Aplicação do AWS Well-Architected Framework para o Amazon Neptune Analytics

AWS Orientação prescritiva

AWS Orientação prescritiva: Aplicação do AWS Well-Architected Framework para o Amazon Neptune Analytics

Table of Contents

Aplicação do AWS Well-Architected Framework para o Amazon Neptune Analytics

Michael Havey, Amazon Web Services ()AWS

Dezembro de 2024 (histórico do documento)

Você pode criar soluções baseadas em gráficos na Amazon Web Services (AWS) usando o Amazon Neptune. O Neptune inclui o Neptune Analytics, um mecanismo de análise gráfica otimizado para memória que pode analisar rapidamente grandes quantidades de dados gráficos para obter insights e encontrar tendências. Ele pode realizar análises em dados em seu cluster de banco de dados Neptune existente ou você pode carregar e analisar dados de conjuntos de dados externos. Este guia fornece orientação prescritiva para aplicar os princípios do AWS Well-Architected Framework ao planejar sua implantação do Neptune Analytics. A aplicação do AWS Well-Architected Framework para o Amazon Neptune abrange o mesmo tópico para um banco de dados Neptune.

O AWS Well-Architected Framework ajuda você a criar infraestruturas seguras, de alto desempenho, resilientes e eficientes para uma variedade de aplicativos e cargas de trabalho. Ele também fornece uma abordagem consistente para você avaliar arquiteturas e implementar projetos escaláveis.

O AWS Well-Architected Framework é construído em torno dos seguintes seis pilares:

- Excelência operacional
- Segurança
- Confiabilidade
- Eficiência de desempenho
- Otimização de custo
- Sustentabilidade

Este guia fornece informações sobre os pilares de design e as melhores práticas do Well-Architected Framework, além de considerações que você deve ter em mente ao implantar o Neptune Analytics no. AWS

Público-alvo

Este guia é destinado a engenheiros de dados, arquitetos de soluções e analistas de dados que projetam e implementam soluções que usam gráficos em AWS.

Objetivos

Este guia pode ajudar você e sua organização a fazer o seguinte:

- Escolha entre as opções de implantação suportadas.
- Siga os padrões de design do AWS Well-Architected que ajudam você a melhorar a resiliência e a segurança.
- Crie suas consultas para otimizar o desempenho e a economia de custos.
- Saiba como ser operacionalmente eficiente ao gerenciar seu gráfico do Neptune Analytics na produção.

Pilar Excelência operacional

O pilar de excelência operacional do AWS Well-Architected Framework se concentra na execução e monitoramento de sistemas e na melhoria contínua de processos e procedimentos. Isso inclui a capacidade de apoiar o desenvolvimento e executar cargas de trabalho de forma eficaz, obter informações sobre suas operações e melhorar continuamente os processos e procedimentos de suporte para agregar valor comercial. Você pode reduzir a complexidade operacional por meio de cargas de trabalho de autorrecuperação, que detectam e solucionam a maioria dos problemas sem intervenção humana. Você pode trabalhar em direção a essa meta seguindo as melhores práticas descritas nesta seção e usando as métricas APIs e os mecanismos do Amazon Neptune Analytics para responder adequadamente quando sua carga de trabalho se desvia do comportamento esperado.

Essa discussão sobre o pilar de excelência operacional se concentra nas seguintes áreas principais:

- Infraestrutura como código (IaC)
- Gerenciamento de alterações
- Estratégias de resiliência
- Gerenciamento de incidentes
- Relatórios de auditoria para conformidade
- Registro em log e monitoramento

Automatize a implantação usando uma abordagem de IaC

As melhores práticas para automatizar a implantação no Neptune usando o IaC incluem o seguinte:

- Aplique o IaC para implantar gráficos do Neptune Analytics e recursos relacionados. Para uma configuração consistente do ambiente, use o suporte ao Neptune Analytics fornecido AWS CloudFormationpela para provisionar gráficos e endpoints privados.
- Use CloudFormation para provisionar instâncias do notebook Neptune na Amazon AI. SageMaker Você pode usar notebooks para consultar e visualizar dados em um gráfico do Neptune Analytics.
- Quando você cria um gráfico do Neptune Analytics a partir de uma fonte existente, como um cluster ou snapshot de banco de dados do Neptune, ou arquivos de dados armazenados no Amazon Simple Storage Service (Amazon S3), monitore a tarefa de importação em massa.

- Automatize os procedimentos operacionais do Neptune Analytics, como redimensionar o gráfico, excluir e capturar o gráfico, restaurar o gráfico a partir de um instantâneo e redefinir e recarregar o gráfico. Use a API Neptune Analytics, que está disponível por meio AWS Command Line Interface do AWS CLI() ou do SDK s.
- Avalie o tempo de atividade necessário do seu gráfico. A análise geralmente é efêmera; o gráfico é necessário somente durante o tempo necessário para executar algoritmos. Se for esse o caso, use o AWS CLI ou SDKs para capturar um instantâneo e excluir o gráfico quando ele não for mais necessário. Em seguida, você poderá restaurá-lo a partir de um snapshot posteriormente, se necessário.
- Armazene cadeias de conexão externamente do seu cliente. Você pode armazenar cadeias de conexão no AWS Secrets ManagerAmazon DynamoDB ou em qualquer local onde elas possam ser alteradas dinamicamente.
- Use tags para adicionar metadados aos seus recursos do Neptune Analytics e monitore o uso com base em tags. As tags ajudam a organizar seus recursos. Por exemplo, você pode aplicar uma tag comum a recursos em um ambiente ou aplicativo específico. Você também pode usar tags para analisar o faturamento do uso de recursos; para obter mais informações, consulte Organização e controle de custos usando tags de alocação de AWS custos no Guia do usuário de AWS faturamento. Além disso, você pode usar condições em suas políticas AWS Identity and Access Management (IAM) para controlar o acesso aos AWS recursos com base nas tags usadas nesse recurso. Faça isso usando a chave de condição global `aws:ResourceTag/tag-key`. Para obter mais informações, consulte Controle do acesso aos AWS recursos no Guia do usuário do IAM.

Design das operações

Adote abordagens para melhorar a forma como você opera os gráficos do Neptune Analytics:

- Mantenha gráficos separados do Neptune Analytics para uso em desenvolvimento, teste e produção. Esses gráficos podem ter conjuntos de dados, usuários e controles operacionais diferentes.
- Mantenha gráficos separados do Neptune Analytics para diferentes usos. Por exemplo, se dois grupos de usuários analíticos precisarem de gráficos separados com cronogramas, modelos, desempenho, disponibilidade SLAs e padrões de uso diferentes, mantenha gráficos separados para cada grupo.

- Prepare os usuários e a equipe operacional para as atualizações de manutenção do Neptune Analytics.

Faça mudanças frequentes, pequenas e reversíveis

As recomendações a seguir se concentram em mudanças pequenas e reversíveis que você pode fazer para minimizar a complexidade e reduzir a probabilidade de interrupção da carga de trabalho:

- Armazene modelos e scripts de IaC em um serviço de controle de origem, como GitHub ou GitLab.

> **Important**
>
> Não armazene AWS credenciais no controle de origem.

- Exija que as implantações de IaC usem um serviço de integração contínua e entrega contínua (CI/CD), como ou. AWS CodeDeployAWS CodeBuild Compile, teste e implante código em um ambiente de não produção do Neptune Analytics antes de promovê-lo a um gráfico de produção.

Implemente a observabilidade para obter insights acionáveis

Obtenha uma compreensão abrangente do comportamento, desempenho, confiabilidade, custo e integridade da carga de trabalho. As recomendações a seguir ajudam você a obter esse nível de compreensão no Neptune Analytics:

- Monitore CloudWatch as métricas da Amazon para o Neptune Analytics. A partir dessas métricas, você pode determinar o tamanho de um gráfico (número de nós, bordas e vetores, mais o tamanho total do byte), a utilização da CPU e as taxas de solicitação e erro de consulta.
- Crie CloudWatch painéis e alarmes para as principais métricas`NumQueuedRequestsPerSec`, como,, `NumOpenCypherRequestsPerSec` `GraphStorageUsagePercentGraphSizeBytes`, e também para `CPUUtilization` as respostas do cliente Neptune encontradas nos registros do seu aplicativo.
- Defina notificações para monitorar a integridade do gráfico do Neptune Analytics, como quando o tamanho do gráfico, a taxa de solicitações ou a utilização da CPU excedem seu limite. Por exemplo, se `GraphStorageUsagePercent` subiu para 90 por cento em um gráfico que você pretende aumentar significativamente, decida se deseja aumentar a capacidade da Unidade de Capacidade de Netuno (m-NCU) otimizada para memória. Se o m-NCU atual for 128,

aumentá-lo para 256 reduzirá o armazenamento em cerca de 45 por cento. Se geralmente `NumQueuedRequestsPerSec` for maior que zero, considere aumentar a capacidade m-NCU para fornecer mais capacidade computacional. Como alternativa, você pode reduzir a simultaneidade do lado do cliente.

