Orientação de multilocação para ISVs execução de bancos de dados Amazon Neptune

AWS Orientação prescritiva

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Orientação prescritiva: Orientação de multilocação para ISVs execução de bancos de dados Amazon Neptune

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

As marcas comerciais e imagens comerciais da Amazon não podem ser usadas no contexto de nenhum produto ou serviço que não seja da Amazon, nem de qualquer maneira que possa gerar confusão entre os clientes ou que deprecie ou desprestigie a Amazon. Todas as outras marcas comerciais que não pertencem à Amazon pertencem a seus respectivos proprietários, que podem ou não ser afiliados, patrocinados pela Amazon ou ter conexão com ela.

Table of Contents

Introdução 1
Modelos de particionamento de dados 3
Modelo de silo 5
  Cluster por inquilino 5
  Orientação de implementação para o modelo de silo 7
Modelo de piscina 9
  Modelo de piscina para LPGs 10
    Estratégia imobiliária 10
    Estratégia de prefixo-rótulo 13
    Estratégia de vários rótulos 15
    Implicações de desempenho para os modelos de GLP 18
  Modelo de piscina para RDF 19
    Opções de consulta SPARQL usando o protocolo HTTP do Graph Store 19
    Isolamento de inquilinos para RDF 20
  Prepare-se para o crescimento 21
    Limitações para cenários de multilocação 22
Modelo híbrido 23
Práticas recomendadas 24
  Atualize seu cluster Neptune com as versões mais recentes 24
  Use deltas em vez de excluir e substituir para ingestão de dados 24
  Modele como os custos do Neptune evoluirão com seus inquilinos 25
  Dimensione seus clusters de acordo com a demanda do cliente 25
Próximas etapas 27
Recursos 28
Colaboradores 29
Histórico do documentos 30
Glossário 31
  # 31
  A 32
  B 35
  C 37
  D 40
  E 45
  F 47

G 49
H 50
eu 51
L 54
M 55
O 59
P 62
Q 65
R 65
S 69
T 73
U 74
V 75
W 75
Z 76
..... lxxviii

Orientação de multilocação para ISVs execução de bancos de dados Amazon Neptune

Amazon Web Services (colaboradores)

Agosto de 2024 (histórico do documento)

A multilocação é uma arquitetura de sistemas de computador em que uma única instância de um aplicativo atende a vários clientes. Cada cliente é chamado de inquilino. Em uma arquitetura multilocatária, essas instâncias do aplicativo operam em um ambiente compartilhado em que cada locatário está fisicamente localizado na mesma infraestrutura, mas está logicamente separado.

Como fornecedor independente de software (ISV), você pode usar o Amazon Neptune para potencializar aplicativos que exigem navegação em dados altamente conectados. Você pode estar gerenciando um aplicativo de software como serviço (SaaS) baseado em nuvem em sua conta e fornecendo assinaturas aos inquilinos. Os inquilinos podem então acessar o serviço pela Internet ou de forma privada. AWS PrivateLink A economia desse modelo funciona para ambas as partes, porque o locatário obtém acesso a um software que é mais barato do que seria para ele comprar, criar e manter. AssimISV, você pode cobrar mais pela assinatura do que custa criar e manter o software. A questão é como você expande sua empresa para vários inquilinos.

A multilocação ISVs oferece importantes benefícios econômicos e operacionais. A arquitetura multilocatária oferece à sua organização um melhor retorno sobre o investimento ()ROI. A multilocação também simplifica os requisitos operacionais para que sua organização possa agir mais rapidamente e reduzir o custo de entrega do software aos seus locatários.

Este documento fornece orientação sobre como executar com eficácia um ISV aplicativo multilocatário usando o Amazon Neptune. Essa orientação é baseada nas melhores práticas adquiridas ao longo de anos ISVs apoiando a entrega bem-sucedida de soluções SaaS a seus clientes. Avaliar essa orientação no contexto das metas e dos princípios arquitetônicos de sua organização ajudará você a encontrar maneiras de otimizar sua solução.

> **Note**
>
> Este documento não fornece uma lista completa das melhores práticas. Ele complementa o documento Applying the AWS Well-Architected Framework for Amazon Neptune fornecendo orientações específicas adicionais para cargas de trabalho de multilocação.

ISV Recomendamos revisar as considerações em ambos os documentos ao projetar sua solução.

Modelos de particionamento de dados SaaS

Um dos desafios dos desenvolvedores de SaaS é projetar padrões de arquitetura para representar e organizar dados em um ambiente multi-locatário. Esses mecanismos e padrões de armazenamento multilocatário são normalmente chamados de particionamento de dados.

Em um ambiente SaaS multilocatário, é importante distinguir entre particionamento de dados e isolamento de inquilinos. Esses conceitos, embora relacionados, não são sinônimos. O particionamento de dados se refere ao método de armazenamento de dados para cada inquilino. No entanto, o particionamento por si só não garante o isolamento do inquilino. Medidas adicionais são necessárias para garantir que os dados de um inquilino permaneçam inacessíveis para outro.

Os três modelos comuns de particionamento de dados em sistemas SaaS multilocatários são silo, pool e híbrido. A escolha de qualquer modelo depende de fatores como os seguintes:

- Conformidade
- Vizinhos barulhentos
- Estratégia de divisão em níveis
- Requisitos operacionais
- Necessidades de isolamento de inquilinos

Além disso, cada tipo de banco de dados disponível AWS normalmente oferece uma coleção exclusiva de modelos de particionamento de dados e isolamento de inquilinos. Ao analisar como os gráficos de locatários podem ser organizados para atender às várias necessidades de sua solução, considere os modelos que o Amazon Neptune fornece.

Muitos ISVs começam seu design em Netuno com uma das seguintes afirmações:

- A ISV solução exige a separação física dos clientes em clusters separados.
- A ISV solução requer construções como bancos de dados nomeados ou esquemas encontrados em sistemas tradicionais de gerenciamento de banco de dados relacional.

Depois de considerar, ISVs perceba que essas afirmações não são verdadeiras porque, em quase todas as cargas de trabalho, cada um de seus clientes tem um gráfico desconectado em seu banco de dados. A implementação da modelagem de dados e das diretrizes de acesso discutidas neste

documento evita que esses limites de dados sejam ultrapassados e mantém a privacidade dos dados do cliente.

Este guia descreve o modelo de silo e o modelo de piscina, mas a maioria ISVs escolhe o modelo de piscina por custos e eficiência operacional. O guia discute brevemente um modelo híbrido que combina aspectos dos modelos de silo e piscina. Alguns ISVs usam um modelo híbrido para seus maiores clientes para acomodar requisitos regulatórios ou de conformidade do tamanho de um gráfico.

