Pilar Otimização de custos

AWS Well-Architected Framework

Julho de 2020

This paper has been archived.

The latest version is now available at:

https://docs.aws.amazon.com/pt_br/wellarchitected/latest/cost-optimization-pillar/welcome.html

Avisos

Os clientes são responsáveis por avaliar as informações neste documento por conta própria. Este documento (a) é fornecido apenas para fins informativos, (b) representa as ofertas e práticas de produtos atuais da AWS, que estão sujeitas a alterações sem aviso prévio e (c) não cria nenhum compromisso ou garantia da AWS e suas afiliadas, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos no “estado em que se encontram”, sem qualquer garantia, declaração ou condição de qualquer tipo, explícita ou implícita. As responsabilidades e obrigações da AWS com seus clientes são regidas por contratos da AWS, e este documento não modifica nem faz parte de nenhum contrato entre a AWS e seus clientes.

Sumário

Resumo

Este whitepaper destaca o pilar Otimização de custos do Well-Architected Framework da Amazon Web Services (AWS). Ele fornece orientações para ajudar os clientes a aplicar as melhores práticas nas áreas de projeto, entrega e manutenção dos ambientes da AWS.

Uma carga de trabalho com custo otimizado utiliza integralmente todos os recursos, alcança um resultado ao menor ponto de preço possível e atende a seus requisitos funcionais. Este whitepaper fornece orientações detalhadas para a criação de recursos dentro da organização, o projeto da carga de trabalho, a seleção dos serviços, a configuração e a operação dos serviços, além da aplicação de técnicas de otimização de custos.

Introdução

O AWS Well-Architected Framework ajuda a entender as decisões tomadas no momento de criar cargas de trabalho na AWS. O Framework fornece as melhores práticas de arquitetura para projetar e operar sistemas confiáveis, seguros, eficientes e econômicos na nuvem. Ele demonstra uma maneira de avaliar consistentemente suas arquiteturas em relação às melhores práticas e identificar áreas de melhoria. Acreditamos que ter as cargas de trabalho bem arquitetadas aumenta muito a probabilidade de sucesso nos negócios.

A estrutura se baseia em cinco pilares:

- Excelência operacional
- Segurança
- Confiabilidade
- Eficiência de performance
- Otimização de custos

Este documento concentra-se no pilar Otimização de custos e em como projetar cargas de trabalho com o uso mais eficaz de serviços e recursos para atingir resultados de negócio com o menor preço.

Você aprenderá a aplicar as melhores práticas do pilar Otimização de custos em sua organização. A otimização de custos pode ser desafiadora em soluções no local, pois é necessário prever a capacidade futura e as necessidades de negócio enquanto realiza processos de aquisição complexos. A adoção das práticas neste documento ajudará sua organização a atingir estes objetivos:

- Praticar o gerenciamento financeiro na nuvem
- Reconhecimento de despesas e uso
- Recursos econômicos
- Gerenciar recursos de demanda e oferta
- Otimizar ao longo do tempo

Este documento destina-se a profissionais de funções de tecnologia e finanças, como diretores de tecnologia (CTOs), diretores financeiros (CFOs), arquitetos, desenvolvedores, controladores financeiros, planejadores financeiros, analistas de negócios e membros da equipe de operações. Este documento não fornece detalhes de implementação ou padrões de arquitetura. No entanto, inclui referências a recursos apropriados.

Otimização de custos

A otimização de custos é um processo contínuo de refinamento e melhoria durante o período do ciclo de vida de uma carga de trabalho. As práticas deste documento ajudam você a criar e operar cargas de trabalho com reconhecimento de custo que alcançam resultados empresariais, minimizando custos e permitindo que sua organização maximize o retorno sobre o investimento.

Princípios do projeto

Considere os seguintes princípios de design para otimização de custos:

Implementar o gerenciamento financeiro na nuvem: para obter sucesso financeiro e acelerar a realização de valor empresarial na nuvem, você deve investir em gerenciamento financeiro na nuvem. Sua organização deve dedicar o tempo e os recursos necessários para criar aptidão nesse novo domínio de gerenciamento de tecnologia e uso. Semelhante à sua aptidão de Segurança ou Operações, você precisa criar aptidão por meio da criação de conhecimento, programas, recursos e processos para ajudá-lo a se tornar uma organização econômica.

Adotar um modelo de consumo: pague apenas pelos recursos de computação que você consome e aumente ou diminua o uso dependendo dos requisitos da empresa. Por exemplo, ambientes de desenvolvimento e teste normalmente são usados apenas por oito horas ao dia durante a semana de trabalho. Você pode parar esses recursos quando eles não estiverem em uso para obter uma economia de custos potencial de 75% (40 horas versus 168 horas).

Medir a eficiência geral: meça o resultado comercial da carga de trabalho e os custos associados à sua entrega. Use esses dados para entender os ganhos obtidos com o aumento da saída, o aumento da funcionalidade e a redução de custos.

Parar de gastar dinheiro em tarefas pesadas genéricas: a AWS faz o trabalho pesado das operações de datacenter, como o armazenamento em rack, o empilhamento e a alimentação de servidores. Ele também elimina a sobrecarga operacional do gerenciamento de sistemas operacionais e aplicativos com serviços gerenciados. Isso permite que você mantenha o foco em seus clientes e projetos de negócios e não na infraestrutura de TI.

Analisar e atribuir gastos: a nuvem facilita a identificação precisa do custo e uso das cargas de trabalho, o que permite a atribuição transparente de custos de TI para fluxos de receita e proprietários de cargas de trabalho individuais. Isso ajuda a medir o retorno sobre o investimento (ROI) e oferece aos proprietários de cargas de trabalho a oportunidade de otimizar seus recursos e reduzir custos.

Definição

Há cinco áreas de foco para otimização de custos na nuvem:

- Praticar o gerenciamento financeiro na nuvem
- Reconhecimento de despesas e uso
- Recursos econômicos
- Gerenciar recursos de demanda e oferta
- Otimizar ao longo do tempo

Semelhante aos outros pilares do Well-Architected Framework, há compensações a serem consideradas para a otimização de custos. Por exemplo, otimizar para velocidade de entrada no mercado ou para o custo. Em alguns casos, é melhor otimizar a velocidade para entrar no mercado rapidamente, enviar novos recursos ou cumprir um prazo, em vez de investir na otimização de custos inicial.

Às vezes, as decisões de projeto são tomadas com base na pressa e não em dados, já que sempre existe a tentação de compensar excessivamente, em vez de dedicar tempo a realizar benchmarking da implantação mais econômica. A compensação excessiva pode levar a implantações com provisionamento excessivo e subotimizadas. Porém, pode ser uma escolha razoável se você precisa transferir rapidamente recursos de seu ambiente no local para a nuvem e então otimizar posteriormente.

Investir na quantidade certa de esforço em uma estratégia de otimização de custos com antecedência permite aproveitar os benefícios econômicos da nuvem de modo mais rápido, garantindo uma adesão consistente às melhores práticas e evitando provisionamento excessivo desnecessário. As seções a seguir fornecem técnicas e melhores práticas para a implementação inicial e contínua do gerenciamento financeiro na nuvem e otimização de custos para suas cargas de trabalho.

Praticar o gerenciamento financeiro na nuvem

O Cloud Financial Management (CFM – Gerenciamento financeiro na nuvem) permite que as organizações obtenham valor comercial e sucesso financeiro à medida que otimizam o custo, o uso e a escala na AWS.

Veja a seguir as melhores práticas de gerenciamento financeiro na nuvem:

- Propriedade funcional
- Parceria financeira e tecnológica
- Orçamentos e previsões de nuvem
- Processos com reconhecimento de custo
- Cultura com reconhecimento de custo
- Quantificar o valor empresarial distribuído por meio da otimização de custos

Propriedade funcional

Estabelecer uma função de otimização de custos: essa função é responsável por estabelecer e manter uma cultura de reconhecimento de custos. Em toda a organização, tal função pode ser exercida por qualquer pessoa ou equipe existente, ou por uma nova equipe de principais partes interessadas em finanças, tecnologia e organização.

A função (individual ou equipe) prioriza e gasta a porcentagem necessária de seu tempo em atividades de gerenciamento e otimização de custos. Para uma organização pequena, a função pode gastar uma porcentagem de tempo menor em comparação com uma função de tempo integral para uma empresa maior.

A função exige uma abordagem multidisciplinada, com recursos de gerenciamento de projetos, ciência de dados, análise financeira e desenvolvimento de software/infraestrutura. A função pode melhorar a eficiência das cargas de trabalho executando otimizações de custo (abordagem centralizada), influenciando as equipes de tecnologia a executar otimizações (descentralizada) ou usando uma combinação de ambas (híbrida). A função pode ser medida em relação à sua capacidade de executar e entregar em relação às metas de otimização de custos (por exemplo, métricas de eficiência da carga de trabalho).

Você deve garantir o patrocínio executivo para essa função. O patrocinador é considerado defensor do consumo de nuvem econômico e oferece suporte ao escalonamento para a função para garantir que as atividades de otimização de custos sejam tratadas com o nível de prioridade definido pela organização. Juntos, o patrocinador e a função garantem que sua organização consuma a nuvem com eficiência e continue a oferecer valor empresarial.

Parceria financeira e tecnológica

Estabelecer uma parceria entre finanças e tecnologia: as equipes de tecnologia inovam mais rapidamente na nuvem devido à redução dos ciclos de aquisição, aprovação e implantação de infraestrutura. Isso pode ser um ajuste para organizações financeiras anteriormente usadas para executar processos demorados e com uso intensivo de recursos para aquisição e implantação de capital em ambientes de datacenter no local, além de alocação de custos apenas na aprovação do projeto.

Estabelecer uma parceria entre as principais partes interessadas em finanças e tecnologia para criar uma compreensão compartilhada dos objetivos organizacionais e desenvolver mecanismos para obter sucesso financeiro no modelo de gastos variáveis da computação em nuvem. As equipes relevantes da sua organização devem estar envolvidas em discussões de custo e uso em todas as fases da jornada para a nuvem, incluindo:

- **Leads financeiros:** CFOs, controladores financeiros, planejadores financeiros, analistas de negócios, aquisições, sourcing e contas a pagar devem compreender o modelo de nuvem de consumo, as opções de compra e o processo de faturamento mensal. Devido às diferenças fundamentais entre a nuvem (como a taxa de alteração no uso, definição de preço com pagamento conforme o uso, definição de preço em camadas, modelos de definição de preço e informações detalhadas de faturamento e uso) em comparação à operação no local, é essencial que a organização financeira entenda como o uso da nuvem pode afetar aspectos empresariais, incluindo processos de aquisição, rastreamento de incentivos, alocação de custos e demonstrações financeiras.

