas as partes de um objeto que podem ser distribuídas entre diferentes fragmentos e nós.

Os fragmentos são gravados em um padrão somente por adição. O comportamento somente por adição significa que a solicitação de um aplicativo para modificar ou atualizar um objeto existente não modificará nem excluirá os dados gravados anteriormente em um fragmento, mas que as novas modificações ou atualizações serão gravadas em um novo fragmento. Portanto, não é necessário fazer um bloqueio de E/S nem invalidação de cache. O projeto somente por adição também simplifica a aplicação de versões dos dados. As versões antigas dos dados são mantidas em fragmentos anteriores. Se a versão do S3 estiver ativada e uma versão mais antiga dos dados for necessária, ela poderá ser recuperada ou restaurada para uma versão anterior usando a API REST do S3.

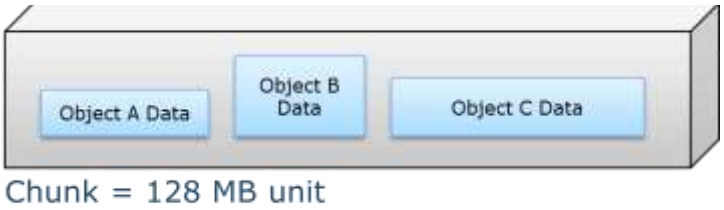

Figura 7 Fragmento de 128 MB armazenando dados de três objetos

A seção *Integridade e proteção de dados* explica como os dados são protegidos no nível do fragmento.

3.4.3 Gerenciamento de dados

O ECS usa um conjunto de tabelas lógicas para armazenar informações relacionadas aos objetos. Por fim, os pares de chave-valor são armazenados em disco, em uma árvore B+ para a indexação rápida de localizações de dados. Ao armazenar o par de chave-valor em uma árvore balanceada e pesquisada, como uma árvore B+, a localização dos dados e dos metadados pode ser acessada rapidamente. O ECS implementa uma árvore de mesclagem estruturada por logs e em dois níveis, na qual existem duas estruturas semelhantes a árvores; uma árvore menor está na memória (tabela da memória) e a principal árvore B+ reside no disco. Em primeiro lugar, a pesquisa de pares de chave-valor ocorre na memória e, depois, na principal árvore B+ do disco, se necessário. As entradas nessas tabelas lógicas, primeiramente, são registradas nos logs de registro que, por sua vez, são gravados em discos em fragmentos com espelhamento triplo. Os registros são usados para monitorar as transações que ainda não foram confirmadas na árvore B+. Depois que cada transação for registrada em um registro, a tabela na memória será atualizada. Depois que a tabela da memória ficar cheia ou depois de determinado período, a mesclagem será classificada ou submetida a dump para a árvore B+ do disco. O número de fragmentos de registro usados pelo sistema é insignificante em comparação com os fragmentos da árvore B+. A Figura 8 ilustra esse processo.

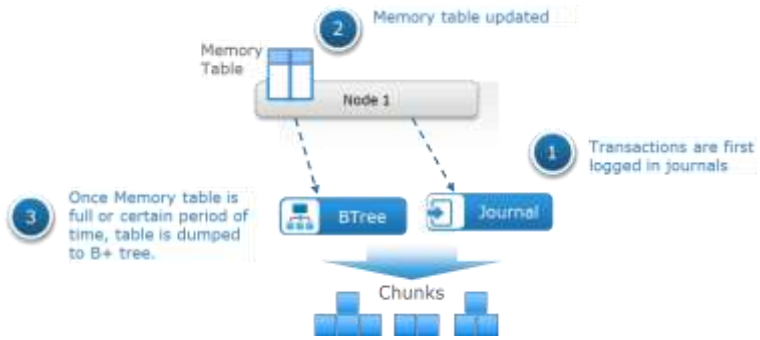

Figura 8 Tabela da memória submetida a dump para a árvore B+

As informações armazenadas na tabela de objetos (OB) são exibidas abaixo na Tabela 2. A tabela OB contém os nomes de objetos e a localização dos fragmentos, com determinado offset e comprimento nesse fragmento. Nessa tabela, o nome do objeto é a chave para o índice e o valor é a localização do fragmento. A camada de índice do Mecanismo de armazenamento é responsável pelo mapeamento de nome a fragmento dos objetos.

Tabela 2 Entradas da tabela de objetos

Nome do objeto	Localização do fragmento
ImgA	• C1:offset:length
FileB	• C2:offset:length • C3:offset:length

A tabela de fragmentos (CT) registra a localização de cada fragmento, conforme detalhado na Tabela 3.

Tabela 3 Entradas da tabela de fragmentos

ID do fragmento	Local
C1	• Node1:Disk1:File1:Offset1:Length • Node2:Disk2:File1:Offset2:Length • Node3:Disk2:File6:Offset:Length

O ECS foi projetado para ser um sistema distribuído, para que o armazenamento e o acesso dos dados sejam distribuídos entre todos os nós. As tabelas usadas para gerenciar os dados e metadados de objetos crescem com o passar do tempo, à medida que o armazenamento é usado e ampliado. As tabelas são divididas em partições e atribuídas a nós diferentes, em que cada nó se torna o proprietário das partições que hospeda para cada uma das tabelas. Para obter a localização de um fragmento, por exemplo, a tabela de registros de partição (PR) é consultada em busca do nó proprietário que conhece a localização do fragmento. Uma tabela PR básica é ilustrada na Tabela 4 abaixo.

Tabela 4 Entradas da tabela de registros de partição

ID de partição	Owner
P1	Nó 1
P2	Nó 2
P3	Nó 3

Se um nó ficar inativo, outros nós assumirão a propriedade de suas partições. As partições são recriadas lendo a raiz da árvore B+ e reproduzindo os registros armazenados em disco. A Figura 9 mostra o failover da propriedade da partição.

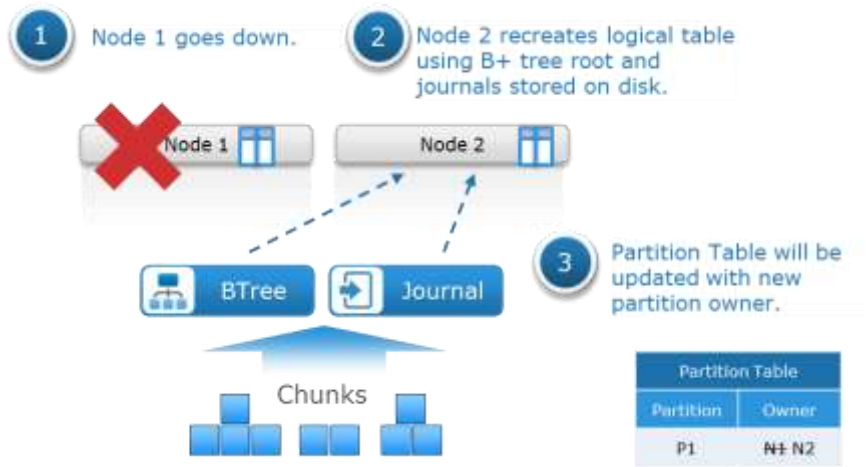

Figura 9 Failover da propriedade da partição

3.4.4 Fluxo de dados

Os serviços de armazenamento estão disponíveis a partir de qualquer nó. Os dados são protegidos por segmentos distribuídos de EC nas unidades, nós e racks. O ECS executa uma função de soma de verificação e armazena o resultado com cada gravação. Se os primeiros bytes de dados forem compactáveis, o ECS compactará os dados. Nas leituras, os dados são descompactados e a soma de verificação armazenada é validada. Apresentamos um exemplo de fluxo de dados de uma gravação em cinco etapas:

1. O client envia a solicitação de criação de objeto para um nó.
2. O nó que atende à solicitação grava os dados do novo objeto em um fragmento do repo (abreviação de repositório).
3. Na gravação bem-sucedida no disco, uma transação de PR ocorre para digitar o nome e a localização do fragmento.
4. O proprietário da partição registra a transação nos logs de registro.
5. Depois que a transação for registrada nos logs, uma confirmação será enviada ao client.

A Figura 10 abaixo mostra um exemplo de fluxo de dados para uma leitura da arquitetura da unidade de disco rígido, como Gen2 e EX300, EX500 e EX3000:

1. Uma solicitação de leitura de objeto é enviada do client ao nó 1.
2. O nó 1 utiliza uma função de hash usando o nome do objeto, para determinar qual nó é o proprietário da partição da tabela lógica em que residem as informações desse objeto. Neste exemplo, o nó 2 é o proprietário e, portanto, o nó 2 fará uma pesquisa nas tabelas lógicas para obter a localização do fragmento. Em alguns casos, a pesquisa pode ocorrer em dois nós diferentes; por exemplo, quando a localização não está armazenada em cache nas tabelas lógicas do nó 2.
3. A partir da etapa anterior, a localização do fragmento é informada ao nó 1 que, em seguida, emitirá uma solicitação de leitura de offset de byes ao nó que contém os dados (neste exemplo, o nó 3) e enviará os dados a ele.
4. O nó 1 envia os dados ao client solicitante.

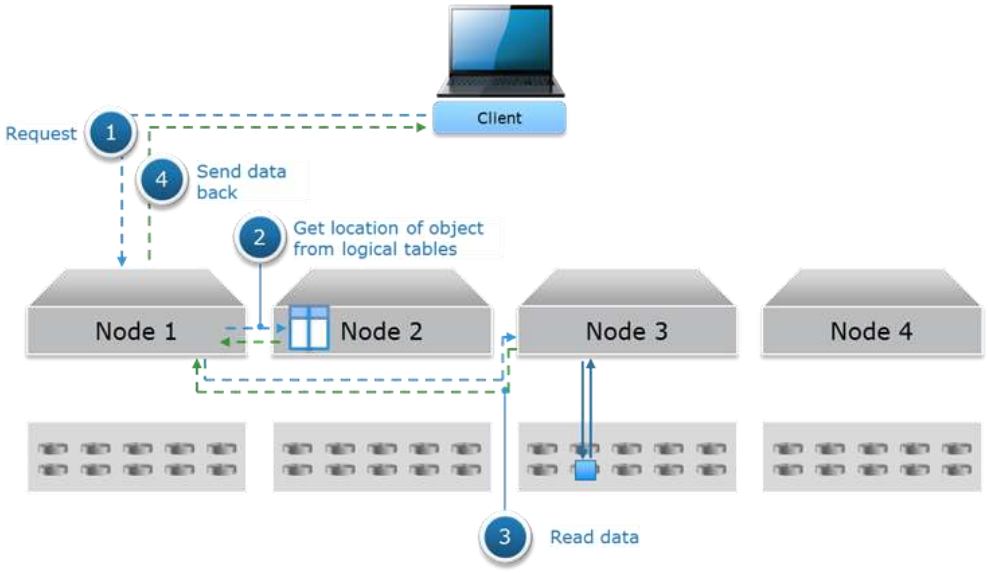

Figura 10 Fluxo de dados de leitura para arquitetura de unidade de disco rígido

A Figura 11 abaixo mostra um exemplo de fluxo de dados para uma leitura da arquitetura all-flash, como o EXF900:

1. Uma solicitação de leitura de objeto é enviada do client ao nó 1.
2. O nó 1 utiliza uma função de hash usando o nome do objeto, para determinar qual nó é o proprietário da partição da tabela lógica em que residem as informações desse objeto. Neste exemplo, o nó 2 é o proprietário e, portanto, o nó 2 fará uma pesquisa nas tabelas lógicas para obter a localização do fragmento. Em alguns casos, a pesquisa pode ocorrer em dois nós diferentes; por exemplo, quando a localização não está armazenada em cache nas tabelas lógicas do nó 2.
3. Na etapa anterior, a localização do fragmento é fornecida ao Nó 1, que, em seguida, lerá os dados diretamente do Nó 3.
4. O nó 1 envia os dados ao client solicitante.

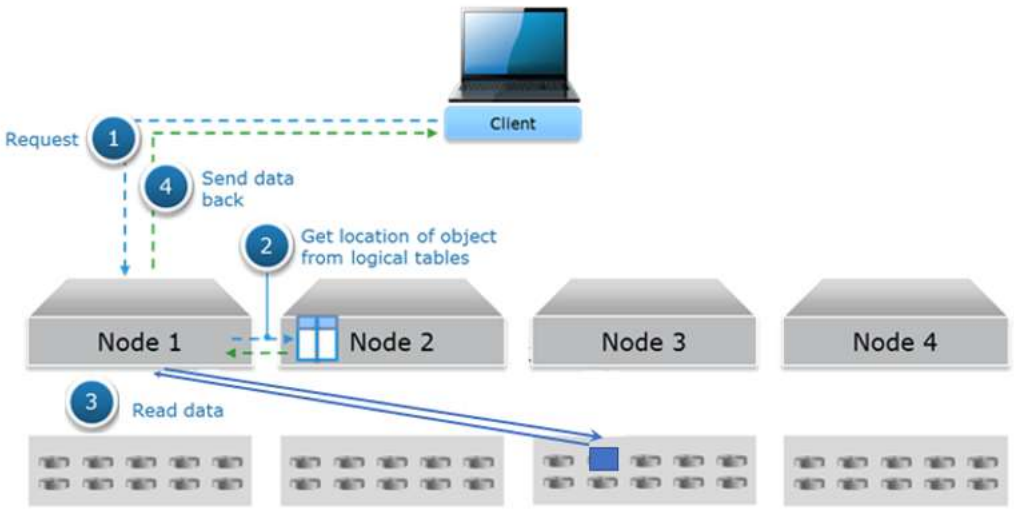

Figura 11 Fluxo de dados de leitura para arquitetura all-flash

Nota: na arquitetura all-flash, como o EXF900, cada nó pode ler dados de outro nó diretamente, diferente da arquitetura da unidade de disco rígido em que cada nó só pode ler o armazenamento de dados por conta própria.

3.4.5 Otimizações de gravação para tamanho do arquivo

Para gravações menores no armazenamento, o ECS usa um método chamado *box-carting* para minimizar o impacto sobre o desempenho. O método box-carting agrega várias gravações menores de 2 MB ou menos na memória e as grava em uma só operação de disco. O box-carting limita o número de viagens de ida e volta ao disco necessárias para processar gravações individuais.

