Melhores práticas para implantar o SQL Server na Amazon EC2

AWS Orientação prescritiva

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

AWS Orientação prescritiva: Melhores práticas para implantar o SQL Server na Amazon EC2

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
Configurar as configurações de computação e de armazenamento 2
Use um tipo de instância otimizada para Amazon EBS 2
Otimize o layout do disco ou a distribuição de arquivos 3
Defina o tamanho da unidade de alocação NTFS para 64 KB 3
Coloque o tempdb em um armazenamento de instâncias 5
Movendo o tempdb para um armazenamento de instâncias 5
Inicializar o armazenamento de instâncias 8
Usando a extensão do grupo de buffer 10
Evite incompatibilidades entre os núcleos da CPU 10
Teste o desempenho dos discos 11
Habilitar a inicialização instantânea de arquivos 12
Bloquear páginas na memória 14
Desative o descarregamento de TCP e as configurações de RSS 16
Determine seus requisitos de IOPS e throughput 17
Use striping para contornar as limitações de IOPS e de throughput 18
Excluir arquivos do SQL Server do software antivírus 18
Configurar o Servidor SQL 19
Configure o tempdb para reduzir a contenção 19
Defina o MAXDOP para o maior desempenho 20
Alterar o limite de custo do paralelismo 22
Otimize para workloads pontuais 23
Use sinalizadores de rastreamento para melhorar o desempenho 23
Instalar os patches mais recentes 24
Limite a memória máxima do servidor para evitar pressão na memória 24
Use o nível mais alto de compatibilidade do banco de dados 26
Controle o número de VLFs 27
Verifique as configurações de crescimento automático do banco de dados 27
Configurando grupos de disponibilidade do Always On 31
Defina RegisterAllProviders o IP como verdadeiro ao usar grupos de disponibilidade Always On 31
Defina HostRecord TTL como 60 ou menos ao usar grupos de disponibilidade Always On 32
Desativar o failback automático para o grupo de clusters Always On 32
Configuração dos backups 33

Melhorias de otimização do banco de dados 34
Reconstruir índices 34
Atualizar estatísticas 35
Otimizando as implantações do SQL Server na Amazon EC2 36
Próximas etapas 37
Recursos adicionais 38
Histórico do documento 40
Glossário 42
..... 42
A 43
B 46
C 48
D 51
E 55
F 57
G 59
H 60
eu 62
L 64
M 65
O 70
P 72
Q 75
R 76
S 79
T 83
U 84
V 85
W 85
Z 86
..... lxxxviii

Melhores práticas para implantar o Microsoft SQL Server na Amazon EC2

Abhishek Soni e Sagar Patel, Amazon Web Services (AWS)

Dezembro de 2023 (histórico do documento)

O objetivo deste guia é garantir uma experiência consistente após a implantação ou migração do Microsoft SQL Server para o Amazon Elastic Compute Cloud (Amazon EC2) na Amazon Web Services (AWS) Cloud. Ele fornece as práticas para recomendadas configurar seu banco de dados e servidor, para ajudar a otimizar sua infraestrutura, ajustar o desempenho e evitar problemas inesperados após a implantação ou a migração.

Este guia é para arquitetos de banco de dados, líderes de sistemas e bancos de dados e administradores que planejam migrar o Microsoft SQL Server de seu ambiente local para a Amazon EC2 ou que desejam otimizar sua nova implantação do SQL Server na Amazon. EC2

A Amazon EC2 fornece capacidade computacional escalável na AWS nuvem. Usar o SQL Server na Amazon EC2 é semelhante à execução do SQL Server no local. A Amazon EC2 oferece controle total sobre sua infraestrutura e seu ambiente de banco de dados. Você se beneficia da escala, do desempenho e da elasticidade da AWS nuvem, mas é responsável por configurar e ajustar todos os componentes, incluindo EC2 instâncias, volumes de armazenamento, sistemas de arquivos, redes e segurança. Este guia fornece informações para ajudar você a otimizar sua configuração e maximizar o desempenho do SQL Server na AWS. Ele discute detalhadamente as configurações do servidor e do armazenamento e as práticas recomendadas. Também explica como automatizar as configurações, quando aplicável, e discute as alterações de configuração no nível do banco de dados.

> **Note**
>
> AWS também oferece opções para mover seu banco de dados SQL Server local para um serviço gerenciado como o Amazon Relational Database Service (Amazon RDS) for SQL Server. Para uma discussão sobre as opções de migração, consulte Estratégia de migração para bancos de dados relacionais no site de Orientação AWS Prescritiva.

Configurar as configurações de computação e de armazenamento

Antes de migrar ou implantar o SQL Server na Amazon EC2, você pode definir suas configurações de EC2 instância e armazenamento para melhorar o desempenho e reduzir seus custos. As seções a seguir fornecem dicas de otimização e melhores práticas.

Tópicos

- Use um tipo de instância otimizada para Amazon EBS
- Otimize o layout do disco ou a distribuição de arquivos
- Defina o tamanho da unidade de alocação NTFS para 64 KB
- Coloque o tempdb em um armazenamento de instâncias
- Evite incompatibilidades entre os núcleos da CPU
- Teste o desempenho dos discos
- Habilitar a inicialização instantânea de arquivos
- Bloquear páginas na memória
- Desative o descarregamento de TCP e as configurações de RSS
- Determine seus requisitos de IOPS e throughput
- Use striping para contornar as limitações de IOPS e de throughput
- Excluir arquivos do SQL Server do software antivírus

Use um tipo de instância otimizada para Amazon EBS

Se seu banco de dados SQL Server lida com cargas de trabalho intensivas de E/S, o provisionamento de instâncias otimizadas do Amazon Elastic Block Store (Amazon EBS) ajudará a melhorar o desempenho.

Uma instância otimizada para o Amazon EBS usa uma pilha de configuração otimizada e fornece capacidade adicional dedicada para E/S do Amazon EBS. Essa otimização proporciona a melhor performance para seus volumes do EBS ao minimizar a contenção entre a E/S do Amazon EBS e outro tráfego de sua instância.

Otimize o layout do disco ou a distribuição de arquivos

Use um volume para arquivos de dados e log, outro volume para cargas de trabalho tempdb e volumes Cold HDD (`sc1`) ou Throughput Optimized HDD (`st1`) para backups.

Se você se deparar com um problema relacionado à E/S e quiser separar as cargas de trabalho dos arquivos de dados e de log, considere usar volumes diferentes. Se suo workload exigir que você separe bancos de dados específicos, considere usar um volume dedicado para cada banco de dados.

Normalmente, o tempdb é o alvo da maior E/S, portanto, se esse workload não for separado, ela poderá se tornar um gargalo. Essa separação também ajuda a isolar o tempdb dos dados e arquivos de log do banco de dados do usuário. Você pode usar armazenamento comparativamente mais barato para backups a fim de otimizar seus custos.