- **Leads de tecnologia:** os líderes de tecnologia (incluindo proprietários de produtos e aplicativos) devem estar cientes dos requisitos financeiros (por exemplo, restrições orçamentárias), bem como dos requisitos de negócios (por exemplo, contratos de nível de serviço). Isso permite que a carga de trabalho seja implementado para atingir os objetivos desejados da organização.

A parceria entre finanças e tecnologia oferece os seguintes benefícios:

- As equipes de finanças e tecnologia têm visibilidade praticamente em tempo real dos custos e do uso.

- As equipes de finanças e tecnologia estabelecem um procedimento operacional padrão para lidar com a variação de gastos na nuvem.

- As partes interessadas nas finanças atuam como consultores estratégicos com relação à forma como o capital é usado para comprar descontos de compromissos (por exemplo, instâncias reservadas ou Savings Plans da AWS) e como a nuvem é usada para expandir a organização.

- As contas a pagar e os processos de aquisição existentes são usados com a nuvem.

- As equipes de finanças e tecnologia colaboram na previsão de custos e uso futuros da AWS para alinhar/criar orçamentos organizacionais.

- Melhor comunicação entre organizações por meio de uma linguagem compartilhada e entendimento comum dos conceitos financeiros.

As partes interessadas adicionais dentro da sua organização que devem ser envolvidas em discussões de custo e uso incluem:

- **Proprietários de unidades de negócios:** os proprietários de unidades de negócios devem compreender o modelo de negócios de nuvem para que possam fornecer orientações tanto para as unidades de negócios quanto para toda a empresa. Esse conhecimento de nuvem é essencial quando há necessidade de prever o crescimento e o uso da carga de trabalho, e ao avaliar opções de compra de longo prazo, como instâncias reservadas ou Savings Plans.
- **Terceiros:** se sua organização usa terceiros (por exemplo, consultores ou ferramentas), certifique-se de que eles estejam alinhados com seus objetivos financeiros e possam demonstrar o alinhamento por meio de seus modelos de engajamento e um retorno sobre o investimento (ROI). Terceiros normalmente contribuirão para o relatório e a análise de qualquer carga de trabalho que gerenciem e fornecerão análise de custo de qualquer carga de trabalho que projetem.

Orçamentos e previsões de nuvem

Estabelecer orçamentos e previsões de nuvem: os clientes usam a nuvem para obter eficiência, velocidade e agilidade, o que cria uma quantidade altamente variável de custo e uso. Os custos podem diminuir com o aumento na eficiência da carga de trabalho ou à medida que novas cargas de trabalho e recursos são implantados. Ou as cargas de trabalho serão escaladas para atender a mais clientes, o que aumenta o uso e os custos da nuvem. Os processos de orçamento organizacional existentes devem ser modificados para incorporar essa variabilidade.

Ajuste os processos de previsão e orçamento existentes para se tornarem mais dinâmicos usando um algoritmo baseado em tendências (usando custos históricos como entradas). Também é possível usar algoritmos baseados em orientadores de negócios (por exemplo, lançamentos de novos produtos ou expansão regional) ou uma combinação de tendências e orientadores de negócios.

Você pode usar o AWS Cost Explorer para prever custos de nuvem diários (até 3 meses) ou mensais (até 12 meses) com base em algoritmos de machine learning aplicados aos seus custos históricos (com base em tendências).

Processos com reconhecimento de custo

Implementar o reconhecimento de custos em seus processos organizacionais: o reconhecimento de custos deve ser implementado em processos organizacionais novos e existentes. Recomendamos reutilizar e modificar processos existentes sempre que possível, o que minimiza o impacto na agilidade e velocidade. As recomendações a seguir ajudarão a implementar o reconhecimento de custos em sua carga de trabalho:

- Certifique-se de que o gerenciamento de alterações inclua uma medição de custo para quantificar o impacto financeiro das alterações. Isso ajuda a abordar de forma proativa as preocupações relacionadas a custos e a destacar as economias de custos.

- Certifique-se de que a otimização de custos seja um componente essencial de seus recursos operacionais. Por exemplo, você pode aproveitar os processos existentes de gerenciamento de incidentes para investigar e identificar a causa raiz das anomalias de custo e uso (excedentes de custo).

- Acelere a economia de custos e a obtenção de valor empresarial por meio da automação ou das ferramentas. Ao pensar sobre o custo da implementação, enquadre a conversa para incluir um componente de ROI para justificar o investimento de tempo ou dinheiro.

- Estenda os programas de treinamento e desenvolvimento existentes para incluir treinamento com reconhecimento de custos em toda a organização. Recomendamos que isso inclua treinamento e certificação contínuos. Isso criará uma organização capaz de autogerenciar custos e uso.

Relatar e notificar sobre otimização de custos e uso: você deve informar regularmente sobre a otimização de custos e uso dentro da sua organização. Você pode implementar sessões dedicadas para otimização de custos ou incluir otimização de custos em seus ciclos regulares de relatórios operacionais para suas cargas de trabalho. O AWS Cost Explorer fornece painéis e relatórios. Você pode acompanhar seu progresso do custo e do uso em relação a orçamentos configurados com os Relatórios do Orçamentos da AWS.

Você também pode usar o Amazon QuickSight com dados do Cost and Usage Report (CUR – Relatório de uso e custos) para fornecer relatórios altamente personalizados com dados mais granulares.

Implemente notificações sobre custo e uso para garantir que alterações no custo e no uso possam ser respondidas rapidamente. O Orçamentos da AWS permite que você forneça notificações em relação a metas. Recomendamos configurar notificações sobre aumentos e diminuições e no custo e no uso das cargas de trabalho.

Monitorar custos e uso proativamente: é recomendável monitorar custos e uso proativamente em sua organização, e não apenas quando há exceções ou anomalias. Painéis altamente visíveis em todo o escritório ou ambiente de trabalho garantem que as principais pessoas tenham acesso às informações necessárias e indicam o foco da organização na otimização de custos. Os painéis visíveis permitem promover ativamente resultados bem-sucedidos e implementá-los em toda a organização.

Cultura com reconhecimento de custo

Criar uma cultura com reconhecimento de custos: implemente alterações ou programas em toda a organização para criar uma cultura com reconhecimento de custos. É recomendável começar aos poucos e, à medida que seus recursos aumentarem e o uso da nuvem por sua organização aumentar, implementar programas grandes e abrangentes.

Uma cultura com reconhecimento de custos permite escalar a otimização de custos e o gerenciamento financeiro na nuvem por meio de melhores práticas executadas de maneira orgânica e descentralizada em toda a organização. Isso cria altos níveis de capacidade em toda a organização com o mínimo de esforço, em comparação com uma abordagem centralizada e de cima para baixo.

Pequenas mudanças na cultura podem ter grandes impactos na eficiência de suas cargas de trabalho atuais e futuras. Exemplos disso incluem:

- Gamificação do custo e do uso em toda a organização. Isso pode ser feito por meio de um painel visível publicamente ou de um relatório que compara custos e uso normalizados entre equipes (por exemplo, custo por carga de trabalho, custo por transação).
- Reconhecimento da eficiência de custos. Recompense realizações de otimização de custos voluntárias ou não solicitadas publicamente ou de forma privada e aprenda com os erros para evitar repeti-los no futuro.
- Crie requisitos organizacionais de cima para baixo para cargas de trabalho a serem executadas em orçamentos predefinidos.

Mantenha-se atualizado com os novos lançamentos de serviços: você pode implementar novos serviços e recursos da AWS para aumentar a eficiência de custos em sua carga de trabalho. Analise regularmente o Blog de notícias da AWS, o Blog de gerenciamento de custos da AWS, e as Novidades da AWS para obter informações sobre lançamentos de novos serviços e recursos.

Quantificar o valor empresarial distribuído por meio da otimização de custos

Quantifique o valor empresarial da otimização de custos: além de relatar economias com base na otimização de custos, é recomendável quantificar o valor adicional entregue. Os benefícios de otimização de custos normalmente são quantificados em termos de custos mais baixos por resultado comercial. Por exemplo, você pode quantificar as economias de custo do Amazon Elastic Compute Cloud (Amazon EC2) sob demanda ao comprar Savings Plans, que reduzem custos e mantêm os níveis de saída da carga de trabalho. Você pode quantificar reduções de custos nos gastos da AWS quando instâncias ociosas do Amazon EC2 são encerradas ou volumes não vinculados do Amazon Elastic Block Store (Amazon EBS) são excluídos.

A quantificação do valor empresarial da otimização de custos permite que você entenda todo o conjunto de benefícios da sua organização. Como a otimização de custos é um investimento necessário, quantificar o valor empresarial permite que você explique o retorno sobre o investimento para as partes interessadas. A quantificação do valor empresarial pode ajudá-lo a ganhar mais participação das partes interessadas em futuros investimentos de otimização de custos e fornece uma estrutura para medir os resultados das atividades de otimização de custos da sua organização.

No entanto, os benefícios da otimização de custos vão além da redução ou da prevenção de custos. Considere a captura de dados adicionais para medir melhorias de eficiência e valor empresarial. Exemplos de melhoria incluem:

- **Execução de melhores práticas de otimização de custos:** por exemplo, o gerenciamento do ciclo de vida de recursos reduz os custos operacionais e de infraestrutura e cria tempo e orçamento inesperado para experimentação. Isso aumenta a agilidade da organização e descobre novas oportunidades de geração de receita.
- **Implementação de automação:** por exemplo, Auto Scaling, que garante elasticidade com esforço mínimo, e aumenta a produtividade da equipe eliminando o trabalho de planejamento de capacidade manual. Para obter mais detalhes sobre resiliência operacional, consulte o whitepaper Pilar Confiabilidade do Well-Architected.
- **Previsão de custos futuros da AWS:** a previsão permite que as partes interessadas financeiras definam expectativas com outras partes interessadas internas e externas da organização, além de ajudar a melhorar a previsibilidade financeira da sua organização. O AWS Cost Explorer pode ser usado para realizar previsões de custo e uso.

Recursos

Consulte os recursos a seguir para saber mais sobre as melhores práticas da AWS para orçamento e previsão de gastos na nuvem.

