Opções de conectividade de rede ativadas AWS para ofertas de SaaS

AWS Recomendações

AWS Recomendações: Opções de conectividade de rede ativadas AWS para ofertas de SaaS

Table of Contents

Opções de conectividade de rede ativadas AWS para ofertas de SaaS

Tomas Sykora e Luca Schumann, Amazon Web Services

Setembro de 2025 (histórico do documento)

Este guia explora cenários comuns para conectar aplicativos de consumo a provedores de software como serviço (SaaS). Ele discute como se conectar a recursos que estão no local Nuvem AWS, em outras nuvens de provedores de serviços de nuvem (CSP) ou em arquiteturas híbridas. Esses cenários incluem o seguinte:

- Expondo serviços da web por HTTPS
- Expondo serviços baseados em TCP
- Usando AWS AppSyncpara implementar publish-subscribe (Pub/Sub) e GraphQL APIs
- Usando AWS recursos para expor WebSockets para aplicativos em tempo real
- Habilitando o acesso bidirecional para comunicação interativa de serviços

Ao se alinharem às melhores práticas abordadas neste guia, os provedores de SaaS podem aumentar a confiança do cliente e oferecer suporte ao acesso escalável, seguro e resiliente às ofertas de SaaS.

Este guia também inclui critérios de autoavaliação para ajudá-lo a avaliar com que sucesso você está atendendo aos requisitos de rede de consumidores para sua oferta de SaaS. Além dos padrões de conectividade, você encontrará comparações abrangentes de serviços de AWS rede, diagramas de arquitetura de alto nível para vários cenários de implantação e orientações práticas sobre como selecionar a abordagem certa com base em seu contexto comercial específico. O guia explora as considerações de segurança para cada opção de rede, discute as armadilhas comuns a serem evitadas e fornece recomendações de implementação que equilibram os requisitos técnicos com a eficiência operacional. Além disso, você encontrará estruturas estratégicas para alinhar suas decisões de rede com seu modelo de negócios, objetivos de crescimento e necessidades de conformidade regulatória.

Público-alvo

Este guia é destinado a provedores de SaaS. Ele ajuda arquitetos de nuvem, gerentes de produto e engenheiros de rede que estão projetando, implementando e otimizando a conectividade de rede para ofertas de SaaS no. Nuvem AWS Para entender os conceitos e recomendações deste guia, você deve estar familiarizado com AWS os fundamentos, os principais conceitos de SaaS e os princípios de rede de alto nível.

Objetivos

Este guia discute as opções de arquitetura de rede e as melhores práticas testadas em campo que ajudam os consumidores a otimizar o acesso às ofertas de SaaS. A implementação das recomendações deste guia oferece suporte ao seguinte:

- Facilidade de integração — ofereça uma jornada descomplicada ao cliente, desde a integração até a produção, para que você possa acelerar o tempo de geração de valor de seus clientes e encurtar o ciclo de reconhecimento de receita.
- Adaptabilidade — Integre-se perfeitamente às infraestruturas de rede existentes de seus clientes, adaptando-se às necessidades em evolução. Isso aprimora a proposta de valor do seu produto.
- Custo total de propriedade — Padronize o acesso à rede para reduzir os custos de mudança e os custos por inquilino. Ao melhorar a consistência da implantação, você também pode reduzir o tempo para realizar a análise ou o reparo da causa raiz.
- Gerenciamento de dependências — entenda as dependências, as implicações de longo prazo e as vantagens e desvantagens das diferentes opções de acesso à rede. Isso ajuda os líderes de produto a tomarem decisões de produto bem informadas.
- Composição e extensibilidade — dissocie o desenvolvimento da funcionalidade principal da infraestrutura operacional. Isso ajuda as equipes de desenvolvimento a se moverem mais rapidamente e se concentrarem na criação de valor para seus clientes.
- Promova a confiança — Ao fornecer acesso resiliente, tolerante a falhas, seguro e escalável às ofertas de SaaS, você pode reduzir os riscos regulatórios e ganhar confiança em sua capacidade de apoiar o crescimento de seus clientes.

Avaliando as decisões de acesso à rede para ofertas de SaaS

Entendendo seu mercado

As decisões que você toma agora sobre redes determinam se a proposta de valor do seu produto SaaS pode ser entregue aos seus clientes. Apesar da importância estratégica dessas decisões, fornecer acesso à sua oferta de SaaS geralmente é percebido como um tópico puramente tecnológico. O risco que essa percepção acarreta inclui ciclos prolongados de reconhecimento de receita, ineficiências operacionais e desalinhamento com a estratégia de negócios. Por exemplo, se a rápida expansão é um objetivo estratégico de negócios, uma luz orientadora do seu processo de tomada de decisão deve ser se as soluções que você está considerando são escaláveis e flexíveis o suficiente para suportar a expansão. Mesmo que você tenha sucesso no crescimento de seus negócios, a sobrecarga operacional não deve se tornar um obstáculo para o crescimento futuro, e uma estrutura de custos desalinhada pode consumir todos os seus lucros.

Por exemplo, considere como as seguintes considerações de mercado afetam os aspectos técnicos do produto, como a rede:

- Se seu modelo de negócios for baseado em assinaturas, é provável que seus clientes prefiram soluções com custos previsíveis e recorrentes em vez de grandes investimentos iniciais.
- Se sua estratégia de negócios tem como alvo clientes corporativos de alto valor, os critérios de segurança, governança e conformidade regulatória determinam se sua oferta de SaaS será mesmo considerada.
- Se seu mercado-alvo é composto principalmente por startups, a facilidade de integração, o tempo de valorização e a adaptabilidade provavelmente são fatores importantes. As startups normalmente priorizam a velocidade e a agilidade. Como precisam criar uma marca e gerar lucros rapidamente, é provável que prefiram soluções que sejam rápidas e fáceis de integrar, que possam ser escaladas de maneira econômica, reduzir a dependência de especialistas e que não restrinjam ciclos preciosos.
- Algumas empresas exigem acesso estável, de alto rendimento e baixa latência. Isso inclui o setor de entretenimento e mídia, manufatura e processamento de transações financeiras. Se esses são seus clientes-alvo, a confiabilidade é sua principal preocupação.

Em todos esses casos, os clientes podem perceber uma oferta de SaaS saudável se o acesso à rede não for perfeito. Se a rede se tornar um obstáculo, isso não dá suporte ao seu caso de negócios. Se seus clientes não conseguem acessar de forma confiável os serviços que você oferece, a proposta de valor de suas ofertas de SaaS é nula.

Entendendo sua função

Seu papel no apoio aos objetivos de negócios depende de quem você é, de quais são seus objetivos individuais e de equipe específicos, de quem são seus clientes e do que é importante para eles. Mesmo que você não faça parte de uma equipe que normalmente interage com os clientes, precisa se preocupar com quem eles são e com o que precisam. As equipes de engenharia e desenvolvimento também devem se preocupar com seus clientes internos, especialmente aqueles com quem interagem regularmente. Normalmente, essas são as equipes de operações e de sucesso do cliente.

Se você faz parte de uma organização de vendas, é essencial que você se comunique com as equipes de produto e engenharia sobre redes, mesmo que seja um tópico aparentemente puro de tecnologia. Compartilhe ideias sobre a estrutura do mercado-alvo. Comunique os pontos problemáticos e as necessidades de seus clientes e parceiros atuais e potenciais. Compartilhe dados e histórias sobre oportunidades perdidas, crescimento previsto por segmento e eventos. Faça perguntas que desafiem a capacidade da sua organização de apoiar o crescimento dos negócios. Isso aumenta o número de oportunidades e melhora a lucratividade de longo prazo do seu negócio. Em última análise, isso ajuda sua organização a financiar a expansão e o desenvolvimento futuros.

Se você faz parte da organização de engenharia, entenda a estratégia comercial da sua organização antes de tentar elaborar uma solução. O alinhamento com a estratégia de negócios ajuda você a escolher as métricas certas para avaliar as diferentes opções de acesso à rede. Também pode evitar um redesenho de rede caro e em grande escala à medida que sua organização cresce. O alinhamento de negócios ajuda sua equipe a proteger e reter os recursos necessários para futuros desafios. O número de funcionários, o orçamento para desenvolvimento profissional ou o acesso à tecnologia de ponta da sua equipe dependerão da sua capacidade de demonstrar o alinhamento comercial. Idealmente, você pode mostrar como suas decisões contribuíram para o sucesso comercial da organização. Portanto, sugerimos que você capture o processo de tomada de decisão, incluindo os critérios de seleção de métricas. Analise periodicamente suas métricas para confirmar se elas estão alinhadas aos objetivos de negócios. Isso pode ajudar sua equipe a obter o crédito que merece. As revisões periódicas também ajudam a validar que sua equipe não está tomando decisões com base em suposições ou motivos históricos obsoletos.

A lista de métricas nas seções a seguir é relevante para o acesso à rede:

- Métricas comerciais e de produto
- Métricas de engenharia que influenciam as decisões de rede

Este guia usa um subconjunto dessas métricas para ajudá-lo a identificar as abordagens ideais de acesso à rede para suas ofertas de SaaS. Escolha as métricas mais importantes e relevantes para sua empresa e, em seguida, avalie as abordagens com base nessas métricas.

Métricas comerciais e de produtos que influenciam as decisões de rede

As equipes comerciais e de produtos usam critérios de sucesso para avaliar se estão atingindo os objetivos de negócios. Esta seção descreve as métricas comerciais ou de produtos que podem ser influenciadas positiva ou negativamente pelas decisões de acesso à rede que sua organização toma.

Use essas métricas e perguntas de autoavaliação para avaliar como sua abordagem de acesso à rede se alinha ao posicionamento comercial e à estratégia de mercado. Essa avaliação ajuda a determinar se suas decisões atuais de rede apoiam a diferenciação de mercado, as vantagens competitivas e as necessidades do público-alvo de sua empresa.

Esta seção contém métricas e perguntas de autoavaliação para os seguintes tópicos:

- Modelo de negócios e posicionamento de mercado
- Mercado total endereçável, taxa de aquisição de novos clientes, crescimento e escalabilidade
- Experiência e retenção de clientes
- Eficiência e desempenho financeiro
- Conformidade regulatória e gerenciamento de riscos
- Estratégia de parceria

Modelo de negócios e posicionamento de mercado

Essas métricas estão relacionadas à posição da sua empresa no mercado, incluindo diferenciação competitiva, alcance de mercado e percepção da marca. É fundamental que você avalie o alinhamento entre a abordagem de acesso à rede e o modelo de negócios. Realize uma avaliação, independentemente de ser baseada em assinatura, baseada no uso, gratuita, hierárquica, de

mercado, com foco na API ou com etiqueta branca. Certifique-se de que o modelo apoie as metas da organização e as metas dos clientes.

Critérios de alta pontuação

A abordagem de acesso à rede se alinha perfeitamente ao modelo de negócios. Isso facilita a adoção e a entrega do serviço. Ele apóia a viabilidade financeira de longo prazo do modelo de negócios e a estrutura de custos é compatível com o crescimento esperado. Isso minimiza qualquer atrito para clientes ou parceiros ao adotar a oferta. Isso aprimora a experiência do usuário e incentiva uma aceitação mais ampla do serviço.

Indicadores de baixa pontuação

A abordagem de acesso à rede selecionada está desalinhada com o modelo de negócios que ela deve suportar. A estrutura de custos e o prazo de entrega para a implantação representam um obstáculo para a adoção no mercado-alvo. A infraestrutura contínua e os custos operacionais inibem quaisquer lucros potenciais. Isso impede o crescimento dos negócios e dificulta a operação na escala pretendida. Como alternativa, as propriedades da abordagem de acesso à rede podem impedir que os clientes considerem o serviço por motivos regulatórios.

Perguntas de autoavaliação

- Quais são as implicações de custo da abordagem de acesso à rede selecionada para implantação inicial e entrega contínua? Quais são os custos fixos e variáveis da abordagem?
- A abordagem de acesso à rede pode ser escalada de forma eficaz e eficiente para atender às demandas de crescimento do modelo de negócios? Considere o tamanho individual do inquilino e o número de inquilinos integrados.
- A abordagem de acesso à rede impõe alguma limitação técnica ou operacional que possa limitar a flexibilidade ou a adaptabilidade do modelo de negócios?
- Para a abordagem de acesso à rede, como o lead time de implantação se alinha à velocidade de comercialização exigida pelo modelo de negócios?

Mercado total endereçável, taxa de aquisição de novos clientes, crescimento e escalabilidade

É fundamental que você avalie o impacto das decisões de rede na capacidade da organização de se expandir para novos mercados, adquirir clientes de forma eficaz e manter a escalabilidade

operacional. Esses fatores afetam as taxas de conversão. Eles também influenciam se a abordagem de acesso à rede suporta a expansão em segmentos de mercado significativos ou se limita a atender apenas tipos específicos de clientes.

Critérios de alta pontuação

A abordagem de acesso à rede ajuda a organização a alcançar uma parte significativa do mercado-alvo ou pode ser combinada de forma eficaz com outras abordagens de rede para ampliar o alcance do mercado. Essa abordagem deve exigir um esforço adicional mínimo de integração. A abordagem oferece suporte a prazos de entrega curtos para implantação, entrada rápida no mercado e expansão. Ele permite um grande número de implantações paralelas. A integração é simples para os clientes, o que reduz as barreiras à adoção e aprimora a experiência do cliente. A abordagem minimiza a sobrecarga operacional, preserva a capacidade operacional e apoia as projeções de crescimento.

Indicadores de baixa pontuação

A abordagem de acesso à rede suporta apenas uma pequena parte do mercado-alvo ou é adequada principalmente para segmentos de nicho que não são priorizados na estratégia de negócios. Ele não complementa efetivamente outras abordagens de acesso à rede já suportadas. Os prazos de implantação atrasam as demandas do mercado, o que limita a expansão do mercado e a aquisição de novos clientes. O modelo de implantação é sequencial, o que aumenta os riscos de gargalos de serviço à medida que a demanda aumenta. Processos complexos de integração detêm clientes em potencial, o que afeta negativamente a taxa de aquisição e as taxas de conversão. Uma sobrecarga operacional significativa diminui a capacidade operacional da organização. Isso se torna um obstáculo para o crescimento projetado.

Para esses indicadores, avalie se a introdução de uma nova abordagem de acesso à rede pode ajudar a organização a alcançar seus objetivos comerciais estratégicos. Considere se a nova abordagem de acesso à rede pode criar novas dependências de produtos ou consumir recursos operacionais sem fornecer os resultados desejados.

Perguntas de autoavaliação

- Há alguma lacuna na abordagem atual que impede você de alcançar segmentos maiores do mercado-alvo?
- Qual é o conjunto mínimo de abordagens padronizadas e não sobrepostas de acesso à rede que você deve suportar para cobrir 70 a 90% do mercado-alvo?

- Que alcance cada abordagem de acesso à rede permite e quais são os aumentos relacionados em métricas importantes, como custos de infraestrutura, ciclos operacionais e dependência de especialistas?
- Como os recursos de implantação e os limites de serviço da infraestrutura de rede se alinham às expectativas de crescimento em seus mercados-alvo?
- A integração da rede cria alguma barreira à entrada de novos clientes? Como eles podem ser abordados para melhorar as taxas de conversão?
- Como a sobrecarga operacional do gerenciamento da rede afeta sua capacidade de crescimento e escalabilidade?
- Quais estratégias você pode implementar para reduzir os prazos de implantação da rede e melhorar a expansão do mercado e a aquisição de clientes?
- Há alguma dependência de recursos especializados que atrasaria a implantação ou a integração com os ecossistemas do cliente?

Experiência e retenção de clientes

As métricas desta seção ajudam você a entender a capacidade da sua organização de adquirir e, o mais importante, reter clientes. Compreender a relação entre as abordagens de acesso à rede e a satisfação do cliente pode ajudar as equipes de produto e engenharia a tomar decisões baseadas em dados.

Critérios de alta pontuação

A abordagem de acesso à rede é confiável e fácil de gerenciar. Ele contribui para resultados de alta satisfação do cliente (CSAT) e Net Promoter Score (NPS). Essas pontuações são indicativas de uma forte reputação da marca e da fidelidade do cliente. Graças à integração perfeita com os ecossistemas existentes de seus clientes, o atrito na adoção é baixo e há uma baixa dependência de especialistas. Sua organização cumpre consistentemente os contratos de nível de serviço (SLAs), o que reforça a confiança do cliente e as obrigações contratuais. Como os clientes desfrutam de serviços estáveis e confiáveis, você tem uma alta retenção de clientes.

Indicadores de baixa pontuação

A integração difícil e o acesso inconsistente aos serviços geralmente causam frustração e feedback negativo do cliente. Isso prejudica a reputação da marca. Novos clientes não conseguem converter planos gratuitos ou de teste para serviços pagos devido à dependência de especialistas ou devido a

tempos prolongados de integração e integração. Falhas frequentes no cumprimento SLAs resultam em penalidades financeiras e perda de credibilidade, reduzindo potencialmente as taxas de retenção de clientes.

Perguntas de autoavaliação

- Como o desempenho da rede (como velocidade, tempo de atividade e latência) afeta diretamente os resultados de CSAT e NPS? Quais melhorias específicas na rede poderiam aumentar essas pontuações?
- Como as métricas atuais de latência e tempo de atividade da rede afetam a experiência inicial do usuário e as taxas de adoção? Quais melhorias específicas no desempenho da rede são necessárias para otimizar essas métricas?
- Há algum problema recorrente nas configurações de rede ou de segurança que complica a integração de novos clientes? Como você pode simplificar esses processos?
- Como a facilidade de configuração do acesso à rede afeta a experiência de integração de novos usuários? Existem pontos de acesso à rede ou prazos de entrega específicos que podem ser otimizados para aprimorar as impressões iniciais do usuário?
- Quais são os desafios de automatizar o provisionamento de serviços de rede para novos clientes. Como você pode ajustar esse processo para melhorar a escalabilidade e a confiabilidade?
- Analise as principais causas das violações recentes do SLA. Eles estavam relacionados à configuração da rede, ao planejamento de capacidade ou a problemas de fornecedores externos?
- Com que frequência os problemas de rede fazem com que você perca os compromissos de SLA? Quais são as falhas mais frequentes relacionadas à rede?
- Quais melhorias no desempenho da rede mostraram o impacto positivo mais significativo na satisfação do cliente no passado?

Eficiência e desempenho financeiro

Essa categoria avalia a saúde financeira e os aspectos de lucratividade do seu negócio, como eficiência de custos, viabilidade a longo prazo, lucratividade, retorno sobre o investimento (ROI) e custo total de propriedade (TCO). Ao simplificar as operações de rede por meio da padronização, você pode reduzir a sobrecarga operacional e os custos de manutenção. Isso apoia os objetivos de crescimento da sua organização.

Critérios de alta pontuação

A estrutura de custos da abordagem de acesso à rede está bem alinhada com o modelo de negócios. Ele apoia o crescimento sustentável e a significativa economia de custos que você obtém aumenta a lucratividade. O acesso eficiente à rede permite a rápida integração do cliente, o que reduz o tempo de entrega de valor e acelera a penetração no mercado. Isso reduz diretamente o ciclo de reconhecimento de receita.

Indicadores de baixa pontuação

Os clientes estão recorrendo à concorrência para acelerar a entrega de seus aplicativos e serviços. Sua organização aumentou os custos operacionais associados a configurações de rede complexas e variadas e prazos de entrega estendidos. A estrutura de custos e o modelo de negócios estão desalinhados, o que pode causar altos custos iniciais para serviços baseados em assinatura. Processos de integração complicados reduzem a penetração no mercado e adiam o reconhecimento da receita.

Perguntas de autoavaliação

- Quais são os prazos atuais para a implantação de novos serviços e como eles afetam o tempo de lançamento no mercado e o reconhecimento da receita?
- Com que eficácia as operações de rede padronizadas reduzem as despesas gerais e os custos de manutenção?
- São necessários recursos especializados para concluir com êxito a integração inicial, operar diariamente, solucionar problemas ou implementar mudanças?
- Quão sustentáveis são os investimentos atuais na rede em termos de avanços tecnológicos? Você está investindo em tecnologias preparadas para o futuro que se alinham aos desenvolvimentos projetados do mercado?
- Com que eficiência você aloca e monitora os custos relacionados ao tráfego e ao uso da rede por locatários individuais?

Conformidade regulatória e gerenciamento de riscos

É fundamental validar a conformidade com os regulamentos relacionados à rede. Isso confirma que você está operando legalmente e pode manter a confiança do cliente. A padronização das operações de rede simplifica o processo de conformidade e promove a consistência em várias jurisdições e regiões geográficas. Essas medidas ajudam você a expandir seus serviços.

Critérios de alta pontuação

As operações de rede seguem consistentemente os padrões legais sem complicações, o que contribui para a expansão do mercado, diminui o atrito na adoção e aumenta a confiança do cliente. A conformidade comprovada com estruturas regulatórias críticas, como a Lei de Resiliência Operacional Digital (DORA) e o Instituto Nacional de Padrões e Tecnologia (NIST), ajuda você a conquistar clientes sensíveis à conformidade regulatória. A visibilidade contínua do seu status de conformidade reduz o tempo necessário para concluir uma auditoria.

Indicadores de baixa pontuação

Lacunas na conformidade da rede causam alto atrito na adoção, atrasos no lançamento do serviço, desafios legais e possíveis multas. Esses desafios levam a planos atrasados ou cancelados de expansão para novos mercados. É difícil manter práticas de conformidade padrão em diferentes jurisdições, e isso afeta a eficiência operacional e a reputação do mercado.

Perguntas de autoavaliação

- Até que ponto suas operações de rede estão alinhadas com as diretrizes regulatórias ou setoriais aplicáveis? O que sua auditoria de conformidade mais recente revelou?
- Como você está mantendo a conformidade com as regulamentações emergentes nas áreas de segurança digital e de rede?
- Quão eficaz é seu processo de documentação e emissão de relatórios para atender aos requisitos de diferentes órgãos reguladores?
- Quais estratégias de gerenciamento de risco você tem em vigor para identificar e abordar possíveis riscos de conformidade antes que eles levem a desafios legais?
- Que nível de treinamento e conscientização sobre conformidade suas equipes de gerenciamento de rede precisam para apoiar suas abordagens de acesso à rede?

Estratégia de parceria

Avalie o quão bem a abordagem de acesso à rede se alinha a um ecossistema de parceiros, plataformas e mercados reconhecidos. Isso é essencial, especialmente se sua estratégia de crescimento depender da escalabilidade por meio de parceiros.

Critérios de alta pontuação

A abordagem de acesso à rede é integrada em todo o ecossistema de seu parceiro. Sua estrutura de custos se alinha bem aos modelos de negócios de seus principais parceiros. Os parceiros possuem as habilidades de rede necessárias para a integração perfeita de suas ofertas de SaaS e podem oferecer acesso e funcionalidade sustentados.

Indicadores de baixa pontuação

A abordagem de acesso à rede selecionada exige habilidades, recursos ou equipamentos especializados que são escassos ou difíceis de adquirir. Ele difere dos protocolos de acesso à rede padrão que são comumente usados por plataformas e mercados. Isso resulta em uma estrutura de custos imprevisível que é difícil de conciliar. A abordagem de acesso à rede está desalinhada com os modelos de negócios de seus principais parceiros.

Perguntas de autoavaliação

- Quais são as implicações de custo da abordagem de acesso à rede para parceiros. Como esses custos se alinham com seus modelos de negócios? Qual lado da integração suporta a maior parte dos custos e quantos ciclos operacionais devem ser investidos?
- Para a abordagem de acesso à rede, há alguma barreira à integração ou manutenção que possa afetar os relacionamentos com parceiros ou a escalabilidade do ecossistema?
- Como a abordagem de acesso à rede pode ser otimizada para melhorar a compatibilidade e a facilidade de integração em todo o ecossistema?

Métricas de engenharia que influenciam as decisões de rede

Assim como as equipes comerciais e de produtos, as equipes de engenharia também usam critérios de sucesso para avaliar se estão atingindo os objetivos de negócios. No entanto, essas métricas são diferentes e se concentram na capacidade da equipe de desenvolver, operar e atender aos requisitos de segurança e conformidade. Esta seção descreve as métricas de engenharia que podem ser influenciadas positiva ou negativamente pelas decisões de acesso à rede que sua organização toma.

Use essas métricas e perguntas de autoavaliação para avaliar sua abordagem atual de acesso à rede em relação aos requisitos comerciais e às capacidades técnicas. Essa avaliação ajuda você a identificar lacunas em sua arquitetura e priorizar melhorias alinhadas aos seus objetivos estratégicos. Ao revisar regularmente esses critérios, você pode garantir que sua estratégia de

acesso à rede continue atendendo às necessidades de seus clientes e aos planos de crescimento de sua organização.

Esta seção contém métricas e perguntas de autoavaliação para as seguintes categorias e tópicos:

Métricas de desenvolvimento relacionadas ao acesso à rede para ofertas de SaaS

Esta seção contém as seguintes métricas:

Frequência de implantação, tempo de implantação e velocidade de sprint

Para otimizar a eficiência do ciclo de desenvolvimento, é essencial que você entenda a influência do provisionamento da pilha de rede na velocidade do sprint.

Critérios de alta pontuação

O provisionamento da pilha de rede é simplificado e automatizado, além de exigir o mínimo de intervenção manual. Isso não afeta significativamente a velocidade do sprint. O provisionamento e a reimplantação da pilha de rede podem ser realizados por qualquer membro da equipe. Isso reduz gargalos e dependências de recursos especializados.

Indicadores de baixa pontuação

É necessário um grande número de pontos históricos para provisionar a pilha de rede. Isso sugere um processo complexo e demorado que prejudica o desenvolvimento de novos recursos. A reimplantação frequente da pilha de rede gera despesas gerais substanciais de tempo e custo. As tarefas de provisionamento de rede exigem conhecimento especializado em engenharia, o que cria gargalos e retarda o ciclo de desenvolvimento.

Perguntas de autoavaliação

- Quais etapas manuais, se houver, estão envolvidas no processo de implantação. Como eles afetam a frequência e o tempo de implantação?
- Como as reversões são tratadas em caso de falhas na implantação. Qual é o impacto deles na frequência de implantação e no tempo de recuperação?
- Quantos pontos históricos são necessários para provisionar a pilha de rede quando você configura novos ambientes?
- Quantos custos adicionais e sobrecarga de tempo estão associados à reimplantação frequente da pilha de rede durante o processo de desenvolvimento?
- O provisionamento da pilha de rede depende de conhecimento especializado em engenharia ou é uma tarefa que pode ser gerenciada por qualquer membro da equipe?

Flexibilidade e entrega de recursos

A abordagem de acesso à rede pode influenciar a capacidade da equipe de engenharia de inovar e implantar novos recursos com eficiência.

Critérios de alta pontuação

A abordagem de acesso à rede oferece a flexibilidade necessária para a implantação rápida e perfeita de recursos. Ele suporta uma ampla variedade de protocolos de comunicação, comunicação

unidirecional e bidirecional e tamanhos de mensagens. Não impõe restrições significativas aos processos de desenvolvimento ou inovação.

Indicadores de baixa pontuação

A abordagem de acesso à rede restringe a capacidade da equipe de implementar novos recursos devido à falta de protocolos de comunicação compatíveis, à inflexibilidade no tamanho das mensagens ou à dependência de tecnologias específicas e recursos especializados relacionados. Isso pode levar a ciclos de desenvolvimento mais lentos e dificultar a evolução do serviço.

Perguntas de autoavaliação

- Como a abordagem de acesso à rede afeta a agilidade da equipe no desenvolvimento e na implantação de novos recursos?
- Há limitações na abordagem de acesso à rede que restringem o suporte de determinados protocolos ou tecnologias de comunicação?
- Como a abordagem facilita ou limita a integração de novas tecnologias e inovações no serviço?
- Como a abordagem de acesso à rede afeta os cronogramas de desenvolvimento e o roteiro do produto?

Alterar a taxa de falha

A abordagem de acesso à rede que você escolher pode afetar a taxa de falha de alteração ao implantar novos serviços ou recursos. Maior controle geralmente significa maior flexibilidade, mas também aumenta o potencial de configurações incorretas, como ao gerenciar uma configuração de roteamento complexa.

Critérios de alta pontuação

Você pode implementar alterações na pilha de rede com risco mínimo de falha. Existem mecanismos de teste suficientes, existem mecanismos de reversão eficientes e o monitoramento eficaz ajuda você a identificar e resolver problemas rapidamente.

Indicadores de baixa pontuação

A abordagem de acesso à rede está sujeita a falhas durante as mudanças. Há opções de teste limitadas, estratégias de implantação complicadas ou recursos insuficientes de monitoramento e solução de problemas. É necessário que várias partes participem das sessões de solução de

problemas. Isso pode aumentar o tempo de inatividade e diminuir a disponibilidade da oferta de SaaS.

Perguntas de autoavaliação

- Quais medidas estão em vigor para mitigar o risco de falha nas alterações ao atualizar a pilha de rede?
- Existem processos completos de teste e validação?
- Com que rapidez o sistema pode se recuperar de uma alteração malsucedida? Existe um processo de reversão eficiente?
- Existem sistemas proativos de monitoramento e alerta para detectar e resolver problemas rapidamente durante e após as mudanças na pilha de rede?
- Qual é a taxa histórica de falhas de alteração nas implantações de pilhas de rede. Que lições foram aprendidas com incidentes anteriores?
- Como a abordagem de acesso à rede facilita ou limita a implementação de mudanças. A abordagem minimiza a interrupção do serviço?
- Qual é o risco de afetar a disponibilidade da oferta de SaaS no ambiente de produção quando você implanta mudanças que envolvem a abordagem de acesso à rede?