Nas gravações de objetos maiores, os nós do ECS podem processar as solicitações de gravação para o mesmo objeto simultaneamente e aproveitar as gravações simultâneas em vários spindles do cluster do ECS. Assim, o ECS pode incluir e armazenar objetos pequenos e grandes com eficiência.

3.4.6 Recuperação de espaço

Gravar fragmentos somente por adição significa que os dados são adicionados ou atualizados mantendo-se os dados gravados originais em seu lugar e, depois, criando novos segmentos de fragmentos que podem ou não ser incluídos no contêiner de fragmentos do objeto original. O benefício da modificação de dados somente por adição é um modelo de acesso aos dados ativo/ativo, que não é prejudicado pelos problemas de bloqueio de arquivos dos file systems tradicionais. Dessa forma, à medida que os objetos são atualizados ou excluídos, os dados dos fragmentos se tornam mais referenciados ou necessários. Dois métodos de coleta de lixo usados pelo ECS para recuperar espaço de fragmentos completos descartados, ou de fragmentos que contenham uma combinação de fragmentos de objetos excluídos e não excluídos que não são mais referenciados, são:

- **Coleta de lixo normal** — quando um fragmento inteiro for lixo, recupere o espaço.
- **Coleta de lixo parcial por mesclagem** — quando um fragmento tiver 66% de lixo, recupere o fragmento mesclando as partes válidas com outros fragmentos parcialmente preenchidos para formar um novo fragmento e recupere espaço.

A coleta de lixo também foi aplicada à API de acesso a serviços de dados do ECS CAS para limpar BLOBs órfãos. Os BLOBs órfãos, que são BLOBs não referenciados identificados nos dados do CAS armazenados no ECS, estarão elegíveis para a recuperação de espaço por meio de métodos normais de coleta de lixo.

3.4.7 Armazenamento em cache de metadados de SSD

Os metadados do ECS são armazenados em árvores B. Cada árvore B pode ter entradas na memória, nas transações de registro e no disco. Para que o sistema tenha uma imagem completa de uma árvore B específica, os três locais são consultados, o que geralmente inclui várias consultas ao disco.

Para minimizar a latência das pesquisas de metadados, um mecanismo opcional de cache baseado em SSD foi implementado no ECS 3.5. O cache contém páginas de árvore B acessadas recentemente. Isso significa que as operações de leitura nas árvores B mais recentes sempre atingirão o cache baseado em SSD e evitarão viagens para discos giratórios.

Aqui estão alguns destaques do novo recurso de armazenamento em cache de metadados de SSD:

- Latência de leitura aprimorada em todo o sistema e TPS (Transactions Per Second) para arquivos pequenos
- Uma unidade flash de 960 GB por nó
- Os novos nós de rede de fabricação incluem a unidade SSD como uma opção
- Os nós de campo existentes, Gen3 e Gen2, podem receber upgrade por meio de kits de upgrade e instalação de autoatendimento
- Unidades SSD podem ser adicionadas enquanto o ECS está on-line
- Melhorias em cargas de trabalho de lógica analítica de arquivos pequenos que exigem leituras rápidas de grandes conjuntos de dados
- Todos os nós de um VDC devem ter SSDs para habilitar esse recurso

O fabric do ECS detecta quando um kit de SSD tiver sido instalado. Isso aciona o sistema para inicializar automaticamente e começar a usar a nova unidade. A Figura 12 mostra o cache do SSD ativado.

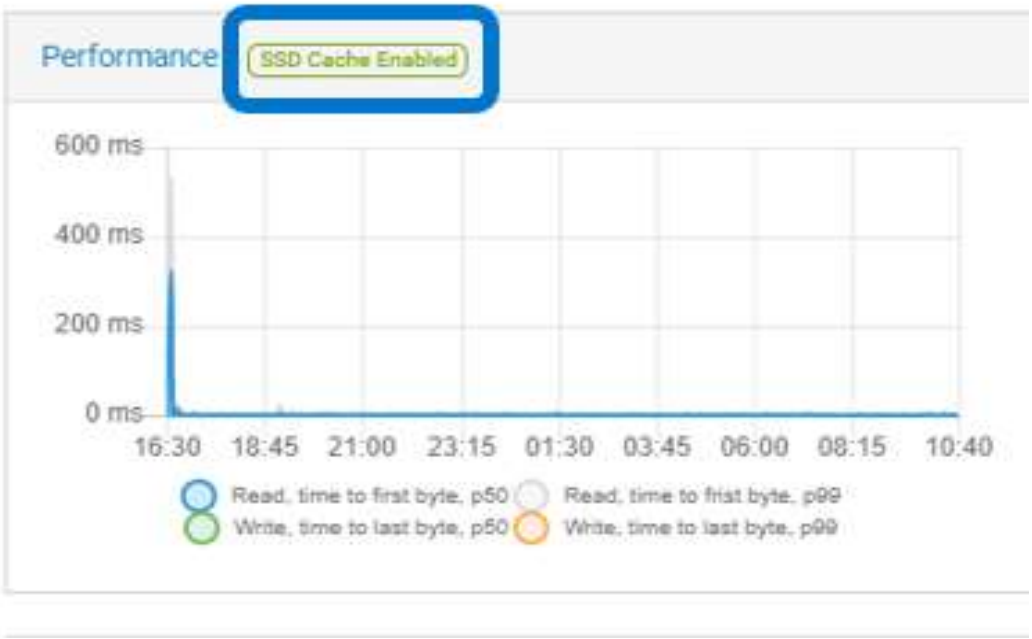

Figura 12 Cache do SSD ativado

O armazenamento em cache de metadados de SSD melhora a listagem de pequenas leituras e buckets. Conforme testamos em nosso laboratório, o desempenho de listagem melhora em 50% com objetos de 10 MB. O desempenho de leitura melhora 35% com objetos de 10 KB e 70% com objetos de 100 KB.

3.4.8 Cloud DVR

O ECS é compatível com o recurso cloud Digital Video Recording (DVR), que atende a um requisito legal de direitos autorais para empresas de cabo e satélite. O requisito é que cada unidade de gravação mapeada para um objeto no ECS precisa ser copiada um número predeterminado de vezes. O número predeterminado de cópias é conhecido como fanout. O número predeterminado de cópias (fan-out) não é realmente um requisito de redundância ou ganho de desempenho, mas sim um requisito legal de direitos autorais para empresas a cabo e satélite. O ECS permite:

- Criar o número fan-out de cópias de objeto criadas no ECS
- A leitura de uma cópia específica
- A exclusão de uma cópia específica
- A exclusão de todas as cópias
- A cópia de uma cópia específica
- A listagem de cópias
- A listagem de buckets de objetos fan-out

O recurso DVR na nuvem pode ser ativado por meio do console de serviços. Na primeira vez, você deve habilitar o recurso DVR na nuvem usando o console de serviços. Depois de habilitar o DVR na nuvem, ele é habilitado por padrão para todos os novos nós.

Execute o comando abaixo no console de serviços para ativar o recurso DVR na nuvem:

```
service-console run Enable_CloudDVR
```

O recurso DVR na nuvem permite APIs, e você pode consultar o *Guia de Acesso a Dados do ECS* para obter mais detalhes

3.5 Fabric

A camada de fabric oferece alertas e recursos de upgrade, organização por clusters, integridade do sistema, gerenciamento de software e gerenciamento de configuração. Ela é responsável por manter os serviços em execução e por gerenciar recursos como os discos, contêineres e a rede. Ela rastreia e reage às alterações no ambiente como detecção de falha e oferece alertas relacionados à integridade do sistema. A camada de fabric tem os seguintes componentes:

- **Node Agent** — gerencia os recursos de host (discos, rede, contêineres etc.) e os processos do sistema.
- **Lifecyle Manager** — gerenciamento do ciclo de vida de aplicativos, que envolve a inicialização de serviços, recuperação, notificação e detecção de falhas.
- **Persistence Manager** — coordena e sincroniza o ambiente distribuído do ECS.
- **Registry** — área de armazenamento de imagens do Docker para o software ECS.
- **Event Library** — mantém o conjunto de eventos que ocorrem no sistema.
- **Hardware Manager** — apresenta informações sobre status e eventos, e provisionamento da camada de hardware para serviços de nível superior. Esses serviços foram integrados para dar suporte a hardware genérico.

3.5.1 Node agent

O Node Agent é um agente leve e escrito em Java que é executado nativamente em todos os nós do ECS. As principais funções incluem o gerenciamento e controle de recursos do host (contêineres Docker, discos, o firewall, a rede) e o monitoramento dos processos do sistema. Exemplos de gerenciamento incluem formatação e montagem de discos, abertura das portas necessárias, garantia de que todos os processos estejam em execução e determinação das interfaces de rede pública e privada. Ele tem um fluxo de eventos que oferece os eventos solicitados a um Lifecycle Manager para indicar eventos que ocorrem no sistema. Uma CLI de fabric é útil para diagnosticar problemas e observar o estado geral do sistema.

3.5.2 Lifecycle manager

O Lifecycle Manager é executado em um subconjunto de três ou cinco nós e gerencia o ciclo de vida dos aplicativos em execução nos nós. Cada Lifecycle Manager é responsável por monitorar vários nós. O principal objetivo é gerenciar todo o ciclo de vida do aplicativo ECS, da inicialização até a implementação, inclusive detecção de falhas, recuperação, notificação e migração. Ele analisa os fluxos do Node Agent e impulsiona o agente para lidar com a situação. Quando um nó fica inativo, ele responde a falhas ou inconsistências no estado do nó restaurando o sistema a um bom estado conhecido. Se uma instância do Lifecycle Manager ficar inativa, outra assumirá seu lugar.

3.5.3 Registro

O Registry contém as imagens do ECS Docker usadas durante a instalação, o upgrade e a substituição dos nós. Um contêiner do Docker chamado *fabric-registry* é executado em um nó do rack do ECS e mantém o repositório de imagens e informações do ECS Docker necessárias para instalações e upgrades. Embora o Registry esteja disponível em um nó por vez, todas as imagens do Docker são armazenadas em cache localmente em todos os nós; portanto, qualquer um deles pode atender ao Registry.

3.5.4 Event library

A Event Library é usada na camada de fabric para expor os fluxos de eventos de ciclo de vida e do Node Agent. Os eventos gerados pelo sistema são persistentes na memória compartilhada e no disco para apresentar informações históricas sobre o estado e a integridade do sistema ECS. Esses fluxos de eventos ordenados podem ser usados para restaurar o sistema para um estado específico, reproduzindo os eventos ordenados armazenados. Alguns exemplos de eventos incluem eventos de nó, como started, stopped ou degraded.

3.5.5 Hardware manager

O Hardware Manager é integrado ao agente de fabric para oferecer suporte ao hardware padrão do setor. O principal objetivo é apresentar informações específicas sobre status e eventos, e fazer o provisionamento da camada de hardware para serviços de nível superior do ECS.

3.6 Infraestrutura

No momento, os nós do equipamento ECS executam o SUSE Linux Enterprise Server 12 para a infraestrutura. Para o software ECS implementado no hardware personalizado padrão do setor, o sistema operacional também pode ser o RedHat Enterprise Linux ou o CoreOS. As implementações personalizadas são feitas por meio de um processo formal de solicitação e validação. O Docker é instalado na infraestrutura para implementar as camadas encapsuladas do ECS. O software ECS é escrito em Java; portanto, a Java Virtual Machine é instalada como parte da infraestrutura.

3.6.1 Docker

O ECS é executado no sistema operacional como um aplicativo Java e é encapsulado em vários contêineres do Docker. Os contêineres são isolados, mas compartilham o hardware e os recursos subjacentes do sistema operacional. Algumas partes do software ECS são executadas em todos os nós e, algumas, são executadas em um ou em alguns nós. Os componentes em execução em um contêiner do Docker incluem:

- **object-main** — contém os recursos e processos relacionados aos serviços de dados, ao mecanismo de armazenamento e aos serviços de portal e de provisionamento. É executado em todos os nós do ECS.
- **fabric-lifecycle** — contém os processos, as informações e os recursos necessários para o monitoramento, gerenciamento de configuração e gerenciamento de integridade no nível do sistema. Um número ímpar de instâncias do fabric-lifecycle estará sempre em execução. Por exemplo, haverá três instâncias em execução em um sistema de quatro nós e cinco instâncias para um sistema de oito nós.
- **fabric-zookeeper** — serviço centralizado para coordenar e sincronizar processos distribuídos, informações sobre configuração, grupos e serviços de nomenclatura. É conhecido como o Persistence Manager e é executado em um número ímpar de nós, por exemplo, cinco em um sistema de oito nós.
- **fabric-registry** — registro das imagens do ECS Docker. Somente uma instância é executada por rack do ECS.

Existem outros processos e ferramentas que são executados fora de um contêiner do Docker, como o agente de nós de fabric e as ferramentas de camada de abstração de hardware. A Figura13 abaixo mostra um exemplo de como os contêineres do ECS podem ser executados em uma implementação de oito nós.

Figura13 Exemplo de agentes e contêineres do Docker em uma implementação de oito nós

A Figura 14 mostra o resultado da linha de comando para o comando *docker ps* em um nó que mostra os quatro contêineres usados pelo ECS no Docker. Uma lista é exibida com todos os serviços relacionados a objetos disponíveis no sistema.

```
admin@hop-u300-11-pub-01:~> sudo docker ps
CONTAINER ID        IMAGE                                                  COMMAND                  CREATED             STATUS
      PORTS               NAMES
7ba30ce42be2        ecs-monitoring/telegraf:3.5.0.0-825.b6b07cf9           "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                    object-telegraf
e22513635cab        ecs-monitoring/grafana:3.5.0.0-825.b6b07cf9            "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                    object-grafana
ee9db1ea40bc        emcvipr/object:3.5.0.0-120417.6a158e139f1              "/opt/vipr/boot/boot."   5 weeks ago         Up 5 weeks
                    object-main
d11a7acd55e5        ecs-monitoring/throttler:3.5.0.0-825.b6b07cf9          "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                    object-throttler
f940267979bb3       ecs-monitoring/fluxd:3.5.0.0-825.b6b07cf9              "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                    object-fluxd
c7b8530a8bb9        caspian/fabric:3.5.0.0-4076.7d40a97                    "./boot.sh lifecycle"    5 weeks ago         Up 5 weeks
                    fabric-lifecycle
bffdf8836853        caspian/fabric-zookeeper:3.5.6.0-99.0356df7            "./boot.sh 1 1=169.2."   5 weeks ago         Up 5 weeks
                    fabric-zookeeper
f4420f7f7d51        caspian/fabric-registry:2.3.1.0-68.10d1aca             "/opt/docker-registr."   5 weeks ago         Up 5 weeks
                    fabric-registry
admin@hop-u300-11-pub-01:~> sudo docker
hop-u300-11-pub-01:/ # cd /opt/storageos/
hop-u300-11-pub-01:/opt/storageos # ls bin/*svc
bin/blobsvc     bin/coordinatorsvc  bin/eventsvc  bin/objcontrolsvc  bin/storagemanagementsvc
bin/cassvc      bin/dataheadsvc     bin/filesvc   bin/objheadsvc     bin/sysvc
bin/controlsvc  bin/ecsportalsvc    bin/hdfssvc   bin/resourcesvc    bin/transformsvc
```

Figura 14 Processos, recursos, ferramentas e binários no contêiner object-main

4 Modelos de hardware do equipamento

Pontos iniciais flexíveis permitem que o ECS seja escalado rapidamente a petabytes e exabytes de dados. Com o mínimo de impacto sobre os negócios, uma solução ECS pode ser escalada linearmente em capacidade e desempenho por meio da adição de nós e discos.

