AWS Estrutura Well-Architected

Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem

Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem: AWS Estrutura Well-Architected

Table of Contents

Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem

Data de publicação: 12 de fevereiro de 2021 (Histórico do documentos)

A recuperação de desastres é o processo de preparação e recuperação de um desastre. Um evento que impede que uma carga de trabalho ou sistema cumpra seus objetivos de negócios em seu local de implantação principal é considerado um desastre. Este paper descreve as melhores práticas para planejar e testar a recuperação de desastres para qualquer carga de trabalho implantada e oferece diferentes abordagens para AWS mitigar riscos e atingir o objetivo de tempo de recuperação (RTO) e o objetivo de ponto de recuperação (RPO) dessa carga de trabalho.

Este whitepaper aborda como implementar a recuperação de desastres para cargas de trabalho em. AWS Consulte Recuperação de desastres de aplicativos locais AWS para obter informações sobre AWS como usar como site de recuperação de desastres para cargas de trabalho locais.

Introdução

Sua carga de trabalho deve desempenhar a função pretendida de forma correta e consistente. Para conseguir isso, você deve arquitetar a resiliência. Resiliência é a capacidade de uma carga de trabalho se recuperar de interrupções na infraestrutura, nos serviços ou nos aplicativos, adquirir dinamicamente recursos de computação para atender à demanda e mitigar interrupções, como configurações incorretas ou problemas transitórios de rede.

A recuperação de desastres (DR) é uma parte importante de sua estratégia de resiliência e diz respeito à forma como sua carga de trabalho responde quando ocorre um desastre (um desastre é um evento que causa um sério impacto negativo em seus negócios). Essa resposta deve ser baseada nos objetivos de negócios de sua organização, que especificam a estratégia de sua carga de trabalho para evitar a perda de dados, conhecida como Objetivo de Ponto de Recuperação (RPO), e reduzir o tempo de inatividade quando sua carga de trabalho não está disponível para uso, conhecida como Objetivo de Tempo de Recuperação (RTO). Portanto, você deve implementar resiliência no design de suas cargas de trabalho na nuvem para atender aos seus objetivos de recuperação (RPO e RTO) para um determinado evento de desastre único. Essa abordagem ajuda sua organização a manter a continuidade dos negócios como parte do Planejamento de Continuidade de Negócios (BCP).

Este paper se concentra em como planejar, projetar e implementar arquiteturas AWS que atendam aos objetivos de recuperação de desastres da sua empresa. As informações compartilhadas aqui são destinadas a pessoas em funções de tecnologia, como diretores de tecnologia (CTOs), arquitetos, desenvolvedores, membros da equipe de operações e pessoas encarregadas de avaliar e mitigar riscos.

Recuperação e disponibilidade de desastres

A recuperação de desastres pode ser comparada à disponibilidade, que é outro componente importante da sua estratégia de resiliência. Enquanto a recuperação de desastres mede os objetivos para eventos únicos, os objetivos de disponibilidade medem os valores médios em um período de tempo.

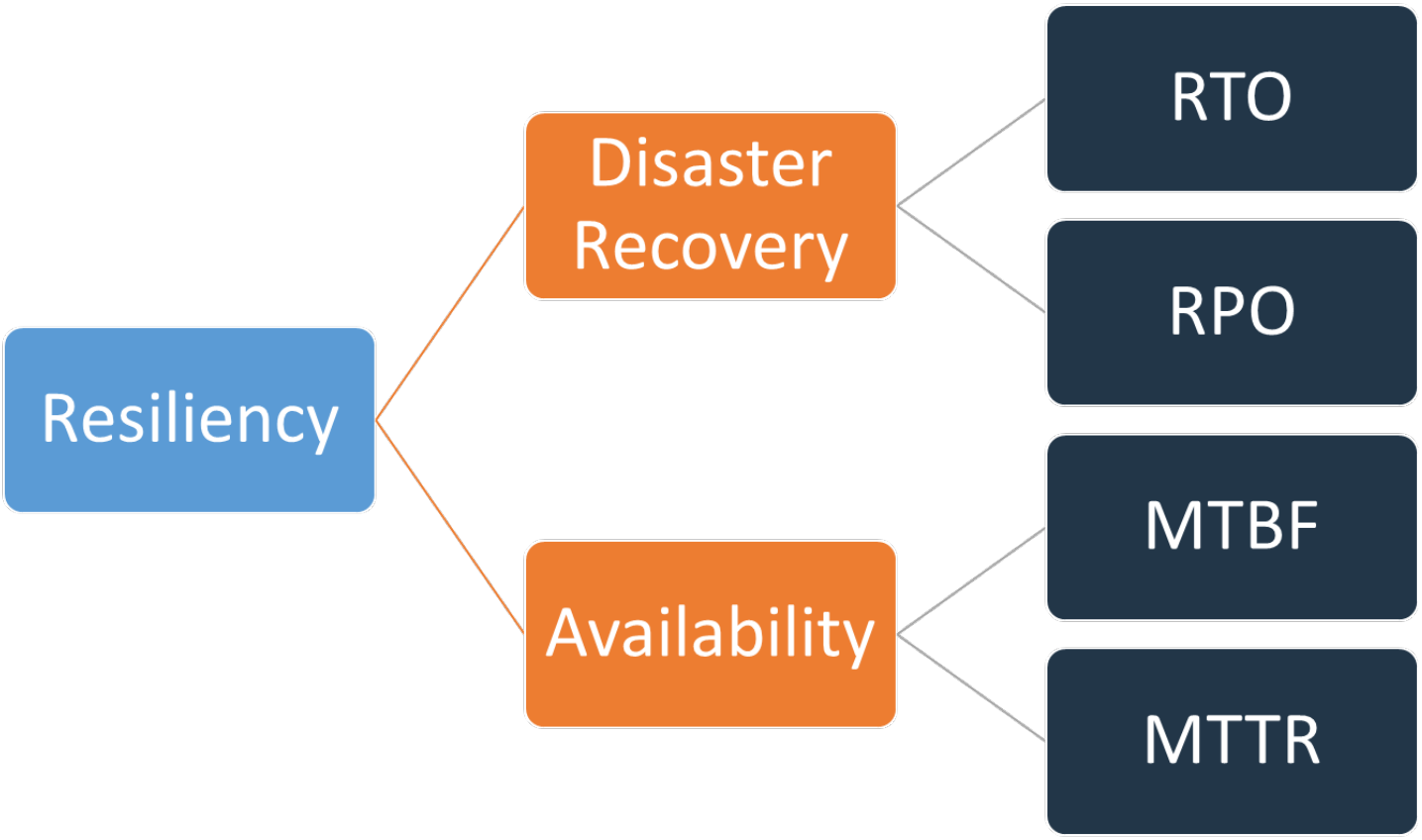

Figura 1 - Objetivos de resiliência

A disponibilidade é calculada usando o tempo médio entre falhas (MTBF) e o tempo médio de recuperação (MTTR):

$$Availability = \frac{Available\ for\ Use\ Time}{Total\ Time} = \frac{MTBF}{MTBF + MTTR}$$

Essa abordagem geralmente é chamada de “nove”, em que uma meta de disponibilidade de 99,9% é chamada de “três noves”.

Para sua carga de trabalho, pode ser mais fácil contar solicitações bem-sucedidas e malsucedidas em vez de usar uma abordagem baseada em tempo. Nesse caso, o seguinte cálculo pode ser usado:

$$Availability = \frac{Successful\ Responses}{Valid\ Requests}$$

A recuperação de desastres se concentra em eventos de desastre, enquanto a disponibilidade se concentra em interrupções mais comuns de menor escala, como falhas de componentes, problemas de rede, bugs de software e picos de carga. O objetivo da recuperação de desastres é a continuidade dos negócios, enquanto a disponibilidade diz respeito à maximização do tempo em que uma carga de trabalho está disponível para realizar a funcionalidade comercial pretendida. Ambos devem fazer parte da sua estratégia de resiliência.

Você é Well-Architected?

O AWS Well-Architected Framework ajuda você a entender os prós e os contras das decisões que você toma ao criar sistemas na nuvem. Os seis pilares do framework permitem a você conhecer as melhores práticas de arquitetura para criar e operar sistemas confiáveis, seguros, econômicos e sustentáveis na nuvem. Usando a AWS Well-Architected Tool, disponível gratuitamente no AWS Management Console, você pode analisar suas cargas de trabalho em relação a essas melhores práticas respondendo a um conjunto de perguntas para cada pilar.

Os conceitos abordados neste whitepaper expandem as melhores práticas contidas no whitepaper do Pilar de Confiabilidade, especificamente na pergunta REL 13, “Como você planeja a recuperação de desastres (DR)?”. Depois de implementar as práticas neste whitepaper, certifique-se de revisar (ou revisar novamente) sua carga de trabalho usando o AWS Well-Architected Tool.

Modelo de responsabilidade compartilhada para resiliência

Resiliência é uma responsabilidade compartilhada entre você AWS e você, o cliente. É importante que você entenda como a recuperação e a disponibilidade de desastres, como parte da resiliência, operam sob esse modelo compartilhado.

Responsabilidade da AWS “Resiliência da nuvem”

A AWS é responsável pela resiliência da infraestrutura que executa todos os serviços oferecidos na nuvem da AWS. Essa infraestrutura compreende o hardware, o software, a rede e as instalações que executam os serviços de nuvem da AWS. A AWS usa esforços comercialmente razoáveis para disponibilizar esses serviços de nuvem da AWS, garantindo que a disponibilidade do serviço atenda ou exceda os acordos de nível de serviço da AWS (SLAs).

A infraestrutura de nuvem global da AWS foi projetada para permitir que os clientes criem arquiteturas de carga de trabalho altamente resilientes. Cada região da AWS é totalmente isolada e consiste em várias zonas de disponibilidade, que são partições de infraestrutura fisicamente isoladas. As zonas de disponibilidade isolam falhas que poderiam afetar a resiliência da workload, impedindo que elas afetem outras zonas na região. Mas, ao mesmo tempo, todas as zonas em uma região da AWS estão interconectadas com redes de alta largura de banda e baixa latência, por meio de fibra metropolitana dedicada e totalmente redundante, fornecendo redes de alta taxa de transferência e baixa latência entre as zonas. Todo o tráfego entre as zonas é criptografado. A performance da rede é suficiente para realizar a replicação síncrona entre as zonas. Quando um aplicativo é particionado AZs, as empresas ficam melhor isoladas e protegidas de problemas como quedas de energia, raios, tornados, furacões e muito mais.

Responsabilidade do cliente “Resiliência na nuvem”

Sua responsabilidade será determinada pelos serviços de nuvem da AWS que você selecionar. Isso determina o volume do trabalho de configuração que você deve executar como parte das suas responsabilidades de resiliência. Por exemplo, um serviço como o Amazon Elastic Compute Cloud (Amazon EC2) exige que o cliente execute todas as tarefas necessárias de configuração e gerenciamento de resiliência. Os clientes que implantam EC2 instâncias da Amazon são responsáveis por implantar EC2 instâncias em vários locais (como zonas de disponibilidade da AWS), implementar a autorrecuperação usando serviços como o Amazon Auto EC2 Scaling, bem como usar as melhores práticas de arquitetura de carga de trabalho resiliente para aplicativos

instalados nas instâncias. Para serviços gerenciados, como Amazon S3 e Amazon DynamoDB, a AWS opera a camada de infraestrutura, o sistema operacional e as plataformas, e os clientes acessam os endpoints para armazenar e recuperar dados. Você é responsável por gerenciar a resiliência dos dados incluindo estratégias de backup, versionamento e replicação.

Implantar sua carga de trabalho em várias zonas de disponibilidade em uma região da AWS faz parte de uma estratégia de alta disponibilidade projetada para proteger cargas de trabalho isolando problemas em uma zona de disponibilidade e usa a redundância das outras zonas de disponibilidade para continuar atendendo às solicitações. Uma arquitetura Multi-AZ também faz parte de uma estratégia de DR desenvolvida para que as workloads sejam mais isoladas e protegidas de problemas como queda de energia, raios, tornados, terremotos, dentre outros. As estratégias de DR também podem fazer uso de várias regiões da AWS. Por exemplo, em uma configuração ativa/passiva, o serviço da carga de trabalho passará de sua região ativa para sua região de DR se a região ativa não puder mais atender às solicitações.