Qualidade do código e desempenho da equipe de engenharia

As abordagens de acesso à rede podem afetar indiretamente a qualidade do código das ofertas de SaaS. A falta de padronização no acesso à rede pode obrigar a equipe de engenharia a oferecer suporte a várias abordagens de integração, o que pode levar a uma base de código inchada. Isso, por sua vez, pode prejudicar a capacidade da equipe de desenvolver a profundidade e o controle sobre a qualidade do código necessários para manter equipes de engenharia de alto desempenho.

Critérios de alta pontuação

A equipe de engenharia mantém o foco graças à modularidade e à reutilização do código em todas as abordagens de acesso à rede suportadas. As abordagens de acesso à rede são compatíveis com os pipelines de implantação existentes e as estratégias de teste automatizadas.

Indicadores de baixa pontuação

O desempenho da equipe de engenharia é reduzido devido à sobrecarga associada à integração e manutenção de muitas abordagens de acesso à rede. Algumas abordagens aumentam

significativamente a complexidade, geram déficit tecnológico ou exigem o desenvolvimento de soluções alternativas para lidar com recursos ausentes ou insuficientes.

Perguntas de autoavaliação

- Como a abordagem de acesso à rede gerencia a variabilidade da rede?
- Você precisa desenvolver código adicional para lidar com interrupções na conectividade?
- Uma nova abordagem de acesso à rede se integra perfeitamente às abordagens existentes ou exige um desenvolvimento personalizado significativo?
- Qual é a extensão da mudança necessária para adotar uma nova abordagem de acesso à rede? A base de código existente e os testes automatizados podem ser usados de forma eficaz?
- É fácil ou difícil implantar ou reimplantar o serviço com a abordagem de acesso à rede selecionada? Isso pode ser feito com frequência? Há alguma dependência de recursos especializados?
- A abordagem de acesso à rede facilita ou complica a adesão aos padrões de codificação e às melhores práticas?
- Como a abordagem afeta time-to-market os novos recursos ou correções?

Redução da dívida técnica

Uma avaliação do impacto de uma abordagem de acesso à rede na dívida técnica deve considerar suas capacidades de escalabilidade, observabilidade e segurança.

Critérios de alta pontuação

A abordagem simplifica efetivamente o gerenciamento da infraestrutura à medida que a base de clientes se expande. Ele oferece recursos robustos de observabilidade. out-of-the-box Isso promove monitoramento e manutenção eficientes.

Indicadores de baixa pontuação

A abordagem de acesso à rede protege inadequadamente os canais de comunicação e carece de ferramentas suficientes para observação métrica qualitativa. Também pode exigir desenvolvimento adicional para o gerenciamento da infraestrutura à medida que a base de clientes aumenta, ou pode exigir soluções alternativas para problemas de confiabilidade.

Perguntas de autoavaliação

- Como a abordagem de acesso à rede influencia a escalabilidade de longo prazo da infraestrutura? Isso facilita o crescimento contínuo com um mínimo de investimento adicional?
- Quão abrangentes são as ferramentas de observabilidade incluídas? Eles permitem o monitoramento proativo e a resolução de problemas?
- Qual é o impacto previsto da abordagem de acesso à rede na manutenção e evolução da base de código ao longo do tempo?
- A abordagem se integra bem à infraestrutura existente e planejada? Isso requer mudanças ou acréscimos significativos?

Escalabilidade, capacidade e desempenho

Para determinar a adequação de uma abordagem de acesso à rede para uma oferta de SaaS, é essencial analisar como ela mantém o desempenho ideal à medida que a demanda aumenta.

Critérios de alta pontuação

A abordagem de acesso à rede facilita perfeitamente a expansão. Ele mantém baixa latência durante o processamento da solicitação e lida com picos de tráfego com eficiência. Ele fornece desempenho consistente, independentemente do aumento dos níveis de tráfego, e não impõe limites operacionais ao crescimento.

Indicadores de baixa pontuação

A abordagem de acesso à rede não é escalável de forma eficaz, possivelmente devido às limitações inerentes à largura de banda ou à capacidade insuficiente da infraestrutura. O provisionamento e o gerenciamento de recursos aumentam a complexidade ou criam dependências. O desempenho do serviço é reduzido devido ao aumento da latência, instabilidade e variabilidade da taxa de transferência, especialmente em condições de rede congestionadas.

Perguntas de autoavaliação

- Como a abordagem de acesso à rede acomoda um número crescente de inquilinos e seus volumes de dados?
- É inerentemente escalável para atender às demandas futuras?
- Quais medidas estão em vigor para garantir que o desempenho seja consistente, mesmo durante períodos de pico de tráfego ou eventos de escalabilidade rápida?

- Como a abordagem lida com a latência e a instabilidade da rede? Existem mecanismos para otimizar a taxa de transferência de dados e minimizar os atrasos?
- A abordagem de acesso à rede pode se adaptar às diferentes condições da rede? Ele pode fornecer uma experiência de inquilino único para cada cliente?
- Qual é o impacto da abordagem de acesso à rede na infraestrutura subjacente? Ela exige atualizações ou mudanças significativas nos sistemas existentes?

Métricas de excelência operacional relacionadas ao acesso à rede para ofertas de SaaS

Esta seção contém as seguintes métricas:

- Resiliência operacional e recuperação de desastres
- Monitoramento do desempenho de serviços e aplicativos

Resiliência operacional e recuperação de desastres

A abordagem de acesso à rede deve ajudar a oferta de SaaS a resistir a vários tipos de interrupções e a se recuperar rapidamente de qualquer desastre.

Critérios de alta pontuação

Planos de recuperação de desastres estabelecidos e testados mostram consistentemente que a abordagem de acesso à rede atende aos requisitos de recuperação de desastres. A abordagem de acesso à rede oferece suporte a configurações de alta disponibilidade e a mecanismos de failover automáticos, rápidos e confiáveis.

Indicadores de baixa pontuação

A abordagem de acesso à rede dificulta a criação de uma estratégia coerente de recuperação de desastres. Você observa tempos de recuperação prolongados após interrupções. Falhas operacionais frequentes da infraestrutura de rede estão afetando a prestação de serviços.

Perguntas de autoavaliação

- Quando foi o último exercício de recuperação de desastres e quais foram os resultados?
- Quanto tempo é necessário para recuperar serviços essenciais após uma interrupção? Que parte da infraestrutura de rede precisa ser reimplantada?

- Quais melhorias podem ser feitas na infraestrutura de rede para agilizar seus planos de recuperação de desastres?
- Existem redundâncias para os componentes de rede mais críticos?
- Você automatizou a possível reimplantação da infraestrutura de rede após uma interrupção crítica?
- Como a abordagem de acesso à rede suporta a tolerância a falhas e a confiabilidade? Existem mecanismos integrados para lidar com interrupções na rede e manter a integridade dos dados?

Monitoramento do desempenho de serviços e aplicativos

A abordagem de acesso à rede pode afetar as ferramentas de monitoramento de desempenho usadas para validar a operação ideal e o tempo de atividade do serviço. Dependendo do serviço, você pode ter acesso a métricas de baixo nível (como taxas de queda de pacotes) ou métricas de nível superior (como duração da sessão). Métricas de baixo nível fornecem uma visão técnica detalhada do comportamento da rede, mas podem ser complexas de interpretar. Por outro lado, métricas de alto nível geralmente oferecem uma maneira mais direta e fácil de avaliar a experiência geral do usuário. Isso ocorre porque eles agregam o impacto das condições subjacentes da rede em indicadores claros da qualidade do serviço.

Critérios de alta pontuação

Ferramentas de monitoramento abrangentes que fornecem informações quase em tempo real estão prontamente disponíveis. Você tem alertas automatizados e sistemas de resposta que tratam de problemas de desempenho. Você pode prever possíveis gargalos ou falhas no serviço antes que eles afetem os usuários.

Indicadores de baixa pontuação

Interrupções frequentes do serviço ou problemas de desempenho acontecem sem serem observados ou corrigidos. A falta de visibilidade do desempenho do serviço resulta em uma resposta lenta aos gargalos de desempenho. É necessário que equipes multipartidárias solucionem problemas de infraestrutura de rede.

Perguntas de autoavaliação

- Quais ferramentas de monitoramento e métricas de infraestrutura de rede estão disponíveis atualmente? Quão eficazes eles são na detecção de anomalias no serviço?
- Com que rapidez você pode identificar e resolver problemas de desempenho?

- Você tem mecanismos que preveem possíveis problemas de desempenho?
- Quais melhorias você pode fazer para aprimorar os recursos de observabilidade?

Métricas de segurança e governança relacionadas ao acesso à rede para ofertas de SaaS

Esta seção contém as seguintes métricas:

Gerenciamento de segurança, conformidade e vulnerabilidade

É fundamental que você avalie os aspectos de segurança da abordagem de acesso à rede, incluindo a conformidade com os padrões de segurança e o gerenciamento de vulnerabilidades.

Critérios de alta pontuação

A abordagem de acesso à rede ajuda sua equipe a aderir às estruturas de segurança, como a International Organization for Standardization (ISO) 27001, System and Organization Controls 2 (SOC 2) ou NIST. Isso facilita a realização de auditorias de segurança regulares. Mecanismos robustos de criptografia e autenticação estão em vigor. As redes são isoladas e somente os recursos necessários são expostos à infraestrutura do cliente. Você pode detectar anomalias de rede quase em tempo real, sem sobrecarga excessiva.

Indicadores de baixa pontuação

A abordagem de acesso à rede está sujeita a violações ou vulnerabilidades de segurança recorrentes e não está em conformidade com os principais padrões de segurança. Você frequentemente observa atrasos na detecção e nas respostas a incidentes de segurança.

Perguntas de autoavaliação

- Há alguma violação de segurança recente relacionada à abordagem de acesso à rede selecionada e o que aprendemos com ela?
- Como sua abordagem de acesso à rede está em conformidade com os padrões globais de segurança?
- Quanto tempo é necessário para detectar e responder às ameaças à segurança? Como o acesso à rede ajuda ou limita essa capacidade?

- Com que frequência as avaliações de segurança são conduzidas sobre as abordagens de acesso à rede? Você pode usar ferramentas comuns para avaliar a segurança da abordagem de acesso à rede ou é necessário um software especializado?
- Qual nível de segurança é inerente à abordagem de acesso à rede e como ela se alinha às melhores práticas e aos requisitos regulatórios do setor?

Visão geral dos serviços AWS de rede para ofertas de SaaS

Esta seção discute os serviços AWS de rede mencionados neste guia. Ele também compara suas capacidades e descreve as considerações de segurança de cada serviço.

Esta seção contém os seguintes tópicos:

- AWS serviços de rede
- Comparando os recursos do serviço
- Características e considerações de segurança

AWS serviços de rede

A seguir estão as Serviços da AWS que são discutidas de forma consistente neste guia.

AWS PrivateLink

AWS PrivateLinké um serviço nativo da nuvem que pode fornecer acesso à sua oferta de SaaS se seus clientes já estiverem operando no. Nuvem AWS Seu cliente se conecta à oferta de SaaS por meio de uma interface VPC endpoint. Essa é uma interface de rede de endpoint que é provisionada em uma ou mais sub-redes na do cliente. Conta da AWS Nos cenários deste guia, o tráfego viaja pela interface VPC endpoint e chega a um Network Load Balancer em sua conta. O Network Load Balancer encaminha o tráfego para o aplicativo SaaS, que você registrou como um serviço de endpoint. Por meio de endpoints de VPC de recursos, também AWS PrivateLink pode ajudá-lo a acessar outros recursos, como bancos de dados.

Amazon VPC Lattice

O Amazon VPC Lattice é um serviço de rede de aplicativos que ajuda os provedores de SaaS a oferecer seus serviços de forma segura e eficiente a clientes que operam em vários países. VPCs Contas da AWS Os clientes acessam sua oferta de SaaS por meio do VPC Lattice, que oferece conectividade de rede consistente, controles de acesso robustos e gerenciamento avançado de tráfego. Nesses cenários, o tráfego flui pelo VPC Lattice para seus serviços de aplicativos registrados. Ele fornece comunicação escalável e segura, independentemente do serviço de computação que você usa.

emparelhamento da VPC

O emparelhamento de VPC é uma conexão de rede entre duas nuvens privadas virtuais (VPCs) que roteia o tráfego entre elas usando endereços ou IPv4 endereços privados. IPv6 O emparelhamento de VPC geralmente é usado entre entidades confiáveis, como aquelas dentro da mesma organização. Seu cliente cria uma solicitação de peering para um de seus VPCs. Quando você aceita, o tráfego pode fluir entre os dois VPCs em qualquer direção. Essa abordagem de conexão se destaca por sua singularidade porque envolve comunicação direta entre duas pessoas VPCs sem nenhum serviço intermediário ou infraestrutura para gerenciar.

AWS Transit Gateway

AWS Transit Gatewayé um hub de trânsito de rede centralizado que pode conectar VPCs conexões de rede privada virtual (VPN), AWS Direct Connect gateways, dispositivos virtuais de terceiros em uma VPC e outros gateways de trânsito. Um gateway de trânsito pode ter uma tabela de rotas diferente para cada anexo. Isso fornece flexibilidade máxima para roteamento e ajuda a isolar as redes. Geralmente é usado para VPCs conectar muitos ou para inspeção centralizada.

AWS Site-to-Site VPN

AWS Site-to-Site VPNpode usar a tecnologia Internet Protocol Security (IPsec) para estabelecer conexões entre redes locais, escritórios remotos, fábricas, outros provedores de nuvem e a rede AWS global. A conexão é estabelecida a partir de um gateway privado virtual ou gateway de trânsito em uma VPC com um gateway de cliente físico ou baseado em software, que pode estar no local ou na Nuvem AWS nuvem de outro CSP. Nuvem AWS A conexão pode ser pela Internet ou por meio de uma AWS Direct Connect conexão física. Também é possível ter uma conexão Site-to-Site VPN acelerada usando AWS Global Accelerator. Uma conexão acelerada direciona o tráfego para uma localização AWS periférica e oferece latência reduzida e desempenho aprimorado.

AWS Direct Connect

AWS Direct Connectestabelece uma conexão privada de alta velocidade entre um data center local e o. Nuvem AWS Ao ignorar a Internet pública, Direct Connect fornece uma conexão de baixa latência mais confiável, segura e consistente com o. Nuvem AWS Os clientes se conectam a um dos Direct Connect locais e, em seguida, escolhem uma conexão hospedada ou dedicada AWS. Embora essa seja uma opção de arquitetura incomum para ofertas de SaaS, ela pode ser adequada para provedores de SaaS que têm poucos consumidores corporativos, mas grandes.

Comparando os recursos do serviço

A tabela a seguir descreve os recursos suportados do Serviços da AWS que são discutidos neste guia. A seguir estão as descrições dos recursos incluídos nesta tabela:

- Intervalos CIDR sobrepostos — Pode conectar duas ou mais redes com intervalos CIDR iguais ou sobrepostos
- Comunicação bidirecional — Pode suportar um canal de comunicação bidirecional para que o consumidor de SaaS possa expor recursos internos, como um banco de dados, ao provedor de SaaS
- IPv6— Pode suportar IPv6 pilha única ou dupla
- Quadro Jumbo — Pode suportar quadros jumbo com um tamanho de quadro de até 8.500 bytes
- Nuvem híbrida — Pode suportar uma conexão com uma rede local
- Multinuvem — Pode suportar uma conexão entre redes em diferentes provedores de serviços em nuvem

Serviço ou abordagem	Intervalos CIDR sobrepostos	Comunicação bidirecional	IPv6	Quadro Jumbo	Nuvem híbrida	Multinuvem
Emparelhamento de VPC	Não	Yes (Sim)	Yes (Sim)	Sim [5]	Não	Não
AWS PrivateLink	Sim	Sim[1]	Yes (Sim)	Yes (Sim)	Não [6]	Não [6]
Amazon VPC Lattice	Sim	Sim[1]	Yes (Sim)	Yes (Sim)	Não [6]	Não [6]
AWS Transit Gateway	Não	Yes (Sim)	Yes (Sim)	Yes (Sim)	Sim [3]	Sim [3]

AWS Site-to-Site VPN	Não	Yes (Sim)	Yes (Sim)	Não	Yes (Sim)	Yes (Sim)
AWS Direct Connect	Não	Yes (Sim)	Yes (Sim)	Sim[2]	Yes (Sim)	Yes (Sim)
Acesso público à internet [4]	Não aplicável	Não	Yes (Sim)	Yes (Sim)	Yes (Sim)	Yes (Sim)

1. Com recursos de VPC no Amazon VPC Lattice
2. Somente para interfaces virtuais privadas e de trânsito
3. Com Site-to-Site VPN ou AWS Direct Connect anexos
4. Como um termo geral para AWS recursos que tornam um aplicativo acessível ao público, como um Application Load Balancer
5. Somente para conexões de emparelhamento dentro de uma Região da AWS
6. Possível por meio de uma conexão preexistente de camada 3 entre os ambientes

Características e considerações de segurança

A tabela a seguir descreve os recursos de segurança do Serviços da AWS que são discutidos neste guia.

- Meios de autenticação — Como você pode garantir que somente seus clientes possam se conectar ao seu serviço. Normalmente, ainda é necessário outro nível de autenticação para solicitações recebidas, especialmente em ambientes compartilhados de locatários.
- Criptografia em trânsito — Descreve se a criptografia em trânsito é fornecida por padrão. A criptografia nativa descreve a criptografia que AWS fornece todo o tráfego dentro VPCs VPCs, entre ou entre data centers. A criptografia suplementar descreve a criptografia que você controla e que pode ser interrompida pelo respectivo serviço.

Serviço ou abordagem	Meios de autenticação	Criptografia em trânsito

Emparelhamento de VPC	Você inicia uma solicitação de emparelhamento para a Conta da AWS VPC e a VPC do seu cliente ou aceita uma solicitação iniciada por ele. Consulte Aceitar ou rejeitar uma conexão de emparelhamento de VPC.	Somente criptografia nativa
AWS PrivateLink	Você escolhe quais Contas da AWS têm permissão para criar endpoints para seu serviço. Essas contas são conhecidas como diretores permitidos. Consulte Aceitar ou rejeitar solicitações de conexão.	Somente criptografia nativa
Amazon VPC Lattice	Você compartilha um serviço ou uma rede de serviços do VPC Lattice com os de seus clientes. Contas da AWS Consulte Compartilhe suas entidades do VPC Lattice.	Criptografia nativa e criptografia TLS suplementar
AWS Transit Gateway	Seu cliente cria uma solicitação de anexo emparelhado a partir deles Conta da AWS, ou você inicia a solicitação. Consulte Transit Gateway emparelhando anexos nos Amazon VPC Transit Gateways.	Criptografia nativa e IPsec criptografia suplementar com um anexo VPN

AWS Site-to-Site VPN	Você usa chaves IPsec pré-compartilhadas ou um certifica do privado no dispositivo do cliente. Veja as opções de autenticação de AWS Site-to-Site VPN túnel.	Criptografia suplementar IPsec
AWS Direct Connect	Seu cliente cria uma solicitaç ão de interface virtual a partir deles Conta da AWS. Veja interfaces Direct Connect virtuais e interfaces virtuais hospedadas.	Criptografia suplementar de camada 2 possível em locais selecionados. Veja os Direct Connect locais.
Acesso público à internet [1]	É necessária uma autentica ção personalizada.	Criptografia TLS suplementar possível

1. Como um termo geral para AWS recursos que tornam um aplicativo acessível ao público, como um Application Load Balancer

Avaliação das opções de acesso à rede para ofertas de SaaS

As métricas que são importantes para sua organização dependerão de quem são seus clientes, de sua estratégia de negócios e de seus objetivos organizacionais. Este guia apresenta métricas que você pode usar para escolher uma abordagem de acesso à rede, mas você deve priorizar aquelas que atendam aos requisitos exclusivos do seu caso de uso.

Esta seção contém os seguintes tópicos:

Métricas de avaliação

Algumas métricas são consistentes entre organizações e casos de uso, e essas são as métricas que podemos ajudar você a avaliar. A seguir estão essas métricas:

- Facilidade de integração — Com que rapidez e facilidade você pode integrar novos clientes?
- Custo total de propriedade (TCO) — Qual é a estrutura de custos? Além dos custos de infraestrutura fixos e variáveis, há importantes considerações de custo adicionais associadas à sobrecarga operacional, dependência de especialistas, custo de implementação de mudanças e conformidade. Para obter mais informações, consulte a seção Custo total de propriedade.
- Escalabilidade — Sua abordagem de acesso à rede é capaz de escalar para apoiar o crescimento da sua empresa? A escalabilidade de sua base de clientes tem importantes considerações arquitetônicas e organizacionais. Pense em como você pode escalar para acomodar de 5 a 100 vezes mais clientes do que você atende atualmente.
- Adaptabilidade — Você consegue implementar mudanças facilmente? As mudanças podem incluir um novo aplicativo, um novo recurso, uma plataforma diferente ou uma rede diferente.
- Isolamento de rede — Quanto da infraestrutura de rede você está expondo aos seus clientes? Você está fornecendo o grau certo de acesso ou está expondo redes inteiras? Se você isolar os recursos da rede mais cedo, será mais fácil fornecer garantias de segurança, privacidade e conformidade posteriormente.

- Observabilidade — Qual é a sua capacidade de detectar falhas ou degradações do serviço? É fácil e rápido identificar o problema? Com que rapidez (e com que sobrecarga) você pode ajudar seus clientes a entender seus pontos de falha e ajudá-los a resolvê-los?
- Tempo de reparo — Qual é o tempo de espera entre a detecção de uma falha ou degradação do serviço e a retomada das operações? Quais são os fatores que afetam essa habilidade?

Outras métricas são exclusivas de sua organização ou oferta porque estão relacionadas às suas operações, estratégia ou metas comerciais. Somente você pode avaliar essas métricas. A seguir estão essas métricas:

- Alinhamento do modelo de negócios — Qual é o seu modelo de negócios e até que ponto as abordagens de acesso individual se alinham a ele?
- Mercado endereçável total (TAM) — Qual é o seu mercado atual e futuro e quão bem ele é coberto pela abordagem de acesso à rede?
- Retorno sobre o investimento (ROI) — Quais melhorias você espera em lucratividade e margens? Os benefícios financeiros esperados são suficientes para atender às suas necessidades de acesso adaptável e flexível aos serviços?
- Conformidade regulatória — Que tipo de requisitos regulatórios se aplicam e em qual mercado?
- Acordos de nível de serviço (SLAs) — Os clientes precisam que sua oferta de SaaS esteja altamente disponível? Que tipo de compromissos você está contratualmente obrigado a cumprir?

Custo total de propriedade

Esta seção explora o custo total de propriedade (TCO), que é uma das métricas de avaliação usadas para comparar as abordagens de acesso à rede. O TCO é uma métrica composta que consiste em custos de infraestrutura fixos e variáveis, despesas gerais operacionais, dependência de especialistas, custo da mudança e custos de conformidade.

A classificação de TCO para cada abordagem de acesso à rede pode variar de acordo com seu caso de uso. Por exemplo, o custo da mudança para um provedor de SaaS com um serviço web simples e cinco inquilinos difere de um provedor de SaaS com um portfólio de produtos complexo e interconectado e centenas ou milhares de inquilinos. Além disso, nem todos os componentes têm o mesmo peso. Por exemplo, contratar um especialista em rede costuma ser mais caro do que os custos de infraestrutura que suportam uma implantação individual de seu serviço. Use os valores na tabela a seguir para orientação inicial e como ponto de referência para uma discussão mais aprofundada.

Abordagem de acesso	Custos fixos de infraestrutura	Custos variáveis de infraestrutura	Sobrecarga operacional	Dependência especializada	Custo da mudança	Custos de conformidade
Emparelhamento de VPC	Nenhum	Nenhum	Alto	Baixo	Alto	Médio
AWS PrivateLink	Baixo	Baixo	Baixo	Nenhum	Baixo	Baixo
Amazon VPC Lattice	Médio	Médio	Baixo	Baixo	Baixo	Baixo
AWS Transit Gateway	Médio	Médio	Baixo	Baixo	Baixo	Médio
AWS Site-to-Site VPN	Médio	Alto	Alto	Médio	Médio	Baixo
AWS Direct Connect	Alto	Médio	Médio	Alto	Alto	Baixo
Acesso público à internet	Baixo	Alto	Médio	Baixo	Baixo	Alto

Custos de emparelhamento de VPC

Não há custo direto de infraestrutura associado a uma conexão de emparelhamento de VPC. Quando o tráfego permanece na mesma zona de disponibilidade, não há cobrança de transferência de dados. No entanto, a sobrecarga operacional pode ser significativa porque o gerenciamento e a complexidade aumentam exponencialmente com cada conexão de peering adicional. Um pouco de

conhecimento básico de rede é suficiente para configurar uma conexão de emparelhamento, mas as mudanças na rede são difíceis de implementar com mais do que um punhado de conexões de peering. Os custos de conformidade são um pouco maiores porque ambas as partes estão expondo uma VPC inteira uma à outra, em vez de serviços individuais.

AWS PrivateLink custos

AWS PrivateLink geralmente é uma solução econômica com pequena sobrecarga operacional. Isso ocorre porque o provedor de SaaS deve gerenciar somente um Network Load Balancer, e o consumidor deve gerenciar somente VPC endpoints. Você pode fazer alterações em ambos os lados de forma transparente, o que reduz a colaboração interorganizacional cara e que consome muitos recursos. Os custos de conformidade tendem a ser baixos porque o provedor de SaaS está expondo apenas os serviços que deseja e não a rede inteira.

Custos do Amazon VPC Lattice

O Amazon VPC Lattice oferece uma estrutura de custos equilibrada com custos moderados de infraestrutura fixos e variáveis. Como uma rede de serviços totalmente gerenciada, ela reduz significativamente a sobrecarga operacional ao automatizar a descoberta de serviços, o gerenciamento de tráfego e os controles de acesso em vários. VPCs Isso simplifica a implantação inicial e o gerenciamento contínuo em comparação com as configurações manuais de rede. Você pode implementar mudanças por meio de controles baseados em políticas sem atualizações complexas de roteamento, o que reduz a dependência de especialistas em rede. Os custos de conformidade tendem a ser mais baixos do que as abordagens tradicionais de rede porque o VPC Lattice fornece controles de acesso refinados e visibilidade abrangente por meio de recursos integrados de monitoramento e registro. Isso pode facilitar a demonstração da conformidade regulatória.

AWS Transit Gateway custos

AWS Transit Gateway tem maiores cobranças horárias e de processamento de dados do que AWS PrivateLink, mas tem uma sobrecarga operacional semelhante. Você deve ter um conhecimento mais profundo do AWS Transit Gateway serviço e do roteamento para configurar corretamente todas as tabelas de rotas. AWS As mudanças na infraestrutura podem exigir atualizações de roteamento ou DNS. Os custos de conformidade são semelhantes aos do peering de VPC porque ambas as partes estão potencialmente expondo sub-redes ou inteiras uma à outra. VPCs AWS Transit Gateway as tabelas de rotas também precisam ser tratadas com cuidado, pois são compartilhadas por vários consumidores e você não deve permitir nenhum tráfego entre elas.

AWS Site-to-Site VPN custos

Como a Site-to-Site VPN basicamente envia tráfego para a Internet, o custo variável é mais alto em comparação com as taxas de transferência de dados. Embora seja um serviço gerenciado de rede privada virtual (VPN), ele vem com uma sobrecarga operacional significativa, especialmente no gateway do cliente. O provisionamento e as operações exigem conhecimento avançado de rede, e as mudanças geralmente exigem a ação de ambas as partes. Os custos de conformidade geralmente são baixos porque as equipes de segurança geralmente pré-aprovam os IPsec túneis sem revisão adicional.

AWS Direct Connect custos

AWS Direct Connect vem com o maior custo de infraestrutura fixa porque é uma conexão física privada diretamente no Nuvem AWS. É necessário conhecimento especializado para configurar e operar uma sessão do Border Gateway Protocol (BGP) (se necessário), operar uma conexão VPN e realizar engenharia de tráfego. Esse serviço reduz o esforço das equipes de segurança porque combina conectividade privada com a opção de ter, além disso, segurança de controle de acesso à mídia (MACsec) e IPsec criptografia.