- Relatar suas métricas de orçamento com relatórios de orçamento
- Previsão com o AWS Cost Explorer
- Treinamento da AWS
- AWS Certification
- Parceiros de Ferramentas de gerenciamento da Nuvem AWS

Reconhecimento de despesas e uso

Entender os custos e os orientadores da sua organização é essencial para gerenciar seus custos e uso com eficácia e identificar oportunidades de redução de custos. Normalmente, as organizações operam várias cargas de trabalho executadas por várias equipes. Essas equipes podem estar em diferentes unidades da organização, cada uma com o próprio fluxo de receita. A capacidade de atribuir custos de recursos a cargas de trabalho, à organização individual ou aos proprietários do produto gera um comportamento eficiente do uso e ajuda a reduzir o desperdício. O monitoramento preciso de custos e uso permite que você entenda como as unidades e os produtos da organização são rentáveis e permite que você tome decisões mais embasadas sobre onde alocar recursos dentro da sua organização.
A conscientização sobre o uso em todos os níveis da organização é essencial para promover mudanças, pois a mudança no uso gera mudanças no custo.

Considere adotar uma abordagem multifacetada para se tornar ciente do seu uso e das suas despesas. Sua equipe precisa reunir dados, analisá-los e então relatá-los. Os principais fatores a serem considerados incluem:

- Governança
- Monitoramento de custo e uso
- Desativação

Governança

Para gerenciar seus custos na nuvem, você deve gerenciar seu uso por meio das áreas de governança abaixo:

Desenvolver políticas organizacionais: o primeiro passo para executar governança é usar os requisitos da sua organização para desenvolver políticas para o uso da nuvem. Essas políticas definem como sua organização usa a nuvem e como os recursos são gerenciados. As políticas devem cobrir todos os aspectos de recursos e cargas de trabalho relacionados ao custo ou uso, incluindo criação, modificação e desativação durante a vida útil do recurso.

As políticas devem ser simples para que sejam facilmente compreendidas e possam ser implementadas com eficácia em toda a organização. Comece com políticas amplas e de alto nível, como em qual região geográfica o uso é permitido ou horários do dia em que os recursos devem estar em execução. Refine gradualmente as políticas para as várias unidades organizacionais e cargas de trabalho. As políticas comuns incluem quais serviços e recursos podem ser usados (por exemplo, armazenamento de menor performance em ambientes de teste/desenvolvimento) e quais tipos de recursos podem ser usados por diferentes grupos (por exemplo, o maior tamanho de recursos em uma conta de desenvolvimento é médio).

Desenvolver objetivos e metas: desenvolva objetivos e metas de custo e uso para a sua organização. Os objetivos fornecem orientações e direcionamento para a sua organização sobre os resultados esperados. As metas fornecem resultados mensuráveis específicos a serem alcançados. Um exemplo de um objetivo é: o uso da plataforma deve aumentar significativamente, com apenas um pequeno aumento (não linear) no custo. Um exemplo de meta é: um aumento de 20% no uso da plataforma, com um aumento de menos de 5% nos custos. Outro objetivo comum é que as cargas de trabalho precisam ser mais eficientes a cada seis meses. A meta acompanhante seria que o custo por saída da carga de trabalho precisa diminuir em 5% a cada 6 meses.

Um objetivo comum para cargas de trabalho na nuvem é aumentar a eficiência da carga de trabalho, que é diminuir o custo por resultado comercial da carga de trabalho ao longo do tempo. É recomendável implementar essa meta para todas as cargas de trabalho e também definir uma meta como um aumento de 5% na eficiência a cada 6 a 12 meses. Isso pode ser obtido na nuvem por meio da criação de recursos na otimização de custos e do lançamento de novos serviços e recursos de serviços.

Estrutura da conta: a AWS tem uma estrutura de conta de um pai para muitos filhos, que é comumente conhecida como conta mestre (o pai, anteriormente pagante)-conta membro (o filho, anteriormente vinculado). Uma melhor prática é sempre ter pelo menos uma conta mestre com uma conta membro, independentemente do tamanho ou uso da sua organização. Todos os recursos de carga de trabalho devem residir somente em contas membro.

Não há uma resposta geral para quantas contas da AWS você deve ter. Avalie seus modelos de custo e operacionais atuais e futuros para garantir que a estrutura de suas contas da AWS reflita os objetivos da sua organização. Algumas empresas criam várias contas da AWS por motivos comerciais, por exemplo:

- O isolamento administrativo e/ou fiscal e de faturamento é necessário entre unidades da organização, centros de custo ou cargas de trabalho específicas.
- Os limites de serviço da AWS são definidos para serem específicos para cargas de trabalho específicas.
- Há um requisito de isolamento e separação entre cargas de trabalho e recursos.

Dentro do AWS Organizations, o faturamento consolidado cria a construção entre uma ou mais contas membro e a conta mestre. As contas membro permitem que você isole e diferencie seu custo e uso por grupos. Uma prática comum é ter contas membro separadas para cada unidade da organização (como finanças, marketing e vendas), ou para cada ciclo de vida do ambiente (como desenvolvimento, teste e produção) ou para cada carga de trabalho (carga de trabalho a, b e c) e, em seguida, agregar essas contas vinculadas usando o faturamento consolidado.

O faturamento consolidado permite que você consolide o pagamento de várias contas membro da AWS em uma única conta mestre, sem deixar de oferecer visibilidade para a atividade de cada conta vinculada. À medida que os custos e o uso são agregados na conta mestre, isso permite que você maximize seus descontos por volume de serviço e maximize o uso de seus descontos de compromisso (Savings Plans e instâncias reservadas) para alcançar os descontos mais altos.

O AWS Control Tower pode instalar e configurar rapidamente várias contas da AWS, garantindo que a governança esteja alinhada com os requisitos da sua organização.

Grupos e funções organizacionais: depois de desenvolver políticas, você pode criar funções e grupos lógicos de usuários em sua organização. Isso permite que você atribua permissões e controle o uso. Comece com agrupamentos de alto nível de pessoas, normalmente isso se alinha a unidades organizacionais e funções de trabalho (por exemplo, administrador de sistemas no departamento de TI ou controlador financeiro). Os grupos juntam pessoas que realizam tarefas semelhantes e precisam de acesso semelhante. As funções definem o que um grupo deve fazer. Por exemplo, um administrador de sistemas em TI requer acesso para criar todos os recursos, mas um membro da equipe de análise só precisa criar recursos de análise.

Controles - Notificações: uma primeira etapa comum na implementação de controles de custo é configurar notificações quando eventos de custo ou uso ocorrerem fora das políticas. Isso permite que você aja rapidamente e verifique se é necessária uma ação corretiva, sem restringir ou afetar negativamente cargas de trabalho ou novas atividades. Depois de conhecer os limites da carga de trabalho e do ambiente, você pode impor a governança. Na AWS, as notificações são realizadas com Orçamentos da AWS, o que permite definir um orçamento mensal para seus custos, uso e descontos de compromisso da AWS (Savings Plans e instâncias reservadas). Você pode criar orçamentos em um nível de custo agregado (por exemplo, todos os custos) ou em um nível mais granular, onde você inclui apenas dimensões específicas, como contas vinculadas, serviços, tags ou zonas de disponibilidade. Você também pode anexar notificações por e-mail aos seus orçamentos, que serão acionadas quando os custos ou o uso atuais ou previstos excederem um limite percentual definido.

Controles - imposição: como uma segunda etapa, você pode aplicar políticas de governança na AWS por meio do AWS Identity and Access Management (IAM), e das Service Control Policies (SCP – Políticas de controle de serviço) do AWS Organizations. O IAM permite que você gerencie com segurança o acesso aos serviços e recursos da AWS. Usando o IAM, você pode controlar quem pode criar e gerenciar recursos da AWS, os tipos de recursos que podem ser criados e onde eles podem ser criados. Isso minimiza a criação de recursos que não são necessários. Use as funções e grupos criados anteriormente e atribua políticas do IAM para impor o uso correto. A SCP oferece controle central sobre o número máximo de permissões disponíveis para todas as contas na sua organização, garantindo que suas contas permaneçam dentro das diretrizes de controle de acesso. As SCPs estão disponíveis somente em uma organização com todos os recursos habilitados, e você pode configurar as SCPs para negar ou permitir ações para contas membro por padrão. Consulte o whitepaper Pilar Segurança do Well-Architected para obter mais detalhes sobre a implementação do gerenciamento de acesso.

Controles - Service Quotas: a governança também pode ser implementada por meio do gerenciamento de Service Quotas. Ao garantir que o Service Quotas seja configurado com o mínimo de sobrecarga e mantido com precisão, você pode minimizar a criação de recursos fora dos requisitos da sua organização. Para conseguir isso, você deve entender a rapidez com que seus requisitos podem mudar, compreender projetos em andamento (criação e desativação de recursos) e considerar a rapidez com que as alterações de cota podem ser implementadas. O Service Quotas pode ser usadas para aumentar suas cotas quando necessário.

Os serviços de gerenciamento de custos da AWS são integrados ao serviço AWS Identity and Access Management (IAM). Você usa o serviço IAM em conjunto com os serviços de gerenciamento de custos para controlar o acesso aos seus dados financeiros e às ferramentas da AWS no console de faturamento.

Acompanhe o ciclo de vida da carga de trabalho: acompanhe todo o ciclo de vida da carga de trabalho. Isso garante que, quando cargas de trabalho ou componentes de carga de trabalho não forem mais necessários, eles possam ser desativados ou modificados. Isso é especialmente útil quando você lança novos serviços ou recursos. As cargas de trabalho e os componentes existentes podem parecer estar em uso, mas devem ser desativados para redirecionar os clientes para o novo serviço. Observe estágios anteriores das cargas de trabalho - depois que uma carga de trabalho está em produção, os ambientes anteriores podem ser desativados ou significativamente reduzidos na capacidade até que sejam necessários novamente.

A AWS fornece uma série de serviços de gerenciamento e governança que você pode usar para o rastreamento do ciclo de vida da entidade. Você pode usar o AWS Config ou o AWS Systems Manager para fornecer um inventário detalhado dos recursos e da configuração da AWS. Recomendamos que você o integre com seus sistemas existentes de gerenciamento de projetos ou ativos para acompanhar projetos e produtos ativos em sua organização. A combinação do seu sistema atual com o conjunto completo de eventos e métricas fornecido pela AWS permite que você crie uma visão de eventos de ciclo de vida significativos e gerencie recursos proativamente para reduzir custos desnecessários.

Consulte o whitepaper Pilar Excelência operacional do Well-Architected para obter mais detalhes sobre a implementação do rastreamento do ciclo de vida da entidade.