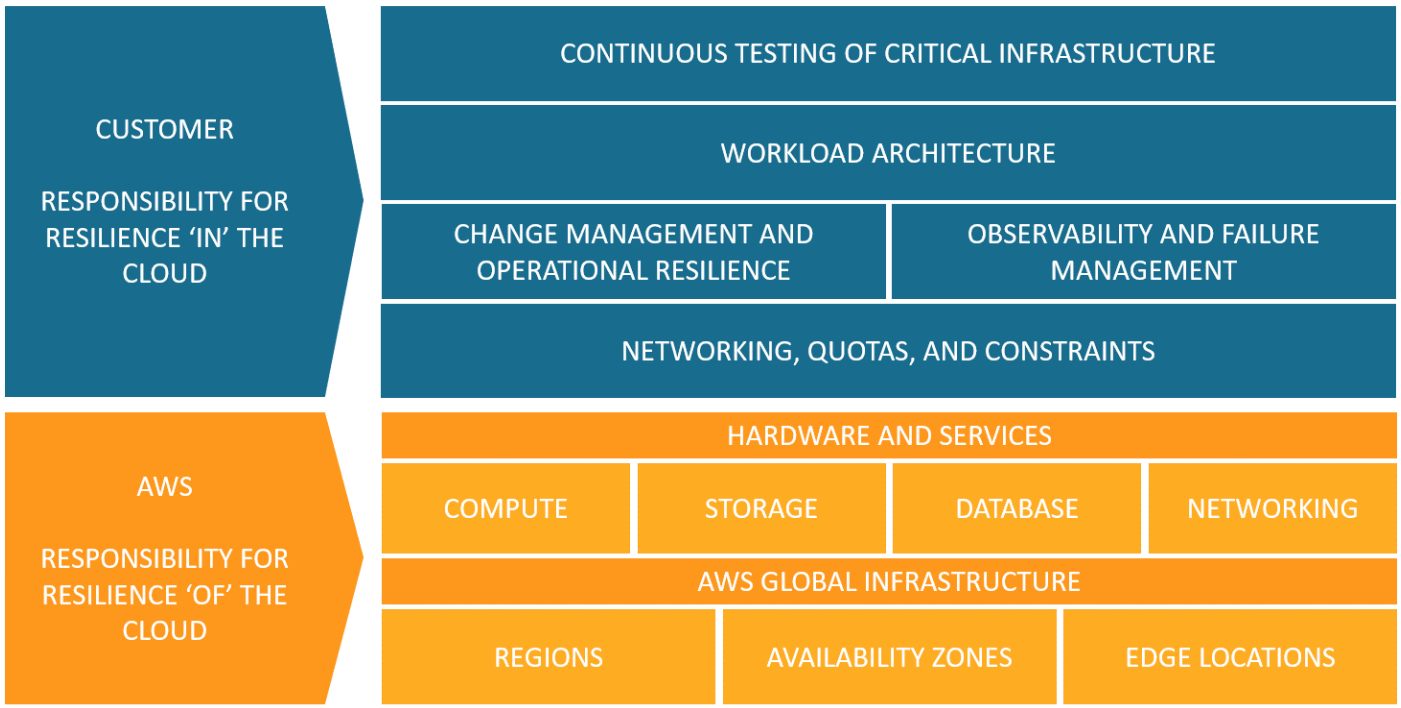

Figura 2 - Resiliência é uma responsabilidade compartilhada entre a AWS e o cliente

O que é um desastre?

Ao planejar a recuperação de desastres, avalie seu plano para essas três categorias principais de desastres:

- Desastres naturais, como terremotos ou inundações
- Falhas técnicas, como falha de energia ou conectividade de rede
- Ações humanas, como configuração incorreta inadvertida ou acesso ou modificação de unauthorized/outside partes

Cada um desses desastres em potencial também terá um impacto geográfico que pode ser local, regional, nacional, continental ou global. Tanto a natureza do desastre quanto o impacto geográfico são importantes ao considerar sua estratégia de recuperação de desastres. Por exemplo, você pode mitigar um problema de inundação local que causa uma interrupção no data center empregando uma estratégia Multi-AZ, já que ela não afetaria mais de uma zona de disponibilidade. No entanto, um ataque aos dados de produção exigiria que você invocasse uma estratégia de recuperação de desastres que executasse o failover dos dados de backup em outra região da AWS.

Alta disponibilidade não é recuperação de desastres

Tanto a disponibilidade quanto a recuperação de desastres dependem de algumas das mesmas práticas recomendadas, como monitoramento de falhas, implantação em vários locais e failover automático. No entanto, a disponibilidade se concentra nos componentes da carga de trabalho, enquanto a recuperação de desastres se concentra em cópias discretas de toda a carga de trabalho. A recuperação de desastres tem objetivos diferentes da disponibilidade, medindo o tempo de recuperação após os eventos de maior escala que se qualificam como desastres. Primeiro, você deve garantir que sua carga de trabalho atenda aos seus objetivos de disponibilidade, pois uma arquitetura altamente disponível permitirá que você atenda às necessidades dos clientes no caso de eventos que afetem a disponibilidade. Sua estratégia de recuperação de desastres exige abordagens diferentes das de disponibilidade, com foco na implantação de sistemas distintos em vários locais, para que você possa executar o failover de toda a carga de trabalho, se necessário.

Você deve considerar a disponibilidade de sua carga de trabalho no planejamento de recuperação de desastres, pois isso influenciará a abordagem adotada. Uma carga de trabalho executada em uma única EC2 instância da Amazon em uma zona de disponibilidade não tem alta disponibilidade. Se um problema de inundação local afetar essa zona de disponibilidade, esse cenário exigirá o failover para outra AZ para atender aos objetivos de DR. Compare esse cenário com uma carga de trabalho altamente disponível implantada em vários sites ativa/ativa, em que a carga de trabalho é implantada em várias regiões ativas e todas as regiões estão atendendo ao tráfego de produção. Nesse caso, mesmo no caso improvável de um grande desastre tornar uma região inutilizável, a estratégia de DR é realizada roteando todo o tráfego para as regiões restantes.

A forma como você aborda os dados também é diferente entre disponibilidade e recuperação de desastres. Considere uma solução de armazenamento que se replica continuamente em outro local para obter alta disponibilidade (como uma carga de active/active trabalho com vários sites). Se um arquivo ou arquivos forem excluídos ou corrompidos no dispositivo de armazenamento primário, essas alterações destrutivas poderão ser replicadas no dispositivo de armazenamento secundário. Nesse cenário, apesar da alta disponibilidade, a capacidade de failover em caso de exclusão ou corrupção de dados será comprometida. Em vez disso, um point-in-time backup também é necessário como parte de uma estratégia de DR.

Plano de Continuidade de Negócios (BCP)

Seu plano de recuperação de desastres deve ser um subconjunto do plano de continuidade de negócios (BCP) da sua organização, não deve ser um documento independente. Não adianta manter metas agressivas de recuperação de desastres para restaurar uma carga de trabalho se os objetivos comerciais dessa carga de trabalho não puderem ser alcançados devido ao impacto do desastre em outros elementos de sua empresa além da carga de trabalho. Por exemplo, um terremoto pode impedir que você transporte produtos comprados em seu aplicativo de comércio eletrônico — mesmo que o DR eficaz mantenha sua carga de trabalho funcionando, seu BCP precisa acomodar as necessidades de transporte. Sua estratégia de DR deve ser baseada nos requisitos, prioridades e contexto dos negócios.

Análise de impacto nos negócios e avaliação de riscos

Uma análise de impacto nos negócios deve quantificar o impacto comercial de uma interrupção em suas cargas de trabalho. Ele deve identificar o impacto sobre os clientes internos e externos da impossibilidade de usar suas cargas de trabalho e o efeito que isso tem em seus negócios. A análise deve ajudar a determinar com que rapidez a carga de trabalho precisa ser disponibilizada e quanta perda de dados pode ser tolerada. No entanto, é importante observar que os objetivos de recuperação não devem ser estabelecidos isoladamente; a probabilidade de interrupção e o custo da recuperação são fatores-chave que ajudam a informar o valor comercial de fornecer recuperação de desastres para uma carga de trabalho.

O impacto nos negócios pode depender do tempo. Talvez você queira considerar isso em seu planejamento de recuperação de desastres. Por exemplo, é provável que a interrupção do sistema de folha de pagamento tenha um impacto muito alto nos negócios pouco antes de todos receberem o pagamento, mas pode ter um impacto baixo logo após todos já terem recebido o pagamento.

Uma avaliação de risco do tipo de desastre e do impacto geográfico, juntamente com uma visão geral da implementação técnica de sua carga de trabalho, determinará a probabilidade de ocorrência de interrupção em cada tipo de desastre.

Para cargas de trabalho altamente críticas, você pode considerar a implantação de infraestrutura em várias regiões com replicação de dados e backups contínuos para minimizar o impacto nos negócios. Para cargas de trabalho menos críticas, uma estratégia válida pode ser não implementar nenhuma recuperação de desastres. E para alguns cenários de desastres, também é válido não ter nenhuma

estratégia de recuperação de desastres implementada como uma decisão informada com base na baixa probabilidade de o desastre ocorrer. Lembre-se de que as zonas de disponibilidade em uma região da AWS já foram projetadas com uma distância significativa entre elas e um planejamento cuidadoso da localização, de forma que os desastres mais comuns afetem apenas uma zona e não as outras. Portanto, uma arquitetura Multi-AZ dentro de uma região da AWS já pode atender a grande parte das suas necessidades de mitigação de riscos.

O custo das opções de recuperação de desastres deve ser avaliado para garantir que a estratégia de recuperação de desastres forneça o nível correto de valor comercial, considerando o impacto e o risco nos negócios.

Com todas essas informações, você pode documentar a ameaça, o risco, o impacto e o custo de diferentes cenários de desastres e as opções de recuperação associadas. Essas informações devem ser usadas para determinar seus objetivos de recuperação para cada uma de suas cargas de trabalho.

Objetivos de recuperação (RTO e RPO)

Ao criar uma estratégia de recuperação de desastres (DR), as organizações geralmente planejam o objetivo de tempo de recuperação (RTO) e o objetivo de ponto de recuperação (RPO).

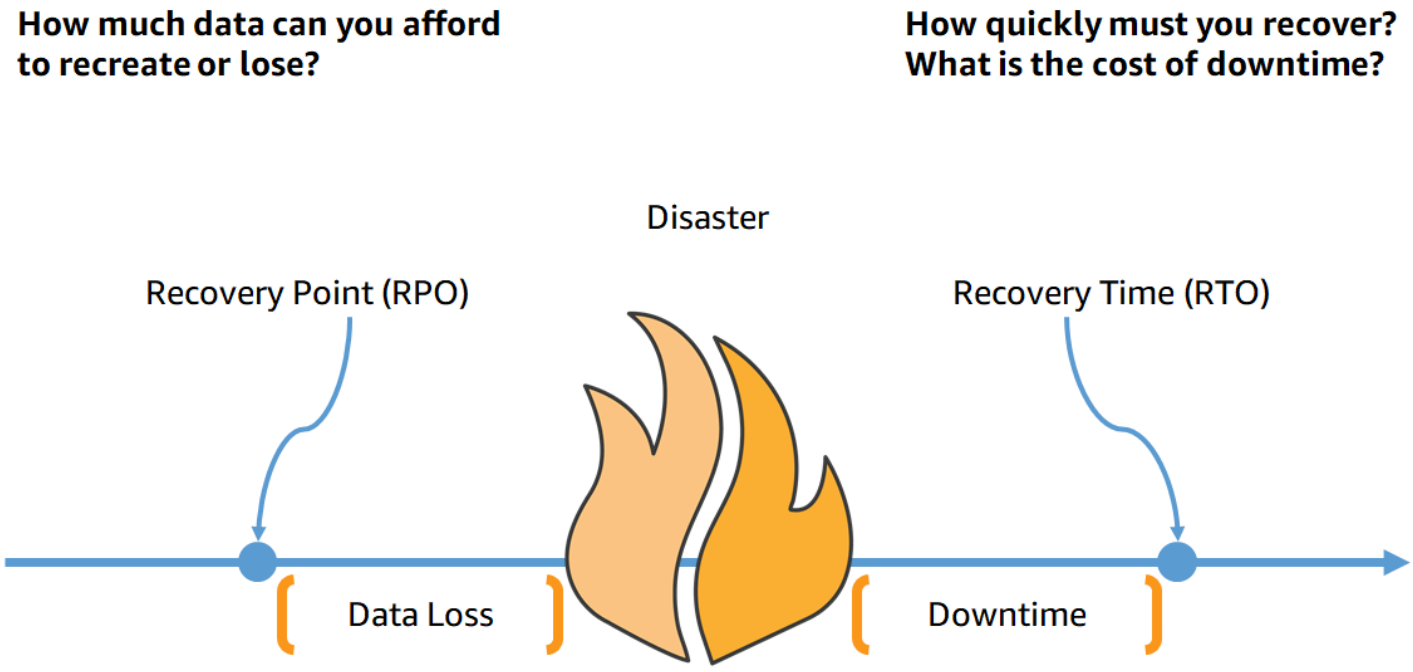

Figura 3 - Objetivos de recuperação

O objetivo de tempo de recuperação (RTO) é o atraso máximo aceitável entre a interrupção do serviço e a restauração do serviço. Esse objetivo determina o que é considerado uma janela de tempo aceitável quando o serviço não está disponível e é definido pela organização.

