Preparação para eventos de infraestrutura

Melhores práticas e diretivas da AWS

Julho de 2017

Avisos

Este documento é fornecido apenas para fins informativos. Ele relaciona as atuais ofertas de produtos e práticas da AWS na data de emissão deste documento, que estão sujeitas a alterações sem aviso prévio. Os clientes são responsáveis por fazer sua própria avaliação independente das informações neste documento e de qualquer uso dos produtos ou serviços da AWS, cada um dos quais é fornecido “como está”, sem garantia de qualquer tipo, expressa ou implícita. Este documento não cria quaisquer garantias, representações, compromissos contratuais, condições ou seguros da AWS, suas afiliadas, fornecedores ou licenciadores. As responsabilidades e obrigações da AWS para com seus clientes são controladas por contratos da AWS, e este documento não modifica nem faz parte de qualquer contrato entre a AWS e seus clientes.

Sumário

Resumo

Este artigo técnico descreve as diretrizes e práticas recomendadas para clientes com cargas de trabalho de produção implantadas no Amazon Web Services (AWS) que desejam projetar e provisionar seus aplicativos baseados em nuvem para lidar com eventos de escalabilidade planejada, como, por exemplo, lançamentos de produtos ou picos de tráfego sazonais, de modo perfeito e dinâmico. Abordamos princípios gerais de projeto e fornecemos orientações e práticas recomendadas específicas em várias áreas conceituais de planejamento de eventos de infraestrutura. Em seguida, descrevemos considerações e práticas de preparação operacional, e atividades pós-evento.

Introdução

A preparação de infraestrutura para eventos tem a ver com projetar e preparar para eventos significativos previstos, que têm um impacto sobre o seu negócio. Esses são eventos durante os quais é extremamente importante que o serviço web da empresa seja confiável, responda rapidamente e seja altamente tolerante a falhas, em todas as condições e alterações em padrões de tráfego. Esses eventos podem incluir a expansão para novas regiões, lançamentos de produtos ou recursos novos, eventos sazonais ou anúncios de negócios, ou eventos de marketing significativos.

Um evento de infraestrutura que não seja adequadamente planejado pode ter um impacto negativo para a reputação, continuidade ou finanças de sua empresa. Falhas em eventos de infraestrutura podem assumir a forma de falhas imprevistas no serviço, degradação de desempenho relacionado a carga, latência de rede, limitações de capacidade de armazenamento, limites de sistema, como taxas de chamada de API, quantidades finitas de endereços IP disponíveis, compreensão insatisfatória dos comportamentos dos componentes de uma pilha de aplicativos devido à insuficiência de monitoramento, dependências imprevistas em um serviço de terceiros ou componente não configurado para redimensionamento, ou alguma outra condição de erro imprevista.

Para minimizar o risco de falhas imprevistas durante um evento importante, as empresas devem investir tempo e recursos para planejar e preparar, treinar os funcionários, projetar e documentar os processos relevantes. A quantidade de investimento no planejamento de eventos de infraestrutura para um determinado aplicativo ou conjunto de aplicativos habilitados para nuvem pode variar de acordo com a complexidade e o alcance global do sistema. Independentemente do escopo ou complexidade da presença de uma empresa na nuvem, os princípios de projeto e a orientação de práticas recomendadas fornecidos neste artigo técnico são os mesmos.

Com o Amazon Web Services (AWS), a sua empresa pode redimensionar sua infraestrutura em preparação para um evento de escalabilidade planejado, de modo dinâmico, adaptável e com pagamento conforme o uso. O rico conjunto de produtos e serviços elásticos e programáveis da Amazon oferece à sua empresa acesso à mesma infraestrutura altamente segura, confiável e rápida que a Amazon utiliza para execução de sua própria rede global, e permite que a sua

empresa adapte-se com agilidade em resposta a seus próprios requisitos de negócios que mudam rapidamente.

Este artigo técnico descreve as melhores práticas e princípios de projeto para orientar o planejamento e a execução de eventos de infraestrutura e mostra como você pode usar os serviços da AWS para garantir que seus aplicativos estejam prontos para aumentar de escala e serem multiplicados de acordo com as necessidades dos negócios.

Planejamento de preparação para evento de infraestrutura

Esta seção descreve o que constitui um evento de infraestrutura planejado e os tipos de atividades que normalmente ocorrem durante tal evento.

O que é um evento de infraestrutura planejado?

Evento de infraestrutura planejado é uma janela de evento prevista e programada, orientada aos negócios, durante a qual é essencial para os negócios manter um serviço web altamente responsivo, dimensionável e tolerante a falhas. Essa necessidade pode ser motivada por campanhas de marketing, eventos de mídia relacionados à linha de negócios da empresa, lançamentos de produtos, expansões geográficas ou qualquer atividade semelhante que resulte em tráfego adicional para os aplicativos baseados na web e para a infraestrutura subjacente da empresa.

O que acontece durante um evento de infraestrutura planejado?

A principal preocupação na maioria dos eventos de infraestrutura planejados é poder adicionar capacidade à sua infraestrutura da web para atender a maiores demandas de tráfego. Em um ambiente tradicional no local físico provisionado com recursos de computação, armazenamento e rede, o departamento de TI de uma empresa teria que provisionar capacidade adicional com base em suas melhores estimativas de um pico máximo teórico. Isso incorre no risco de provisionamento de capacidade insuficiente e a empresa pode sofrer perda de negócios devido à sobrecarga dos servidores web, tempos de resposta lentos e outros erros em tempo de execução.

Dentro da nuvem da AWS, a infraestrutura é programável e elástica. Isso significa que ela pode ser provisionada rapidamente em resposta à demanda em tempo real. Também significa que ela pode ser configurada para responder às métricas do sistema de modo automatizado, inteligente e dinâmico – aumentando ou diminuindo recursos como clusters de servidores web, throughput provisionado, capacidade de armazenamento, núcleos de computação disponíveis, número de fragmentos de streaming e assim por diante, conforme necessário.

Além disso, muitos serviços da AWS são totalmente gerenciados. Esses serviços incluem armazenamento, banco de dados, análise, aplicativos e serviços de implantação. Isso significa que os clientes da AWS não precisam se preocupar com as complexidades de configuração desses serviços para o caso de tráfego intenso. Os serviços totalmente gerenciados da AWS são projetados para oferecer escalabilidade e alta disponibilidade.

Normalmente, na preparação para um evento de infraestrutura planejado, os clientes da AWS realizam uma análise do sistema para avaliar sua arquitetura de aplicativos e disponibilidade operacional, considerando a escalabilidade e a tolerância a falhas. As estimativas de tráfego são levadas em conta e comparadas com o desempenho em atividades normais de negócios, e métricas e estimativas da capacidade adicional necessária são determinadas. Quaisquer potenciais gargalos e dependências upstream e downstream de terceiros são identificados e resolvidos. Se o evento planejado incluir expansão territorial ou introdução de novos públicos, a questão geográfica também será considerada. A expansão para outras regiões ou zonas de disponibilidade da AWS é realizada antes do evento planejado. Uma análise das configurações dinâmicas do sistema AWS do cliente, como Auto Scaling, balanceamento de carga, roteamento geográfico, alta disponibilidade e medidas de failover, também é realizada para assegurar que elas estão configuradas para lidar corretamente com o espera o aumento de volume e de taxa de transações esperado. Configurações estáticas, como limites de recursos do AWS e localização dos servidores de origem da Content Delivery Network (CDN, Rede de entrega de conteúdo) também são consideradas e modificadas de acordo com a necessidade.

Mecanismos de monitoramento e notificação também são analisados e aprimorados, conforme necessário, para fornecer transparência em tempo real dos eventos enquanto eles ocorrem e para análise posterior, após a conclusão do evento planejado.

Durante o evento planejado, os clientes da AWS, também poderão abrir casos de suporte com a AWS se houver necessidade de solução de problemas ou suporte em tempo real, como no caso de um servidor inoperante. Se uma rápida resposta for necessária, os clientes que assinam o plano Enterprise Support da AWS terão a flexibilidade adicional de poder falar com engenheiros de suporte imediatamente e abrir casos com gravidade crítica.

Após o evento, os recursos da AWS são projetados para serem reduzidos automaticamente para níveis apropriados aos níveis de tráfego, ou continuar a aumentar de escala, conforme ditado pelos eventos.

Princípios de projeto

A preparação para eventos planejados começa com um bom projeto no início de qualquer implementação de uma pilha de aplicativos ou carga de trabalho baseadas na nuvem.

Cargas de trabalho separadas

Um bom projeto é essencial para o gerenciamento eficaz das cargas de trabalho de eventos planejados, tanto em níveis de tráfego normais quanto elevados. Certifique-se, desde o início, de projetar agrupamentos funcionais separados e independentes dos recursos centrados em um aplicativo ou produto específico da empresa. Esta seção descreve as várias dimensões desse objetivo de projeto.

Adicionar tags

As tags são usadas para identificar e organizar os recursos. Elas são um componente essencial do gerenciamento de recursos de infraestrutura durante um evento de infraestrutura planejado. Na AWS, as tags são etiquetas de importância essencial, gerenciadas pelo cliente, aplicadas a um recurso gerenciado individual, como um balanceador de carga ou uma instância da Amazon Elastic Compute Cloud (EC2). Consultando as tags bem definidas que foram anexadas aos recursos da AWS, você pode facilmente identificar quais são os recursos de sua infraestrutura geral que compõem a carga de trabalho do seu evento planejado. Em seguida, usando essas informações, você pode fazer o trabalho de análise para sua preparação. As tags também podem ser usadas para fins de alocação de custos.