Monitorar custos e uso

Permita que as equipes atuem em seu custo e uso por meio de visibilidade detalhada da carga de trabalho. A otimização de custos começa com uma compreensão granular da discriminação de custo e uso, da capacidade de modelar e prever gastos, uso e recursos futuros e da implementação de mecanismos suficientes para alinhar custos e uso aos objetivos da sua organização. Veja a seguir as áreas necessárias para monitorar seu custo e uso:

Configurar fontes de dados detalhadas: habilite a granularidade por hora no Cost Explorer e crie um Cost and Usage Report (CUR - Relatório de custo e uso). Essas fontes de dados oferecem a visualização mais precisa do custo e do uso em toda a organização. O CUR fornece granularidade de uso diário ou por hora, taxas, custos e atributos de uso para todos os serviços da AWS cobráveis. Todas as dimensões possíveis estão no CUR, incluindo: marcação, localização, atributos de recurso e IDs de conta.

Configure seu CUR com as seguintes personalizações:

- Incluir IDs de recurso
- Atualizar automaticamente o CUR
- Granularidade por hora
- Controle de versões: substituir relatório existente
- Integração de dados: Athena (formato Parquet e compactação)

Use o AWS Glue para preparar os dados para análise e use o Amazon Athena para executar análises de dados usando SQL para consultar os dados. Você também pode usar o Amazon QuickSight para criar visualizações personalizadas e complexas e distribuí-las em toda a organização.

Identificar categorias de atribuição de custos: trabalhe com sua equipe financeira e outras partes interessadas relevantes para compreender os requisitos de como os custos devem ser alocados em sua organização. Os custos da carga de trabalho devem ser alocados durante todo o ciclo de vida, incluindo desenvolvimento, teste, produção e desativação. Entenda como os custos incorridos para o aprendizado, o desenvolvimento da equipe e a criação de ideias são atribuídos na organização. Isso pode ser útil para alocar corretamente contas usadas para essa finalidade para orçamentos de treinamento e desenvolvimento, em vez de orçamentos genéricos de custo de TI.

Estabelecer métricas de carga de trabalho: entenda como a saída da carga de trabalho é medida em relação ao sucesso empresarial. Cada carga de trabalho normalmente tem um pequeno conjunto de saídas principais que indicam performance. Se você tiver uma carga de trabalho complexa com muitos componentes, poderá priorizar a lista ou definir e rastrear métricas para cada componente. Trabalhe com suas equipes para entender quais métricas usar. Essa unidade será usada para compreender a eficiência da carga de trabalho ou o custo de cada saída de negócios.

Atribuir significados organizacionais ao custo e uso: implemente a marcação na AWS para adicionar informações da organização aos seus recursos, que serão adicionadas às suas informações de custo e uso. Uma tag é um par de chave-valor - a chave é definida e deve ser exclusiva em toda a organização, e o valor é exclusivo para um grupo de recursos. Um exemplo de um par de chave-valor é a chave Environment, com um valor de Production. Todos os recursos no ambiente de produção terão esse par de chave-valor. A marcação permite categorizar e rastrear seus custos com informações relevantes e significativas da organização. Você pode aplicar tags que representam categorias da organização (como centros de custo, nomes de aplicativos, projetos ou proprietários) e identificar cargas de trabalho e características de cargas de trabalho (como teste ou produção) para atribuir seus custos e uso em toda a organização.

Quando você aplica tags a seus recursos da AWS (como instâncias do EC2 ou buckets do Amazon S3) e ativa as tags, a AWS adiciona essas informações aos relatórios de custo e uso. Você pode gerar relatórios e realizar análises em recursos marcados e não marcados para permitir maior conformidade com políticas internas de gerenciamento de custos e garantir a atribuição precisa.

Criar e implementar um padrão de marcação da AWS em todas as contas da organização permite que você gerencie e administre seus ambientes da AWS de maneira consistente e uniforme. Use políticas de tags no AWS Organizations para definir regras de como as tags podem ser usadas em recursos da AWS em suas contas no AWS Organizations. As políticas de tags permitem que você adote facilmente uma abordagem padronizada para marcar recursos da AWS.

O AWS Tag Editor permite adicionar, excluir e gerenciar tags de vários recursos.

As Cost Categories da AWS permitem que você atribua significado da organização aos seus custos, sem exigir tags nos recursos. Você pode mapear suas informações de custo e uso para estruturas internas exclusivas da organização. Você define regras de categoria para mapear e categorizar custos usando dimensões de faturamento, como contas e tags. Isso fornece outro nível de capacidade de gerenciamento, além da marcação. Você também pode mapear contas e tags específicas para vários projetos.

Configurar ferramentas de faturamento e otimização de custos: para modificar o uso e ajustar os custos, cada pessoa em sua organização deve ter acesso às suas informações de custo e uso. Recomendamos que todas as cargas de trabalho e equipes tenham as seguintes ferramentas configuradas ao usar a nuvem:

- **Relatórios:** resumo de todas as informações de custo e uso.
- **Notificações:** forneça notificações quando o custo ou o uso estiverem fora dos limites definidos.
- **Estado atual:** configure um painel mostrando os níveis atuais de custo e uso. O painel deve estar disponível em um local altamente visível dentro do ambiente de trabalho (semelhante a um painel de operações do).
- **Tendências:** forneça o recurso para mostrar a variabilidade de custo e uso ao longo do período de tempo necessário, com a granularidade necessária.
- **Previsões:** forneça o recurso para mostrar custos futuros estimados.
- **Rastreamento:** mostra o custo e o uso atuais em relação a metas ou objetivos configurados.
- **Análise:** disponibilize a capacidade para que os membros da equipe realizem análises personalizadas e detalhadas até a granularidade horária, com todas as dimensões possíveis.

Você pode usar ferramentas nativas da AWS, como o AWS Cost Explorer, o Orçamentos da AWS, e o Amazon Athena com o QuickSight para fornecer esse recurso. Você também pode usar ferramentas de terceiros. No entanto, você deve garantir que os custos dessas ferramentas forneçam valor à sua organização.

Alocar custos com base nas métricas de carga de trabalho: a otimização de custos está fornecendo resultados de negócios com o menor preço, que só pode ser alcançado ao alocar custos de carga de trabalho por métricas de carga de trabalho (medidas pela eficiência da carga de trabalho). Monitore as métricas de carga de trabalho definidas por meio de arquivos de log ou outro monitoramento de aplicativos. Combine esses dados com os custos da carga de trabalho, que podem ser obtidos examinando os custos com um valor de tag específico ou ID de conta. É recomendável executar essa análise no nível por hora. Sua eficiência normalmente mudará se você tiver alguns componentes de custo estático (por exemplo, um banco de dados de back-end em execução 24 horas por dia, 7 dias por semana) com uma taxa de solicitações variável (por exemplo, picos de uso entre 9h e 17h, com poucas solicitações à noite). Entender a relação entre os custos estáticos e variáveis ajudará você a concentrar suas atividades de otimização.

Recursos de desativação

Depois de gerenciar uma lista de projetos, funcionários e recursos de tecnologia ao longo do tempo, você poderá identificar quais recursos não estão mais sendo usados e quais projetos não têm mais um proprietário.

Acompanhar recursos durante sua vida útil: desative recursos de carga de trabalho que não são mais necessários. Um exemplo comum são os recursos usados para testes, após a conclusão do teste, os recursos podem ser removidos. Rastrear recursos com tags (e executar relatórios sobre essas tags) ajudará você a identificar ativos para desativação. Usar tags é uma maneira eficaz de rastrear recursos, rotulando o recurso com sua função ou uma data conhecida em que ele pode ser desativado. Os relatórios podem ser executados nessas tags. Os valores de exemplo para marcação de recursos são "testes de featureX" para identificar a finalidade do recurso em termos de ciclo de vida da carga de trabalho.

Implementar um processo de desativação: implemente um processo padronizado em toda a organização para identificar e remover recursos não utilizados. O processo deve definir a frequência das pesquisas e os processos para remover o recurso para garantir que todos os requisitos da organização sejam atendidos.

Recursos de desativação: a frequência e o esforço para pesquisar recursos não utilizados devem refletir as possíveis economias, portanto, uma conta com um custo pequeno deve ser analisada com menos frequência do que uma conta com custos maiores. Pesquisas e eventos de desativação podem ser acionados por alterações de estado na carga de trabalho, como um produto que termina a vida útil ou é substituído. Pesquisas e eventos de desativação também podem ser acionados por eventos externos, como alterações nas condições de mercado ou encerramento do produto.

Recursos de desativação automaticamente: use a automação para reduzir ou remover os custos associados do processo de desativação. Projetar sua carga de trabalho para executar a desativação automatizada reduzirá os custos gerais da carga de trabalho durante sua vida útil. Você pode usar o AWS Auto Scaling para executar o processo de desativação. Você também pode implementar código personalizado usando a API ou o SDK para desativar recursos de carga de trabalho automaticamente.

Recursos

Consulte os recursos a seguir para saber mais sobre as melhores práticas da AWS para conscientização sobre gastos.

- Estratégias de marcação da AWS
- Ativar tags de alocação de custos definidas pelo usuário
- AWS Billing and Cost Management
- Blog de gerenciamento de custos
- Estratégia de faturamento de várias contas
- AWS SDK e ferramentas
- Melhores práticas de marcação
- Laboratórios do Well-Architected - Fundamentos de custo
- Laboratórios do Well-Architected - Conscientização de despesas

Recursos econômicos

Usar os serviços, os recursos e as configurações adequados para suas cargas de trabalho é essencial para economias de custo. Considere o seguinte ao criar recursos econômicos:

- Avaliar o custo ao selecionar serviços
- Selecione o tipo de recurso, o tamanho e o número corretos
- Selecione o melhor modelo de definição de preço
- Planejar a transferência de dados

Você pode usar os arquitetos de soluções da AWS, as soluções da AWS, as arquiteturas de referência da AWS e os parceiros do APN para ajudá-lo a escolher uma arquitetura baseada no que você aprendeu.

Avaliar o custo ao selecionar serviços

Identificar os requisitos da organização: ao selecionar serviços para sua carga de trabalho, é fundamental compreender suas prioridades da organização. Verifique se você tem um equilíbrio entre custo e outros pilares do Well-Architected, como performance e confiabilidade. Uma carga de trabalho totalmente otimizada para custo é a solução mais alinhada aos requisitos da sua organização, não necessariamente o menor custo. Reúna-se com todas as equipes da sua organização para coletar informações, como produtos, negócios, técnicas e financeiras.