Os modelos de hardware do equipamento ECS são caracterizados pela geração do hardware. A série de terceira geração do equipamento, conhecida como 3ª geração ou série EX, inclui três modelos de hardware. Esta seção apresenta uma visão geral de alto nível da série EX. Para obter detalhes completos, consulte o *Guia de Hardware do ECS Série EX*.

As informações sobre o hardware do ECS Appliance de primeira e segunda geração estão disponíveis no *Guia de hardware do Dell EMC ECS Série D e Série U*.

4.1 Série EX

Os modelos do dispositivo série EX se baseia nos servidores e comutadores padrão da Dell. As ofertas da série são:

- **EX300** — O EX300 tem uma capacidade bruta inicial de 60 TB. É a plataforma de armazenamento perfeita para aplicativos nativos da nuvem e iniciativas de transformação digital do cliente. O EX300 é ideal para modernizar implementações Centera. Ainda mais importante, o EX300 pode ser escalado de modo econômico para capacidades maiores. Fornece 12 unidades por nó e opções de disco de 1 TB, 2 TB, 4 TB, 8 TB e 16 TB (todas iguais no nó)
- **EX500** — O EX500 é o equipamento de edição mais recente, que visa a proporcionar economia com densidade. Com opções para 12 ou 24 unidades e opções de disco de 8 TB, 12 TB e 16 TB (todas iguais no nó). O cluster varia de 480 TB a 6,1 PB por rack. Essa série oferece uma opção versátil para empresas de médio porte que buscam oferecer suporte a casos de uso de aplicativos modernos e/ou arquivo morto.
- **EX3000** — O EX3000 tem uma capacidade máxima de 11,5 PB de armazenamento bruto por rack, 30 a 90 unidades por nó, discos de 12 TB ou 16 TB e pode crescer para exabytes em vários locais, oferecendo uma solução escalável de data center que é ideal para cargas de trabalho com maior espaço ocupado por dados. Esses nós estão disponíveis em duas configurações diferentes, conhecidas como EX3000S e EX3000D. O EX3000S é um chassi de nó único e o EX3000D é um chassi de nó duplo. Esses nós de alta densidade têm discos intercambiáveis. Eles começam com um mínimo de trinta discos por nó. 30 unidades por nó do ECS é o ponto em que os ganhos de desempenho com a adição de mais unidades diminuem. Com 30 ou mais unidades em cada nó como mínimo, as expectativas de desempenho são semelhantes em todos os nós do EX3000, independentemente do número de unidades.
- **EXF900** — O EXF900 é uma solução de armazenamento em objeto all-flash de nós hiperconvergentes para implementações de ECS de baixa latência e alto IOPS. Com opções para 12 ou 24 unidades, opções de unidade SSD NVMe de 3,84 TB (o driver de SSD NVMe de 7,68 TB será compatível quando o hardware estiver disponível). Essa plataforma começa em uma configuração mínima de 230 TB brutos e é dimensionada para 1,4 PB brutos por rack. A Figura 15 mostra um nó do EXF900.

Figura 15 Nó do EXF900

Nota: O recurso cache de leitura SSD não se aplica ao EXF900; o DVR em nuvem não é compatível com o EXF900; o Tech Refresh não é compatível com o EXF900; o EXF900 não pode coexistir com nenhum outro hardware não EXF900 em um VDC; o EXF900 não pode coexistir com nenhum outro hardware não EXF900 no GEO (todos os locais devem ser EXF900).

As opções de capacidade inicial da série EX permitem que os clientes comecem uma implementação do ECS com apenas a capacidade necessária e a aumentem com facilidade à medida que as necessidades mudarem no futuro. Consulte a *Specification Sheet do ECS Appliance* para obter mais detalhes sobre os equipamentos da série EX, que também detalha os equipamentos anteriores séries U e D de 2ª geração.

Não há suporte às atualizações pós-implementação dos nós da série EX. Entre eles, estão:

- Alteração da CPU.
- Ajuste da capacidade da memória.
- Upgrade do tamanho do disco rígido.

4.2 Conexões de rede do equipamento

A partir do lançamento dos dispositivos da série EX, é usado um par redundante de switches de gerenciamento dedicados de back-end. Ao mudar para um novo equipamento, agora, o ECS pode adotar um modo de configuração de comutação de front e back-end.

Os equipamentos EX300, EX500 e EX3000 usam o Dell EMC S5148F para o par de switches de front-end e para o par de switches de back-end. O equipamento EXF900 usa o Dell EMC S5248F para o par de switches de front-end e para o par de switches de back-end e o S5232F para o switch de back-end de agregação. Note que os clientes têm a opção de usar seus próprios switches de front-end em vez dos switches Dell EMC.

4.2.1 S5148F - switches públicos de front-end

Dois switches Ethernet opcionais Dell EMC S5148F de 25 GbE e 1U podem ser obtidos para a conexão de rede ou o cliente pode oferecer seu próprio par de HA de 10 GbE ou 25 GbE para a conectividade de front-end. Geralmente, os switches públicos são chamados de *hare* e *rabbit* ou apenas de front-end.

Aviso: é obrigatório ter conexões da rede do cliente com os dois switches front-end (rabbit e hare) para manter a arquitetura de alta disponibilidade do ECS Appliance. Se o cliente optar por não se conectar à rede da maneira necessária para alta disponibilidade, não haverá garantia de alta disponibilidade dos dados para o uso deste produto.

Esses comutadores oferecem 48 portas SFP28 de 25 GbE e 6 portas QSFP28 de 100 GbE. Mais detalhes sobre esses dois tipos de porta são:

- SFP28 é uma versão aprimorada de SFP+
  - SFP+ oferece suporte a até 16 GB/s e SFP28 oferece suporte a até 28 Gb/s
  - Mesmo formato
  - Compatibilidade reversa com os módulos SFP+
- QSFP28 é uma versão aprimorada de QSFP+
  - QSFP+ oferece suporte a até 4 faixas de 16 GB/s; QSFP28 oferece suporte a até 4 faixas de 28 Gb/s
    - QSFP+ agregou faixas para obter Ethernet de 40 GB/s
    - QSFP28 agregou faixas para obter Ethernet de 100 GB/s
  - Mesmo formato
  - Compatibilidade reversa com os módulos QSFP+
  - Pode ser dividido em quatro faixas individuais de SFP28

Nota: dois cabos LAG de 100 GbE são oferecidos com os switches públicos Dell EMC S5148F de 25 GbE. As organizações que oferecem seus próprios comutadores públicos devem oferecer cabos requeridos de conexão externa, LAG ou SFPs.

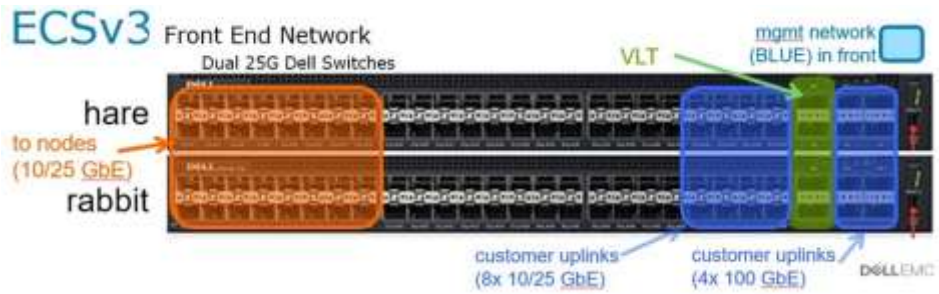

Figura 16 Designação e uso de portas de switch de rede de front-end

A Figura 16 acima oferece uma representação visual de como as portas devem ser usadas para ativar o tráfego de nós do ECS, bem como as portas de uplink do cliente. Esse é o padrão em todas as implementações.

4.2.2 S5148F - switches privados de back-end

Os switches Ethernet requeridos Dell EMC S5148F de 25 GbE e 1U com 48 portas SFP de 25 GbE e 6 portas de uplink de 100 GbE são incluídos em todos os racks do ECS. Geralmente, eles são chamados de *fox* e *hound* ou switches de back-end, e são responsáveis pela rede de gerenciamento. Nas versões futuras do ECS, os comutadores de back-end também vão oferecer separação de rede para o tráfego de replicação. O principal objetivo da rede privada é o gerenciamento e o console remotos, inicialização PXE para o gerenciador de instalação e permitir o provisionamento e gerenciamento em todo o rack e cluster. A Figura 17 mostra uma visão frontal de dois switches Dell de 25 GbE.

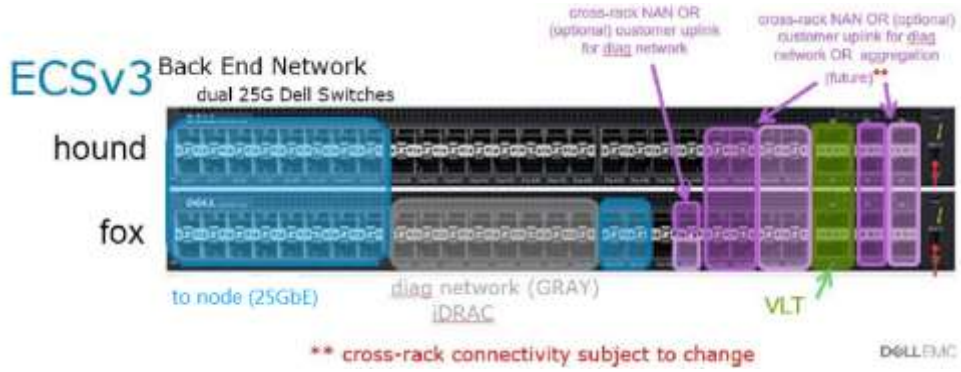

Figura 17 Designação e uso de portas de switch de rede de back-end

O diagrama acima oferece uma representação visual de como as portas devem ser usadas para ativar o tráfego de gerenciamento e as portas de diagnóstico do ECS. Essas alocações de portas são padrão em todas as implementações. As possíveis portas de uso futuro são observadas em roxo; no entanto, esse uso está sujeito a alterações no futuro.

4.2.3 S5248F - switches públicos de front-end

A Dell EMC oferece um par de HA opcional de switches front-end S5248F de 25 GbE para conexões de rede do cliente com o rack. Tem dois cabos de Virtual Link Trunking (VLT) de 200 GbE (QSFP28-DD) por par de HA. Esses switches são chamados de switches Hare e Rabbit. A Figura 18 mostra uma representação visual de como as portas devem ser usadas para ativar o tráfego de nós do ECS, bem como as portas de uplink do cliente.

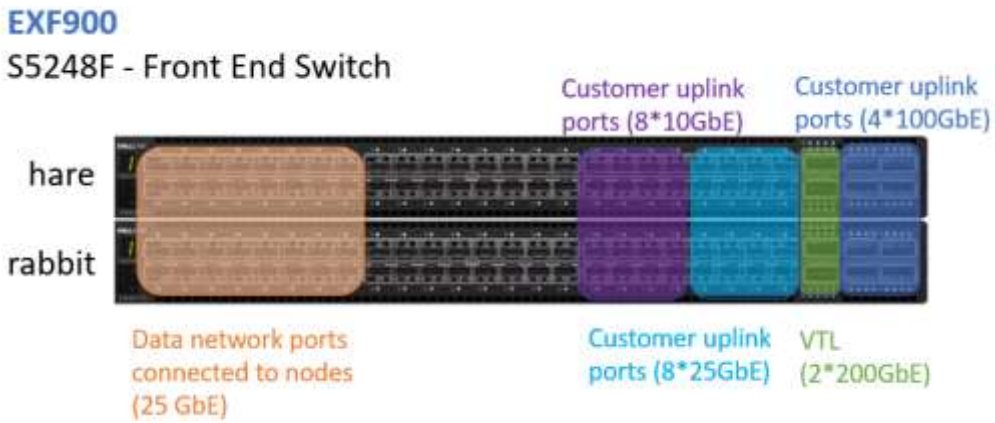

Figura 18 Designação e uso de portas de switch de rede de front-end

4.2.4 S5248F - switches privados de back-end

A Dell EMC oferece dois switches de back-end S5248F de 25 GbE com dois cabos de VLT de 200 GbE (QSFP28-DD). Esses switches são chamados de switches Hound e Fox. Todos os cabos do iDRAC de nós e todas as conexões de cabos de gerenciamento de comutadores de front-end são roteados para o comutador Fox. A Figura 19 oferece uma representação visual de como as portas devem ser usadas para ativar o tráfego de gerenciamento e as portas de diagnóstico do ECS. Essas alocações de portas são padrão em todas as implementações.