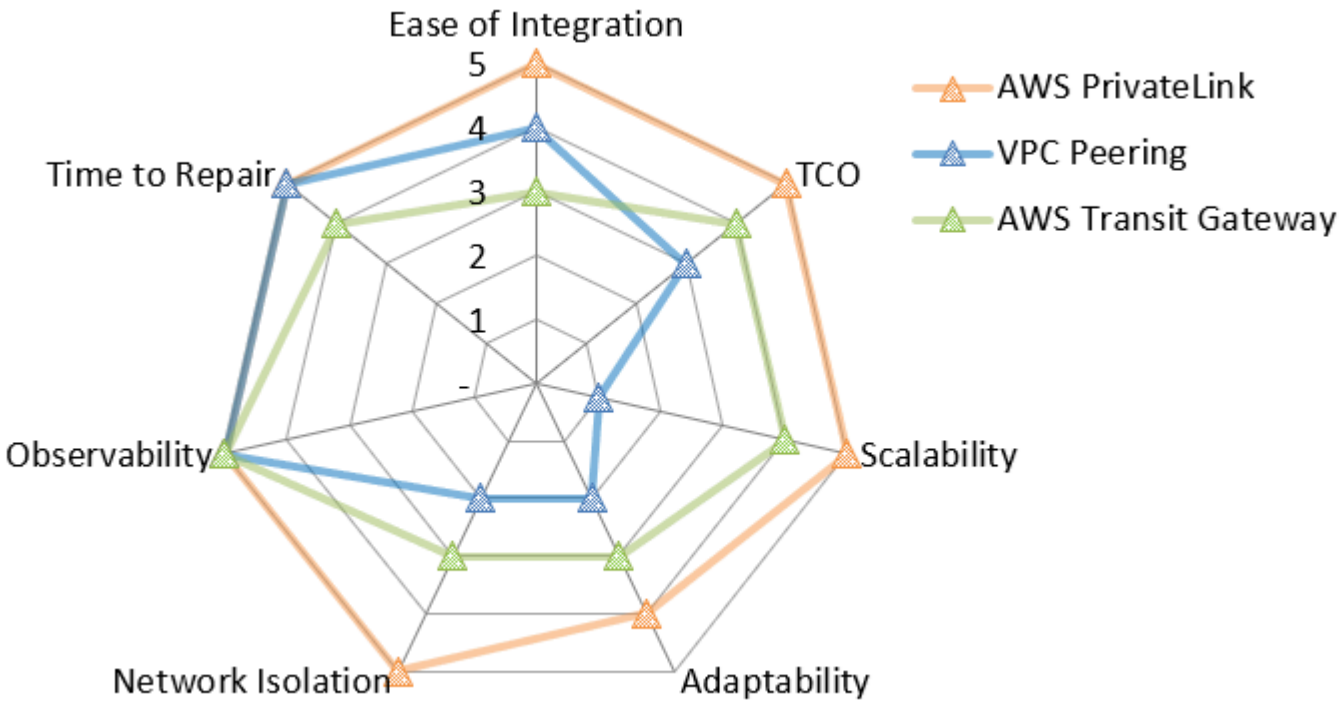
Custos de acesso público à Internet

O acesso público à Internet se refere aos AWS recursos que você pode usar para tornar um aplicativo acessível ao público, como um Application Load Balancer. Para essa abordagem, há custos variáveis vinculados ao fornecimento de acesso aos seus serviços, incluindo cobranças pela transferência de dados para a Internet. A sobrecarga operacional e os custos de conformidade podem ser significativos porque você está expondo o serviço à Internet e precisará de mecanismos adicionais de segurança e autenticação. No entanto, não há roteamento complexo envolvido e nenhuma das partes precisa saber detalhes sobre a infraestrutura uma da outra.

Mapa de valores de rede

Para ajudar você a ter uma visão geral e tomar decisões informadas, este guia inclui um mapa de valores de rede para cada cenário. Como as classificações diferem de cenário para cenário, o mesmo serviço pode ter pontuações diferentes em dois cenários. Os mapas de valores são gráficos de radar, em que uma pontuação perfeita hipotética seria cinco em todas as categorias.

Por exemplo, a imagem a seguir mostra um exemplo de gráfico de radar. Inclui somente as métricas que podemos ajudar a avaliar. Recomendamos que você crie seu próprio mapa de valores que inclua as métricas adicionais que somente você pode avaliar.



Cenários de acesso à rede para ofertas de SaaS no Nuvem AWS

Esta seção aborda diferentes opções de acesso à rede para suas ofertas de SaaS no. Nuvem AWS Ele discute as abordagens da perspectiva de seu consumidor, que pode ter necessidades de conectividade dentro do Nuvem AWS, de data centers locais ou de outros provedores de serviços em nuvem ()CSPs. Além disso, talvez seja necessário oferecer suporte ao acesso de vários tipos de ambientes de consumidores.

Compreender os requisitos de conectividade de rede nesses diversos ambientes é essencial para criar uma estratégia de acesso abrangente. Suas decisões de arquitetura devem levar em conta os diversos modelos de segurança, expectativas de desempenho e restrições técnicas, mantendo a eficiência operacional. A abordagem correta fornece conectividade segura e confiável que se adapta ao crescimento de seus negócios e minimiza a complexidade da implementação e a sobrecarga contínua de gerenciamento.

Ao avaliar as opções de acesso à rede, considere como cada abordagem afeta seu custo total de propriedade, incluindo não apenas os custos de infraestrutura, mas também as despesas operacionais e os requisitos de conformidade. Algumas abordagens se destacam em escalabilidade, mas podem introduzir complexidade, enquanto outras priorizam a facilidade de integração em detrimento do isolamento da rede. As capacidades e os recursos técnicos de seus consumidores também desempenham um papel importante na determinação da solução mais adequada.

Para consumidores do Nuvem AWS, serviços como o AWS PrivateLink oferecem vantagens significativas em segurança e escalabilidade. Os consumidores locais podem se beneficiar do AWS Direct Connect desempenho consistente ou da Site-to-Site VPN para uma conectividade econômica. Os cenários de várias nuvens geralmente exigem uma análise cuidadosa dos desafios de interoperabilidade, e você pode usar arquiteturas de VPC de trânsito para padronizar os padrões de acesso. Em todos os casos, seu design deve antecipar o crescimento futuro do consumidor e do tráfego para que sua arquitetura de rede permaneça resiliente e adaptável à medida que sua oferta de SaaS evolui.

Esta seção contém os seguintes cenários:

- Consumidores de SaaS que operam em AWS
- Consumidores de serviços que operam no local
- Consumidores de SaaS que operam em outros provedores de serviços em nuvem

- Oferecendo suporte a ambientes híbridos

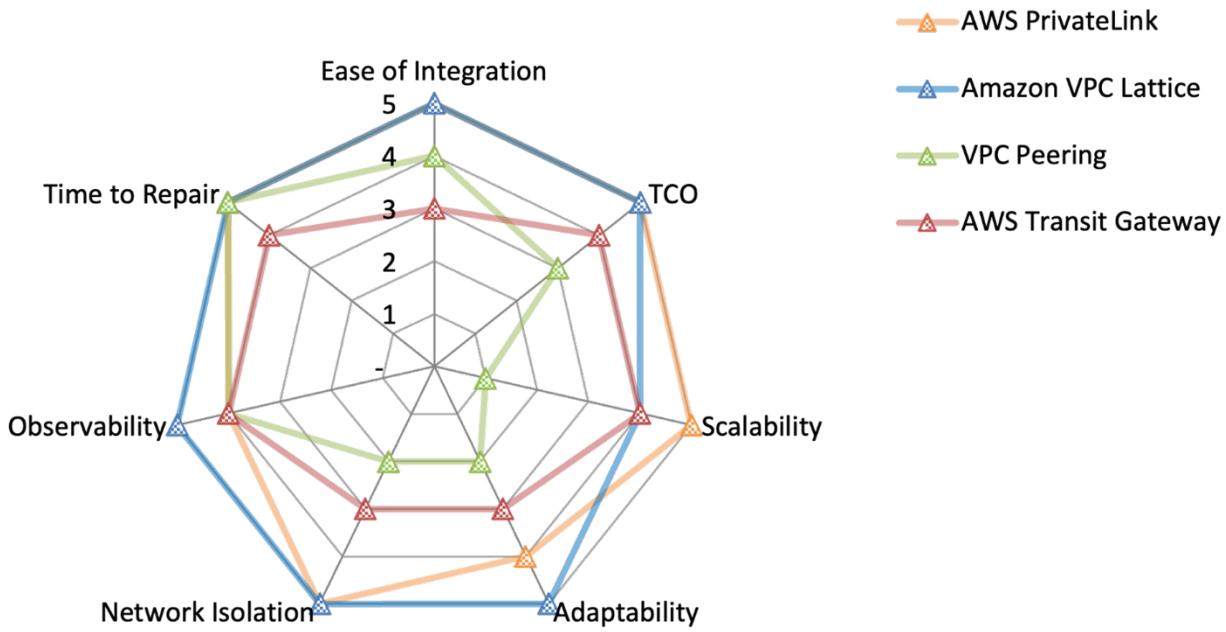
Consumidores de SaaS que operam em AWS

Esta seção discute as opções de conectividade se você e seus consumidores estiverem operando no Nuvem AWS. Esse cenário oferece a maior flexibilidade porque muitos se integram de Serviços da AWS forma nativa e porque ambas as partes têm acesso a todo o AWS service (Serviço da AWS) portfólio.

Esta seção aborda as seguintes abordagens de acesso à rede:

- Integrando com AWS PrivateLink
- Compartilhando um serviço Amazon VPC Lattice
- Criação de conexões de emparelhamento de VPC
- Conectando-se VPCs com AWS Transit Gateway

O mapa de valores de rede a seguir resume a pontuação de cada uma dessas opções em cada métrica de avaliação. Para obter mais informações sobre as métricas de avaliação, consulte Métricas de avaliação neste guia. No mapa, cinco representa a melhor pontuação, como o menor TCO, o melhor isolamento de rede ou o menor tempo de reparo. Para obter mais informações sobre como ler esse gráfico de radar, consulte Mapa de valores de rede este guia.



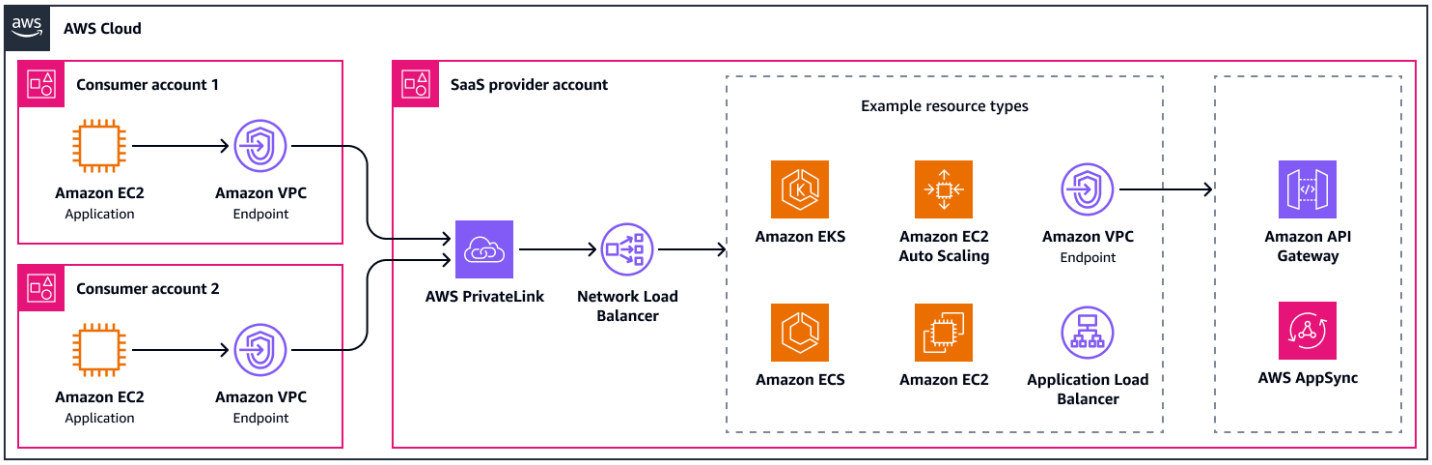
O gráfico do radar mostra os seguintes valores.

Métrica de avaliação	AWS PrivateLink	Amazon VPC Lattice	emparelhamento da VPC	AWS Transit Gateway
Facilidade de integração	5	5	4	3
TCO	5	5	3	4
Escalabilidade	5	4	1	4
Adaptabilidade	4	5	2	3
Isolamento de rede	5	5	2	3
Observabilidade	4	5	4	4
Hora de reparar	5	5	5	4

Integrando com AWS PrivateLink

AWS PrivateLinké a forma mais nativa da nuvem de integrar uma oferta de SaaS. Os provedores de SaaS podem hospedar seus aplicativos por trás de um Network Load Balancer. O Network Load Balancer se integra diretamente a um Application Load Balancer, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service(Amazon EKS) e grupos Auto Scaling. Também é possível rotear o tráfego do Network Load Balancer para os endpoints VPC de interface na conta do provedor de SaaS. Isso ajuda você a usar uma API para acessar aplicativos, como por meio do Amazon API Gateway ou AWS AppSync. Se seu aplicativo exigir acesso a recursos no ambiente do cliente que não têm carga balanceada, como um banco de dados, você pode usar endpoints de VPC de recursos.

AWS PrivateLink suporta uma largura de banda de até 100 Gbps por zona de disponibilidade.
O diagrama a seguir mostra uma configuração básica com algumas integrações possíveis. Ele conecta duas contas de consumidor à conta do provedor de SaaS por meio de. AWS PrivateLink Há endpoints de serviço nas contas dos consumidores e um Network Load Balancer na conta do provedor de SaaS.



Estes são os benefícios desta abordagem:

- Facilidade de integração: Nenhuma alteração na tabela de rotas é necessária
- Facilidade de integração: você pode oferecer serviços de endpoint por meio de AWS Marketplace
- Facilidade de integração: os endpoints VPC oferecem suporte a nomes DNS amigáveis
- Escalabilidade: pode ser escalado para milhares de consumidores de SaaS
- Adaptabilidade: Support para intervalos CIDR sobrepostos
- Adaptabilidade: Support for IPv6
- Adaptabilidade: suporte entre regiões
- TCO: AWS PrivateLink é um serviço totalmente gerenciado, portanto, requer menos esforço operacional
- Isolamento de rede: benefício de segurança para o consumidor de SaaS porque o tráfego não pode ser iniciado pelo provedor de SaaS
- Isolamento de rede: benefício de segurança para o provedor de SaaS porque ele não está expondo uma sub-rede ou VPC inteira

A seguir estão as desvantagens dessa abordagem:

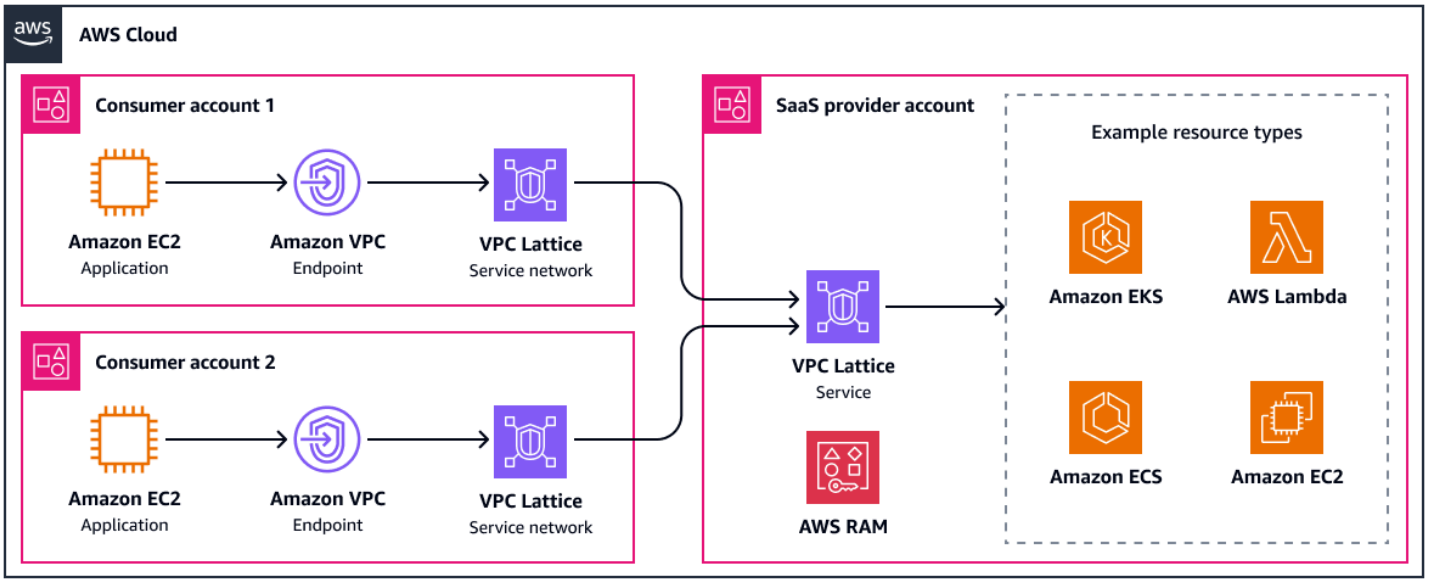
- Adaptabilidade: o provedor de SaaS deve usar as mesmas zonas de disponibilidade que o consumidor
- Adaptabilidade: Support somente para conexões iniciadas pelo cliente, e endpoints VPC de recursos são necessários para a comunicação iniciada pelo serviço
- Adaptabilidade: o Network Load Balancer é a única integração direta para AWS PrivateLink

Compartilhando um serviço Amazon VPC Lattice

Para usar o Amazon VPC Lattice como uma opção de conectividade para seu aplicativo SaaS, primeiro você cria um ou mais serviços VPC Lattice que representam os componentes do seu aplicativo SaaS. Você configura ouvintes e regras de roteamento para direcionar o tráfego para seus destinos de back-end, como instâncias, contêineres ou funções do Amazon EC2. AWS Lambda Para obter mais informações, consulte Conectando serviços Saas em uma rede de serviços VPC LatticeAWS (postagem no blog). Em termos de conceito, isso é quase o mesmo que configurar um Application Load Balancer. Em seguida, você compartilha seu serviço SaaS de forma segura com clientes Contas da AWS ou organizações usando AWS Resource Access Manager (AWS RAM), especificando quais permissões eles têm. Depois que os clientes aceitarem o compartilhamento de recursos, eles poderão associar seu serviço SaaS às redes de serviços VPC Lattice existentes ou recém-criadas para permitir a comunicação. service-to-service

Cada serviço VPC Lattice pode suportar até 10 Gbps e 10.000 solicitações por segundo por zona de disponibilidade. Ao implementar políticas de autenticação, seus clientes podem ter um controle refinado sobre quais serviços e recursos podem acessar o aplicativo SaaS. Você pode usar gateways de recursos para acessar recursos que exigem uma conexão TCP. Por exemplo, pode ser um cluster do Amazon EKS que você gerencia ou pode ser um recurso gerenciado pelo cliente que seu aplicativo precisa acessar. Para obter mais informações sobre o uso de gateways de recursos para ofertas de SaaS, consulte Estender os recursos de SaaS Contas da AWS usando o AWS PrivateLink suporte para recursos de VPC (postagem no blog).AWS

O diagrama a seguir mostra uma configuração de alto nível do VPC Lattice com alguns exemplos de integrações. Ele usa redes de serviços gerenciadas pelo cliente para acessar o aplicativo SaaS.



Estes são os benefícios desta abordagem:

- Facilidade de integração: Nenhuma alteração na tabela de rotas é necessária
- Facilidade de integração: descoberta de serviços pronta para uso
- Escalabilidade: pode ser escalado para milhares de consumidores de SaaS
- Adaptabilidade: Support para intervalos CIDR sobrepostos
- Adaptabilidade: Support for IPv6
- Adaptabilidade: integra-se a qualquer serviço de AWS computação como um serviço VPC Lattice
- TCO: O VPC Lattice é um serviço totalmente gerenciado, portanto, exige menos esforço operacional
- TCO: balanceamento de carga integrado com roteamento de tráfego avançado
- Isolamento de rede: autorização refinada com políticas de autenticação
- Isolamento de rede: benefício de segurança para o consumidor de SaaS porque o tráfego não pode ser iniciado pelo provedor de SaaS
- Isolamento de rede: benefício de segurança para o provedor de SaaS porque você não está expondo uma sub-rede ou VPC inteira

A seguir estão as desvantagens dessa abordagem:

- Adaptabilidade: Support somente para conexões iniciadas pelo cliente, e gateways de recursos são necessários para a comunicação iniciada pelo serviço
- Adaptabilidade: sem suporte entre regiões

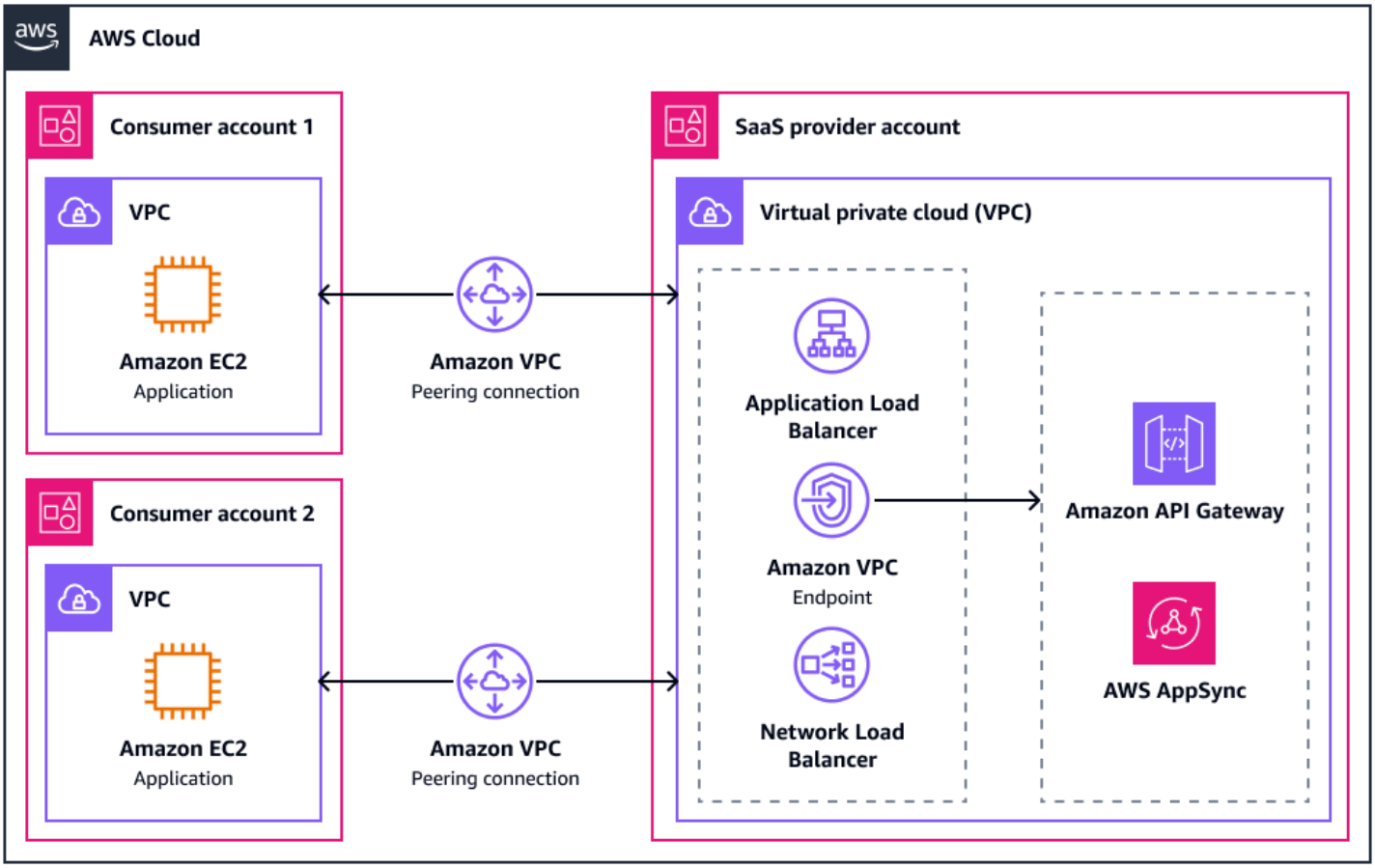
Criação de conexões de emparelhamento de VPC

Quando você usa o emparelhamento de VPC para conectar a VPC do provedor de SaaS à VPC do consumidor, ambas as partes podem iniciar conexões. Isso requer a configuração adequada de grupos de segurança, firewalls e listas de controle de acesso à rede (NACLs) em ambas as contas. Caso contrário, o tráfego indesejado poderá entrar na rede por meio da conexão de peering. Você pode usar grupos de segurança para referenciar grupos de segurança a partir do peering VPCs. Isso pode ajudar você a controlar o acesso ao seu aplicativo porque os grupos de segurança da lista de permissões fornecem um controle de acesso mais explícito e granular em comparação com os endereços IP da lista de permissões.

Com o emparelhamento de VPC, a oferta de SaaS pode ser alcançada por meio de um serviço ou recurso implantado na VPC. A maioria dos aplicativos SaaS está por trás de um Application Load Balancer ou Network Load Balancer. AWS AppSync private APIs ou Amazon API Gateway private APIs são outros pontos de entrada comuns para aplicativos SaaS porque podem ser um alvo em uma conexão de peering por meio de endpoints de interface VPC.

Depois de estabelecer uma conexão de emparelhamento, você deve atualizar as tabelas de rotas VPCs em ambas as contas para definir a conexão de emparelhamento como o próximo salto para o respectivo intervalo CIDR. Essa solução é recomendada somente para provedores de SaaS que têm poucos consumidores, pois gerenciar várias conexões de peering rapidamente se torna muito complexo.

O diagrama a seguir mostra uma configuração básica com algumas integrações possíveis. VPCs em duas contas de consumidores, tenha uma conexão de emparelhamento com uma VPC na conta do provedor de SaaS.



Estes são os benefícios desta abordagem:

- Hora de reparar: nenhum ponto único de falha na comunicação

- Escalabilidade: sem limitações de largura de banda em relação ao peering de VPC
- TCO: Sem custo de conexão de emparelhamento ou tráfego pela conexão de emparelhamento dentro da mesma zona de disponibilidade
- TCO: Sem infraestrutura para gerenciar
- Adaptabilidade: Support for IPv6
- Adaptabilidade: suporte para emparelhamento entre regiões

A seguir estão as desvantagens dessa abordagem:

- Adaptabilidade: Não há suporte para roteamento transitivo
- Adaptabilidade: Não há suporte para intervalos CIDR sobrepostos
- Escalabilidade: escalabilidade limitada (máximo de 125 conexões emparelhadas por VPC)
- TCO: a complexidade cresce exponencialmente com cada conexão de peering adicional
- TCO: despesas gerais de gerenciamento de tabelas de rotas, conexões de emparelhamento em si, regras de grupos de segurança e inspeção de tráfego
- Isolamento de rede: controles de segurança rígidos são necessários VPCs porque ambas as partes estão expostas

Conectando-se VPCs com AWS Transit Gateway

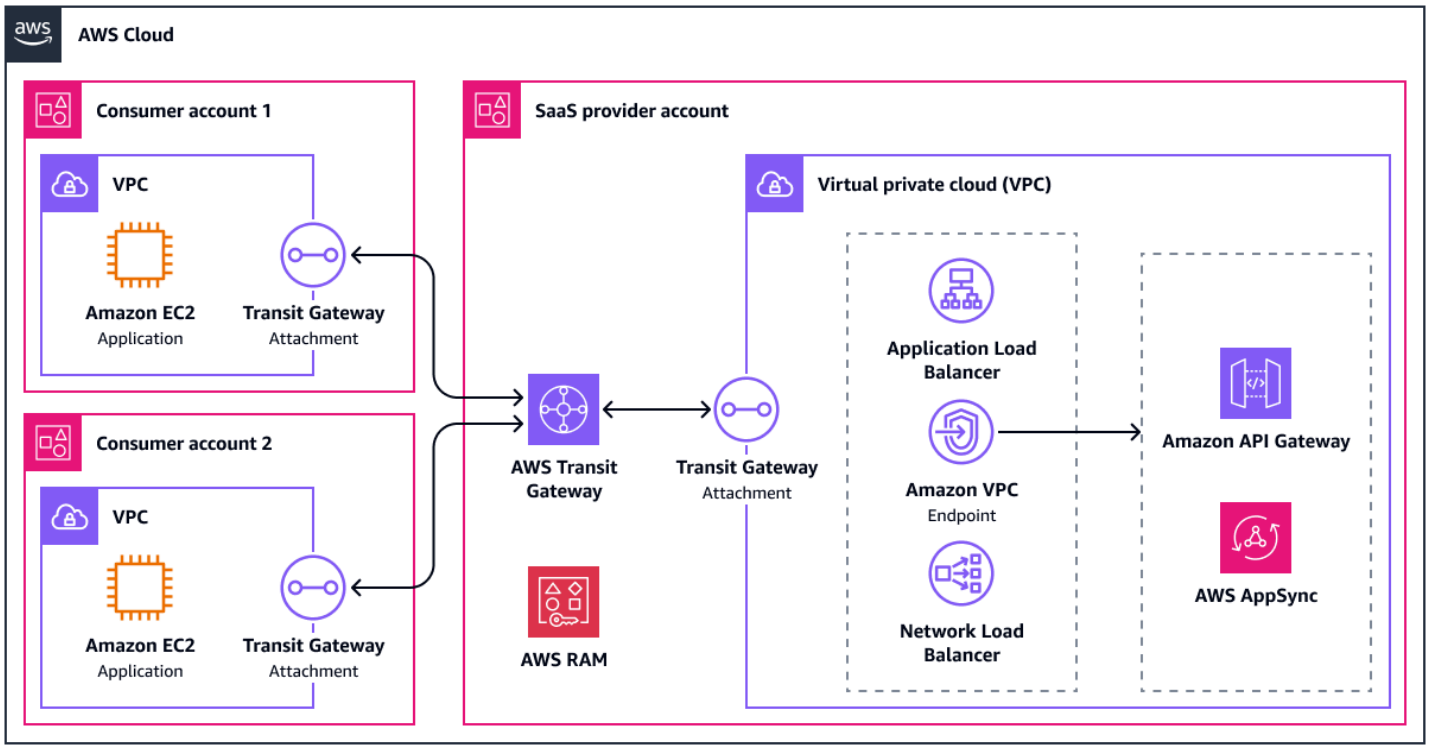
Quando você se conecta VPCs AWS Transit Gateway, ele cria anexos de VPC e implanta interfaces de rede nas sub-redes de cada zona de disponibilidade que devem rotear o tráfego de e para a VPC. É recomendável ter uma `/28` sub-rede dedicada em cada zona de disponibilidade para o anexo VPC. Para obter mais informações, consulte as melhores práticas de design do Amazon VPC Transit Gateways. Eles VPCs precisam de uma tabela de rotas atualizada para enviar tráfego pela interface de rede implantada, e as tabelas de rotas do Transit Gateway precisam ser atualizadas adequadamente. Em uma configuração multilocatária, você deseja que a VPC do provedor de SaaS tenha uma rota para a de todos os consumidores. VPCs O consumidor VPCs deve ter uma rota somente para a VPC do provedor de SaaS.