Defina o tamanho da unidade de alocação NTFS para 64 KB

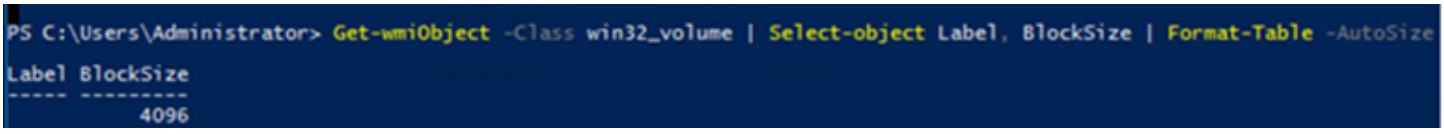
A unidade atômica de armazenamento no SQL Server é uma página com tamanho de 8 KB. Oito páginas fisicamente contíguas formam uma extensão (que tem 64 KB de tamanho). O SQL Server usa extensões para armazenar dados. Portanto, em uma máquina SQL Server, o tamanho da unidade de alocação NTFS para hospedar arquivos de banco de dados SQL (incluindo tempdb) deve ser 64 KB.

Para verificar o tamanho do cluster (alocação de NTFS) de suas unidades, você pode usar PowerShell ou a linha de comando.

Usando PowerShell:

```
Get-wmiObject -Class win32_volume | Select-object Label, BlockSize | Format-Table –
AutoSize
```

A ilustração a seguir mostra um exemplo de saída de PowerShell.



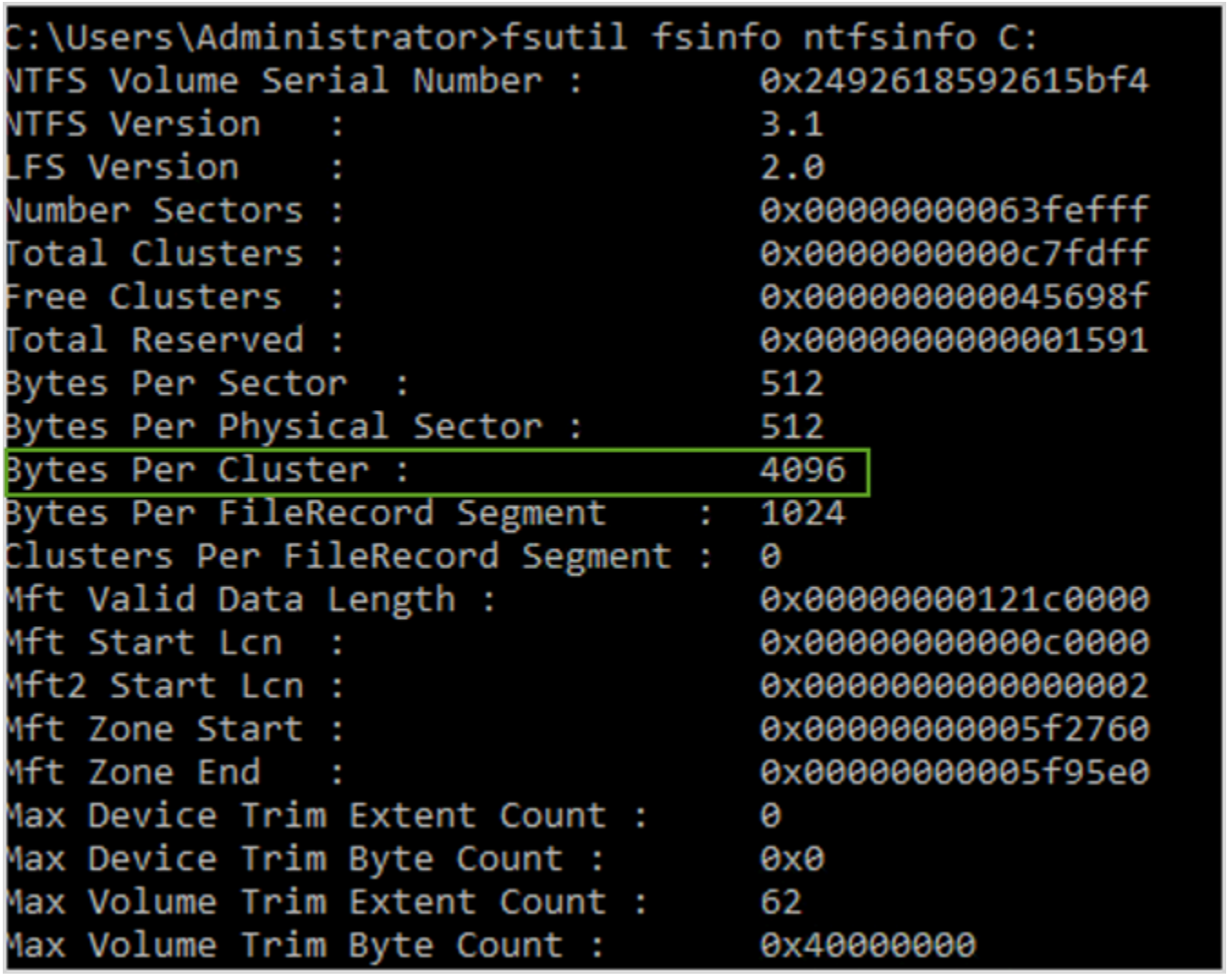
Ou use:

```
$wmiQuery = "SELECT Name, Label, BlockSize FROM win32_volume WHERE FileSystem='NTFS'"
Get-wmiObject -Query $wmiQuery -ComputerName '.' | Sort-Object Name | Select-Object
 Name, Label, BlockSize
```

Usando a linha de comando:

```
$ fsutil fsinfo ntfsinfo C:
```

A ilustração a seguir mostra a saída de exemplo da linha de comando. O valor Bytes por cluster exibe o tamanho do formato em bytes. O exemplo de saída mostra 4096 bytes. Para as unidades que hospedam arquivos de banco de dados do SQL Server, esse valor deve ser 64 KB.



Em alguns casos, o desempenho do SQL Server não depende do tamanho do bloco quando você usa o armazenamento SSD na Amazon EC2. Para obter mais informações, consulte a postagem do

blog Os clientes AWS se beneficiam do tamanho de bloco de 64 KB para armazenamento do SQL Server?

Coloque o tempdb em um armazenamento de instâncias

Ao usar um armazenamento de EC2 instâncias da Amazon, use o volume de armazenamento de instâncias para tempdb. Um armazenamento de instância fornece armazenamento temporário (efêmero) em nível de bloco para sua instância. Recomendamos que você coloque o tempdb em um volume de armazenamento de instâncias por dois motivos: velocidade e custo. O Tempdb é normalmente o banco de dados mais usado, portanto, ele se beneficia da unidade mais rápida disponível. Outro benefício de colocar o tempdb em um armazenamento de instâncias é a economia de custos, porque você não é cobrado separadamente pela E/S em relação ao armazenamento de instâncias.