As tags podem ser usadas para organizar instâncias EC2, imagens Amazon Machine Image (AMI), balanceadores de carga, grupos de segurança, recursos do Amazon Relational Database Service (RDS), recursos da Amazon Virtual Private Cloud (VPC), verificações de estado do Amazon Route 53 e buckets do Amazon Simple Storage Service (S3), por exemplo.

Para obter mais informações sobre as estratégias eficazes de colocação de tags, consulte Estratégias de adição de tags da AWS.[1]

Para obter exemplos de como criar e gerenciar tags e colocá-las em grupos de recursos, consulte Grupos de recursos e adição de tags para AWS.[2]

Acoplamento fraco

Ao definir uma arquitetura para a nuvem, você deve projetar cada componente de sua pilha de aplicativos para operar da forma mais independente possível dos outros. Assim, as cargas de trabalho baseadas na nuvem terão a vantagem de resiliência e escalabilidade.

Você pode reduzir as interdependências entre os componentes de uma pilha de aplicativos baseada em nuvem projetando cada componente como uma caixa preta com interfaces bem definidas para entradas e saídas (por exemplo, APIs RESTful). Quando os componentes não são aplicativos, e sim serviços que juntos formam um aplicativo, isso é conhecido como uma arquitetura de *microsserviços*. Para a comunicação e a coordenação entre componentes de aplicativo, você pode usar mecanismos de notificação orientados por eventos, como as filas de mensagens do AWS, para transmitir mensagens entre os componentes, conforme mostrado na figura 1.

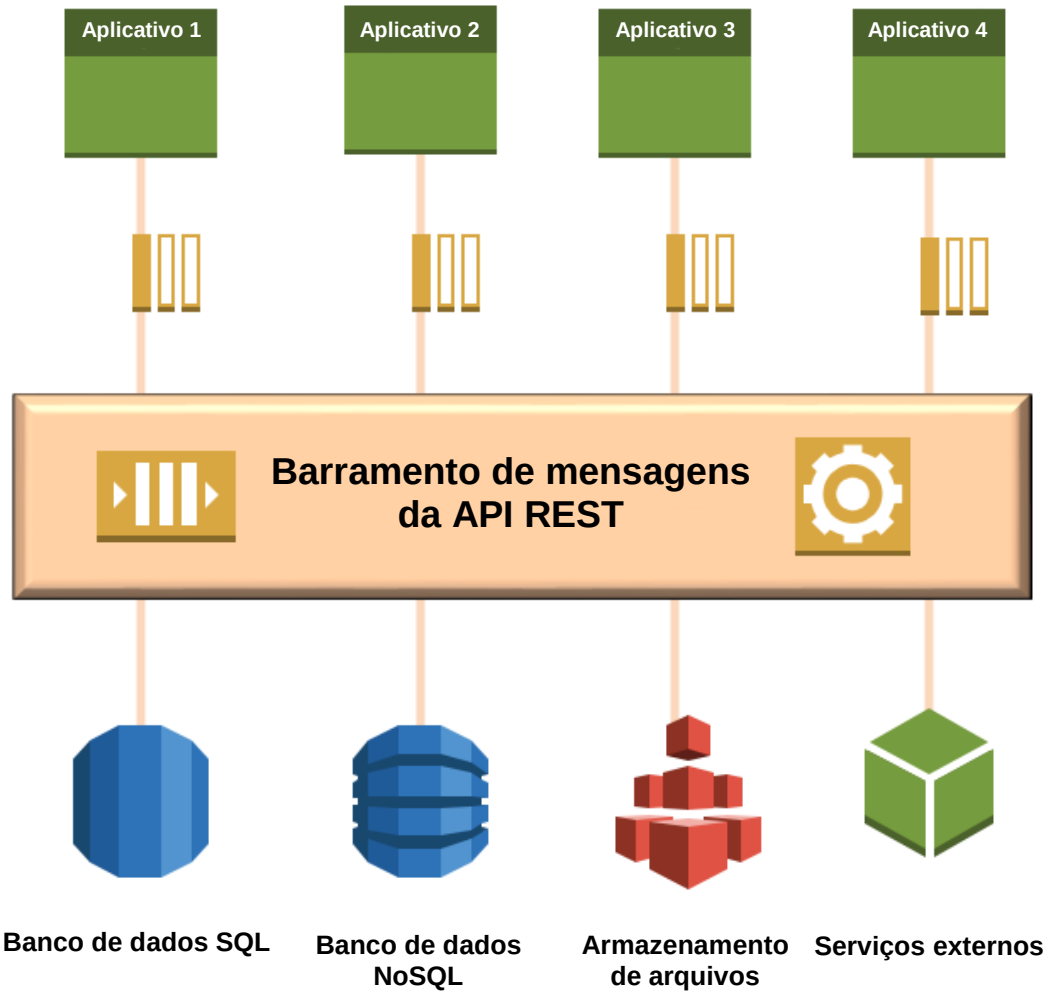

Figura 1. Acoplamento fraco usando interfaces RESTful e filas de mensagens

Usando mecanismos como esses, uma alteração ou uma falha em um componente tem muito menos chance de cascatear para outros componentes. Por exemplo, se um servidor de pilha de aplicativos de vários níveis deixa de responder, os aplicativos que estão acoplados de forma fraca podem ser projetados para contornar o nível que não está respondendo ou alternar para transações alternativas em modo degradado.

Componentes de aplicativos acoplados de forma fraca usando filas de mensagens intermediárias também podem ser mais facilmente projetados para integração assíncrona. Como os componentes de um aplicativo não empregam comunicação direta ponto a ponto, mas usam uma camada de mensagens persistente e intermediária (por exemplo, uma fila do Amazon Simple Queue Service (SQS) ou um mecanismo de dados de streaming, como o Amazon

Kinesis Streams), eles podem suportar aumentos repentinos na atividade em um componente enquanto os componentes downstream processam a fila de entrada. Ou, se houver uma falha de um componente, as mensagens persistirão nas filas ou streams até que o componente com falha possa ser recuperado.

Para obter mais informações sobre serviços de enfileiramento e notificação de mensagem oferecidos pela AWS, consulte o Amazon Simple Queue Service.[3]

Serviços, não servidores

Serviços gerenciados e endpoints de serviço livram você da preocupação com segurança ou acesso, backpus ou restaurações, gerenciamentos de correções ou controle de alterações, configuração de monitoramento ou relatório e da necessidade de administrar muitos dos detalhes de gerenciamento dos sistemas tradicionais. É possível pré-provisionar esses recursos de nuvem para alta disponibilidade e resiliência, usando várias configurações de zona de disponibilidade (ou, em alguns casos, várias regiões). Eles podem ser aumentados ou diminuídos, frequentemente sem necessidade de tempo de inatividade, e você pode configurá-los rapidamente por meio do AWS Management Console ou de chamadas de API/CLI.

Os serviços gerenciados e os endpoints de serviço podem ser usados para potencializar pilhas de aplicativos do cliente com recursos como sistemas de bancos de dados relacionais e NoSQL, armazéns de dados, notificação de eventos, armazenamento de arquivos e objetos, streaming em tempo real, análise de big data, aprendizagem de máquina, pesquisa, transcodificação e muitos outros. Endpoint é um URL que serve como ponto de entrada para um serviço da AWS. Por exemplo, https://dynamodb.us-west-2.amazonaws.com é um ponto de entrada para o serviço Amazon DynamoDB.

Usando serviços gerenciados e seus endpoints de serviço, você pode aproveitar o poder de recursos prontos para produção como parte de sua solução de projeto para lidar com o aumento de volume, de alcance e de taxas de transações que ocorre durante um evento de infraestrutura planejado. Você não precisa provisionar e administrar seus próprios servidores que executam as mesmas funções que os serviços gerenciados.

Para obter mais informações sobre os endpoints de serviço da AWS, consulte Regiões e endpoints de AWS.[4] Consulte também Amazon EMR,[5] Amazon RDS[6] e Amazon ECS[7] para obter exemplos de serviços gerenciados que têm endpoints.

Arquiteturas sem servidor

A utilização do AWS Lambda é outra estratégia que pode atender eficazmente à necessidade de responder a cargas de processamento que se alteram dinamicamente durante um evento de infraestrutura planejado. O Lambda é uma plataforma de computação sem servidor, baseada em eventos. É um serviço chamado dinamicamente que executa código Python, Node.js ou Java em resposta a eventos (por meio de notificações) e gerencia automaticamente os recursos de computação especificados por esse código. O Lambda não exige o pré-provisionamento de recursos de computação do Amazon EC2. O Amazon Simple Notification Service (Amazon SNS) pode ser configurado para disparar as funções do Lambda. Para obter mais detalhes sobre o Amazon SNS, consulte Amazon Push Notification Service.[8]

As funções de servidor do Lambda podem executar código que acessa ou chama outros serviços da AWS, como operações do banco de dados, transformações de dados, recuperação de objetos e arquivos ou, até mesmo, dimensionamento das operações em resposta a eventos externos ou métricas de carga do sistema internas. O AWS Lambda também pode gerar novas notificações ou eventos próprios e até mesmo iniciar outras funções do Lambda.