Aprenda com todas as falhas operacionais

Uma infraestrutura de autorrecuperação é um esforço de longo prazo que se desenvolve em iterações à medida que problemas raros ocorrem ou as respostas não são tão eficazes quanto o desejado. A adoção das seguintes práticas impulsiona o foco em direção a essa meta:

- Promova a melhoria aprendendo com todas as falhas.
- Compartilhe o que foi aprendido entre as equipes e a organização. Se várias equipes da sua organização usarem o Neptune, crie uma sala de bate-papo ou um grupo de usuários comum para compartilhar aprendizados e melhores práticas.

Use recursos de registro para monitorar atividades não autorizadas ou anômalas

Use o registro para observar padrões anômalos de desempenho e atividade. Considere as seguintes práticas recomendadas:

- O Neptune Analytics suporta o registro de ações do plano de controle usando. AWS CloudTrail Para obter mais informações, consulte Como registrar chamadas da API Neptune Analytics usando. AWS CloudTrail Por meio desse recurso, você pode acompanhar a criação, atualização e exclusão dos recursos do Neptune Analytics. Para monitoramento e alertas robustos, você também pode integrar CloudTrail eventos com o Amazon CloudWatch Logs. Para aprimorar sua análise da atividade do serviço Neptune Analytics e identificar alterações nas atividades de Conta da AWS um, você pode CloudTrail consultar registros usando o Amazon Athena. Por exemplo, é possível usar consultas para identificar tendências e isolar ainda mais a atividade por atributos, como endereço IP de origem ou usuário.
- Você também pode usar CloudTrail para habilitar o registro de atividades do plano de dados do Neptune Analytics, como execuções de consultas. Você pode ver quais consultas estão sendo executadas, sua frequência e sua origem. Por padrão, CloudTrail não registra eventos de dados.

Há cobranças adicionais para eventos de dados. Para obter mais informações, consulte Preços do AWS CloudTrail.

- Você também pode registrar suas chamadas de aplicativo no Neptune Analytics no plano de controle ou no plano de dados. Por exemplo, se você usar o AWS SDK para Python (Boto3)para fazer consultas, poderá ativar o registro em nível de depuração para obter um rastreamento das consultas no console ou no arquivo. Isso é útil durante o desenvolvimento. Também recomendamos que você capture e registre exceções do seu aplicativo.

Pilar Segurança

A segurança na nuvem é a maior prioridade em AWS. Como AWS cliente, você se beneficia de uma arquitetura de data center e rede criada para atender aos requisitos das organizações mais sensíveis à segurança. A segurança é uma responsabilidade compartilhada entre a AWS e você. O modelo de responsabilidade compartilhada descreve isto como segurança da nuvem e segurança na nuvem.

- Segurança da nuvem — AWS é responsável por proteger a infraestrutura que funciona Serviços da AWS no Nuvem AWS. AWS também fornece serviços que você pode usar com segurança. Auditores terceirizados testam e verificam regularmente a eficácia da AWS segurança como parte dos programas de AWS conformidade. Para saber mais sobre os programas de conformidade que se aplicam ao Neptune, AWS consulte Serviços no escopo por programa de conformidade.
- Segurança na nuvem — Sua responsabilidade é determinada pelo AWS service (Serviço da AWS) que você usa. Você também é responsável por outros fatores, incluindo a confidencialidade dos dados, os requisitos da empresa e as leis e regulamentos aplicáveis. Para obter mais informações sobre privacidade de dados, consulte Privacidade de dados FAQs. Para obter informações sobre proteção de dados na Europa, consulte o Modelo de Responsabilidade AWS Compartilhada e a postagem no blog do GDPR.

O pilar de segurança do AWS Well-Architected Framework ajuda você a entender como aplicar o modelo de responsabilidade compartilhada ao usar o Neptune Analytics. Os tópicos a seguir explicam como configurar o Neptune Analytics para atender aos seus objetivos de segurança e conformidade. Você também aprende a usar outros Serviços da AWS que ajudam a monitorar e proteger seus recursos do Neptune Analytics. O pilar de segurança inclui as seguintes áreas principais de foco:

- Segurança de dados
- Segurança de rede
- Autenticação e autorização

Implemente a segurança de dados

Vazamentos e violações de dados colocam seus clientes em risco e podem causar um impacto negativo substancial em sua empresa. As práticas recomendadas a seguir ajudam a proteger os dados de seus clientes contra exposição inadvertida e maliciosa:

- Nomes de gráficos, tags, funções do IAM e outros metadados não devem conter informações confidenciais ou sigilosas, pois esses dados podem aparecer em registros de faturamento ou diagnóstico.
- URIs ou links para servidores externos armazenados como dados no Neptune não devem conter informações de credenciais para validar solicitações.
- Um gráfico do Neptune Analytics é criptografado em repouso. Você pode usar a chave padrão ou uma chave AWS Key Management Service (AWS KMS) de sua escolha para criptografar o gráfico. Você também pode criptografar snapshots e dados que são exportados para o Amazon S3 durante a importação em massa. Você pode remover a criptografia quando a importação for concluída.
- Ao usar a linguagem OpenCypher, pratique técnicas adequadas de validação e parametrização de entrada para evitar a injeção de SQL e outras formas de ataques. Evite criar consultas que usem concatenação de strings com entrada fornecida pelo usuário. Use consultas parametrizadas ou instruções preparadas para passar com segurança os parâmetros de entrada para o banco de dados gráfico. Para obter mais informações, consulte Exemplos de consultas parametrizadas do OpenCypher na documentação do Neptune.

Proteja suas redes

Você pode habilitar um gráfico do Neptune Analytics para conectividade pública para que ele possa ser acessado de fora de uma nuvem privada virtual (VPC). Essa conectividade está desativada por padrão. O gráfico exige autenticação do IAM. O chamador deve obter uma identidade e ter permissões para usar o gráfico. Por exemplo, para executar uma consulta OpenCypher, o chamador precisaria ter permissões de leitura, gravação ou exclusão no gráfico específico.

Você também pode criar endpoints privados para o gráfico para acessar o gráfico de dentro de uma VPC. Ao criar o endpoint, você especifica a VPC, as sub-redes e os grupos de segurança para restringir o acesso e chamar o gráfico.

Para proteger seus dados em trânsito, o Neptune Analytics impõe conexões SSL por meio de HTTPS ao gráfico. Para obter mais informações, consulte Proteção de dados no Neptune Analytics na documentação do Neptune Analytics.

Implemente autenticação e autorização

As chamadas para um gráfico do Neptune Analytics exigem autenticação do IAM. O chamador deve obter uma identidade e ter permissões suficientes para realizar a ação no gráfico. Para obter

descrições das ações da API e das permissões necessárias, consulte a documentação da API Neptune Analytics. Você pode aplicar verificações de condições para restringir o acesso por tag.

A autenticação do IAM usa o AWS protocolo Signature Version 4 (SigV4). Para simplificar o uso do seu aplicativo, recomendamos que você use um AWS SDK. Por exemplo, em Python, use o cliente Boto3 para o Neptune Graph, que abstrai o SigV4.

Quando você carrega dados no gráfico, o carregamento em lote usa as credenciais do IAM do chamador. O chamador deve ter permissões para baixar dados do Amazon S3 com a relação de confiança configurada para que o Neptune Analytics possa assumir a função de carregar os dados no gráfico a partir dos arquivos do Amazon S3.