O Tempdb é recriado sempre que você reinicia o SQL Server, portanto, interromper ou encerrar uma instância não causará perda de dados. No entanto, um volume de armazenamento de instâncias é perdido quando a máquina virtual é iniciada em outro host, porque o disco temporário está conectado localmente à máquina, portanto, planeje com cuidado.

Quando você usa um volume de armazenamento de instâncias:

- Inicialize o volume antes do início do serviço SQL Server. Caso contrário, o procedimento de startup do SQL Server falhará.
- Conceda permissões (controle total) no volume de armazenamento da instância explicitamente para a conta de startup do SQL Server.

Movendo o tempdb para um armazenamento de instâncias

Para mover o tempdb para um volume de armazenamento de instâncias:

1. No Windows, execute `diskmgmt.msc` como administrador para abrir o utilitário do sistema de gerenciamento de disco.
2. Inicialize um novo disco.
3. Abra o menu de contexto (clique com o botão direito do mouse) do disco e escolha Novo volume simples.
4. Complete as instruções, usando as seguintes configurações para formatar o volume:

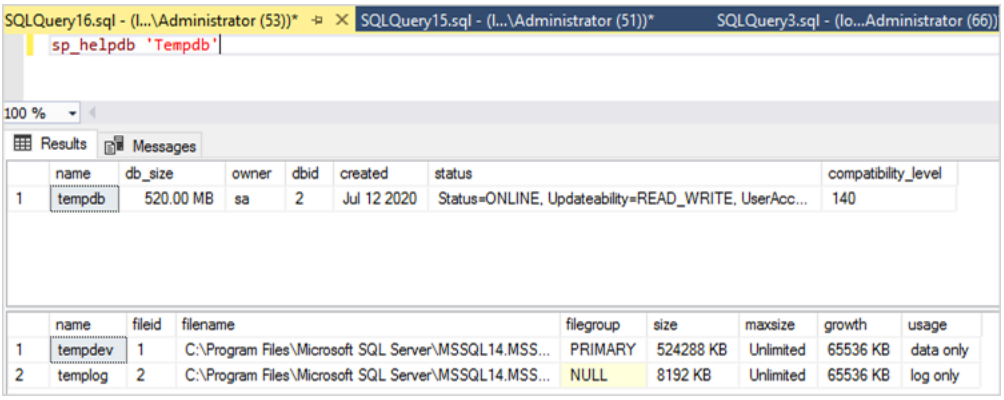
- Sistema de arquivos: NTFS
- Tamanho da unidade de alocação: 64K
- Etiqueta do volume: tempdb

Para obter mais informações, consulte Documentação de gerenciamento de disco de disco no site da Microsoft.

5. Conecte-se à instância do SQL Server e execute o comando a seguir para observar o nome do arquivo lógico e físico do banco de dados tempdb:

```
$ sp_helpdb 'tempdb'
```

O exemplo a seguir mostra o comando e a saída.

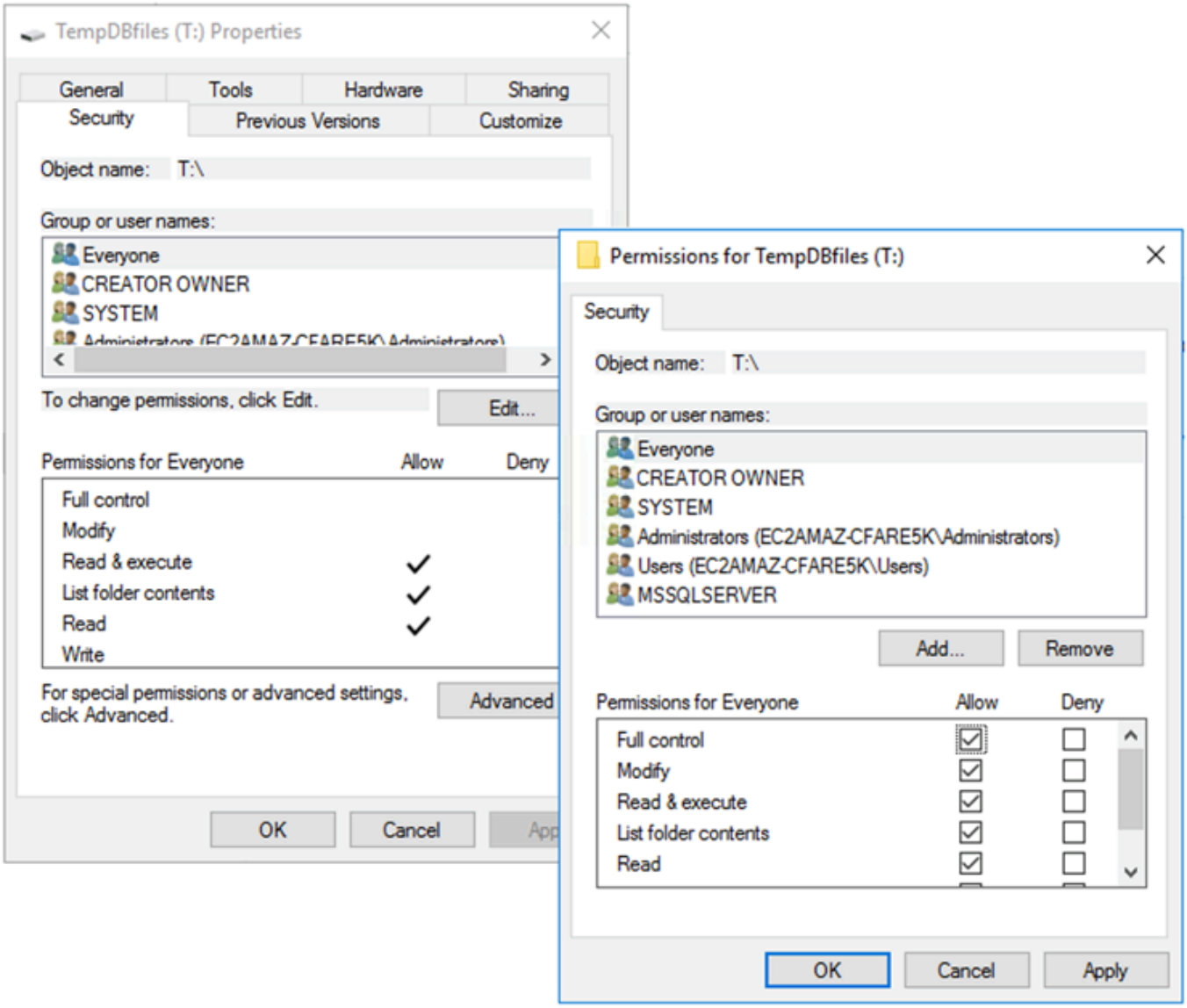


6. Mova o arquivo tempdb para o novo local. Lembre-se de definir todos os arquivos do banco de dados tempdb com o mesmo tamanho inicial. O exemplo de script do SQL Server a seguir move os arquivos tempdb para a unidade T e define os arquivos de dados para o mesmo tamanho.

```
USE master
GO
ALTER DATABASE TempDB MODIFY FILE (NAME = tempdev, FILENAME = 'T:\tempdb.mdf',SIZE
 = 524288KB)
GO
ALTER DATABASE TempDB MODIFY FILE (NAME = temp2, FILENAME = 'T:
\tempdb_mssql_2.ndf',SIZE = 524288KB)
GO
ALTER DATABASE TempDB MODIFY FILE (NAME = temp3, FILENAME = 'T:
\tempdb_mssql_3.ndf',SIZE = 524288KB)
GO
```