Analisar todos os componentes da carga de trabalho: execute uma análise completa de todos os componentes da carga de trabalho. Garantir o equilíbrio entre o custo da análise e as possíveis economias na carga de trabalho durante seu ciclo de vida. Você deve encontrar o impacto atual e o possível impacto futuro do componente. Por exemplo, se o custo do recurso proposto for de 10 USD/mês, e sob as cargas previstas não excederem 15 USD/mês, gastar um dia de esforço para reduzir custos em 50% (5 USD por mês) poderá exceder o benefício potencial durante a vida útil do sistema. Usar uma estimativa baseada em dados mais rápida e eficiente criará o melhor resultado geral para esse componente.

As cargas de trabalho podem mudar ao longo do tempo. O conjunto certo de serviços pode não ser ideal se a arquitetura da carga de trabalho ou o uso mudar. A análise para seleção de serviços deve incorporar estados de carga de trabalho e níveis de uso atuais e futuros. A implementação de um serviço para o estado ou uso futuro da carga de trabalho pode reduzir os custos gerais ao reduzir ou remover o esforço necessário para fazer alterações futuras.

O AWS Cost Explorer e o CUR podem analisar o custo de uma prova de conceito (PoC) ou um ambiente em execução. Você também pode usar a Calculadora Mensal da AWS ou a Calculadora de definição de preço da AWS para estimar os custos da carga de trabalho.

Serviços gerenciados: os serviços gerenciados eliminam a sobrecarga operacional e administrativa da manutenção de um serviço, o que permite que você se concentre na inovação. Além disso, como serviços gerenciados operam em escala da nuvem, eles podem oferecer menor custo por transação ou serviço.

Considere a economia de tempo que permitirá que sua equipe se concentre na aposentadoria de recursos de endividamento técnico, inovação e agregação de valor. Por exemplo, talvez você precise transferir rapidamente seu ambiente no local para a nuvem e otimizar mais tarde. Vale a pena explorar as economias que você poderia obter usando serviços gerenciados que removem ou reduzem os custos de licença.

Geralmente, os serviços gerenciados têm atributos que você pode definir para garantir capacidade suficiente. Você deve definir e monitorar esses atributos para que sua capacidade em excesso seja mínima e a performance seja maximizada. Você pode modificar os atributos do AWS Managed Services usando o Console de Gerenciamento da AWS ou os SDKs e as APIs da AWS para alinhar as necessidades de recursos às mudanças na demanda. Por exemplo, você pode aumentar ou diminuir o número de nós em um cluster Amazon EMR (ou um cluster Amazon Redshift) para aumentar ou reduzir a escala.

Você também pode unir várias instâncias em um recurso da AWS para habilitar usos de maior densidade. Por exemplo, você pode provisionar vários pequenos bancos de dados em uma única instância de banco de dados do Amazon Relational Database Service (Amazon RDS). Conforme o uso aumenta, você pode migrar um dos bancos de dados para uma instância de banco de dados RDS dedicada usando um processo de snapshot e restauração.

Ao provisionar cargas de trabalho em serviços gerenciados, você deve compreender os requisitos de ajuste da capacidade do serviço. Esses requisitos geralmente são tempo, esforço e qualquer impacto na operação normal da carga de trabalho. O recurso provisionado deve permitir tempo para que as alterações ocorram, provisionar a sobrecarga necessária para permitir isso. O trabalho contínuo necessário para modificar os serviços pode ser reduzido a praticamente zero usando APIs e SDKs integrados a ferramentas de sistema e monitoramento como o Amazon CloudWatch.

O Amazon Relational Database Service (RDS), o Amazon Redshift, e o Amazon ElastiCache oferecem um serviço de banco de dados gerenciado. O Amazon Athena, o Amazon Elastic Map Reduce (EMR), e o Amazon Elasticsearch oferecem um serviço de análise gerenciado.

O AWS Managed Services (AMS) é um serviço que opera a infraestrutura da AWS em nome de clientes e parceiros empresariais. Ele fornece um ambiente seguro e compatível no qual você pode implantar suas cargas de trabalho. O AMS usa modelos operacionais de nuvem empresarial com automação para permitir que você atenda aos requisitos da sua organização, migre para a nuvem mais rapidamente e reduza seus custos de gerenciamento constantes.

Serviços sem servidor ou no nível do aplicativo: você pode usar serviços sem servidor ou no nível do aplicativo, como AWS Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS), e Amazon Simple Email Service (Amazon SES). Esses serviços eliminam a necessidade de gerenciar um recurso e fornecem a função de execução de código, serviços de enfileiramento e entrega de mensagens. O outro benefício é que eles escalam a performance e o custo de acordo com o uso, permitindo a alocação e a atribuição eficientes de custos.

Para obter mais informações sobre o Serverless, consulte o whitepaper Well-Architected Serverless Application.

Analisar a carga de trabalho para uso diferente ao longo do tempo: à medida que a AWS lança novos serviços e recursos, os serviços ideais para sua carga de trabalho podem mudar. O esforço necessário deve refletir possíveis benefícios. A frequência da análise da carga de trabalho depende dos requisitos da sua organização. Se for uma carga de trabalho com custo significativo, implementar novos serviços mais cedo maximizará a economia de custos, portanto, uma revisão mais frequente poderá ser vantajosa. Outro trigger para revisão é a alteração nos padrões de uso. Alterações significativas no uso podem indicar que serviços alternativos seriam mais ideais. Por exemplo, para taxas de transferência de dados mais altas, um serviço de conexão direta pode ser mais barato do que uma VPN e fornecer a conectividade necessária. Preveja o possível impacto das alterações de serviço para que você possa monitorar esses triggers de nível de uso e implementar os serviços mais econômicos mais cedo.

Custos de licenciamento: o custo das licenças de software pode ser eliminado com o uso de software de código aberto. Isso pode ter impacto significativo nos custos da carga de trabalho à medida que o tamanho da carga de trabalho é dimensionado. Meça os benefícios do software licenciado em relação ao custo total para garantir que você tenha a carga de trabalho mais otimizada. Modele todas as alterações no licenciamento e como elas afetariam seus custos de carga de trabalho. Se um fornecedor alterar o custo da sua licença de banco de dados, investigue como isso afeta a eficiência geral da sua carga de trabalho. Considere anúncios históricos de definição de preço de seus fornecedores para tendências de alterações de licenciamento em seus produtos. Os custos de licenciamento também podem ser dimensionados independentemente do throughput ou do uso, como licenças que escalam por hardware (licenças vinculadas à CPU). Essas licenças devem ser evitadas porque os custos podem aumentar rapidamente sem resultados correspondentes.

Você pode usar o AWS License Manager para gerenciar as licenças de software na sua carga de trabalho. Você pode configurar regras de licenciamento e aplicar as condições necessárias para ajudar a evitar violações de licenciamento e também reduzir custos devido a excedentes de licença.

Selecione o tipo de recurso, o tamanho e o número corretos

Ao selecionar o melhor tipo de recurso, tamanho e número de recursos, você atende aos requisitos técnicos com o recurso de menor custo. As atividades de dimensionamento correto levam em conta todos os recursos de uma carga de trabalho, todos os atributos de cada recurso individual e o esforço envolvido na operação de dimensionamento correto. O dimensionamento correto pode ser um processo iterativo, acionado por alterações em padrões de uso e fatores externos, como quedas de preço da AWS ou novos tipos de recursos da AWS. O dimensionamento correto também pode ser único se o custo do esforço para dimensionar corretamente, ponderar as economias potenciais durante a vida útil da carga de trabalho.

Na AWS, há várias abordagens diferentes:

- Executar modelagem de custos
- Selecionar tamanho com base em métricas ou dados
- Selecionar tamanho automaticamente (com base em métricas)

Modelagem de custos: execute a modelagem de custos para sua carga de trabalho e cada um de seus componentes para entender o equilíbrio entre recursos e encontrar o tamanho correto para cada recurso na carga de trabalho, dado um nível específico de performance. Realize atividades de referência para a carga de trabalho sob diferentes cargas previstas e compare os custos. O esforço de modelagem deve refletir o benefício potencial. Por exemplo, o tempo gasto é proporcional ao custo do componente ou à economia prevista. Para obter as melhores práticas, consulte a seção Review do whitepaper Performance Efficiency Pillar of the AWS Well-Architected Framework.

O AWS Compute Optimizer pode ajudar na modelagem de custos para a execução de cargas de trabalho. Ele fornece recomendações de dimensionamento correto para recursos de computação com base no uso histórico. Essa é a fonte de dados ideal para recursos de computação, pois é um serviço gratuito e utiliza Machine Learning para fazer várias recomendações, dependendo dos níveis de risco. Você também pode usar o Amazon CloudWatch e o CloudWatch Logs com logs personalizados como fontes de dados para operações de dimensionamento correto para outros serviços e componentes de carga de trabalho.

Veja a seguir as recomendações para dados e métricas de modelagem de custo:

- O monitoramento deve refletir com precisão a experiência do usuário final. Selecione a granularidade correta para o período e escolha com cuidado o máximo ou o 99º percentil, em vez da média.
- Selecione a granularidade correta para o período de análise necessário para cobrir todos os ciclos de carga de trabalho. Por exemplo, se uma análise de duas semanas for realizada, talvez você esteja deixando passar um ciclo de alta utilização, o que pode levar a subprovisionamento.

Métricas ou seleção baseada em dados: selecione o tamanho ou o tipo do recurso com base na carga de trabalho e nas características do recurso; por exemplo, computação, memória, throughput ou gravação intensiva. Essa seleção geralmente é feita usando a modelagem de custo, uma versão anterior da carga de trabalho (como uma versão local), usando a documentação ou usando outras fontes de informações sobre a carga de trabalho (whitepapers, soluções publicadas).

Seleção automática baseada em métricas: crie um loop de comentários dentro da carga de trabalho que usa métricas ativas da carga de trabalho em execução para fazer alterações nessa carga de trabalho. Você pode usar um serviço gerenciado, como o AWS Auto Scaling, que você configura para executar as operações de dimensionamento certas para você. A AWS também fornece APIs, SDKs e recursos que permitem que os recursos sejam modificados com o mínimo de esforço. Você pode programar uma carga de trabalho para interromper e iniciar uma instância do EC2 para permitir uma alteração de tamanho ou tipo de instância. Isso fornece os benefícios do dimensionamento correto e, ao mesmo tempo, remove quase todo o custo operacional necessário para fazer a alteração.

Alguns serviços da AWS têm seleção automática de tipo ou tamanho, como S3 Intelligent-Tiering. O S3 Intelligent-Tiering move automaticamente seus dados entre dois níveis de acesso: acesso frequente e acesso infrequente, com base em seus padrões de uso.