A importação em massa pode ser realizada durante a criação do gráfico (pela equipe de infraestrutura) ou em um gráfico vazio existente (pela equipe de engenharia de dados que tem permissões para iniciar as tarefas de importação). Em ambos os casos, o Neptune Analytics assume a função do IAM que o chamador fornece como entrada. Essa função lhe dá permissão para ler e listar o conteúdo da pasta Amazon S3 onde os dados de entrada são armazenados.

Pilar Confiabilidade

O pilar de confiabilidade do AWS Well-Architected Framework engloba a capacidade de uma carga de trabalho realizar a função pretendida de forma correta e consistente, quando se espera que o faça. Isso inclui a capacidade de operar e testar a carga de trabalho durante todo o seu ciclo de vida.

Uma workload confiável começa com as decisões iniciais de projeto que envolvem tanto o software quanto a infraestrutura. Suas escolhas de arquitetura afetam o comportamento da carga de trabalho em todos os pilares do Well-Architected. Para maior confiabilidade, há padrões específicos que você deve seguir, conforme discutido nesta seção.

O pilar de confiabilidade se concentra nas seguintes áreas principais:

- Arquitetura de carga de trabalho, incluindo cotas de serviço e padrões de implantação
- Gerenciamento de alterações
- Gerenciamento de falhas

Entenda as cotas de serviço do Neptune

Sua AWS conta tem cotas padrão (anteriormente chamadas de limites) para cada uma. AWS service (Serviço da AWS) A menos que especificado de outra forma, cada cota é específica da região. Você pode solicitar aumentos para algumas cotas, mas não para todas.

Para encontrar cotas para o Neptune Analytics, abra o console Service Quotas. No painel de navegação, escolha e, em seguida Serviços da AWS, selecione Amazon Neptune Analytics. Preste atenção às cotas no número de gráficos e instantâneos, na memória máxima provisionada para um gráfico e nas taxas de solicitação de API.

Se a memória máxima provisionada não for suficiente para seu conjunto de dados, avalie quais tipos de nós e bordas são essenciais para o uso analítico pretendido. Carregue um subconjunto dos dados para que a análise seja possível dentro de uma capacidade provisionada permitida. Muitas cargas de trabalho de análise, especialmente aquelas que executam algoritmos gráficos, precisam apenas da topologia com um conjunto limitado de propriedades, em vez do gráfico transacional completo. (Para uma discussão sobre as diferenças entre cargas de trabalho transacionais e analíticas, consulte a seção Pilar de eficiência de desempenho.)

Se o número máximo de gráficos não for suficiente para o uso pretendido:

- Considere combinar gráficos que tenham usos semelhantes.
- Avalie quantos gráficos precisam ser executados em um determinado momento. Se você tiver um caso de uso de análise efêmero, capture e exclua um gráfico quando ele não for mais necessário. Isso reduz o número de gráficos em relação à cota.
- Considere gráficos de provisionamento em diferentes. Contas da AWS

Entenda os padrões de implantação do Neptune

Entenda os seguintes pontos de decisão ao planejar a implantação de um gráfico do Neptune Analytics:

- Semeadura: decida se deseja criar um gráfico vazio ou carregar dados nele no momento da criação com dados do Amazon S3, de um cluster de banco de dados Neptune existente ou de um snapshot de banco de dados Neptune existente.

  Recomendação: se a origem for um cluster ou snapshot do Neptune, você deverá carregar seus dados no momento da criação do gráfico. Se a fonte for o Amazon S3, carregue os dados no momento da criação se o esforço de carregá-los for significativo e melhor executado como uma atividade de provisionamento de infraestrutura. Se você preferir carregar dados como uma atividade de engenharia de dados ou de aplicação, crie um gráfico vazio e carregue dados do Amazon S3 posteriormente.
- Capacidade: estime a capacidade provisionada necessária para um gráfico, considerando o tamanho dos dados e o uso esperado do aplicativo.

  Recomendação: no momento da criação, especifique a memória máxima provisionada para limitar o tamanho do gráfico. Essa configuração é obrigatória. Você pode alterar a capacidade posteriormente, se necessário.
- Disponibilidade e tolerância a falhas: decida se as réplicas são necessárias para garantir a disponibilidade. Uma réplica atua como um modo de espera para recuperação em caso de falha no gráfico. Um gráfico com réplicas se recupera mais rápido do que um gráfico sem réplicas. Considere também por quanto tempo o gráfico é necessário, se é apenas para análises efêmeras e, em caso afirmativo, quando será removido.

  Recomendação: determine os requisitos de disponibilidade, como por quanto tempo o gráfico pode ficar indisponível e quando ele pode ser removido, antes de criar um gráfico.

- Rede e segurança: determine se você precisa de conectividade pública, privada ou ambas e se deseja criptografar seus dados.

  Recomendação: entenda os requisitos organizacionais, como se a conectividade pública é permitida e onde os aplicativos cliente gráficos serão implantados, antes de criar um gráfico.
- Backups e recuperação: determine se os instantâneos devem ser criados e, em caso afirmativo, quando ou sob quais condições. Considere se sua organização tem requisitos de recuperação de desastres (DR).

  Recomendação: criar instantâneos é uma atividade manual. Decida quando criar instantâneos e considere seus requisitos de DR antes de criar um gráfico.

Gerencie e escale os clusters do Neptune

Um gráfico do Neptune Analytics consiste em uma única instância otimizada para memória. A capacidade (m-NCU) da instância é definida no momento da criação. A instância pode ser escalada verticalmente aumentando a capacidade provisionada por meio de uma ação administrativa; a capacidade provisionada também pode ser reduzida. As réplicas são alvos passivos de failover, portanto, não aumentam a escala de um gráfico. Nesse aspecto, uma réplica gráfica difere de uma réplica de leitura do banco de dados Neptune, que é uma instância ativa em um cluster do Neptune que pode processar operações de leitura de aplicativos.

As réplicas têm custos. O preço da réplica é baseado na taxa m-NCU do gráfico. Por exemplo, se um gráfico for provisionado para 128 m-NCU e tiver uma única réplica, o custo será o dobro de um gráfico equivalente sem réplicas.

Na análise, há dois motivos principais para aumentar a escala:

- Para fornecer mais memória e CPU para consultas e algoritmos analíticos, porque a consulta individual é cara, o algoritmo gráfico a ser executado é inerentemente complexo e requer mais recursos de acordo com sua entrada, ou a taxa de solicitações simultâneas é alta. Se essas consultas encontrarem out-of-memory erros, aumentar a escala é uma solução razoável.
- Para suportar um tamanho de gráfico maior do que o planejado. Por exemplo, se a capacidade provisionada atual for de 128 M-NCU para suportar 60 GB de dados de origem e você precisar de mais 40 GB de dados de origem, um aumento para 256 M-NCU é garantido.

Monitore CloudWatch métricas do Neptune Analytics, `NumQueuedRequestsPerSec` como,, `CPUUtilization` e `NumOpenCypherRequestsPerSec` `GraphStorageUsagePercentGraphSizeBytes`, para determinar se o dimensionamento é necessário. Você pode atualizar a configuração de um gráfico por meio do console ou SDKs. AWS CLI(Para exemplos e melhores práticas, consulte a seção Pilar de excelência operacional.)

Gerencie backups e eventos de failover

Use réplicas para garantir que um gráfico permaneça disponível em caso de falha. Um gráfico usa persistência baseada em registros para confirmar alterações nas zonas de disponibilidade em um. Região da AWS A réplica funciona como um modo de espera dinâmico e tem acesso a esses dados. Se houver uma falha, o gráfico retoma as operações na réplica. O aplicativo continua usando o mesmo endpoint para se conectar ao gráfico. Solicitações em andamento durante a falha geram erros com uma exceção de serviço indisponível. Considere usar uma nova tentativa com padrão de recuo no código do aplicativo para capturar o erro e tente novamente após um breve intervalo. Novas solicitações feitas durante o failover ficam na fila e podem ter uma latência maior.

Se nenhuma réplica estiver configurada e o gráfico falhar, o Neptune Analytics se recuperará do armazenamento durável, mas a recuperação demorará mais porque o Neptune precisa reinicializar os recursos.