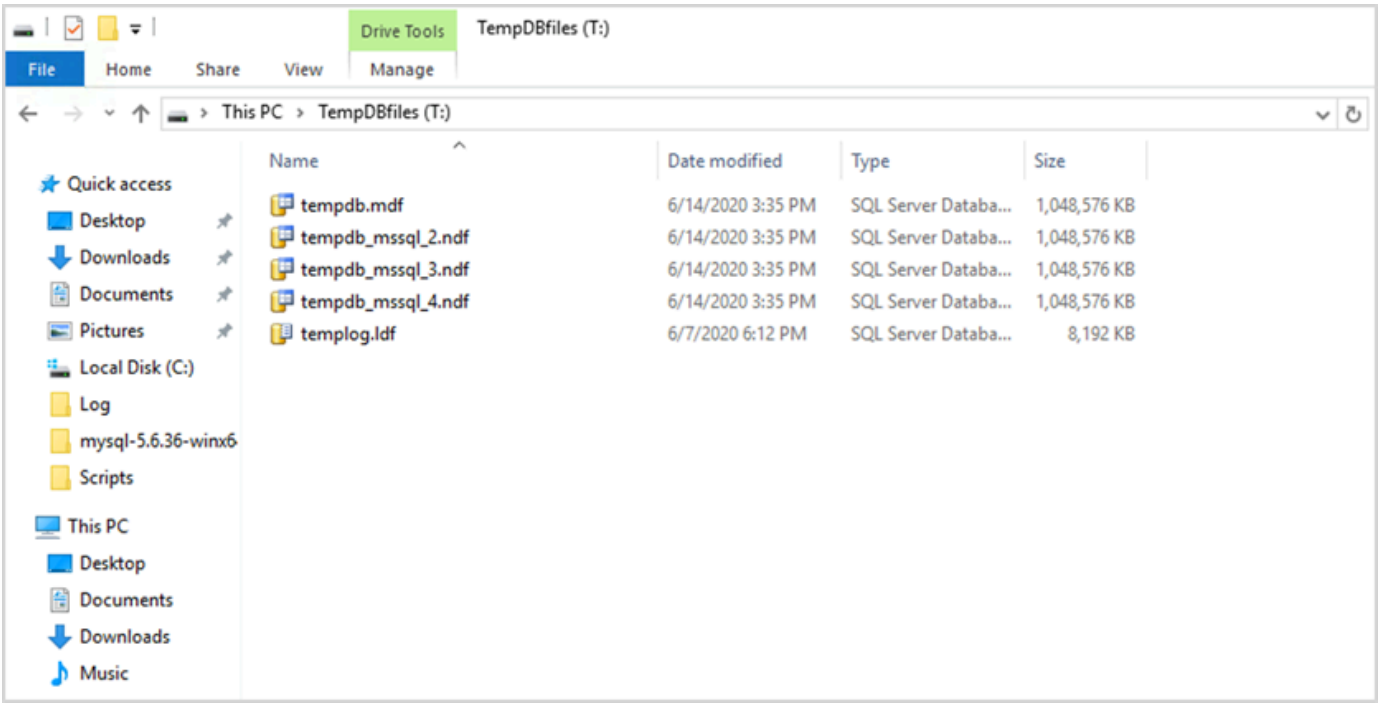
```
ALTER DATABASE TempDB MODIFY FILE (NAME = temp4, FILENAME = 'T:
\tempdb_mssql_4.ndf',SIZE = 524288KB)
GO
ALTER DATABASE TempDB MODIFY FILE (NAME = templog, FILENAME = 'T:\templog.ldf')
GO
```

7. Conceda à conta de startup do SQL Server permissões para o novo local do banco de dados tempdb, para que ele possa criar os arquivos tempdb, conforme mostrado na captura de tela a seguir.



8. Reinicie o SQL Server para usar o novo local para tempdb.

   Você verá os arquivos tempdb criados no novo local, como mostrado na captura de tela a seguir.



9. Exclua os arquivos tempdb do local antigo.

Para garantir que o volume de armazenamento da instância seja inicializado antes do início do SQL Server em caso de reinicialização ou início/parada da instância, siga as etapas na próxima seção. Caso contrário, o startup do SQL Server falhará porque o tempdb não foi inicializado.

Inicializar o armazenamento de instâncias

Para inicializar o armazenamento de dados:

1. Abra o Gerenciador de Serviços do Windows (`services.msc`) e defina o SQL Server e seus serviços dependentes (por exemplo, SQL Server Agent) para serem iniciados manualmente. (Você usará um script para iniciá-lo quando o volume de armazenamento da instância estiver pronto.)
2. Crie um PowerShell script para passar para a EC2 instância da Amazon como dados do usuário. Este script faz o seguinte:
   - Detecta armazenamento efêmero e cria uma unidade tempdb para ele (unidade T no exemplo).
   - Atualiza o disco temporário se a EC2 instância parar e reiniciar.

- Concede à conta de startup do SQL Server controle total do volume tempdb recém-inicializado. O exemplo pressupõe uma instância padrão, então ele usa o `NT SERVICE\MSSQLSERVER`. Para uma instância nomeada, isso normalmente é `NT SERVICE\MSSQL$<InstanceName>` por padrão.
- Salva o script em um volume local (`c:\scripts` no exemplo) e atribui a ele um nome de arquivo (`InstanceStoreMapping.ps1`).
- Cria uma tarefa programada usando o Agendador de Tarefas do Windows. Essa tarefa executa o PowerShell script na inicialização.
- Inicia o SQL Server e o SQL Server Agent após as ações anteriores.