O AWS Lambda oferece a capacidade de exercer um controle mais preciso sobre as operações de escalabilidade durante um evento de infraestrutura planejado. Por exemplo, o Lambda pode ser usado para estender a funcionalidade das operações de Auto Scaling para executar ações como notificar os sistemas de terceiros de que eles também precisam se redimensionar ou para adicionar mais interfaces de rede a novas instâncias à medida que elas são provisionadas. Consulte Usando o AWS Lambda com ganchos de ciclo vida de Auto Scaling[9] para obter exemplos de como usar o Lambda para personalizar as operações de dimensionamento.

Para obter mais informações sobre o AWS Lambda, consulte O que é o AWS Lambda?[10]

Automação

Auto Scaling

Um componente essencial do planejamento de eventos de infraestrutura é o Auto Scaling. Ser capaz de aumentar e diminuir automaticamente a capacidade de um aplicativo para mais ou para menos, de acordo com condições

predefinidas, ajuda a manter a disponibilidade do aplicativo durante as variações de padrões de tráfego e volume que ocorrem em um evento de infraestrutura planejada.

A AWS fornece Auto Scaling para muitos de seus recursos, incluindo instâncias EC2, capacidade de banco de dados, contêineres, etc.

O Auto Scaling pode ser usado para aumentar ou reduzir agrupamentos de instâncias, como um grupo de servidores que compreende um aplicativo baseado em nuvem, para que eles sejam dimensionados automaticamente segundo critérios especificados. O Auto Scaling também pode ser usado para manter um número fixo de instâncias, mesmo quando o estado de uma instância se deteriora. Essa escalabilidade e manutenção automáticas do número de instâncias são a funcionalidade central do serviço de Auto Scaling.

O Auto Scaling mantém o número de instâncias que você especifica executando verificações periódicas da integridade das instâncias do grupo. Se o estado de uma instância se deteriora, o grupo encerra essa instância e inicia outra instância para substituí-la.

Pode-se usar políticas de Auto Scaling para aumentar ou reduzir automaticamente o número de instâncias EC2 em execução em um grupo de servidores para atender mudanças de condições. Quando a política de escalabilidade está em vigor, o grupo de Auto Scaling ajusta a capacidade desejada do grupo e inicia ou encerra instâncias, conforme necessário, seja dinamicamente ou de acordo com uma agenda, se existir um fluxo e refluxo conhecido e previsível de tráfego.

Reinício e recuperação

Um elemento de projeto importante em qualquer evento de infraestrutura planejado é ter procedimentos e automação implantados para lidar com instâncias ou servidores comprometidos e poder recuperá-los ou reiniciá-los sem interrupções.

Instâncias EC2 podem ser configuradas para recuperação automática quando uma verificação de status do sistema do hardware subjacente apresentar falha. A instância será reinicializada (em hardware novo, se necessário), mas manterá sua ID de instância, endereço IP, endereços IP do Elastic, anexos de volume do Amazon Elastic Block Store (EBS) e outros detalhes de configuração. Para obter

mais informações sobre a recuperação automática de instâncias EC2, consulte Recuperação automática do Amazon EC2.[11]

Gerenciamento/orquestração de configuração

A incorporação de ferramentas de gerenciamento e orquestração de configuração para gerenciamento do estado de recursos individuais e implantação de pilha de aplicativo é parte essencial de uma estratégia de eventos de infraestrutura planejados robusta, confiável e responsiva.

As ferramentas de configuração normalmente lidam com o provisionamento e a configuração de instâncias de servidor, balanceadores de carga, Auto Scaling, implementação de aplicativos individuais e monitoramento da integridade dos aplicativos. Elas também permitem a integração com serviços adicionais, como bancos de dados, volumes de armazenamento e camadas de armazenamento em cache.

As ferramentas de orquestração, uma camada de abstração acima do gerenciamento da configuração, fornecem os meios para especificação dos relacionamentos desses vários recursos, o que permite que os clientes provisionem e gerenciem vários recursos como uma infraestrutura de aplicativo em nuvem unificada, sem se preocuparem com as dependências de recursos.

Como essas ferramentas definem e descrevem recursos individuais e seus relacionamentos como código, esse código pode ser controlado por versão, facilitando a possibilidade de reverter para versões anteriores ou de experimentar novas ramificações de código para fins de teste e desenvolvimento. Também é possível definir orquestrações e configurações otimizadas para um evento de infraestrutura e reverter para a configuração padrão após o evento.

A Amazon Web Services recomenda as seguintes ferramentas para possibilitar hardware como implementações e orquestrações de código:

- **AWS Config com regras de configuração** ou um parceiro de configuração da AWS para fornecer um inventário detalhado, visual e pesquisável dos recursos, histórico de configuração e conformidade da configuração de recursos da AWS.

- **AWS CloudFormation** ou ferramentas de orquestração de recursos da AWS de terceiros para gerenciar provisionamento de recursos, atualização e cancelamento do AWS.
- **AWS OpsWorks, Elastic Beanstalk** ou ferramentas de gerenciamento de configuração de servidor de terceiros para gerenciar alterações de configuração de sistema operacional (SO) e aplicativos.

Consulte Gerenciamento de configuração de infraestrutura para obter mais detalhes sobre maneiras de gerenciar hardware como código.[12]

Diversidade/resiliência

Removendo pontos únicos de falha e gargalos

Ao planejar um evento de infraestrutura, você deve analisar suas pilhas de aplicativos para verificar se há algum ponto único de falha (SPOF) ou gargalo de desempenho. Existe alguma instância única de um servidor, volume de dados, banco de dados, gateway NAT ou balanceador de carga que, se falhar, poderá fazer com que um aplicativo inteiro ou de partes significativas de um aplicativo parem de funcionar?

Em segundo lugar, à medida que o aplicativo baseado na nuvem receber mais tráfego ou volume de transações, existe alguma parte da infraestrutura que encontrará um limite físico ou restrição, como largura de banda da rede ou ciclos de processamento de CPU quando o volume de dados aumentar ao longo do caminho de fluxo de dados?

Esses riscos, uma vez identificados, podem ser minimizados de vários modos.

Projeto à prova de falhas

Como mencionado anteriormente, usar acoplamento fraco e filas de mensagens com interfaces RESTful é uma boa estratégia para conseguir resiliência contra falhas de recursos individuais ou flutuações em tráfego ou volume de transações. Outra dimensão do projeto resiliente é configurar os componentes do aplicativo para terem o mínimo de monitoração de estado possível.

Os aplicativos sem estado não exigem conhecimento de transações anteriores e têm pouca dependência de outros componentes do aplicativo. Eles não armazenam informações de sessão. Um aplicativo sem estado pode ser expandido horizontalmente, como um membro de um pool ou cluster, já que

qualquer solicitação pode ser tratada por qualquer instância dentro do pool ou cluster. Você pode simplesmente adicionar mais recursos, conforme necessário, usando o Auto Scaling e critérios de verificação de estado para lidar programaticamente com requisitos flutuantes de computação, capacidade e throughput. Quando um aplicativo é projetado sem monitoração de estado, ele, potencialmente, pode ser refatorado para uma arquitetura sem servidor, usando as funções do Lambda em vez de instâncias EC2. As funções do Lambda também têm capacidade interna de dimensionamento dinâmico.

Na situação em que o recurso de um aplicativo, como um servidor web, não pode deixar de ter dados de estado das transações, você deve considerar a possibilidade de projetar o aplicativo de modo que as partes dele com monitoração de estado fiquem separadas dos servidores em si. Por exemplo, um cookie HTTP, ou dados de estado equivalentes, podem ser armazenados em um banco de dados, como o DynamoDB, ou em um bucket do S3 ou volume do EBS.

Se você tiver um fluxo de trabalho complexo, de várias etapas, em que exista a necessidade de rastrear o estado atual de cada etapa do fluxo de trabalho, o Amazon Simple Workflow Service (SWF) pode ser usado para armazenar de forma centralizada o histórico de execução e tornar essas cargas de trabalho sem estado.

Empregar o processamento distribuído é outra medida para garantir a resiliência. Para casos de uso que exigem o processamento de grandes quantidades de dados em tempo hábil, em que um único recurso de computação não pode atender a necessidade, você pode projetar suas cargas de trabalho para que as tarefas e os dados sejam particionados em fragmentos menores e executados em paralelo em um cluster de recursos de computação. O processamento distribuído é sem estado, uma vez que os nós independentes em que os dados e tarefas particionados estão sendo processados podem falhar. Neste caso, a reinicialização automática de tarefas com falha em outro nó do cluster de processamento distribuído é automaticamente controlada pelo mecanismo de agendamento de processamento distribuído.

A AWS oferece uma variedade de mecanismos de processamento de dados distribuídos, o Amazon EMR, o Amazon Athena e o Amazon Machine Learning, e cada um deles é um serviço gerenciado que oferece endpoints e protege você de qualquer complexidade que envolva a aplicação de correções, manutenção, escalabilidade, failover, etc.

Para o processamento em tempo real de dados de streaming, o Amazon Kinesis Streams pode particionar os dados em vários fragmentos que podem ser processados por vários consumidores desses dados, como as funções do Lambda ou as instâncias EC2.

Para obter mais informações sobre esses tipos de cargas de trabalho, consulte Opções de análise de big data no AWS.[13]

Multizona e multirregião

Os serviços da AWS são hospedados em vários locais em todo o mundo. Esses locais são compostos de regiões e zonas de disponibilidade. Região é uma área geográfica específica. Cada região tem vários locais isolados, que são conhecidos como zonas de disponibilidade. A AWS oferece aos clientes a capacidade de colocar recursos, tais como instâncias e dados, em vários locais.