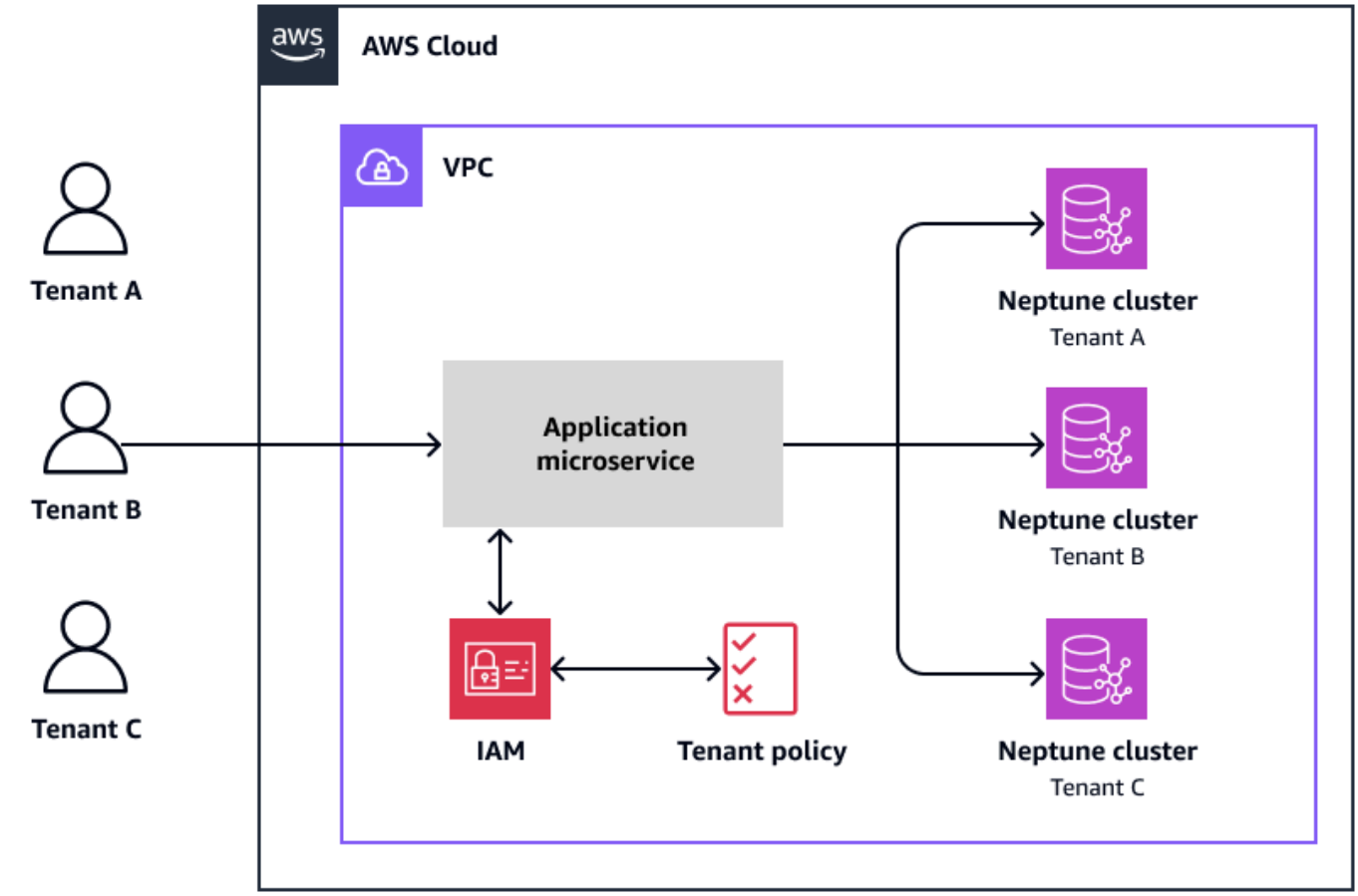
multilocação

Alguns ambientes SaaS multilocatários podem exigir que os dados dos locatários sejam implantados em recursos totalmente separados devido aos requisitos regulatórios e de conformidade. Em alguns casos, grandes clientes precisam de clusters dedicados para reduzir o impacto de vizinhos ruidosos. Nessas situações, você pode aplicar o modelo de silo.

No modelo de silo, o armazenamento dos dados do inquilino é totalmente isolado de qualquer outro dado do inquilino. Todas as construções usadas para representar os dados do inquilino são consideradas fisicamente exclusivas desse cliente, o que significa que cada inquilino geralmente terá armazenamento, monitoramento e gerenciamento distintos. Cada inquilino também terá uma chave AWS Key Management Service (AWS KMS) separada para criptografia. No Amazon Neptune, um silo é um cluster por inquilino.

Cluster por inquilino

Você pode implementar um modelo de silo com o Neptune tendo um inquilino por cluster. O diagrama a seguir mostra três locatários acessando um microsserviço de aplicativo em uma nuvem privada virtual (VPC), com um cluster separado para cada inquilino.



Cada cluster tem seu endpoint individual para ajudar a garantir pontos de acesso distintos para interação e gerenciamento eficientes de dados. Ao colocar cada inquilino em seu próprio cluster, você cria um limite bem definido entre os locatários, garantindo aos clientes que seus dados sejam isolados com sucesso dos dados de outros locatários. Esse isolamento também é atraente para soluções SaaS que têm restrições regulatórias e de segurança rígidas. Além disso, quando cada inquilino tem seu próprio cluster, você não precisa se preocupar com vizinhos barulhentos, onde um inquilino impõe uma carga que pode afetar adversamente a experiência de outros inquilinos.

Embora o modelo de cluster-per-tenant silo tenha vantagens, ele também apresenta desafios de gerenciamento e agilidade. A natureza distribuída desse modelo torna mais difícil agregar e avaliar a atividade do inquilino e a saúde operacional de todos os inquilinos. A implantação também se torna mais desafiadora porque a configuração de um novo inquilino agora exige o provisionamento de um cluster separado. A atualização se torna mais desafiadora em ambientes com uma camada de cliente compartilhada quando as atualizações e versões do cliente estão fortemente acopladas à atualização do banco de dados.

O Neptune oferece suporte a clusters provisionados e sem servidor. Avalie se a carga de trabalho do seu aplicativo é melhor gerenciada por instâncias provisionadas ou sem servidor. Em geral, se sua carga de trabalho tiver um nível constante de demanda, as instâncias provisionadas serão mais econômicas. O Serverless é otimizado para workloads exigentes e altamente variáveis, com uso intenso do banco de dados por curtos intervalos de tempo, seguidos por longos períodos de atividade leve ou nenhuma atividade.

Ao usar um cluster provisionado pelo Neptune por locatário, você deve selecionar um tamanho de instância que se aproxime da carga máxima da demanda do seu locatário. Essa dependência de um servidor também tem um impacto em cascata na eficiência de escalabilidade e no custo do seu ambiente SaaS. Embora a meta do SaaS seja dimensionar dinamicamente com base na carga real do inquilino, um cluster provisionado pelo Neptune exige que você provisione em excesso para compensar períodos mais pesados de uso e picos nas cargas. O provisionamento excessivo aumenta o custo por inquilino. Além disso, à medida que o uso do inquilino muda com o tempo, a ampliação ou redução do cluster deve ser aplicada separadamente para cada inquilino.

A equipe do Neptune geralmente desaconselha um modelo de silo devido ao maior custo incorrido por recursos ociosos e às complexidades operacionais adicionais. No entanto, se cargas de trabalho altamente regulamentadas ou sensíveis exigirem esse isolamento adicional, os clientes podem estar dispostos a pagar o custo adicional.

Orientação de implementação para o modelo de silo

Para implementar um modelo de cluster-per-tenant isolamento de silos, crie AWS Identity and Access Management (IAM) políticas de acesso a dados. Essas políticas controlam o acesso aos clusters Neptune dos inquilinos, garantindo que os inquilinos possam acessar somente o cluster Neptune contendo seus próprios dados. Vincule a IAM política de cada inquilino a uma IAM função. O microsserviço do aplicativo então usa a IAM função para gerar credenciais temporárias refinadas usando o `AssumeRole` método de (). AWS Security Token Service AWS STS Essas credenciais, que têm acesso somente ao cluster Neptune desse inquilino, são usadas para se conectar ao cluster Neptune do inquilino.