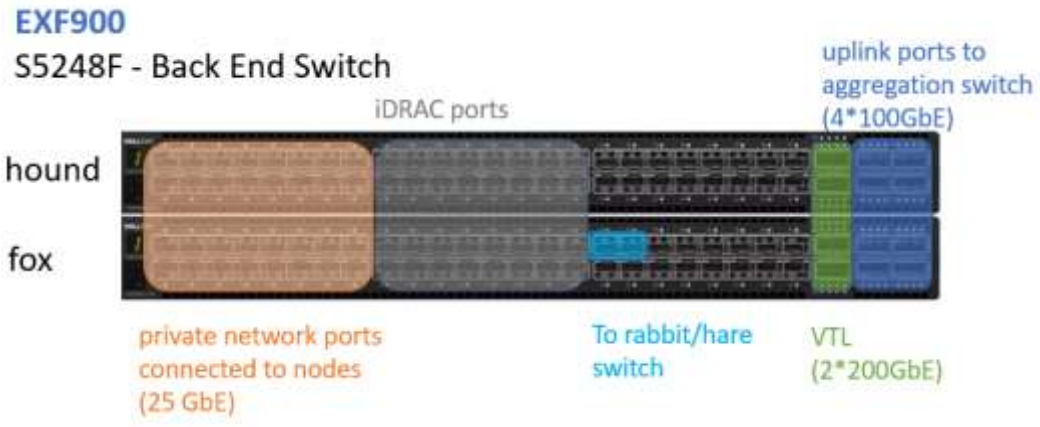

Figura 19 Designação e uso de portas de switch de rede de back-end

4.2.5 S5232 - switch de agregação

A Dell EMC oferece dois switches de agregação de back-end S5232F de 100 GbE (AGG1 e AGG2) com quatro cabos de VLT de 100 GbE. Esses switches são chamados de switches Falcon e Eagle. Na Figura 20 abaixo, todas as portas rotuladas indicam as designações das portas. Esta configuração permite conectar sete racks de nós do EXF900.

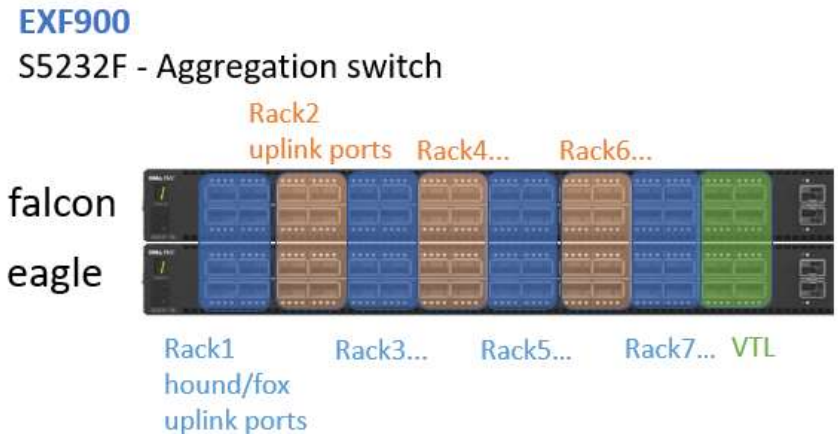

Figura 20 Designação e uso de portas de switch de rede de front-end

Para obter mais informações sobre rede e cabeamento, consulte o *Guia de Hardware do ECS Série EX*.

5 Separação de rede

O ECS oferece suporte à segregação de diferentes tipos de tráfego de rede para proporcionar isolamento de desempenho e segurança. Os tipos de tráfego que podem ser separados incluem:

- Gerenciamento
- Replicação
- Dados

Há um modo de operação chamado *modo de separação de rede*. Nesse modo, cada nó pode ser configurado no nível do sistema operacional com até três endereços IP, ou redes lógicas, para cada um dos diferentes tipos de tráfego. Esse recurso foi projetado para oferecer a flexibilidade de criar três redes lógicas separadas para gerenciamento, replicação e dados, ou combiná-las para criar duas redes lógicas, para que o gerenciamento de instâncias e o tráfego de replicação esteja em uma rede lógica e o tráfego de dados, em outra. Uma segunda rede de dados lógicos para tráfego somente de CAS pode ser configurada, permitindo a separação do tráfego de CAS de outros tipos de tráfego de dados, como S3.

A implementação de separação de rede pelo ECS exige que todo o tráfego de rede lógica seja associado aos serviços e às portas. Por exemplo, os serviços do portal do ECS se comunicam pelas portas 80 ou 443; portanto, essas portas e serviços estarão vinculados à rede lógica de gerenciamento. Uma segunda rede de dados pode ser configurada; no entanto, ela serve apenas para o tráfego de CAS. A Tabela 5 abaixo destaca os serviços fixos para um tipo de rede lógica. Para obter uma lista completa dos serviços associados às portas, consulte o mais recente *Guia de Configuração de Segurança do ECS*.

Tabela 5 Serviços para mapeamento de rede lógica

Serviços	Rede lógica	Identifier
IU na Web e API, SSH, DNS, NTP, AD, SMTP	Gerenciamento	public.mgmt
Dados do client	Dados	public.data
	Dados somente de CAS	public.data2
Dados de replicação	Replicação	public.repl
SRS (Dell EMC Secure Remote Services)	Com base no fato de o SRS Gateway da rede estar conectado	public.data ou public.mgmt

Nota: o ECS 3.6 permite o acesso aos dados do S3 na rede de dados (padrão) e na rede data2 (embora o S3 não esteja ativado por padrão na data2). Para habilitar o acesso aos dados do S3 na rede data2, o public.data é obrigatório. Entre em contato com o suporte remoto do ECS.

A separação de rede pode ser obtida logicamente usando diferentes endereços IP, virtualmente usando VLANs diferentes ou fisicamente usando cabos diferentes. O comando *setrackinfo* é usado para configurar os endereços IP e as VLANs. A configuração de VLAN no nível do comutador ou do client é responsabilidade do cliente. Para a separação da rede física, os clientes precisam enviar uma solicitação de qualificação do produto (RPQ), entrando em contato com o Dell EMC Global Business Services. Para obter mais informações sobre separação de rede, consulte o white paper *Rede e Práticas Recomendadas do ECS,* que oferece uma visualização de alto nível da separação de rede.

6 Security

A segurança do ECS é implementada nos níveis de administração, transporte e dados. A autenticação de usuários e administradores é obtida por meio do Active Directory, de métodos de LDAP, Keystone ou diretamente no portal do ECS. A segurança em nível de dados é feita por meio de HTTPS para dados em movimento e/ou de criptografia no servidor para dados em repouso.

6.1 Autenticação

O ECS oferece suporte aos métodos de autenticação do Active Directory, LDAP, Keystone e IAM para oferecer acesso para gerenciar e configurar o ECS; no entanto, existem limitações, como exibido na Tabela 6. Para obter mais informações sobre segurança, consulte o mais recente *Guia de Configuração de Segurança do ECS*.

Tabela 6 Métodos de autenticação compatíveis

Método de autenticação	**Compatível**
Active Directory	• Suporte a grupos do AD como usuários de gerenciamento • Suporte a grupos do AD para métodos de autoprovisionamento de usuários de objeto, usando chaves de autoatendimento por meio da API • Suporte a vários domínios
LDAP	• Os usuários de gerenciamento podem ser autenticados individualmente por LDAP • Os grupos do LDAP NÃO são compatíveis com usuários de gerenciamento • O LDAP é compatível com usuários de objeto (chaves de autoatendimento via API) • Suporte a vários domínios.
Keystone	• As políticas de RBAC ainda não são compatíveis. • Não há suporte a tokens sem escopo • Não há suporte a vários servidores Keystone por sistema ECS
IAM	• Oferece federação de identidade e single sign-on (SSO) por meio dos padrões SAML 2.0 • Disponível somente por meio do protocolo S3

6.2 Autenticação de serviços de dados

O acesso a objetos usando APIs RESTful é protegido por HTTPS (TLS v1.2). As solicitações recebidas são autenticadas usando métodos definidos, como código de autenticação de mensagens com base em hash (HBAC), Kerberos ou autenticação de tokens. A Tabela 7 abaixo apresenta os diferentes métodos usados para cada protocolo.

Tabela 7 Autenticação de serviços de dados

Protocolos		**Métodos de autenticação**
Objeto	S3	V2 (HMAC-SHA1), V4 (HMAC-SHA256)
	Swift	Token — Keystone v2 e v3 (com escopo, UUID, tokens de PKI), SWAuth v1
	Atmos	HMAC-SHA1
	CAS	Arquivo PEA de chave secreta
File	HDFS	Kerberos
	NFS	Kerberos, AUTH_SYS

6.3 Criptografia de dados em repouso (D@RE)

Muitas vezes, os requisitos de conformidade exigem o uso de criptografia para proteger os dados gravados nos discos. No ECS, a criptografia pode ser ativada nos níveis de namespace e bucket. Os principais recursos da D@RE do ECS são:

- Criptografia em repouso, nativa e com pouco contato — facilmente ativada, configuração simples
- CIPHERs (AES-256 CTR) usados
- Criptografia de chave pública RSA com 2.048 bits de comprimento
- Suporte a gerenciamento de chaves externas (EKM) em nível de cluster:
  - Gemalto SafeNet
  - IBM Security Key Lifecycle Manager
- Rodízio de chaves
- Suporte à semântica da criptografia S3 usando cabeçalhos HTTP, como *x-amz-server-side-encryption*
- Conformidade com FIPS 140-2 com padrões de segurança criptográfica do governo dos EUA

Nota: O modo FIPS 140-2 impõe o uso de algoritmos somente aprovados dentro D@RE; a conformidade com FIPS 140-2 é apenas para o módulo D@RE, não para todo o produto ECS.

O ECS usa uma hierarquia de chaves para criptografar e descriptografar os dados. O gerenciador de chaves nativo armazena uma chave privada comum para todos os nós a fim de descriptografar a chave principal. Na configuração do EKM, a chave principal é oferecida pelo EKM. As chaves oferecidas pelo EKM residem somente na memória do ECS. Elas nunca são armazenadas no armazenamento persistente do ECS.

Em um ambiente replicado regionalmente, quando um novo sistema do ECS se une a uma federação existente, a chave principal é extraída usando a chave pública-privada do sistema existente e criptografada usando o novo par de chave pública-privada gerado do novo sistema que se uniu à federação. A partir deste momento, a chave principal é global e conhecida para os dois sistemas na federação. Ao usar o EKM, todos os sistemas federados recuperam a chave principal do sistema de gerenciamento de chaves.

6.3.1 Rodízio de chaves

O ECS oferece suporte à alteração de chaves de criptografia. Isso pode ser feito periodicamente para limitar o volume de dados protegidos por um conjunto específico de chaves de criptografia de chaves (KEK) ou em resposta a uma possível perda ou comprometimento. Um registro de KEK de rodízio é usado em combinação com outras chaves principais para criar chaves de encapsulamento virtual para proteger as chaves de criptografia de dados (DEK) e as KEKs de namespace.

As chaves de rodízio são geradas e oferecidas nativamente e mantidas por um EKM. O ECS usa a chave de rodízio atual para criar chaves de encapsulamento virtual para proteger qualquer DEK ou KEK, independentemente de o gerenciamento de chaves ser feito de modo nativo ou externo.

Durante as gravações, o ECS encapsula a DEK gerada aleatoriamente usando uma chave de encapsulamento virtual criada usando o bucket e a chave de rodízio ativa.

Como parte do rodízio de chaves, o ECS encapsula novamente todos os registros de KEK de namespace com uma nova KEK principal virtual criada a partir da nova chave de rodízio, do contexto do segredo associado e da chave principal ativa. Isso é feito para proteger o acesso aos dados protegidos pelas chaves de rodízio anteriores.

O uso de um EKM afeta o caminho de leitura/gravação dos objetos criptografados. O rodízio de chaves permite a proteção de dados extra, usando chaves de encapsulamento virtual para DEKs e KEKs de namespace. As chaves de encapsulamento virtual não são persistentes e são geradas de duas hierarquias independentes de chaves persistentes. Com o uso do EKM, a chave de rodízio não é armazenada no ECS e adiciona mais detalhes à segurança dos dados. Principalmente, nós adicionamos os novos registros de KEK e atualizamos os IDs ativos, mas nunca excluímos nada.

Os pontos adicionais a serem considerados em relação ao rodízio de chaves no ECS são:

- O processo de rodízio de chaves altera apenas a chave de rodízio atual. A chave principal existente e as chaves de namespace e bucket não são alteradas durante o processo de rodízio de chaves.
- Não há suporte ao rodízio de chaves de namespace ou de bucket; no entanto, o escopo de rodízio está no nível do cluster, de modo que todos os novos objetos criptografados do sistema serão afetados.
- Os dados existentes não são criptografados novamente por meio do rodízio de chaves.
- O ECS não oferece suporte ao rodízio de chaves durante falhas elétricas.
  - TSO durante o rodízio: a tarefa de rodízio de chaves será suspensa, até que o sistema saia da TSO.
  - A PSO está em andamento. O ECS deve sair de uma PSO para que o rodízio de chaves seja ativado. Se uma PSO ocorrer durante o rodízio, o rodízio apresentará falha imediatamente.
- A criptografia de bucket não é necessária para a criptografia de objetos por meio do S3.
- Os metadados indexados de objetos do client utilizados como uma chave de pesquisa não são criptografados.

Consulte o mais recente *Guia de Configuração de Segurança do ECS* para obter mais informações sobre D@RE, EKM e rodízio de chaves.

6.4 IAM de ECS

O Gerenciamento de acesso e identidades (IAM) do ECS permite que você tenha controle e acesso seguro aos recursos do ECS S3. Essa funcionalidade garante que cada solicitação de acesso a um recurso do ECS seja identificada, autenticada e autorizada. O IAM de ECS permite que o administrador adicione usuários, funções e grupos. O administrador também pode restringir o acesso adicionando políticas às entidades do IAM de ECS.

Nota: o IAM de ECS é para uso somente com S3. Ele não permite buckets habilitados para CAS ou filesystem.