Crie instantâneos do gráfico. (O Neptune Analytics não tira instantâneos automáticos.) Se o gráfico for modificado regularmente após a criação, tire instantâneos frequentes para capturar seu estado atual. Exclua instantâneos antigos se a restauração para um momento anterior não for necessária.

Você pode compartilhar instantâneos com outras contas e entre Regiões da AWS si. Se você tiver requisitos de DR, considere se a restauração do gráfico em uma região diferente de um snapshot atende aos requisitos de objetivo de tempo de recuperação (RTO) e objetivo de ponto de recuperação (RPO).

Pilar Eficiência de performance

O pilar de eficiência de desempenho do AWS Well-Architected Framework se concentra em como otimizar o desempenho ao ingerir ou consultar dados. A otimização do desempenho é um processo incremental e contínuo do seguinte:

- Confirmando os requisitos de negócios
- Medindo o desempenho da carga de trabalho
- Identificação de componentes com baixo desempenho
- Ajustando componentes para atender às suas necessidades de negócios

O pilar de eficiência de desempenho fornece diretrizes específicas para casos de uso que podem ajudá-lo a identificar o modelo de dados gráficos e as linguagens de consulta corretos a serem usados. Também inclui as melhores práticas a serem seguidas ao ingerir e consumir dados do Neptune Analytics.

O pilar de eficiência de desempenho se concentra nas seguintes áreas principais:

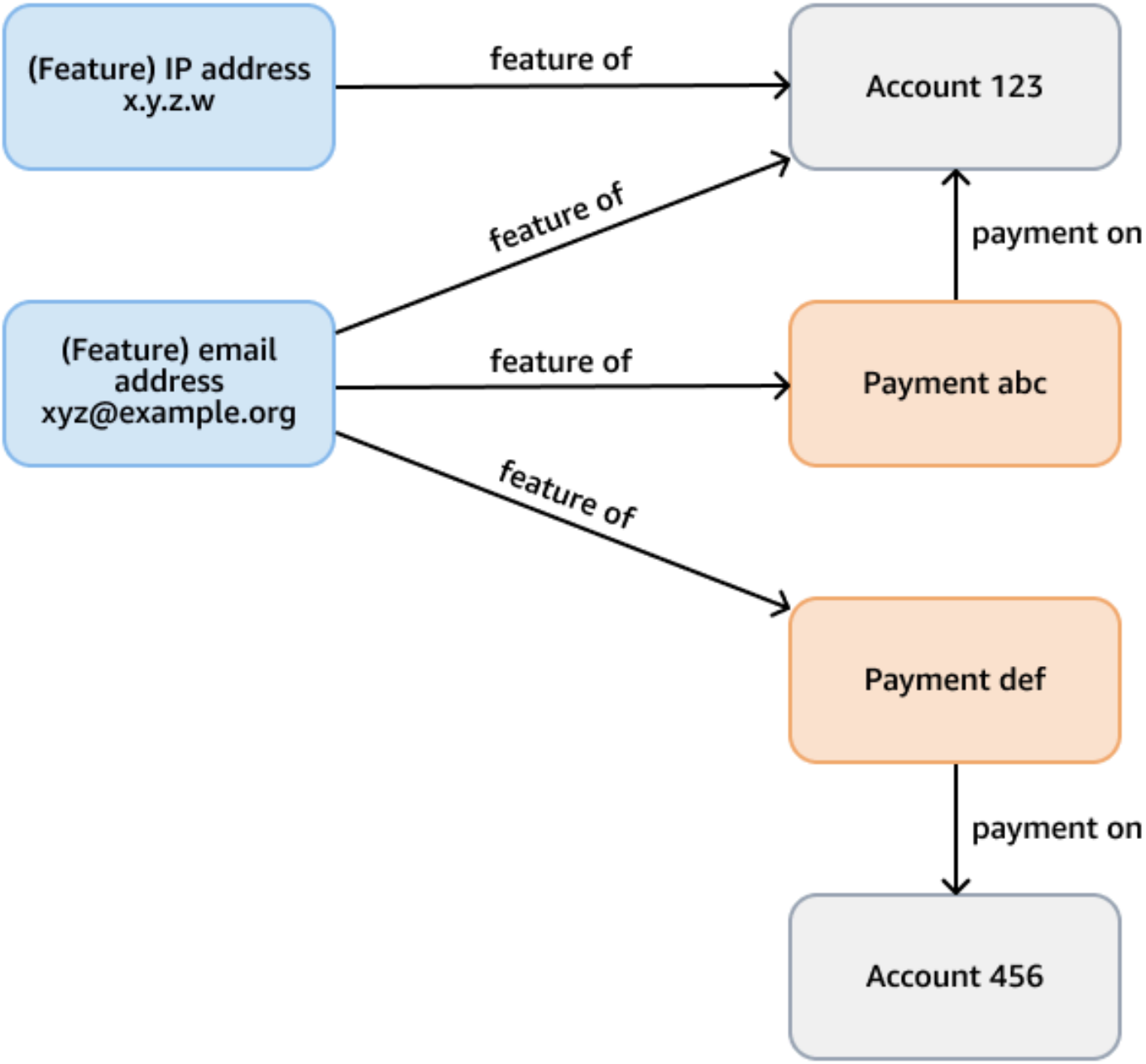
- Modelagem gráfica
- Otimização de consultas
- Dimensionamento correto do gráfico
- Otimização de gravação

Entenda a modelagem gráfica para análise

O guia Applying the AWS Well-Architected Framework for Amazon Neptune discute a modelagem gráfica para eficiência de desempenho. As decisões de modelagem que afetam o desempenho incluem escolher quais nós e bordas são necessários IDs, seus rótulos e propriedades, a direção das bordas, se os rótulos devem ser genéricos ou específicos e, em geral, com que eficiência o mecanismo de consulta pode navegar pelo gráfico para processar consultas comuns.

Essas considerações também se aplicam ao Neptune Analytics; no entanto, é importante distinguir entre padrões de uso transacional e analítico. Um modelo gráfico eficiente para consultas em um banco de dados transacional, como um banco de dados Neptune, talvez precise ser reformulado para fins analíticos.

Por exemplo, considere um gráfico de fraude em um banco de dados do Neptune, cujo objetivo é verificar padrões fraudulentos em pagamentos com cartão de crédito. Esse gráfico pode ter nós que representam contas, pagamentos e recursos (como endereço de e-mail, endereço IP, número de telefone) da conta e do pagamento. Esse gráfico conectado suporta consultas como percorrer um caminho de comprimento variável que começa com um determinado pagamento e leva vários passos para encontrar recursos e contas relacionados. A figura a seguir mostra esse gráfico.



O requisito analítico pode ser mais específico, como encontrar comunidades de contas vinculadas por um recurso. Você pode usar o algoritmo de componentes fracamente conectados (WCC)

para essa finalidade. Executá-lo em relação ao modelo do exemplo anterior é ineficiente, pois ele precisa atravessar vários tipos diferentes de nós e bordas. O modelo no diagrama a seguir é mais eficiente. Ele vincula `account` os nós a uma `shares feature` vantagem se as próprias contas — ou os pagamentos das contas — compartilharem um recurso. Por exemplo, `Account 123` tem o recurso `xyz@example.org` de e-mail e `Account 456` usa esse mesmo e-mail para um pagamento (`Payment def`).

A complexidade computacional do WCC é`O(|E|logD)`: onde `|E|` está o número de arestas no gráfico e `D` o diâmetro (o comprimento do caminho mais longo) que conecta os nós. Como o modelo transacional omite nós e bordas não essenciais, ele otimiza o número de arestas e o diâmetro e reduz a complexidade do algoritmo WCC.

Ao usar o Neptune Analytics, use os algoritmos e as consultas analíticas necessárias. Se necessário, reformule o modelo para otimizar essas consultas. Você pode remodelar o modelo antes de carregar dados no gráfico ou escrever consultas que modifiquem os dados existentes no gráfico.

Otimizar consultas

Siga estas recomendações para otimizar as consultas do Neptune Analytics:

- Use consultas parametrizadas e o cache do plano de consulta, que é ativado por padrão. Quando você usa o cache do plano, o mecanismo prepara a consulta para uso posterior, desde que a consulta seja concluída em 100 milissegundos ou menos, o que economiza tempo em invocações subsequentes.
- Para consultas lentas, execute um plano de explicação para identificar gargalos e fazer as melhorias necessárias.
- Se você usar a pesquisa de similaridade vetorial, decida se incorporações menores produzem resultados de similaridade precisos. Você pode criar, armazenar e pesquisar incorporações menores com mais eficiência.