O script a seguir é do segundo laboratório do Workshop do grupo de disponibilidade do MS-SQL com algumas alterações. Copie o script no campo Dados do usuário ao iniciar a EC2 instância e personalize-o conforme necessário.

```
<powershell>
# Create pool and virtual disk for TempDB using the local NVMe, ReFS 64K, T: Drive
    $NVMe = Get-PhysicalDisk | ? { $_.CanPool -eq $True -and $_.FriendlyName -eq "NVMe
 Amazon EC2 NVMe"}
    New-StoragePool -FriendlyName TempDBPool -StorageSubsystemFriendlyName "Windows
 Storage*" -PhysicalDisks $NVMe
    New-VirtualDisk -StoragePoolFriendlyName TempDBPool -FriendlyName TempDBDisk -
ResiliencySettingName simple -ProvisioningType Fixed -UseMaximumSize
    Get-VirtualDisk -FriendlyName TempDBDisk | Get-Disk | Initialize-Disk -Passthru
 | New-Partition -DriveLetter T -UseMaximumSize | Format-Volume -FileSystem ReFS -
AllocationUnitSize 65536 -NewFileSystemLabel TempDBfiles -Confirm:$false
    # Script to handle NVMe refresh on start/stop instance
    $InstanceStoreMapping = {
    if (!(Get-Volume -DriveLetter T)) {
        #Create pool and virtual disk for TempDB using mirroring with NVMe
        $NVMe = Get-PhysicalDisk | ? { $_.CanPool -eq $True -and $_.FriendlyName -eq
 "NVMe Amazon EC2 NVMe"}
        New-StoragePool -FriendlyName TempDBPool -StorageSubsystemFriendlyName "Windows
 Storage*" -PhysicalDisks $NVMe
        New-VirtualDisk -StoragePoolFriendlyName TempDBPool -FriendlyName TempDBDisk -
ResiliencySettingName simple -ProvisioningType Fixed -UseMaximumSize
        Get-VirtualDisk -FriendlyName TempDBDisk | Get-Disk | Initialize-Disk -Passthru
 | New-Partition -DriveLetter T -UseMaximumSize | Format-Volume -FileSystem ReFS -
AllocationUnitSize 65536 -NewFileSystemLabel TempDBfiles -Confirm:$false
         #grant SQL Server Startup account full access to the new drive
```

```
        $item = gi -literalpath "T:\"
        $acl = $item.GetAccessControl()
        $permission="NT SERVICE\MSSQLSERVER","FullControl","Allow"
        $rule = New-Object System.Security.AccessControl.FileSystemAccessRule
 $permission
        $acl.SetAccessRule($rule)
        $item.SetAccessControl($acl)
        #Restart SQL so it can create tempdb on new drive
        Stop-Service SQLSERVERAGENT
        Stop-Service MSSQLSERVER
        Start-Service MSSQLSERVER
        Start-Service SQLSERVERAGENT
        }
    }
    New-Item -ItemType Directory -Path c:\Scripts
    $InstanceStoreMapping | set-content c:\Scripts\InstanceStoreMapping.ps1
# Create a scheduled task on startup to run script if required (if T: is lost)
    $action = New-ScheduledTaskAction -Execute 'Powershell.exe' -Argument 'c:\scripts
\InstanceStoreMapping.ps1'
    $trigger =  New-ScheduledTaskTrigger -AtStartup
    Register-ScheduledTask -Action $action -Trigger $trigger -TaskName "Rebuild
 TempDBPool" -Description "Rebuild TempDBPool if required" -RunLevel Highest -User
 System
</powershell>
```

Usando a extensão do grupo de buffer

Se você planeja usar a extensão do buffer pool, você também pode considerar colocá-la em um volume efêmero. No entanto, é altamente recomendável testá-lo minuciosamente antes da implementação. Evite usar o mesmo volume para a extensão do buffer pool e o tempdb.

> **Note**
>
> Embora a extensão do buffer pool possa ser útil em alguns casos, ela não substitui a RAM. Antes de decidir usá-lo, consulte os detalhes fornecidos no site da Microsoft.

Evite incompatibilidades entre os núcleos da CPU

Escolher um servidor que tenha um número maior de núcleos do que as coberturas de sua licença pode causar distorção da CPU e desperdício de energia da CPU. Isso se deve ao mapeamento entre

os núcleos lógicos e reais. Quando você usa o SQL Server com uma Licença de Acesso para Cliente (CAL), alguns agendadores serão `VISIBLE ONLINE` e os demais serão `VISIBLE OFFLINE`. Isso pode levar a problemas de desempenho com topologias de acesso não uniforme à memória (NUMA), porque os nós do agendador não estão sendo utilizados de maneira ideal.

Por exemplo, se você executar o SQL Server em uma instância `m5.24xlarge`, ele detectará dois soquetes com 24 núcleos e 48 processadores lógicos por soquete, o que resultará em um total de 96 processadores lógicos. Se você tiver uma licença para apenas 48 núcleos, verá uma mensagem semelhante à seguinte no log de erros do SQL Server:

2020-06-08 12:35:27.37 Server O SQL Server detectou 2 soquetes com 24 núcleos por soquete e 48 processadores lógicos por soquete, 96 processadores lógicos no total; usando 48 processadores lógicos baseados no licenciamento do SQL Server. Esta é uma mensagem informativa; nenhuma ação do usuário é necessária.

Se você observar uma diferença entre o total de núcleos e o número de núcleos usados pelo SQL Server, verifique se há desequilíbrio no uso da CPU ou use um tipo de servidor que tenha o mesmo número de núcleos que sua licença suporta.

CPU skew: para o tipo de instância em nosso exemplo (`m5.24xlarge`), o SQL Server cria oito nós NUMA por padrão. Somente quatro desses nós (ID do nó pai 0,1,2,3) têm agendadores com o status `VISIBLE ONLINE`. Os horários restantes são todos `VISIBLE OFFLINE`. Essa disparidade entre os programadores pode levar à degradação do desempenho.

Para verificar as informações e o status do agendador, use:

```
$ select * from sys.dm_os_schedulers
```

Se você quiser usar uma instância de servidor que tenha um número maior de núcleos do que a sua licença do SQL Server suporta, considere personalizar o número de núcleos seguindo as instruções em Especificação de opções de CPU para sua instância na documentação da Amazon EC2 .

Teste o desempenho dos discos

Recomendamos que você verifique o desempenho do disco usando uma ferramenta como DiskSpd. Essa ferramenta fornece uma estimativa da velocidade do disco antes de você executar testes específicos do SQL Server. É muito importante avaliar o desempenho do disco, porque um volume do EBS funciona de forma diferente de uma SAN tradicional em um ambiente on-premises. A falta de

testes de desempenho adequados pode levar a um desempenho lento inesperado após a migração. Você também pode executar testes personalizados com DiskSpd.