O IAM de ECS apresenta os componentes a seguir:

- **Gerenciamento de contas** — permite gerenciar identidades IAM em cada namespace, como usuários, grupos e funções
- **Gerenciamento de acessos** — o acesso é gerenciado criando políticas e anexando-as a identidades ou recursos do IAM
- **Federação de identidade** — a identidade é estabelecida e autenticada pelo SAML (Security Assertion Markup Language). Depois que a identidade for estabelecida, você usará o Secure Token Service para obter credenciais temporárias que serão usadas para acessar o recurso
- **Secure Token Service** — permite solicitar credenciais temporárias para acesso entre contas aos recursos e também para usuários autenticados usando a autenticação SAML de um provedor de identidade empresarial ou serviço de diretório

Ao usar o IAM, você pode controlar quem é autenticado e autorizado a usar os recursos do ECS ao criar e gerenciar

- **Usuários** — um usuário do IAM representa uma pessoa ou um aplicativo no namespace que pode interagir com os recursos do ECS
- **Grupos** — um grupo do IAM é um conjunto de usuários do IAM. Use os grupos para especificar permissões para um conjunto de usuários do IAM
- **Funções** — a função do IAM é uma identidade que pode ser assumida por qualquer pessoa que exija a função. Uma função é semelhante a um usuário, é uma identidade com políticas de permissão que determinam o que a identidade pode e não pode fazer
- **Políticas** — Uma política de IAM é um documento em formato JSON que define as permissões para uma função. Atribua e anexe políticas a usuários do IAM, grupos de IAM e funções do IAM.
- **Provedor SAML** — SAML (Security Assertion Markup Language) é um padrão aberto para a troca de autenticação e dados de autorização entre um provedor de identidade e um provedor de serviços. O provedor SAML no ECS é usado para estabelecer confiança entre um provedor de identidade (IdP) compatível com SAML e o ECS

Cada sistema ECS é alocado com uma conta IAM do ECS. Essa conta dá suporte a vários namespaces e tem entidades de IAM relacionadas que são definidas no próprio namespace.

- Os namespaces individuais dão suporte ao gerenciamento de contas usando as entidades IAM do ECS, como usuários, funções e grupos.
- Políticas, permissões, lista de controle de acesso (ACL) associadas às entidades do IAM de ECS e os recursos do ECS S3 dão suporte ao gerenciamento do acesso aos recursos do IAM de ECS.
- O IAM de ECS dá suporte ao acesso entre contas usando Security Assertion Markup Language (SAML) e funções.
- IAM do ECS é compatível com a chave de acesso da AWS (Amazon Web Services) para acessar IAM e S3 no ECS.

Consulte o *Guia de Segurança do ECS* mais recente para obter mais informações sobre IAM do ECS

6.5 Object tagging

A marcação de objetos permite a categorização dos objetos, atribuindo etiquetas a objetos individuais. Um único objeto pode ter várias etiquetas associadas a ele, habilitando a categorização multidimensional.

Uma etiqueta pode descrever algum tipo de informações confidenciais, como um registro de integridade, ou você pode marcar um objeto em relação a determinado produto que pode ser categorizado como confidencial. A marcação é um sub-recurso de um objeto que tem um ciclo de vida integrado às operações do objeto. Você pode adicionar etiquetas a novos objetos ao carregá-los ou adicionar etiquetas a objetos existentes. É aceitável usar etiquetas para rotular objetos que contêm dados confidenciais, como Informações de identificação pessoal (PII) ou Informações de saúde protegidas (PHI). As etiquetas não devem conter informações confidenciais, pois podem ser visualizadas sem ter a permissão de leitura real de um objeto.

6.5.1 Informações adicionais sobre a marcação de objetos

Esta seção fornece informações sobre a marcação de objetos no IAM, marcação de objetos com políticas de bucket, manuseio de marcação de objetos durante TSO/PSO e marcação de objetos durante o gerenciamento do ciclo de vida do objeto. Considerações adicionais:

- Marcação de objetos no IAM
  - A função principal da marcação de objetos como sistema de categorização é fornecida quando está integrada às políticas do IAM. Isso permite que o administrador configure permissões de usuário específicas. Por exemplo, o administrador pode adicionar uma política que permite que todos acessem objetos com uma determinada etiqueta. Outra opção é configurar e conceder permissões aos usuários, que podem gerenciar as etiquetas em objetos específicos. Outro aspecto importante da marcação de objetos é como e onde as etiquetas são conservadas. Isso é importante, porque tem um impacto direto sobre vários aspectos do sistema.
- Marcação de objetos com políticas de bucket
  - A marcação de objetos permite que você categorize os objetos. Além disso, a marcação é integrada a várias políticas. A política de gerenciamento do ciclo de vida permite a configuração em nível de bucket. As versões anteriores do ECS são compatíveis com Expiration, Abort Incomplete Uploads e Deletion of Expired Object Tagging Delete Marker. O filtro pode incluir várias condições, até mesmo uma condição baseada em etiqueta. Cada etiqueta na condição de filtro precisa corresponder à chave e ao valor.
- Marcação de objetos durante TSO/PSO
  - A marcação de objetos é outro conjunto de entradas nos metadados do sistema. Nenhum manuseio especial é obrigatório durante TSO/PSO. Há um limite definido no número de etiquetas que podem ser associadas a cada objeto, no tamanho dos metadados do sistema junto com a marcação do objeto e também nos limites de memória.
- Marcação de objetos durante o gerenciamento do ciclo de vida do objeto
  - A marcação de objetos faz parte dos metadados do sistema e é manuseada simultaneamente com o manuseio de metadados do sistema, durante o gerenciamento do ciclo de vida. A Lógica de expiração e o Scanner de exclusão do ciclo de vida exigem o conhecimento das políticas baseadas em etiquetas. As etiquetas de objeto habilitam o gerenciamento do ciclo de vida do objeto refinado, no qual você pode especificar um filtro baseado em etiquetas, além de um prefixo de nome de chave, em uma regra de ciclo de vida.

Consulte o mais recente *Guia de Configuração de Segurança do ECS* para obter mais informações sobre marcação de objetos.

7 Integridade e proteção dos dados

Para a integridade dos dados, o ECS utiliza somas de verificação. As somas de verificação são criadas durante as operações de gravação e são armazenadas com os dados. As somas de verificação na leitura são calculadas e comparadas com a versão armazenada. Uma tarefa de segundo plano verifica proativamente as informações de soma de verificação.

Para proteção dos dados, o ECS utiliza espelhamento triplo para fragmentos de registro e esquemas separados de EC para fragmentos de *repo* (dados de repositório do usuário) e *btree* (árvore B+).

A codificação de eliminação oferece proteção de dados aprimorada a partir de uma falha de disco, nó e rack, de uma forma eficiente em termos de armazenamento, em comparação com os esquemas convencionais de proteção. O mecanismo de armazenamento do ECS implementa a correção de erros Reed Solomon usando dois esquemas:

- 12+4 (padrão) — o fragmento é dividido em 12 segmentos de dados. Quatro segmentos de codificação (paridade) são criados.
- 10+2 (arquivamento estático) — o fragmento é dividido em 10 segmentos de dados. Dois segmentos de codificação são criados.

Usando o padrão de 12+4, os 16 segmentos resultantes são dispersos entre os nós do site local. Os dados e os segmentos de codificação de cada fragmento são distribuídos igualmente entre os nós do cluster. Por exemplo, com 8 nós, cada nó tem 2 segmentos (de um total de 16). O mecanismo de armazenamento pode reconstruir um fragmento a partir de qualquer um dos 12 segmentos do total de 16.

O ECS requer um mínimo de seis nós para a opção de arquivamento estático, na qual um esquema de 10+2 é usado em vez do 12+4. A EC é interrompida quando o número de nós cai abaixo do mínimo necessário para o esquema de EC.

Quando um fragmento está cheio ou depois de um período definido, ele é selado, a paridade é calculada e os segmentos de codificação são gravados nos discos do domínio de falha. Os dados dos fragmentos permanecem como uma cópia única que consiste em 16 segmentos (12 dados, 4 códigos) dispersos em todo o cluster. O ECS usa apenas os segmentos de código para a reconstrução de fragmentos quando ocorre uma falha.

Quando a infraestrutura subjacente de um VDC for alterada no nível do nó ou do rack, as camadas de fabric detectarão a alteração e acionarão um scanner de rebalanceamento como uma tarefa de segundo plano. O scanner calcula o melhor layout para os segmentos de EC nos domínios de falha para cada fragmento, usando a nova topologia. Se o novo layout oferecer uma melhor proteção que o layout existente, o ECS redistribuirá os segmentos de EC em uma tarefa de segundo plano. Essa tarefa causa impacto mínimo sobre o desempenho do sistema; no entanto, haverá um aumento no tráfego entre os nós durante o rebalanceamento. Também ocorrerá o balanceamento das partições da tabela lógica para os novos nós e, daqui em diante, os fragmentos recém-criados de registro e árvore B+ serão alocados igualmente nos nós novos e antigos. A redistribuição aprimora a proteção local, aproveitando todos os recursos da infraestrutura.

Nota: recomenda-se não aguardar até que a plataforma de armazenamento esteja totalmente completa para adicionar unidades ou nós. Um limite razoável de utilização do armazenamento é de 70%, considerando-se a taxa de inclusão diária e o tempo esperado para solicitação, entrega e integração das unidades/nós adicionados.

7.1 Conformidade

Para atender aos requisitos de conformidade do setor e corporativos (norma 17a-4(f) da SEC) para armazenamento de dados, o ECS implementou os seguintes elementos:

- **Fortalecimento da plataforma** — o fortalecimento aborda as vulnerabilidades de segurança do ECS, como o bloqueio da plataforma para desativar o acesso a nós ou ao cluster, todas as portas não essenciais (por exemplo, *ftpd, sshd*) são fechadas, log de auditoria completo para comandos sudo e suporte a SRS (Dell EMC Secure Remote Services) para desligar o acesso remoto aos nós.
- **Relatórios de conformidade** — um agente do sistema informa o status de conformidade do sistema, em que *Good* indica a conformidade e *Bad* indica a não conformidade.
- **Retenção de registros e regras baseadas em políticas** — capacidade de limitar as alterações dos registros ou aos dados sob retenção usando políticas, período e regras.
- **Advanced Retention Management (ARM)** — para atender aos requisitos de conformidade do Centera, um conjunto de regras de retenção foi definido somente para CAS.
  - **Retenção baseada em eventos** — permite períodos de retenção que começam quando o evento especificado ocorre.
  - **Retenção legal** — permite a prevenção temporária contra a exclusão de dados sujeita a ações judiciais.
  - **Controle mín./máx.** — configuração por bucket do período mínimo e máximo padrão de retenção.

A conformidade é ativada no nível do namespace. Os períodos de retenção são configurados no nível do bucket. Os requisitos de conformidade certificam a plataforma e, por isso, o recurso de conformidade só está disponível para o ECS em execução no hardware do equipamento. Para obter informações sobre como ativar e configurar a conformidade no ECS, consulte o atual *Guia de Acesso a Dados do ECS* e o mais recente *Guia do Administrador do ECS*.

8 Implementação

O ECS pode ser implementado como uma instância de um ou vários locais. Os componentes modulares de uma implementação do ECS incluem:

- **Data center virtual (VDC)** — um cluster, também conhecido geralmente como local ou região geograficamente distinta, composto por um conjunto de infraestrutura do ECS gerenciado por uma só instância de fabric.
- **Pool de armazenamento (SP)** — podemos pensar nos SPs como um subconjunto de nós e seu armazenamento associado que pertence a um VDC. Um nó pode pertencer somente a um SP. A EC é configurada no nível de SP, com um esquema de 12+4 ou 10+2. Um SP pode ser usado como uma ferramenta para separar fisicamente os dados entre clients ou grupos de clients que acessam o armazenamento no ECS.
- **Grupo de replicação (RG)** — os RGs definem onde o conteúdo do SP é protegido e os locais de onde os dados podem ser acessados. Um RG com um só local membro, às vezes, é chamado de RG local. Os dados são sempre protegidos localmente, onde são gravados em relação a falhas de disco, nó e rack. Geralmente, os RGs com dois ou mais são chamados de RGs globais. Os RGs globais abrangem até 8 VDCs e oferecem proteção contra falhas de disco, nó, rack e local. Um VDC pode pertencer a vários RGs.
- **Namespace** — um namespace é conceitualmente o mesmo que um grupo de usuários no ECS. A principal característica de um namespace é que os usuários de um namespace não podem acessar os objetos de outro namespace.
- **Buckets** — os buckets são contêineres de objetos criados em um namespace e, às vezes, são considerados um contêiner lógico para subgrupos de usuários. No S3, os contêineres são chamados de buckets, um termo que foi adotado pelo ECS. No Atmos, o equivalente a um bucket é um subtenant; no Swift, o equivalente a um bucket é um contêiner e, no CAS, um bucket é um pool do CAS. Os buckets são recursos globais do ECS. Cada bucket é criado em um namespace e cada namespace é criado em um RG.

O ECS aproveita os seguintes sistemas de infraestrutura:

- **DNS** — (obrigatório) pesquisas diretas e inversas exigidas para cada nó do ECS.
- **NTP** — (obrigatório) servidor Network Time Protocol.
- **SMTP** - (opcional) Simple Mail Transfer Protocol Server para envio de alertas e geração de relatórios.
- **DHCP** — (opcional) necessário ao atribuir endereços IP via DHCP.
- **Provedores de autenticação** — (opcional) os administradores do ECS podem ser autenticados usando grupos do LDAP e Active Directory. Os usuários de objeto podem ser autenticados usando o Keystone. Os provedores de autenticação não são necessários para o ECS. O ECS tem funcionalidade de gerenciamento de usuários locais integrada; no entanto, observe que os usuários criados localmente não são replicados entre os VDCs.
- **Balanceador de carga** — (obrigatório se exigido pelo fluxo de trabalho; caso contrário, é opcional) a carga do client deve ser distribuída entre os nós para utilizar efetivamente todos os recursos disponíveis no sistema. Se um equipamento ou serviço dedicado de balanceador de carga for necessário para gerenciar a carga entre nós do ECS, ele deverá ser considerado um requisito. Os desenvolvedores que escrevem aplicativos usando o SDK ECS S3 podem aproveitar a funcionalidade integrada de balanceador de carga. Os balanceadores de carga sofisticados podem considerar fatores adicionais, como a carga informada de um servidor, tempos de resposta, status ativo/inativo, número de conexões ativas e localização geográfica. O cliente é responsável pelo gerenciamento do tráfego do client e pela determinação dos requisitos de acesso. Independentemente do método, existem algumas opções básicas que, em geral, são consideradas, como a alocação manual de IP, rodízio do DNS, balanceamento de carga no client, equipamentos de balanceador de carga e balanceadores de carga regionais. A seguir, há breves descrições de cada um desses métodos:
  - **Alocação manual de IP** — os endereços IP são distribuídos manualmente aos aplicativos. Geralmente, isso não é recomendado, pois pode não distribuir a carga nem oferecer tolerância a falhas.