Há basicamente quatro estratégias de DR discutidas neste paper: backup e restauração, luz piloto, espera quente e vários locais active/active (consulte Opções de recuperação de desastres na nuvem). No diagrama a seguir, a empresa determinou seu RTO máximo permitido, bem como o limite do que eles podem gastar em sua estratégia de restauração de serviços. Dados os objetivos do negócio, as estratégias de DR Pilot Light ou Warm Standby satisfarão tanto o RTO quanto os critérios de custo.

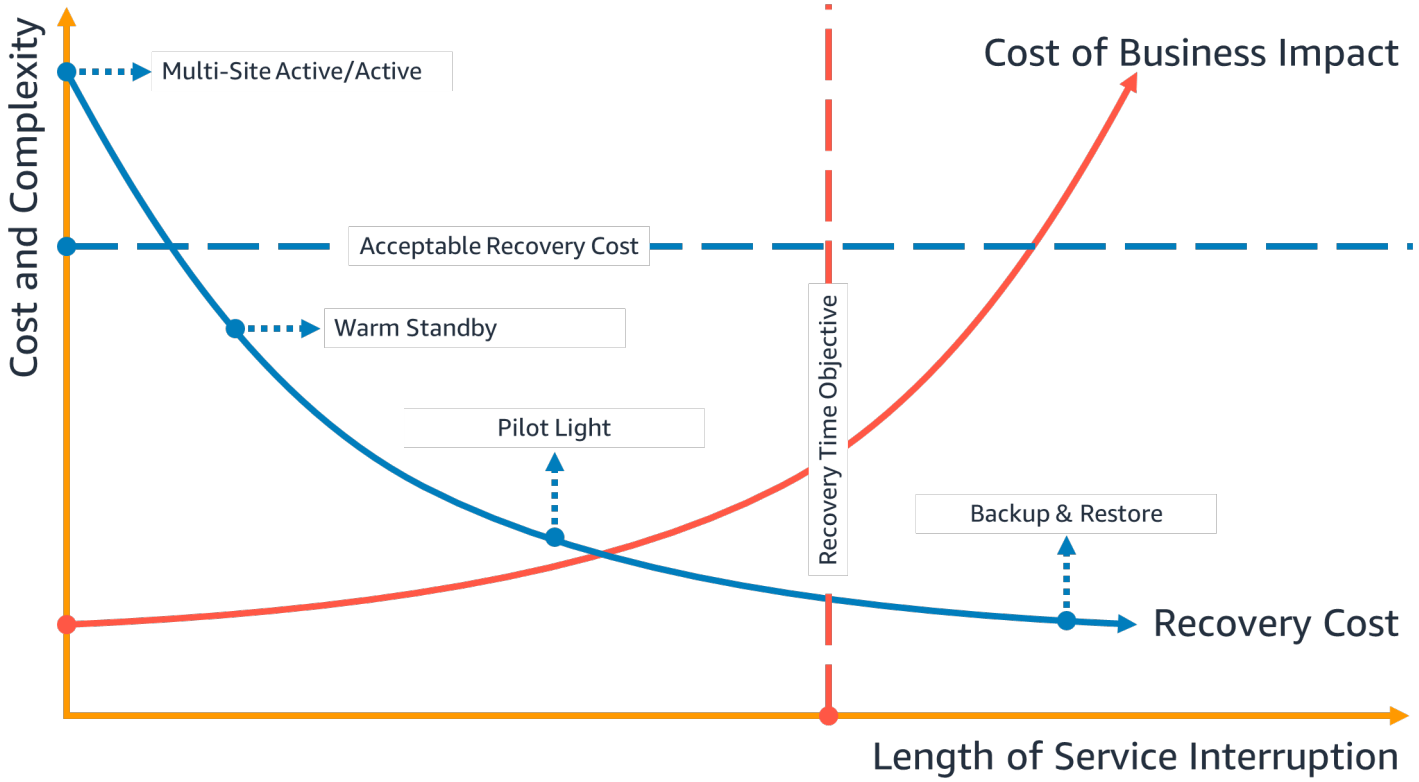

Figura 4 - Objetivo do tempo de recuperação

O objetivo de ponto de recuperação (RPO) é o tempo máximo aceitável desde o último ponto de recuperação de dados. Esse objetivo determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço e é definido pela organização.

No diagrama a seguir, a empresa determinou seu RPO máximo permitido, bem como o limite do que pode gastar em sua estratégia de recuperação de dados. Das quatro estratégias de DR, a estratégia Pilot Light ou a estratégia Warm Standby DR atendem aos critérios de RPO e custo.

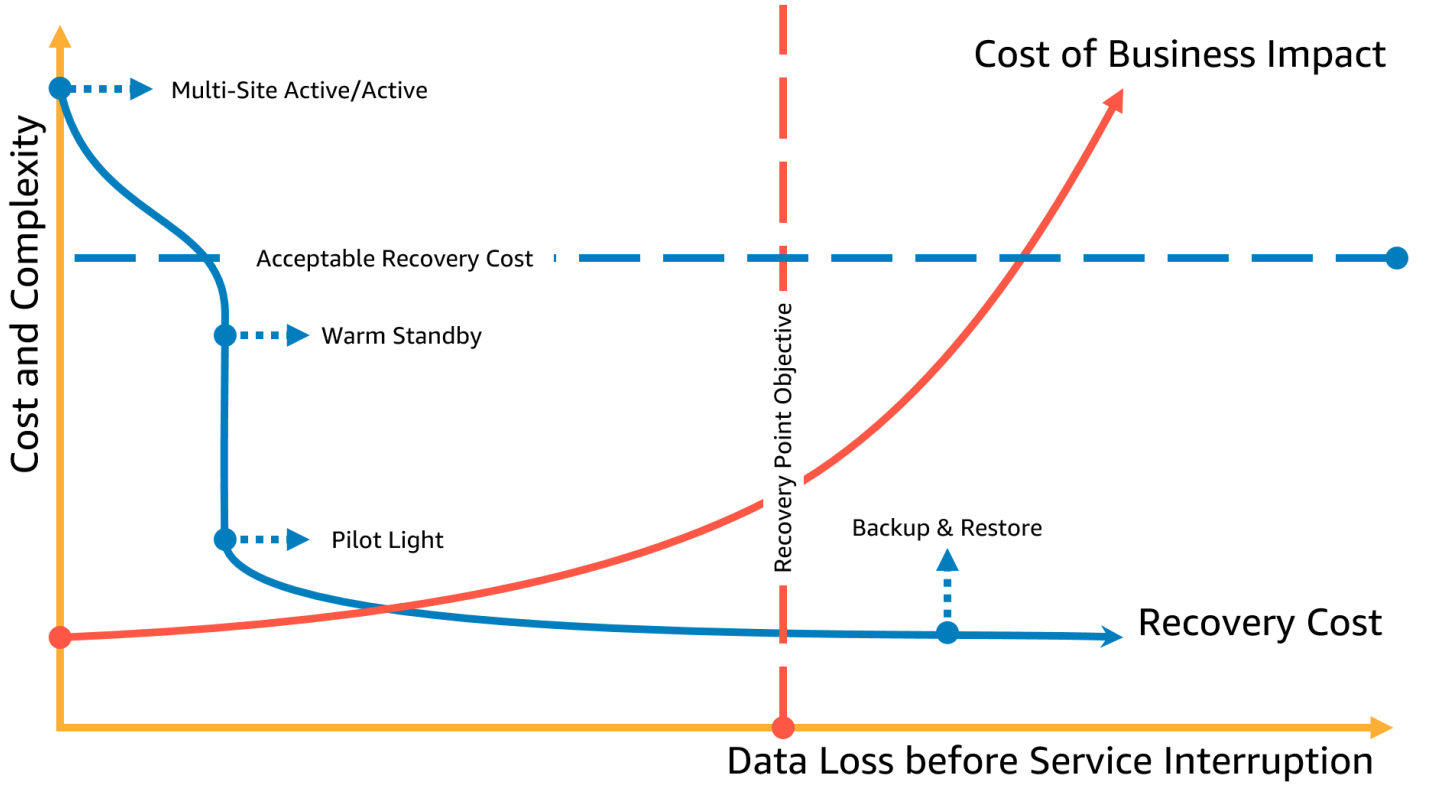

Figura 5 - Objetivo do ponto de recuperação

> **Note**
>
> Se o custo da estratégia de recuperação for maior do que o custo da falha ou perda, a opção de recuperação não deve ser implementada, a menos que haja um fator secundário, como requisitos regulatórios. Considere estratégias de recuperação de custo variável ao fazer essa avaliação.

A recuperação de desastres é diferente na nuvem

As estratégias de recuperação de desastres evoluem com a inovação técnica. Um plano de recuperação de desastres no local pode envolver o transporte físico de fitas ou a replicação de dados para outro local. Sua organização precisa reavaliar o impacto comercial, o risco e o custo de suas estratégias anteriores de recuperação de desastres para cumprir seus objetivos de DR na AWS. A recuperação de desastres na nuvem da AWS inclui as seguintes vantagens em relação aos ambientes tradicionais:

- Recupere-se rapidamente de um desastre com complexidade reduzida
- Testes simples e repetíveis permitem que você teste com mais facilidade e frequência
- A menor sobrecarga de gerenciamento diminui a carga operacional
- Oportunidades para automatizar, diminuir as chances de erro e melhorar o tempo de recuperação

A AWS permite que você troque a despesa de capital fixo de um datacenter de backup físico pela despesa operacional variável de um ambiente do tamanho certo na nuvem, o que pode reduzir significativamente os custos.

Para muitas organizações, a recuperação de desastres no local foi baseada no risco de interrupção de uma carga de trabalho ou cargas de trabalho em um data center e na recuperação de dados copiados ou replicados em um data center secundário. Quando as organizações implantam cargas de trabalho na AWS, elas podem implementar uma carga de trabalho bem arquitetada e confiar no design da infraestrutura de nuvem global da AWS para ajudar a mitigar o efeito dessas interrupções. Consulte o whitepaper AWS Well-Architected Framework — Reliability Pillar para obter mais informações sobre as melhores práticas arquitetônicas para projetar e operar cargas de trabalho confiáveis, seguras, eficientes e econômicas na nuvem. Use o AWS Well-Architected Toolpara revisar suas cargas de trabalho periodicamente para garantir que elas sigam as melhores práticas e orientações do Well-Architected Framework. A ferramenta está disponível gratuitamente no AWS Management Console.

Se suas cargas de trabalho estão na AWS, você não precisa se preocupar com a conectividade do data center (com exceção da sua capacidade de acessá-la), energia, ar condicionado, supressão de incêndio e hardware. Tudo isso é gerenciado para você e você tem acesso a várias zonas de disponibilidade isoladas de falhas (cada uma composta por um ou mais data centers distintos).