- Siga as melhores práticas documentadas para usar o OpenCypher no Neptune Analytics. Por exemplo, use mapas nivelados em uma cláusula UNWIND e especifique rótulos de borda sempre que possível.
- Ao usar um algoritmo gráfico, entenda as entradas e saídas do algoritmo, sua complexidade computacional e, em geral, como ele funciona.
  - Antes de chamar um algoritmo gráfico, use uma `MATCH` cláusula para minimizar o conjunto de nós de entrada. Por exemplo, para limitar a pesquisa abrangente (BFS) a partir dos nós, siga os exemplos fornecidos na documentação do Neptune Analytics.
  - Filtre as etiquetas dos nós e das bordas, se possível. Por exemplo, o BFS tem parâmetros de entrada para filtrar a passagem para um rótulo de nó específico (`vertexLabel`) ou rótulos de borda específicos (). `edgeLabels`
  - Use parâmetros delimitadores, como `maxDepth` para limitar os resultados.
  - Experimente com o `concurrency` parâmetro. Experimente com um valor de 0, que usa todos os encadeamentos de algoritmo disponíveis para paralelizar o processamento. Compare isso com a execução de um único encadeamento definindo o parâmetro como 1. Um algoritmo pode ser concluído mais rapidamente em um único encadeamento, especialmente em entradas menores, como pesquisas superficiais, nas quais o paralelismo não oferece nenhuma redução mensurável no tempo de execução e pode gerar sobrecarga.
  - Escolha entre tipos similares de algoritmos. Por exemplo, Bellman-Ford e delta-stepping são algoritmos de caminho mais curto de fonte única. Ao testar com seu próprio conjunto de dados, experimente os dois algoritmos e compare os resultados. O Delta-Stepping geralmente é mais rápido do que o Bellman-Ford devido à sua menor complexidade computacional. No entanto, o desempenho depende do conjunto de dados e dos parâmetros de entrada, principalmente do `delta` parâmetro.

Otimize as gravações

Siga estas práticas para otimizar as operações de gravação no Neptune Analytics:

- Busque a maneira mais eficiente de carregar dados em um gráfico. Ao carregar dados no Amazon S3, use a importação em massa se os dados tiverem mais de 50 GB de tamanho. Para dados menores, use o carregamento em lote. Se você receber out-of-memory erros ao executar o carregamento em lote, considere aumentar o valor de m-NCU ou dividir a carga em várias solicitações. Uma forma de fazer isso é dividir os arquivos em vários prefixos no bucket do S3. Nesse caso, chame o carregamento em lote separadamente para cada prefixo.

- Use a importação em massa ou o carregador de lotes para preencher o conjunto inicial de dados gráficos. Use as operações transacionais de criação, atualização e exclusão do OpenCypher somente para pequenas alterações.
- Use a importação em massa ou o carregador de lotes com uma simultaneidade de 1 (thread único) para ingerir incorporações no gráfico. Tente carregar as incorporações antecipadamente usando um desses métodos.
- Avalie a dimensão das incorporações vetoriais necessárias para uma pesquisa precisa de similaridade em algoritmos de pesquisa de similaridade vetorial. Use uma dimensão menor, se possível. Isso resulta em uma velocidade de carregamento mais rápida para incorporações.
- Use algoritmos de mutação para lembrar os resultados algorítmicos, se necessário. Por exemplo, o algoritmo de centralidade de mutação de grau encontra o grau de cada nó de entrada e grava esse valor como uma propriedade do nó. Se as conexões ao redor desses nós não mudarem posteriormente, a propriedade manterá o resultado correto. Não há necessidade de executar o algoritmo novamente.
- Use a ação administrativa de redefinição de gráfico para limpar todos os nós, bordas e incorporações se precisar recomeçar. Eliminar todos os nós, bordas e incorporações usando uma consulta OpenCypher não é viável se o gráfico for grande. Uma única consulta suspensa em um grande conjunto de dados pode atingir o tempo limite. Conforme o tamanho aumenta, o conjunto de dados leva mais tempo para ser removido e o tamanho da transação aumenta. Por outro lado, o tempo para concluir uma redefinição gráfica é praticamente constante, e a ação oferece a opção de criar um instantâneo antes de executá-lo.

Gráficos do tamanho certo

O desempenho geral depende da capacidade provisionada de um gráfico do Neptune Analytics. A capacidade é medida em unidades chamadas Unidades de Capacidade de Netuno otimizadas para memória (m-). NCUs Certifique-se de que seu gráfico tenha tamanho suficiente para suportar o tamanho do gráfico e as consultas. Observe que o aumento da capacidade não necessariamente melhora o desempenho de uma consulta individual.

Se possível, crie o gráfico importando dados de uma fonte existente, como o Amazon S3 ou de um cluster ou snapshot existente do Neptune. Você pode colocar limites na capacidade mínima e máxima. Você também pode alterar a capacidade provisionada em um gráfico existente.

Monitore CloudWatch métricas como`NumQueuedRequestsPerSec`,`NumOpenCypherRequestsPerSec`,

`GraphStorageUsagePercent`GraphSizeBytes, e avalie `CPUUtilization` se o gráfico está no tamanho certo. Determine se é necessária mais capacidade para suportar o tamanho e a carga do gráfico. Para obter mais informações sobre como interpretar algumas dessas métricas, consulte a seção Pilar da excelência operacional.

Pilar Otimização de custos

O pilar de otimização de custos do AWS Well-Architected Framework se concentra em evitar custos desnecessários. As recomendações a seguir podem ajudá-lo a atender aos princípios de design de otimização de custos e às melhores práticas arquitetônicas do Neptune Analytics.

O pilar de otimização de custos se concentra nas seguintes áreas principais:

- Compreender os gastos ao longo do tempo e controlar a alocação de fundos
- Seleção de recursos do tipo e quantidade corretos
- Dimensionamento para atender às necessidades de negócios sem gastar demais

Entenda os padrões de uso e os serviços necessários

Antes de adotar o Neptune Analytics, avalie se seu caso de uso é adequado para análise gráfica.

- Bancos de dados gráficos: um banco de dados gráfico como o Neptune é uma boa opção para sua carga de trabalho se seu modelo de dados tiver uma estrutura gráfica perceptível e suas consultas precisarem explorar relacionamentos e atravessar vários saltos. Um banco de dados gráfico não é adequado para os seguintes padrões:
  - Principalmente consultas de salto único. Nesse caso de uso, considere se seus dados podem ser melhor representados como atributos de um objeto.
  - Dados JSON ou binários de objetos grandes (blob) armazenados como propriedades.

- Análise gráfica: o Neptune Analytics é um mecanismo de banco de dados de análise gráfica que pode analisar rapidamente grandes quantidades de dados gráficos na memória para obter insights e encontrar tendências. Você pode armazenar e consultar dados gráficos em um banco de dados do Neptune e em um gráfico do Neptune Analytics. Um banco de dados Neptune é mais adequado para necessidades de processamento transacional on-line escalável (OLTP). O Neptune Analytics é melhor para cargas de trabalho de análise efêmeras. Você pode usar os dois em combinação carregando dados do seu banco de dados Neptune orientado a transações em um gráfico do Neptune Analytics para executar análises desses dados. Quando a análise estiver concluída, você poderá remover o gráfico do Neptune Analytics. Para uma comparação mais detalhada, consulte Quando usar o Neptune Analytics e quando usar o Neptune Database na documentação do Neptune Analytics.

Determine, com atenção ao custo, a melhor forma de preencher seu gráfico do Neptune Analytics.