Você deve projetar seus aplicativos para que eles sejam distribuídos entre várias zonas de disponibilidade e regiões. Em conjunto com a distribuição e replicação de recursos entre zonas de disponibilidade e regiões, você deve projetar seus aplicativos usando balanceamento de carga e mecanismos de failover, de modo que suas pilhas de aplicativos redirecionem automaticamente fluxos de dados e tráfego para esses locais alternativos em caso de uma falha.

Balanceamento de carga

Com o serviço Elastic Load Balancing (ELB, balanceamento elástico de carga), um conjunto de servidores de aplicativos pode ser conectado a um balanceador de carga e ainda ser distribuído entre várias zonas de disponibilidade. Quando as instâncias EC2 em uma zona de disponibilidade específica localizada atrás de um balanceador de carga apresentam falha nas verificações de estado, o balanceador de carga interrompe o envio de tráfego para esses nós. Quando combinado com o Auto Scaling, o número de nós íntegros é automaticamente rebalanceado com as outras zonas de disponibilidade, sem necessidade de intervenção manual.

Também é possível ter o balanceamento de carga em várias regiões usando o Amazon Route 53 e algoritmos de roteamento de DNS baseado em latência. Consulte o Roteamento baseado latência para obter mais informações.[14]

Estratégias de descarte de carga

O conceito de *descarte de carga* em infraestruturas baseadas em nuvem consiste em usar proxy ou redirecionar o tráfego para outro lugar a fim de diminuir a pressão sobre sistemas principais. Em alguns casos, a estratégia de descarte de carga pode ser um exercício de triagem, em que você escolhe eliminar determinados fluxos de tráfego ou reduzir a funcionalidade dos seus aplicativos para reduzir a carga de processamento e poder atender a, pelo menos, um subconjunto das solicitações recebidas.

Há várias técnicas que podem ser usadas para descarte de carga. O roteamento de DNS baseado em latência é um método. Outro método é usar o armazenamento em cache. O cache pode ocorrer perto do aplicativo, usando uma camada de armazenamento em cache na memória, como o Amazon ElastiCache. Como alternativa, você pode usar uma camada de cache que está mais próxima da extremidade do usuário, usando uma rede de distribuição de conteúdo global, como o Amazon CloudFront.

Para obter mais informações sobre o ElastiCache e o CloudFront, consulte o manual de introdução do ElastiCache[15] e do Amazon CloudFront CDN.[16]

Otimização de custo

Comparação entre reservada, spot e sob demanda

A capacidade de controlar os custos de provisionamento dos recursos na nuvem está associada de perto à capacidade de provisionar recursos na nuvem dinamicamente, com base em métricas de sistemas e outros critérios de desempenho e verificação de integridade. Com o Auto Scaling, a utilização de recursos pode corresponder de perto às necessidades de processamento e armazenamento reais, minimizando despesas desnecessárias e recursos subutilizados.

Outra dimensão de controle dos custos na nuvem é a possibilidade de escolher entre instâncias sob demanda, instâncias reservadas (IRs) ou instâncias spot. Há também a possibilidade de capacidade de reserva para o DynamoDB.

Com as instâncias sob demanda, você paga apenas pelas instâncias EC2 que usa. As instâncias sob demanda permitem que você pague pela capacidade de computação por hora, sem compromissos de longo prazo.

As instâncias reservadas do Amazon EC2 oferecem um desconto significativo (até 75%) em comparação com os preços das instâncias sob demanda e fornecem uma reserva de capacidade, quando usadas em uma zona de disponibilidade específica. No entanto, a não ser pela reserva de disponibilidade e do desconto, não há diferença funcional entre as instâncias reservadas e as instâncias sob demanda.

As instâncias spot permitem que você ofereça sugestões de preço para adquirir capacidade de computação excedente do Amazon EC2. As instâncias spot geralmente estão disponíveis com desconto em comparação com os preços de instâncias sob demanda, o que reduz significativamente o custo de execução de seus aplicativos baseados na nuvem.

Quando se projeta para a nuvem, alguns casos de uso são mais adequados para o uso de instâncias spot do que outros. Por exemplo, como as instâncias spot podem ser retiradas a qualquer momento, quando o preço da oferta fica acima da sua proposta, você deve considerar a execução de instâncias spot apenas para pilhas de aplicativos relativamente sem estado e dimensionadas horizontalmente. Para aplicativos com estado ou cargas de processamento caras, pode ser melhor usar instâncias reservadas ou instâncias sob demanda. Para aplicativos de missão crítica em que limitações de capacidade são inadmissíveis, as instâncias reservadas são a melhor escolha.

Consulte Instâncias reservadas[17] e Instâncias spot[18] para obter mais detalhes.

Processo de gerenciamento de eventos

O planejamento de eventos de infraestrutura é uma atividade de grupo, que envolve os desenvolvedores de aplicativos, administradores e partes interessadas da empresa. Semanas antes de um evento de infraestrutura, você deve estabelecer um ritmo de reuniões recorrentes envolvendo os principais técnicos que são responsáveis e operam cada um dos principais componentes de infraestrutura do serviço web.

Programação de evento de infraestrutura

O planejamento um evento de infraestrutura deve começar várias semanas antes da data do evento. Um cronograma típico no ciclo de vida de eventos planejados é mostrado na figura 2.

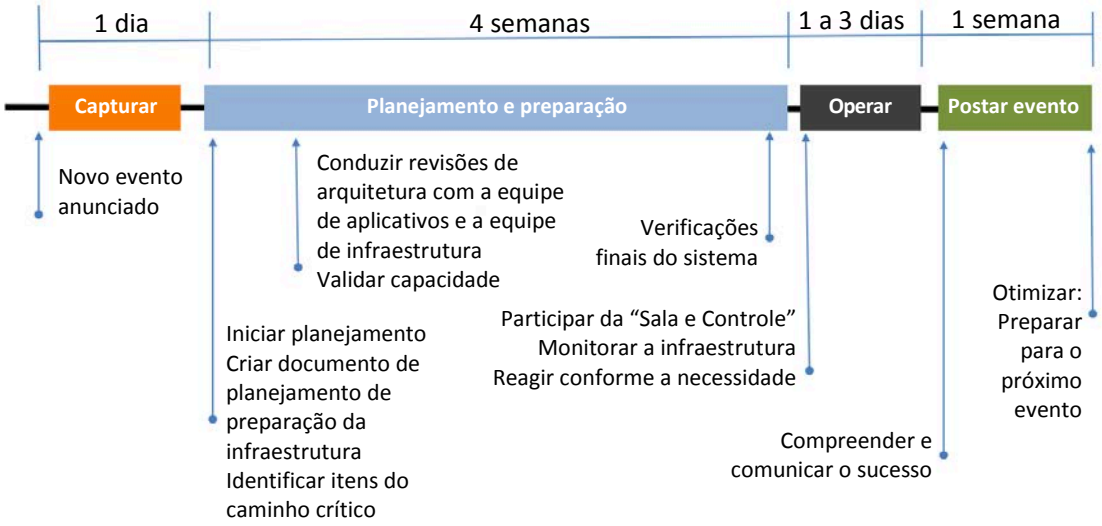

Figura 2. Cronograma típico de eventos de infraestrutura

Planejamento e preparação

Agenda

Recomendamos a seguinte agenda de atividades nas semanas anteriores a um evento de infraestrutura:

Semana 1:

- Nomear uma equipe para conduzir o planejamento e a engenharia para o evento de infraestrutura.
- Conduzir reuniões entre as partes interessadas para entender os parâmetros do evento (escala, duração, tempo, alcance geográfico, cargas de trabalho afetadas) e os critérios de sucesso.
- Envolver todos os parceiros e fornecedores downstream ou upstream.

Semana 2-3:

- Analisar a arquitetura e fazer ajustes de acordo com a necessidade.
- Executar uma análise operacional, fazer ajustes de acordo com a necessidade.
- Seguir as práticas recomendadas descritas neste documento e nas referências de pé de página.

- Identificar riscos e desenvolver planos de atenuação.
- Desenvolver um runbook do evento planejado.

Semana 4:

- Analisar todos os serviços de fornecedor na nuvem que exigem redimensionamento com base na carga esperada.
- Verificar limites de serviço e aumentar os limites conforme necessário.
- Configurar painel de monitoramento e alertas em limites predefinidos.

Análise da arquitetura

Uma parte essencial de sua preparação para um evento de infraestrutura é uma análise da arquitetura da pilha de aplicativos que sofrerá o aumento de tráfego. A finalidade da análise é verificar e identificar possíveis áreas de risco para a escalabilidade ou a confiabilidade do aplicativo e identificar as oportunidades para otimização antes do evento.

A AWS fornece aos seus clientes do Enterprise Support uma estrutura para analisar pilhas de aplicativos do cliente, baseada em cinco pilares de projeto. São eles: segurança, confiabilidade, eficiência de desempenho, otimização de custos e excelência operacional, como descrito abaixo.