Habilitar a inicialização instantânea de arquivos

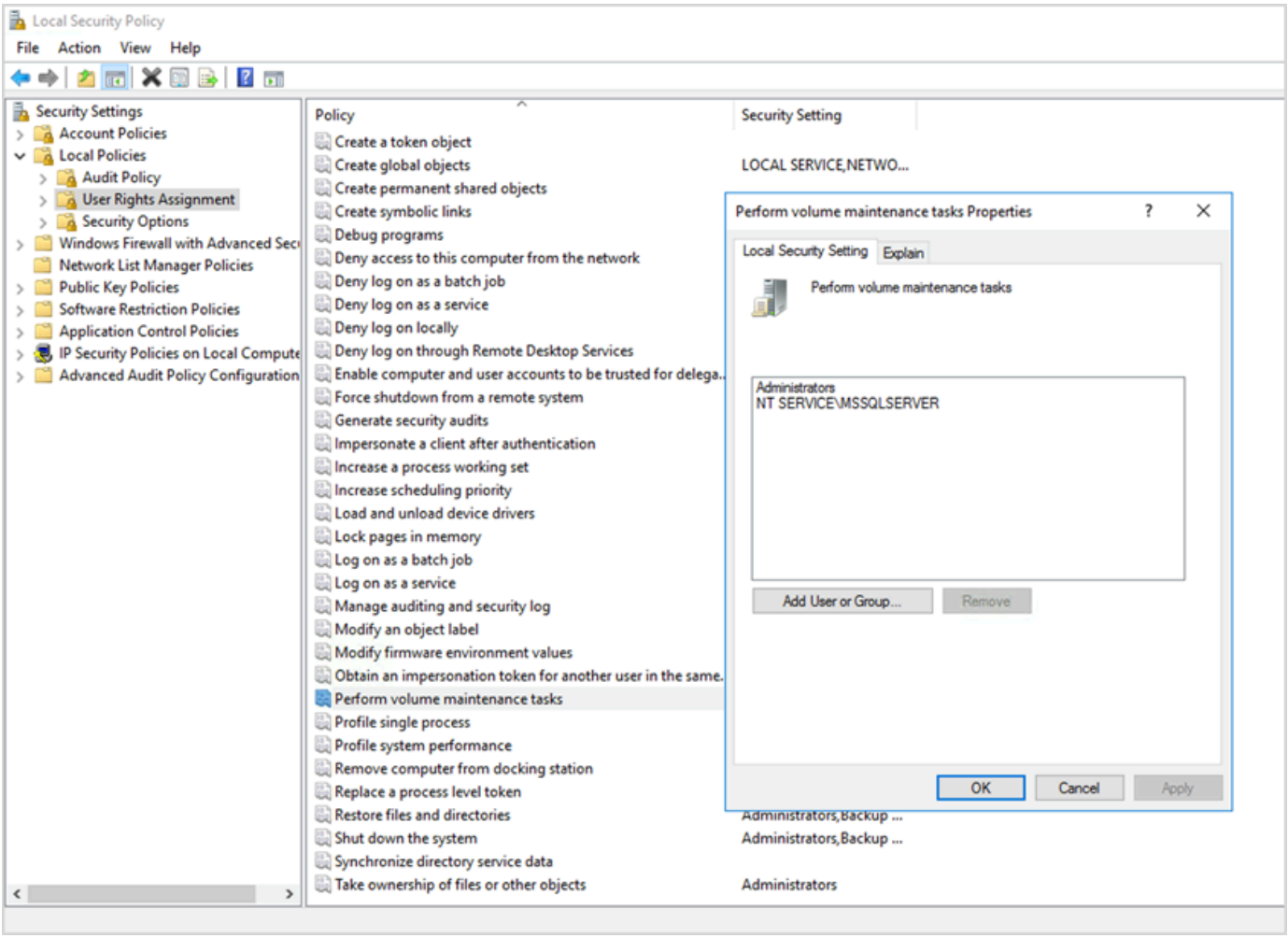
No SQL Server, use a configuração Realizar tarefas de manutenção de volume para habilitar a inicialização instantânea de arquivos, a menos que você esteja seguindo restrições regulatórias. Essa opção aumenta significativamente o desempenho do crescimento automático de arquivos.

Essa configuração ignora as operações de zerar os arquivos de dados. Ou seja, os arquivos de dados não são preenchidos com valores zero (`0x0`) quando são inicializados, o que pode levar muito tempo. O conteúdo atual no disco é sobrescrito somente quando novos dados são gravados no disco.

> **Note**
>
> Os arquivos de log não se beneficiam da inicialização instantânea de arquivos.

Para habilitar a inicialização instantânea de arquivos:

1. Na tela Iniciar, execute o `secpol.msc` para abrir o console da Política de segurança local.
2. Escolha Políticas locais, Atribuição de direitos ao usuário, Executar tarefas de manutenção e adicione a conta de serviço do SQL Server, conforme mostrado na captura de tela a seguir.



3. Reinicie o servidor X para que as alterações entrem em vigor.

Para obter mais informações sobre inicialização instantânea de arquivos, consulte a Documentação doSQL Server no site da Microsoft.

> **Observações de segurança**
>
> Quando você usa a inicialização instantânea de arquivos, o disco é sobrescrito somente quando novos dados são gravados nos arquivos, então é possível ler o conteúdo excluído. Enquanto a unidade está conectada à instância, a lista de controle de acesso discricionário (DACL) no arquivo reduz o risco de divulgação de informações, pois permite acesso somente à conta de serviço do SQL Server e ao administrador local. No entanto, quando o arquivo é desanexado, ele pode ser acessado. Se a divulgação do conteúdo excluído for uma

preocupação, você deve desabilitar a inicialização instantânea de arquivos para a instância do SQL Server.

Bloquear páginas na memória

Ative a opção Bloquear páginas na memória para a conta de startup do SQL Server para garantir que o sistema operacional não corte o conjunto de trabalho do SQL Server.

Para verificar se essa opção está habilitada, use a seguinte consulta SQL:

```
SELECT sql_memory_model, sql_memory_model_desc
FROM sys.dm_os_sys_info;
```