Região única da AWS

Para um evento de desastre baseado na interrupção ou perda de um datacenter físico, a implementação de uma carga de trabalho altamente disponível em várias zonas de disponibilidade em uma única região da AWS ajuda a mitigar desastres naturais e técnicos. O backup contínuo de dados nessa única região pode reduzir o risco de ameaças humanas, como um erro ou atividade não autorizada que pode resultar em perda de dados. Cada região da AWS é composta por várias zonas de disponibilidade, cada uma isolada de falhas nas outras zonas. Cada zona de disponibilidade, por sua vez, consiste em um ou mais data centers físicos distintos. Para isolar melhor os problemas impactantes e obter alta disponibilidade, você pode particionar cargas de trabalho em várias zonas na mesma região. As zonas de disponibilidade são projetadas para redundância física e fornecem resiliência, permitindo desempenho ininterrupto, mesmo em caso de falta de energia, tempo de inatividade da Internet, inundações e outros desastres naturais. Consulte a infraestrutura de nuvem global da AWS para descobrir como a AWS faz isso.

Ao implantar em várias zonas de disponibilidade em uma única região da AWS, sua carga de trabalho fica mais protegida contra falhas em um único (ou até mesmo vários) datacenters. Para maior garantia com sua implantação em uma única região, você pode fazer backup dos dados e da configuração (incluindo a definição da infraestrutura) em outra região. Essa estratégia reduz o escopo do seu plano de recuperação de desastres para incluir apenas backup e restauração de dados. Aproveitar a resiliência multirregional fazendo backup em outra região da AWS é simples e barato em relação às outras opções multirregionais descritas na seção a seguir. Por exemplo, fazer backup no Amazon Simple Storage Service (Amazon S3) dá acesso à recuperação imediata de seus dados. No entanto, se sua estratégia de DR para partes de seus dados tiver requisitos mais flexíveis de tempos de recuperação (de minutos a horas), o uso do Amazon S3 Glacier ou do Amazon S3 Glacier Deep Archive reduzirá significativamente os custos de sua estratégia de backup e recuperação.

Algumas cargas de trabalho podem ter requisitos regulamentares de residência de dados. Se isso se aplica à sua carga de trabalho em uma localidade que atualmente tem apenas uma região da AWS, além de projetar cargas de trabalho Multi-AZ para alta disponibilidade, conforme discutido acima, você também pode usar as AZs dentro dessa região como locais discretos, o que pode ser útil para atender aos requisitos de residência de dados aplicáveis à sua carga de trabalho nessa região. As estratégias de DR descritas nas seções a seguir usam várias regiões da AWS, mas também podem ser implementadas usando zonas de disponibilidade em vez de regiões.

Várias regiões da AWS

Para um evento de desastre que inclua o risco de perder vários datacenters a uma distância significativa um do outro, você deve considerar as opções de recuperação de desastres para mitigar desastres naturais e técnicos que afetam uma região inteira na AWS. Todas as opções descritas nas seções a seguir podem ser implementadas como arquiteturas multirregionais para proteção contra esses desastres.

Opções de recuperação de desastres na nuvem

As estratégias de recuperação de desastres disponíveis para você na AWS podem ser amplamente categorizadas em quatro abordagens, desde o baixo custo e a baixa complexidade de fazer backups até estratégias mais complexas usando várias regiões ativas. Active/passive as estratégias usam um site ativo (como uma região da AWS) para hospedar a carga de trabalho e fornecer tráfego. O site passivo (como uma região diferente da AWS) é usado para recuperação. O site passivo não fornece tráfego ativamente até que um evento de failover seja acionado.

É fundamental avaliar e testar regularmente sua estratégia de recuperação de desastres para que você tenha confiança em invocá-la, caso seja necessário. Use o AWS Resilience Hub para validar e monitorar continuamente a resiliência de suas AWS cargas de trabalho, inclusive se é provável que você atinja suas metas de RTO e RPO.

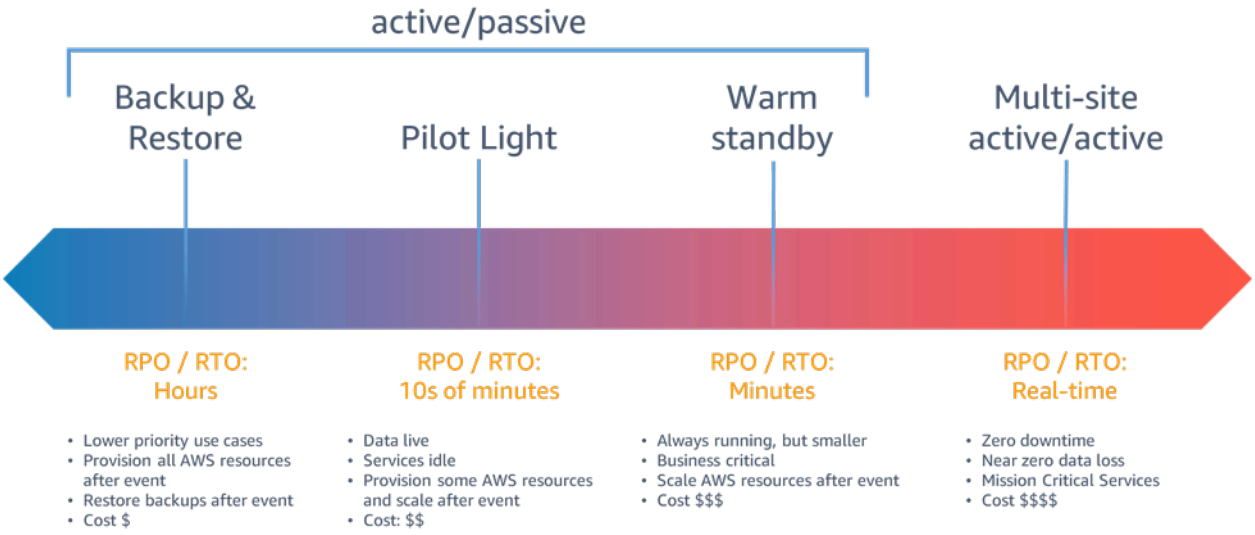

Figura 6 - Estratégias de recuperação de desastres

Para um evento de desastre baseado na interrupção ou perda de um data center físico para uma carga de trabalho bem arquitetada e altamente disponível, você pode precisar apenas de uma abordagem de backup e restauração para a recuperação de desastres. Se sua definição de desastre vai além da interrupção ou perda de um data center físico para a de uma região ou se você estiver sujeito aos requisitos regulatórios que o exijam, considere Pilot Light, Warm Standby ou Multi-Site Active/Active.

Ao escolher sua estratégia e os recursos da AWS para implementá-la, lembre-se de que, dentro da AWS, geralmente dividimos os serviços no plano de dados e no plano de controle. O plano de

dados é responsável por prestar serviço em tempo real, enquanto os ambientes de gerenciamento são usados para configurar o ambiente. Para máxima resiliência, você deve usar somente operações de plano de dados como parte de sua operação de failover. Isso ocorre porque os planos de dados normalmente têm metas de projeto de maior disponibilidade do que os planos de controle.

Backup e restauração

O backup e a restauração são uma abordagem adequada para mitigar a perda ou a corrupção de dados. Essa abordagem também pode ser usada para mitigar um desastre regional replicando dados para outras regiões da AWS ou para mitigar a falta de redundância para cargas de trabalho implantadas em uma única zona de disponibilidade. Além dos dados, você deve reimplantar a infraestrutura, a configuração e o código do aplicativo na região de recuperação. Para permitir que a infraestrutura seja reimplantada rapidamente sem erros, você deve sempre implantar usando a infraestrutura como código (IaC) usando serviços como AWS CloudFormationou o. AWS Cloud Development Kit (AWS CDK) Sem o IaC, pode ser complexo restaurar cargas de trabalho na região de recuperação, o que resultará em maiores tempos de recuperação e possivelmente excederá seu RTO. Além dos dados do usuário, certifique-se também de fazer backup do código e da configuração, incluindo Amazon Machine Images (AMIs) que você usa para criar EC2 instâncias da Amazon. Você pode usar AWS CodePipelinepara automatizar a reimplantação do código e da configuração do aplicativo.

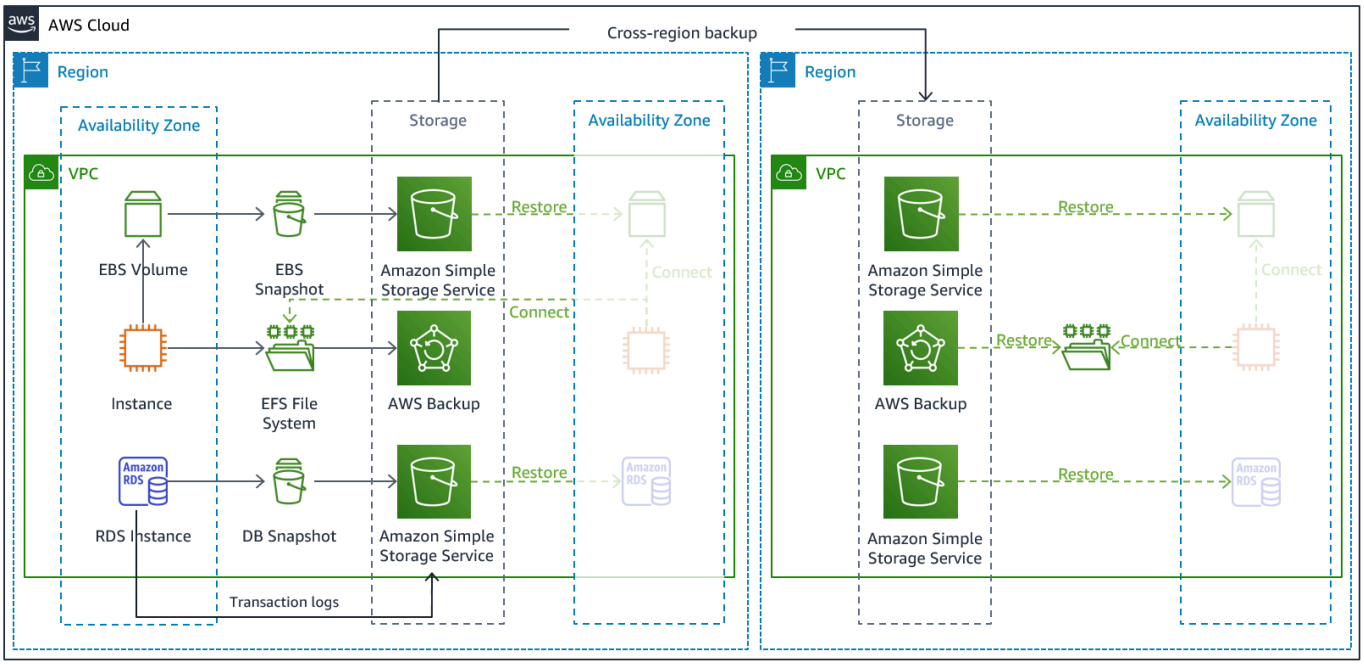

Figura 7 - Arquitetura de backup e restauração

Serviços da AWS

Seus dados de carga de trabalho exigirão uma estratégia de backup que seja executada periodicamente ou contínua. A frequência com que você executa o backup determinará seu ponto de recuperação alcançável (que deve se alinhar para atender ao seu RPO). O backup também deve oferecer uma maneira de restaurá-lo até o momento em que foi feito. O backup com point-in-time recuperação está disponível por meio dos seguintes serviços e recursos:

- Instantâneo do Amazon Elastic Block Store (Amazon EBS)
- Backup do Amazon DynamoDB
- Snapshot do Amazon RDS
- Snapshot de banco de dados Amazon Aurora
- Backup do Amazon EFS (ao usar AWS Backup)
- Snapshot do Amazon Redshift
- Snapshot do Amazon Neptune
- Amazon DocumentDB
- Amazon FSx para Windows File Server, Amazon FSx para Lustre, Amazon FSx para NetApp ONTAP e Amazon FSx para OpenZFS

Para o Amazon Simple Storage Service (Amazon S3), você pode usar o Amazon S3 Cross-Region Replication (CRR) para copiar objetos de forma assíncrona para um bucket do S3 na região de DR continuamente, ao mesmo tempo em que fornece versionamento para os objetos armazenados para que você possa escolher seu ponto de restauração. A replicação contínua de dados tem a vantagem de ser o menor tempo (quase zero) para fazer backup de seus dados, mas pode não proteger contra eventos de desastre, como corrupção de dados ou ataques maliciosos (como exclusão não autorizada de dados), bem como backups. point-in-time A replicação contínua é abordada na seção Serviços da AWS para Pilot Light.