- Importe dados gráficos em massa que são armazenados em um bucket do S3. Recomendamos essa opção se seus dados foram previamente preparados para carregamento em massa em um banco de dados Neptune, ou se você já tem, ou pode produzir prontamente, os dados a serem analisados em CSV ou em outros formatos compatíveis que a importação em massa exija. Você pode executar a importação em massa como parte do procedimento de criação do gráfico. Você pode colocar limites na capacidade mínima e máxima. Você também pode executar a importação em um gráfico vazio criado anteriormente e monitorar a tarefa de importação enquanto ela é executada.
- Você pode criar um gráfico vazio e, em seguida, preenchê-lo por meio de uma consulta OpenCypher usando o carregamento em lote. Essa opção é ideal se os dados a serem carregados estiverem armazenados no Amazon S3 e tiverem menos de 50 GB.
- Você pode preencher o gráfico a partir dos dados em seu cluster de banco de dados Neptune (compatível com o Neptune Database versão 1.3.0 ou posterior). A intenção desse padrão é executar análises em dados que estão atualmente em seu banco de dados gráfico. Mesmo que o banco de dados tenha sido preenchido inicialmente por meio de carregamento em massa, ele pode ter mudado significativamente desde então. Para importar do banco de dados, o Neptune Analytics clona seu banco de dados e exporta dados do clone para um bucket do S3. Esse procedimento gera custos: principalmente os custos do banco de dados Neptune para executar o clone e os custos do Amazon S3 para armazenar e consumir os dados exportados. O clone é removido quando a exportação é concluída. Você pode excluir os dados exportados no Amazon S3.
- Você pode preencher o gráfico a partir do snapshot de um cluster de banco de dados Neptune. Isso é semelhante à opção anterior, exceto que a origem é um instantâneo do banco de dados. Para importar de um snapshot, o Neptune Analytics primeiro restaura o snapshot em um novo cluster de banco de dados e depois exporta os dados para um bucket do S3. Esse procedimento gera custos: principalmente os custos do banco de dados Neptune para executar o cluster restaurado e os custos do Amazon S3 para armazenar e consumir os dados exportados.
- Você também pode realizar consultas do OpenCypher para criar, atualizar ou excluir dados usando transações compatíveis com atomicidade, consistência, isolamento e durabilidade (ACID) no gráfico. Recomendamos essa abordagem como forma de fazer pequenas atualizações, mas não como forma de semear o gráfico.

Se os dados necessários para análise já estiverem armazenados no Amazon S3, recomendamos a importação em massa ou o carregamento em lote. Eles são mais econômicos do que preencher o gráfico a partir de um cluster ou snapshot do banco de dados Neptune.

Selecione recursos com atenção ao custo

Os preços do Neptune Analytics usam uma unidade conhecida como Unidade de Capacidade de Netuno com otimização de memória (m-NCU). Há um custo fixo por hora para executar um gráfico com um determinado m-NCU. Um gráfico pode ter réplicas para failover, e essas réplicas também geram custos horários de m-NCU.

Recomendamos as seguintes práticas recomendadas para estimar a capacidade, limitar os custos e monitorar os custos em relação ao desempenho:

- Se possível, crie o gráfico importando dados de uma fonte existente: dados armazenados no Amazon S3 ou em um cluster ou snapshot existente do Neptune. Isso economiza seu esforço porque o Neptune Analytics realiza o trabalho pesado de semear o gráfico e você pode especificar uma capacidade máxima limitada.
- Você pode alterar a capacidade provisionada em um gráfico existente.
- Quando o gráfico não for mais necessário, você poderá criar um instantâneo e excluir o gráfico. Se precisar usá-lo novamente, você pode restaurar o gráfico a partir do instantâneo.
- Você pode escolher o número de réplicas ao criar o gráfico. Defina o valor de acordo com seus requisitos de disponibilidade de análise. Economize custos minimizando essa configuração. O valor máximo de 2 permite duas instâncias de réplica em zonas de disponibilidade separadas. O valor mínimo de 0 significa que o Neptune Analytics não executará uma réplica. No entanto, a recuperação é mais rápida quando uma réplica está disponível. Para obter uma explicação sobre falha e recuperação de gráficos, consulte a seção Pilar de confiabilidade.
- Monitore as despesas do Neptune Analytics para períodos de cobrança atuais e passados usando. Gerenciamento de Faturamento e Custos da AWS
- Monitore as métricas do Neptune Analytics CloudWatch para, `NumQueuedRequestsPerSec` especialmente,`NumOpenCypherRequestsPerSec`,,`GraphStorageUsagePercent`, `CPUUtlization` e`GraphSizeBytes`, para avaliar se a capacidade provisionada está dimensionada adequadamente para o gráfico. Determine se uma capacidade menor pode acomodar a taxa de solicitação observada, o uso da CPU e o tamanho do gráfico.

- Se você precisar de um endpoint privado para seu gráfico, preste atenção aos custos de endpoints elásticos IPs de nuvem privada virtual (VPC), gateways NAT ou outros custos relacionados à VPC. Para saber mais, consulte os preços do Amazon VPC e os preços da Amazon EC2.
- Talvez você queira executar uma ou mais instâncias do notebook Neptune para fornecer uma interface de cliente para ajudar desenvolvedores e analistas a consultar e visualizar o gráfico (consulte os preços do Neptune Workbench). Para minimizar os custos, compartilhe a instância entre os usuários e crie pastas de caderno separadas para cada usuário. Encerre a instância quando ela não estiver em uso. Para uma abordagem para automatizar o desligamento, consulte a postagem do AWS blog Automatize a parada e a inicialização dos recursos ambientais do Amazon Neptune usando tags de recursos.

Pilar Sustentabilidade

O pilar de sustentabilidade do AWS Well-Architected Framework se concentra em minimizar os impactos ambientais da execução de cargas de trabalho na nuvem. Os principais tópicos incluem um modelo de responsabilidade compartilhada para sustentabilidade, compreensão do impacto e maximização do uso para minimizar os recursos necessários e reduzir os impactos posteriores.

O pilar de sustentabilidade contém as seguintes áreas de foco principais:

- Seu impacto
- Metas de sustentabilidade
- Uso maximizado
- Antecipando e adotando ofertas de software novas e mais eficientes
- Uso de serviços gerenciados
- Redução do impacto a jusante

Este guia se concentra em entender seu impacto. Para obter mais informações sobre os outros princípios de design de sustentabilidade, consulte o AWS Well-Architected Framework.

Suas escolhas e requisitos têm um impacto no meio ambiente. Se você puder escolher uma Regiões da AWS que tenha menor intensidade de carbono e se seus requisitos refletirem as necessidades reais da carga de trabalho em vez de apenas maximizar o tempo de atividade e a durabilidade, a sustentabilidade da carga de trabalho aumentará. As próximas seções discutem as melhores práticas e considerações que terão um impacto ambiental positivo se adotadas em seu projeto de carga de trabalho e operações contínuas.

Considere sua Região da AWS seleção

Alguns Regiões da AWS estão perto de projetos de energia renovável da Amazon ou localizados onde a rede tem uma intensidade de carbono publicada que é menor do que outros. Considere o impacto da sustentabilidade em regiões que podem ser viáveis para sua carga de trabalho e faça referência cruzada à sua lista com as regiões em que o Neptune Analytics está disponível.

Otimize o consumo

Minimize o consumo do Neptune Analytics praticando o seguinte:

- A análise geralmente é efêmera. O gráfico é necessário apenas para o tempo de execução de algoritmos e registro dos resultados. Se for esse o caso, capture e exclua o gráfico quando ele não for mais necessário. Você pode restaurá-lo a partir de um snapshot posteriormente, se necessário.
- Se a carga de trabalho for efêmera e você tiver a flexibilidade de decidir quando executar a análise, considere as day-to-day tendências no consumo de energia. A demanda por eletricidade é maior em determinados períodos. Se você estiver nos Estados Unidos da América, consulte as métricas sobre o consumo diário de eletricidade no site da Administração de Informações de Energia (EIA) dos EUA. Execute cargas de trabalho fora dos períodos de pico da sua região, se possível.
- Se a carga de trabalho não for efêmera, mas precisar estar disponível somente por períodos limitados, exclua o gráfico e restaure-o de um snapshot quando necessário. Se sua disponibilidade seguir um cronograma, automatize o processo de restauração por meio de scripts para que o gráfico fique pronto no horário programado.
- Se os dados forem somente para leitura ou não tiverem sido alterados desde o último instantâneo, não os capture novamente antes da exclusão.
- Pare os cadernos Neptune quando eles não estiverem em uso.
- Monitore CloudWatch métricas como`NumQueuedRequestsPerSec`,`NumOpenCypherRequestsPerSec`, `GraphStorageUsagePercentGraphSizeBytes`, e avalie `CPUUtilization` se o gráfico está superdimensionado. Determine se uma capacidade de instância menor pode acomodar a taxa de solicitação observada, o uso da CPU e o tamanho do gráfico.