Tabela 1: Pilares dos aplicativos bem projetados

Nome do pilar	Definição do pilar	Área de Interesse relevante
Segurança	A capacidade de proteger informações, sistemas e ativos e, ao mesmo tempo, oferecer valor para os negócios por meio de avaliações de risco e estratégias de atenuação.	Gerenciamento de identidade, criptografia, monitoramento, registro, gerenciamento de chaves, instâncias dedicadas, conformidade, governança
Confiabilidade	A capacidade de um sistema recuperar falhas de serviço de infraestrutura ou adquirir dinamicamente recursos de computação para atender à demanda e atenuar interrupções, como configurações incorretas ou problemas de rede transiente.	Limites de serviço, várias zonas e regiões de disponibilidade, escalabilidade, verificação/monitoramento de integridade, backup e recuperação de desastres, rede, automação de autocorreção
Eficiência de desempenho	A capacidade de usar recursos de computação com eficiência para atender aos requisitos do sistema e manter essa eficiência quando a demanda mudar e as tecnologias evoluírem.	Serviços certos da AWS, utilização de recursos, arquitetura de armazenamento, armazenamento em cache, requisitos de latência

Nome do pilar	Definição do pilar	Área de Interesse relevante
Otimização de custo	A capacidade de evitar ou eliminar custos desnecessários ou recursos abaixo do ideal.	Instâncias spot/reservadas, ajuste de ambiente, seleção de serviço, ajuste de volume, gerenciamento de conta, faturamento consolidado, recursos de descontratação
Excelência operacional	A capacidade de executar e monitorar sistemas para fornecer valor de negócios e melhorar continuamente os processos e procedimentos de suporte.	Runbooks, guias estratégicos, CI/CD, dias de jogo, infraestrutura como código, RCAs

Uma lista de verificação detalhada de itens de análise de arquitetura, que pode ser usada para analisar uma pilha de aplicativos baseada no AWS, está disponível no apêndice deste artigo técnico.

Revisão operacional

Além de uma análise de arquitetura, que é mais concentrada nos componentes de projeto do aplicativo, você deve analisar as suas operações em nuvem e as suas práticas de gerenciamento para avaliar como está lidando com o gerenciamento de suas cargas de trabalho na nuvem. O objetivo da análise é identificar lacunas e problemas operacionais e tomar medidas para minimizá-los antes do evento.

A AWS oferece uma análise de operações em nuvem para seus clientes de suporte empresarial, que pode ser uma ferramenta valiosa para se preparar para um evento de infraestrutura. A análise se concentra em avaliar as seguintes áreas:

- Preparação – Você deve ter a combinação correta de estrutura organizacional, processos e tecnologia. Você deve ter funções e responsabilidades claramente definidas para a equipe que gerencia a sua pilha de aplicativos. Os processos devem ser definidos com antecedência para se alinharem com o evento. Os procedimentos devem ser automatizados sempre que possível.
- Monitoramento – O monitoramento eficaz mede o desempenho dos aplicativos. O monitoramento é essencial para detectar anomalias antes que se tornem problemas e fornece oportunidades para minimizar o impacto de eventos adversos.

- Operações – Atividades operacionais precisam ser executadas em tempo hábil e de maneira confiável, aproveitando a automação sempre que possível e, ao mesmo tempo, lidar com eventos operacionais inesperados que exigem escalações.
- Otimização – Realizar uma análise posterior usando as métricas coletadas, tendências operacionais e lições aprendidas para capturar e relatar oportunidades de aperfeiçoamento durante eventos futuros. Otimização mais preparação criam um loop de feedback para resolver problemas operacionais e impedir que eles se repitam.

Compreender os limites de serviços da AWS

Durante um evento de infraestrutura planejado, é crucial evitar que sejam excedidos os limites de serviço que possam ser impostos por um provedor de nuvem ao aumentar ou reduzir um aplicativo ou carga de trabalho.

Os provedores de serviços de nuvem normalmente têm limites para os diferentes recursos que você pode usar. Esses limites normalmente são impostos por conta e por região. Os recursos afetados incluem instâncias, volumes, streams, invocações sem servidor, snapshots, número de VPCs, regras de segurança e assim por diante. Eles pretendem atuar como medida de segurança contra código desviado ou agentes inescrupulosos que tentam abusar dos recursos e como um controle para ajudar a minimizar o risco de cobrança.

Alguns limites de serviço são elevados automaticamente com o tempo à medida que você expande sua presença na nuvem, embora a maioria desses serviços exija que você solicite aumentos de limite abrindo um caso de suporte. Embora alguns limites de serviço possam ser aumentados por meio de casos de suporte, outros serviços têm limites que não podem ser alterados.

A AWS fornece aos clientes de suporte Enterprise e Business, o Trusted Advisor, que fornece um painel de verificação de limite para que os clientes possam gerenciar proativamente todos os limites de serviço.

Para obter mais informações sobre os limites dos vários serviços da AWS e como verificá-los, consulte Limites dos serviços da AWS[19] e Trusted Advisor.[20]

Compreender padrões

Linhas de base

Você deve documentar os valores de "recuperação de integridade" para as principais métricas antes do início de um evento de infraestrutura. Isso ajuda a determinar quando um aplicativo/serviço retorna com segurança para os níveis normais após a conclusão/final do evento. Por exemplo, identificar que a taxa de transações normal através de um balanceador de carga é de 2.500 solicitações por segundo ajudará a determinar quando é seguro começar a relaxar os procedimentos após o evento.

Fluxos de dados e dependências

Compreender como os dados fluem através dos componentes de um aplicativo ajuda a identificar possíveis gargalos e dependências. Os níveis dos aplicativos ou componentes que são os consumidores dos dados em um fluxo de dados estão dimensionados e configurados corretamente para fazer o autodimensionamento adequado se os níveis ou os componentes de uma pilha de aplicativos produtores de dados aumentarem? Em caso de falha de um componente, os dados podem ser colocados em fila até que o componente se recupere? Há algum provedor ou consumidor de dados downstream ou upstream que possa ser redimensionado em resposta ao seu evento?

Proporcionalidade

Outra consideração para analisar na preparação para um evento de infraestrutura é a proporcionalidade do redimensionamento exigidos pelos vários componentes de uma pilha de aplicativos. Essa proporcionalidade é sempre de um para um. Por exemplo, um aumento de dez vezes nas transações por segundo através de um balanceador de carga pode exigir vinte vezes mais capacidade de armazenamento ou número de fragmentos de streaming ou número de operações de leitura e gravação no banco de dados, devido ao processamento que possa estar ocorrendo no aplicativo de frente.

Plano de comunicação

Antes do evento, você deve desenvolver um plano de comunicação. Obtenha uma lista das partes interessadas e grupos de suporte e identifique quem deve ser contatado em vários estágios de um evento, em vários cenários, como no início do evento, durante o evento, no final do evento, na análise pós-evento, contatos de emergência, contatos durante situações de solução de problemas, etc.

Pessoas e grupos para ser contatados podem incluir os seguintes:

- Partes interessadas
- Gerentes de operações
- Desenvolvedores
- Equipes de suporte
- Equipes do provedor de serviços em nuvem
- Equipe do centro de operações de rede (NOC)

À medida que prepara uma lista de contatos internos, você também deve desenvolver uma lista de contatos com as partes interessadas externas envolvidas na entrega contínua do aplicativo. Essas partes interessadas incluem os parceiros e fornecedores que suportam os principais componentes da pilha, fornecedores downstream e upstream que fornecem serviços externos, feeds de dados, serviços de autenticação e assim por diante.

Esta lista de contatos externos também deve incluir o seguinte:

- Fornecedores de hospedagem de infraestrutura
- Fornecedores de telecomunicações
- Parceiros de streaming de dados em tempo real
- Contatos de marketing de RP
- Parceiros de publicidade
- Consultores técnicos envolvidos com a engenharia do serviço

Peça as seguintes informações de cada fornecedor:

- Pontos de contato em tempo real durante todo o evento
- Contato de suporte crítico e processo de escalação
- Nome, número de telefone e endereço de e-mail
- Confirmação de que os contatos técnicos estarão disponíveis em tempo real

Os clientes da AWS assinantes do Enterprise Support também têm gerentes técnicos de conta (TAMs, Technical Account Managers) designados para a sua conta que podem coordenar e garantir que a equipe de suporte dedicada da AWS esteja ciente e preparada para fornecer suporte ao evento. Os TAMs também de serviço durante o evento, presentes na sala de comando e disponíveis para conduzir escalações de suporte, se necessário.

Preparação do centro de operações de rede (NOC)

Antes do evento, você deve instruir a equipe de desenvolvimento e/ou operações de desenvolvedor de operações que crie um painel de métricas em tempo real que monitore cada componente crítico do serviço web em produção durante o evento. Idealmente, o painel de controle deve apresentar as métricas atualizadas automaticamente a cada minuto ou em qualquer intervalo adequado e eficaz durante o evento.

Considere monitorar os seguintes componentes:

- Utilização de recursos de cada servidor (CPU, disco e memória)
- Tempo de resposta do serviço web
- Métricas de tráfego da web (usuários, visualizações de página, sessões)
- Tráfego da web por região de visitante (segmentos de clientes globais)
- Utilização de servidor de banco de dados
- Funis de conversão de fluxo de marketing, como taxas de conversão e porcentagem de perda
- Logs de erro de aplicativos
- Monitoramento de canários

O Amazon CloudWatch oferece um meio de coletar a maioria dessas métricas dos recursos da AWS em uma única janela usando os painéis personalizados do CloudWatch. Além disso, o CloudWatch oferece a capacidade de importar métricas personalizadas no CloudWatch sempre que o AWS não estiver fornecendo uma métrica automaticamente. Consulte a seção Monitoramento deste artigo para obter mais detalhes sobre os recursos e ferramentas de monitoramento do AWS.