- **Rodízio do DNS** — uma entrada de DNS é criada e inclui todos os endereços IP dos nós. Os clients consultam o DNS para resolver nomes de domínio completos dos serviços do ECS e são respondidos com os endereços IP de um nó aleatório. Isso pode oferecer um pseudobalanceamento de carga. Esse método pode não oferecer tolerância a falhas pois, geralmente, a intervenção manual é usada para remover endereços IP de nós com falha do DNS. Problemas de time-to-Live (TTL) podem ser encontrados com esse método. Algumas implementações do servidor DNS podem armazenar em cache as pesquisas do DNS por um período; assim, os clients que se conectam em um intervalo próximo podem se vincular ao mesmo endereço IP, reduzindo o volume de distribuição de carga para os nós de dados. Não é recomendável usar o DNS para distribuir o tráfego no método de rodízio.
- **Balanceamento de carga** — os balanceadores de carga são a abordagem mais comum para distribuir a carga do client. Os clients podem enviar o tráfego a um balanceador de carga que o recebe e o encaminha a um nó íntegro do ECS. As verificações de integridade proativas ou o estado da conexão são usados para verificar a disponibilidade de cada nó para os chamados. Os nós indisponíveis são removidos do uso até que transmitam uma verificação de integridade. O descarregamento do processamento SSL com uso intenso de CPU pode ser usado para liberar esse recurso no ECS.
- **Balanceamento de carga regional** — aproveita o DNS para rotear as pesquisas a um equipamento como o Riverbed SteelApp, por exemplo, que usa o IP regional ou outro mecanismo para determinar o melhor local ao qual fazer o roteamento do client.

8.1 Implementação em local único

Durante uma implementação inicial em um só local ou um só cluster, primeiramente, os nós são adicionados a um SP. Os SPs são contêineres lógicos de nós físicos. A configuração do SP envolve a seleção do número mínimo necessário de nós disponíveis e a seleção do esquema de EC padrão de 12+4 ou de arquivamento estático de 10+2. Os níveis de alerta críticos podem ser definidos durante a configuração inicial do SP; no entanto, no futuro, o esquema de EC não poderá ser alterado após a inicialização do SP. O primeiro SP criado é designado como o SP do sistema e é usado para armazenar os metadados do sistema. O SP do sistema não pode ser excluído.

Geralmente, os clusters contêm um ou dois SPs, como exibido na Figura 21, um para cada esquema de EC; no entanto, se uma organização exigir a separação física dos dados, SPs adicionais serão usados para implementar os limites.

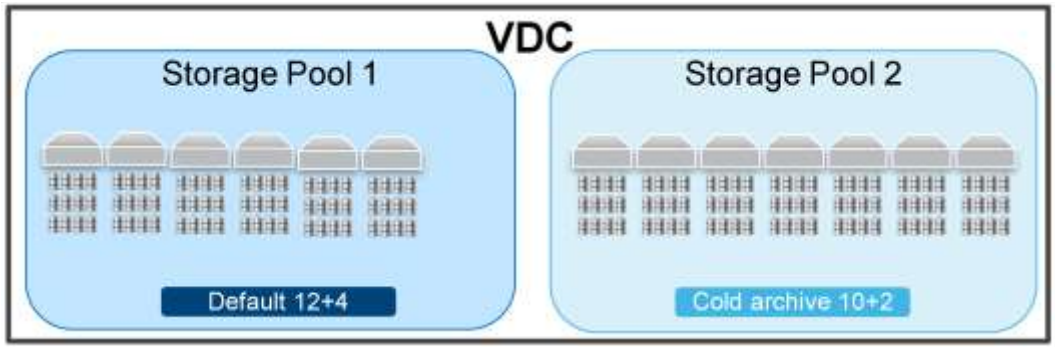

Figura 21 VDC com dois pools de armazenamento, cada um configurado com um esquema de EC diferente

Após a inicialização do primeiro SP, um VDC pode ser criado. A configuração do VDC envolve a designação de endpoints de replicação e gerenciamento. Observe que, embora a inicialização do SP do sistema seja necessária antes da criação de VDC, a configuração do VDC não atribui SPs, mas sim os endereços IP dos nós.

Depois que um VDC é criado, os RGs são configurados. RGs são recursos globais com uma configuração que envolve designar pelo menos um VDC, por si só, na instalação do local inicial ou único, junto a um dos SPs do VDC. Um RG com um só VDC membro protege os dados localmente no nível do disco, nó e rack. A próxima seção apresenta mais detalhes sobre os RGs, incluindo implementações em vários locais.

Namespaces são recursos globais criados e atribuídos a um RG. Nas políticas de retenção em nível de namespace, são definidos os administradores de namespace, cotas e conformidade. O acesso durante a falha elétrica (ADO) pode ser configurado no nível de namespace, que é abordado na próxima seção. Geralmente, ele fica no nível de namespace, em que os tenants são organizados. Tenants podem ser uma instância de aplicativo, um grupo de negócios, usuários ou equipes, ou qualquer outro agrupamento que faça sentido para a organização.

Buckets são recursos globais que podem abranger vários locais. A criação de buckets envolve a atribuição deles a um namespace e a um RG. No nível do bucket, a propriedade e o acesso ao CAS ou aos arquivos são ativados. A Figura 22 abaixo mostra um SP de um VDC, com um namespace que contém dois buckets.

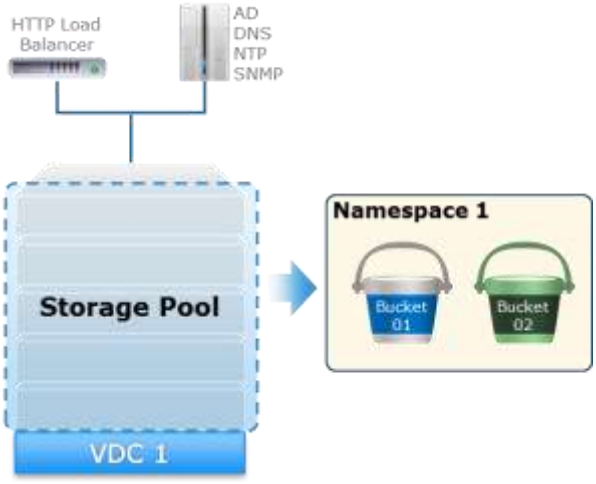

Figura 22 Exemplo de implementação em local único

8.2 Implementação em vários locais

Uma implementação em vários locais, também conhecida como um ambiente federado ou ECS federado, pode abranger até oito VDCs. Os dados são replicados no ECS no nível de fragmentos. Os nós que participam de um RG enviam seus dados locais de modo assíncrono para um ou todos os outros locais. Os dados são criptografados usando AES256, antes de serem enviados pela WAN por HTTP. Os principais benefícios reconhecidos ao fazer a federação de vários VDCs são:

- Consolidação de iniciativas de gerenciamento de vários VDCs em um só recurso lógico
- Proteção no nível do local, além de localmente em nível de nó, disco e rack
- Acesso regionalmente distribuído ao armazenamento, de modo altamente consistente e ativo em todos os lugares

Esta seção sobre a implementação em vários locais descreve os recursos específicos do ECS federado, como:

- **Consistência de dados** — por padrão, o ECS oferece um serviço de armazenamento altamente consistente.
- **Grupos de replicação** — contêineres globais usados para designar limites de acesso e proteção.
- **Armazenamento regional em cache** — otimização de fluxos de trabalho de acesso a site remoto em implementações em vários locais.
- **ADO** — comportamento de acesso do client durante uma interrupção temporária do local (TSO).

8.2.1 Consistência de dados

O ECS é um sistema altamente consistente que usa a propriedade para manter uma versão autorizada de cada namespace, bucket e objeto. A propriedade é atribuída ao VDC em que o namespace, o bucket ou o objeto é criado. Por exemplo, se um namespace NS1 for criado no VDC1, o VDC1 será proprietário do NS1 e será responsável por manter a versão autorizada dos buckets dentro do NS1. Se um bucket B1 for criado no VDC2 dentro do NS1, o VDC2 será proprietário do B1 e será responsável por manter a versão autorizada do conteúdo do bucket, bem como do VDC proprietário de cada objeto. Da mesma forma, se um objeto O1 for criado dentro do B1 no VDC3, o VDC3 será proprietário do O1 e será responsável por manter a versão autorizada do O1 e dos metadados associados.

A resiliência da proteção de dados em vários locais ocorre à custa da maior sobrecarga de proteção de armazenamento e do maior consumo de largura de banda da WAN. As consultas do índice são necessárias quando um objeto é acessado ou atualizado a partir de um local que não é proprietário do objeto. Da mesma forma, as pesquisas de índice na WAN também são necessárias para recuperar informações, como uma lista autorizada de buckets em um namespace ou de objetos em um bucket, que são propriedade de um site remoto.

Entender como o ECS usa a propriedade para monitorar os dados de maneira autorizada no nível de namespace, bucket e objeto ajuda os administradores e proprietários de aplicativos a tomar decisões sobre a configuração de seu ambiente para acesso.

8.2.2 Grupo ativo de replicação

Durante a criação do RG, uma configuração *Replicate to All Sites* está disponível; ela é desativada, por padrão, ou pode ser ativada, o que habilita esse recurso. A replicação de dados em todos os locais significa que os dados gravados individualmente em cada VDC são replicados em todos os outros VDCs membros do RG. Por exemplo, uma instância de X número de locais do ECS federado com um RG ativo configurado para replicar dados em todos os locais resultará em X vezes de sobrecarga de proteção, ou X * 1,33 (ou 1,2 na EC de arquivamento estático) de sobrecarga total de proteção de dados. A replicação em todos os locais pode fazer sentido especialmente para conjuntos de dados menores, em que o acesso local é importante. A desativação dessa configuração significa que todos os dados gravados em cada VDC serão replicados em outro VDC. O local principal, onde o objeto é criado, e o local que armazena a cópia replicada, protegem os dados localmente usando o esquema de EC atribuído ao SP local. Ou seja, somente os dados originais são replicados na WAN, e não quaisquer segmentos associados de codificação de EC.

Os dados armazenados em um RG ativo podem ser acessados pelos clients por meio de qualquer VDC membro do RG disponível. A Figura 23 abaixo mostra um exemplo de um ECS federado criado usando VDC1, VDC2 e VDC3. Dois RGs são exibidos: o RG1 tem um só membro, VDC1, e o RG2 tem todos os três VDCs como membros. Três buckets são exibidos: B1, B2 e B3.

Neste exemplo, os clients que estão acessando:

- O VDC1 têm acesso a todos os buckets
- O VDC2 e VDC3 têm acesso apenas aos buckets B2 e B3.

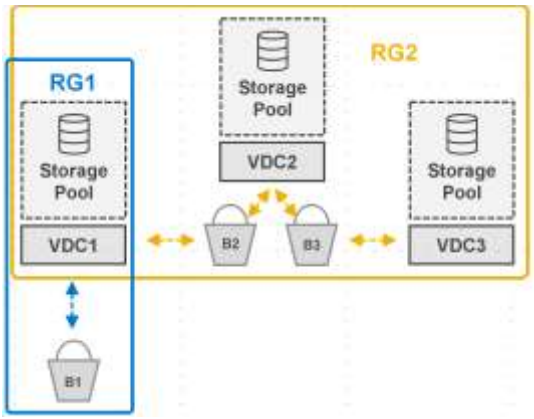

Figura 23 Acesso em nível de bucket, por local, com grupos de replicação em um e vários locais

8.2.3 Grupo passivo de replicação

Um RG passivo tem três VDCs membros. Dois dos VDCs são designados como ativos e podem ser acessados pelos clients. O terceiro VDC é designado como passivo e usado somente como destino de replicação. O local passivo é usado somente para fins de recuperação e não permite acesso direto ao client. Os benefícios da replicação regionalmente passiva são:

- Diminuição da sobrecarga da proteção de armazenamento, aumentando o potencial de operações de XOR
- Controle, em nível de administrador, da localização usada para o armazenamento somente de replicação

Figura 24 mostra um exemplo de uma configuração regionalmente passiva, em que o VDC 1 e o VDC 2 são locais principais (de origem) que replicam seus dados (fragmentos) no destino de replicação, VDC 3.

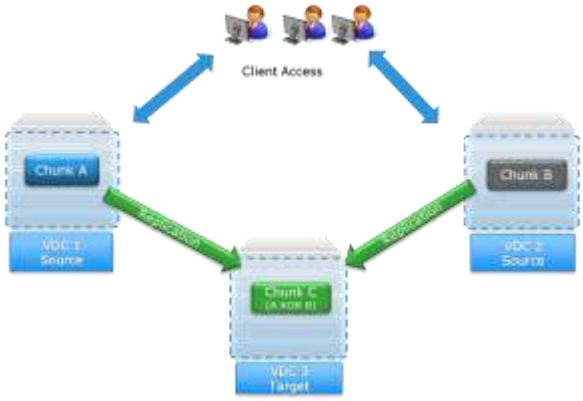

Figura 24 Caminhos de replicação e acesso do client para o grupo de replicação regionalmente passiva

O acesso por vários locais aos dados altamente consistentes é feito usando a propriedade de namespace, bucket e objeto entre os locais membros do RG. As consultas de índice na WAN entre locais são necessárias quando o acesso à API é originado de um VDC que não é proprietário das construções lógicas necessárias. As pesquisas na WAN são usadas para determinar a versão autorizada dos dados. Assim, se um objeto criado no local 1 for lido a partir do local 2, uma pesquisa na WAN será necessária para consultar o VDC proprietário do objeto, local 1, para verificar se os dados do objeto que foram replicados no local 2 são a versão mais recente dos dados. Se o local 2 não tiver a versão mais recente, ele buscará os dados necessários do local1; caso contrário, usará os dados replicados anteriormente nele. Isso é ilustrado na Figura 25 abaixo.