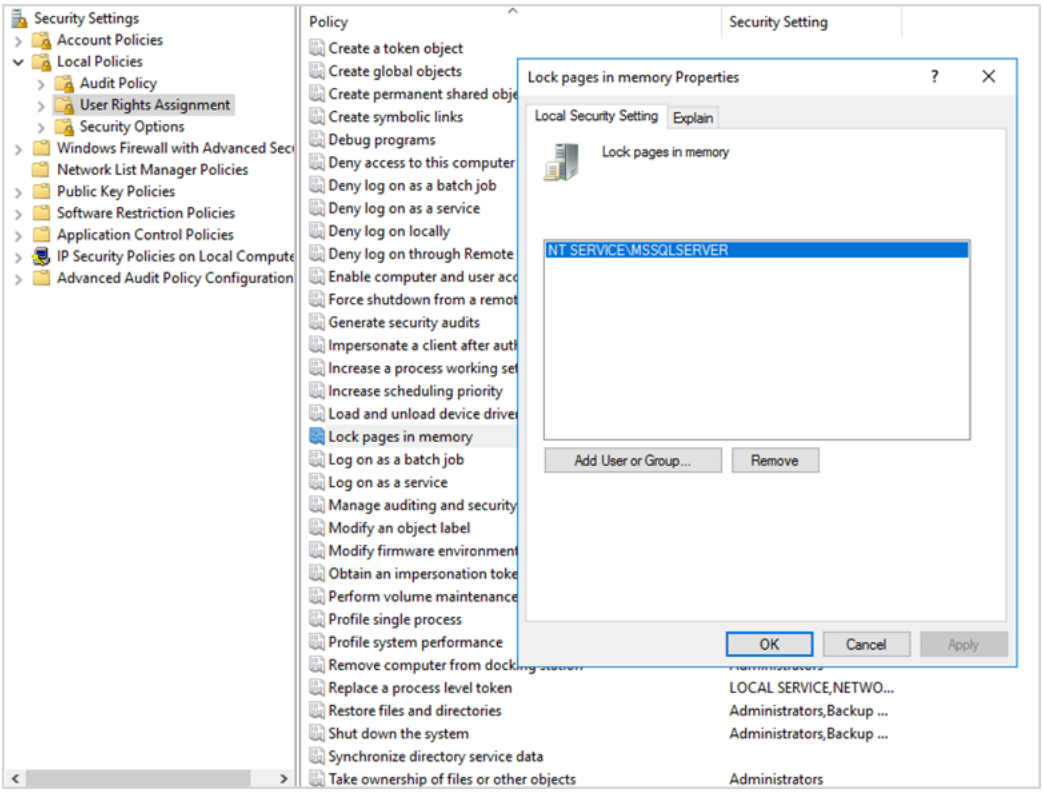
Saída:

```
sql_memory_model     sql_memory_model_desc
1                   CONVENTIONAL

"CONVENTIONAL" means it’s not enabled.
```

Para ativar a opção Bloquear páginas na memória:

1. Na tela Iniciar, execute o `secpol.msc` para abrir o console da Política de segurança local.
2. Escolha Políticas locais, Atribuição de direitos ao usuário, Bloquear páginas na memória e adicione a conta de serviço do SQL Server, conforme mostrado na captura de tela a seguir.



3. Reinicie o servidor X para que as alterações entrem em vigor.
4. Use a consulta SQL a seguir para confirmar se a opção Bloquear páginas na memória está ativada:

```
SELECT sql_memory_model, sql_memory_model_desc
FROM sys.dm_os_sys_info;
```

Saída:

```
sql_memory_model     sql_memory_model_desc
2                    LOCK_PAGES

"LOCK_PAGES" means it’s enabled.
```

Para obter mais informações sobre o modelo de memória do SQL Server, consulte `sql_memory_model` e `sql_memory_model_desc` na documentação sys.dm_os_sys_info no site da Microsoft.

Desative o descarregamento de TCP e as configurações de RSS

Se você observar problemas de conectividade aleatórios, como erros de nível de transporte ou erros de transmissão de pacotes ao executar cargas de trabalho SQL, talvez queira desabilitar o descarregamento de TCP e as configurações de RSS.

- O descarregamento de TCP (recurso Descarga de chaminé TCP) descarrega o processamento de pacotes TCP/IP do processador para o adaptador de rede, para liberar a CPU para outras tarefas.
- O escalonamento do lado do recebimento (RSS) ajuda a distribuir o processamento do tráfego de entrada da rede em sistemas multiprocessados. Ele equilibra a carga do processamento da rede de forma eficiente entre os CPUs.

Para verificar suas configurações atuais, no prompt de comando, execute o comando netsh:

```
$ netsh int tcp show global
```

Veja a saída desse comando. Neste exemplo, o Estado de dimensionamento do lado de recepção e o Estado de descarga de chaminé estão ambos desativados.

```
C:\Users\Administrator>netsh int tcp show global
Querying active state...

TCP Global Parameters
----------------------------------------------
Receive-Side Scaling State          : disabled
Chimney Offload State               : disabled
NetDMA State                        : disabled
Direct Cache Access (DCA)           : disabled
Receive Window Auto-Tuning Level    : normal
Add-On Congestion Control Provider  : none
ECN Capability                      : enabled
RFC 1323 Timestamps                 : disabled
Initial RTO                         : 3000
Receive Segment Coalescing State    : enabled
Non Sack Rtt Resiliency             : disabled
Max SYN Retransmissions             : 2
TCP Fast Open                       : disabled
```

Para obter informações de descarga de tarefas sobre uma conexão específica, no prompt de comando, execute:

```
netstat -t
```

e verifique o valor da coluna de estado de descarga.

Para desabilitar o descarregamento de TCP e o RSS para Windows Server 2008 e 2012, execute estes comandos no prompt de comando:

```
netsh int ip set global taskoffload=disabled
netsh int tcp set global chimney=disabled
netsh int tcp set global rss=disabled
netsh int tcp set global netdma=disabled
```

Para saber mais sobre as configurações de capacidade, consulte .

- Recursos de descarga de chaminé TCP, dimensionamento do lado de recepção e acesso à memória direta da rede e Introdução ao dimensionamento do lado de recepção no site da Microsoft
- Descarregamento de TCP na documentação da Amazon EC2
- Solução de problemas de drivers fotovoltaicos na documentação da Amazon EC2

> **Important**
>
> Não use IPsec Task Offload ou TCP Chimney Offload. De acordo com a documentação da Microsoft, esses recursos de descarga estão obsoletos no Windows Server 2016 e podem não ter suporte em versões futuras. O uso desses recursos pode afetar adversamente o desempenho.

Determine seus requisitos de IOPS e throughput

Use o Monitor de Desempenho do Windows para obter informações sobre IOPS e taxa de transferência.

Para abrir o Windows Performance Monitor, execute perfmon no prompt de comando. Os dados de IOPS e de throughput são fornecidos pelos seguintes contadores de desempenho:

- Disco reads/sec + disk writes/sec = IOPS
- Leitura do disco bytes/sec + disk write bytes/sec = taxa de transferência

Recomendamos que você obtenha os dados de IOPS e de throughput para o horário de pico de uso e também durante um ciclo de workload típico, para obter uma boa estimativa de seus requisitos. Certifique-se de que o tipo de instância escolhido para o SQL Server ofereça suporte a esses requisitos de E/S.

É importante acertar essa estimativa. Caso contrário, você pode provisionar seus recursos em excesso, o que pode resultar em recursos subutilizados, ou provisionar insuficientemente seus recursos, o que pode resultar em graves problemas de desempenho.

Use striping para contornar as limitações de IOPS e de throughput

Quando seu aplicativo SQL Server exigir mais do que o máximo de IOPS e throughput disponíveis em um volume do EBS, considere reduzir seu volume do EBS para superar essas limitações.

Os volumes striping (RAID) ajuda você a atender aos requisitos de IOPS e throughput ,e é limitado pelo máximo de IOPS e largura de banda suportados por uma instância específica. Para obter mais informações sobre as opções de distribuição, consulte a configuração de RAID na documentação da Amazon EC2. Você também pode usar Espaços de Armazenamento em um servidor autônomo. Para obter mais informações, consulte a Documentação da Microsoft.