O seguinte trecho de código trecho de código mostra um exemplo de política baseada em dadosIAM:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
```

```
      "Effect": "Allow",
      "Action": [
        "neptune-db:ReadDataViaQuery",
        "neptune-db:WriteDataViaQuery"
      ],
      "Resource": "arn:aws:neptune-db:<region>:<account-id>:tenant-1-cluster/*",
      "Condition": {
        "ArnEquals": {
          "aws:PrincipalArn": "arn:aws:iam::<account-id>:role/tenant-role-1"
        }
      }
    }
  ]
}
```

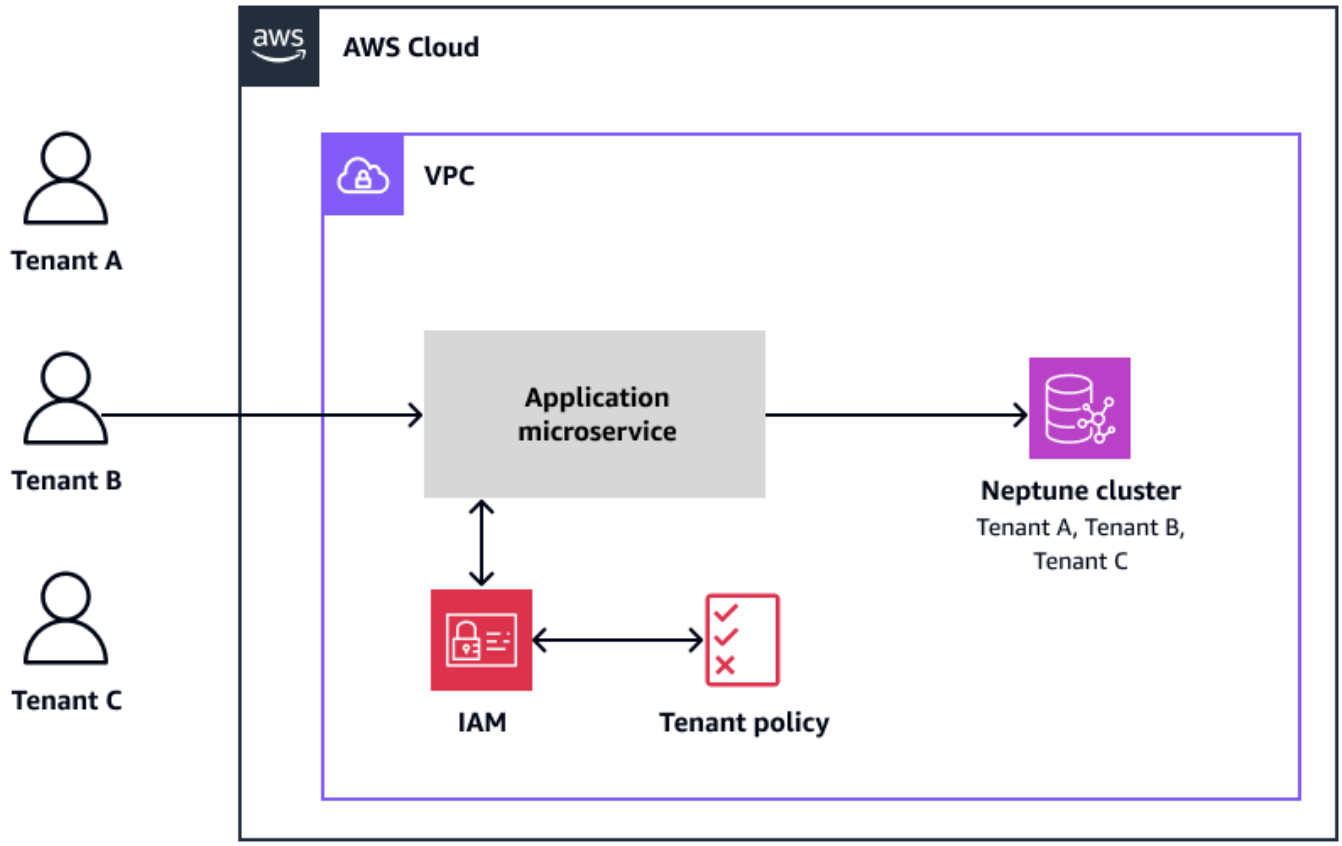
O código fornece um exemplo de inquilino,`tenant-1`, com acesso de consulta de leitura e gravação ao respectivo cluster Neptune. O `Condition` elemento garante que somente a entidade chamadora (a principal), que assumiu a `tentant-1` IAM função (`tenant-role-1`), tenha permissão para acessar o cluster `tenant-1` Neptune.

Multilocação em modelo de piscina

Às vezes, não é necessário ou viável implementar o modelo de silo devido ao custo ou à sobrecarga operacional:

- Talvez você não tenha os recursos para manter um cluster individual por inquilino.
- Talvez não seja necessário separar fisicamente os dados de cada inquilino, e uma separação lógica é suficiente para atender às suas necessidades e requisitos de conformidade.

O diagrama a seguir mostra o modelo de pool, com os dados do inquilino colocados em um único cluster do Amazon Neptune e todos os inquilinos compartilham um banco de dados comum.



Esse modelo de isolamento de pool reduz a sobrecarga de gerenciamento e pode melhorar a eficiência operacional porque há menos clusters para gerenciar. Além disso, os recursos computacionais podem ser compartilhados entre vários clientes, em vez de permanecerem ociosos durante os períodos de inatividade do cliente.

Quando você usa o modelo de pool, há duas maneiras de modelar dados. Sua abordagem depende se você está criando um gráfico de propriedades rotuladas (LPG) ou um gráfico com o Resource Description Framework (RDF).

Modelo de piscina para gráficos de propriedades rotulados

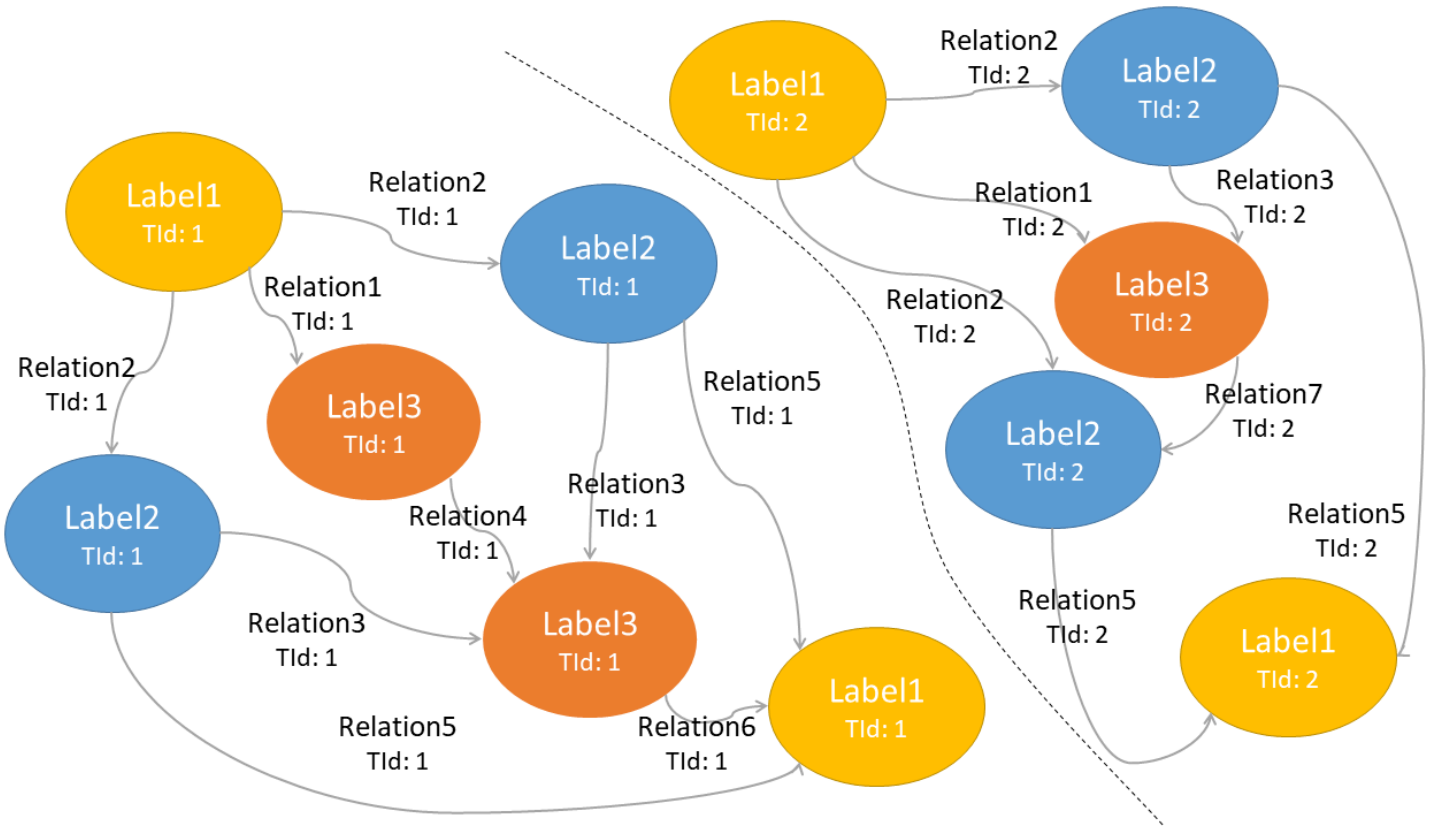
Há três abordagens diferentes para o modelo de pool LPGs no Amazon Neptune:

- Estratégia de propriedade – Escolha a estratégia de propriedade quando precisar priorizar o uso de construções de biblioteca estabelecidas, como a linguagem Apache TinkerPop Gremlin, em detrimento do desempenho. PartitionStrategy
- Estratégia de prefixo-rótulo – Recomendamos a estratégia de prefixo-rótulo para a maioria dos cenários com base no desempenho e na limitação dos efeitos de vizinhos ruidosos.
- Estratégia de rótulos múltiplos – A estratégia de rótulos múltiplos melhorou o desempenho da estratégia de prefixo-rótulo. Ele também suporta a execução de consultas que abrangem todos os inquilinos em um cluster (por exemplo, consultas ISV para geração de relatórios ou monitoramento de todos os inquilinos).

Estratégia imobiliária

Com LPGs, os usuários podem adicionar propriedades de pares de valores-chave aos nós, vértices e bordas. Para obter a separação lógica, a maioria dos clientes modela intuitivamente isso como uma propriedade exclusiva em cada nó e borda com uma chave de propriedade comum do inquilino. A chave da propriedade do inquilino representa todos os inquilinos que possuem o nó. O identificador do inquilino é um valor exclusivo que identifica um inquilino individual.

O diagrama a seguir mostra esse modelo. Os dois subgráficos desconectados têm vários nós e bordas rotulados, com a chave da propriedade do locatário representada por. `TId` Cada nó e aresta de um subgráfico tem um `TId` valor de`1`. No outro subgráfico, cada nó e aresta tem um `TId` valor de`2`.



Nos gráficos de propriedades rotulados, há duas maneiras de gerenciar isso. A linguagem de consulta Gremlin oferece a biblioteca PartitionStrategytransversal para ajudar a gerenciar o particionamento dos dados. O código no exemplo a seguir espera que cada nó e borda tenham uma propriedade chamadaTId:

```
strategy1 = new PartitionStrategy(partitionKey: "TId", writePartition: "1",
 readPartitions: ["1"])
strategy2 = new PartitionStrategy(partitionKey: "TId", writePartition: "2",
 readPartitions: ["2"])
```

Quando novos nós ou bordas são gravados, a propriedade `"TId"` é adicionada com um valor de `"1"` ou`"2"`, dependendo se `strategy1` ou `strategy2` foi selecionada. Para o cliente com `"TId"` of`"1"`, você usa`strategy1`. O exemplo a seguir mostra a gravação de dados desse cliente:

```
g.withStrategies(strategy1).addV("Label1").property("Value", "123456").property(id,
 "Item_1")
```

Para consultas de leitura, um filtro para `"TId == '1'"` ou `"TId == '2'"` é adicionado a cada nó ou passagem de borda usando `strategy1` ou`strategy2`, respectivamente. Essas estratégias de

partição simplificam seu código, mas não são necessárias. A vantagem de usar a estratégia é que ela pode ser injetada em um nível de autorização e passada para o código de nível inferior que forma a consulta. Isso separa o código que determina o identificador do cliente (`TId`) da lógica da consulta.

O código de exemplo a seguir mostra uma consulta Gremlin para ler dados:

```
g.withStrategies(strategy1).V().hasLabel("Label1")
```

O código anterior é equivalente ao exemplo a seguir:

```
g.V().hasLabel("Label1").has("TId", "1")
```

Da mesma forma, ao gravar dados usando o Gremlin, você pode usar a seguinte consulta:

```
g.withStrategies(strategy1).addV("Label1").property("Value").property(id, "Item_1")
```

O código anterior é equivalente ao exemplo a seguir, que não usa a estratégia de partição e, portanto, exige que a `"TId"` propriedade seja escrita explicitamente:

```
g.addV("Label1").property("TId", "1").property("Value").property(id, "Item_1")
```

No OpenCypher, essas bibliotecas não existem. Você é responsável por escrever e modificar suas consultas para adicionar o identificador do inquilino como uma propriedade nos nós e bordas. Por exemplo:

```
CREATE (n:Item {`~id`: 'Item_1', Value: '123456', TId: '1'})
CREATE (n:Item {`~id`: 'Item_2', Value: '123456', TId: '2'})
```

Observe a semelhança entre o código Gremlin sem a estratégia de partição. Em seguida, você pode ler o nó escrito na primeira `CREATE` declaração usando o seguinte código:

```
MATCH (n:Item {TId: '1'})
RETURN n
--or
MATCH (n:Item)
WHERE n.TId == '1'
RETURN n
```

Você pode escolher a estratégia de propriedade quando quiser usar construções nativas do TinkerPop Gremlin, como. PartitionStrategy No entanto, esse modelo tem desvantagens de desempenho no Amazon Neptune em comparação com a estratégia de prefixo-rótulo. Para uma discussão sobre essas desvantagens de desempenho, consulte a seção Implicações de desempenho para os modelos de GLP.

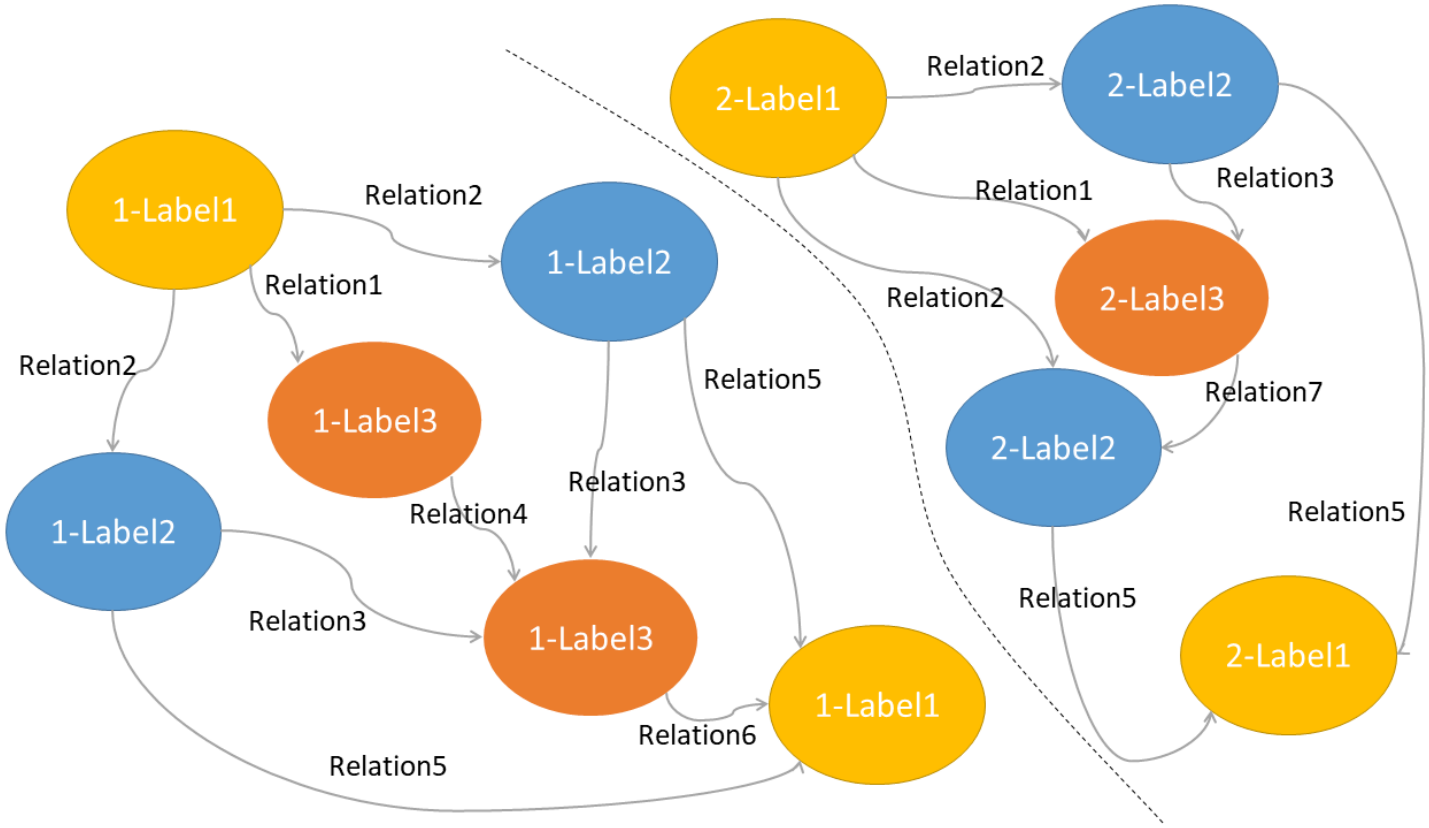
Se as condições a seguir se aplicarem, considere modelar a estratégia da propriedade somente nos nós, não nas bordas:

- Seu gráfico tem muito mais bordas do que rótulos.
- Cada inquilino é um gráfico desconectado.
- Você acessa o gráfico somente usando nós como ponto de partida, não rótulos.

Estratégia de prefixo-rótulo

Se o desempenho for uma das principais preocupações, é altamente recomendável considerar a estratégia de prefixo-rótulo em vez da estratégia da propriedade.

Na estratégia prefix-label, você rotula cada nó com uma combinação de identificador de inquilino e rótulo de nó. Por exemplo, se o inquilino tiver um identificador de `"1"` e o rótulo do nó for`"Label1"`, você especifica o rótulo do nó como`"1-Label1"`. O diagrama a seguir mostra dois subgráficos desconectados que usam esse modelo.



Ao gravar dados no Gremlin, você pode adicionar um número de identificação ao rótulo de qualquer nó:

```
g.addV("1-Label1")
g.addV("2-Label6")
```

Ao consultar esse gráfico, você pode verificar a existência desse prefixo em um nó:

```
g.V().hasLabel("1-Label1")
```

No OpenCypher, você pode gravar dados usando uma declaração: `CREATE`

```
CREATE (n:`1-Label1` {`~id`: 'Item_1', Value: 'XYZ123456'})
```

Para consultar os dados que você escreveu no OpenCypher, use o código a seguir:

```
MATCH n= (:`1-Label1`)
RETURN n
```

A estratégia de prefixo-rótulo pressupõe que todos os nós sejam atribuídos a um ou mais inquilinos e que as permissões não sejam atribuídas no escopo periférico. Evite usar essa estratégia em rótulos de borda, pois isso causará um grande número de predicados e afetará negativamente o desempenho de Neptune.

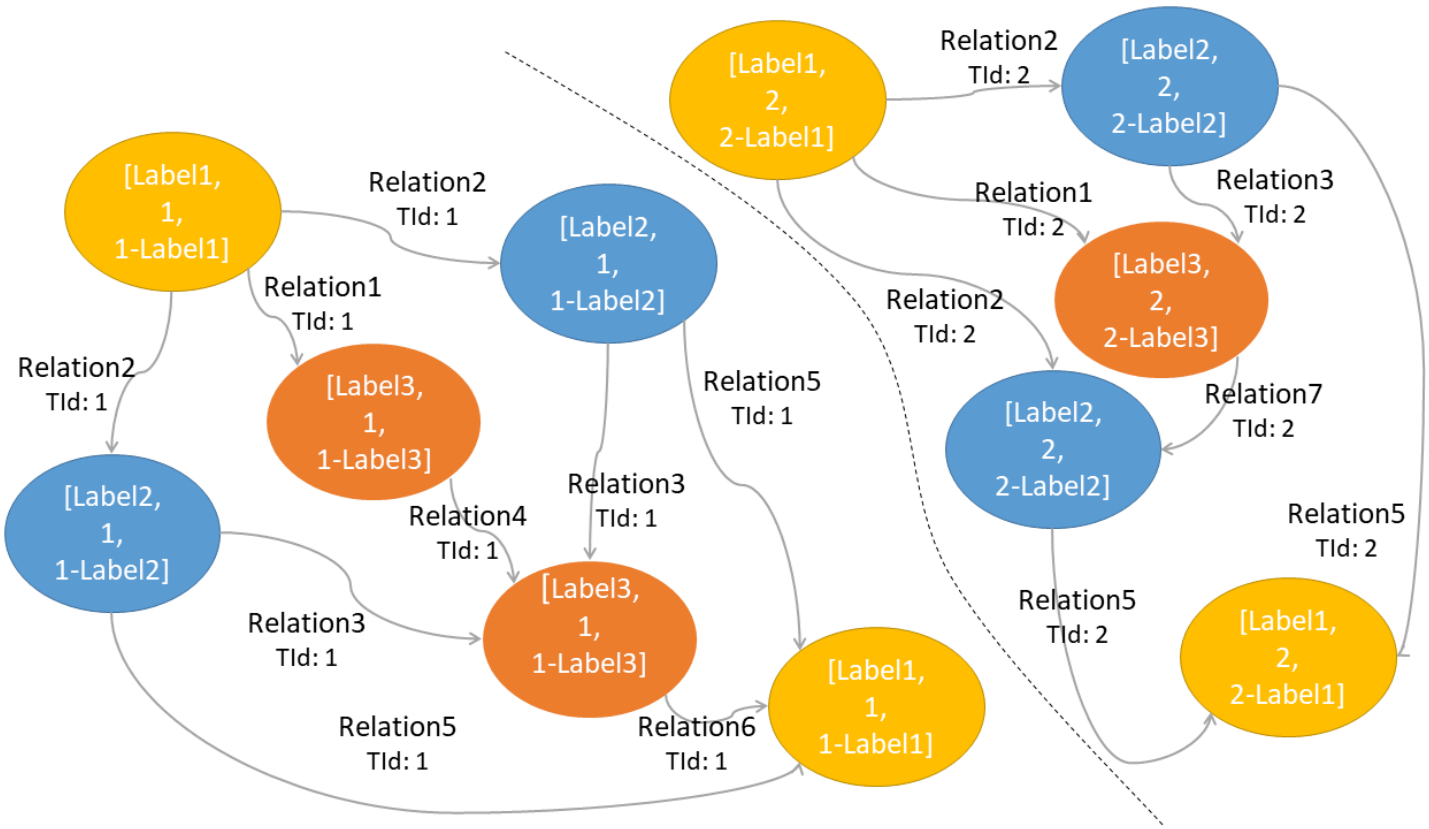
Há duas desvantagens principais na abordagem do rótulo de prefixo. Primeiro, é difícil executar qualquer consulta que abranja os locatários. Um exemplo é uma consulta que conta todos os nós de um determinado rótulo para geração de relatórios ou monitoramento. Se esse for seu caso de uso, considere combinar essa estratégia com a estratégia de vários rótulos. Para obter mais informações sobre a combinação de estratégias, consulte a seção Modelo híbrido.

Em segundo lugar, a estratégia de prefixo-rótulo exige controles que imponham a aplicação adequada do prefixo apropriado a cada consulta para evitar o vazamento de dados. No entanto, essa estratégia é a opção mais eficiente para cargas de trabalho que exigem consultas de baixa latência, e é altamente recomendável. A seção Implicações de desempenho para modelos de GLP fornece exemplos de por que essa é a estratégia mais eficiente.

Estratégia de vários rótulos

A terceira opção é usar uma estratégia de vários rótulos. Para essa abordagem, você adiciona rótulos extras a cada nó no gráfico. Por exemplo, se você precisar filtrar todos os dados de um determinado inquilino, adicione a etiqueta de ID do inquilino. Se você precisar filtrar todos os dados de um determinado rótulo, independentemente do locatário, adicione esse rótulo. O diagrama a seguir mostra a estratégia de vários rótulos aplicada usando três rótulos para cada nó.

Agora você pode acessar o gráfico usando três padrões diferentes:



- Filtre `Label1` para retornar todos os nós de `Label1` todos os inquilinos.
- Filtre `1` para retornar todos os nós do inquilino 1.
- Filtre `1-Label1` para retornar todos os nós somente para o locatário 1 com etiqueta`Label1`.

Pois LPGs, há duas maneiras de implementar isso.

No Gremlin, você pode usar a estratégia de travessia chamada SubgraphStrategypara limitar o escopo de todas as consultas a apenas vértices com um rótulo específico, como: `"Label1"`