Preparação do runbook

Você deve desenvolver um runbook em preparação para o evento de infraestrutura. *Runbook* é um manual operacional contendo uma compilação de procedimentos e operações que seus operadores executarão durante o evento. Runbooks de eventos podem ser desenvolvidos a partir dos runbooks existentes usados para operações de rotina e tratamento de exceções. Normalmente, um runbook contém procedimentos para iniciar, parar, supervisionar e depurar um sistema. Ele também deve descrever os procedimentos para lidar com eventos e contingências inesperados.

Um runbook deve incluir as seguintes seções:

- **Detalhes do evento**: Faz uma breve descrição do evento, critérios de sucesso, cobertura de mídia, de evento e detalhes de contato com das principais partes interessadas, no cliente e na AWS.
- **Lista de serviços da AWS**: Enumera todos os serviços da AWS para serem usados durante o evento. Além disso, a carga esperada nesses serviços, as regiões afetadas e as IDs de conta.
- **Revisão de arquitetura e aplicativos**: Documenta os resultados de testes de carga, os pontos de estresse no projeto do aplicativo e da infraestrutura, resiliência medidas para a carga de trabalho, os pontos únicos de falha e possíveis gargalos.
- **Revisão operacional**: Destaca a configuração de monitoramento, critérios de integridade, mecanismos de notificação e os procedimentos de restauração de serviço.
- **Lista de verificação da preparação**: Inclui considerações como verificações de limites de serviço pré-aquecimento dos componentes da pilha de aplicativos, como balanceadores de carga, pré-provisionamento de recursos, como fragmentos de stream, partições do DynamoDB, partições do S3 e assim por diante. Para obter mais informações, consulte a lista de verificação detalhada da análise arquitetural no apêndice deste artigo técnico.

Monitorar

Plano de monitoramento

O monitoramento de banco de dados, aplicativos e sistema operacional é crucial para garantir um evento bem-sucedido. Sistemas de monitoramento

abrangentes devem ser configurados para que você possa detectar eficazmente e responder imediatamente a sérios incidentes durante o evento de infraestrutura. Em um nível alto, uma estratégia de monitoramento eficaz garante que ferramentas de monitoramento sejam instrumentadas no nível adequado para um aplicativo, de acordo com sua importância para os negócios. Uma estratégia de gerenciamento de incidentes eficaz incorpora os dados de monitoramento do cliente e da AWS às suas ferramentas e processos de gerenciamento de eventos e incidentes. A implementação de um plano de monitoramento que coletivamente reúna dados de monitoramento de todos os segmentos da sua solução da AWS ajuda muito na depuração de uma falha complexa, se tal ocorrer.

O plano de monitoramento deve considerar as seguintes questões:

- Que ferramentas de monitoramento e painéis devem ser configuradas para o evento?
- Quais são os objetivos de monitoramento e os limites permitidos? Que eventos disparam ações?
- Quais os recursos e métricas desses recursos a serem monitorados e com que frequência eles serão consultados?
- Quem vai realizar as tarefas de monitoramento? Quais são os alertas de monitoramento implantados? Quem será alertado?
- Que planos de correção foram configurados para falhas comuns e esperadas? E os eventos inesperados?
- O que é o processo de escalação em caso de falha?

As seguintes ferramentas de monitoramento da AWS podem ser usadas como parte dessa estratégia:

- **Amazon CloudWatch**: Uma solução pronta para as métricas de painel, monitoramento, alertas e provisionamento automatizado do AWS.
- **Métricas personalizadas do Amazon CloudWatch**: Usado para coleta de métricas de sistema operacional e aplicativo de negócios. A API do Amazon CloudWatch permite a coleta de praticamente qualquer tipo de métrica personalizada.

- **Saúde das instâncias do Amazon EC2**: Usado para visualização de verificações de status e para programar eventos para suas instâncias com base em seu status, como a reinicialização automática ou o reinício de uma instância.
- **Amazon SNS**: Usado para a configuração, operação e envio de notificações orientadas por eventos.
- **AWS X-Ray**: Ajuda na depuração e na análise de aplicativos distribuídos e arquitetura de microsserviços, analisando os dados que fluem através dos componentes do sistema.
- **Amazon Elasticsearch Service**: Usado para coleta de log centralizado e análise de log em tempo real. Para a rápida, detecção heurística de problemas.
- **Ferramentas de terceiros:** Usadas para análise em tempo real e monitoramento e visibilidade totais da pilha.
- **Ferramentas de monitoramento de sistema operacional padrão**: Usadas para monitoramento no nível do sistema operacional.

Para obter mais detalhes sobre as ferramentas de monitoramento AWS, consulte Monitoramento automatizado e manual.[21] Consulte também Usando os painéis do Amazon CloudWatch[22] e Publicando métricas personalizadas.[23]

Notificações

Um elemento operacional crucial de seu projeto para eventos de infraestrutura é a configuração de alertas e notificações para integração com suas soluções de monitoramento. Esses alertas e notificações podem ser usados com serviços como o AWS Lambda para disparar ações com base no alerta. Automatizar respostas para eventos operacionais é um elemento importante para alcançar a atenuação, reversão e recuperação com o máximo de capacidade de resposta.

As ferramentas também devem ser implementadas para monitorar centralmente cargas de trabalho e criar alertas e notificações apropriados com base em logs e métricas disponíveis relacionadas aos principais indicadores operacionais. Isso inclui alertas e notificações para anomalias fora dos limites aceitáveis, bem como falhas de serviço ou de componentes. O ideal é que, quando os limites de baixo desempenho forem ultrapassados ou ocorrerem falhas, o sistema tenha sido projetado para se autorreparar, ou aumentar ou reduzir automaticamente em resposta a essas notificações e alertas.

Como observado anteriormente, a AWS oferece serviços (Amazon SQS e Amazon SNS) para garantir alertas e notificações apropriadas em resposta a eventos operacionais não planejados, bem como para permitir respostas automatizadas.

Preparação operacional (dia de evento)

Execução do plano

No dia do evento, a equipe principal envolvida com o evento de infraestrutura deve estar em uma teleconferência monitorando os painéis ao vivo. Os runbooks devem estar totalmente desenvolvidos e disponíveis. Certifique-se de que o plano de comunicação esteja bem definido e seja conhecido por toda a equipe de suporte e partes interessadas e que um plano de contingência esteja implantado.

Sala de comando

Durante o evento, tenha uma ponte de conferência ao vivo aberta com os seguintes participantes:

- A principal equipe responsável por aplicativos e operações
- A equipe de liderança de operações
- Recursos técnicos de parceiros externos diretamente envolvidos com entrega técnica
- Partes interessadas de negócios

Durante a maior parte do evento, a conversa dessa ponte de conferência deve ser mínima. Se um evento operacional adverso surgir, as pessoas-chave que podem responder ao evento já estarão nesta ponte, prontas para ações e consultas.

Relatórios da liderança

Durante o evento, envie um e-mail por hora para as principais partes interessadas de liderança. Esta atualização deve incluir o seguinte:

- Resumo do status: Verde (como previsto), Amarelo (problemas encontrados), Vermelho (problema sério)
- Atualização das principais métricas

- Problemas encontrados, status do plano de remediação, hora estimada de conclusão da remediação
- O número de telefone da ponte de conferência da sala de controle (no caso de alguém querer participar)

Na conclusão do evento, um e-mail de resumo final deve ser enviado, em um formato semelhante.

Plano de contingência

Cada etapa do processo de preparação para o evento deve ter um plano de reversão correspondente que tenha sido verificado em um ambiente de teste.

Considere as seguintes perguntas ao preparar um plano de reversão:

- Quais são os piores cenários que podem ocorrer durante o evento?
- Que tipo de evento teria um impacto negativo para as relações públicas?
- Que componentes e serviços de terceiros podem falhar durante o evento?
- Que métricas devem ser monitoradas que indicariam que um cenário ruim está ocorrendo?
- O que é o plano de reversão para cada cenário possível?
- Quanto tempo cada processo de reversão levará? Qual é o objetivo de ponto de recuperação (RPO) e o objetivo de tempo de recuperação (RTO) aceitável? (Consulte Usando o AWS para recuperação de desastres[24] para ler mais sobre esses conceitos.)

Considere os seguintes tipos de reversão:

- **Implantação azul/verde**: Se estiver implantando um novo aplicativo ambiente de produção, mantenha o ambiente de produção online e disponível para voltar para ele rapidamente.
- **Piloto aquecido:** Inicie um ambiente mínimo em uma segunda região que possa ser redimensionada rapidamente, se necessário. Em caso de falha da região principal, rapidamente redimensione a região de backup e transfira o tráfego para ela.

- **Páginas de erro do modo de manutenção**: Verifique os recursos e os gatilhos de página de erro em cada camada de seu serviço web. Esteja preparado para inserir uma mensagem de erro mais específica nessas páginas de erro, de acordo com a necessidade.

Teste e documente cada plano de reversão para cada possível cenário de falha.

Atividades pós-evento

Análise pós-evento

A análise pós-evento muitas vezes é esquecida porque os clientes estão normalmente ansiosos por retornar às operações normais. No entanto, recomendamos que você exija uma análise pós-evento como parte do ciclo de vida de gerenciamento de qualquer evento de infraestrutura. As análises pós-evento permitem que você colabore com cada equipe envolvida e identifique as áreas que podem exigir mais otimização, como procedimentos operacionais, detalhes de implementação, failover e procedimentos de recuperação, etc. Isso é especialmente relevante se uma pilha de aplicativos sofreu interrupções durante o evento. A análise pós-evento também ajuda a fornecer documentação se houver necessidade de desenvolver documentos de análise de causa-raiz (RCA).