AWS Backupfornece um local centralizado para configurar, programar e monitorar os recursos de backup da AWS para os seguintes serviços e recursos:

- Volumes do Amazon Elastic Block Store (Amazon EBS)
- EC2Instâncias da Amazon

- Bancos de dados Amazon Relational Database Service (Amazon RDS) (incluindo bancos de dados Amazon Aurora)
- Tabelas do Amazon DynamoDB
- Sistemas de arquivos Amazon Elastic File System (Amazon EFS)
- Volumes do AWS Storage Gateway
- Amazon FSx para Windows File Server, Amazon FSx para Lustre, Amazon FSx para NetApp ONTAP e Amazon FSx para OpenZFS

AWS Backup suporta a cópia de backups entre regiões, como para uma região de recuperação de desastres.

Como uma estratégia adicional de recuperação de desastres para seus dados do Amazon S3, habilite o controle de versão de objetos do S3. O controle de versão de objetos protege seus dados no S3 das consequências das ações de exclusão ou modificação, mantendo a versão original antes da ação. O controle de versão de objetos pode ser uma mitigação útil para desastres do tipo erro humano. Se você estiver usando a replicação do S3 para fazer backup de dados na sua região de DR, então, por padrão, quando um objeto é excluído no bucket de origem, o Amazon S3 adiciona um marcador de exclusão somente no bucket de origem. Essa abordagem protege os dados na região de DR contra exclusões maliciosas na região de origem.

Além dos dados, você também deve fazer backup da configuração e da infraestrutura necessárias para reimplantar sua carga de trabalho e atingir seu objetivo de tempo de recuperação (RTO). AWS CloudFormationfornece Infraestrutura como Código (IaC) e permite que você defina todos os recursos da AWS em sua carga de trabalho para que você possa implantar e reimplantar de forma confiável em várias contas e regiões da AWS. Você pode fazer backup de EC2 instâncias da Amazon usadas pela sua carga de trabalho como Amazon Machine Images (AMIs). A AMI é criada a partir de instantâneos do volume raiz da sua instância e de quaisquer outros volumes do EBS anexados à sua instância. Você pode usar essa AMI para executar uma versão restaurada da EC2 instância. Uma AMI pode ser copiada dentro ou entre regiões. Ou você pode usar AWS Backuppara copiar backups entre contas e para outras regiões da AWS. O recurso de backup entre contas ajuda a proteger contra eventos de desastres que incluem ameaças internas ou comprometimento da conta. AWS Backup também adiciona recursos adicionais para EC2 backup — além dos volumes individuais do EBS da instância, AWS Backup também armazena e rastreia os seguintes metadados: tipo de instância, nuvem privada virtual (VPC) configurada, grupo de segurança, função do IAM, configuração de monitoramento e tags. No entanto, esses metadados adicionais são usados somente ao restaurar o EC2 backup na mesma região da AWS.

Todos os dados armazenados na região de recuperação de desastres como backups devem ser restaurados no momento do failover. AWS Backup oferece capacidade de restauração, mas atualmente não permite a restauração programada ou automática. Você pode implementar a restauração automática na região de DR usando o AWS SDK APIs para AWS Backup solicitar. Você pode configurar isso como um trabalho recorrente regular ou acionar a restauração sempre que um backup for concluído. A figura a seguir mostra um exemplo de restauração automática usando o Amazon Simple Notification Service (Amazon AWS LambdaSNS) e. Implementar uma restauração periódica programada de dados é uma boa ideia, pois a restauração de dados do backup é uma operação do plano de controle. Se essa operação não estivesse disponível durante um desastre, você ainda teria armazenamentos de dados operacionais criados a partir de um backup recente.

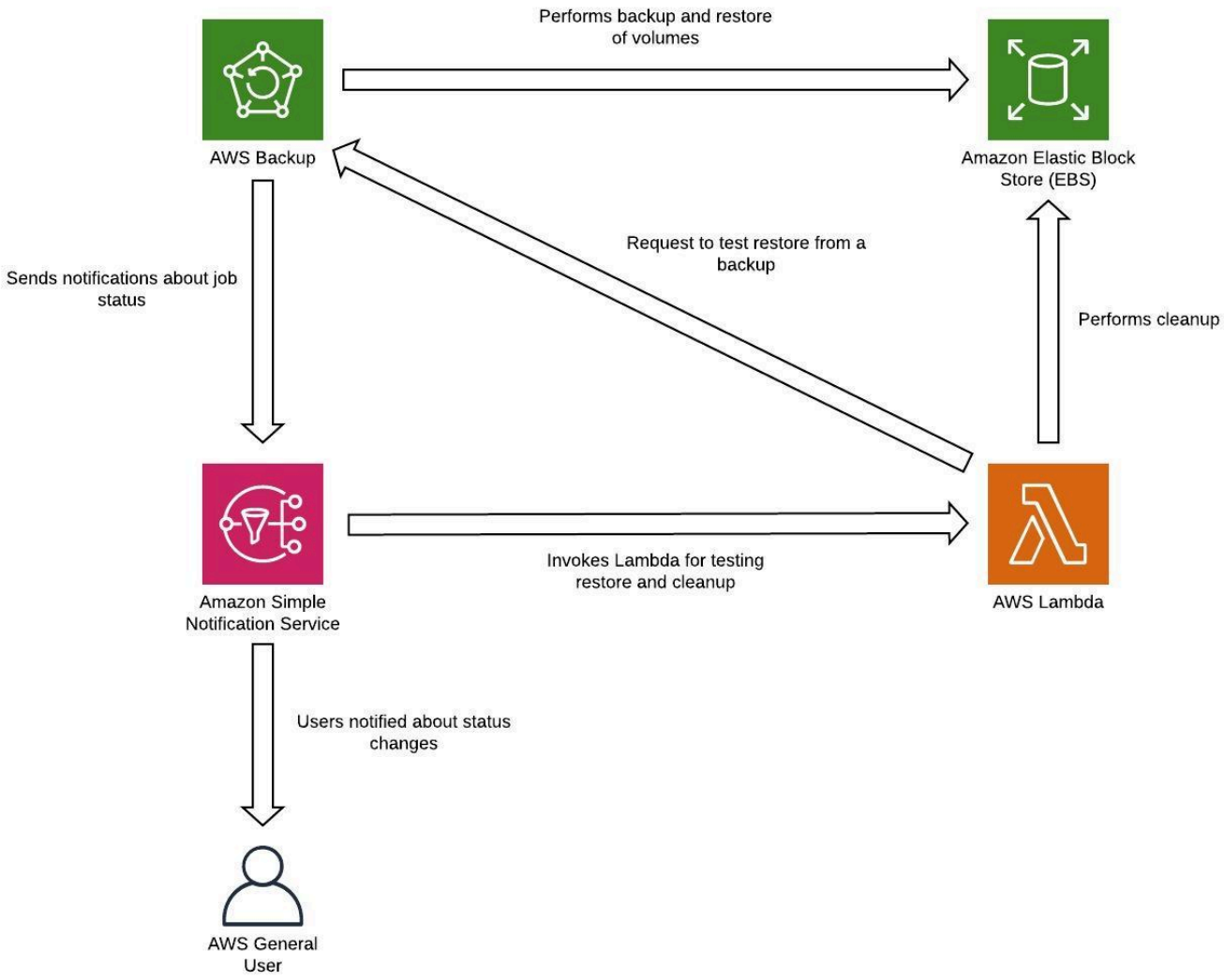

Figura 8 - Restaurando e testando backups

> **Note**
>
> Sua estratégia de backup deve incluir o teste dos backups. Consulte a seção Testando a recuperação de desastres para obter mais informações. Consulte o AWS Well-Architected Lab: Testando o Backup and Restore of Data para ver uma demonstração prática da implementação.

Luz piloto

Com a abordagem piloto leve, você replica seus dados de uma região para outra e provisiona uma cópia da sua infraestrutura principal de carga de trabalho. Os recursos necessários para permitir a replicação e o backup, como bancos de dados e armazenamento de objetos, estão sempre ativos. Outros elementos, como servidores de aplicativos, são carregados com o código e as configurações do aplicativo, mas são “desligados” e são usados somente durante testes ou quando o failover de recuperação de desastres é invocado. Na nuvem, você tem a flexibilidade de desprovisionar recursos quando não precisar deles e provisioná-los quando precisar. Uma prática recomendada para “desligado” é não implantar o recurso e, em seguida, criar a configuração e os recursos para implantá-lo (“ativar”) quando necessário. Diferentemente da abordagem de backup e restauração, sua infraestrutura principal está sempre disponível e você sempre tem a opção de provisionar rapidamente um ambiente de produção em grande escala ativando e expandindo seus servidores de aplicativos.

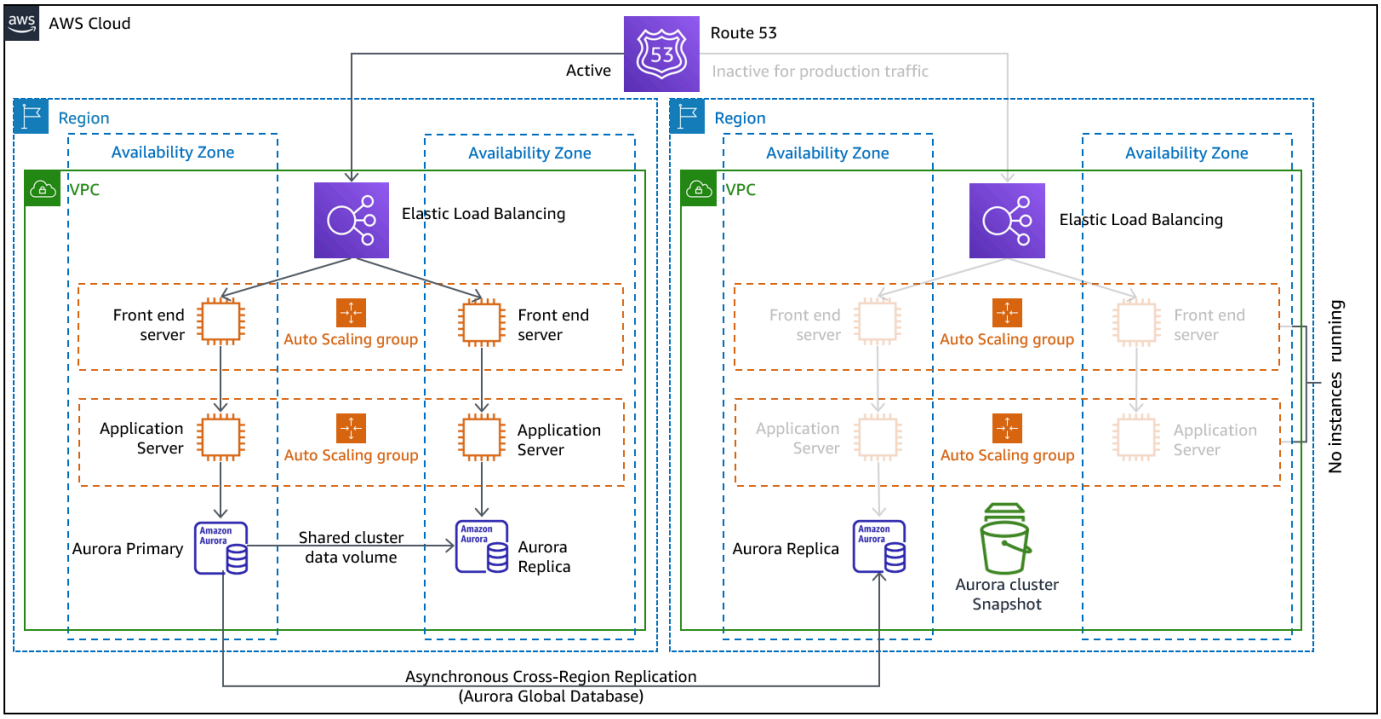

Figura 9: Arquitetura da luz piloto

Uma abordagem piloto leve minimiza o custo contínuo da recuperação de desastres, minimizando os recursos ativos, e simplifica a recuperação no momento de um desastre, pois os principais requisitos de infraestrutura estão todos prontos. Essa opção de recuperação exige que você altere sua abordagem de implantação. Você precisa fazer alterações na infraestrutura principal em cada região e implantar alterações na carga de trabalho (configuração, código) simultaneamente em cada região. Essa etapa pode ser simplificada automatizando suas implantações e usando a infraestrutura como código (IaC) para implantar a infraestrutura em várias contas e regiões (implantação completa da infraestrutura na região primária e implantação de infraestrutura reduzida/desativada nas regiões de DR). É recomendável usar uma conta diferente por região para fornecer o mais alto nível de isolamento de recursos e segurança (no caso de credenciais comprometidas também fazerem parte de seus planos de recuperação de desastres).