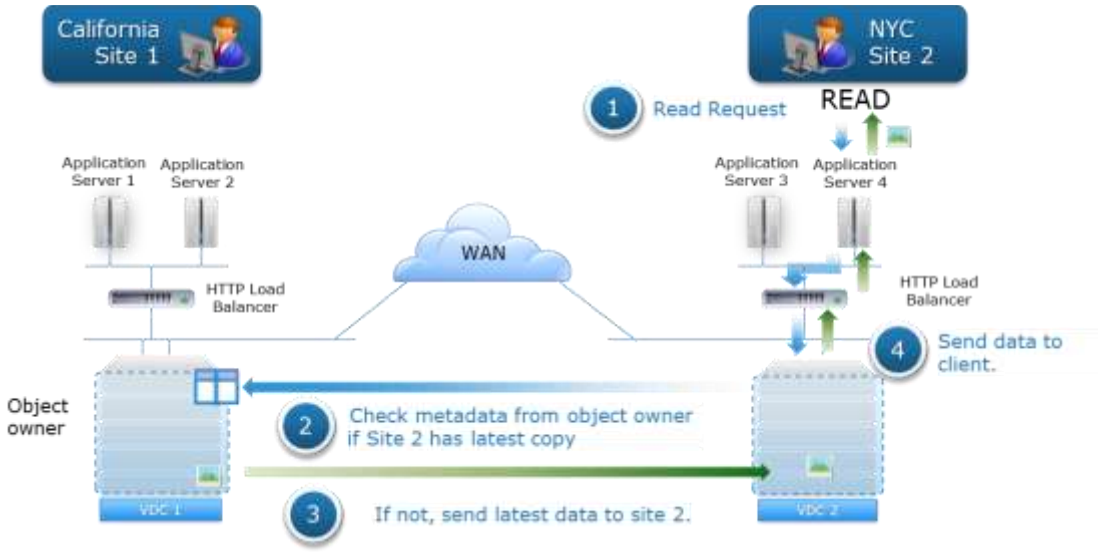

Figura 25 Solicitação de leitura para o VDC não proprietário aciona a pesquisa na WAN do VDC proprietário do objeto

O fluxo de dados das gravações em um ambiente replicado regionalmente no qual dois locais estão atualizando o mesmo objeto é exibido na Figura 26. Neste exemplo, o local 1 foi inicialmente criado e é proprietário do objeto. O objeto foi codificado para eliminação e as transações de registro relacionadas foram gravadas no disco do local 1. O fluxo de dados de uma atualização do objeto recebido no local 2 é o seguinte:

1. Primeiramente, o local 2 grava os dados localmente.
2. O local 2 atualiza os metadados de forma sincronizada (gravação de registro) com o proprietário do objeto, local 1, e aguarda a confirmação da atualização de metadados do local 1.
3. O local 1 confirma a gravação de metadados no local 2.
4. O local 2 confirma a gravação no client.

Nota: o local 2 faz a replicação assíncrona dos dados no local 1, o local proprietário do objeto, normalmente. Se os dados devem ser oferecidos a partir do local 1 antes de serem replicados nele a partir do local 2, o local 1 recuperará os dados diretamente do local 2.

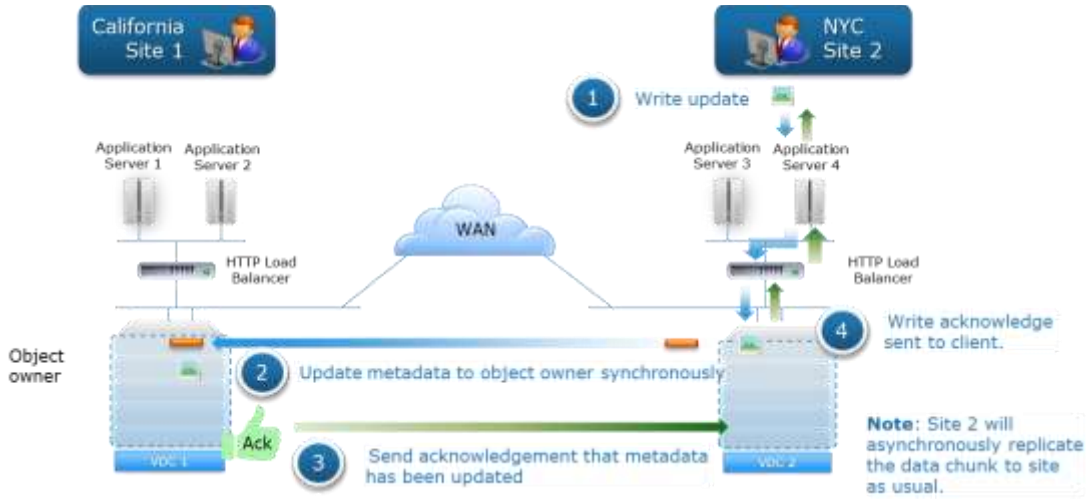

Figura 26 Atualização do fluxo de dados do mesmo objeto em um ambiente replicado regionalmente

Em cenários de leitura e gravação de um ambiente replicado regionalmente, há latência envolvida na leitura e atualização dos metadados e na recuperação de dados a partir do site proprietário do objeto.

Nota: a partir do ECS 3.4, você pode remover um VDC de um grupo de replicação (RG) em uma federação de vários VDC sem afetar o VDC ou outros RGs associados ao VDC. Remover o VDC do RG não inicia mais o PSO (interrupção permanente do local). Remover um VDC do RG inicia a recuperação.

Consulte o *Guia do Administrador do ECS* mais recente para obter mais informações sobre o grupo de replicação

8.2.4 Dados remotos com armazenamento regional em cache

O ECS otimiza os tempos de resposta para acessar os dados armazenados em sites remotos por meio do armazenamento em cache local dos objetos lidos na WAN. Isso pode ser útil para padrões de acesso por vários locais, em que os dados são frequentemente obtidos de um local remoto ou não proprietário. Considere um ambiente replicado regionalmente com três locais: VDC1, VDC2 e VDC3, em que um objeto é gravado no VDC1 e a cópia de replicação do objeto é armazenada no VDC2. Nesse cenário, para atender a uma solicitação de leitura recebida no VDC3, para o objeto criado no VDC1 e replicado no VDC2, os dados do objeto devem ser enviados ao VDC3 a partir do VDC1 ou do VDC2. O armazenamento regional em cache dos dados remotos acessados com frequência ajuda a reduzir os tempos de resposta. Um algoritmo Least Recently Used é usado para o armazenamento em cache. O tamanho do cache regional é ajustado quando a infraestrutura de hardware, como discos, nós e racks, é adicionada a um SP replicado regionalmente.

8.2.5 Comportamento durante a interrupção do local

Geralmente, a falha elétrica temporária do local (TSO) se refere a uma falha de conectividade da WAN ou de um local inteiro, como durante um desastre natural. O ECS usa mecanismos de heartbeat para detectar e lidar com falhas elétricas temporárias do local. O acesso do client e a disponibilidade de operações da API nos níveis de namespace, bucket e objeto durante uma TSO são regidos pelas seguintes opções de ADO configuradas no nível de namespace e de bucket:

- **Off (padrão)** — uma sólida consistência é mantida durante uma interrupção temporária.
- **On** — um acesso eventualmente consistente é oferecido durante uma interrupção temporária do local.

A consistência de dados durante uma TSO é implementada no nível do bucket. A configuração é definida no nível do namespace, o que define a configuração padrão de ADO em vigor para o ADO durante a criação do novo bucket, e pode ser substituída na criação de novos buckets, o que significa que a TSO pode ser configurada para alguns buckets e não para outros.

8.2.5.1 Acesso durante a paralisação (ADO) não habilitado

Por padrão, o ADO não é ativado e uma sólida consistência é mantida. Todas as solicitações de API do client em que dados autorizados de namespace, bucket ou objeto são necessários, mas estão temporariamente indisponíveis, apresentarão falha. As operações de leitura, criação, atualização e exclusão de objetos, bem como a listagem de buckets que não propriedade de um local on-line, apresentarão falha. Além disso, as operações de criação e edição de buckets, usuários e namespaces também apresentarão falha.

Conforme mencionado anteriormente, o proprietário do local inicial do bucket, namespace e objeto é o local em que o recurso foi criado pela primeira vez. Durante uma TSO, determinadas operações poderão falhar se o proprietário do local do recurso não estiver acessível. Os destaques das operações permitidas ou não permitidas durante uma falha elétrica temporária do local são:

- A criação, exclusão e atualização de buckets, namespaces, usuários de objeto, provedores de autenticação, RGs e mapeamentos de usuários e grupos do NFS não são permitidas de nenhum local.
- Listar os buckets de um namespace será permitido se o local proprietário do namespace estiver disponível.

O HDFS/NFS permite que os buckets que pertencem ao local inacessível sejam somente leitura.

8.2.5.2 Habilitado para ADO

Em um bucket habilitado para ADO, durante uma TSO, o serviço de armazenamento oferece respostas eventualmente consistentes. Nesse cenário, as leituras e, opcionalmente, as gravações de um local secundário (não proprietário) são aceitas e atendidas. Além disso, uma gravação em um local secundário durante uma TSO fará com que o local secundário aproprie-se do objeto. Isso permite que o VDC continue lendo e gravando objetos dos buckets em um namespace compartilhado. Por fim, a nova versão do objeto se tornará a versão autorizada do objeto durante a conciliação posterior à TSO, mesmo que outro aplicativo atualize o objeto no VDC proprietário.

Embora muitas operações de objeto continuem durante uma falha elétrica da rede, determinadas operações não são permitidas, como a criação de novos buckets, namespaces ou usuários. Quando a conectividade de rede entre os dois VDCs for restaurada, o mecanismo de heartbeat detectará automaticamente a conectividade, restaurará o serviço e conciliará os objetos dos dois VDCs. Se o mesmo objeto for atualizado nos VDCs A e B, a cópia do VDC não proprietário será a cópia autorizada. Portanto, se um objeto de propriedade do VDC B for atualizado nos VDCs A e B durante a sincronização, a cópia no VDC A será a cópia autorizada que é mantida, e a outra cópia terá a reversão de sua referência e ficará disponível para a recuperação de espaço.

Quando mais de dois VDCs fizerem parte de um RG, e se a conectividade de rede for interrompida entre um VDC e os outros dois, as operações de gravação/atualização/apropriação continuarão normalmente com dois VDCs; no entanto, o processo para responder às solicitações de leitura é mais complexo, como descrito abaixo.

Se um aplicativo solicitar um objeto que pertence a um VDC que não está acessível, o ECS enviará a solicitação ao VDC com a cópia secundária do objeto. No entanto, a cópia secundária do local pode ter sido sujeita a uma operação de contração de dados, que é uma operação de XOR entre dois conjuntos de dados diferentes que produz um novo conjunto de dados. Assim, primeiramente, o VDC do local secundário deve recuperar os fragmentos do objeto incluído na operação de XOR original e deve executar a operação

de XOR desses fragmentos com a cópia de recuperação. Essa operação exibirá o conteúdo do fragmento originalmente armazenado no VDC com falha. Os fragmentos do objeto recuperado podem ser remontados e devolvidos. Quando os fragmentos são reconstruídos, eles também são armazenados em cache para que o VDC possa responder mais rapidamente às solicitações subsequentes. Observe que a reconstrução é um processo demorado. Quanto mais VDCs houver em um RG, mais fragmentos deverão ser recuperados de outros VDCs e, portanto, a reconstrução do objeto levará mais tempo.

Se ocorrer um desastre, o VDC pode ficar totalmente irrecuperável. O ECS trata o VDC irrecuperável como uma falha temporária no local. Se a falha for permanente, o administrador do sistema deverá fazer permanentemente o failover do VDC da federação para iniciar o processamento do failover; o que inicia a ressincronização e a reproteção dos objetos armazenados no VDC que apresentou falha. As tarefas de recuperação são executadas como um processo de segundo plano. Você pode analisar o progresso da recuperação no portal do ECS.

Uma opção adicional de bucket está disponível para o ADO *somente leitura (RO).* Ela garante que a propriedade do objeto nunca seja alterada e remove a chance de conflitos que, de outra forma, podem ser causados por atualizações de objeto nos locais com falha e on-line durante uma falha elétrica temporária no local. A desvantagem do ADO RO é que, durante uma falha elétrica temporária no local, nenhum novo objeto poderá ser criado e nenhum objeto existente no bucket poderá ser atualizado até que todos os locais fiquem on-line novamente. A opção de ADO RO está disponível somente durante a criação do bucket, mas não pode ser modificada posteriormente. Por padrão, essa opção fica desativada.

Tabela 8 Tolerância a falhas em vários locais

Modelo de falha	Tolerância
Ambiente replicado regionalmente	Falha de até um local

8.3 Tolerância a falhas

O ECS foi projetado para tolerar várias situações de falha de equipamentos usando domínios de falha. O intervalo das condições de falha abrange um escopo diversificado, que inclui:

- Falha de um só disco rígido em um único nó
- Falha de vários discos rígidos em um único nó
- Vários nós com falha de um só disco rígido
- Vários nós com várias falhas de disco rígido
- Falha de único nó
- Várias falhas de nó
- Perda de comunicação com um VDC replicado
- Perda de todo um VDC replicado

Em uma configuração de local único, de dois locais ou de replicação regional, o impacto da falha depende da quantidade e do tipo de componentes afetados. No entanto, em cada nível, o ECS oferece mecanismos para se defender do impacto das falhas de componentes. Muitos desses mecanismos já foram discutidos neste artigo, mas são analisados aqui e na Figura 27 para mostrar como eles são aplicados à solução. Entre eles, estão:

- Disk failure
  - Os segmentos de EC ou as cópias de réplica do mesmo fragmento não são armazenadas no mesmo disco

  - Cálculo de soma de verificação nas operações de gravação e leitura
  - Verificador de consistência em segundo plano verificando novamente as somas de verificação

- Node failure

  - Distribuir segmentos ou cópias de réplica de um fragmento igualmente entre os nós de um VDC
  - O fabric do ECS mantém os serviços em execução e gerencia os recursos, como discos e rede.
  - Particionar registros e tabelas protegidos por failover de propriedade de partições entre os nós.

- Falha do rack no VDC

  - Distribuir os segmentos das cópias de réplica de um fragmento igualmente entre os racks de um VDC.
  - Uma instância de registro do fabric é executada em cada rack e pode ser reiniciada em qualquer outro nó do mesmo rack, caso o nó apresente falha.