Selecione o melhor modelo de definição de preço

Executar a modelagem de custo da carga de trabalho: considere os requisitos dos componentes da carga de trabalho e entenda os possíveis modelos de definição de preço. Defina o requisito de disponibilidade do componente. Determine se há vários recursos independentes que executam a função na carga de trabalho e quais são os requisitos da carga de trabalho ao longo do tempo. Compare o custo dos recursos usando o modelo de definição de preço sob demanda padrão e outros modelos aplicáveis. Leve em consideração possíveis alterações nos recursos ou componentes da carga de trabalho.

Executar análises regulares no nível da conta: a execução de uma modelagem de custo regular garante que as oportunidades de otimização em várias cargas de trabalho possam ser implementadas. Por exemplo, se várias cargas de trabalho usarem sob demanda, em um nível agregado, o risco de alteração será menor, e a implementação de um desconto baseado em compromisso atingirá um custo geral mais baixo. É recomendável realizar análises em ciclos regulares de duas semanas a um mês. Isso permite que você faça pequenas compras de ajuste, para que a cobertura de seus modelos de definição de preço continue a evoluir com suas cargas de trabalho dinâmicas e seus componentes.

Use a ferramenta de recomendações do AWS Cost Explorer para encontrar oportunidades de descontos de compromisso.

Para encontrar oportunidades para cargas de trabalho spot, use uma visualização por hora do uso geral e procure períodos regulares de uso ou elasticidade variáveis.

Modelos de definição de preço: a AWS tem vários modelos de definição de preço que permitem que você pague pelos seus recursos da maneira mais econômica que atenda às necessidades da sua organização. A seção a seguir descreve cada modelo de compra:

- Sob demanda
- Spot
- Descontos de compromisso - Savings Plans
- Descontos de compromisso - Instâncias reservadas/capacidade
- Seleção geográfica
- Acordos e definição de preço de terceiros

Sob demanda: este é o modelo de definição de preço padrão, com pagamento conforme o uso. Quando você usa recursos (por exemplo, instâncias do EC2 ou serviços como o DynamoDB sob demanda), você paga uma taxa fixa e não tem compromissos de longo prazo. Você pode aumentar ou diminuir a capacidade de seus recursos ou serviços com base nas demandas de seu aplicativo. Sob demanda tem uma taxa horária, mas, dependendo do serviço, pode ser cobrado em incrementos de 1 segundo (por exemplo, instâncias do AWS Lambda ou do EC2 do Linux). Sob demanda é recomendado para aplicativos com cargas de trabalho de curto prazo (por exemplo, um projeto de quatro meses), com picos periódicos ou cargas de trabalho imprevisíveis que não podem ser interrompidas. Sob demanda também é adequado para cargas de trabalho, como ambientes de pré-produção, que exigem tempos de execução ininterruptos, mas não são executados por tempo suficiente para um desconto de compromisso (Savings Plans ou instâncias reservadas).

Spot: uma instância spot é uma capacidade computacional extra do EC2 disponível com descontos de até 90% em preços sob demanda, sem necessidade de compromissos de longo prazo. Com as instâncias spot, você pode reduzir significativamente o custo de execução dos aplicativos ou escalar a capacidade computacional do aplicativo para o mesmo orçamento. Ao contrário das instâncias sob demanda, as instâncias spot poderão ser interrompidas com um aviso de 2 minutos se o EC2 precisar da capacidade de volta ou se o preço da instância spot exceder o preço configurado. Em média, as instâncias spot são interrompidas em menos de 5% do tempo.

Spot é ideal quando há uma fila ou buffer implementado ou quando há vários recursos trabalhando de forma independente para processar as solicitações (por exemplo, processamento de dados do Hadoop). Normalmente, essas cargas de trabalho são tolerantes a falhas, sem estado e flexíveis, como processamento em lotes, big data e análises, ambientes conteinerizados e computação de alta performance (HPC). Cargas de trabalho não críticas, como ambientes de teste e desenvolvimento, também são candidatas ao spot.

O spot também é integrado a vários serviços da AWS, como grupos de Auto Scaling do EC2 (ASGs), Elastic MapReduce (EMR), Elastic Container Service (ECS) e AWS Batch.

Quando uma instância spot precisa ser recuperada, o EC2 envia um aviso de dois minutos por meio de um aviso de interrupção de instância spot entregue por meio do CloudWatch Events, bem como nos metadados da instância. Durante esse período de dois minutos, seu aplicativo pode usar o tempo para salvar seu estado, drenar contêineres em execução, fazer upload de arquivos de log finais ou se remover de um load balancer. Ao final dos dois minutos, você tem a opção de hibernar, interromper ou encerrar a instância spot.

Considere as seguintes melhores práticas ao adotar instâncias spot em suas cargas de trabalho:

- **Definir o preço máximo como a taxa sob demanda:** isso garante que você pagará a taxa spot atual (o preço mais barato disponível) e nunca pagará mais do que a taxa sob demanda. As taxas atuais e históricas estão disponíveis por meio do console e da API.
- **Ser flexível no maior número possível de tipos de instância:** seja flexível na família e no tamanho do tipo de instância, para melhorar a probabilidade de atender aos requisitos de capacidade pretendidos, obter o menor custo possível e minimizar o impacto das interrupções.
- **Ser flexível sobre onde sua carga de trabalho será executada:** a capacidade disponível pode variar de acordo com a zona de disponibilidade. Isso melhora a probabilidade de atender à capacidade pretendida ao tocar em vários grupos de capacidade sobressalente e fornece o menor custo possível.
- **Projetar para continuidade:** projete suas cargas de trabalho serem do tipo sem estado e tolerante a falhas, para que, se parte da capacidade do EC2 for interrompida, isso não afete a disponibilidade ou a performance da carga de trabalho.
- Recomendamos o uso de instâncias spot em combinação com planos sob demanda e Savings Plans/instâncias reservadas para maximizar a otimização de custos da carga de trabalho com a performance.

Descontos de compromisso - Savings Plans: a AWS fornece várias maneiras de reduzir seus custos reservando ou comprometendo-se a usar uma determinada quantidade de recursos e recebendo uma taxa com desconto para seus recursos. Um Savings Plan permite que você faça um compromisso de gastos por hora por um ou três anos e receba preços com desconto em todos os seus recursos. Os Savings Plans oferecem descontos para serviços de computação da AWS, como EC2, Fargate e Lambda. Ao fazer o compromisso, você paga esse valor de compromisso a cada hora, e ele é subtraído do uso sob demanda com a taxa de desconto. Por exemplo, você se compromete com 50 USD por hora e tem 150 USD por hora de uso sob demanda. Considerando a definição de preço dos Savings Plans, seu uso específico tem uma taxa de desconto de 50%. Portanto, seu compromisso de 50 USD cobre 100 USD de uso sob demanda. Você pagará 50 USD (compromisso) e 50 USD de uso sob demanda restante.

Os Compute Savings Plans são os mais flexíveis e oferecem um desconto de até 66%. Eles se aplicam automaticamente em zonas de disponibilidade, tamanho de instância, família de instâncias, sistema operacional, locação, região e serviço de computação.

Os Instance Savings Plans têm menos flexibilidade, mas fornecem uma taxa de desconto mais alta (até 72%). Eles se aplicam automaticamente em zonas de disponibilidade, tamanho de instância, família de instâncias, sistema operacional e locação.

Existem três opções de pagamento:

- **Sem pagamento adiantado:** não há pagamento adiantado; você paga uma taxa horária reduzida a cada mês para o total de horas do mês.
- **Pagamento adiantado parcial:** fornece uma taxa de desconto mais alta do que Sem pagamento adiantado. Parte do uso é paga antecipadamente. Em seguida, você paga uma taxa horária reduzida menor a cada mês referente ao total de horas do mês.
- **Pagamento adiantado integral:** o uso de todo o período é pago antecipadamente, e nenhum outro custo é incorrido durante o restante do período de vigência pelo uso coberto pelo compromisso.

Você pode aplicar qualquer combinação dessas três opções de compra em suas cargas de trabalho.

Os Savings Plans se aplicam primeiro ao uso na conta em que foram comprados, da porcentagem de desconto mais alta para a mais baixa e, em seguida, ao uso consolidado em todas as outras contas, da porcentagem de desconto mais alta para a mais baixa.

É recomendável comprar todos os Savings Plans em uma conta sem uso ou recursos, como a conta mestre. Isso garante que o Savings Plan se aplique às taxas de desconto mais altas em todo o seu uso, maximizando o valor do desconto.

As cargas de trabalho e o uso normalmente mudam com o passar do tempo. É recomendável adquirir continuamente pequenas quantidades de compromissos com Savings Plans ao longo do tempo. Isso garante que você mantenha altos níveis de cobertura para maximizar seus descontos, e seus planos sempre atendam aos requisitos de carga de trabalho e organização.

Não defina uma meta de cobertura em suas contas, devido à variação do desconto que é possível. A baixa cobertura não indica necessariamente um alto potencial de economia. Você pode ter uma baixa cobertura em sua conta, mas se seu uso for composto de instâncias pequenas, com um sistema operacional licenciado, a economia potencial poderá ser tão baixa quanto alguns %. Em vez disso, acompanhe e monitore as possíveis economias disponíveis na ferramenta de recomendação do Savings Plan. Analise com frequência as recomendações dos Savings Plans no Cost Explorer (execute análises regulares) e continue a comprar compromissos até que as economias estimadas estejam abaixo do desconto necessário para a organização. Por exemplo, acompanhe e monitore se seus possíveis descontos permaneceram abaixo de 20%, se for além de que uma compra deve ser feita.

Monitore a utilização e a cobertura, mas apenas para detectar alterações. Não aponte para uma porcentagem de utilização específica ou porcentagem de cobertura, pois isso não necessariamente escala com economias. Certifique-se de que uma compra de Savings Plans resulte em um aumento na cobertura e, se houver diminuição na cobertura ou utilização, garanta que eles sejam quantificados e conhecidos. Por exemplo, você migra um recurso de carga de trabalho para um tipo de instância mais recente, o que reduz a utilização de um plano existente, mas o benefício de performance supera a redução de economia.

Descontos de compromisso - Instâncias reservadas/compromisso: de forma semelhante aos Savings Plans, as instâncias reservadas oferecem descontos de até 72% para um compromisso de executar uma quantidade mínima de recursos. As instâncias reservadas estão disponíveis para RDS, Elasticsearch, ElastiCache, Amazon Redshift e DynamoDB. O Amazon CloudFront e o AWS Elemental MediaConvert também oferecem descontos quando você faz compromissos de uso mínimo. No momento, as instâncias reservadas estão disponíveis para o EC2. No entanto, os Savings Plans oferecem os mesmos níveis de desconto com maior flexibilidade e sem sobrecarga de gerenciamento.