Com essa abordagem, você também deve evitar um desastre de dados. A replicação contínua de dados protege você contra alguns tipos de desastres, mas pode não protegê-lo contra corrupção ou destruição de dados, a menos que sua estratégia também inclua o controle de versões dos dados armazenados ou opções de recuperação. point-in-time Você pode fazer backup dos dados replicados na região do desastre para criar point-in-time backups nessa mesma região.

Serviços da AWS

Além de usar os serviços da AWS abordados na seção Backup e restauração para criar point-in-time backups, considere também os seguintes serviços para sua estratégia piloto.

Para fins piloto, a replicação contínua de dados em bancos de dados ativos e armazenamentos de dados na região de DR é a melhor abordagem para baixo RPO (quando usada em adição aos point-in-time backups discutidos anteriormente). A AWS fornece replicação de dados assíncrona, contínua e entre regiões para dados usando os seguintes serviços e recursos:

- Replicação do Amazon Simple Storage Service (Amazon S3)
- Réplicas de leitura do Amazon RDS
- Bancos de dados globais Amazon Aurora
- Tabelas globais do Amazon DynamoDB
- Cluster globais do Amazon DocumentDB
- Armazenamento de dados global para Amazon ElastiCache (Redis OSS)

Com a replicação contínua, as versões de seus dados estão disponíveis quase imediatamente em sua região de DR. Os tempos reais de replicação podem ser monitorados usando recursos de serviço como o S3 Replication Time Control (S3 RTC) para objetos do S3 e recursos de gerenciamento dos bancos de dados globais Amazon Aurora.

Ao fazer o failover para executar sua read/write carga de trabalho na região de recuperação de desastres, você deve promover uma réplica de leitura do RDS para se tornar a instância primária. Para instâncias de banco de dados diferentes do Aurora, o processo leva alguns minutos para ser concluído e a reinicialização faz parte do processo. Para replicação entre regiões (CRR) e failover com o RDS, o uso do banco de dados global Amazon Aurora oferece várias vantagens. O banco de dados global usa uma infraestrutura dedicada que deixa seus bancos de dados totalmente disponíveis para atender seu aplicativo e pode ser replicado para a região secundária com latência típica de menos de um segundo (e dentro de uma região da AWS é muito menos de 100 milissegundos). Com o banco de dados global Amazon Aurora, se sua região principal sofrer uma degradação de desempenho ou uma interrupção, você pode promover que uma das regiões secundárias assuma responsabilidades de leitura/gravação em menos de um minuto, mesmo no caso de uma interrupção regional completa. Você também pode configurar o Aurora para monitorar o tempo de atraso do RPO de todos os clusters secundários para garantir que pelo menos um cluster secundário permaneça dentro da janela de RPO de destino.

Uma versão reduzida de sua infraestrutura principal de carga de trabalho com menos ou menos recursos deve ser implantada em sua região de DR. Usando AWS CloudFormation, você pode definir sua infraestrutura e implantá-la de forma consistente em todas as contas da AWS e em todas as regiões da AWS. AWS CloudFormation usa pseudoparâmetros predefinidos para identificar a conta da AWS e a região da AWS na qual ela está implantada. Portanto, você pode implementar a lógica de condição em seus CloudFormation modelos para implantar somente a versão reduzida de sua infraestrutura na região de DR. Por EC2 exemplo, implantações, uma Amazon Machine Image (AMI) fornece informações como configuração de hardware e software instalado. Você pode implementar um pipeline do Image Builder que cria o que AMIs você precisa e copiá-lo para suas regiões primária e de backup. Isso ajuda a garantir que esses dourados AMIs tenham tudo o que você precisa para reimplantar ou expandir sua carga de trabalho em uma nova região, no caso de um evento de desastre. As EC2 instâncias da Amazon são implantadas em uma configuração reduzida (menos instâncias do que na sua região principal). Para expandir a infraestrutura para suportar o tráfego de produção, consulte Amazon Auto EC2 Scaling na seção Warm Standby.

Para uma active/passive configuração como a luz piloto, todo o tráfego vai inicialmente para a região principal e muda para a região de recuperação de desastres se a região principal não estiver mais disponível. Essa operação de failover pode ser iniciada de forma manual ou automática. O failover iniciado automaticamente com base em verificações de saúde ou alarmes deve ser usado com cautela. Mesmo usando as melhores práticas discutidas aqui, o tempo de recuperação e o ponto de recuperação serão maiores que zero, incorrendo em alguma perda de disponibilidade e de dados. Se você falhar quando não precisar (alarme falso), você incorrerá nessas perdas. Portanto, o failover iniciado manualmente é usado com frequência. Nesse caso, você ainda deve automatizar as etapas para failover para que a inicialização manual ocorra com o apertar um botão.

Há várias opções de gerenciamento de tráfego a serem consideradas ao usar AWS os serviços.

Uma opção é usar o Amazon Route 53. Usando o Amazon Route 53, você pode associar vários endpoints IP em uma ou mais regiões da AWS a um nome de domínio do Route 53. Em seguida, você pode rotear o tráfego para o endpoint apropriado sob esse nome de domínio. No failover, você precisa mudar o tráfego para o endpoint de recuperação e para longe do endpoint primário. As verificações de saúde do Amazon Route 53 monitoram esses endpoints. Usando essas verificações de integridade, você pode configurar o failover de DNS iniciado automaticamente para garantir que o tráfego seja enviado somente para endpoints íntegros, o que é uma operação altamente confiável feita no plano de dados. Para implementar isso usando o failover iniciado manualmente, você pode usar o Amazon Application Recovery Controller (ARC). Com o ARC, você pode criar verificações de saúde do Route 53 que, na verdade, não verificam a integridade, mas atuam como interruptores liga/desliga sobre os quais você tem controle total. Usando a AWS CLI ou o AWS SDK, você pode

criar um script de failover usando essa API de plano de dados altamente disponível. Seu script ativa essas opções (as verificações de saúde do Route 53) dizendo ao Route 53 que envie tráfego para a região de recuperação em vez da região primária. Outra opção para o failover iniciado manualmente que alguns usaram é usar uma política de roteamento ponderado e alterar os pesos das regiões primária e de recuperação para que todo o tráfego vá para a região de recuperação. No entanto, esteja ciente de que essa é uma operação de plano de controle e, portanto, não é tão resiliente quanto a abordagem do plano de dados usando o Amazon Application Recovery Controller (ARC).

Outra opção é usar AWS Global Accelerator. Usando AnyCast IP, você pode associar vários endpoints em uma ou mais regiões da AWS com o mesmo endereço IP público estático ou endereços. AWS Global Accelerator em seguida, encaminha o tráfego para o endpoint apropriado associado a esse endereço. As verificações de integridade do Global Accelerator monitoram os endpoints. Usando essas verificações de saúde, AWS Global Accelerator verifica a integridade de seus aplicativos e encaminha automaticamente o tráfego do usuário para o endpoint íntegro do aplicativo. Para o failover iniciado manualmente, você pode ajustar qual endpoint recebe tráfego usando discagens de tráfego, mas observe que essa é uma operação de plano de controle. O Global Accelerator oferece latências mais baixas para o endpoint do aplicativo, pois usa a extensa rede de borda da AWS para colocar tráfego no backbone da rede da AWS o mais rápido possível. O Global Accelerator também evita problemas de cache que podem ocorrer com sistemas DNS (como o Route 53).

A Amazon CloudFront oferece failover de origem, em que, se uma determinada solicitação para o endpoint primário falhar, CloudFront encaminha a solicitação para o endpoint secundário. Diferentemente das operações de failover descritas anteriormente, todas as solicitações subsequentes ainda vão para o endpoint primário, e o failover é feito a cada solicitação.

AWS Recuperação flexível de desastres

AWS O Elastic Disaster Recovery (DRS) replica continuamente aplicativos hospedados no servidor e bancos de dados hospedados no servidor de qualquer fonte para AWS usar a replicação em nível de bloco do servidor subjacente. O Elastic Disaster Recovery permite que você use uma região Nuvem AWS como alvo de recuperação de desastres para uma carga de trabalho hospedada localmente ou em outro provedor de nuvem e seu ambiente. Ele também pode ser usado para recuperação de desastres de cargas de trabalho AWS hospedadas se elas consistirem somente em aplicativos e bancos de dados hospedados EC2 (ou seja, não no RDS). O Elastic Disaster Recovery usa a estratégia Pilot Light, mantendo uma cópia dos dados e recursos “desligados” em uma Amazon Virtual Private Cloud (Amazon VPC) usada como área de preparação. Quando um evento de failover

é acionado, os recursos em estágios são usados para criar automaticamente uma implantação de capacidade total no Amazon VPC de destino usado como local de recuperação.

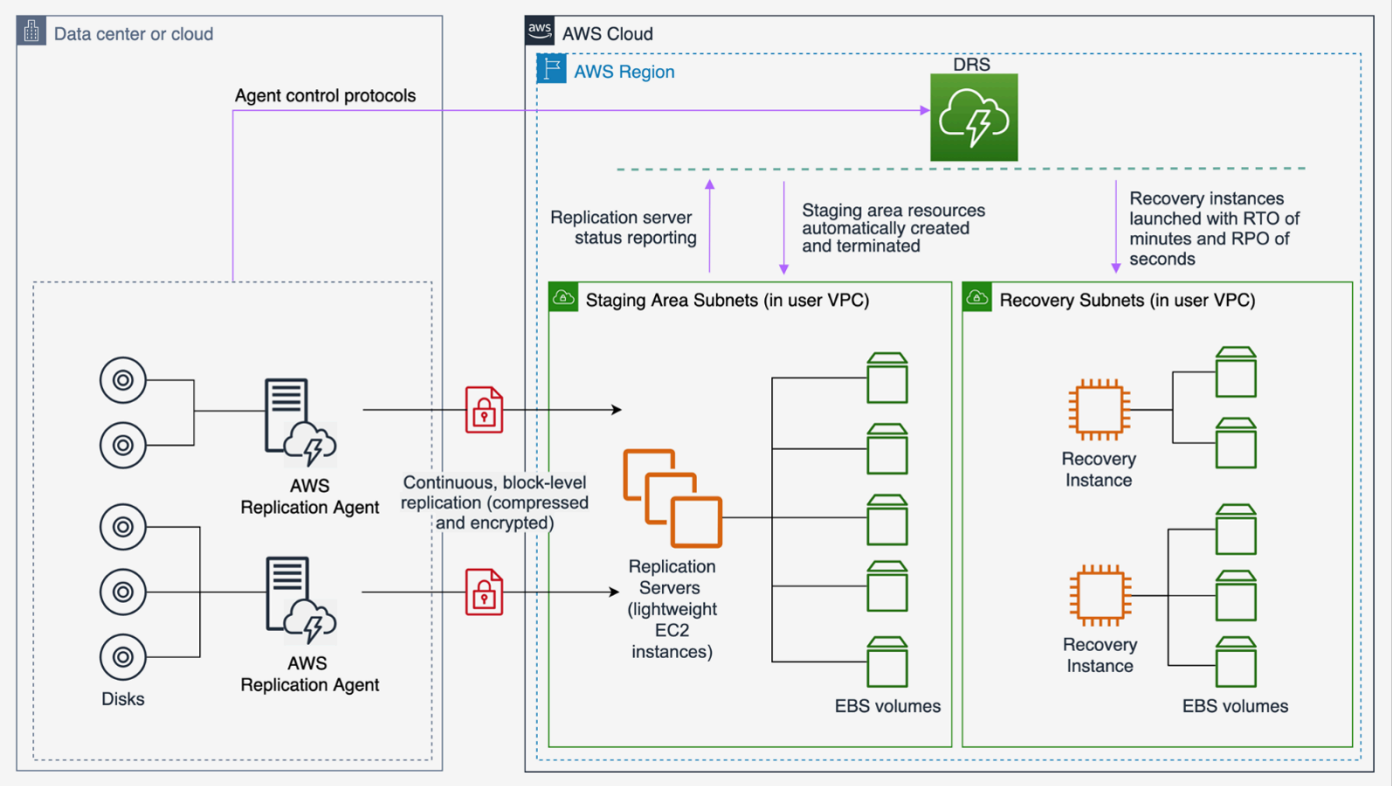

Figura 10 - Arquitetura AWS elástica de recuperação de desastres

Standby passivo

A abordagem de standby passivo envolve garantir que haja uma cópia reduzida, mas totalmente funcional, do seu ambiente de produção em outra região. Essa abordagem estende o conceito de luz piloto e diminui o tempo de recuperação, já que a workload está sempre ativa em outra região. Essa abordagem também permite que você realize testes ou implemente testes contínuos com mais facilidade para aumentar a confiança em sua capacidade de se recuperar de um desastre.