Otimize os padrões de desenvolvimento e arquitetura de software

Para evitar desperdícios, otimize seus modelos e consultas e compartilhe recursos computacionais para usar todos os recursos disponíveis nas instâncias e clusters do Neptune. As melhores práticas específicas incluem:

- Otimize consultas e crie gráficos de invocações de algoritmos. Use consultas parametrizadas e use o cache do plano de consulta, que é ativado por padrão. Para consultas lentas, execute um plano de explicação para fazer melhorias. Se você usar a pesquisa por similaridade vetorial, decida se incorporações menores produzem resultados de similaridade precisos, pois incorporações menores podem ser criadas, armazenadas e pesquisadas com mais eficiência. Antes de chamar um algoritmo gráfico, use uma `MATCH` cláusula para minimizar o conjunto de nós de entrada. Filtre as etiquetas dos nós e das bordas, se possível.

- Busque a maneira mais eficiente de carregar dados no gráfico. Se você carregar dados no Amazon S3, use a importação em massa se os dados tiverem mais de 50 GB de tamanho. Use o carregamento em lote para dados menores.
- Peça aos desenvolvedores que compartilhem instâncias do notebook Neptune em vez de cada um criar sua própria instância. Crie pastas de caderno separadas para cada desenvolvedor em uma única instância do Jupyter. Encerre a instância quando ela não estiver em uso.

Recursos

Referências

- AWS Well-Architected
- AWS Documentação do Well-Architected Framework
- Aplicação do AWS Well-Architected Framework para o Amazon Neptune
- Melhores práticas do Neptune Analytics

Postagens e vídeos no blog

- Publicações AWS do blog do Neptune (Database Blog)
- Automatize a interrupção e a inicialização dos recursos ambientais do Amazon Neptune usando tags de recursos (Database Blog)AWS
- Lanches Amazon Neptune (vídeos curtos) YouTube

Treinamento

- Conceitos básicos do Amazon Neptune
- Crie com o Amazon Neptune
- Modelagem de dados para Amazon Neptune
- Workshop de análise do Amazon Neptune

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	—	20 de dezembro de 2024

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:

- Refatorar/rearquitetar: mova uma aplicação e modifique sua arquitetura aproveitando ao máximo os recursos nativos de nuvem para melhorar a agilidade, a performance 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 edição compatível com o Amazon Aurora PostgreSQL.
- 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: Migre seu banco de dados Oracle local para o Amazon Relational Database Service (Amazon RDS) for Oracle no. 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 a 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: Migre seu banco de dados Oracle local para o Oracle em uma EC2 instância no. 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 local para um serviço em 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

ABAC

Consulte controle de acesso baseado em atributos.

serviços abstratos

Veja os 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 no qual 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.

função agregada

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 a inteligência artificial.

AIOps

Veja as 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 aplicativos

Uma abordagem de segurança que permite o uso somente de aplicativos aprovados 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 AIOps é usado na estratégia de AWS migração, 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 ruim

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

BCP

Veja o 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 deployment (implantação azul/verde)

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

bot

Um aplicativo de software que executa tarefas automatizadas pela Internet e simula a atividade ou interação humana. Alguns bots são úteis ou benéficos, como rastreadores da Web que indexam informações na Internet. Alguns outros bots, conhecidos como bots ruins, têm como objetivo perturbar ou causar danos a indivíduos ou organizações.

botnet

Redes de bots infectadas por malware e sob o controle de uma única parte, conhecidas como pastor de bots ou operador de bots. As redes de bots 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 em vidro quebrado

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 vidro na orientação do Well-Architected AWS .

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

Consulte Estrutura de adoção da AWS nuvem.

implantação canária

O lançamento lento e incremental de uma versão para usuários finais. Quando estiver confiante, você implanta a nova versão e substituirá a versão atual em sua totalidade.

CCoE

Veja o Centro de Excelência em Nuvem.

CDC

Veja a captura de dados de alterações.

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 a integração e a 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.

criptografia no lado do cliente

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

Centro de excelência em 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 publicações CCo E no Blog de Estratégia Nuvem AWS Empresarial.

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem geralmente está conectada à tecnologia de computação de ponta.

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 quando migram para o Nuvem AWS:

- Projeto: executar alguns projetos relacionados à nuvem para fins de prova de conceito e aprendizado
- Fundação — Fazer investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma landing zone, definir um CCo E, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- Reinvenção: otimizar produtos e serviços e inovar na nuvem

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

CMDB

Consulte o 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 GitHub ouBitbucket 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 pipeline de CI/CD 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 da IA que usa aprendizado de máquina para analisar e extrair informações de formatos visuais, como imagens e vídeos digitais. Por exemplo, a Amazon SageMaker AI fornece algoritmos de processamento de imagem para CV.

desvio de configuração

Para uma carga de trabalho, uma alteração de configuração em relação ao estado esperado. Isso pode fazer com que a carga de trabalho 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

Um conjunto 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 is commonly described as a pipeline. CI/CDpode 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 no AWS Well-Architected Framework. 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.

malha de dados

Uma estrutura arquitetônica 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 que oferece suporte à inteligência comercial, como análises. 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

Consulte a 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.

defense-in-depth

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 defense-in-depth abordagem 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 o 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 são comumente 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 desastre. Para obter mais informações, consulte Recuperação de desastres de cargas de trabalho em AWS: Recuperação na nuvem no AWS Well-Architected Framework.

DML

Veja a 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, Design orientado por domínio: lidando com a complexidade no coração do software (Boston: Addison-Wesley Professional, 2003). Para obter informações sobre como usar o design orientado por domínio com o padrão strangler fig, consulte Modernizar incrementalmente os serviços web herdados do Microsoft ASP.NET (ASMX) usando contêineres e o Amazon API Gateway.

DR

Veja a recuperação de desastres.

detecção de deriva

Rastreando desvios de uma configuração básica. 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 o mapeamento do fluxo de valor do desenvolvimento.

E

EDA

Veja a 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 à computação em nuvem, a computação de ponta 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 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. Os sistemas big-endian armazenam o byte mais significativo antes. Os sistemas little-endian armazenam o byte menos significativo antes.

endpoint

Veja o endpoint do 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 corporativos (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 pipeline de CI/CD, 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 o planejamento de recursos corporativos.

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. Ele armazena dados quantitativos sobre as operações comerciais. Normalmente, uma tabela de fatos contém dois tipos de colunas: aquelas que contêm medidas e aquelas que contêm uma chave externa para uma tabela de dimensões.

falham rapidamente

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

limite de isolamento de 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 Limites de isolamento de AWS falhas.

ramificação de recursos

Veja a filial.

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.

solicitação rápida

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. Solicitações rápidas podem ser eficazes para tarefas que exigem formatação, raciocínio ou conhecimento de domínio específicos. Veja também a solicitação zero-shot.

FGAC

Veja o 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 por meio da captura de dados alterados para migrar 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 o modelo da fundação.

modelo de fundação (FM)

Uma grande rede neural de aprendizado profundo que vem treinando em grandes conjuntos de dados generalizados e não rotulados. 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 básicos.

G

IA generativa

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

bloqueio geográfico

Veja as 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 baseado em troncos é a abordagem moderna e preferida.

imagem dourada

Um instantâneo de um sistema ou software usado como modelo para implantar novas instâncias desse sistema ou software. Por exemplo, na manufatura, uma imagem dourada 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 governar recursos, políticas e conformidade em todas as unidades organizacionais ()OUs. 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, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector e verificações personalizadas AWS Lambda .

H

HA

Veja a 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 retenção

Uma parte dos dados históricos rotulados que são retidos de um conjunto de dados usado para treinar um modelo de aprendizado de máquina. Você pode usar dados de retenção para avaliar o desempenho do modelo comparando as previsões do modelo com os dados de retenção.