As instâncias reservadas oferecem as mesmas opções de definição de preço, sem adiantamento, pagamento adiantado parcial e pagamento adiantado, e os mesmos períodos de vigência de um ou três anos.

As instâncias reservadas podem ser adquiridas em uma região ou em uma zona de disponibilidade específica. Elas fornecem uma reserva de capacidade quando compradas em uma zona de disponibilidade.

O EC2 oferece RIs conversíveis, no entanto, Savings Plans devem ser usados para todas as instâncias do EC2 devido à maior flexibilidade e redução dos custos operacionais.

O mesmo processo e métricas devem ser usados para rastrear e fazer compras de instâncias reservadas. É recomendável não rastrear a cobertura de RI em todas as suas contas. Também é recomendável que a % de utilização não seja monitorada ou rastreada. Em vez disso, visualize o relatório de utilização no Cost Explorer e use a coluna de economia líquida na tabela. Se a economia líquida for um valor negativo significativamente grande, você deverá tomar medidas para corrigir a RI não utilizada.

Frota do EC2: a Frota do EC2 é um recurso que permite definir uma capacidade computacional de destino e, em seguida, especificar os tipos de instância e o equilíbrio de instâncias sob demanda e spot para a frota. A Frota do EC2 executará automaticamente a combinação de recursos de menor preço para atender à capacidade definida.

Seleção geográfica: quando você arquiteta suas soluções, uma melhor prática é buscar colocar recursos computacionais mais perto dos usuários para fornecer menor latência e uma sólida soberania de dados. Para públicos globais, você deve usar vários locais para atender a essas necessidades. Você deve selecionar a localização geográfica que minimiza seus custos.

A infraestrutura da Nuvem AWS é criada em torno de regiões e zonas de disponibilidade. Região é um local físico do mundo onde há várias zonas de disponibilidade. As zonas de disponibilidade consistem em um ou mais datacenters separados, cada um com energia, rede e conectividade redundantes, alojados em instalações distintas.

Cada região da AWS opera dentro das condições do mercado local, e a definição de preço dos recursos é diferente em cada região. Escolha uma região específica para operar um componente de sua solução completa para que você possa operar ao menor preço possível globalmente. Você pode usar a Calculadora Mensal da AWS para estimar os custos da carga de trabalho em várias regiões.

Acordos e definição de preço de terceiros: quando você utiliza soluções ou serviços de terceiros na nuvem, é importante que as estruturas de definição de preço estejam alinhadas aos resultados da otimização de custos. A definição de preço deve ser dimensionada de acordo com os resultados e o valor que fornece. Um exemplo disso é um software que leva uma porcentagem das economias que ele fornece, quanto mais você economiza (resultado), mais ele cobra. Contratos que escalam com sua fatura normalmente não estão alinhados com a otimização de custos, a menos que forneçam resultados para cada parte da sua fatura específica. Por exemplo, uma solução que fornece recomendações para o EC2 e cobra uma porcentagem de toda a sua fatura aumentará se você usar outros serviços para os quais ela não oferece nenhum benefício. Outro exemplo é um serviço gerenciado que é cobrado a uma porcentagem do custo dos recursos que são gerenciados. Um tamanho de instância maior pode não exigir necessariamente mais esforço de gerenciamento, mas será cobrado mais. Certifique-se de que essas disposições de definição de preço de serviços incluam um programa de otimização de custos ou recursos em seu serviço para promover a eficiência.

Planejar a transferência de dados

Uma vantagem da nuvem é que ela é um serviço de rede gerenciado. Não há mais a necessidade de gerenciar e operar uma frota de switches, roteadores e outros equipamentos de rede associados. Os recursos de rede na nuvem são consumidos e pagos da mesma forma que você paga pela CPU e pelo armazenamento - você paga apenas pelo que usa. O uso eficiente de recursos de rede é necessário para otimização de custos na nuvem.

Executar modelagem de transferência de dados: entenda onde a transferência de dados ocorre na carga de trabalho, o custo da transferência e o benefício associado. Isso permite que você tome uma decisão embasada para modificar ou aceitar a decisão arquitetônica. Por exemplo, você pode ter uma configuração de várias zonas de disponibilidade na qual replicar dados entre as zonas de disponibilidade. Você modela o custo da estrutura e decide que esse é um custo aceitável (semelhante ao pagamento por computação e armazenamento em ambas as zonas de disponibilidade) para alcançar a confiabilidade e a resiliência necessárias.

Modele os custos em diferentes níveis de uso. O uso da carga de trabalho pode mudar ao longo do tempo, e diferentes serviços podem ser mais econômicos em diferentes níveis.

Use o AWS Cost Explorer ou o Cost and Usage Report (CUR - Relatório de custo e uso) para compreender e modelar seus custos de transferência de dados. Configure uma prova de conceito (PoC) ou teste sua carga de trabalho e execute um teste com uma carga simulada realista. Você pode modelar seus custos em diferentes demandas de carga de trabalho.

Otimizar a transferência de dados: a arquitetura para transferência de dados garante que você minimize os custos de transferência de dados. Isso pode envolver usar redes de entrega de conteúdo para colocar os dados mais perto dos usuários ou usar links de rede dedicados de seu local para a AWS. Você também pode usar a otimização de WAN e a otimização de aplicativos para reduzir a quantidade de dados transferidos entre componentes.

Selecionar serviços para reduzir custos de transferência de dados: o Amazon CloudFront é uma rede global de entrega de conteúdo que entrega dados com baixa latência e altas velocidades de transferência. Ele armazena dados em cache em pontos de presença no mundo inteiro, o que reduz a carga sobre seus recursos. Ao usar o CloudFront, você pode reduzir o trabalho administrativo para entregar conteúdo a grandes números de usuários globalmente com latência mínima.

O AWS Direct Connect permite estabelecer uma conexão de rede dedicada com a AWS. Isso pode reduzir os custos de rede, aumentar a largura de banda e fornecer uma experiência de rede mais consistente do que conexões baseadas em Internet.

A VPN da AWS permite estabelecer uma conexão segura e privada entre sua rede privada e a rede global da AWS. Ele é ideal para pequenos escritórios ou parceiros de negócios porque oferece conectividade rápida e fácil, além de ser um serviço totalmente gerenciado e elástico.

Os VPC endpoints permitem conectividade entre os serviços da AWS em redes privadas e podem ser usados para reduzir os custos de transferência de dados pública e gateways NAT. Os VPC endpoints do gateway não têm cobranças por hora e oferecem suporte ao Amazon S3 e ao Amazon DynamoDB. Os VPC endpoints de interface são fornecidos pelo AWS PrivateLink e têm uma taxa horária e custo de uso por GB.

Recursos

Consulte os recursos a seguir para saber mais sobre as melhores práticas da AWS para recursos econômicos.

- AWS Managed Services: Enterprise Transformation Journey Video
- Análise de custos com o Cost Explorer
- Acesso a recomendações de instância reservada
- Conceitos básicos das recomendações de dimensionamento correto
- Melhores práticas de instâncias spot

- Frotas spot
- Como funcionam as instâncias reservadas
- Infraestrutura global da AWS
- Consultor de instância spot
- Well-Architected Labs - Recursos econômicos

Gerenciar recursos de demanda e oferta

Quando você passa para a nuvem, paga apenas pelo que precisa. Você pode fornecer recursos para atender à demanda da carga de trabalho no momento em que eles são necessários, eliminando a necessidade de provisionamento em excesso dispendioso e desperdiçador. Você também pode modificar a demanda usando um controle de utilização, um buffer ou uma fila para suavizar a demanda e atendê-la com menos recursos.

Os benefícios econômicos da oferta just-in-time devem ser equilibrados em relação à necessidade de provisionar para compensar falhas de recursos, alta disponibilidade e tempo de provisionamento. Dependendo de sua demanda (fixa ou variável), planeje criar métricas e automação que garantam que o gerenciamento de seu ambiente seja mínimo, mesmo conforme você ajusta a escala. Ao modificar a demanda, você deve saber o atraso aceitável e máximo que a carga de trabalho pode permitir.

Na AWS, você pode usar várias abordagens diferentes para gerenciar a demanda e fornecer recursos. As seções a seguir descrevem como usar essas abordagens:

- Analisar a carga de trabalho
- Gerenciar demanda
- Oferta baseada em demanda
- Oferta baseada em tempo

Analisar a carga de trabalho: conheça os requisitos da carga de trabalho. Os requisitos da organização devem indicar os tempos de resposta da carga de trabalho para solicitações. O tempo de resposta pode ser usado para determinar se a demanda é gerenciada ou se a oferta de recursos será alterada para atender à demanda.

A análise deve incluir a previsibilidade e a repetibilidade da demanda, a taxa de alteração na demanda e a quantidade de alteração na demanda. Certifique-se de que a análise seja realizada durante um período longo o suficiente para incorporar qualquer variação sazonal, como processamento de fim de mês ou picos de fim de ano.

Certifique-se de que o esforço de análise reflita os possíveis benefícios da implementação da escalabilidade. Observe o custo total esperado do componente e quaisquer aumentos ou diminuições no uso e no custo durante a vida útil da carga de trabalho.

Você pode usar o AWS Cost Explorer ou o Amazon QuickSight com o CUR ou os logs do aplicativo para executar uma análise visual da demanda da carga de trabalho.

Gerenciar demanda

Gerenciar demanda - controle de utilização: se a origem da demanda tiver capacidade de repetição, você poderá implementar o controle de utilização. O controle de utilização informa à origem que, se ela não puder atender à solicitação no momento atual, deverá tentar novamente mais tarde. A origem aguardará um período e, em seguida, tentará novamente a solicitação. A implementação do controle de utilização tem a vantagem de limitar a quantidade máxima de recursos e custos da carga de trabalho. Na AWS, você pode usar o Amazon API Gateway para implementar o controle de utilização. Consulte o whitepaper Pilar Confiabilidade do Well-Architected para obter mais detalhes sobre a implementação do controle de utilização.

Gerenciar demanda - baseada em buffer: semelhante ao controle de utilização, um buffer adia o processamento de solicitações, permitindo que aplicativos executados em diferentes taxas se comuniquem com eficácia. Uma abordagem baseada em buffer usa uma fila para aceitar mensagens (unidades de trabalho) de produtores. As mensagens são lidas pelos consumidores e processadas, permitindo que as mensagens sejam executadas na taxa que atenda aos requisitos de negócios dos consumidores. Você não precisa se preocupar com os produtores que precisam lidar com problemas de controle de utilização, como durabilidade de dados e pressão contrária (onde os produtores ficam lentos porque o consumidor está correndo lentamente).