O Transit Gateway está altamente disponível por design. Ele oferece suporte ao monitoramento com VPC Flow Logs, e a largura de banda máxima para um anexo do Transit Gateway é de 100 Gbps por zona de disponibilidade. Assim como o peering de VPC, essa abordagem permite a referência de grupos de segurança entre VPCs, o que simplifica o controle de acesso entre os ambientes.

Há duas opções principais para conectar os consumidores à sua oferta de SaaS com o Transit Gateway.

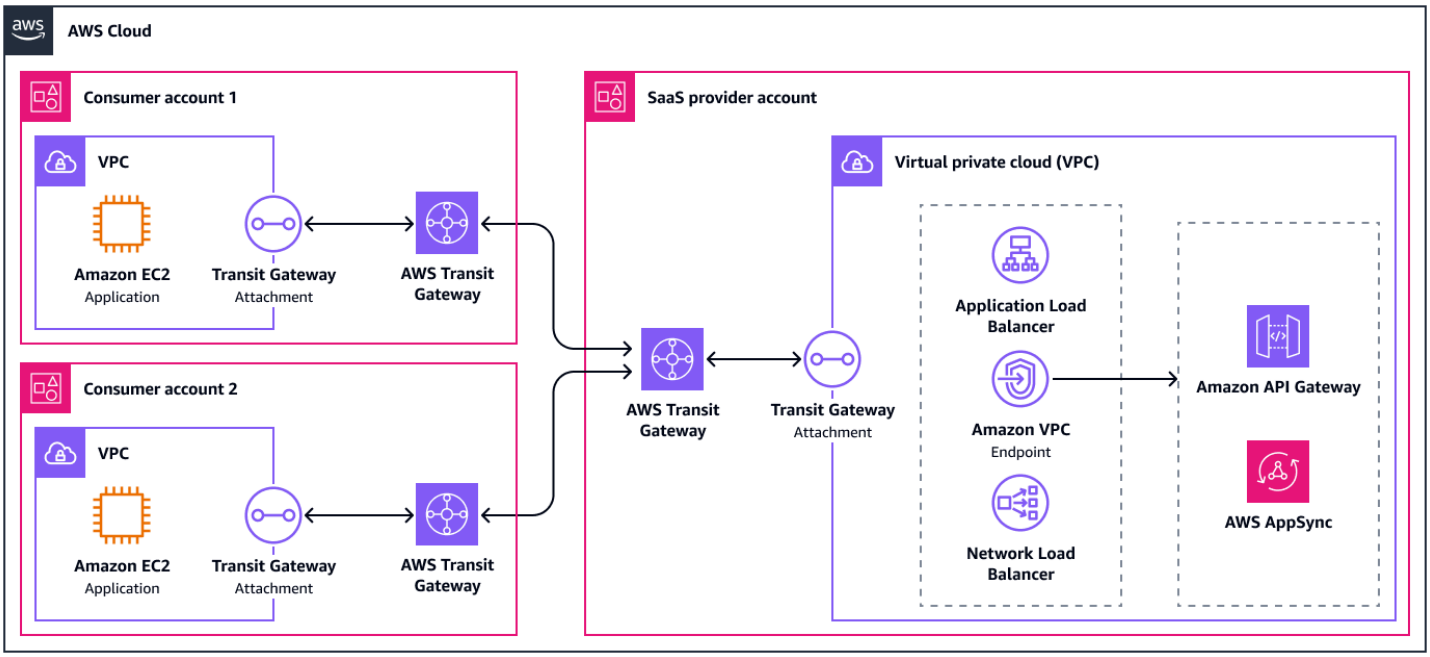
Opção 1: usando RAM

Na primeira opção, o provedor de serviços compartilha o Transit Gateway com os consumidores usando AWS Resource Access Manager (AWS RAM). Isso permite que os consumidores implantem os anexos da VPC em suas próprias contas. O diagrama a seguir mostra essa opção em um alto nível.



Opção 2: Gateways de trânsito pareados

A segunda opção é emparelhar seu gateway de trânsito com um gateway de trânsito nas contas dos consumidores. Isso proporciona aos consumidores mais flexibilidade, pois agora eles podem controlar totalmente as tabelas de rotas em seu gateway de trânsito. Por exemplo, eles poderiam configurar uma inspeção centralizada entre o serviço e suas cargas de trabalho. Uma desvantagem dessa opção é que somente o roteamento estático entre gateways de trânsito é suportado. O diagrama a seguir mostra essa opção em um alto nível.



Estes são os benefícios desta abordagem:

- Escalabilidade: Support para até 5.000 anexos
- Escalabilidade: um lugar para gerenciar e monitorar todos os conectados VPCs
- Adaptabilidade: o Transit Gateway também pode se conectar a VPNs Direct Connect gateways e dispositivos SD-WAN de terceiros
- Adaptabilidade: arquitetura flexível, como adicionar uma VPC de inspeção
- Adaptabilidade: Support para roteamento transitivo
- Adaptabilidade: pode emparelhar gateways de trânsito intra-regionais e inter-regionais
- Adaptabilidade: Support for IPv6
- TCO: AWS Transit Gateway é um serviço totalmente gerenciado, portanto, requer menos esforço operacional
- TCO: o TCO cresce linearmente com cada conexão adicional de gateway de trânsito

A seguir estão as desvantagens dessa abordagem:

- Facilidade de integração: a configuração de roteamento requer conhecimento avançado de rede
- Adaptabilidade: Não há suporte para intervalos CIDR sobrepostos
- TCO: despesas gerais de gerenciamento de entradas de tabelas de rotas, regras de grupos de segurança e inspeção de tráfego

- Segurança: controles de segurança rígidos são necessários porque ambas as VPCs partes estão expostas

Consumidores de serviços que operam no local

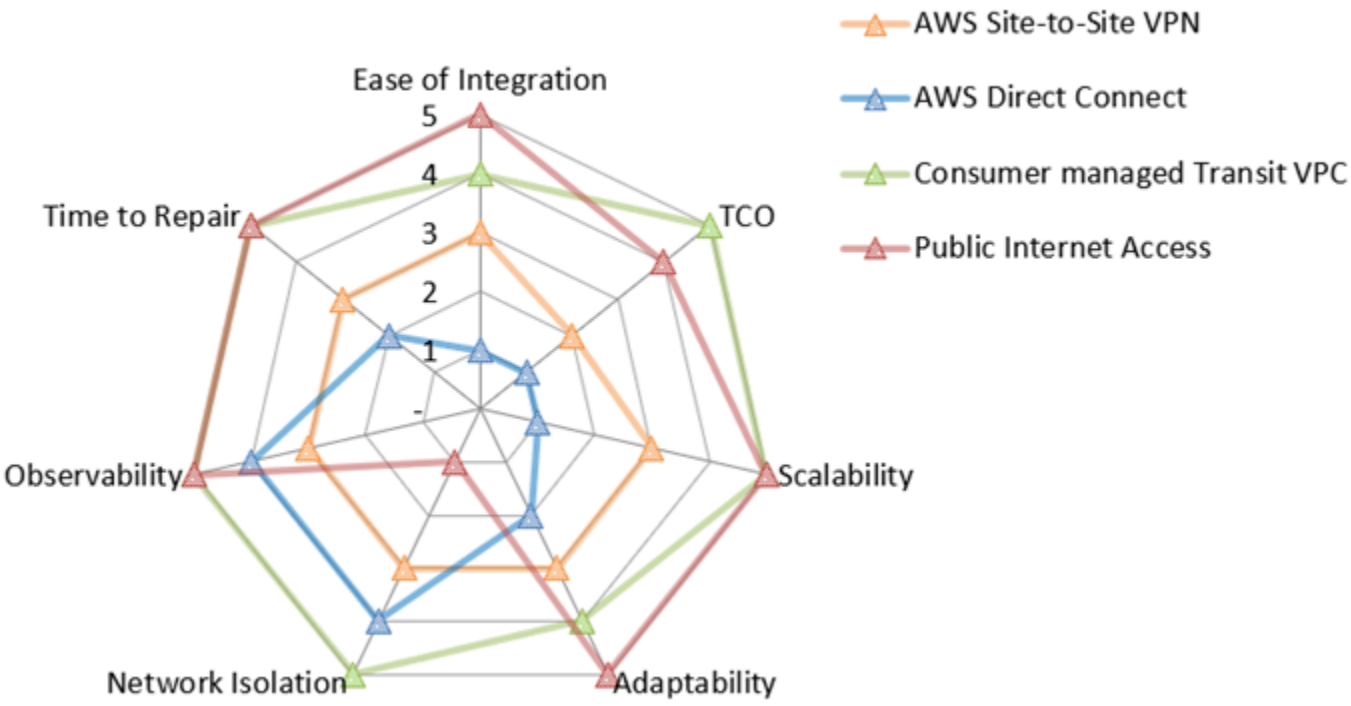
Esta seção discute as opções de conectividade entre cargas de trabalho SaaS nos data centers locais e Nuvem AWS nos data centers. Muitos consumidores com requisitos locais, especialmente no nível corporativo, veem a nuvem como uma extensão de sua rede física e querem refletir isso em sua arquitetura. Isso significa conectividade privada com a oferta de SaaS na nuvem, seja por meio de túneis lógicos ou até mesmo por meio de uma conexão física privada. Outros consumidores aceitarão a conectividade por meio da Internet pública, o que também é discutido nesta seção.

Esta seção aborda as seguintes abordagens de acesso à rede:

- Conectando-se com AWS Site-to-Site VPN
- Conectando-se com AWS Direct Connect
- Conexão com uma arquitetura VPC de trânsito
- Conectando-se pela Internet pública

O mapa de valores de rede a seguir resume a pontuação de cada uma dessas opções em cada métrica de avaliação. Para obter mais informações sobre as métricas de avaliação, consulte Métricas de avaliação neste guia. No mapa, cinco representa a melhor pontuação, como o menor TCO, o melhor isolamento de rede ou o menor tempo de reparo. Para obter mais informações sobre como ler esse gráfico de radar, consulte Mapa de valores de rede este guia.

> **Note**
>
> A opção VPC de trânsito gerenciada pelo provedor está excluída porque as pontuações dependem muito de quais serviços estão sendo operados.



O gráfico do radar mostra os seguintes valores.

Métrica de avaliação	AWS Site-to-Site VPN	AWS Direct Connect	VPC de trânsito gerenciada pelo consumidor	Acesso público à internet
Facilidade de integração	3	1	4	5
TCO	2	1	5	4
Escalabilidade	3	1	5	5
Adaptabilidade	3	2	4	5
Isolamento de rede	3	4	5	1
Observabilidade	3	4	5	5
Hora de reparar	3	2	5	5

Conectando-se com AWS Site-to-Site VPN

AWS Site-to-Site VPNas conexões podem terminar em um gateway privado virtual ou em um gateway de trânsito. Um gateway privado virtual é o endpoint VPN no AWS lado da sua conexão Site-to-Site VPN que pode ser conectado a uma única VPC. Um gateway de trânsito é um hub de trânsito que pode ser usado para interconectar várias VPCs redes locais. Ele também pode ser usado como um endpoint VPN para o AWS lado da conexão Site-to-Site VPN. Esta seção discute as duas opções.

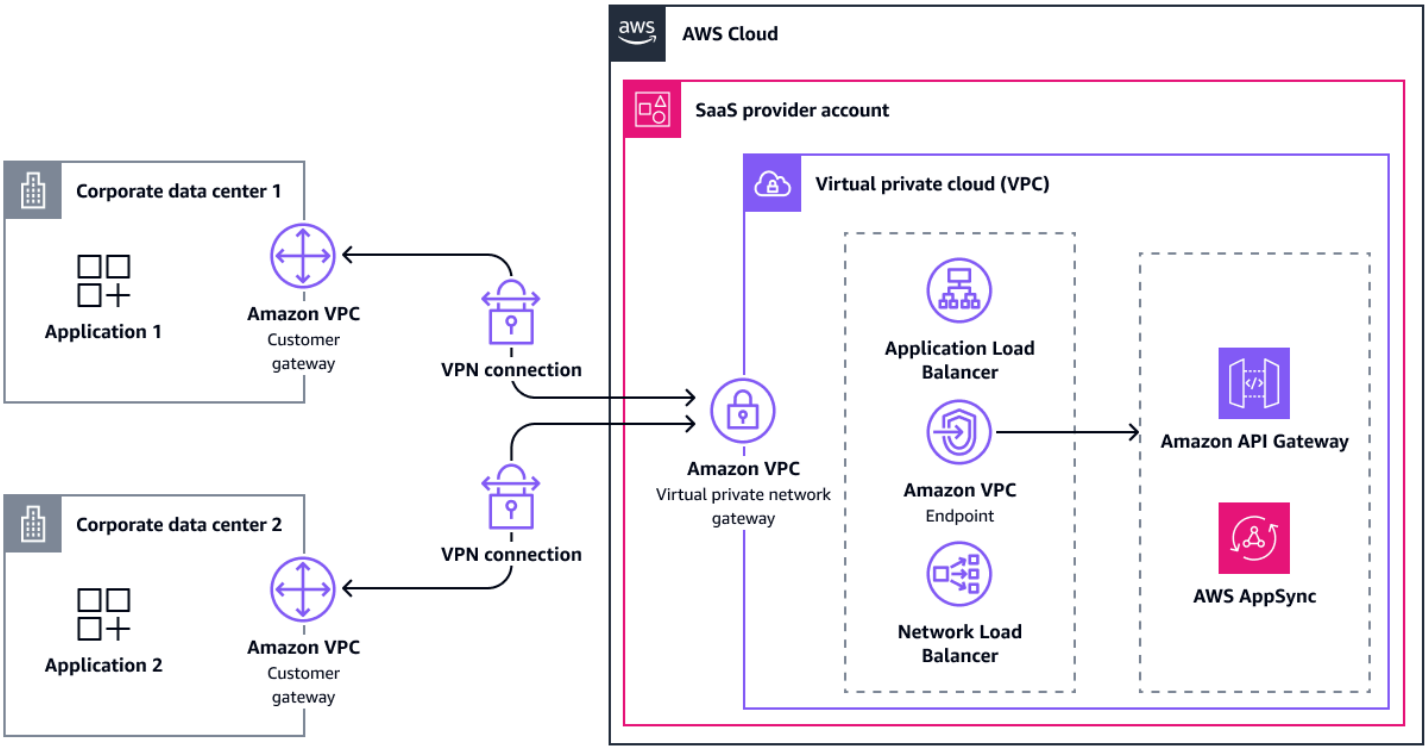
Conexão por meio de um gateway privado virtual

Depois de criar um gateway privado virtual, você o anexa à VPC que contém sua oferta de SaaS. Em seguida, você ativa a propagação de rotas para propagar as rotas de VPN para a tabela de rotas da VPC. Essas rotas podem ser rotas estáticas ou dinâmicas anunciadas pelo BGP.

Para alta disponibilidade, uma conexão Site-to-Site VPN tem dois túneis VPN que terminam em duas zonas de disponibilidade na lateral. AWS Se um ficar indisponível, o segundo túnel pode assumir o controle. Um único túnel permite uma largura de banda máxima de 1,25 Gbps. Como os gateways privados virtuais não oferecem suporte ao roteamento multicaminho (ECMP) de custo igual, você pode usar somente um túnel por vez.

Para aumentar a tolerância a falhas, você pode configurar uma segunda conexão VPN com um segundo gateway físico do cliente. Depois que a conexão é estabelecida, o consumidor pode acessar recursos na VPC do provedor de SaaS.

O diagrama a seguir mostra essa arquitetura.



Estes são os benefícios desta abordagem:

- Hora do reparo: failover gerenciado para o túnel VPN secundário
- Observabilidade: integração para monitoramento ativo gerenciado usando o Network Synthetic Monitor
- Facilidade de integração: suporte de roteamento dinâmico por meio do BGP
- Adaptabilidade: compatibilidade com a maioria dos equipamentos de rede locais
- Adaptabilidade: suporte IPv6
- TCO: AWS Site-to-Site VPN é um serviço totalmente gerenciado, portanto, requer menos esforço operacional
- TCO: sem custo para gateways virtuais, embora haja cobranças pelos dois IPv4 endereços públicos em cada um
- Isolamento de rede: permite comunicação privada segura pela Internet

A seguir estão as desvantagens dessa abordagem:

- Facilidade de integração: o consumidor deve configurar o gateway do cliente

- Escalabilidade: a falta de suporte a ECMP limita a largura de banda a 1,25 Gbps por gateway virtual
- Escalabilidade: escalabilidade limitada devido ao aumento da complexidade da rede e da sobrecarga operacional
- Adaptabilidade: IPv6 suporte somente para os endereços IP internos dos túneis VPN
- Adaptabilidade: Sem roteamento transitivo
- TCO: sobrecarga operacional para manter, gerenciar e configurar várias conexões VPN para o provedor de SaaS

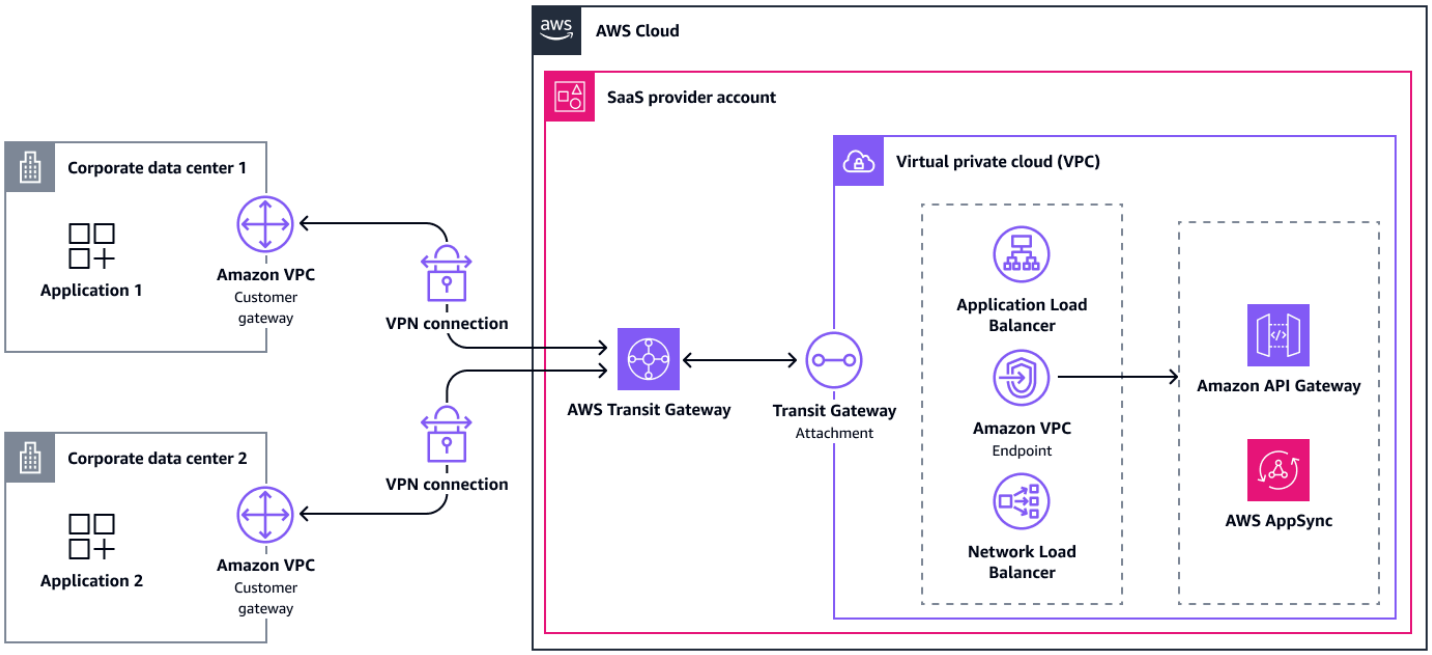
Conexão por meio de um gateway de trânsito

As conexões por meio de gateways de trânsito são semelhantes às dos gateways virtuais. No entanto, há algumas diferenças que você deve ter em mente.

Primeiro, as rotas para o anexo VPN podem ser propagadas automaticamente na tabela de rotas do Transit Gateway, mas você deve adicionar manualmente as rotas ao anexo VPCs.

Comparado a um gateway virtual, o Transit Gateway suporta ECMP. Se o gateway do cliente suportar ECMP, ele poderá usar os dois túneis para atingir uma taxa de transferência máxima total de 2,5 Gbps. Você pode estabelecer várias conexões entre a mesma rede local e o gateway de trânsito. Usando essa abordagem, você pode aumentar a largura de banda máxima em até 2,5 Gbps por conexão.

O diagrama a seguir mostra essa arquitetura.



Estes são os benefícios desta abordagem:

- Hora do reparo: failover gerenciado para o túnel VPN secundário
- Observabilidade: integração para monitoramento ativo gerenciado usando o Network Synthetic Monitor
- Facilidade de integração: suporte de roteamento dinâmico por meio do BGP
- Escalabilidade: o suporte a ECMP permite escalar a taxa de transferência da VPN para atender aos grandes requisitos de largura de banda
- Escalabilidade: grande número de conexões VPN suportadas por um único gateway de trânsito (até quase 5.000)
- Escalabilidade: um lugar para gerenciar e monitorar todas as conexões VPN
- Adaptabilidade: compatibilidade com a maioria dos equipamentos de rede locais
- Adaptabilidade: suporte IPv6
- Adaptabilidade: herde a flexibilidade do AWS Transit Gateway
- TCO: AWS Transit Gateway é um serviço totalmente gerenciado, portanto, requer menos esforço operacional
- TCO: sem custo para gateways virtuais, embora haja cobranças pelos dois IPv4 endereços públicos em cada um
- Isolamento de rede: permite comunicação privada segura pela Internet

A seguir estão as desvantagens dessa abordagem:

- Facilidade de integração: o consumidor deve configurar o gateway do cliente
- Escalabilidade: escalabilidade limitada devido ao aumento da complexidade da rede e da sobrecarga operacional
- Adaptabilidade: IPv6 suporte somente para os endereços IP internos dos túneis VPN
- TCO: sobrecarga operacional para manter, gerenciar e configurar várias conexões VPN para o provedor de SaaS
- TCO: taxas extras pelo uso de AWS Transit Gateway
- TCO: complexidade adicional no gerenciamento das tabelas de rotas do gateway de trânsito

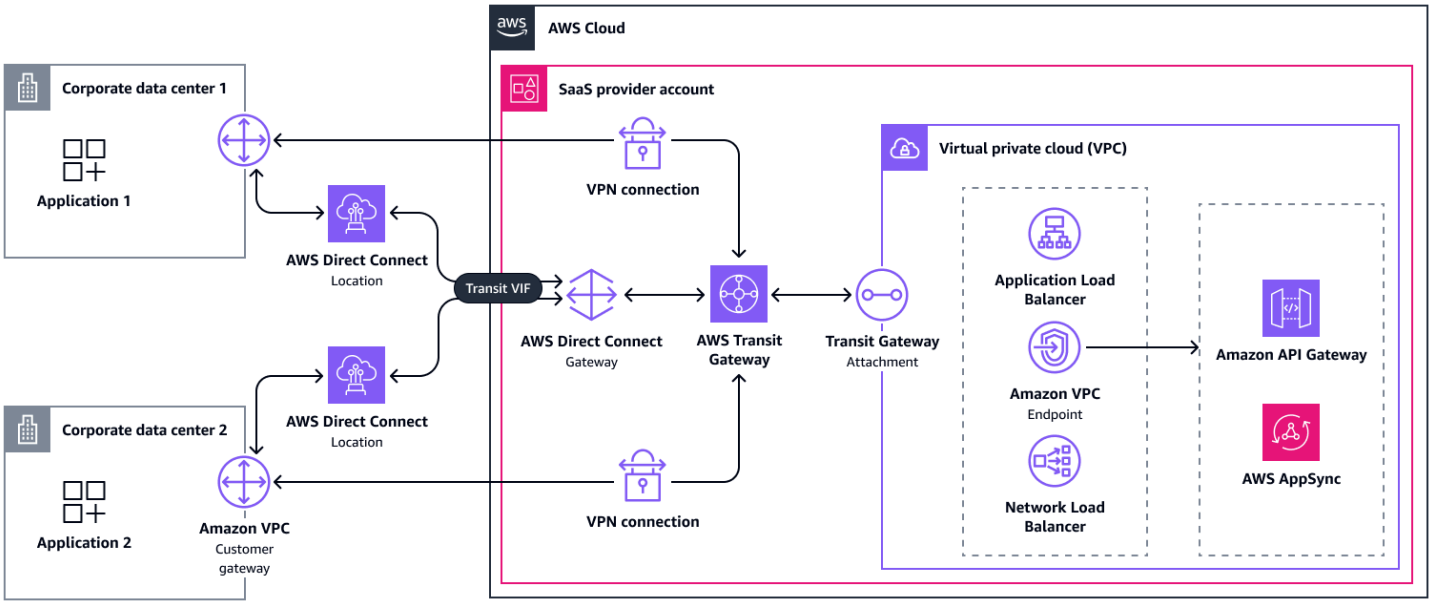
Conectando-se com AWS Direct Connect

AWS Direct Connectconecta sua rede interna a um Direct Connect local por meio de um cabo de fibra óptica Ethernet padrão. Diferentemente das outras opções de arquitetura, uma conexão dedicada não pode ser estabelecida em alguns minutos. Em vez disso, esse processo pode levar vários dias se todos os requisitos forem atendidos. Caso contrário, pode levar mais tempo. Portanto, sugerimos que você entre em contato com sua equipe de AWS contas ou AWS Support peça ajuda com essa abordagem. Opcionalmente, você pode escolher uma conexão hospedada fornecida por um AWS parceiro e compartilhada com outros clientes. Independentemente disso, a arquitetura é a mesma. Você pode escolher Direct Connect porque reduz a latência, melhora a largura de banda ou está em conformidade com os requisitos regulamentares.

Para usar a Direct Connect conexão, os consumidores devem criar uma interface virtual pública, privada ou de trânsito. Há diferentes opções de arquitetura disponíveis. A mais flexível para conectar vários locais locais ao Nuvem AWS é uma interface virtual de trânsito conectada a um Direct Connect gateway. Um Direct Connect gateway é um componente lógico global que permite ao provedor de serviços conectar até seis gateways de trânsito a ele. Além disso, você pode conectar até 30 interfaces virtuais ao gateway. Para escalar, você pode criar Direct Connect gateways adicionais. Na conta do provedor de SaaS, os gateways de trânsito então se conectam ao VPCs, conforme descrito anteriormente.

Os consumidores podem se conectar usando de uma a quatro Direct Connect conexões de um total de um ou dois Direct Connect locais, dependendo do nível de resiliência desejado. Para obter mais informações, consulte Configurar Direct Connect para máxima resiliência. Uma AWS Site-to-Site VPN conexão pela Internet também pode servir como um caminho de backup de baixo custo para uma Direct Connect conexão. As conexões Direct Connect dedicadas suportadas podem ser usadas

MACsecpara criptografar o link na camada 2 entre o Direct Connect local e o data center. É comum ter uma conexão Site-to-Site VPN para maior confidencialidade dos dados. A conexão Site-to-Site VPN pode terminar no gateway de trânsito usando um anexo VPN normal. O diagrama a seguir mostra essa arquitetura.