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 a 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 a Internet das Coisas industrial.

infraestrutura imutável

Um modelo que implanta uma nova infraestrutura para cargas de trabalho de produção em vez de atualizar, corrigir ou modificar a infraestrutura existente. Infraestruturas imutáveis são inerentemente mais consistentes, confiáveis e previsíveis do que infraestruturas mutáveis. Para obter mais informações, consulte as melhores práticas de implantação usando infraestrutura imutável no Well-Architected AWS Framework.

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 AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo 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 em conectividade, dados em tempo real, automação, análise e IA/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 industrial das coisas (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 Criando uma estratégia de transformação digital industrial da Internet das Coisas (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 AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo 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

Consulte 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

Consulte a biblioteca de informações de TI.

ITSM

Veja o 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.

modelo de linguagem grande (LLM)

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

migração de grande porte

Uma migração de 300 servidores ou mais.

LBAC

Veja controle de acesso baseado em etiquetas.

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 um modelo de linguagem grande.

ambientes inferiores

Veja o 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 a filial.

malware

Software projetado para comprometer a segurança ou a privacidade do computador. O malware pode interromper os sistemas do computador, vazar informações confidenciais ou obter acesso não autorizado. Exemplos de malware incluem vírus, worms, ransomware, cavalos de Tróia, 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 abstratos.

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

Consulte Migration Acceleration Program.

mecanismo

Um processo completo no qual 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 Construindo mecanismos no AWS Well-Architected Framework.

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 o sistema de execução de manufatura.

Transporte de telemetria de enfileiramento de mensagens (MQTT)

Um protocolo de comunicação leve machine-to-machine (M2M), baseado no padrão de publicação/assinatura, para dispositivos de IoT com recursos limitados.

microsserviço

Um serviço pequeno e independente que se comunica de forma bem definida APIs e normalmente é de propriedade de equipes pequenas e independentes. 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 leveza. APIs 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

Equipes multifuncionais que simplificam a migração de workloads 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 a Amazon EC2 com o AWS 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 o. 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 carga de trabalho para o. Nuvem AWS Para obter mais informações, consulte a entrada de 7 Rs neste glossário e consulte Mobilize sua organização para acelerar migrações em grande escala.

ML

Veja o aprendizado de máquina.

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 Estratégia para modernizar aplicativos no 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 Avaliação da prontidão para modernização de aplicativos no. 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.

MAPA

Consulte Avaliação do portfólio de migração.

MQTT

Consulte 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 cargas de trabalho de produção. Para melhorar a consistência, confiabilidade e previsibilidade, o AWS Well-Architected Framework recomenda o uso de infraestrutura imutável como uma prática recomendada.

O

OAC

Veja o controle de acesso de origem.

CARVALHO

Veja a identidade de acesso de origem.

OCM

Veja o gerenciamento de mudanças 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 a integração de operações.

OLA

Veja o contrato em 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

Consulte Comunicação de processo aberto — Arquitetura unificada.

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

Um protocolo de comunicação machine-to-machine (M2M) para automação industrial. O 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 melhores práticas 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) no Well-Architected AWS Framework.

tecnologia operacional (OT)

Sistemas de hardware e software que funcionam com o ambiente físico para controlar operações, equipamentos e infraestrutura industriais. Na manufatura, a integração dos sistemas OT e de tecnologia da informação (TI) 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 exigida 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 S3 Regiões da AWS, criptografia do lado do servidor com AWS KMS (SSE-KMS) e solicitações dinâmicas ao bucket S3. `PUT` `DELETE`

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 a análise de prontidão operacional.

OT

Veja a 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 AWS Segurança recomenda configurar sua conta de rede com entrada, saída e inspeção VPCs para proteger a interface bidirecional entre seu aplicativo 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 as 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

Consulte controlador lógico programável.

AMEIXA

Veja o gerenciamento do ciclo de vida do produto.

política

Um objeto que pode definir permissões (consulte a política baseada em identidade), especificar as condições de acesso (consulte a política baseada em recursos) ou definir as permissões máximas para todas as contas em uma organização em AWS Organizations (consulte a 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. Para obter mais informações, consulte Habilitar a persistência de dados em microsserviços.

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 `WHERE` cláusula.

pressão de predicados

Uma técnica de otimização de consulta 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 o desempenho 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 de engenharia de sistema que leva em consideração a privacidade em todo o processo de desenvolvimento.

zonas hospedadas privadas

Um contêiner que contém informações sobre como você deseja que o Amazon Route 53 responda às consultas de DNS para um domínio e seus subdomínios em um 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 projetado para impedir a implantação de recursos não compatíveis. Esses controles examinam 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 o design, desenvolvimento e lançamento, passando pelo crescimento e maturidade, até o declínio e a remoção.

ambiente de produção

Veja o ambiente.

controlador lógico programável (PLC)

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

encadeamento imediato

Usando a 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 no qual outros microsserviços possam se inscrever. 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, responsável, consultado, informado (RACI).

RAG

Consulte Geração Aumentada de 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, responsável, consultado, informado (RACI).

RCAC

Veja o 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.

rearquiteta

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 mais informações, consulte Especificar o que Regiões da AWS sua conta pode usar.

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 um aplicativo de resistir ou se recuperar de interrupções. Alta disponibilidade e recuperação de desastres são considerações comuns ao planejar a resiliência no. Nuvem AWS Para obter mais informações, consulte Nuvem AWS Resiliência.

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.

aposentar-se

Veja 7 Rs.

Geração Aumentada de Recuperação (RAG)

Uma tecnologia de IA generativa na qual 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.

alternância

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

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 o objetivo do ponto de recuperação.

RTO

Veja o objetivo do 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 AWS Management Console 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 a política de controle de serviços.

secret

Em AWS Secrets Manager, informações confidenciais ou restritas, como uma senha ou credenciais de usuário, que você armazena de forma criptografada. Ele 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 O que há em um segredo do Secrets Manager? na documentação do Secrets Manager.

segurança por design

Uma abordagem de engenharia de sistemas que leva em conta 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 principais de controles de segurança: preventivos, detectivos, responsivos e proativos.

fortalecimento 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 correção de uma instância EC2 da Amazon ou a rotação 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 em AWS Organizations. SCPs defina barreiras ou estabeleça limites nas ações que um administrador pode delegar a usuários ou funções. Você pode usar SCPs como listas de permissão ou listas 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 medida de um aspecto de desempenho de um serviço, como taxa de erro, disponibilidade ou taxa de transferência.

objetivo de nível de serviço (SLO)

Uma métrica alvo que representa a integridade de um serviço, conforme medida 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.

SIEM

Veja informações de segurança e sistema de gerenciamento de eventos.

ponto único de falha (SPOF)

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

SLA

Veja o contrato de nível de serviço.

ESGUIO

Veja o indicador de nível de serviço.

SLO

Veja o objetivo do nível de serviço.

split-and-seed modelo

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 Abordagem em fases para modernizar aplicativos no. Nuvem AWS

CUSPE

Veja um único ponto de falha.

esquema de estrelas

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 uso 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 Modernizar incrementalmente os serviços Web herdados do Microsoft ASP.NET (ASMX) 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 de supervisão 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 o desempenho. Você pode usar o Amazon CloudWatch Synthetics para criar esses testes.

prompt do sistema

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

T

tags

Pares de valores-chave que atuam como metadados para organizar seus recursos. AWS As tags podem ajudar você a gerenciar, identificar, organizar, pesquisar e filtrar recursos. 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 o 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.

gateway de trânsito

Um hub de trânsito de rede que você pode usar para interconectar sua rede com VPCs a rede local. 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. Para obter mais informações, consulte o guia Como quantificar a incerteza em sistemas de aprendizado profundo.

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 o 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 da VPC

Uma conexão entre duas VPCs que permite rotear o 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 back-end.

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.

MINHOCA

Veja escrever uma vez, ler muitas.

WQF

Consulte Estrutura de qualificação AWS da carga de trabalho.

escreva uma vez, leia muitas (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, geralmente malware, que tira proveito de uma vulnerabilidade de dia zero.

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.

aviso zero-shot

Fornecer a um LLM instruções para realizar uma tarefa, mas sem exemplos (fotos) que possam ajudar a orientá-la. O LLM deve usar seu conhecimento pré-treinado para lidar com a tarefa. A

eficácia da solicitação zero depende da complexidade da tarefa e da qualidade da solicitação. Veja também a solicitação de algumas fotos.

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.