Processo de desaceleração

Imediatamente após a conclusão do evento de infraestrutura o processo de desaceleração deve começar. Durante esse período, é aconselhável continuar o monitoramento dos aplicativos e serviços relevantes para garantir que o tráfego voltou aos para níveis normais de produção. Use qualquer painel de integridade criado durante a fase de preparação para verificar a normalização do tráfego e das taxas de transações. Os períodos de desaceleração de alguns eventos podem ser lineares e diretos, enquanto outros podem experimentar reduções irregulares ou mais graduais de volume. Alguns padrões de tráfego podem persistir. Por exemplo, a recuperação de um pico de tráfego geralmente requer procedimentos simples de desaceleração, enquanto algo como a implementação de um aplicativo ou a expansão para uma nova região geográfica podem ter efeitos de longa duração que exigem você monitore cuidadosamente novos padrões de tráfego e torne o monitoramento adicional parte da pilha de aplicativos permanente.

Em algum momento após a conclusão do evento, você deve determinar se é seguro interromper as operações de gerenciamento do evento. Consulte os

valores "normais" documentados anteriormente para as principais métricas para ajudar a determinar quando declarar que um evento está concluído ou encerrado. Recomendamos dividir as atividades de desaceleração em duas ramificações, que podem ter diferentes cronogramas. Concentre a primeira ramificação no gerenciamento operacional do evento, como envio de comunicações às partes interessadas parceiros internos e externos e redefinição dos limites de serviço. Concentre a segunda ramificação nos aspectos técnicos da desaceleração, como procedimentos de redimensionamento, validação da integridade do ambiente e critérios de arquitetura, para determinar se as alterações arquiteturais devem ser revertidas ou confirmadas.

O cronograma associado a cada uma dessas ramificações pode variar de acordo com a natureza do evento, as principais métricas e conforto do cliente. Descrevemos algumas tarefas comuns associadas a cada ramificação na tabela a seguir para ajudá-lo a determinar o gerenciamento adequado da hora de encerramento para um evento.

Tabela 2: Tarefas operacionais de desaceleração

Tarefa	Descrição
Comunicação	Notificação às partes interessadas internas e externas de que o evento terminou. A comunicação da hora de encerramento deve ser alinhada com a definição da conclusão do evento. Use métricas de “volta à integridade” para determinar quando é apropriado encerrar a comunicação. Como alternativa, você pode encerrar a comunicação em camadas. Por exemplo, você pode encerrar a ponte da sala de comando, mas deixar os procedimentos de escalação de eventos intactos no caso de falhas pós-evento.
Limites de serviço/contenção de custos	Embora possa ser tentador manter um limite de serviço elevado após um evento, lembre-se de que os limites de serviço também são usados como rede de segurança. Os limites de serviço protegem você e seus custos, evitando o excesso de utilização do serviço, seja isso uma conta comprometida ou automação mal configurada.
Relatórios e análise	Coleta de dados e comparação com as métricas do evento, acompanhadas por narrativas analíticas mostrando padrões, tendências, áreas problemáticas, procedimentos bem-sucedidos, procedimentos pontuais, cronograma do evento e se os critérios de sucesso foram ou não atendidos devem ser desenvolvidas e distribuídas para todas as partes internas identificadas no plano de comunicações. Uma análise de custos detalhada também deve ser desenvolvida para mostrar as despesas operacionais do suporte ao evento.
Tarefas de otimização	As organizações corporativas evoluem ao longo do tempo à medida que continuam a melhorar suas operações. A otimização operacional requer a constante coleta de métricas, tendências operacionais e lições aprendidas de eventos para descobrir oportunidades de aprimoramento. A otimização se vincula de volta com a

Tarefa	Descrição
	preparação para formar um loop de feedback para resolver problemas operacionais e impedir que eles se repitam.

Tabela 3: As tarefas técnicas da desaceleração

Tarefa	Descrição
Limites de serviço/contenção de custos	Embora possa ser tentador manter limites de serviço elevados após um evento, lembre-se de que os limites de serviço também são usados como rede de segurança. Os limites de serviço protegem você e seus custos, evitando o excesso de utilização do serviço, seja por atividade mal intencionada resultante de uma conta comprometida ou através de automação mal configurada.
Procedimentos de redução	Reverter recursos que foram aumentados durante a fase de preparação. Esses itens são exclusivos de sua arquitetura, mas os exemplos a seguir são comuns: Tamanho da instância EC2/RDS Configuração de Auto Scaling Capacidade reservada IOPS provisionado
Validação de integridade do ambiente	Compare com as métricas da linha de base e analise a integridade de produção para verificar se, após o evento e depois dos procedimentos de redução, os sistemas afetados estão informando um comportamento normal.
Disposição de alterações arquiteturais	Pode valer a pena manter algumas alterações feitas na preparação para o evento, dependendo da natureza do evento e da observação de métricas operacionais. Por exemplo, a expansão para uma nova região geográfica permanente pode exigir um aumento permanente de recursos nessa região ou a elevação de certos limites de serviço ou parâmetros de configuração, como o número de partições em um banco de dados ou fragmentos em um stream de PIOPS em um volume, pode ser uma medida de ajuste de desempenho que deve ser mantida.

Otimizar

Talvez o componente mais importante do gerenciamento de eventos de infraestrutura seja a análise pós-evento e a identificação de desafios operacionais e arquiteturais observados e as oportunidades de aprimoramento. Eventos de infraestrutura de são eventos únicos. Eles podem ser sazonais ou coincidir com novas versões de um aplicativo ou podem ser parte do crescimento da empresa, medida que ela se expande para novos mercados e territórios. Assim, cada evento de infraestrutura é uma oportunidade para observar, melhorar e preparar de forma mais eficaz para o próximo.

Conclusão

A AWS fornece as peças, na forma de produtos e serviços elásticos e programáveis, que sua empresa pode montar para suportar praticamente qualquer escala de carga de trabalho. Com as diretrizes e as práticas recomendadas da AWS para eventos de infraestrutura, junto com o nosso conjunto completo de serviços altamente disponíveis, a sua empresa pode desenhar e preparar grandes eventos de negócios e garantir que as demandas de escalabilidade possam ser atendidas perfeita e dinamicamente, assegurando resposta rápida e alcance global.

Colaboradores

As seguintes organizações e pessoas contribuíram neste documento:

- Presley Acuna, gerente de Enterprise Support da AWS
- Kurt Gray, arquiteto de soluções globais da AWS
- Michael Bozek, gerente técnico de conta sênior da AWS
- Rovan Omar, gerente técnico de conta da AWS
- Will Badr, gerente técnico de conta sênior da AWS
- Eric Blankenship, gerente técnico de conta sênior da AWS
- Greg Bur, gerente técnico de conta da AWS
- Lista Hesse, gerente técnico de conta sênior da AWS
- Hasan Khan, gerente técnico de conta sênior da AWS
- Varun Bakshi, gerente técnico de conta sênior da AWS

Outras leituras

Para leituras adicionais sobre práticas recomendadas de arquitetura e operação, consulte Listas de verificação operacionais para AWS.[25] Recomendamos a leitura de Estrutura bem arquitetada de AWS[26] para uma abordagem estruturada da avaliação das pilhas de entrega de aplicativos baseada em nuvem. A AWS oferece Infrastructure Event Management (IEM) como uma oferta de suporte premium para os clientes que desejam um envolvimento mais direto do gerente técnico de conta e dos engenheiros de suporte da AWS no

projeto, planejamento e nas operações no dia do evento. Para obter mais detalhes sobre a oferta de suporte premium IEM da AWS, consulte Infrastructure Event Management.[27]

Apêndice

Lista de verificação de análise arquitetural detalhada

Sim-Não-N/D	Segurança
☐-☐-☐	Fazemos um rodízio de nossas chaves de acesso e senhas de usuário e credenciais do AWS Identity and Access Management (IAM) para os recursos envolvidos no nosso aplicativo no máximo a cada três meses, de acordo com as práticas recomendadas de segurança da AWS. Aplicamos política de senha em todas as contas e usamos o hardware ou dispositivos de autenticação multifatorial (MFA) virtuais.
☐-☐-☐	Temos processos e controles internos de segurança para controlar acesso de privilégio mínimo exclusivo e baseado em função às APIs do AWS utilizando o IAM.
☐-☐-☐	Removemos qualquer informação confidencial ou sensível, incluindo pares de chaves de instâncias públicas/privadas embutidas e revisamos todos os arquivos de chave de SSH autorizados de todas as Amazon Machine Images (AMIs) personalizadas.
☐-☐-☐	Usamos funções do IAM para instâncias EC2 como conveniente, em vez de incorporar credenciais dentro das AMIs.
☐-☐-☐	Segregamos os privilégios administrativos do IAM dos privilégios de usuários regulares, criando uma função administrativa no IAM e restringindo as ações de IAM de outras funções.
☐-☐-☐	Aplicamos as correções de segurança mais recentes em nossa instâncias EC2 para Windows e Linux. Usamos controles de acesso de sistema operacional, incluindo regras de grupo de segurança do Amazon EC2, listas de controle de acesso à rede VPC, reforço do sistema operacional, proteção do firewall baseado no host, detecção/prevenção de invasões, monitoramento de configuração de software e inventário de host.
☐-☐-☐	Garantimos que a conectividade de rede para dentro e para fora do AWS e dos ambientes corporativos da organização use um transporte de protocolos de criptografia.
☐-☐-☐	Aplicamos uma solução de gerenciamento centralizado de log e auditoria para identificar e analisar padrões de acesso incomuns ou qualquer ataque inescrupuloso ao ambiente.
☐-☐-☐	Temos gerenciamento de eventos e incidentes de segurança, correlação e processos de relatório implementados.
☐-☐-☐	Garantimos que não haja acesso irrestrito a recursos do AWS em nenhum de nossos grupos de segurança.
☐-☐-☐	Usamos um protocolo seguro (HTTPS ou SSL), políticas de segurança atualizadas

Sim-Não-N/D	Segurança
	protocolo de cifras para conexão de front-end (cliente para balanceador de carga). As solicitações são criptografadas entre os clientes e o balanceador de carga, o que é mais seguro.
☐-☐-☐	Configuramos nosso conjunto de registros de recursos Amazon Route 53 MX para ter um conjunto de registros de recursos TXT que contém um valor Sender Policy Framework (SPF) correspondente para especificar os servidores autorizados a enviar e-mail pelo nosso domínio.