Estes são os benefícios desta abordagem:

- Observabilidade: integração para monitoramento ativo gerenciado usando o Network Synthetic Monitor
- Escalabilidade: Support para maior taxa de transferência de largura de banda
- Adaptabilidade: suporte IPv6
- TCO: potencial para reduzir a transferência de dados
- TCO: experiência de rede consistente
- Isolamento de rede: conectividade privada que pode atender aos requisitos regulatórios

A seguir estão as desvantagens dessa abordagem:

- Facilidade de integração: tempo e esforço manual para configurar
- Escalabilidade: escalabilidade limitada além de dezenas de Direct Connect conexões porque há várias cotas para rastrear
- Adaptabilidade: as opções de configuração dependem dos locais disponíveis Direct Connect
- TCO: a Direct Connect manutenção programada pode causar tempo de inatividade que exige ação

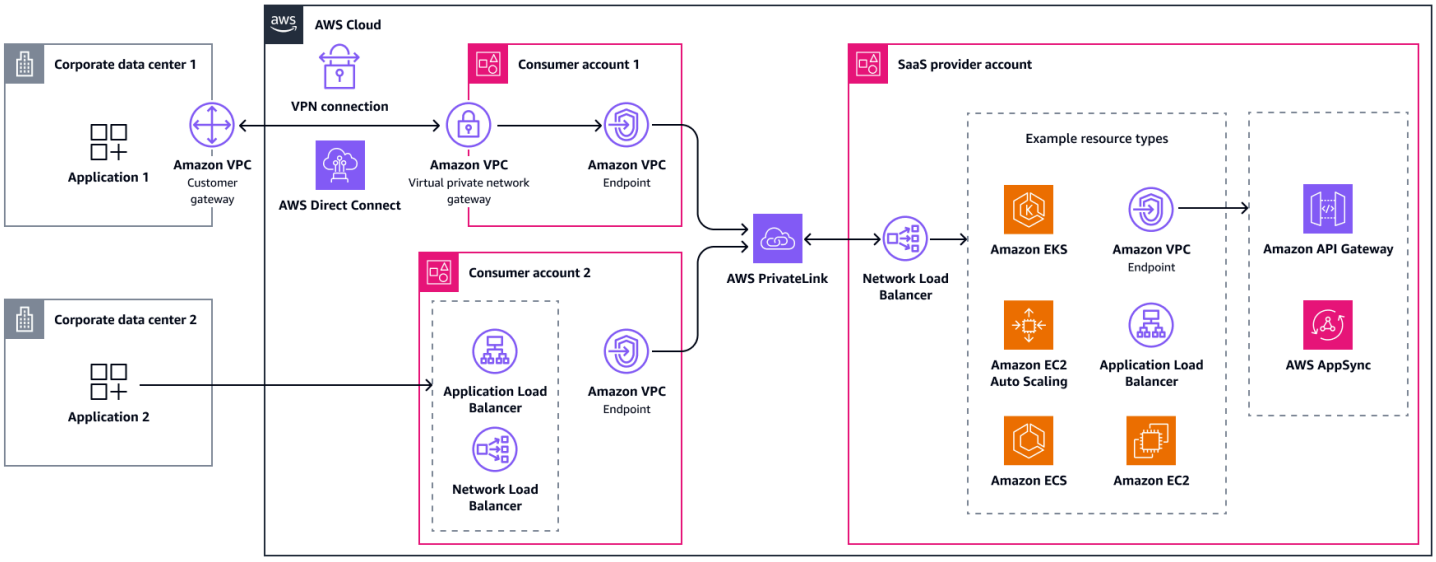
Conexão com uma arquitetura VPC de trânsito

O Transit VPC é uma opção de arquitetura que oferece flexibilidade aos consumidores sobre como se conectar AWS e permite que os provedores de SaaS se beneficiem de ter acesso unificado a seus serviços por meio de. AWS PrivateLink O consumidor se conecta localmente a uma VPC de trânsito que contém somente um ponto de entrada (como um gateway privado virtual) e um endpoint VPC de interface, que é um recurso. AWS PrivateLink O trânsito VPCs deve ser de propriedade do provedor de SaaS ou dos consumidores. Esta seção discute as duas opções.

Você pode criar a VPC e as sub-redes de trânsito com intervalos CIDR compatíveis com o data center local. Se precisarem de conectividade privada, os consumidores podem se conectar a essa VPC por meio AWS Direct Connect de ou. AWS Site-to-Site VPN Você também pode configurar o acesso à conta de trânsito da Internet pública usando um Application Load Balancer ou Network Load Balancer que aponta para o VPC endpoint.

VPC de trânsito gerenciada pelo consumidor

Nessa abordagem, o provedor de SaaS deixa o gerenciamento do trânsito para VPCs os consumidores. Do ponto de vista técnico, a arquitetura do provedor de SaaS é a mesma da conexão direta com os consumidores. Nuvem AWS AWS PrivateLink Do ponto de vista das vendas e do produto, é um esforço adicional, porque alguns consumidores Contas da AWS ainda não o fizeram. Eles podem hesitar em abrir e operar uma conta. O provedor de SaaS deve orientar seus consumidores sobre como criar Contas da AWS e conectar seu data center local. O diagrama a seguir mostra uma combinação de acesso público e privado, em que os consumidores são donos do transporte público VPCs.



Estes são os benefícios desta abordagem:

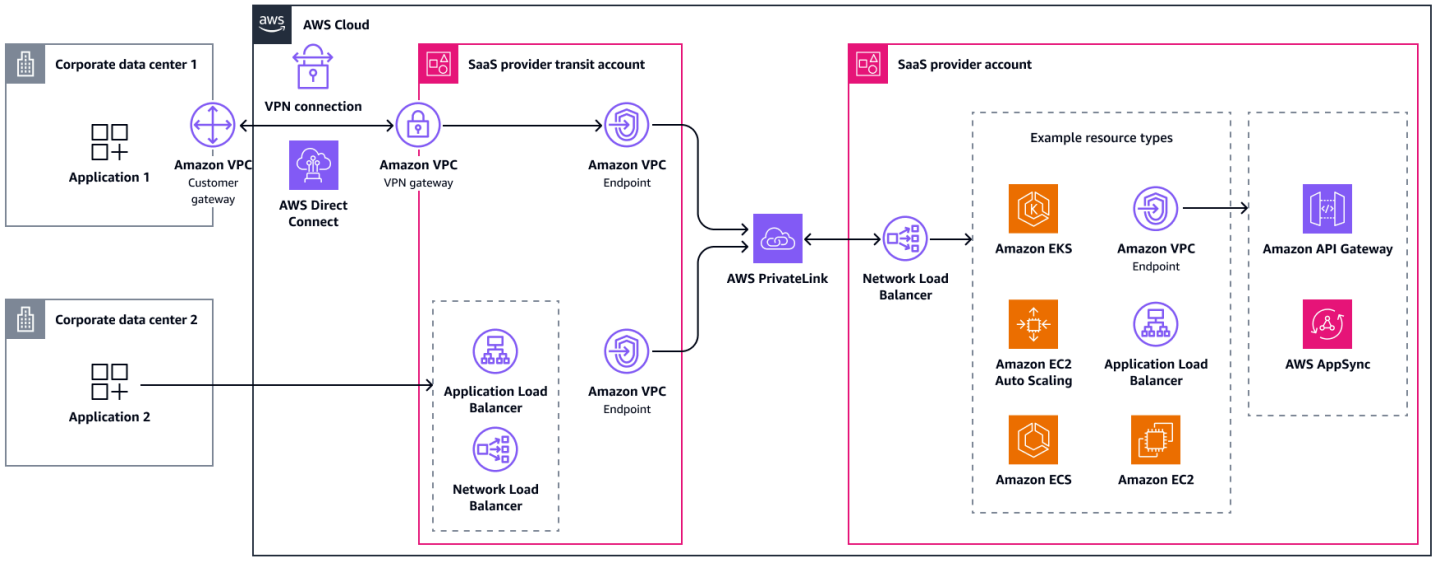
- Tempo de reparo: a sobrecarga operacional é em grande parte transferida para os consumidores de SaaS
- Adaptabilidade: os consumidores de SaaS podem escolher entre diferentes opções de acesso
- Adaptabilidade: Sem conflitos de alcance de CIDR, mesmo ao usar Site-to-Site VPN ou Direct Connect
- Todas as métricas: o provedor de serviços herda os benefícios AWS PrivateLink

A seguir estão as desvantagens dessa abordagem:

- Facilidade de integração: os consumidores de SaaS precisam de pelo menos um Conta da AWS
- TCO: uma VPC de trânsito é uma arquitetura, não um serviço totalmente gerenciado, portanto, exige mais esforço operacional

VPC de trânsito gerenciada pelo provedor

Essa abordagem usa as mesmas tecnologias, mas os limites e as responsabilidades da conta mudam. Aqui, o provedor de SaaS possui o trânsito VPCs, de preferência em uma conta separada da oferta de SaaS. Essa dissociação reduz custos, reduz riscos e permite que a conta de transporte público seja dimensionada de forma independente. Para ambientes que exigem um alto grau de isolamento, você pode criar uma separação adicional entre os locatários usando uma sub-rede ou criando uma VPC de trânsito separada para cada consumidor. Os consumidores podem então escolher como se conectar à VPC de trânsito. Essa abordagem oferece mais opções para expandir o mercado endereçável total, mas tem um TCO mais alto para o provedor de SaaS devido à necessidade de operar e monitorar componentes arquitetônicos adicionais.



Estes são os benefícios desta abordagem:

- Adaptabilidade: os consumidores de SaaS podem escolher entre diferentes opções de acesso
- Adaptabilidade: os consumidores de SaaS não precisam ter um Conta da AWS
- Adaptabilidade: Sem conflitos de alcance de CIDR, mesmo ao usar Site-to-Site VPN ou Direct Connect

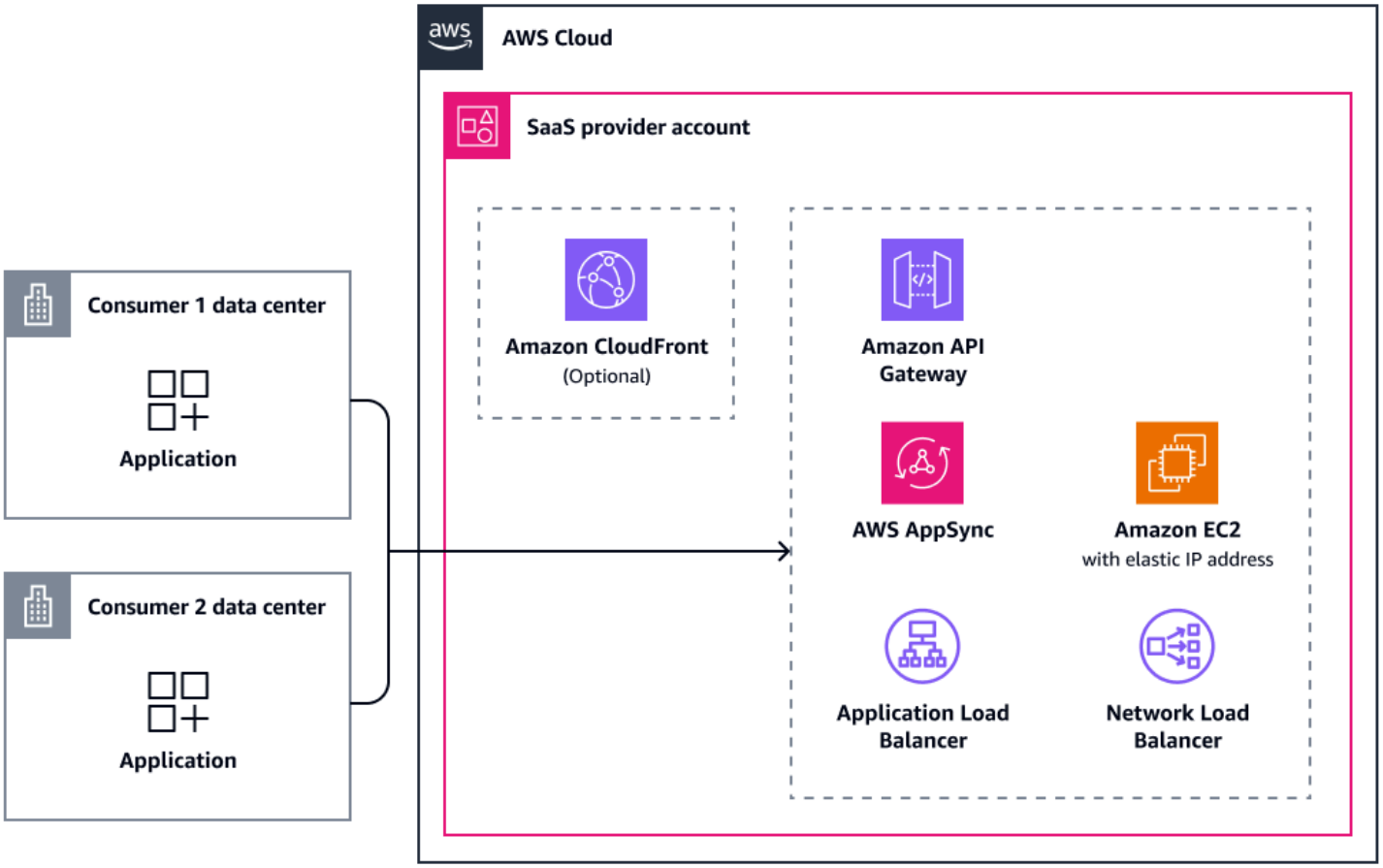
A seguir estão as desvantagens dessa abordagem:

- TCO: uma VPC de trânsito é uma arquitetura, não um serviço totalmente gerenciado, portanto, exige mais esforço operacional
- TCO: o provedor de SaaS precisa operar e monitorar componentes arquitetônicos adicionais

Conectando-se pela Internet pública

O acesso público à Internet também é uma opção válida para fornecer acesso a uma oferta de SaaS, embora não ofereça conectividade privada no sentido tradicional. Alguns consumidores ainda podem preferir uma abordagem de acesso público porque ela não requer infraestrutura de rede adicional entre eles e o provedor de SaaS. Ele reduz a complexidade, o custo e o tempo de integração em troca de uma maior superfície de ataque. Mecanismos fortes de autenticação e autorização podem ajudar a mitigar o aumento do nível de ameaça, e você deve sempre criptografar o tráfego. Ainda é recomendável que você tenha uma camada adicional de segurança nesse cenário, como usando AWS WAF.

A arquitetura nesse cenário é simples. O consumidor se conecta a um host público (o provedor de SaaS) pela Internet. O aplicativo pode ser hospedado diretamente em uma instância pública do Amazon Elastic Compute Cloud (Amazon EC2) com um endereço IP elástico. A opção preferida é hospedá-lo por trás de um Application Load Balancer ou serviço similar. Para melhor desempenho e armazenamento em cache de ativos estáticos, você pode usar uma rede de distribuição de conteúdo, como a Amazon CloudFront. Para servir um aplicativo com latência mínima em dois endereços IP Anycast estáticos globais, você pode colocá-lo AWS Global Acceleratorna frente de uma instância do Amazon EC2, Network Load Balancer ou Application Load Balancer. Além disso CloudFront, os Application Load Balancers e o Amazon API Gateway se integram ao AWS WAF. AWS AppSync
O diagrama a seguir fornece uma visão geral das opções de conectividade de acesso público à Internet.



A tabela a seguir descreve os protocolos e integrações compatíveis com esse cenário.

Serviço ou recurso	IPv6	AWS WAF integração	Pode ser um endpoint do Global Accelerator

Amazon CloudFront	Compatível	Compatível	Não compatível
Amazon API Gateway	Compatível	Compatível	Não compatível
AWS AppSync	Suporte parcial	Compatível	Não compatível
Amazon EC2 com um endereço IP elástico	Compatível	Não compatível	Compatível
Application Load Balancer	Compatível	Compatível	Compatível
Network Load Balancer	Compatível	Não compatível	Compatível

Estes são os benefícios desta abordagem:

- Facilidade de integração: simplicidade e acessibilidade
- Escalabilidade: escala ilimitada
- Adaptabilidade: Nenhum conflito de intervalo CIDR é possível
- Adaptabilidade: suporte CloudFront

A seguir estão as desvantagens dessa abordagem:

- Isolamento de rede: sem conectividade privada
- Isolamento de rede: são necessárias fortes medidas de segurança

Outros benefícios e desvantagens se aplicam, dependendo dos serviços que você escolher.

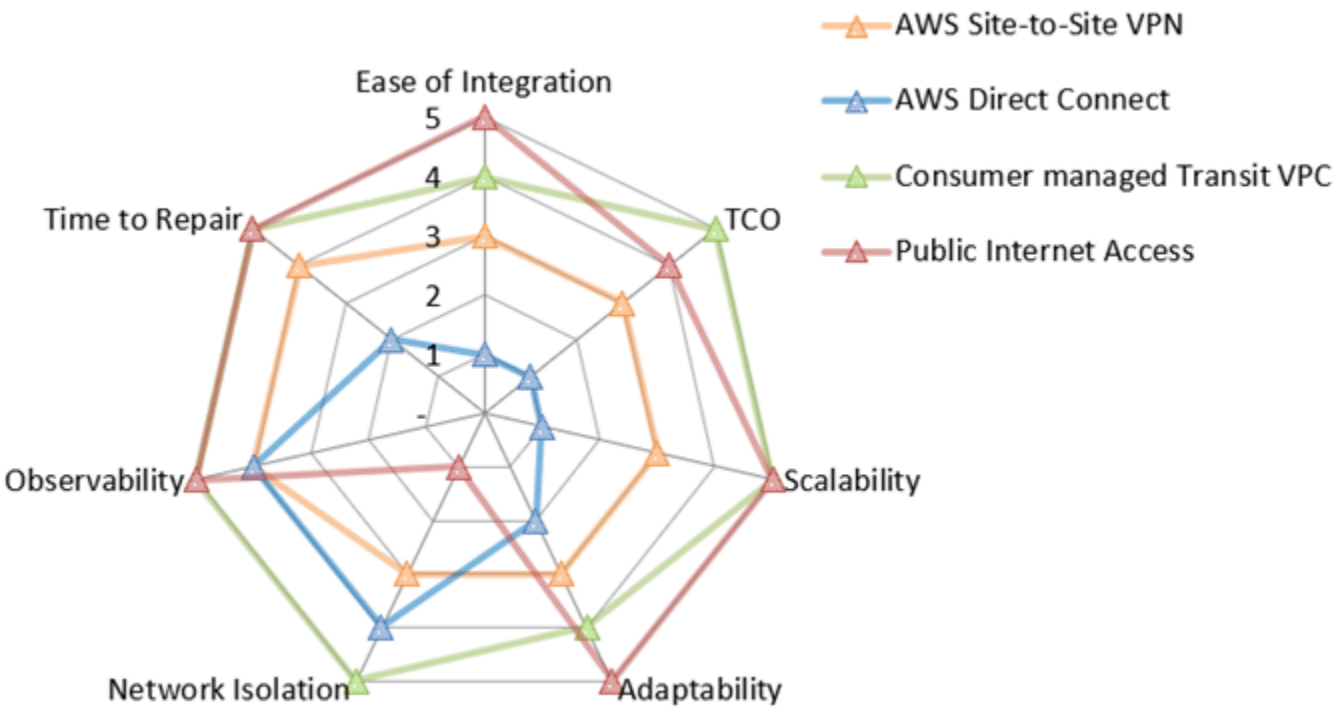
Consumidores de SaaS que operam em outros provedores de serviços em nuvem

Esse cenário descreve soluções para consumidores em outros provedores de serviços em nuvem (CSPs). Esse cenário compartilha alguns pontos em comum com conexões com data centers locais. Na verdade, todas as opções de conectividade para ambientes locais são igualmente válidas para consumidores em outros ambientes CSPs, até mesmo uma conexão privada com alguns AWS Direct

Connect CSPs é possível. A maioria CSPs oferece documentação e suporte sobre como se conectar ao Nuvem AWS por meio de AWS Site-to-Site VPN ou AWS Direct Connect.

Ao escolher a Site-to-Site VPN, os consumidores podem se beneficiar de gateways gerenciados ou recursos similares de seus respectivos CSP. Os consumidores não precisam necessariamente configurá-los sozinhos, como no cenário local. Isso influencia algumas das métricas da Site-to-Site VPN, como melhorias no tempo de reparo e na observabilidade. Isso ocorre porque as duas extremidades da conexão agora são gerenciadas.

O mapa de valores de rede a seguir resume a pontuação de cada uma dessas opções em cada métrica de avaliação. É muito semelhante ao mapa de valores de rede para conexões locais, embora os valores para Site-to-Site VPN sejam diferentes. Para obter mais informações sobre as métricas de avaliação, consulte Métricas de avaliação este guia. No mapa, cinco representa a melhor pontuação, como o menor TCO, o melhor isolamento de rede ou o menor tempo de reparo. Para obter mais informações sobre como ler esse gráfico de radar, consulte Mapa de valores de rede este guia.



O gráfico do radar mostra os seguintes valores.

Métrica de avaliação	AWS Site-to-Site VPN	AWS Direct Connect	VPC de trânsito gerenciada pelo consumidor	Acesso público à internet

Facilidade de integração	3	1	4	5
TCO	3	1	5	4
Escalabilidade	3	1	5	5
Adaptabilidade	3	2	4	5
Isolamento de rede	3	4	5	1
Observabilidade	4	4	5	5
Hora de reparar	4	2	5	5

Oferecendo suporte a ambientes híbridos

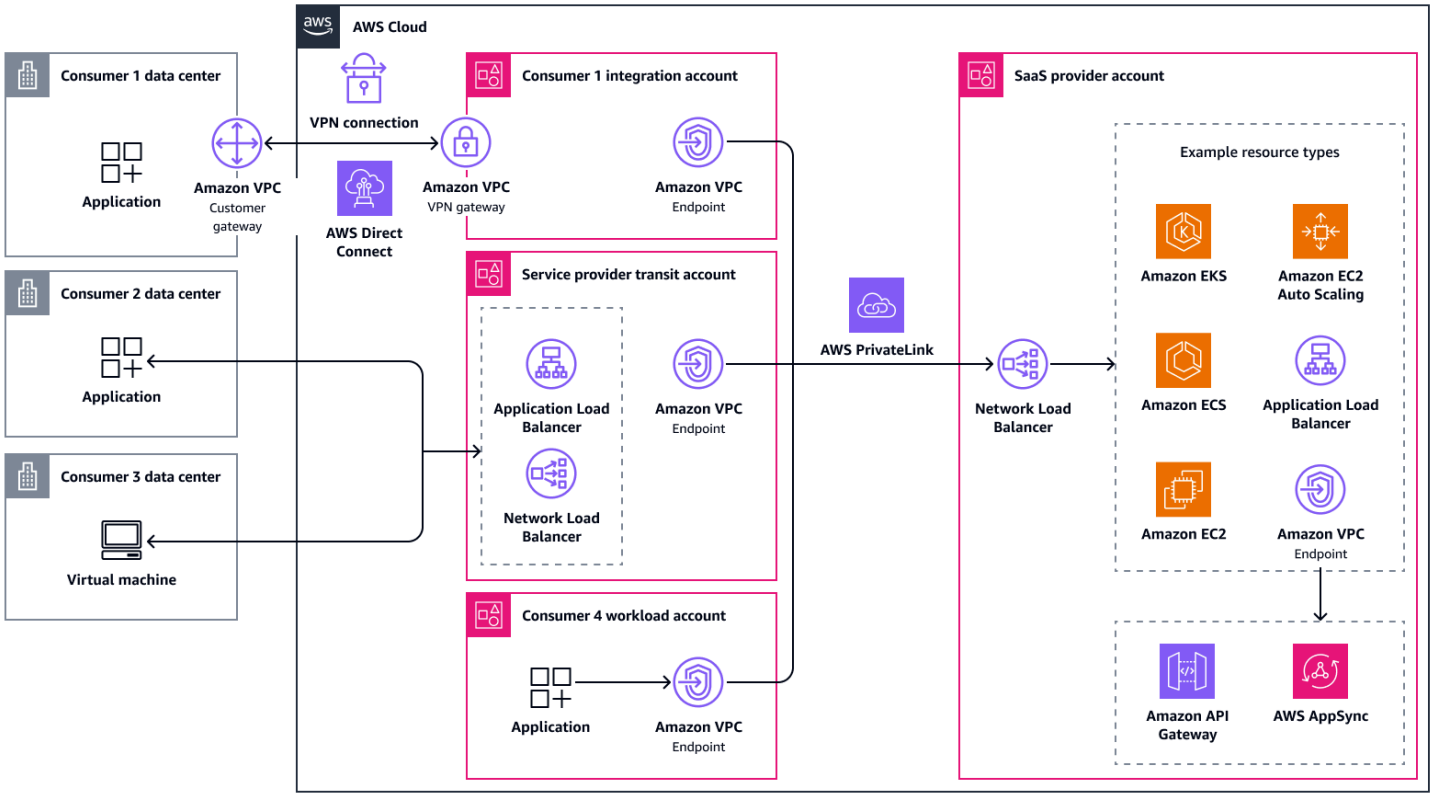
É comum que os consumidores venham de ambientes diferentes, cada um com suas próprias restrições técnicas e de segurança. Alguns clientes podem operar inteiramente em data centers locais que exigem conectividade segura pela Internet ou por meio de links de rede dedicados. Outros podem já estar executando cargas de trabalho internamente AWS e esperar caminhos de rede privada de baixa latência. Um terceiro grupo pode confiar em outro CSPs, onde a conectividade deve unir diferentes redes em nuvem.

Independentemente disso, você deve buscar acesso padronizado à rede ao seu aplicativo SaaS para simplificar sua arquitetura e reduzir a complexidade operacional. Duas das abordagens apresentadas anteriormente — acesso público à Internet e trânsito VPCs — funcionam bem nesses cenários. O acesso público à Internet oferece o caminho de integração mais rápido com configuração mínima para seus clientes. O transporte público VPCs oferece acesso mais controlado e privado, geralmente usando AWS PrivateLink.

Ao projetar sua oferta de SaaS, você pode adotar um único modelo de acesso à rede ou combinar várias abordagens em uma oferta hierárquica. Por exemplo, você pode oferecer uma camada de implantação de acesso público para clientes que priorizam a facilidade de conexão e a rápida integração, e você pode oferecer uma camada de implantação de acesso privado para clientes que têm requisitos rígidos de conformidade ou controle de segurança. Esses níveis vêm com diferentes perfis de custo, desempenho e risco. Também é possível combinar as duas abordagens

em uma única arquitetura. Nesse caso, certifique-se de ter medidas de segurança fortes para que os caminhos públicos e privados permaneçam isolados.

O diagrama a seguir mostra uma abordagem de acesso híbrido, na qual os consumidores têm a opção de se conectar de forma privada a partir de seu data center ou CSP, publicamente ou diretamente AWS PrivateLink (se tiverem cargas de trabalho no). Nuvem AWS



Cenários avançados de acesso à rede para ofertas de SaaS no Nuvem AWS

As arquiteturas discutidas na Cenários de acesso à rede para ofertas de SaaS no Nuvem AWS seção devem ajudá-lo a encontrar uma solução para a maioria dos casos de uso. No entanto, existem alguns cenários que têm requisitos técnicos específicos. Muitos estão além do escopo deste guia.

Esta seção discute os seguintes requisitos e considerações técnicas avançadas:

- Comunicação bidirecional
- TCP, UDP e protocolos proprietários

Comunicação bidirecional

Em alguns casos, os aplicativos exigem tráfego bidirecional para operar conforme o esperado. Casos de uso comuns são webhooks ou serviços de notificação. Geralmente, você pode conseguir isso tendo uma WebSocket conexão entre o servidor e o cliente. Essa conexão mantém a sessão TCP aberta e permite que ambos os participantes enviem tráfego pela conexão. A maioria dos serviços discutidos neste guia oferece suporte nativo WebSocket, incluindo Network Load Balancers, Application Load Balancers, Amazon API Gateway e AWS AppSync (por meio de AWS PrivateLink endpoints privados em tempo real).

Em outros casos, um aplicativo do lado do provedor de SaaS pode precisar acessar recursos do lado do consumidor, como um banco de dados. Quando você se conecta por meio de canais bidirecionais, como uma AWS Site-to-Site VPN conexão, isso não é um problema.

Por outro lado, AWS PrivateLink o Elastic Load Balancing suporta somente tráfego unidirecional. Se você usar esses serviços, deverá configurar outro caminho de rede para o tráfego que começa na sua oferta de SaaS. Por exemplo, isso pode ser uma AWS PrivateLink conexão adicional que vai na direção inversa.

TCP, UDP e protocolos proprietários

Muitos aplicativos são servidos por meio de HTTP ou HTTPS, mas não todos. Alguns podem usar outros protocolos de camada 7 além do TCP, como o Message Queuing Telemetry Support (MQTT).

Outros podem até usar o UDP para atender aos consumidores. Em casos raros, os serviços usam protocolos proprietários que devem ser transmitidos dentro de pacotes (camada 3). Para esses cenários, é importante entender quais serviços oferecem suporte à sua oferta de SaaS.

Para serviços de camada 3, você pode usar AWS PrivateLink balanceadores de carga de rede, ambos compatíveis com todo o tráfego TCP e UDP.

Para serviços de camada 7, os Application Load Balancers e a Amazon CloudFront oferecem suporte a HTTP WebSocket, HTTPS e chamadas de procedimentos remotos do Google (gRPC). Da mesma forma, o Amazon API Gateway e AWS AppSync cada um deles oferecem suporte a HTTP, HTTPS WebSocket e. A Amazon CloudFront é o único serviço que atualmente oferece suporte a HTTP/3.

Você pode usar o Amazon VPC Lattice para conectar aplicativos de camada 7 e recursos de camada 3. Ele suporta passagem HTTP, HTTPS, gRPC, TCP e TLS.

Se o aplicativo puder fornecer tráfego somente na camada 3, é fundamental que você use os principais serviços de AWS rede, como, AWS Transit Gateway AWS Direct Connect AWS Site-to-Site VPN, e emparelhamento de VPC. O tráfego deve então ser roteado diretamente do consumidor de SaaS para a camada computacional da oferta de SaaS.