Na AWS, você pode escolher entre vários serviços para implementar uma abordagem de buffering. O Amazon SQS é um serviço gerenciado que fornece filas que permitem que um único consumidor leia mensagens individuais. O Amazon Kinesis oferece um stream que permite a muitos consumidores lerem as mesmas mensagens.

Ao criar uma arquitetura com uma abordagem baseada em buffer, certifique-se de arquitetar sua carga de trabalho para atender à solicitação no tempo necessário e de lidar com solicitações duplicadas de trabalho.

Dynamic Supply

Oferta baseada em demanda: aproveite a elasticidade da nuvem para fornecer recursos para atender à demanda em constante mudança. Aproveite as APIs ou os recursos de serviço para variar programaticamente a quantidade de recursos de nuvem em sua arquitetura dinamicamente. Isso permite que você ajuste a escala de componentes em sua arquitetura e aumente automaticamente o número de recursos durante picos de demanda para manter a performance e reduzir a capacidade quando a demanda diminui para reduzir os custos.

O Auto Scaling ajuda você a ajustar sua capacidade para manter uma performance estável e previsível pelo menor custo possível. É um serviço totalmente gerenciado e gratuito que se integra às instâncias do Amazon EC2 e às frotas spot, ao Amazon ECS, ao Amazon DynamoDB e ao Amazon Aurora.

O Auto Scaling oferece descoberta automática de recursos para ajudar a encontrar recursos na sua carga de trabalho que possam ser configurados, tem estratégias de escalabilidade incorporadas para otimizar performance, custos ou um equilíbrio entre os dois, além de oferecer escalabilidade preditiva para ajudar com picos que ocorrem regularmente.

O Auto Scaling pode implementar escalabilidade manual, programada ou baseada em demanda. Você também pode usar métricas e alarmes do Amazon CloudWatch para acionar eventos de escalabilidade para sua carga de trabalho. As métricas típicas podem ser métricas padrão do Amazon EC2, como utilização de CPU, throughput de rede e latência de solicitação/resposta observada pelo ELB. Quando possível, você deve usar uma métrica que seja indicativa da experiência do cliente, normalmente essa é uma métrica personalizada que pode se originar do código do aplicativo em sua carga de trabalho.

Ao arquitetar com uma abordagem baseada em demanda, tenha em mente dois pontos essenciais. Primeiro, entenda a rapidez com que você deve provisionar novos recursos. Segundo, entenda que o tamanho da margem entre oferta e demanda mudará. Você deve estar pronto para lidar com a taxa de alteração na demanda e também estar pronto para falhas de recursos.

O Elastic Load Balancing (ELB) ajuda você a escalar distribuindo a demanda entre vários recursos. À medida que você implementa mais recursos, você os adiciona ao load balancer para atender à demanda. O AWS ELB tem suporte para instâncias do EC2, contêineres, endereços IP e funções Lambda.

Oferta baseada em tempo: uma abordagem baseada em tempo alinha a capacidade de recursos à demanda previsível ou bem definida por tempo. Essa abordagem costuma não depender dos níveis de utilização dos recursos. Uma abordagem baseada em tempo garante que os recursos estejam disponíveis no momento específico em que são necessários e podem ser fornecidos sem nenhum atraso devido a procedimentos de inicialização e verificações do sistema ou de consistência. Usando uma abordagem baseada em tempo, você pode fornecer recursos adicionais ou aumentar a capacidade durante períodos ocupados.

Você pode usar o Auto Scaling programado para implementar uma abordagem baseada em tempo. As cargas de trabalho podem ser programadas para expandir ou reduzir em horários definidos (por exemplo, o início do horário comercial), garantindo assim que os recursos estejam disponíveis quando os usuários ou a demanda chegarem.

Você também pode aproveitar as APIs e os SDKs da AWS e o AWS CloudFormation para provisionar e desativar automaticamente ambientes inteiros conforme necessário. Essa abordagem é adequada para ambientes de desenvolvimento ou teste que são executados apenas nos períodos ou horários comerciais definidos.

Você pode usar APIs para ajustar a escala dos recursos dentro de um ambiente (ajuste de escala vertical). Por exemplo, você pode escalar uma carga de trabalho de produção alterando o tamanho ou a classe da instância. Isso pode ser feito interrompendo e iniciando a instância e selecionando a classe ou o tamanho da instância diferente. Essa técnica também pode ser aplicada a outros recursos, como Volumes elásticos do EBS, que podem ser modificados para aumentar o tamanho, ajustar a performance (IOPS) ou alterar o tipo de volume durante o uso.

Ao arquitetar com uma abordagem baseada em tempo, tenha em mente dois pontos essenciais. Primeiro, qual é a consistência do padrão de uso? Segundo, qual será o impacto se o padrão mudar? Você pode aumentar a precisão das previsões monitorando suas cargas de trabalho e usando inteligência de negócios. Se você vir alterações significativas no padrão de uso, poderá ajustar os tempos para garantir que a cobertura seja fornecida.

Oferta dinâmica: você pode usar o AWS Auto Scaling, ou incorporar escalabilidade no código com a API ou SDKs da AWS. Isso reduz os custos gerais da carga de trabalho removendo o custo operacional de fazer alterações manualmente em seu ambiente e pode ser executado muito mais rapidamente. Isso garantirá que o recurso da carga de trabalho corresponda melhor à demanda a qualquer momento.

Recursos

Consulte os recursos a seguir para saber mais sobre as melhores práticas da AWS para gerenciar a demanda e fornecer recursos.

- Controle de utilização do API Gateway
- Conceitos básicos do Amazon SQS
- Conceitos básicos do Amazon EC2 Auto Scaling

Otimizar ao longo do tempo

Na AWS, você otimiza ao longo do tempo analisando novos serviços e implementando-os em sua carga de trabalho.

Analise e implemente novos serviços

À medida que a AWS lança novos serviços e recursos, é uma melhor prática analisar suas decisões de arquitetura atuais para garantir que elas permaneçam econômicas. Conforme seus requisitos mudam, seja agressivo na desativação de recursos, componentes e cargas de trabalho de que não precisa mais. Considere o seguinte para ajudá-lo a otimizar ao longo do tempo:

- Desenvolver um processo de análise da carga de trabalho
- Analisar e implementar serviços

Desenvolver um processo de análise de carga de trabalho: para garantir que você sempre tenha a carga de trabalho mais econômica, você deve revisar regularmente a carga de trabalho para saber se há oportunidades de implementar novos serviços, recursos e componentes. Para garantir que você atinja custos gerais mais baixos, o processo deve ser proporcional à quantidade potencial de economia. Por exemplo, as cargas de trabalho que representam 50% do seu gasto geral devem ser analisadas com mais frequência e mais precisão do que as cargas de trabalho que representam 5% do seu gasto geral. Leve em consideração quaisquer fatores externos ou volatilidade. Se a carga de trabalho atender a uma área geográfica ou segmento de mercado específico e houver previsão de mudanças nessa área, revisões mais frequentes poderão resultar em economias de custos. Outro fator em análise é o esforço para implementar alterações. Se houver custos significativos em testes e validação de alterações, as revisões devem ser menos frequentes.

Leve em consideração o custo de longo prazo da manutenção de componentes e recursos obsoletos e na incapacidade de implementar novos recursos neles. O custo atual de testes e validação pode exceder o benefício proposto. No entanto, ao longo do tempo, o custo de fazer a mudança pode aumentar significativamente à medida que a lacuna entre a carga de trabalho e as tecnologias atuais aumenta, resultando em custos ainda maiores. Por exemplo, o custo da migração para uma nova linguagem de programação pode não ser econômico no momento. No entanto, em cinco anos, o custo de pessoas com qualificações nessa linguagem pode aumentar e, devido ao crescimento da carga de trabalho, você estaria movendo um sistema ainda maior para a nova linguagem, exigindo ainda mais esforço do que anteriormente.

Divida sua carga de trabalho em componentes, atribua o custo do componente (uma estimativa é suficiente) e liste os fatores (por exemplo, esforço e mercados externos) ao lado de cada componente. Use esses indicadores para determinar uma frequência de revisão para cada carga de trabalho. Por exemplo, você pode ter servidores web como um alto custo, baixo esforço de alteração e altos fatores externos, resultando em alta frequência de revisão. Um banco de dados central pode ser de custo médio, alto esforço de alteração e baixos fatores externos, resultando em uma média frequência de análise.

Analisar a carga de trabalho e implementar serviços: para obter os benefícios de novos serviços e recursos da AWS, você deve executar o processo de análise em suas cargas de trabalho e implementar novos serviços e recursos, conforme necessário. Por exemplo, você pode revisar suas cargas de trabalho e substituir o componente de mensagens pelo Amazon Simple Email Service (SES). Isso remove o custo de operação e manutenção de uma frota de instâncias e, ao mesmo tempo, fornece toda a funcionalidade a um custo reduzido.

Conclusão

A otimização de custos e o gerenciamento financeiro na nuvem são um esforço contínuo. Você deve trabalhar regularmente com suas equipes de finanças e tecnologia, analisar sua abordagem arquitetônica e atualizar sua seleção de componentes.

A AWS ajuda você a minimizar o custo enquanto cria implantações altamente resilientes, responsivas e adaptáveis. Para realmente otimizar o custo de sua implantação, aproveite as ferramentas, as técnicas e as melhores práticas discutidas neste documento.

Colaboradores

Os colaboradores desse documento incluem:

- Philip Fitzsimons, gerente sênior de Well-Architected, Amazon Web Services
- Nathan Besh, líder de custo de Well-Architected, Amazon Web Services
- Levon Stepanian, Amazon Web Services
- Keith Jarrett, líder de desenvolvimento de negócios – otimização de custos
- PT Ng, arquiteto comercial, Amazon Web Services
- Arthur Basbaum, gerente desenvolvedor de negócios, Amazon Web Services
- Jarman Hauser, arquiteto comercial, Amazon Web Services

Leitura adicional

Para obter informações adicionais, consulte:

- AWS Well-Architected Framework

Revisões do documento

Data	Descrição
Abril de 2020	Atualizado para incorporar CFM, novos serviços e integração com o Well-Architected também.
Julho de 2018	Atualizado para refletir alterações à AWS e incorporar aprendizados de análises com clientes.
Novembro de 2017	Atualizado para refletir alterações à AWS e incorporar aprendizados de análises com clientes.
Novembro de 2016	Primeira publicação