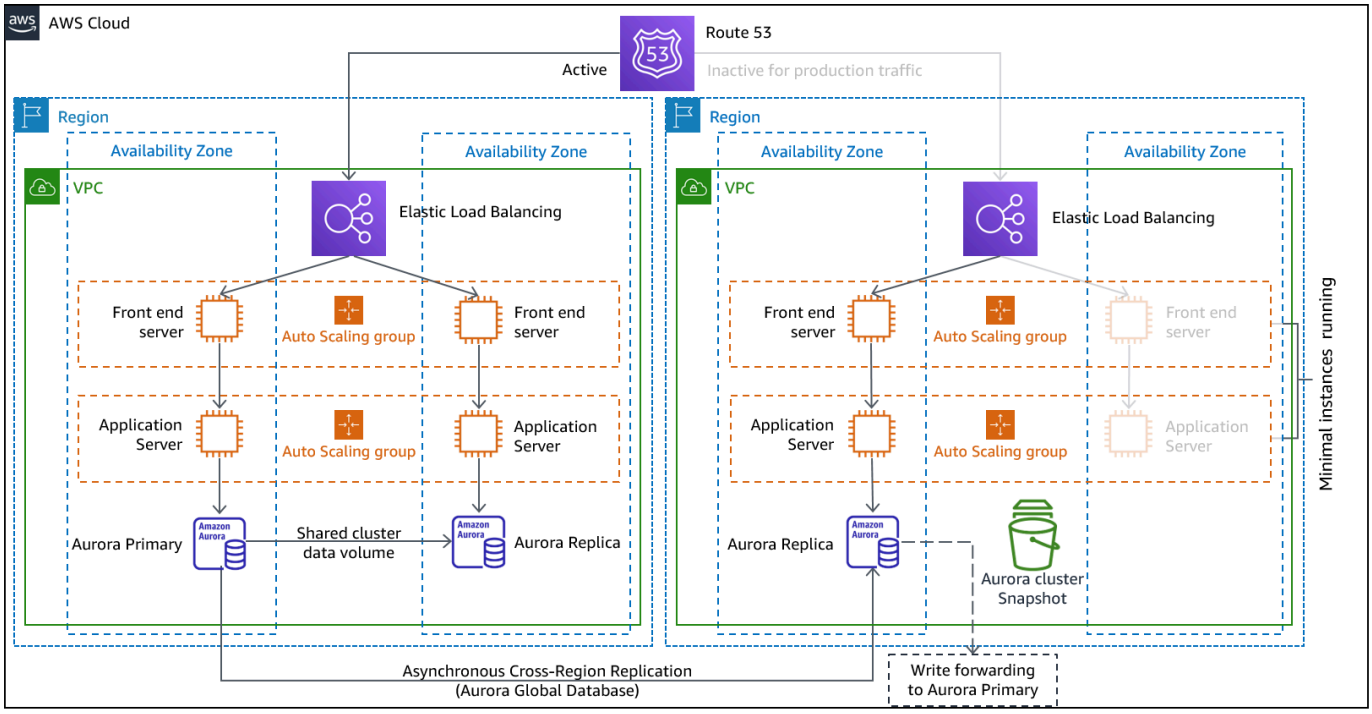

Figura 11 - Arquitetura de espera quente

Nota: Às vezes, a diferença entre a luz piloto e a espera quente pode ser difícil de entender. Ambos incluem um ambiente em sua região de DR com cópias dos ativos de sua região principal. A diferença é que a luz piloto não pode processar solicitações sem que uma ação adicional seja tomada primeiro, enquanto o modo de espera aquecido pode lidar com o tráfego (em níveis de capacidade reduzidos) imediatamente. A abordagem piloto exige que você “ligue” os servidores, possivelmente implante uma infraestrutura adicional (não essencial) e aumente a escala, enquanto o modo de espera aquecido exige apenas que você aumente a escala (tudo já está implantado e em execução). Use suas necessidades de RTO e RPO para ajudá-lo a escolher entre essas abordagens.

Serviços da AWS

Todos os serviços da AWS cobertos por backup e restauração e piloto também são usados em espera quente para backup de dados, replicação de dados, roteamento de active/passive tráfego e implantação de infraestrutura, incluindo instâncias. EC2

O Amazon EC2 Auto Scaling é usado para escalar recursos, incluindo EC2 instâncias da Amazon, tarefas do Amazon ECS, taxa de transferência do Amazon DynamoDB e réplicas do Amazon Aurora em uma região da AWS. O Amazon EC2 Auto Scaling escala a implantação da EC2 instância em

todas as zonas de disponibilidade dentro de uma região da AWS, fornecendo resiliência dentro dessa região. Use o Auto Scaling para expandir sua região de DR até a capacidade total de produção, como parte de uma estratégia piloto de luz ou espera quente. Por exemplo, para EC2, aumente a configuração de capacidade desejada no grupo Auto Scaling. Você pode ajustar essa configuração manualmente por meio do AWS Management Console, automaticamente por meio do AWS SDK ou reimplantando seu AWS CloudFormation modelo usando o novo valor de capacidade desejado. Você pode usar AWS CloudFormation parâmetros para facilitar a reimplantação do CloudFormation modelo. Garanta que as cotas de serviço em sua região de DR estejam definidas como altas o suficiente para não limitar sua escalabilidade até a capacidade de produção.

Como o Auto Scaling é uma atividade do plano de controle, depender dela diminuirá a resiliência de sua estratégia geral de recuperação. É uma troca. Você pode optar por provisionar capacidade suficiente para que a região de recuperação possa lidar com toda a carga de produção conforme implantada. Essa configuração estaticamente estável é chamada de hot standby (consulte a próxima seção). Ou você pode optar por provisionar menos recursos, o que custará menos, mas depender do Auto Scaling. Algumas implementações de DR implantarão recursos suficientes para lidar com o tráfego inicial, garantindo um baixo RTO e, em seguida, confiarão no Auto Scaling para acelerar o tráfego subsequente.

Multissite ativa/ativa

Você pode executar sua carga de trabalho simultaneamente em várias regiões como parte de uma estratégia ativa/ativa de vários sites ou ativa/passiva em espera ativa. O multisite active/active atende ao tráfego de todas as regiões nas quais está implantado, enquanto o hot standby atende ao tráfego somente de uma única região, e as outras regiões são usadas apenas para recuperação de desastres. Com uma active/active abordagem de vários sites, os usuários podem acessar sua carga de trabalho em qualquer uma das regiões em que ela está implantada. Essa abordagem é a mais complexa e cara para a recuperação de desastres, mas pode reduzir seu tempo de recuperação para quase zero na maioria dos desastres com as escolhas tecnológicas e a implementação corretas (no entanto, a corrupção de dados pode precisar depender de backups, o que geralmente resulta em um ponto de recuperação diferente de zero). O hot standby usa uma active/passive configuração em que os usuários são direcionados somente para uma única região e as regiões de DR não recebem tráfego. A maioria dos clientes acha que, se quiserem criar um ambiente completo na segunda região, faz sentido usá-lo ativo/ativo. Como alternativa, se você não quiser usar as duas regiões para lidar com o tráfego de usuários, o Warm Standby oferece uma abordagem mais econômica e operacionalmente menos complexa.

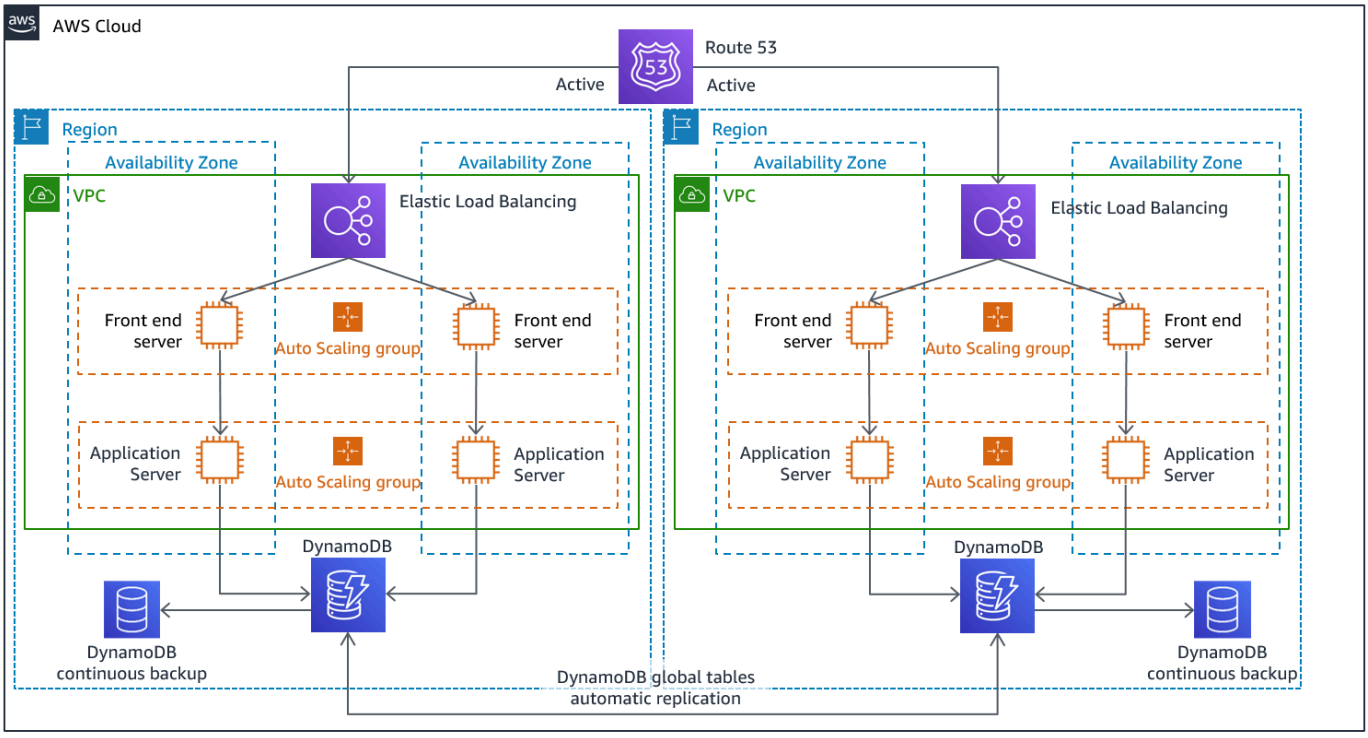

Figura 12 - active/active Arquitetura de vários sites (altere um caminho ativo para inativo para hot standby)

Com a necessidade de uma abordagem de vários locais active/active, because the workload is running in more than one Region, there is no such thing as failover in this scenario. Disaster recovery testing in this case would focus on how the workload reacts to loss of a Region: Is traffic routed away from the failed Region? Can the other Region(s) handle all the traffic? Testing for a data disaster is also required. Backup and recovery are still required and should be tested regularly. It should also be noted that recovery times for a data disaster involving data corruption, deletion, or obfuscation will always be greater than zero and the recovery point will always be at some point before the disaster was discovered. If the additional complexity and cost of a multi-site active/active (ou de espera ativa) para manter tempos de recuperação quase nulos, esforços adicionais devem ser feitos para manter a segurança e evitar erros humanos para mitigar os desastres humanos.