Antipadrões para acesso à rede no Nuvem AWS

Um antipadrão é uma solução frequentemente usada para um problema recorrente em que a solução é contraproducente, ineficaz ou menos eficaz do que uma alternativa. As opções de design mencionadas nesta seção geralmente funcionam, mas apresentam desvantagens significativas. Se possível, eles devem ser evitados porque existem alternativas melhores.

Esta seção discute os seguintes antipadrões e desafios:

- Incompatibilidade da zona de disponibilidade com AWS PrivateLink
- AWS Site-to-Site VPN conexões entre Contas da AWS

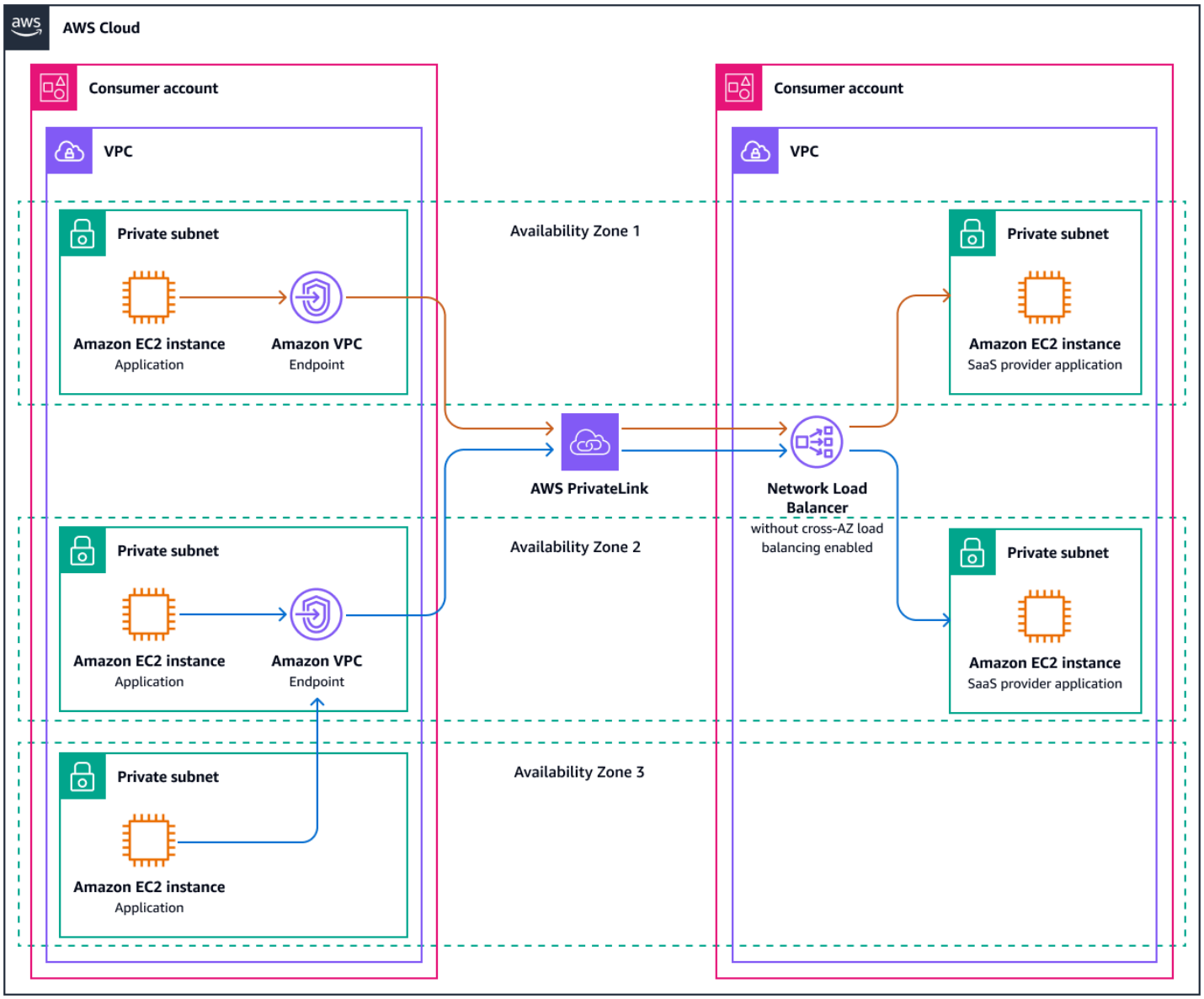
Incompatibilidade da zona de disponibilidade com AWS PrivateLink

Ao fornecer acesso a um aplicativo por meio de AWS PrivateLink, os consumidores de SaaS podem criar endpoints VPC de interface somente nas zonas de disponibilidade em que o aplicativo é implantado. Por exemplo, se o aplicativo for implantado em `use1-az1` `euse1-az2`, o consumidor não poderá implantar um VPC endpoint em. `use1-az3` Recomendamos que você implante a oferta de SaaS em todas as zonas de disponibilidade. A maioria Regiões da AWS tem três zonas de disponibilidade, embora algumas tenham mais. Para obter uma lista abrangente, consulte Regiões e zonas de disponibilidade. Considere o número de zonas de disponibilidade ao escolher uma Região da AWS.

> **Note**
>
> Os nomes das zonas de disponibilidade são diferentes dos da zona de disponibilidade IDs. Para obter mais informações, consulte Zona de disponibilidade IDs para seus AWS recursos.

Se um provedor de SaaS optar por não implantar em todas as zonas de disponibilidade, haverá algumas consequências. Suponha que a oferta de SaaS esteja implantada em `use1-az1` `euse1-az2`, mas o consumidor esteja usando todas as três zonas de disponibilidade, inclusive. `use1-az3` Os endpoints VPC da interface são implantados no lado do consumidor `use1-az1` `euse1-az2`, agora, o aplicativo `use1-az3` precisa acessar um desses endpoints. Em primeiro lugar, o tráfego deve ser permitido das sub-redes nas zonas de disponibilidade incomparáveis para os respectivos VPC endpoints. O consumidor pode decidir usar o nome AWS PrivateLink DNS regional, que pode ser resolvido em qualquer um dos endpoints da VPC e que distribui uniformemente o tráfego entre os

dois. Ou o consumidor pode optar por enviar tráfego diretamente para um endpoint, comouse1-az2. Isso resulta em 67% do tráfego chegando ao lado do provedor use1-az2 e 33% entrando. use1-az1 A figura a seguir mostra esse cenário.



Com um número significativo de consumidores e uma distribuição desigual do tráfego, uma carga de trabalho pode ter problemas de capacidade em uma zona de disponibilidade e estar abaixo da capacidade em outra. Para resolver esse problema, o provedor de SaaS pode decidir balancear uniformemente a carga do tráfego do seu lado, habilitando o balanceamento de carga entre zonas no Network Load Balancer. Isso acarreta custos adicionais.

Se apenas uma zona de disponibilidade for correspondida pelo provedor de serviços, todo o tráfego entrará em um único endpoint. Isso cria um desequilíbrio ainda maior. Como resultado, a oferta de

SaaS não está mais altamente disponível para o consumidor. Não importa para o consumidor se o aplicativo é servido em zonas de disponibilidade adicionais que ele mesmo não está usando. Na pior das hipóteses, um provedor de SaaS pode não ser capaz de atender um consumidor que não usa nenhuma das mesmas zonas de disponibilidade.

No caso raro de não haver uma opção viável para o provedor de SaaS provisionar seu aplicativo em todas as zonas de disponibilidade, também é possível criar uma sub-rede somente nas zonas de disponibilidade ausentes e, em seguida, estender o serviço para essas zonas de disponibilidade vazias. O balanceamento de carga entre zonas pode então distribuir o tráfego de entrada pelos endpoints reais do aplicativo nas outras zonas de disponibilidade.

AWS Site-to-Site VPN conexões entre Contas da AWS

Às vezes, as empresas que migram de ambientes locais para a nuvem tentam elevar e deslocar toda a rede. Isso pode causar problemas porque há diferenças significativas entre as práticas de rede local e na nuvem. Se essa mudança de mentalidade não acontecer, coisas como AWS Site-to-Site VPN conexões de uma VPC para outra VPC podem acontecer. Essa abordagem não tira proveito dos serviços de rede desenvolvidos especificamente no Nuvem AWS, que simplificam o gerenciamento e melhoram o desempenho. A adaptação aos designs nativos da nuvem ajuda a reduzir a sobrecarga operacional e resulta em uma conectividade mais confiável e escalável entre eles. VPCs

Se você está pensando em fornecer essa opção de conectividade como provedor de SaaS, pergunte a si mesmo ou ao consumidor por que ela AWS Site-to-Site VPN deve ser usada. Em seguida, retroceda a partir desses requisitos para encontrar uma melhor opção de conectividade. A seção Comparando os recursos do serviço deste guia contém uma matriz que você pode usar para ajudar a identificar opções. Em seguida, você pode examinar as seções relevantes deste guia para encontrar uma abordagem arquitetônica que aborde seu caso de uso.

Próximas etapas

Este guia descreveu várias abordagens de acesso à rede em diferentes cenários e descreve as vantagens e desvantagens de cada arquitetura. Você deve entender por que escolher uma abordagem de acesso à rede não deve ser uma discussão puramente tecnológica. O alinhamento entre negócios e tecnologia é essencial. As próximas etapas e recomendações a seguir podem ajudá-lo a avaliar e padronizar sua estratégia de arquitetura de rede avaliando os recursos atuais, analisando as necessidades do mercado e implementando controles de governança.

Esta seção contém os seguintes tópicos:

- Avaliando a arquitetura e os recursos atuais
- Análise de mercado e clientes
- Alinhamento estratégico
- Padronização
- Governança
- Repetição

Avaliando a arquitetura e os recursos atuais

Analise a arquitetura de rede atual em relação às fontes de dados relevantes, como a estrutura de autoavaliação deste guia, os requisitos regulatórios atuais e o estado atual do mercado (tanto em termos de seu cliente quanto de uma análise competitiva). Por exemplo, considere usar o AWS Well-Architected Framework, que se baseia em décadas de experiência na execução de sistemas de produção em grande escala no. Nuvem AWS

Analise todas as possíveis exceções, decisões pontuais e históricas do produto. Seja curioso, desafie-os e não assuma automaticamente sua validade. Os requisitos do cliente de anos atrás podem não ser mais válidos. Suposições desafiadoras criam oportunidades para simplificar e reduzir a complexidade de sua arquitetura.

Em termos simples, documente as observações para que elas possam ser acessadas e compreendidas por diversas funções em sua organização. Capture onde o estado atual difere do estado alvo, qual é o estado alvo, o impacto e quando as observações foram feitas. O registro dessas informações ajuda suas organizações a tomar decisões com base em dados atualizados.

Análise de mercado e clientes

Reúna informações sobre as tendências do mercado. Atualmente, qual é a forma preferida dos consumidores de acessar ofertas de SaaS como a sua? Você ainda está atendendo seus clientes onde eles estão? As coortes ou o comportamento dos clientes mudaram? Seus executivos direcionaram o navio para um novo mercado, uma geografia com requisitos regulatórios específicos ou um novo nível de clientes? Seu negócio ou modelo operacional mudou? Por exemplo, você está pensando em colocar uma etiqueta branca em seus serviços? Seu plano de crescimento inclui trabalhar com parceiros para que seu serviço esteja disponível para os clientes quando eles se conectarem com esses parceiros?

Alinhamento estratégico

Quando você entender suas capacidades atuais, arquitetura, mercado e clientes atuais, convoque uma reunião de alinhamento estratégico. Com as partes interessadas relevantes de produtos, negócios e tecnologia, desafie quais requisitos ainda são válidos e quais novos requisitos precisam ser considerados. Encontre oportunidades para reduzir a complexidade eliminando requisitos que não são mais necessários. Isso não é um projeto feito por um comitê; a equipe de engenharia precisa preparar e possuir a arquitetura real e os detalhes da implementação. No entanto, essa reunião deve esclarecer por que esse é o conjunto de requisitos que maximiza os benefícios para seus clientes e sua organização.

Padronização

Para atrair clientes, pode ser tentador deixar que cada um escolha livremente como se conectar ao seu serviço. Afinal, qualquer solução pode funcionar tecnicamente, e você também pode ter o conhecimento e os recursos para gerenciar e operar todas elas. Isso pode funcionar bem até certo ponto, mas à medida que sua empresa cresce, fica difícil gerenciá-la. Sua pilha de observabilidade precisa suportar métricas de várias soluções, e os engenheiros de confiabilidade do seu site também precisam ser capazes de entendê-las. Você precisa de up-to-date documentação para cada abordagem de conectividade. As principais mudanças em seu aplicativo precisam ser avaliadas em relação a cada abordagem de acesso que você está oferecendo. Você precisa escrever e manter automações e infraestrutura como código (IaC) para cada abordagem de acesso. A sobrecarga adicional de não padronizar o acesso ao seu serviço deve ser comparada à flexibilidade que você deseja oferecer aos seus clientes.

Se você precisar de uma estrela do norte para orientar sua tomada de decisão, sugerimos a padronização. A padronização de como seus clientes interagem com os serviços que você fornece é normalmente a ação mais impactante que você pode tomar para melhorar muitas métricas de sucesso em sua organização. A padronização torna mais fácil para as equipes de produto entenderem a estrutura de custos de seus serviços e tomarem decisões de produto baseadas em dados. É mais fácil para as equipes de operações solucionar problemas e automatizar partes do processo de solução de problemas em um ambiente desenvolvido, implementado e operado de acordo com padrões predefinidos. Ele pode ajudá-lo a detectar anomalias, comportamentos inesperados ou ações de um agente mal-intencionado. A padronização também reduz a dívida técnica. São necessários menos ciclos para que as equipes de engenharia testem e implementem as mudanças na produção. Também pode aumentar sua velocidade de entrada no mercado, melhorar o sucesso da integração de autoatendimento e reduzir o risco regulatório.

Portanto, sugerimos que você também analise quaisquer itens pontuais que possam estar em vigor hoje. Quantifique o número de ciclos operacionais que você gasta dando suporte aos clientes existentes. Compare seus resultados com dados históricos e avalie se sua abordagem atual se expande para os próximos anos. Sempre que houver necessidade de se desviar dos padrões, desafie os requisitos por trás dessas solicitações. Avalie o impacto e equilibre os benefícios imediatos com os compromissos de longo prazo.

Nos casos em que a personalização é inevitável, mas está em conflito com seus padrões, considere um modelo de responsabilidade compartilhada. Nesse modelo, seus produtos são amplamente protegidos das alterações solicitadas e a personalização acontece em um ambiente minimalista e dedicado. Para ver um exemplo, consulte a Conexão com uma arquitetura VPC de trânsito seção.

Governança

Para a conformidade com os requisitos regulatórios e com seus próprios padrões internos, a governança é essencial. Com a governança adequada, você pode controlar onde e como aplicar os padrões. Você também estabelece dois controles para detectar divergências em relação aos padrões e informar os proprietários dos recursos sobre as ações corretivas necessárias. AWS Organizations, AWS Config, AWS CloudTrail, e AWS Control Towersão algumas das muitas Serviços da AWS que podem ajudá-lo a gerenciar e controlar suas cargas de trabalho no. Nuvem AWS

Repetição

Usando o que aprendeu com seus esforços iniciais, configure um processo leve e repetível para se manter alinhado no futuro. Defina de quais funções você precisa de informações, com que

frequência, quão precisos os dados precisam ser, como os dados serão compartilhados e quem agirá com base neles.

Recursos

AWS documentação

- Integração de serviços de terceiros na Nuvem AWS(OrientaçãoAWS prescritiva)
- Autorização de SaaS multilocatário e controle de acesso à API (orientação prescritiva)AWS
- Gerencie locatários em vários produtos SaaS em um único plano de controleAWS (orientação prescritiva)
- O que AWS Direct Connecté (Direct Connect documentação)
- O que é o AWS PrivateLink? (Documentação da Amazon VPC)
- O que AWS Site-to-Site VPNé (AWS Site-to-Site VPN documentação)
- O que AWS Transit Gatewayé (Documentação da Amazon VPC)
- O que é peering de VPC? (Documentação da Amazon VPC)

Outros AWS recursos

- Opções de conectividade da Amazon Virtual Private Cloud (AWS Whitepaper)
- AWS re:Invent 2021 - Como escolher o balanceador de carga certo para suas cargas de trabalho () AWS YouTube
- O que é SaaS? (AWS site)
- AWS Programa SaaS Factory (programa)AWS Partner
- Orientação para arquiteturas multilocatárias na AWS (AWS Biblioteca de soluções)

Histórico do documento

A tabela a seguir descreve alterações significativas feitas neste guia. Se desejar receber notificações sobre futuras atualizações, inscreva-se em um feed RSS.

Alteração	Descrição	Data
Publicação inicial	—	12 de setembro de 2025

AWS Glossário de orientação prescritiva

A seguir estão os termos comumente usados em estratégias, guias e padrões fornecidos pela Orientação AWS Prescritiva. Para sugerir entradas, use o link Fornecer feedback no final do glossário.

Números

7 Rs

Sete estratégias comuns de migração para mover aplicações para a nuvem. Essas estratégias baseiam-se nos 5 Rs identificados pela Gartner em 2011 e consistem em:

- Refactor/re-architect — mova um aplicativo e modifique sua arquitetura aproveitando ao máximo os recursos nativos da nuvem para melhorar a agilidade, o desempenho e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migre seu banco de dados Oracle local para a Amazon PostgreSQL-Compatible Aurora Edition.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape]):: mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Amazon Relational Database Service (Amazon RDS) para Oracle na Nuvem AWS.
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: Migre seu sistema de gerenciamento de relacionamento com o cliente (CRM) para o. Salesforce.com
- Redefinir a hospedagem (mover sem alterações [lift-and-shift]): mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: migrar seu banco de dados Oracle on-premises para o Oracle em uma instância do EC2 na Nuvem AWS.
- Realocar (mover o hipervisor sem alterações [hypervisor-level lift-and-shift]): mover a infraestrutura para a nuvem sem comprar novo hardware, reescrever aplicações ou modificar suas operações existentes. Você migra servidores de uma plataforma on-premises para um serviço de nuvem para a mesma plataforma. Exemplo: Migrar um Microsoft Hyper-V aplicativo para o. AWS
- Reter (revisitar): mantenha as aplicações em seu ambiente de origem. Isso pode incluir aplicações que exigem grande refatoração, e você deseja adiar esse trabalho para um

momento posterior, e aplicações antigas que você deseja manter porque não há justificativa comercial para migrá-las.

- Retirar: desative ou remova aplicações que não são mais necessárias em seu ambiente de origem.

A

A2A () Agent-to-Agent

Um protocolo com estado para colaboração entre agentes, apoiando a delegação de tarefas e a transferência de estados.

ABAC

Consulte controle de acesso baseado em atributo.

serviços abstraídos

Veja serviços gerenciados.

ACID

Veja atomicidade, consistência, isolamento, durabilidade.

migração ativa-ativa

Um método de migração de banco de dados no qual os bancos de dados de origem e de destino são mantidos em sincronia (por meio de uma ferramenta de replicação bidirecional ou operações de gravação dupla), e ambos os bancos de dados lidam com transações de aplicações conectadas durante a migração. Esse método oferece suporte à migração em lotes pequenos e controlados, em vez de exigir uma substituição única. É mais flexível, mas exige mais trabalho do que a migração ativa-passiva.

migração ativa-passiva

Um método de migração de banco de dados em que os bancos de dados de origem e de destino são mantidos em sincronia, mas somente o banco de dados de origem manipula as transações das aplicações conectadas, enquanto os dados são replicados no banco de dados de destino. O banco de dados de destino não aceita nenhuma transação durante a migração.

Agente

Um sistema de IA que pode raciocinar, planejar e realizar ações de forma autônoma usando ferramentas para atingir metas.

Agente Ops

Práticas operacionais para criar, testar, implantar e executar agentes de IA na produção em grande escala.

AGGREGATE FUNCTION

Uma função SQL que opera em um grupo de linhas e calcula um único valor de retorno para o grupo. Exemplos de funções agregadas incluem `SUM` e `MAX`.

AI

Veja inteligência artificial.

AIOps

Veja operações de inteligência artificial.

anonimização

O processo de excluir permanentemente informações pessoais em um conjunto de dados. A anonimização pode ajudar a proteger a privacidade pessoal. Dados anônimos não são mais considerados dados pessoais.

antipadrões

Uma solução frequentemente usada para um problema recorrente em que a solução é contraproducente, ineficaz ou menos eficaz do que uma alternativa.

controle de aplicações

Uma abordagem de segurança que permite o uso somente de aplicações aprovadas para ajudar a proteger um sistema contra malware.

portfólio de aplicações

Uma coleção de informações detalhadas sobre cada aplicação usada por uma organização, incluindo o custo para criar e manter a aplicação e seu valor comercial. Essas informações são fundamentais para o processo de descoberta e análise de portfólio e ajudam a identificar e priorizar as aplicações a serem migradas, modernizadas e otimizadas.

inteligência artificial (IA)

O campo da ciência da computação que se dedica ao uso de tecnologias de computação para desempenhar funções cognitivas normalmente associadas aos humanos, como aprender, resolver problemas e reconhecer padrões. Para obter mais informações, consulte O que é inteligência artificial?

operações de inteligência artificial (AIOps)

O processo de usar técnicas de machine learning para resolver problemas operacionais, reduzir incidentes operacionais e intervenção humana e aumentar a qualidade do serviço. Para obter mais informações sobre como as AIOps são usadas na estratégia de migração para a AWS , consulte o guia de integração de operações.

criptografia assimétrica

Um algoritmo de criptografia que usa um par de chaves, uma chave pública para criptografia e uma chave privada para descriptografia. É possível compartilhar a chave pública porque ela não é usada na descriptografia, mas o acesso à chave privada deve ser altamente restrito.

atomicidade, consistência, isolamento, durabilidade (ACID)

Um conjunto de propriedades de software que garantem a validade dos dados e a confiabilidade operacional de um banco de dados, mesmo no caso de erros, falhas de energia ou outros problemas.

controle de acesso por atributo (ABAC)

A prática de criar permissões minuciosas com base nos atributos do usuário, como departamento, cargo e nome da equipe. Para obter mais informações, consulte ABAC AWS na documentação AWS Identity and Access Management (IAM).

fonte de dados autorizada

Um local onde você armazena a versão principal dos dados, que é considerada a fonte de informações mais confiável. Você pode copiar dados da fonte de dados autorizada para outros locais com o objetivo de processar ou modificar os dados, como anonimizá-los, redigi-los ou pseudonimizá-los.

Zona de disponibilidade

Um local distinto dentro de um Região da AWS que está isolado de falhas em outras zonas de disponibilidade e fornece conectividade de rede barata e de baixa latência a outras zonas de disponibilidade na mesma região.

AWS Estrutura de adoção da nuvem (AWS CAF)

Uma estrutura de diretrizes e melhores práticas AWS para ajudar as organizações a desenvolver um plano eficiente e eficaz para migrar com sucesso para a nuvem. AWS O CAF organiza a orientação em seis áreas de foco chamadas perspectivas: negócios, pessoas, governança, plataforma, segurança e operações. As perspectivas de negócios, pessoas e governança têm

como foco habilidades e processos de negócios; as perspectivas de plataforma, segurança e operações concentram-se em habilidades e processos técnicos. Por exemplo, a perspectiva das pessoas tem como alvo as partes interessadas que lidam com recursos humanos (RH), funções de pessoal e gerenciamento de pessoal. Nessa perspectiva, o AWS CAF fornece orientação para desenvolvimento, treinamento e comunicação de pessoas para ajudar a preparar a organização para a adoção bem-sucedida da nuvem. Para obter mais informações, consulte o site da AWS CAF e o whitepaper da AWS CAF.

AWS Estrutura de qualificação da carga de trabalho (AWS WQF)

Uma ferramenta que avalia as cargas de trabalho de migração do banco de dados, recomenda estratégias de migração e fornece estimativas de trabalho. AWS O WQF está incluído com AWS Schema Conversion Tool ()AWS SCT. Ela analisa esquemas de banco de dados e objetos de código, código de aplicações, dependências e características de performance, além de fornecer relatórios de avaliação.

B

bot malicioso

Um bot destinado a causar disrupção ou danos a indivíduos ou organizações.

BCP

Veja planejamento de continuidade de negócios

gráfico de comportamento

Uma visualização unificada e interativa do comportamento e das interações de recursos ao longo do tempo. É possível usar um gráfico de comportamento com o Amazon Detective para examinar tentativas de login malsucedidas, chamadas de API suspeitas e ações similares. Para obter mais informações, consulte Dados em um gráfico de comportamento na documentação do Detective.

sistema big-endian

Um sistema que armazena o byte mais significativo antes. Veja também endianness.

classificação binária

Um processo que prevê um resultado binário (uma de duas classes possíveis). Por exemplo, seu modelo de ML pode precisar prever problemas como “Este e-mail é ou não é spam?” ou “Este produto é um livro ou um carro?”

filtro de bloom

Uma estrutura de dados probabilística e eficiente em termos de memória que é usada para testar se um elemento é membro de um conjunto.

blue/green implantação

Uma estratégia de implantação em que você cria dois ambientes separados, mas idênticos. Você executa a versão atual da aplicação em um ambiente (azul) e a nova versão da aplicação no outro ambiente (verde). Essa estratégia ajuda você a reverter rapidamente com o mínimo de impacto.

bot

Uma aplicação de software que executa tarefas automatizadas na internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como crawlers da web que indexam informações na internet. Outros bots, conhecidos como bots maliciosos, têm como objetivo causar disrupção ou danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como bot herder ou operador de bots. Os botnets são o mecanismo mais conhecido para escalar bots e seu impacto.

ramo

Uma área contida de um repositório de código. A primeira ramificação criada em um repositório é a ramificação principal. Você pode criar uma nova ramificação a partir de uma ramificação existente e, em seguida, desenvolver recursos ou corrigir bugs na nova ramificação. Uma ramificação que você cria para gerar um recurso é comumente chamada de ramificação de recurso. Quando o recurso estiver pronto para lançamento, você mesclará a ramificação do recurso de volta com a ramificação principal. Para obter mais informações, consulte Sobre filiais (GitHub documentação).

Acesso de emergência

Em circunstâncias excepcionais e por meio de um processo aprovado, um meio rápido para um usuário obter acesso a um Conta da AWS que ele normalmente não tem permissão para acessar. Para obter mais informações, consulte o indicador Implementar procedimentos de quebra de vidros na AWS Well-Architected orientação.

estratégia brownfield

A infraestrutura existente em seu ambiente. Ao adotar uma estratégia brownfield para uma arquitetura de sistema, você desenvolve a arquitetura de acordo com as restrições dos sistemas e da infraestrutura atuais. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

cache do buffer

A área da memória em que os dados acessados com mais frequência são armazenados.

capacidade de negócios

O que uma empresa faz para gerar valor (por exemplo, vendas, atendimento ao cliente ou marketing). As arquiteturas de microsserviços e as decisões de desenvolvimento podem ser orientadas por recursos de negócios. Para obter mais informações, consulte a seção Organizados de acordo com as capacidades de negócios do whitepaper Executar microsserviços conteinerizados na AWS.

planejamento de continuidade de negócios (BCP)

Um plano que aborda o impacto potencial de um evento disruptivo, como uma migração em grande escala, nas operações e permite que uma empresa retome as operações rapidamente.

C

CAF

Veja AWS Cloud Adoption Framework.

implantação canário

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substitui a versão atual por completo.

CCoE

Veja Centro de Excelência da Nuvem.

CDC

Veja captura de dados de alteração.

captura de dados de alterações (CDC)

O processo de rastrear alterações em uma fonte de dados, como uma tabela de banco de dados, e registrar metadados sobre a alteração. É possível usar o CDC para várias finalidades, como auditar ou replicar alterações em um sistema de destino para manter a sincronização.

engenharia do caos

Introduzir intencionalmente falhas ou eventos disruptivos para testar a resiliência de um sistema. Você pode usar AWS Fault Injection Service (AWS FIS) para realizar experimentos que estressam suas AWS cargas de trabalho e avaliar sua resposta.

CI/CD

Veja integração e entrega contínuas.

classificação

Um processo de categorização que ajuda a gerar previsões. Os modelos de ML para problemas de classificação predizem um valor discreto. Os valores discretos são sempre diferentes uns dos outros. Por exemplo, um modelo pode precisar avaliar se há ou não um carro em uma imagem.

Desenvolvedor cidadão

Um usuário corporativo que cria aplicativos de IA usando plataformas sem code/low código sem habilidades técnicas especializadas.

criptografia no lado do cliente

Criptografia de dados localmente, antes que o alvo os AWS service (Serviço da AWS) receba.

Centro de Excelência da Nuvem (CCoE)

Uma equipe multidisciplinar que impulsiona os esforços de adoção da nuvem em toda a organização, incluindo o desenvolvimento de práticas recomendas de nuvem, a mobilização de recursos, o estabelecimento de cronogramas de migração e a liderança da organização em transformações em grande escala. Para obter mais informações, consulte as postagens do CCoE no blog de estratégia Nuvem AWS corporativa.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem é normalmente conectada à tecnologia de computação de borda.

modelo operacional em nuvem

Em uma organização de TI, o modelo operacional usado para criar, amadurecer e otimizar um ou mais ambientes de nuvem. Para obter mais informações, consulte Criar seu modelo operacional de nuvem.

estágios de adoção da nuvem

As quatro fases pelas quais as organizações normalmente passam ao migrar para a Nuvem AWS:

- Projeto: executar alguns projetos relacionados à nuvem para fins de prova de conceito e aprendizado
- Fundação: realizar investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma zona de pouso, definir um CCoE, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Re-invention — Otimizando produtos e serviços e inovando na nuvem

Esses estágios foram definidos por Stephen Orban na postagem do blog The Journey Toward Cloud-First & the Stages of Adoption no blog Nuvem AWS Enterprise Strategy. Para obter informações sobre como eles se relacionam com a estratégia de AWS migração, consulte o guia de preparação para migração.