```
g.withStrategies(
    new SubgraphStrategy(
        vertices=hasLabel("Label1")
    )
)
```

Ao contrário PartitionStrategy, SubgraphStrategy afeta somente a leitura de dados, não a gravação de dados. Para gravar os dados, atribua manualmente os rótulos em cada consulta:

```
g.addV("Label1").property("Value","XYZ123456")
```

```
.addV("Label2").property("Value","XYZ123456")
```

Ao ler os dados, você pode usar SubgraphStrategy para consultar todos os nós com`"Label1"`:

```
g.withStrategies(
new SubgraphStrategy(vertices=.hasLabel("Label1"))
).
V().has("Value","XYZ123456")
```

Neptune retorna somente o primeiro registro, que `"Label1"` tem um valor de. `"XYZ123456"` É equivalente à consulta a seguir, que não usa SubgraphStrategy:

```
g.V().hasLabel("Label1").hasValue("XYZ123456")
```

Nessa consulta básica, parece que SubgraphStrategy é mais complexo de usar. Lembre-se de que suas bibliotecas podem fornecer uma instância g com a estratégia já definida. Os desenvolvedores não precisam garantir que os filtros adequados sejam aplicados:

```
def getGraphTraversal():
return g.withStrategies(new SubgraphStrategy(vertices=.hasLabel("Label1"))

getGraphTraversal().has("Value","XYZ123456")
```

As bibliotecas do OpenCypher não têm essas construções, então você deve criar vários rótulos para cada nó:

```
CREATE (n:`1`:`Label1`:`1-Label1` {`~id`: 'Item_1', Value: '12345'})
```

Ao usar esses rótulos para filtrar um subgráfico, você pode retornar nós que têm o rótulo do cliente que você está procurando ou que compartilham um relacionamento com outro nó que tenha esse rótulo:

```
MATCH n=(:Label1:`1`)
// or
MATCH n=(:`1-Label1`)
```

A estratégia de vários rótulos oferece a maior flexibilidade para consultar nós por tipo (`Label1`) ou locatário (1), ou para usar a estratégia mais eficiente de prefixo-rótulo quando o desempenho é mais importante (). `1-Label1`

A principal desvantagem dessa estratégia é que cada rótulo é um objeto extra armazenado em seu gráfico. Um objeto é um nó, borda ou uma propriedade em um nó ou borda em LPGs. A velocidade de ingestão é medida e limitada por objetos por segundo, e os custos de armazenamento dependem do número de gigabytes consumidos. Isso significa que objetos extras podem ter um impacto mensurável em grande escala.

Implicações de desempenho para os modelos de GLP

O curso Modelagem de dados do AWS Skill Builder para o Amazon Neptune descreve detalhadamente os aspectos internos do modelo de dados do Neptune e as implicações da modelagem, mas resumiremos as considerações importantes para esses projetos aqui. Considere ter três inquilinos (T1, T2, T3) em um único cluster de Neptune. Esses inquilinos têm os seguintes atributos:

- O locatário 1 (T1) tem 100 milhões de nós no total e 10 milhões são do tipo Item.
- O locatário 2 (T2) tem um total de 10 milhões de nós e 1 milhão são do tipo Item.
- O locatário 3 (T3) tem 100 milhões de nós no total e 1 milhão são do tipo Item.

Execute uma consulta que recuperará os itens do Locatário 3 usando a estratégia de propriedade. O Neptune inspeciona as estatísticas de duas chamadas de índice:

- Onde `tenant property key=T3` tem 100 milhões de resultados
- Onde `label = Item` tem 12 milhões de resultados (10 milhões do T1 + 1 milhão do T2 + 1 milhão do T3)

O otimizador de consultas Neptune determina que a última consulta é melhor aplicada primeiro (12 milhões de resultados) e, em seguida, inspeciona cada item. `tenant property key=T3` Você recupera 12 milhões de itens para encontrar 1 milhão de resultados.

Observe o impacto ruidoso dessa consulta na vizinhança. Se você tivesse 100 milhões de nós de item por inquilino, a primeira consulta teria 300 milhões de resultados em vez de 12 milhões (isso é excessivamente simplificado para fins ilustrativos). O otimizador Neptune pode ter aplicado uma ordem diferente de operações).

Em seguida, considere a estratégia de prefixo-rótulo. Faça uma única chamada de índice where`label=T3-Item`, que retorna 1 milhão de resultados. Isso obtém o mesmo resultado da

estratégia imobiliária, mas recupera 11 milhões de registros a menos. Além disso, você não se preocupa mais com vizinhos barulhentos porque o rótulo não se sobrepõe no índice.

A estratégia de vários rótulos não melhora diretamente o desempenho da consulta em relação à estratégia de propriedade. A filtragem pelo valor da propriedade é comparável à filtragem pelo valor do rótulo quando o espaço de pesquisa também é comparável. Em vez disso, a estratégia de vários rótulos oferece mais flexibilidade. A estratégia de vários rótulos fornece desempenho equivalente à estratégia de prefixo-rótulo para ou para o rótulo. `label=T3` `T3-Item` A estratégia de vários rótulos fornece desempenho equivalente à estratégia de propriedade para. `label=Item` O benefício é oferecer suporte a uma variedade de padrões de acesso.

Modelo de piscina para RDF

O Resource Description Framework (RDF) tem um conceito de gráficos nomeados, que fornece uma maneira lógica de separar dados. No Amazon Neptune, você tem um gráfico nomeado padrão e gráficos nomeados definidos pelo usuário. Você pode criar quantos gráficos nomeados quiser. Coletivamente, eles são chamados de conjunto de dados RDF. Todos os gráficos nomeados, padrão ou definidos pelo usuário, são definidos por um Identificador de Recursos Internacionalizado (IRI) dentro do conjunto de dados RDF. Em Neptune, a menos que um usuário tenha declarado um gráfico nomeado ao gravar dados, todos os triplos são considerados parte do gráfico nomeado padrão.

Há vários casos de uso para gráficos nomeados:

- Particionamento e isolamento de dados
- Proveniência dos dados
- Versionamento
- Inferência

Este guia se concentra no caso de uso do particionamento de dados. Recomendamos criar um gráfico nomeado definido pelo usuário para cada inquilino.

Opções de consulta SPARQL usando o protocolo HTTP do Graph Store

As consultas de exemplo a seguir usam o Protocolo SPARQL e a Linguagem de Consulta RDF (SPARQL) e o Protocolo HTTP do Graph Store para consultar ou criar um gráfico nomeado para um inquilino.

- `HTTP GET`– Para recuperar um gráfico específico de um inquilino:

```
curl --request GET 'https://your-neptune-endpoint:port/sparql/gsp/?graph=http%3A//
www.example.com/named/tenant1'
```

- `HTTP PUT`– Para criar ou substituir um gráfico nomeado específico por uma carga especificada na solicitação:

```
curl --request PUT -H "Content-Type: text/turtle" \  --data-raw "@prefix ex: http://
example.com/ . ex:subject ex:predicate ex:object ." \
'https://your-neptune-endpoint:port/sparql/gsp/?graph=http%3A//www.example.com/named/
tenant1'
```

  Em RDF, um objeto é triplo.

- `HTTP POST`– Para criar um novo gráfico nomeado, se não existir, ou mesclar com um gráfico existente:

```
curl --request POST -H "Content-Type: text/turtle" \
--data-raw "@prefix ex: http://example.com/ . ex:subject ex:predicate ex:object ." \

'https://your-neptune-endpoint:port/sparql/gsp/?graph=http%3A//www.example.com/named/
tenant1'
```

Isolamento de inquilinos para RDF

Para o isolamento lógico dos dados com as proteções necessárias na camada do aplicativo, crie um mapeamento entre o locatário e os gráficos nomeados definidos pelo usuário. Ao projetar a multilocação para um conjunto de dados RDF, esteja ciente dos seguintes aspectos do RDF e do SPARQL:

- Em Neptune, quando você consulta sem especificar um gráfico nomeado, ele recupera todas as triplas que correspondem ao padrão em todos os gráficos nomeados no banco de dados.
- No RDF, não há restrições em relação às conexões entre nós de gráficos com nomes diferentes. Por exemplo, no diagrama anterior, um nó em `:G1` pode ser conectado a um nó em: `G2` por meio de uma borda.

Por exemplo, se um usuário final de um determinado locatário enviar uma consulta à API, a API deverá validar os seguintes requisitos antes de enviar a consulta ao banco de dados Neptune:

- Qualquer consulta com escopo em um único inquilino deve especificar um gráfico nomeado. Caso contrário, você corre o risco de vazar dados entre os inquilinos.
- As consultas de atualização ou exclusão devem sempre especificar um gráfico nomeado.
- Os nós em cada lado de uma borda ou relacionamento devem sempre pertencer ao gráfico nomeado correto.

Para obter informações adicionais sobre as melhores práticas, consulte a documentação do Neptune.

Prepare-se para o crescimento

Quando você usa o modelo de pool com sucesso, você acaba superando o tamanho de um único cluster de Neptune. Os inquilinos crescem, ou o número de inquilinos aumenta, e a taxa de ingestão de dados necessária em todos os seus clientes excede a capacidade do cluster. Quando isso ocorrer, você precisará dividir seus clientes em vários clusters. Projete essa configuração com antecedência, em vez de tentar adaptá-la posteriormente. Mesmo que sua escala inicial seja usar apenas um único cluster, simule os componentes que você precisará para rotear locatários em vários clusters no futuro quando atingir essa escala.

Se sua solução exigir mais recursos com base no tamanho do seu inquilino, prepare-se também para o crescimento dele. Se vários clientes em um único cluster crescerem significativamente, esse cluster pode não atender mais às suas necessidades. Crie uma estratégia para mover inquilinos para outro cluster ou dividir um cluster existente em dois usando o recurso de clonagem do Amazon Neptune DB.

Familiarize-se com o Protocolo Copy-on-Write Neptune, que pode economizar dinheiro ao implementar a clonagem de banco de dados. Se você dividir um cluster devido a gargalos de ingestão, talvez seja mais eficiente não excluir dados dos clusters, desde que suas políticas permitam isso. Os dois clusters compartilharão uma página de dados se ela não for alterada, mas não se a página de dados tiver sido modificada (porque alguns dados nela foram excluídos).

> **Note**
>
> Esta orientação se aplica à versão mais recente do Netuno no momento da redação deste artigo, que é a versão 1.3.1 do Netuno. Essa orientação pode mudar em versões futuras à medida que a camada de armazenamento Neptune evolui.

Limitações para cenários de multilocação

Esteja ciente de que alguns recursos do Neptune não foram criados para cenários de multilocação. Os inquilinos não devem ter acesso direto aos endpoints do Neptune em um modelo de pool porque essas estratégias de multilocação não são aplicadas no nível do banco de dados. Sempre mantenha algum tipo de proxy entre seus clientes e o endpoint Neptune que aplique os designs descritos neste documento. Exemplos desse proxy incluem o seguinte:

- Anexando os filtros de etiquetas na sua camada de cliente
- Ter uma API que mapeia o token de autenticação para um ID de inquilino e injeta esse filtro na consulta

Essa orientação também se aplica ao fornecimento de acesso direto aos clientes a recursos como os notebooks gráficos Neptune, o Neptune Graph-Explorer ou o Neptune Streams.

Multilocação do Model

As soluções SaaS geralmente usam uma combinação de modelos de silo e reserva. Vários fatores influenciam a decisão de quando e como empregar modelos de silo e piscina no mesmo ambiente.

Um desses fatores é a hierarquização, em que uma solução SaaS oferece experiências únicas para cada camada de inquilinos. Por exemplo, se seus níveis forem Gratuito, Padrão e Premium, os dados do inquilino do nível Gratuito poderão ser armazenados em um cluster compartilhado do Neptune usando um modelo de pool. Para seus inquilinos de nível Standard e Premium, você pode usar um modelo de cluster-per-tenant silo.

Além disso, alguns provedores de SaaS têm a capacidade de criar sua solução de pool em um cluster compartilhado do Amazon Neptune como base. Posteriormente, eles podem criar um cluster Neptune separado para locatários que precisam de armazenamento em silos, geralmente devido a exigências regulatórias e de conformidade.

Embora isso possa adicionar um nível de complexidade à sua camada de acesso a dados e ao seu perfil de gerenciamento, também pode oferecer à sua empresa uma maneira de hierarquizar sua oferta para atender aos requisitos do cliente.

Práticas operacionais para ISVs

Muitas das diretrizes desta seção são as melhores práticas para todos os clientes, mas elas têm um significado adicional paraISVs.

Atualize seu cluster Neptune com as versões mais recentes

Nas notas de lançamento do Amazon Neptune, você pode ver que cada versão traz várias correções de erros, melhorias de desempenho e novos recursos. Mantenha seus clusters do Neptune na versão mais recente possível.

Se você encontrar um bug não descoberto anteriormente em sua carga de trabalho e seu cluster estiver na versão mais recente, os engenheiros do Neptune poderão criar um patch privado para seu cluster (se necessário e você quiser). O patch pode durar até a próxima versão, quando essa correção estará disponível ao público em geral. Para ajudar a atualizar seus clusters para a versão mais recente, use a solução Neptune Blue/Green.

Use deltas em vez de excluir e substituir para ingestão de dados