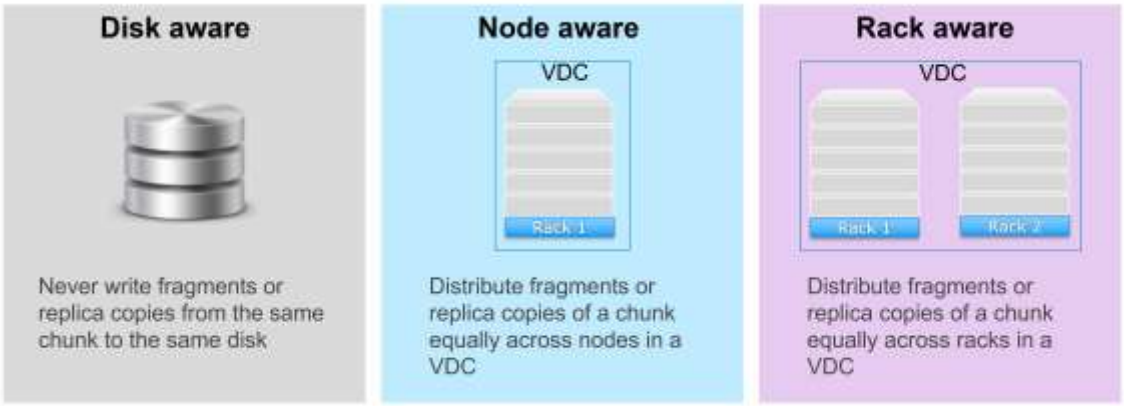

Figura 27 Mecanismos de proteção nos níveis de disco, nó e rack

O gráfico abaixo define o tipo e o número de falhas de componentes para os quais cada esquema de EC oferece proteção, em relação à configuração básica de rack. A Tabela 9 destaca a importância de considerar o impacto dos domínios de falha de proteção sobre os dados gerais e a disponibilidade de serviço, em termos de número de nós necessários em cada esquema de EC.

Tabela 9 Proteção da codificação de eliminação em domínios de falha

Esquema de EC	Nº de nós no VDC	Nº de fragmentos por nó	Dados de EC protegidos contra...
12+4 Padrão	5 ou menos	4	• Perda de até quatro discos ou • Perda de um nó
	6 ou 7	3	• Perda de até quatro discos ou • Perda de um nó e de um disco de um segundo nó
	8 ou mais	2	• Perda de até quatro discos ou • Perda de dois nós ou • Perda de um nó e de dois discos
	16 ou mais	1	• Perda de quatro nós ou • Perda de três nós e discos de um nó adicional ou • Perda de dois nós e discos de até dois nós diferentes ou • Perda de um nó e de discos de até três nós diferentes ou • Perda de quatro discos de quatro nós diferentes
10+2 Cold Storage	11 ou menos	2	• Perda de até dois discos ou • Perda de um nó
	12 ou mais	1	• Perda de qualquer número de discos de dois nós diferentes ou • Perda de dois nós

8.4 Automação da substituição de disco

A partir do ECS 3.5, os clientes podem substituir discos com falha pelo Dell EMC Services usando um fluxo de trabalho intuitivo do portal do ECS (IU na Web). O recurso fornece:

- Resolução do tipo “faça você mesmo” das falhas da unidade
- Menor tempo para correção de falhas
- Flexibilidade operacional e economia de TCO

A página de manutenção no portal do ECS oferece visibilidade de administrador para todos os discos em cada nó. Quando uma unidade falha, o sistema inicia automaticamente a recuperação. Todos os tipos de recursos na unidade são recuperados e, quando a unidade está pronta para ser removida do nó, o portal do ECS exibirá o botão Replace, conforme mostrado na Figura 28 .

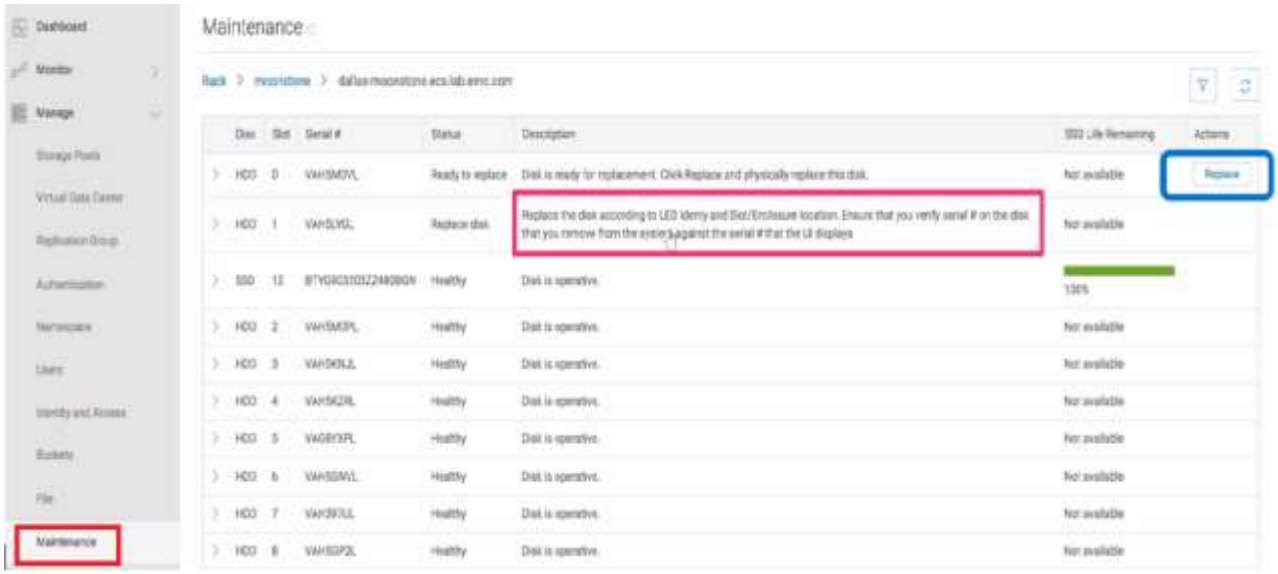

Figura 28 Automação da substituição de disco

Nota: apenas uma unidade pode ser substituída por vez. para evitar a substituição da unidade errada.

8.5 Tech Refresh

O Tech Refresh é um engajamento direcionado da Dell EMC Professional Services disponível a partir do ECS 3.5 para remover, sem interrupções, nós de hardware mais antigos dos clusters do ECS usando o recurso de software incorporado. É uma operação eficiente e de baixo consumo de recursos que pode ser acelerada com precisão. Esse recurso reduz a sobrecarga anteriormente associada ao descomissionamento do hardware do ECS.

O Tech Refresh inclui três partes:

- **Extensão de nó**: adição de nós Gen3 ao cluster existente
- **Migração de recursos**: movimentação de todos os recursos dos nós existentes para nós Gen3
- **Evacuação de nós**: limpeza dos nós antigos e remoção desses nós do cluster

A Dell EMC Professional Services deve estar envolvida na manutenção de Tech Refresh. Consulte o *Guia de Tech Refresh do ECS* mais recente para obter mais informações sobre o Tech Refresh.

9 Sobrecarga da proteção de armazenamento

Cada VDC membro de um RG é responsável por sua própria proteção por EC dos dados em nível local. Ou seja, os dados são replicados, mas nem todos os segmentos de codificação relacionados também são replicados. Embora a EC tenha mais eficiência de armazenamento que outras formas de proteção, como espelhamento total das unidades de cópia, ela provoca uma sobrecarga inerente de custo de armazenamento em nível local. No entanto, quando for necessário ter cópias secundárias replicadas fora do local e que todos os locais tenham acesso aos dados quando um só local se tornar indisponível, os custos de armazenamento se tornarão maiores que ao usar métodos tradicionais de proteção de cópia de dados entre locais. Isso é especialmente válido quando dados exclusivos são distribuídos entre três ou mais locais.

O ECS oferece um mecanismo em que a eficiência da sobrecarga da proteção de armazenamento pode aumentar à medida que três ou mais locais são federados. Em um ambiente replicado com dois VDCs, o ECS replica os fragmentos a partir do VDC principal ou proprietário para um site remoto, a fim de oferecer alta disponibilidade e resiliência. Não há como evitar o custo total da sobrecarga de proteção de uma cópia completa dos dados em uma implementação de ECS federado com dois locais.

Agora, considere três VDCs em um ambiente com vários locais: VDC1, VDC2 e VDC3, em que cada VDC tem dados exclusivos replicados nele a partir de cada um dos outros VDCs. O VDC2 e o VDC3 podem enviar uma cópia de seus dados ao VDC1 para proteção. Portanto, o VDC1 teria seus próprios dados originais, além de replicar os dados do VDC2 e do VDC3. Isso significa que o VDC1 armazenaria o equivalente ao triplo do volume de dados gravados em seu próprio local.

Nessa situação, o ECS pode executar uma operação de XOR dos dados do VDC2 e do VDC3 armazenados localmente no VDC1. Essa operação matemática compara as quantidades iguais de fragmentos de dados exclusivos e gera um resultado em um novo fragmento que contém características suficientes dos dois fragmentos de dados originais para possibilitar a restauração de qualquer um dos dois conjuntos originais. Então, onde antes havia três conjuntos exclusivos de fragmentos de dados armazenados no VDC1, consumindo o triplo da capacidade disponível, agora, há apenas dois — o conjunto de dados local original e as cópias de proteção reduzidas pela operação XOR.

Nesse mesmo cenário, se o VDC3 se tornar indisponível, o ECS poderá reconstruir os fragmentos de dados do VDC3 usando cópias de fragmentos recuperadas do VDC2 e os dados (C1 ⊕ C2) do VDC3 armazenados localmente no VDC1. Esse princípio se aplica a todos os três locais que participam do RG e depende de todos os três VDCs terem conjuntos de dados exclusivos. A Figura 29 mostra um cálculo de XOR com dois locais fazendo a replicação a um terceiro local.

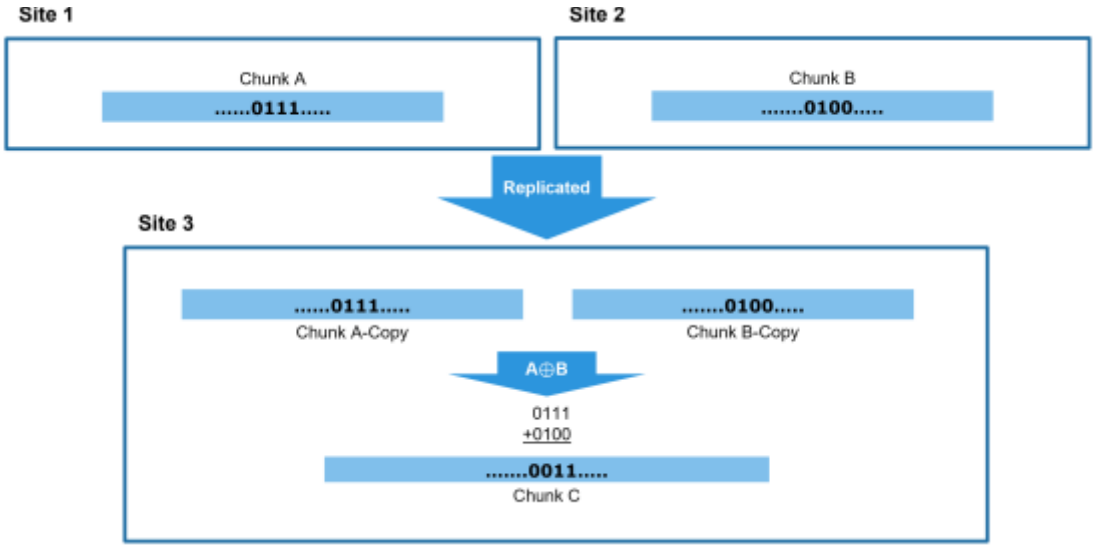

Figura 29 Eficiência da proteção de dados por XOR

Se os acordos de nível de serviço de negócios exigirem velocidades ideais de acesso de leitura, mesmo em caso de falha total no local, a configuração replicate to all sites forçará o ECS a reverter às cópias completas dos dados replicados que serão armazenados em todos os locais. Como esperado, isso aumentará os custos de armazenamento em proporção ao número dos VDCs que participam do RG. Portanto, uma configuração de três locais seria revertida para o triplo de sobrecarga da proteção de armazenamento. A configuração Replicate to All Sites está disponível durante a criação do RG e não pode ser habilitada e desabilitada.

À medida que o número de locais federados aumenta, a otimização da operação de XOR é mais eficiente na redução da sobrecarga da proteção de armazenamento, devido à replicação. A Tabela 10 apresenta informações sobre a sobrecarga da proteção de armazenamento com base no número de locais para EC normal de 12+4 e EC de arquivamento estático de 10+2, ilustrando como o ECS pode se tornar mais eficiente em termos de armazenamento à medida que mais locais são vinculados.

Nota: para reduzir a sobrecarga de dados replicados em três e até oito locais, os dados exclusivos deverão ser gravados de modo relativamente igual em cada local. Ao gravar dados em volumes iguais nos sites, cada local terá um número semelhante de fragmentos de réplica. Os números semelhantes de fragmentos de réplica em cada local resultarão em um número semelhante de operações de XOR que podem ocorrer em cada local. A eficiência máxima de armazenamento em vários locais é obtida por meio da redução do número máximo de fragmentos de réplica armazenados usando a operação de XOR.

Tabela 10 Sobrecarga da proteção de armazenamento

Nº de locais no RG	**EC 12+4**	**EC 10+2**
1	1,33	1,2
2	2,67	2.4
3	2	1.1.8
4	1,77	1,6
5	1,67	1,5
6	1,60	1,44
7	1,55	1,40
8 (nº máx. de locais no RG)	1,52	1,37

10 Conclusão

As organizações estão enfrentando volumes de dados e custos de armazenamento cada vez maiores, principalmente no espaço de nuvem pública. A arquitetura de scale-out e regionalmente distribuída do ECS oferece uma plataforma de nuvem no local que pode ser escalada a exabytes de dados com um *custo total de propriedade* significativamente menor que o do armazenamento em nuvem pública. O ECS é uma excelente solução devido a sua versatilidade, hiperescalabilidade, recursos avançados e uso de hardware genérico.

A Recursose suporte técnico

O foco do site Dell.com.br/support é atender às necessidades dos clientes com serviços e suporte comprovados.

Os documentos e vídeos técnicos sobre armazenamento oferecem expertise que ajuda a garantir o sucesso do cliente em plataformas de armazenamento Dell EMC.