CMDB

Veja banco de dados de gerenciamento de configuração.

repositório de código

Um local onde o código-fonte e outros ativos, como documentação, amostras e scripts, são armazenados e atualizados por meio de processos de controle de versão. Os repositórios de nuvem comuns incluem o GitHub ou o Bitbucket Cloud. Cada versão do código é chamada de ramificação. Em uma estrutura de microsserviços, cada repositório é dedicado a uma única peça de funcionalidade. Um único CI/CD pipeline pode usar vários repositórios.

cache frio

Um cache de buffer que está vazio, não está bem preenchido ou contém dados obsoletos ou irrelevantes. Isso afeta a performance porque a instância do banco de dados deve ler da memória principal ou do disco, um processo que é mais lento do que a leitura do cache do buffer.

dados frios

Dados que raramente são acessados e geralmente são históricos. Ao consultar esse tipo de dados, consultas lentas geralmente são aceitáveis. Mover esses dados para níveis ou classes de armazenamento de baixo desempenho e menos caros pode reduzir os custos.

visão computacional (CV)

Um campo de IA que usa machine learning para analisar e extrair informações de formatos visuais, como vídeos e imagens digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Em uma workload, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a workload se torne incompatível e, normalmente, é gradual e não intencional.

banco de dados de gerenciamento de configuração (CMDB)

Um repositório que armazena e gerencia informações sobre um banco de dados e seu ambiente de TI, incluindo componentes de hardware e software e suas configurações. Normalmente, os dados de um CMDB são usados no estágio de descoberta e análise do portfólio da migração.

pacote de conformidade

Uma coleção de AWS Config regras e ações de remediação que você pode montar para personalizar suas verificações de conformidade e segurança. Você pode implantar um pacote de conformidade como uma entidade única em uma Conta da AWS região ou em uma organização usando um modelo YAML. Para obter mais informações, consulte Pacotes de conformidade na documentação. AWS Config

integração contínua e entrega contínua (CI/CD)

O processo de automatizar os estágios de origem, criação, teste, preparação e produção do processo de lançamento do software. CI/CD é comumente descrito como um pipeline. CI/CD pode ajudá-lo a automatizar processos, melhorar a produtividade, melhorar a qualidade do código e entregar com mais rapidez. Para obter mais informações, consulte Benefícios da entrega contínua. CD também pode significar implantação contínua. Para obter mais informações, consulte Entrega contínua versus implantação contínua.

CV

Veja visão computacional.

D

dados em repouso

Dados estacionários em sua rede, por exemplo, dados que estão em um armazenamento.

classificação de dados

Um processo para identificar e categorizar os dados em sua rede com base em criticalidade e confidencialidade. É um componente crítico de qualquer estratégia de gerenciamento de riscos de segurança cibernética, pois ajuda a determinar os controles adequados de proteção e retenção para os dados. A classificação de dados é um componente do pilar de segurança na AWS Well-Architected Estrutura. Para obter mais informações, consulte Classificação de dados.

desvio de dados

Uma variação significativa entre os dados de produção e os dados usados para treinar um modelo de ML ou uma alteração significativa nos dados de entrada ao longo do tempo. O desvio de dados pode reduzir a qualidade geral, a precisão e a imparcialidade das previsões do modelo de ML.

dados em trânsito

Dados que estão se movendo ativamente pela sua rede, como entre os recursos da rede.

data mesh

Um framework de arquitetura que fornece propriedade de dados distribuída e descentralizada com gerenciamento e governança centralizados.

minimização de dados

O princípio de coletar e processar apenas os dados estritamente necessários. Praticar a minimização de dados no Nuvem AWS pode reduzir os riscos de privacidade, os custos e a pegada de carbono de sua análise.

perímetro de dados

Um conjunto de proteções preventivas em seu AWS ambiente que ajudam a garantir que somente identidades confiáveis acessem recursos confiáveis das redes esperadas. Para obter mais informações, consulte Construindo um perímetro de dados em. AWS

pré-processamento de dados

A transformação de dados brutos em um formato que seja facilmente analisado por seu modelo de ML. O pré-processamento de dados pode significar a remoção de determinadas colunas ou linhas e o tratamento de valores ausentes, inconsistentes ou duplicados.

proveniência dos dados

O processo de rastrear a origem e o histórico dos dados ao longo de seu ciclo de vida, por exemplo, como os dados foram gerados, transmitidos e armazenados.

titular dos dados

Um indivíduo cujos dados estão sendo coletados e processados.

data warehouse

Um sistema de gerenciamento de dados compatível com business intelligence, como analytics. Os data warehouses geralmente contêm grandes quantidades de dados históricos e geralmente são usados para consultas e análises.

linguagem de definição de dados (DDL)

Instruções ou comandos para criar ou modificar a estrutura de tabelas e objetos em um banco de dados.

linguagem de manipulação de dados (DML)

Instruções ou comandos para modificar (inserir, atualizar e excluir) informações em um banco de dados.

DDL

Veja linguagem de definição de banco de dados.

deep ensemble

A combinação de vários modelos de aprendizado profundo para gerar previsões. Os deep ensembles podem ser usados para produzir uma previsão mais precisa ou para estimar a incerteza nas previsões.

Aprendizado profundo

Um subcampo do ML que usa várias camadas de redes neurais artificiais para identificar o mapeamento entre os dados de entrada e as variáveis-alvo de interesse.

defesa completa

Uma abordagem de segurança da informação na qual uma série de mecanismos e controles de segurança são cuidadosamente distribuídos por toda a rede de computadores para proteger a confidencialidade, a integridade e a disponibilidade da rede e dos dados nela contidos. Ao adotar essa estratégia AWS, você adiciona vários controles em diferentes camadas da AWS Organizations estrutura para ajudar a proteger os recursos. Por exemplo, uma abordagem de defesa aprofundada pode combinar autenticação multifatorial, segmentação de rede e criptografia.

administrador delegado

Em AWS Organizations, um serviço compatível pode registrar uma conta de AWS membro para administrar as contas da organização e gerenciar as permissões desse serviço. Essa conta é chamada de administrador delegado para esse serviço Para obter mais informações e uma lista de serviços compatíveis, consulte Serviços que funcionam com o AWS Organizations na documentação do AWS Organizations .

implantação

O processo de criar uma aplicação, novos recursos ou correções de código disponíveis no ambiente de destino. A implantação envolve a implementação de mudanças em uma base de código e, em seguida, a criação e execução dessa base de código nos ambientes da aplicação

ambiente de desenvolvimento

Veja ambiente.

controle detectivo

Um controle de segurança projetado para detectar, registrar e alertar após a ocorrência de um evento. Esses controles são uma segunda linha de defesa, alertando você sobre eventos de segurança que contornaram os controles preventivos em vigor. Para obter mais informações, consulte Controles detectivos em Como implementar controles de segurança na AWS.

mapeamento do fluxo de valor de desenvolvimento (DVSM)

Um processo usado para identificar e priorizar restrições que afetam negativamente a velocidade e a qualidade em um ciclo de vida de desenvolvimento de software. O DVSM estende o processo de mapeamento do fluxo de valor originalmente projetado para práticas de manufatura enxuta. Ele se concentra nas etapas e equipes necessárias para criar e movimentar valor por meio do processo de desenvolvimento de software.

gêmeo digital

Uma representação virtual de um sistema real, como um prédio, fábrica, equipamento industrial ou linha de produção. Os gêmeos digitais oferecem suporte à manutenção preditiva, ao monitoramento remoto e à otimização da produção.

tabela de dimensões

Em um esquema em estrela, uma tabela menor que contém atributos de dados sobre dados quantitativos em uma tabela de fatos. Os atributos da tabela de dimensões geralmente são campos de texto ou números discretos que se comportam como texto. Esses atributos normalmente são usados para restringir consultas, filtrar e rotular conjuntos de resultados.

desastre

Um evento que impede que uma workload ou sistema cumpra seus objetivos de negócios em seu local principal de implantação. Esses eventos podem ser desastres naturais, falhas técnicas ou o resultado de ações humanas, como configuração incorreta não intencional ou ataque de malware.

Recuperação de desastres (RD)

A estratégia e o processo que você usa para minimizar o tempo de inatividade e a perda de dados causados por um disastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem na AWS Well-Architected estrutura.

DML

Veja linguagem de manipulação de banco de dados.

design orientado por domínio

Uma abordagem ao desenvolvimento de um sistema de software complexo conectando seus componentes aos domínios em evolução, ou principais metas de negócios, atendidos por cada componente. Esse conceito foi introduzido por Eric Evans em seu livro Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como você pode usar o design orientado por domínio com o padrão strangler fig, consulte Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental usando contêineres e o Amazon API Gateway.

DR

Veja recuperação de desastres.

Detecção da oscilação

Rastreamento de desvios de uma configuração de linha de base. Por exemplo, você pode usar AWS CloudFormation para detectar desvios nos recursos do sistema ou AWS Control Tower para detectar mudanças em seu landing zone que possam afetar a conformidade com os requisitos de governança.

DVSM

Veja mapeamento do fluxo de valor de desenvolvimento.

E

EDA

Veja análise exploratória de dados.

EDI

Veja intercâmbio eletrônico de dados.

computação de borda

A tecnologia que aumenta o poder computacional de dispositivos inteligentes nas bordas de uma rede de IoT. Quando comparada com a computação em nuvem, a computação de borda pode reduzir a latência da comunicação e melhorar o tempo de resposta.

intercâmbio eletrônico de dados (EDI)

A troca automatizada de documentos comerciais entre organizações. Para obter mais informações, consulte O que é EDI (Intercâmbio eletrônico de dados)?.

criptografia

Um processo de computação que transforma dados de texto simples, legíveis por humanos, em texto cifrado.

chave de criptografia

Uma sequência criptográfica de bits aleatórios que é gerada por um algoritmo de criptografia. As chaves podem variar em tamanho, e cada chave foi projetada para ser imprevisível e exclusiva.

endianismo

A ordem na qual os bytes são armazenados na memória do computador. Big-endian os sistemas armazenam primeiro o byte mais significativo. Little-endian os sistemas armazenam primeiro o byte menos significativo.

endpoint

Veja endpoint de serviço.

serviço de endpoint

Um serviço que pode ser hospedado em uma nuvem privada virtual (VPC) para ser compartilhado com outros usuários. Você pode criar um serviço de endpoint com AWS PrivateLink e conceder permissões a outros diretores Contas da AWS ou a AWS Identity and Access Management (IAM). Essas contas ou entidades principais podem se conectar ao serviço de endpoint de maneira privada criando endpoints da VPC de interface. Para obter mais informações, consulte Criar um serviço de endpoint na documentação do Amazon Virtual Private Cloud (Amazon VPC).

planejamento de recursos empresariais (ERP)

Um sistema que automatiza e gerencia os principais processos de negócios (como contabilidade, MES e gerenciamento de projetos) para uma empresa.

criptografia envelopada

O processo de criptografar uma chave de criptografia com outra chave de criptografia. Para obter mais informações, consulte Criptografia de envelope na documentação AWS Key Management Service (AWS KMS).

ambiente

Uma instância de uma aplicação em execução. Estes são tipos comuns de ambientes na computação em nuvem:

- ambiente de desenvolvimento: uma instância de uma aplicação em execução que está disponível somente para a equipe principal responsável pela manutenção da aplicação. Ambientes de desenvolvimento são usados para testar mudanças antes de promovê-las para ambientes superiores. Esse tipo de ambiente às vezes é chamado de ambiente de teste.
- ambientes inferiores: todos os ambientes de desenvolvimento para uma aplicação, como aqueles usados para compilações e testes iniciais.
- ambiente de produção: uma instância de uma aplicação em execução que os usuários finais podem acessar. Em um CI/CD pipeline, o ambiente de produção é o último ambiente de implantação.

- ambientes superiores: todos os ambientes que podem ser acessados por usuários que não sejam a equipe principal de desenvolvimento. Isso pode incluir um ambiente de produção, ambientes de pré-produção e ambientes para testes de aceitação do usuário.

epic

Em metodologias ágeis, categorias funcionais que ajudam a organizar e priorizar seu trabalho. Os epics fornecem uma descrição de alto nível dos requisitos e das tarefas de implementação. Por exemplo, os épicos de segurança AWS da CAF incluem gerenciamento de identidade e acesso, controles de detetive, segurança de infraestrutura, proteção de dados e resposta a incidentes. Para obter mais informações sobre epics na estratégia de migração da AWS , consulte o guia de implementação do programa.

ERP

Veja planejamento de recursos empresariais.

análise exploratória de dados (EDA)

O processo de analisar um conjunto de dados para entender suas principais características. Você coleta ou agrega dados e, em seguida, realiza investigações iniciais para encontrar padrões, detectar anomalias e verificar suposições. O EDA é realizado por meio do cálculo de estatísticas resumidas e da criação de visualizações de dados.

F

tabela de fatos

A tabela central em um esquema em estrela. Ela armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: as que contêm medidas e as que contêm uma chave externa para uma tabela de dimensões.

Antecipar-se à falha

Uma filosofia que usa testes frequentes e incrementais para reduzir o ciclo de vida do desenvolvimento. É uma parte essencial de uma abordagem ágil.

delimitação de isolamento contra falhas

No Nuvem AWS, um limite, como uma zona de disponibilidade, Região da AWS um plano de controle ou um plano de dados, que limita o efeito de uma falha e ajuda a melhorar a resiliência das cargas de trabalho. Para obter mais informações, consulte AWS Fault Isolation Boundaries.

ramificação de recursos

Veja ramificação.

recursos

Os dados de entrada usados para fazer uma previsão. Por exemplo, em um contexto de manufatura, os recursos podem ser imagens capturadas periodicamente na linha de fabricação.

importância do recurso

O quanto um recurso é importante para as previsões de um modelo. Isso geralmente é expresso como uma pontuação numérica que pode ser calculada por meio de várias técnicas, como Shapley Additive Explanations (SHAP) e gradientes integrados. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

transformação de recursos

O processo de otimizar dados para o processo de ML, incluindo enriquecer dados com fontes adicionais, escalar valores ou extrair vários conjuntos de informações de um único campo de dados. Isso permite que o modelo de ML se beneficie dos dados. Por exemplo, se a data “2021-05-27 00:15:37” for dividida em “2021”, “maio”, “quinta” e “15”, isso poderá ajudar o algoritmo de aprendizado a aprender padrões diferenciados associados a diferentes componentes de dados.

prompt few shot

Fornecer a um LLM um pequeno número de exemplos que demonstram a tarefa e o resultado desejado antes de solicitar que ele execute uma tarefa semelhante. Essa técnica é uma aplicação do aprendizado contextual, em que os modelos aprendem com exemplos (fotos) incorporados aos prompts. Few-shot a solicitação pode ser eficaz para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também prompts zero-shot.

FGAC

Veja controle de acesso refinado.

Controle de acesso refinado (FGAC)

O uso de várias condições para permitir ou negar uma solicitação de acesso.

migração flash-cut

Um método de migração de banco de dados que usa replicação contínua de dados via captura de dados de alteração para migrar os dados no menor tempo possível, em vez de usar uma abordagem em fases. O objetivo é reduzir ao mínimo o tempo de inatividade.

FM

Veja modelo de base.

modelo de base (FM)

Uma grande rede neural de aprendizado profundo que treina em grandes conjuntos de dados generalizados e não rotulados. Os FMs são capazes de realizar uma ampla variedade de tarefas gerais, como entender a linguagem, gerar texto e imagens e conversar em linguagem natural. Para obter mais informações, consulte O que são modelos de base?.

Gateway FM

Um intermediário centralizado que controla e normaliza o acesso aos modelos de fundação. Também conhecido como gateway LLM.

G

IA generativa

Um subconjunto de modelos de IA que foram treinados em grandes quantidades de dados e que podem usar um simples prompt de texto para criar novos artefatos e conteúdo, como imagens, vídeos, texto e áudio. Para obter mais informações, consulte O que é IA generativa?.

bloqueio geográfico

Veja restrições geográficas.

restrições geográficas (bloqueio geográfico)

Na Amazon CloudFront, uma opção para impedir que usuários em países específicos acessem distribuições de conteúdo. É possível usar uma lista de permissões ou uma lista de bloqueios para especificar países aprovados e banidos. Para obter mais informações, consulte Restringir a distribuição geográfica do seu conteúdo na CloudFront documentação.

Fluxo de trabalho do GitFlow

Uma abordagem na qual ambientes inferiores e superiores usam ramificações diferentes em um repositório de código-fonte. O fluxo de trabalho do Gitflow é considerado legado, e o fluxo de trabalho trunk-based é a abordagem moderna e preferencial.

golden image

Um snapshot de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma golden image pode ser usada para

provisionar software em vários dispositivos e ajudar a melhorar a velocidade, a escalabilidade e a produtividade nas operações de fabricação de dispositivos.

estratégia greenfield

A ausência de infraestrutura existente em um novo ambiente. Ao adotar uma estratégia greenfield para uma arquitetura de sistema, é possível selecionar todas as novas tecnologias sem a restrição da compatibilidade com a infraestrutura existente, também conhecida como brownfield. Se estiver expandindo a infraestrutura existente, poderá combinar as estratégias brownfield e greenfield.

barreira de proteção

Uma regra de alto nível que ajuda a gerenciar recursos, políticas e conformidade em todas as unidades organizacionais (UOs). Barreiras de proteção preventivas impõem políticas para garantir o alinhamento a padrões de conformidade. Elas são implementadas usando políticas de controle de serviço e limites de permissões do IAM. Barreiras de proteção detectivas detectam violações de políticas e problemas de conformidade e geram alertas para remediação. Eles são implementados usando AWS Config, AWS Security Hub CSPM, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

grades de proteção (IA)

Mecanismos de segurança que filtram, validam e restringem as entradas e saídas dos agentes para ajudar a garantir um comportamento de IA responsável e seguro.

H

HA

Veja alta disponibilidade.

migração heterogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que usa um mecanismo de banco de dados diferente (por exemplo, Oracle para Amazon Aurora). A migração heterogênea geralmente faz parte de um esforço de redefinição da arquitetura, e converter o esquema pode ser uma tarefa complexa. O AWS fornece o AWS SCT para ajudar nas conversões de esquemas.

alta disponibilidade (HA)

A capacidade de uma workload operar continuamente, sem intervenção, em caso de desafios ou desastres. Os sistemas AH são projetados para realizar o failover automático, oferecer consistentemente desempenho de alta qualidade e lidar com diferentes cargas e falhas com impacto mínimo no desempenho.

modernização de historiador

Uma abordagem usada para modernizar e atualizar os sistemas de tecnologia operacional (OT) para melhor atender às necessidades do setor de manufatura. Um historiador é um tipo de banco de dados usado para coletar e armazenar dados de várias fontes em uma fábrica.

dados de hold-out

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de machine learning. Você pode usar dados de hold-out para avaliar a performance do modelo comparando as predições do modelo com os dados de retenção.

humano no circuito (HiTL)

Um padrão de fluxo de trabalho em que a execução do agente é pausada para análise e aprovação humana em pontos críticos de decisão.

migração homogênea de bancos de dados

Migrar seu banco de dados de origem para um banco de dados de destino que compartilha o mesmo mecanismo de banco de dados (por exemplo, Microsoft SQL Server para Amazon RDS para SQL Server). A migração homogênea geralmente faz parte de um esforço de redefinição da hospedagem ou da plataforma. É possível usar utilitários de banco de dados nativos para migrar o esquema.

dados quentes

Dados acessados com frequência, como dados em tempo real ou dados translacionais recentes. Esses dados normalmente exigem uma camada ou classe de armazenamento de alto desempenho para fornecer respostas rápidas às consultas.

hotfix

Uma correção urgente para um problema crítico em um ambiente de produção. Devido à sua urgência, um hotfix geralmente é feito fora do fluxo de trabalho típico de uma DevOps versão.

período de hipercuidados

Imediatamente após a substituição, o período em que uma equipe de migração gerencia e monitora as aplicações migradas na nuvem para resolver quaisquer problemas. Normalmente, a duração desse período é de 1 a 4 dias. No final do período de hipercuidados, a equipe de migração normalmente transfere a responsabilidade pelas aplicações para a equipe de operações de nuvem.

eu

IaC

Veja infraestrutura como código.

Política baseada em identidade

Uma política anexada a um ou mais diretores do IAM que define suas permissões no Nuvem AWS ambiente.

aplicação ociosa

Uma aplicação que tem um uso médio de CPU e memória entre 5 e 20% em um período de 90 dias. Em um projeto de migração, é comum retirar essas aplicações ou retê-las on-premises.

IIoT

Veja Internet das Coisas Industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para workloads de produção em vez de atualizar, aplicar patches ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e preditivas do que infraestruturas mutáveis. Para obter mais informações, consulte as melhores práticas de implantação usando infraestrutura imutável na AWS Well-Architected Estrutura.

VPC de entrada (admissão)

Em uma arquitetura de AWS várias contas, uma VPC que aceita, inspeciona e roteia conexões de rede de fora de um aplicativo. A Arquitetura de referência de segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente apenas alguns microsserviços ou usuários para o novo sistema. Depois de verificar se tudo está funcionando corretamente, mova os microsserviços ou usuários adicionais de forma incremental até poder descomissionar seu sistema herdado. Essa estratégia reduz os riscos associados a migrações de grande porte.

Indústria 4.0

Um termo que foi introduzido por Klaus Schwab em 2016 para se referir à modernização dos processos de fabricação por meio de avanços na conectividade, dados em tempo real, automação, análise e. AI/ML

infraestrutura

Todos os recursos e ativos contidos no ambiente de uma aplicação.

Infraestrutura como código (IaC)

O processo de provisionamento e gerenciamento da infraestrutura de uma aplicação por meio de um conjunto de arquivos de configuração. A IaC foi projetada para ajudar você a centralizar o gerenciamento da infraestrutura, padronizar recursos e escalar rapidamente para que novos ambientes sejam reproduzíveis, confiáveis e consistentes.

Internet das Coisas Industrial (IIoT)

O uso de sensores e dispositivos conectados à Internet nos setores industriais, como manufatura, energia, automotivo, saúde, ciências biológicas e agricultura. Para obter mais informações, consulte Construir uma estratégia de transformação digital para a Internet das Coisas Industrial (IIoT).

VPC de inspeção

Em uma arquitetura de AWS várias contas, uma VPC centralizada que gerencia as inspeções do tráfego de rede entre VPCs (na mesma ou em diferentes Regiões da AWS), a Internet e as redes locais. A Arquitetura de referência de segurança da AWS recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

Internet das coisas (IoT)

A rede de objetos físicos conectados com sensores ou processadores incorporados que se comunicam com outros dispositivos e sistemas pela Internet ou por uma rede de comunicação local. Para obter mais informações, consulte O que é IoT?

interpretabilidade

Uma característica de um modelo de machine learning que descreve o grau em que um ser humano pode entender como as previsões do modelo dependem de suas entradas. Para obter mais informações, consulte Interpretabilidade do modelo de aprendizado de máquina com AWS.

IoT

Veja Internet das Coisas.

Biblioteca de informações de TI (ITIL)

Um conjunto de práticas recomendadas para fornecer serviços de TI e alinhar esses serviços a requisitos de negócios. A ITIL fornece a base para o ITSM.

Gerenciamento de serviços de TI (ITSM)

Atividades associadas a design, implementação, gerenciamento e suporte de serviços de TI para uma organização. Para obter informações sobre a integração de operações em nuvem com ferramentas de ITSM, consulte o guia de integração de operações.

ITIL

Veja biblioteca de informações de TI.

ITSM

Veja gerenciamento de serviços de TI.

L

controle de acesso baseado em etiqueta (LBAC)

Uma implementação do controle de acesso obrigatório (MAC) em que os usuários e os dados em si recebem explicitamente um valor de etiqueta de segurança. A interseção entre a etiqueta de segurança do usuário e a etiqueta de segurança dos dados determina quais linhas e colunas podem ser vistas pelo usuário.

zona de pouso

Uma landing zone é um AWS ambiente bem arquitetado, com várias contas, escalável e seguro. Um ponto a partir do qual suas organizações podem iniciar e implantar rapidamente workloads e aplicações com confiança em seu ambiente de segurança e infraestrutura. Para obter mais informações sobre zonas de pouso, consulte Configurar um ambiente da AWS com várias contas seguro e escalável.

grande modelo de linguagem (LLM)

Um modelo de IA de aprendizado profundo pré-treinado em uma grande quantidade de dados. Um LLM pode realizar várias tarefas, como responder a perguntas, resumir documentos, traduzir texto para outros idiomas e completar frases. Para obter mais informações, consulte O que é grande modelo de linguagem (LLM)?.

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em rótulo.

privilégio mínimo

A prática recomendada de segurança de conceder as permissões mínimas necessárias para executar uma tarefa. Para obter mais informações, consulte Aplicar permissões de privilégios mínimos na documentação do IAM.

mover sem alterações (lift-and-shift)

Veja 7 Rs.

sistema little-endian

Um sistema que armazena o byte menos significativo antes. Veja também endianness.

LLM

Veja grande modelo de linguagem.

ambientes inferiores

Veja ambiente.

M

machine learning (ML)

Um tipo de inteligência artificial que usa algoritmos e técnicas para reconhecimento e aprendizado de padrões. O ML analisa e aprende com dados gravados, por exemplo, dados da Internet das Coisas (IoT), para gerar um modelo estatístico baseado em padrões. Para obter mais informações, consulte Machine learning.

ramificação principal

Veja ramificação.

Malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações sensíveis ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Troia, spyware e keyloggers.

Serviços gerenciados

Serviços da AWS para o qual AWS opera a camada de infraestrutura, o sistema operacional e as plataformas, e você acessa os endpoints para armazenar e recuperar dados. O Amazon Simple Storage Service (Amazon S3) e o Amazon DynamoDB são exemplos de serviços gerenciados. Eles também são conhecidos como serviços abstraídos.

sistema de execução de manufatura (MES)

Um sistema de software para rastrear, monitorar, documentar e controlar processos de produção que convertem matérias-primas em produtos acabados no chão de fábrica.

MAP

Veja Programa de Aceleração da Migração.

MCP

Consulte Protocolo de contexto do modelo.

Protocolo de contexto para modelos (MCP)

Um protocolo sem estado para comunicação entre agentes e ferramentas.

Servidor MCP

Um serviço que expõe uma ou mais ferramentas por meio do Model Context Protocol.

mecanismo

Um processo completo em que você cria uma ferramenta, impulsiona a adoção da ferramenta e, em seguida, inspeciona os resultados para fazer ajustes. Um mecanismo é um ciclo que se reforça e se aprimora à medida que opera. Para obter mais informações, consulte Criação de mecanismos na AWS Well-Architected estrutura.

conta de membro

Todos, Contas da AWS exceto a conta de gerenciamento, que fazem parte de uma organização em AWS Organizations. Uma conta só pode ser membro de uma organização de cada vez.

MES

Veja sistema de execução de manufatura.

Transporte de Telemetria de Enfileiramento de Mensagens (MQTT)

Um protocolo de comunicação leve, máquina a máquina (M2M), baseado no padrão, para dispositivos de IoT com recursos publish/subscribelimitados.

microsserviço

Um serviço pequeno e independente que se comunica por meio de APIs bem definidas e normalmente pertence a equipes pequenas e autônomas. Por exemplo, um sistema de seguradora pode incluir microsserviços que mapeiam as capacidades comerciais, como vendas ou marketing, ou subdomínios, como compras, reclamações ou análises. Os benefícios dos microsserviços incluem agilidade, escalabilidade flexível, fácil implantação, código reutilizável e resiliência. Para obter mais informações, consulte Integração de microsserviços usando serviços sem AWS servidor.

arquitetura de microsserviços

Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microsserviço. Esses microsserviços se comunicam por meio de uma interface bem definida usando APIs leves. Cada microsserviço nessa arquitetura pode ser atualizado, implantado e escalado para atender à demanda por funções específicas de uma aplicação. Para obter mais informações, consulte Implementação de microsserviços em. AWS

Programa de Aceleração da Migração (MAP)

Um AWS programa que fornece suporte de consultoria, treinamento e serviços para ajudar as organizações a criar uma base operacional sólida para migrar para a nuvem e ajudar a

compensar o custo inicial das migrações. O MAP inclui uma metodologia de migração para executar migrações legadas de forma metódica e um conjunto de ferramentas para automatizar e acelerar cenários comuns de migração.

migração em escala

O processo de mover a maior parte do portfólio de aplicações para a nuvem em ondas, com mais aplicações sendo movidas em um ritmo mais rápido a cada onda. Essa fase usa as práticas recomendadas e lições aprendidas nas fases anteriores para implementar uma fábrica de migração de equipes, ferramentas e processos para agilizar a migração de workloads por meio de automação e entrega ágeis. Esta é a terceira fase da estratégia de migração para a AWS.

fábrica de migração

Cross-functional equipes que simplificam a migração de cargas de trabalho por meio de abordagens automatizadas e ágeis. As equipes da fábrica de migração geralmente incluem operações, analistas e proprietários de negócios, engenheiros de migração, desenvolvedores e DevOps profissionais que trabalham em sprints. Entre 20 e 50% de um portfólio de aplicações corporativas consiste em padrões repetidos que podem ser otimizados por meio de uma abordagem de fábrica. Para obter mais informações, consulte discussão sobre fábricas de migração e o guia do Cloud Migration Factory neste conjunto de conteúdo.

metadados de migração

As informações sobre a aplicação e o servidor necessárias para concluir a migração. Cada padrão de migração exige um conjunto de metadados de migração diferente. Exemplos de metadados de migração incluem a sub-rede, o grupo de segurança e AWS a conta de destino.

padrão de migração

Uma tarefa de migração repetível que detalha a estratégia de migração, o destino da migração e a aplicação ou o serviço de migração usado. Exemplo: rehospede a migração para o Amazon EC2 AWS com o Application Migration Service.