Você pode usar várias técnicas para ingerir ou gravar dados no Neptune. Muitos clientes tentam simplificar a ingestão de dados excluindo e reinserindo o gráfico sempre que uma alteração é recebida no feed. Eles podem adicionar uma `last-modified` propriedade a cada nó e verificar periodicamente os nós que não foram modificados desde uma data especificada e excluí-los. Embora essas técnicas simplifiquem o processo de ingestão de dados, elas têm implicações de longo prazo na saúde e na escalabilidade do seu cluster Neptune.

Primeiro, Neptune usa a codificação de strings no dicionário. A menos que você especifique explicitamente os IDs nós e bordas, o Neptune gera GUID um representado como uma string para o ID e armazena essa string no dicionário. Se você estiver constantemente excluindo e adicionando objetos, o gerado automaticamente IDs causará inchaço no dicionário.

Em segundo lugar, Netuno se expande para ingerir cerca de 120 K objetos por segundo, no máximo. Se você excluir e adicionar objetos continuamente, consumirá grande parte dessa largura de banda em objetos que basicamente não estão mudando. Isso limita o número de inquilinos que você pode hospedar em um cluster, exige instâncias de gravação maiores nos clusters e exige mais operações de E/S. Todos esses fatores aumentam seus custos.

É altamente recomendável que você desenvolva uma forma de calcular o verdadeiro delta do que foi alterado em vez de usar os métodos delete e add. No entanto, algumas fontes de dados não são propícias para isso (por exemplo, API chamadas que retornam o estado atual ou eventos que não rastreiam exatamente o que mudou). Se sua fonte de dados bruta não for propícia à identificação de alterações, use seus processos de extração, transformação e carregamento (ETL) para calcular o delta. Por exemplo, você pode manter instantâneos de cada captura de dados anterior no formato Parquet, usar AWS Glue para calcular as diferenças entre esses instantâneos e enviar somente as diferenças para Netuno.

Modele como os custos do Neptune evoluirão com seus inquilinos

Se você usa um modelo de silo, pool ou híbrido, seus custos de nuvem aumentarão de acordo com o tamanho de seus inquilinos. Os locatários que exigem mais conexões simultâneas precisam de instâncias maiores ou mais réplicas de leitura do que aqueles com menos conexões simultâneas. O mesmo se aplica aos locatários que exigem uma ingestão de dados mais rápida.

Os três componentes do custo do cluster Neptune são tamanho (e número) da instância, tamanho dos dados (GB/mês) e operações de E/S (por milhão). Embora esses custos geralmente sejam específicos da carga de trabalho, eles escalam de acordo com o tamanho e o volume de dados e podem ser medidos usando AWS ferramentas. Acompanhe e entenda as economias de escala em relação aos principais indicadores do tamanho de seus inquilinos, incluindo como seus tamanhos variam ao longo do tempo. Se a imprevisibilidade de suas cobranças de E/S afetar suas margens, considere escolher o armazenamento otimizado de E/S Neptune para obter um custo mais previsível.

Dimensione seus clusters de acordo com a demanda do cliente

Não existe uma fórmula testada ou comprovada para dimensionar corretamente o tamanho da instância do Neptune. A documentação do Neptune fornece orientação, mas há muitas variáveis para recomendar um mapeamento direto. Essas variáveis incluem, mas não estão limitadas ao seguinte:

- Modelo de dados
- formato de dados
- Simultaneidade da consulta
- Complexidade da consulta.

Planeje testes para determinar o tamanho ideal para suas cargas de trabalho e perfis de inquilinos. Em geral, recomendamos o uso de instâncias provisionadas para obter eficiência de custos e previsibilidade. Se suas metas de experiência do cliente priorizam a escalabilidade ideal em relação aos custos, considere usar instâncias Neptune Serverless para garantir uma experiência mais consistente, independentemente das flutuações da carga de trabalho.

Se as cargas de trabalho de leitura do locatário tiverem uma variabilidade significativa em seus altos e baixos, combine as instâncias do Neptune Serverless com o auto-scaling do Neptune. Geralmente, leva de 10 a 15 minutos para que uma nova réplica de leitura fique on-line depois de inicializada. Isso significa que o auto-scaling sozinho pode lidar com mudanças prolongadas no tráfego, mas não é suficiente para alterar rapidamente os picos de atividade. Ao combinar o escalonamento automático do Neptune Serverless e do Neptune, você pode aumentar ou reduzir as instâncias e escalar o número de réplicas de leitura para dentro e para fora.

Se seus locatários tiverem perfis de carga de trabalho ou contratos de nível de serviço (SLAs) significativamente diferentes, considere usar endpoints personalizados e réplicas de leitura dedicadas para direcionar o tráfego para instâncias otimizadas para esse tráfego. A otimização pode incluir um tamanho diferente da instância, padrões de consulta específicos ou pré-aquecimento do cache do buffer.

Próximas etapas

Se você está apenas começando sua jornada de implementação do Amazon Neptune para seu aplicativo de ISV multilocação, pense mais no modelo desejado. Mudar o modelo será mais caro posteriormente em sua jornada.

Se você estiver no início da viagem, verifique se está usando o melhor modelo para suas necessidades e se está seguindo as orientações desse modelo.

Planejar antecipadamente. Quando você está no início de sua jornada, é tentador adiar o trabalho de fragmentar clientes em clusters ou otimizar seus ETL processos para fornecer o delta das mudanças em vez de excluir e adicionar novamente vértices e arestas. À medida que você escala, essas decisões podem afetar negativamente o desempenho e o custo.

Por fim, se você já está no início de sua jornada, essa orientação pode garantir que sua arquitetura é ideal ou fornecer mudanças para aprimorá-la.

Se você tiver dúvidas sobre esta orientação ou precisar de mais ajuda, entre em contato com sua Conta da AWS equipe e solicite uma sessão com um especialista do Neptune.

Recursos

- Documentação do Amazon Neptune
- Modelagem de dados para Amazon Neptune (curso)
- Aplicação do AWS Well-Architected Framework para o Amazon Neptune
- Well-Architected Framework do SaaS Lens
- Orientação para arquiteturas multilocatárias sobre AWS
- Estratégias de isolamento de inquilinos de SaaS: isolando recursos em um ambiente multilocatário
- Documentação do Apache TinkerPop
- SPARQL

Colaboradores

Os colaboradores deste guia incluem:

- Brian O'Keefe, diretor da Guerra Mundial de Netuno, SSA AWS
- Veeresham Gande, gerente técnico sênior de contas, AWS
- Dana Owens, arquiteta de soluções para startups, AWS
- Nima Seifi, arquiteta de soluções para startups, AWS

Histórico do documento

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

Alteração	Descrição	Data
Publicação inicial	—	3 de setembro 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.