Sim-Não-N/D	Confiabilidade
☐-☐-☐	Implantamos nosso aplicativo em um grupo de instâncias EC2 que são implantadas em um grupo de Auto Scaling para garantir escalabilidade horizontal automática com base em um plano de escalabilidade predefinido. Saiba mais.
☐-☐-☐	Usamos uma verificação de integridade do Elastic Load Balancing em nossa configuração de grupo de Auto Scaling para garantir que o grupo de Auto Scaling atue na integridade as instâncias EC2 subjacentes. (Aplicável somente se você usar balanceadores de carga em grupos de Auto Scaling.)
☐-☐-☐	Implantamos componentes críticos de nossos aplicativos em várias zonas de disponibilidade, estão replicando adequadamente dados entre regiões. Testamos como falha dentro desses componentes afeta a disponibilidade do aplicativo usando o Elastic Load Balancing, o Amazon Route 53, ou qualquer ferramenta de terceiros apropriada.
☐-☐-☐	Na camada de banco de dados, implantamos nossas instâncias do Amazon RDS em várias zonas de disponibilidade para aprimorar a disponibilidade dos bancos de dados fazendo a replicação síncrona para uma instância em espera em uma zona de disponibilidade diferente.
☐-☐-☐	Temos processos definidos para failover automático ou manual em caso de qualquer interrupção ou degradação do desempenho.
☐-☐-☐	Usamos registros CNAME para mapear nosso nome DNS para nossos serviços. NÃO usamos registros A.
☐-☐-☐	Configuramos um valor de time-to-live (TTL) menor para o nosso conjunto de registros do Amazon Route 53. Isso evita atrasos quando os resolvedores de DNS solicitam registros de DNS atualizados ao rotear o tráfego. (Por exemplo, isso pode ocorrer quando o failover de DNS detecta e responde a uma falha de um dos endpoints.)
☐-☐-☐	Temos pelo menos dois túneis VPN configurados para fornecer redundância no caso de queda de serviço ou manutenção planejada dos dispositivos no endpoint do AWS.
☐-☐-☐	Usamos o AWS Direct Connect e temos duas conexões Direct Connect configurados em todos os momentos para fornecer redundância no caso de um dispositivo não estar disponível. As conexões são provisionadas em diferentes locais do Direct Connect para fornecer redundância no caso de um local estar indisponível. Também configuramos a conectividade com o nosso gateway privado virtual para termos várias interfaces virtuais configuradas em várias conexões e locais do Direct Connect.

Sim-Não-N/D	Confiabilidade
☐-☐-☐	Usamos instâncias do Windows e nos certificamos de que estamos usando os mais recentes drivers de PV. O driver de PV ajuda a otimizar o desempenho de driver e minimizar os riscos de segurança e problemas de tempo de execução. Também nos certificamos de que agente do EC2Config esteja executando a versão mais recente em nossa instância do Windows.
☐-☐-☐	Fazemos snapshots de nossos volumes do Amazon Elastic Block Store (EBS), a fim de garantir uma recuperação para um ponto no tempo em caso de falha.
☐-☐-☐	Usamos volumes do Amazon EBS separados para o sistema operacional e o aplicativo/banco de dados, onde apropriado.
☐-☐-☐	Aplicamos as mais recentes correções de kernel, software e drivers em todas as instâncias do Linux.

Sim-Não-N/D	Eficiência de desempenho
☐-☐-☐	Testamos totalmente nossos componentes de aplicativos hospedados da AWS, incluindo testes de desempenho, antes de colocá-los no ar. Além disso, realizamos testes de carga para garantir que usamos o tamanho de instância EC2, o número de IOPS, tamanho da instância de banco de dados do RDS, etc, corretos.
☐-☐-☐	Executamos um relatório de verificação de uso em relação aos limites de serviços e nos certificamos de que o uso atual dos serviços AWS é igual ou menor que 80% dos limites de serviço. Saiba mais
☐-☐-☐	Usamos a rede de entrega/distribuição de conteúdo (CDN) para utilizar o armazenamento em cache para nosso aplicativo (Amazon CloudFront) e como uma maneira de otimizar a entrega de conteúdo e a distribuição automática do conteúdo para o ponto de presença mais próximo do usuário.
☐-☐-☐	Sabemos que alguns cabeçalhos de solicitações de HTTP dinâmico que o Amazon CloudFront recebe (usuário-agente, dados, etc.) podem afetar o desempenho, reduzindo a taxa de acertos do cache e aumentando a carga na origem. Saiba mais
☐-☐-☐	Garantimos que a taxa de transferência máxima de uma instância EC2 é maior do que o throughput máximo agregado dos volumes do EBS. Também usamos instâncias otimizadas para EBS com volumes do EBS PIOPS para obter o desempenho esperado dos volumes.
☐-☐-☐	Garantimos que o projeto não tenha nenhum gargalo na infraestrutura ou um ponto de estresse no banco de dados ou no projeto do aplicativo.
☐-☐-☐	Implementamos o monitoramento de recursos de aplicativos e configuramos alertas com base em qualquer quebra de desempenho usando o Amazon CloudWatch ou ferramentas de terceiros parceiros.
☐-☐-☐	O nosso projeto evita usar um grande número de regras em qualquer grupo de segurança anexado a nossas instâncias de aplicativo. Um grande número de regras em um grupo de segurança pode degradar o desempenho.

Sim-Não-N/D	Otimização de custo
☐-☐-☐	Entendemos que o evento de infraestrutura pode envolver algum excesso de capacidade provisionada que precisa ser resolvido após o evento para evitar custos desnecessários.
☐-☐-☐	Usamos o dimensionamento certo para todos os nossos componentes de infraestrutura, incluindo tamanho de instância EC2, tamanho de instância de banco de dados do RDS, tamanho e número de nós de cluster de cache, tamanho e número de nós de cluster do Redshift e tamanho do volume de EBS.
☐-☐-☐	Usamos instâncias spot quando é conveniente. As instâncias spot são ideais para cargas de trabalho que têm hora de início e de fim flexíveis. Os casos de uso comuns para instâncias spot são: Processamento de lotes, geração de relatórios e cargas de trabalho de computação de alto desempenho.
☐-☐-☐	Temos requisitos mínimos de capacidade de aplicativo previsíveis e aproveitamos a vantagem das instâncias reservadas. . As instâncias reservadas permitem que você reserve capacidade de computação do Amazon EC2 em troca de um desconto substancial em comparação com os preços por hora de instâncias sob demanda.

Notes

[1] https://aws.amazon.com/answers/account-management/aws-tagging-strategies/

[2] https://aws.amazon.com/blogs/aws/resource-groups-and-tagging/

[3] https://aws.amazon.com/sqs/

[4] http://docs.aws.amazon.com/general/latest/gr/rande.html

[5] https://aws.amazon.com/emr/

[6] https://aws.amazon.com/rds/

[7] https://aws.amazon.com/ecs/

[8] https://aws.amazon.com/sns/

[9] https://aws.amazon.com/blogs/compute/using-aws-lambda-with-auto-scaling-lifecycle-hooks/

[10] http://docs.aws.amazon.com/lambda/latest/dg/welcome.html

[11] https://aws.amazon.com/blogs/aws/new-auto-recovery-for-amazon-ec2/

[12] https://aws.amazon.com/answers/configuration-management/aws-infrastructure-configuration-management/

[13] https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS%20.pdf

[14] http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html#routing-policy-latency

[15] https://aws.amazon.com/elasticache/

[16] https://aws.amazon.com/cloudfront/

[17] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-on-demand-reserved-instances.html

[18] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-instances.html

[19] https://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html

[20] https://aws.amazon.com/about-aws/whats-new/2014/07/31/aws-trusted-advisor-security-and-service-limits-checks-now-free/

[21] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/monitoring_automated_manual.html

[22] http://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch_Dashboards.html

[23] http://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/publishingMetrics.html

[24] https://aws.amazon.com/blogs/aws/new-whitepaper-use-aws-for-disaster-recovery/

[25] http://media.amazonwebservices.com/AWS_Operational_Checklists.pdf

[26] http://d0.awsstatic.com/whitepapers/architecture/AWS_Well-Architected_Framework.pdf

[27] https://aws.amazon.com/premiumsupport/iem/