Avaliação de Portfólio para Migração (MPA)

Uma ferramenta on-line que fornece informações para validar o caso de negócios para migrar para a Nuvem AWS. O MPA fornece avaliação detalhada do portfólio (dimensionamento correto do servidor, preços, comparações de TCO, análise de custos de migração), bem como planejamento de migração (análise e coleta de dados de aplicações, agrupamento de aplicações, priorização de migração e planejamento de ondas). A ferramenta MPA (requer login) está disponível gratuitamente para todos os AWS consultores e consultores parceiros da APN.

Avaliação de Preparação para Migração (MRA)

O processo de obter insights sobre o status de prontidão de uma organização para a nuvem, identificar pontos fortes e fracos e criar um plano de ação para fechar as lacunas identificadas, usando o CAF. AWS Para mais informações, consulte o guia de preparação para migração. A MRA é a primeira fase da estratégia de migração para a AWS.

estratégia de migração

A abordagem usada para migrar uma workload para a Nuvem AWS. Para obter mais informações, veja a entrada 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja machine learning.

modernização

Transformar uma aplicação desatualizada (herdada ou monolítica) e sua infraestrutura em um sistema ágil, elástico e altamente disponível na nuvem para reduzir custos, ganhar eficiência e aproveitar as inovações. Para obter mais informações, consulte Strategy for modernizing applications in the Nuvem AWS.

avaliação de preparação para modernização

Uma avaliação que ajuda a determinar a preparação para modernização das aplicações de uma organização. Ela identifica benefícios, riscos e dependências e determina o quão bem a organização pode acomodar o estado futuro dessas aplicações. O resultado da avaliação é um esquema da arquitetura de destino, um roteiro que detalha as fases de desenvolvimento e os marcos do processo de modernização e um plano de ação para abordar as lacunas identificadas. Para obter mais informações, consulte Evaluating modernization readiness for applications in the Nuvem AWS.

aplicações monolíticas (monólitos)

Aplicações que são executadas como um único serviço com processos fortemente acoplados. As aplicações monolíticas apresentam várias desvantagens. Se um recurso da aplicação apresentar um aumento na demanda, toda a arquitetura deverá ser escalada. Adicionar ou melhorar os recursos de uma aplicação monolítica também se torna mais complexo quando a base de código cresce. Para resolver esses problemas, é possível criar uma arquitetura de microsserviços. Para obter mais informações, consulte Decompor monólitos em microsserviços.

MPA

Veja Avaliação do Portfólio para Migração.

MQTT

Veja Transporte de Telemetria de Enfileiramento de Mensagens.

classificação multiclasse

Um processo que ajuda a gerar previsões para várias classes (prevendo um ou mais de dois resultados). Por exemplo, um modelo de ML pode perguntar “Este produto é um livro, um carro ou um telefone?” ou “Qual categoria de produtos é mais interessante para este cliente?”

infraestrutura mutável

Um modelo que atualiza e modifica a infraestrutura existente para workloads de produção. Para melhorar a consistência, confiabilidade e previsibilidade, a AWS Well-Architected Estrutura recomenda o uso de infraestrutura imutável como uma prática recomendada.

O

OAC

Veja controle de acesso de origem.

OAI

Veja identidade de acesso de origem.

OCM

Veja gerenciamento de alterações organizacionais.

migração offline

Um método de migração no qual a workload de origem é desativada durante o processo de migração. Esse método envolve tempo de inatividade prolongado e geralmente é usado para workloads pequenas e não críticas.

OI

Veja integração de operações.

Ola

Veja acordo de nível operacional.

migração online

Um método de migração no qual a workload de origem é copiada para o sistema de destino sem ser colocada offline. As aplicações conectadas à workload podem continuar funcionando durante a migração. Esse método envolve um tempo de inatividade nulo ou mínimo e normalmente é usado para workloads essenciais para a produção.

OPC-UA

Veja Open Process Communications - Unified Architecture.

Comunicação de processo aberto - Arquitetura unificada (OPC-UA)

Um protocolo de comunicação máquina a máquina (M2M) para automação industrial. OPC-UA fornece um padrão de interoperabilidade com esquemas de criptografia, autenticação e autorização de dados.

acordo de nível operacional (OLA)

Um acordo que esclarece o que os grupos funcionais de TI prometem oferecer uns aos outros para apoiar um acordo de serviço (SLA).

análise de prontidão operacional (ORR)

Uma lista de verificação de perguntas e práticas recomendadas associadas que ajudam você a entender, avaliar, prevenir ou reduzir o escopo de incidentes e possíveis falhas. Para obter mais informações, consulte Operational Readiness Reviews (ORR) na AWS Well-Architected Estrutura.

tecnologia operacional (TO)

Sistemas de hardware e software que trabalham com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas de tecnologia da informação (TI) e tecnologia operacional (TO) é o foco principal das transformações da Indústria 4.0.

integração de operações (OI)

O processo de modernização das operações na nuvem, que envolve planejamento de preparação, automação e integração. Para obter mais informações, consulte o guia de integração de operações.

trilha organizacional

Uma trilha criada por ela AWS CloudTrail registra todos os eventos de todos Contas da AWS em uma organização em AWS Organizations. Essa trilha é criada em cada Conta da AWS que faz parte da organização e monitora a atividade em cada conta. Para obter mais informações, consulte Criação de uma trilha para uma organização na CloudTrail documentação.

gerenciamento de alterações organizacionais (OCM)

Uma estrutura para gerenciar grandes transformações de negócios disruptivas de uma perspectiva de pessoas, cultura e liderança. O OCM ajuda as organizações a se prepararem e fazerem a transição para novos sistemas e estratégias, acelerando a adoção de alterações, abordando questões de transição e promovendo mudanças culturais e organizacionais. Na estratégia de AWS migração, essa estrutura é chamada de aceleração de pessoas, devido à velocidade de mudança necessária nos projetos de adoção da nuvem. Para obter mais informações, consulte o guia do OCM.

controle de acesso de origem (OAC)

Em CloudFront, uma opção aprimorada para restringir o acesso para proteger seu conteúdo do Amazon Simple Storage Service (Amazon S3). O OAC oferece suporte a todos os buckets do S3 Regiões da AWS, à criptografia do lado do servidor com AWS KMS (SSE-KMS) e à dinâmica `PUT` e `DELETE` às solicitações ao bucket do S3.

Identidade do acesso de origem (OAI)

Em CloudFront, uma opção para restringir o acesso para proteger seu conteúdo do Amazon S3. Quando você usa o OAI, CloudFront cria um principal com o qual o Amazon S3 pode se autenticar. Os diretores autenticados podem acessar o conteúdo em um bucket do S3 somente por meio de uma distribuição específica. CloudFront Veja também OAC, que fornece um controle de acesso mais granular e aprimorado.

ORR

Veja análise de prontidão operacional.

OT

Veja tecnologia operacional.

VPC de saída (egresso)

Em uma arquitetura de AWS várias contas, uma VPC que gerencia conexões de rede que são iniciadas de dentro de um aplicativo. A Arquitetura de referência de segurança da AWS

recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

P

limite de permissões

Uma política de gerenciamento do IAM anexada a entidades principais do IAM para definir as permissões máximas que o usuário ou perfil podem ter. Para obter mais informações, consulte Limites de permissões na documentação do IAM.

Informações de identificação pessoal (PII)

Informações que, quando visualizadas diretamente ou combinadas com outros dados relacionados, podem ser usadas para inferir razoavelmente a identidade de um indivíduo. Exemplos de PII incluem nomes, endereços e informações de contato.

PII

Veja informações de identificação pessoal.

manual

Um conjunto de etapas predefinidas que capturam o trabalho associado às migrações, como a entrega das principais funções operacionais na nuvem. Um manual pode assumir a forma de scripts, runbooks automatizados ou um resumo dos processos ou etapas necessários para operar seu ambiente modernizado.

PLC

Veja controlador lógico programável.

PLM

Veja gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (veja política baseada em identidade), especificar condições de acesso (veja políta baseada em recurso) ou definir as permissões máximas para todas as contas em uma organização no AWS Organizations (veja política de controle de serviços).

persistência poliglota

Escolher de forma independente a tecnologia de armazenamento de dados de um microsserviço com base em padrões de acesso a dados e outros requisitos. Se seus microsserviços tiverem a mesma tecnologia de armazenamento de dados, eles poderão enfrentar desafios de implementação ou apresentar baixa performance. Os microsserviços serão implementados com mais facilidade e alcançarão performance e escalabilidade melhores se usarem o armazenamento de dados mais bem adaptado às suas necessidades.

avaliação do portfólio

Um processo de descobrir, analisar e priorizar o portfólio de aplicações para planejar a migração. Para obter mais informações, consulte Avaliar a preparação para a migração.

predicado

Uma condição de consulta que retorna `true` ou `false`, normalmente localizada em uma cláusula `WHERE`.

pushdown de predicados

Uma técnica de otimização de consultas de banco de dados que filtra os dados na consulta antes da transferência. Isso reduz a quantidade de dados que devem ser recuperados e processados do banco de dados relacional e melhora a performance das consultas.

controle preventivo

Um controle de segurança projetado para evitar que um evento ocorra. Esses controles são a primeira linha de defesa para ajudar a evitar acesso não autorizado ou alterações indesejadas em sua rede. Para obter mais informações, consulte Controles preventivos em Como implementar controles de segurança na AWS.

principal (entidade principal)

Uma entidade AWS que pode realizar ações e acessar recursos. Essa entidade geralmente é um usuário raiz para um Conta da AWS, uma função do IAM ou um usuário. Para obter mais informações, consulte Entidade principal em Termos e conceitos de perfis na documentação do IAM.

Privacidade por design

Uma abordagem em engenharia de sistemas que leva em consideração a privacidade em todo o processo de desenvolvimento.

zonas hospedadas privadas

Um contêiner que armazena informações sobre como você quer que o Amazon Route 53 responda a consultas ao DNS para um domínio e seus subdomínios dentro de uma ou mais VPCs. Para obter mais informações, consulte Como trabalhar com zonas hospedadas privadas na documentação do Route 53.

controle proativo

Um controle de segurança desenvolvido para evitar a implantação de recursos não conformes. Esses controles verificam os recursos antes de serem provisionados. Se o recurso não estiver em conformidade com o controle, ele não será provisionado. Para obter mais informações, consulte o guia de referência de controles na AWS Control Tower documentação e consulte Controles proativos em Implementação de controles de segurança em AWS.

gerenciamento do ciclo de vida do produto (PLM)

O gerenciamento de dados e processos de um produto em todo o seu ciclo de vida, desde a concepção, o desenvolvimento e o lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja ambiente.

controlador lógico programável (PLC)

Na manufatura, um computador altamente confiável e adaptável que monitora as máquinas e automatiza os processos de fabricação.

encadeamento de prompts

Uso da saída de um prompt do LLM como entrada para o próximo prompt para gerar respostas melhores. Essa técnica é usada para dividir uma tarefa complexa em subtarefas, ou para refinar ou expandir iterativamente uma resposta preliminar. Isso ajuda a melhorar a precisão e a relevância das respostas de um modelo e permite resultados mais granulares e personalizados.

pseudonimização

O processo de substituir identificadores pessoais em um conjunto de dados por valores de espaço reservado. A pseudonimização pode ajudar a proteger a privacidade pessoal. Os dados pseudonimizados ainda são considerados dados pessoais.

publish/subscribe (pub/sub)

Um padrão que permite comunicações assíncronas entre microsserviços para melhorar a escalabilidade e a capacidade de resposta. Por exemplo, em um MES baseado em microsserviços, um microsserviço pode publicar mensagens de eventos em um canal em que outros microsserviços possam assinar. O sistema pode adicionar novos microsserviços sem alterar o serviço de publicação.

Q

plano de consulta

Uma série de etapas, como instruções, usadas para acessar os dados em um sistema de banco de dados relacional SQL.

regressão de planos de consultas

Quando um otimizador de serviço de banco de dados escolhe um plano menos adequado do que escolhia antes de uma determinada alteração no ambiente de banco de dados ocorrer. Isso pode ser causado por alterações em estatísticas, restrições, configurações do ambiente, associações de parâmetros de consulta e atualizações do mecanismo de banco de dados.

R

Matriz RACI

Veja responsável, aprovador, consultado, informado (RACI).

RAG

Veja geração aumentada via recuperação.

ransomware

Um software mal-intencionado desenvolvido para bloquear o acesso a um sistema ou dados de computador até que um pagamento seja feito.

Matriz RASCI

Veja responsável, aprovador, consultado, informado (RACI).

RCAC

Veja controle de acesso por linha e coluna.

réplica de leitura

Uma cópia de um banco de dados usada somente para leitura. É possível encaminhar consultas para a réplica de leitura e reduzir a carga no banco de dados principal.

Redefinir arquitetura

Veja 7 Rs.

objetivo de ponto de recuperação (RPO).

O máximo período de tempo aceitável desde o último ponto de recuperação de dados. Isso determina o que é considerado uma perda aceitável de dados entre o último ponto de recuperação e a interrupção do serviço.

objetivo de tempo de recuperação (RTO)

O máximo atraso aceitável entre a interrupção e a restauração do serviço.

refatorar

Veja 7 Rs.

Região

Uma coleção de AWS recursos em uma área geográfica. Cada um Região da AWS é isolado e independente dos outros para fornecer tolerância a falhas, estabilidade e resiliência. Para obter informações, consulte Specify which Regiões da AWS your account can use.

regressão

Uma técnica de ML que prevê um valor numérico. Por exemplo, para resolver o problema de “Por qual preço esta casa será vendida?” um modelo de ML pode usar um modelo de regressão linear para prever o preço de venda de uma casa com base em fatos conhecidos sobre a casa (por exemplo, a metragem quadrada).

redefinir a hospedagem

Veja 7 Rs.

versão

Em um processo de implantação, o ato de promover mudanças em um ambiente de produção.

realocar

Veja 7 Rs.

redefinir a plataforma

Veja 7 Rs.

recomprar

Veja 7 Rs.

resiliência

A capacidade de uma aplicação de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência na Nuvem AWS. Para obter mais informações, consulte Nuvem AWS Resilience.

política baseada em recurso

Uma política associada a um recurso, como um bucket do Amazon S3, um endpoint ou uma chave de criptografia. Esse tipo de política especifica quais entidades principais têm acesso permitido, ações válidas e quaisquer outras condições que devem ser atendidas.

matriz responsável, accountable, consultada, informada (RACI)

Uma matriz que define as funções e responsabilidades de todas as partes envolvidas nas atividades de migração e nas operações de nuvem. O nome da matriz é derivado dos tipos de responsabilidade definidos na matriz: responsável (R), responsabilizável (A), consultado (C) e informado (I). O tipo de suporte (S) é opcional. Se você incluir suporte, a matriz será chamada de matriz RASCI e, se excluir, será chamada de matriz RACI.

controle responsivo

Um controle de segurança desenvolvido para conduzir a remediação de eventos adversos ou desvios em relação à linha de base de segurança. Para obter mais informações, consulte Controles responsivos em Como implementar controles de segurança na AWS.

reter

Veja 7 Rs.

Retirada

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa em que um LLM faz referência a uma fonte de dados autorizada que está fora de suas fontes de dados de treinamento antes de gerar uma resposta. Por exemplo, um modelo RAG pode realizar uma pesquisa semântica na base de conhecimento ou nos dados personalizados de uma organização. Para obter mais informações, consulte O que é RAG (geração aumentada via recuperação)?.

alternância

O processo de atualizar periodicamente um segredo para dificultar o acesso de um invasor às credenciais.

controle de acesso por linha e coluna (RCAC)

O uso de expressões SQL básicas e flexíveis que tenham regras de acesso definidas. O RCAC consiste em permissões de linha e máscaras de coluna.

RPO

Veja objetivo de ponto de recuperação.

RTO

Veja objetivo de tempo de recuperação.

runbook

Um conjunto de procedimentos manuais ou automatizados necessários para realizar uma tarefa específica. Eles são normalmente criados para agilizar operações ou procedimentos repetitivos com altas taxas de erro.

S

SAML 2.0

Um padrão aberto que muitos provedores de identidade (IdPs) usam. Esse recurso permite o login único federado (SSO), para que os usuários possam fazer login Console de gerenciamento da AWS ou chamar as operações da AWS API sem que você precise criar um usuário no IAM para todos em sua organização. Para obter mais informações sobre a federação baseada em SAML 2.0, consulte Sobre a federação baseada em SAML 2.0 na documentação do IAM.

SCADA

Veja controle de supervisão e aquisição de dados.

SCP

Veja política de controle de serviço.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Consiste no valor secreto e em seus metadados. O valor secreto pode ser binário, uma única string ou várias strings. Para obter mais informações, consulte What's in a Secrets Manager secret? na documentação do Secrets Manager.

segurança desde a concepção

Uma abordagem em engenharia de sistemas que leva em consideração a segurança em todo o processo de desenvolvimento.

controle de segurança

Uma barreira de proteção técnica ou administrativa que impede, detecta ou reduz a capacidade de uma ameaça explorar uma vulnerabilidade de segurança. Existem quatro tipos primários de controles de segurança: preventivos, detectivos, responsivos e proativos.

hardening da segurança

O processo de reduzir a superfície de ataque para torná-la mais resistente a ataques. Isso pode incluir ações como remover recursos que não são mais necessários, implementar a prática recomendada de segurança de conceder privilégios mínimos ou desativar recursos desnecessários em arquivos de configuração.

sistema de gerenciamento de eventos e informações de segurança (SIEM)

Ferramentas e serviços que combinam sistemas de gerenciamento de informações de segurança (SIM) e gerenciamento de eventos de segurança (SEM). Um sistema SIEM coleta, monitora e analisa dados de servidores, redes, dispositivos e outras fontes para detectar ameaças e violações de segurança e gerar alertas.

automação de resposta de segurança

Uma ação predefinida e programada projetada para responder ou remediar automaticamente um evento de segurança. Essas automações servem como controles de segurança responsivos ou detectivos que ajudam você a implementar as melhores práticas AWS de segurança. Exemplos de ações de resposta automatizada incluem a modificação de um grupo de segurança da VPC, a aplicação de patches em uma instância do Amazon EC2 ou a alternância de credenciais.

Criptografia do lado do servidor

Criptografia dos dados em seu destino, por AWS service (Serviço da AWS) quem os recebe.

política de controle de serviços (SCP)

Uma política que fornece controle centralizado sobre as permissões de todas as contas em uma organização no AWS Organizations. As SCPs definem barreiras de proteção ou estabelecem limites para as ações que um administrador pode delegar a usuários ou perfis. É possível usar SCPs como listas de permissão ou de negação para especificar quais serviços ou ações são permitidos ou proibidos. Para obter mais informações, consulte Políticas de controle de serviço na AWS Organizations documentação.

service endpoint (endpoint de serviço)

O URL do ponto de entrada para um AWS service (Serviço da AWS). Você pode usar o endpoint para se conectar programaticamente ao serviço de destino. Para obter mais informações, consulte Endpoints do AWS service (Serviço da AWS) na Referência geral da AWS.

acordo de serviço (SLA)

Um acordo que esclarece o que uma equipe de TI promete fornecer aos clientes, como tempo de atividade e performance do serviço.

indicador de nível de serviço (SLI)

Uma avaliação de um aspecto de performance de um serviço, como taxa de erro, disponibilidade ou throughput.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme avaliado por um indicador de nível de serviço.

modelo de responsabilidade compartilhada

Um modelo que descreve a responsabilidade com a qual você compartilha AWS pela segurança e conformidade na nuvem. AWS é responsável pela segurança da nuvem, enquanto você é responsável pela segurança na nuvem. Para obter mais informações, consulte o Modelo de responsabilidade compartilhada.

Inteligência artificial sombria

Aplicativos de IA não autorizados criados ou usados fora dos canais controlados dentro de uma organização.

SIEM

Veja sistema de gerenciamento de eventos e informações de segurança.

ponto único de falha (SPOF)

Uma falha em um único componente crítico de uma aplicação que pode interromper o sistema.

SLA

Veja cacordo de serviço.

SLI

Veja indicador de nível de serviço.

SLO

Veja objetivo de nível de serviço.

modelo dividir e semear

Um padrão para escalar e acelerar projetos de modernização. À medida que novos recursos e lançamentos de produtos são definidos, a equipe principal se divide para criar novas equipes de produtos. Isso ajuda a escalar os recursos e os serviços da sua organização, melhora a produtividade do desenvolvedor e possibilita inovações rápidas. Para obter mais informações, consulte Phased approach to modernizing applications in the Nuvem AWS.

SPOF

Veja ponto único de falha.

esquema em estrela

Uma estrutura organizacional de banco de dados que usa uma grande tabela de fatos para armazenar dados transacionais ou medidos e usa uma ou mais tabelas dimensionais menores para armazenar atributos de dados. Essa estrutura foi projetada para ser usada em um data warehouse ou para fins de inteligência comercial.

padrão strangler fig

Uma abordagem à modernização de sistemas monolíticos que consiste em reescrever e substituir incrementalmente a funcionalidade do sistema até que o sistema herdado possa ser desativado. Esse padrão usa a analogia de uma videira que cresce e se torna uma árvore estabelecida e, eventualmente, supera e substitui sua hospedeira. O padrão foi apresentado por Martin Fowler

como forma de gerenciar riscos ao reescrever sistemas monolíticos. Para ver um exemplo de como aplicar esse padrão, consulte Modernizando os serviços web legados da Microsoft ASP.NET (ASMX) de forma incremental usando contêineres e o Amazon API Gateway.

sub-rede

Um intervalo de endereços IP na VPC. Cada sub-rede fica alocada em uma única zona de disponibilidade.

controle supervisório e aquisição de dados (SCADA)

Na manufatura, um sistema que usa hardware e software para monitorar ativos físicos e operações de produção.

symmetric encryption (criptografia simétrica)

Um algoritmo de criptografia que usa a mesma chave para criptografar e descriptografar dados.

testes sintéticos

Testar um sistema de forma que simule as interações do usuário para detectar possíveis problemas ou monitorar a performance. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

Uma técnica para fornecer contexto, instruções ou orientações a um LLM a fim de direcionar seu comportamento. Os prompts do sistema ajudam a definir o contexto e a estabelecer regras para interações com os usuários.

T

tags

Key-value pares que atuam como metadados para organizar seus AWS recursos. As tags podem ajudar você a gerenciar, identificar, organizar, pesquisar e filtrar recursos da . Para obter mais informações, consulte Marcar seus recursos do AWS.

variável-alvo

O valor que você está tentando prever no ML supervisionado. Ela também é conhecida como variável de resultado. Por exemplo, em uma configuração de fabricação, a variável-alvo pode ser um defeito do produto.

lista de tarefas

Uma ferramenta usada para monitorar o progresso por meio de um runbook. Uma lista de tarefas contém uma visão geral do runbook e uma lista de tarefas gerais a serem concluídas. Para cada tarefa geral, ela inclui o tempo estimado necessário, o proprietário e o progresso.

ambiente de teste

Veja ambiente.

treinamento

O processo de fornecer dados para que seu modelo de ML aprenda. Os dados de treinamento devem conter a resposta correta. O algoritmo de aprendizado descobre padrões nos dados de treinamento que mapeiam os atributos dos dados de entrada no destino (a resposta que você deseja prever). Ele gera um modelo de ML que captura esses padrões. Você pode usar o modelo de ML para obter previsões de novos dados cujo destino você não conhece.

ferramenta

Uma função ou API que um agente pode invocar para realizar operações em sistemas externos.

gateway de trânsito

Um hub de trânsito de rede que pode ser usado para interconectar as VPCs e as redes on-premises. Para obter mais informações, consulte O que é um gateway de trânsito na AWS Transit Gateway documentação.

fluxo de trabalho baseado em troncos

Uma abordagem na qual os desenvolvedores criam e testam recursos localmente em uma ramificação de recursos e, em seguida, mesclam essas alterações na ramificação principal. A ramificação principal é então criada para os ambientes de desenvolvimento, pré-produção e produção, sequencialmente.

Acesso confiável

Conceder permissões a um serviço que você especifica para realizar tarefas em sua organização AWS Organizations e em suas contas em seu nome. O serviço confiável cria um perfil vinculado ao serviço em cada conta, quando esse perfil é necessário, para realizar tarefas de gerenciamento para você. Para obter mais informações, consulte Usando AWS Organizations com outros AWS serviços na AWS Organizations documentação.

tuning (ajustar)

Alterar aspectos do processo de treinamento para melhorar a precisão do modelo de ML. Por exemplo, você pode treinar o modelo de ML gerando um conjunto de rótulos, adicionando rótulos e repetindo essas etapas várias vezes em configurações diferentes para otimizar o modelo.

equipe de duas pizzas

Uma pequena DevOps equipe que você pode alimentar com duas pizzas. Uma equipe de duas pizzas garante a melhor oportunidade possível de colaboração no desenvolvimento de software.

U

incerteza

Um conceito que se refere a informações imprecisas, incompletas ou desconhecidas que podem minar a confiabilidade dos modelos preditivos de ML. Há dois tipos de incertezas: a incerteza epistêmica é causada por dados limitados e incompletos, enquanto a incerteza aleatória é causada pelo ruído e pela aleatoriedade inerentes aos dados.

tarefas indiferenciadas

Também conhecido como trabalho pesado, trabalho necessário para criar e operar um aplicativo, mas que não fornece valor direto ao usuário final nem oferece vantagem competitiva. Exemplos de tarefas indiferenciadas incluem aquisição, manutenção e planejamento de capacidade.

ambientes superiores

Veja ambiente.

V

aspiração

Uma operação de manutenção de banco de dados que envolve limpeza após atualizações incrementais para recuperar armazenamento e melhorar a performance.

controle de versões

Processos e ferramentas que rastreiam mudanças, como alterações no código-fonte em um repositório.

emparelhamento de VPC

Uma conexão entre duas VPCs que permite rotear tráfego usando endereços IP privados. Para ter mais informações, consulte O que é emparelhamento de VPC? na documentação da Amazon VPC.

Vulnerabilidade

Uma falha de software ou hardware que compromete a segurança do sistema.

W

cache quente

Um cache de buffer que contém dados atuais e relevantes que são acessados com frequência. A instância do banco de dados pode ler do cache do buffer, o que é mais rápido do que ler da memória principal ou do disco.

dados mornos

Dados acessados raramente. Ao consultar esse tipo de dados, consultas moderadamente lentas geralmente são aceitáveis.

função de janela

Uma função SQL que executa um cálculo em um grupo de linhas que se relacionam de alguma forma com o registro atual. As funções de janela são úteis para processar tarefas, como calcular uma média móvel ou acessar o valor das linhas com base na posição relativa da linha atual.

workload

Uma coleção de códigos e recursos que geram valor empresarial, como uma aplicação voltada para o cliente ou um processo de backend.

workstreams

Grupos funcionais em um projeto de migração que são responsáveis por um conjunto específico de tarefas. Cada workstream é independente, mas oferece suporte aos outros workstreams do projeto. Por exemplo, o workstream de portfólio é responsável por priorizar aplicações, planejar ondas e coletar metadados de migração. O workstream de portfólio entrega esses ativos ao workstream de migração, que então migra os servidores e as aplicações.

WORM

Veja gravação única e várias leituras.

WQF

Veja AWS Workload Qualification Framework.

gravação única e várias leituras (WORM)

Um modelo de armazenamento que grava dados uma única vez e evita que os dados sejam excluídos ou modificados. Os usuários autorizados podem ler os dados quantas vezes forem necessárias, mas não podem alterá-los. Essa infraestrutura de armazenamento de dados é considerada imutável.

Z

exploração de dia zero

Um ataque, normalmente malware, que tira proveito de uma vulnerabilidade zero-day.

vulnerabilidade de dia zero

Uma falha ou vulnerabilidade não mitigada em um sistema de produção. Os agentes de ameaças podem usar esse tipo de vulnerabilidade para atacar o sistema. Os desenvolvedores frequentemente ficam cientes da vulnerabilidade como resultado do ataque.

prompt zero shot

Fornecer a um LLM instruções para realizar uma tarefa, mas sem exemplos (shots) que possam ajudar a orientá-lo. O LLM deve usar seu conhecimento pré-treinado para lidar com a tarefa. A eficácia dos prompts zero-shot depende da complexidade da tarefa e da qualidade do prompt. Veja também prompts few-shot.

aplicação zumbi

Uma aplicação que tem um uso médio de CPU e memória inferior a 5%. Em um projeto de migração, é comum retirar essas aplicações.

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