Serviços da AWS

Todos os serviços da AWS cobertos por backup e restauração, iluminação piloto e espera quente também são usados aqui para backup de point-in-time dados, replicação de dados, roteamento de active/active tráfego e implantação e escalabilidade da infraestrutura, incluindo instâncias. EC2

Para os active/passive cenários discutidos anteriormente (Pilot Light e Warm Standby), tanto o Amazon Route 53 quanto o Amazon AWS Global Accelerator podem ser usados para rotear o tráfego de rede para a região ativa. Para a active/active estratégia aqui, esses dois serviços também permitem a definição de políticas que determinam quais usuários acessam qual endpoint regional ativo. Com AWS Global Accelerator você define uma discagem de tráfego para controlar a porcentagem de tráfego que é direcionada para cada endpoint do aplicativo. O Amazon Route 53 suporta essa abordagem percentual e também várias outras políticas disponíveis, incluindo políticas baseadas em geoproximidade e latência. O Global Accelerator aproveita automaticamente a extensa rede de servidores de borda da AWS para integrar o tráfego ao backbone da rede da AWS o mais rápido possível, resultando em menores latências de solicitação.

A replicação assíncrona de dados com essa estratégia permite um RPO quase zero. Os serviços da AWS, como o banco de dados global Amazon Aurora, usam uma infraestrutura dedicada que deixa seus bancos de dados totalmente disponíveis para atender seu aplicativo e podem ser replicados em até cinco regiões secundárias com latência típica de menos de um segundo. With active/passive strategies, writes occur only to the primary Region. The difference with active/active está projetando como a consistência dos dados com gravações em cada região ativa é tratada. É comum projetar leituras de usuários para serem atendidas na região mais próxima a eles, conhecida como leitura local. Com as gravações, você tem várias opções:

- Uma estratégia global de gravação direciona todas as gravações para uma única região. Em caso de falha dessa região, outra região seria promovida a aceitar escritos. O banco de dados global Aurora é uma boa opção para gravação global, pois oferece suporte à sincronização com réplicas de leitura em todas as regiões, e você pode promover uma das regiões secundárias para assumir read/write responsabilidades em menos de um minuto. O Aurora também oferece suporte ao encaminhamento de gravação, o que permite que clusters secundários em um banco de dados global do Aurora encaminhem instruções SQL que realizam operações de gravação para o cluster primário.
- Uma estratégia local de gravação direciona as gravações para a região mais próxima (assim como as leituras). As tabelas globais do Amazon DynamoDB possibilitam essa estratégia, permitindo leitura e gravação de todas as regiões em que sua tabela global está implantada. As tabelas globais do Amazon DynamoDB usam um último escritor para vencer a reconciliação entre atualizações simultâneas.
- Uma estratégia de gravação particionada atribui gravações a uma região específica com base em uma chave de partição (como ID de usuário) para evitar conflitos de gravação. A replicação do Amazon S3 configurada bidirecionalmente pode ser usada para esse caso e atualmente oferece suporte à replicação entre duas regiões. Ao implementar essa abordagem, certifique-se de ativar

a sincronização de modificação da réplica nos buckets A e B para replicar as alterações dos metadados da réplica, como listas de controle de acesso a objetos (ACLs), tags de objetos ou bloqueios de objetos nos objetos replicados. Você também pode configurar se deseja ou não replicar marcadores de exclusão entre buckets em suas regiões ativas. Além da replicação, sua estratégia também deve incluir point-in-time backups para se proteger contra eventos de corrupção ou destruição de dados.

AWS CloudFormation é uma ferramenta poderosa para aplicar a infraestrutura implantada de forma consistente entre contas da AWS em várias regiões da AWS. AWS CloudFormation StackSetsamplia essa funcionalidade ao permitir que você crie, atualize ou exclua CloudFormation pilhas em várias contas e regiões com uma única operação. Embora AWS CloudFormation use YAML ou JSON para definir infraestrutura como código, AWS Cloud Development Kit (AWS CDK)permite definir infraestrutura como código usando linguagens de programação familiares. Seu código é convertido e CloudFormation , em seguida, é usado para implantar recursos na AWS.

Detecção

É importante saber o mais rápido possível que suas cargas de trabalho não estão fornecendo os resultados comerciais que deveriam oferecer. Dessa forma, você pode declarar rapidamente um desastre e se recuperar de um incidente. Para objetivos agressivos de recuperação, esse tempo de resposta, juntamente com as informações apropriadas, é fundamental para atingir os objetivos de recuperação. Se seu objetivo de tempo de recuperação for de uma hora, você precisará detectar o incidente, notificar a equipe apropriada, engajar seus processos de escalonamento, avaliar as informações (se houver) sobre o tempo esperado de recuperação (sem executar o plano de DR), declarar um desastre e se recuperar em uma hora.

> **Note**
>
> Se as partes interessadas decidirem não invocar a DR, mesmo que o RTO esteja em risco, reavalie os planos e objetivos da DR. A decisão de não invocar os planos de DR pode ser porque os planos são inadequados ou há falta de confiança na execução.

É fundamental considerar a detecção, notificação, escalonamento, descoberta e declaração de incidentes em seu planejamento e objetivos para fornecer objetivos realistas e alcançáveis que forneçam valor comercial.

A AWS publica nossa maioria das up-to-the-minute informações sobre disponibilidade de serviços no Service Health Dashboard. Verifique a qualquer momento para obter informações de status atuais ou assine um feed RSS para ser notificado sobre interrupções em cada serviço individual. Se você estiver enfrentando um problema operacional em tempo real com um de nossos serviços que não é exibido no Service Health Dashboard, você pode criar uma Solicitação de Suporte.

AWS Health DashboardFornece informações sobre AWS Health eventos que podem afetar sua conta. As informações são apresentadas de duas formas: um painel que mostra eventos recentes e futuros organizados por categoria, e um log de eventos completo que mostra todos os eventos dos últimos 90 dias.

Para os requisitos de RTO mais rigorosos, você pode implementar o failover automatizado com base em verificações de integridade. Crie verificações de integridade que sejam representativas da experiência do usuário e baseadas em indicadores-chave de desempenho. As verificações profundas de saúde exercem as principais funcionalidades de sua carga de trabalho e vão além das

verificações superficiais de batimentos cardíacos. Use verificações de saúde detalhadas com base em vários sinais. Tenha cuidado com essa abordagem para não acionar alarmes falsos, pois falhar quando não é necessário pode, por si só, introduzir riscos de disponibilidade.

Teste da recuperação de desastres

Teste a implementação da recuperação de desastres para validar a implementação e teste regularmente o failover na região de DR da sua carga de trabalho para garantir que o RTO e o RPO sejam atendidos.

Um padrão a ser evitado é desenvolver caminhos de recuperação que raramente são executados. Por exemplo, você pode ter um datastore secundário utilizado para consultas somente leitura. Quando você grava em um datastore e o datastore primário falha, pode ser necessário fazer o failover para o repositório de dados secundário. Se você não testar esse failover com frequência, poderá descobrir que suas suposições sobre as capacidades do datastore secundário são incorretas. A capacidade do secundário, que pode ter sido suficiente quando você testou pela última vez, pode não ser mais capaz de tolerar a carga nesse cenário, ou as cotas de serviço na região secundária podem não ser suficientes.

Nossa experiência mostrou que a única recuperação de erro que funciona é o caminho testado com frequência. Essa é a razão pela qual é melhor ter um pequeno número de caminhos de recuperação.

Você pode estabelecer padrões de recuperação e testá-los regularmente. Se você tiver um caminho de recuperação complexo ou crítico, ainda precisará executar regularmente essa falha na produção para validar se o caminho de recuperação funciona.

Gerencie o desvio de configuração na região de DR. Garanta que sua infraestrutura, dados e configuração estejam conforme necessário na região de DR. Por exemplo, verifique se as AMIs cotas de serviço são up-to-date.

Você pode usá-lo AWS Configpara monitorar e registrar continuamente suas configurações de recursos da AWS. AWS Config pode detectar desvios e acionar o AWS Systems Manager Automation para corrigir alarmes de desvio e aumento. AWS CloudFormationtambém pode detectar desvios nas pilhas que você implantou.

Conclusão

Os clientes são responsáveis pela disponibilidade de seus aplicativos na nuvem. É importante definir o que é um desastre e ter um plano de recuperação de desastres que reflita essa definição e o impacto que ela pode ter nos resultados comerciais. Crie o Recovery Time Objective (RTO) e o Recovery Point Objective (RPO) com base na análise de impacto e nas avaliações de risco e, em seguida, escolha a arquitetura apropriada para mitigar os desastres. Garanta que a detecção de desastres seja possível e oportuna — é vital saber quando os objetivos estão em risco. Certifique-se de ter um plano e valide o plano com testes. Os planos de recuperação de desastres que não foram validados correm o risco de não serem implementados devido à falta de confiança ou à falha em cumprir os objetivos de recuperação de desastres.

Colaboradores

Os colaboradores deste documento incluem:

- Alex Livingstone, líder prático de operações de nuvem, AWS Enterprise Support
- Seth Eliot, arquiteto principal de soluções de confiabilidade, Amazon Web Services

Outras fontes de leitura

Para obter informações adicionais, consulte:

- AWS Centro de Arquitetura
- Pilar de confiabilidade, AWS Well-Architected Framework
- Lista de verificação do plano de recuperação de desastres
- Implementando verificações de saúde
- Arquitetura de recuperação de desastres (DR) na AWS, parte I: Estratégias para recuperação na nuvem
- Arquitetura de recuperação de desastres (DR) na AWS, parte II: Backup e restauração com recuperação rápida
- Arquitetura de recuperação de desastres (DR) na AWS, parte III: luz piloto e espera quente
- Arquitetura de recuperação de desastres (DR) na AWS, parte IV: ativo/ativo em vários sites
- Criar mecanismos de recuperação de desastres usando o Amazon Route 53
- Minimizar dependências em um plano de recuperação de desastres
- Laboratórios práticos de recuperação de AWS desastres Well-Architected
- AWS Implementações de soluções: Multi-Region Application Architecture
- AWS re:Invent 2018: Padrões de arquitetura para aplicativos ativo-ativos em várias regiões (09-R2) ARC2

Histórico do documento

Para ser notificado sobre atualizações desse whitepaper, inscreva-se no feed RSS.

Alteração	Descrição	Data
Atualizações menores	Correções de bugs e várias pequenas alterações por toda parte.	1.º de abril de 2022
Whitepaper atualizado	Pequenas atualizações editoriais.	21 de março de 2022
Whitepaper atualizado	Informações adicionadas sobre o plano de dados e o plano de controle. Foram adicionados mais detalhes sobre como implementar o active/passive failover. Substituiu a CloudEndure recuperação de desastres pela recuperação de desastres AWS elástica	17 de fevereiro de 2022
Atualização secundária	AWS Well-Architected Tool informações adicionadas.	11 de fevereiro de 2022
Publicação inicial	Whitepaper publicado pela primeira vez.	12 de fevereiro de 2021

Avisos

Os clientes são responsáveis por fazer uma avaliação independente das informações contidas neste documento. Este documento: (a) serve apenas para fins informativos, (b) representa as práticas e ofertas atuais de produtos da AWS, que estão sujeitas a alterações sem aviso prévio, e (c) não cria nenhum compromisso ou garantia por parte da AWS e de seus afiliados, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos “no estado em que se encontram”, sem garantias, representações ou condições de qualquer tipo, expressas ou implícitas. As responsabilidades e as obrigações da AWS para com os clientes são controladas por contratos da AWS, e este documento não faz parte nem modifica nenhum contrato entre a AWS e seus clientes.

AWS Glossário

Para obter a AWS terminologia mais recente, consulte o AWS glossário na Glossário da AWS Referência.

As traduções são geradas por tradução automática. Em caso de conflito entre o conteúdo da tradução e da versão original em inglês, a versão em inglês prevalecerá.