Excluir arquivos do SQL Server do software antivírus

Ao definir as configurações do software antivírus, certifique-se de excluir os arquivos e diretórios do SQL Server da verificação de vírus. Para obter detalhes e uma lista de arquivos e diretórios a serem excluídos, consulte Como escolher software antivírus para execução em computadores que rodam o SQL Server no site da Microsoft.

Se você não excluir esses arquivos do SQL Server, eles poderão ser corrompidos ou colocados em quarentena pelo software antivírus quando o SQL Server precisar usá-los. Não excluir esses arquivos também pode causar problemas de desempenho.

Configurar o Servidor SQL

Esta seção fornece as melhores práticas para configurar seus bancos de dados do SQL Server para ajustar o desempenho, evitar armadilhas comuns e atender aos requisitos de segurança e disponibilidade. Você pode implementar essas mudanças antes ou depois de migrar seus bancos de dados para a Amazon EC2. As seções a seguir fornecem dicas de configuração e melhores práticas.

Tópicos

- Configure o tempdb para reduzir a contenção
- Defina o MAXDOP para o maior desempenho
- Alterar o limite de custo do paralelismo
- Otimize para workloads pontuais
- Use sinalizadores de rastreamento para melhorar o desempenho
- Instalar os patches mais recentes
- Limite a memória máxima do servidor para evitar pressão na memória
- Use o nível mais alto de compatibilidade do banco de dados
- Controle o número de VLFs
- Verifique as configurações de crescimento automático do banco de dados

Configure o tempdb para reduzir a contenção

Recomendamos que você configure o tempdb com vários arquivos de dados do mesmo tamanho e com o mesmo fator de crescimento.

Em um servidor de banco de dados ocupado, que usa muito o tempdb, você pode notar um bloqueio severo quando o servidor está sob uma carga pesada. Você pode notar que as tarefas estão aguardando recursos de espera que apontam para páginas em tempdb. Essas páginas podem ser páginas de Page Free Space (PFS) e Shared Global Allocation Map (SGM) que têm o formato 2:*x*: *x* (por exemplo, 2:1:1 ou 2:1:2).

Para melhorar a simultaneidade do tempdb, você pode aumentar o número de arquivos de dados no tempdb para maximizar a largura de banda do disco e reduzir a contenção nas estruturas de alocação. Veja aqui algumas diretrizes:

- Se o número de processadores lógicos for igual ou menor que 8: use o mesmo número de arquivos de dados e processadores lógicos.
- Se o número de processadores lógicos for maior que 8: use 8 arquivos de dados.

Se a contenção persistir, aumente o número de arquivos de dados em múltiplos de 4 até que a contenção seja corrigida e até o número de processadores lógicos no servidor. Isso ajudará a evitar a contenção do SGAM no tempdb. Se você estiver usando o SQL Server 2014 ou uma versão anterior, também precisará habilitar o trace flag 1118. Esse sinalizador força as alocações de páginas em extensões uniformes em vez de extensões mistas, o que minimiza as digitalizações na página SGAM e reduz a contenção.

Começando com o SQL Server 2016 (13.x), esse comportamento é controlado pelas opções `AUTOGROW_SINGLE_FILE` e `AUTOGROW_ALL_FILES` do `ALTER DATABASE`. Por exemplo:

```
alter database <database name> MODIFY FILEGROUP [PRIMARY] AUTOGROW_ALL_FILES
```

Para obter mais informações sobre a configuração dessas opções, consulte a documentação do Microsoft SQL Server.

Defina o MAXDOP para o maior desempenho

O grau máximo de paralelismo (MAXDOP) é uma opção de configuração do servidor para executar o SQL Server em vários. CPUs Ele controla o número de processadores usados para executar uma única instrução na execução paralela do plano. O valor padrão é 0, que permite que o SQL Server use todos os processadores disponíveis. Isso pode afetar o desempenho e não é ideal para a maioria dos casos de uso.

Use as diretrizes a seguir ao configurar o valor MAXDOP para o SQL Server.

Nós NUMA	Processadores lógicos	Valor MAXDOP
Único	≤ 8	4, 2 ou número de núcleos (para um ou dois núcleos)
Único	> 8	8, 4, ou 2
Vários	≤ 16	8, 4, ou 2

Nós NUMA	Processadores lógicos	Valor MAXDOP
Vários	> 16	16, 8, 4, ou 2

> **Note**
>
> Definir o MAXDOP como 2, 4 ou 8 geralmente fornece os melhores resultados na maioria dos casos de uso. Recomendamos que você teste sua workload e monitore qualquer tipo de espera relacionada ao paralelismo, como `CXPACKET`.

Você pode usar a consulta a seguir para reunir a configuração NUMA atual para o SQL Server 2016 e versionamentos posteriores:

```
select @@SERVERNAME,
SERVERPROPERTY('ComputerNamePhysicalNetBIOS'),
cpu_count,
hyperthread_ratio,
softnuma_configuration,
softnuma_configuration_desc,
socket_count,
numa_node_count
from
sys.dm_os_sys_info
```

em que:

- `cpu_count`refere-se ao número de lógicos CPUs no sistema.
- `hyperthread_ratio` é a proporção do número de núcleos que são expostos por um processador físico.
- `softnuma_configuration` é `0`, `1`, ou `2`:
  - `0 (OFF)`: default
  - `1 (automated)`: soft-NUMA
  - `2 (manual)`: soft-NUMA
- `softnuma_configuration_desc` é `OFF`, `ON`, ou `MANUAL`:
  - `OFF` indica que o atributo soft-NUMA está desativado.

  - `ON` indica que o SQL Server decide automaticamente os tamanhos dos nós NUMA.
  - `MANUAL` indica que o soft-NUMA está configurado manualmente.
- `socket_count` é o número de soquetes do processador.
- `numa_node_count` é o número de nós NUMA disponíveis no sistema.

Para verificar o valor atual do MAXDOP, use:

```
$ sp_configure 'max_degree_of_parallelism'
```

Para obter mais informações sobre o MAXDOP, consulte a documentação do Microsoft SQL Server.

Alterar o limite de custo do paralelismo

O limite de custo do paralelismo determina quais consultas são candidatas à execução paralela. O valor padrão dessa propriedade é 5, o que significa que o otimizador muda para um plano paralelo se o custo de um plano serial for maior que 5 (que se refere a uma unidade de custo abstrata, não ao tempo estimado). Recomendamos que você configure esta propriedade como um número maior.

O valor padrão era apropriado quando os processadores tinham preços altos, a capacidade de processamento era baixa e o processamento de consultas era mais lento do que agora. Atualmente, os processadores são muito mais rápidos. Como resultado, consultas comparativamente menores (por exemplo, considerando um limite de custo de 32) não se beneficiarão muito da execução paralela, especialmente devido à sobrecarga associada à coordenação da execução paralela.

Na maioria dos casos, um limite de custo da configuração de paralelismo no valor de 50 é um bom ponto de partida. Veja um exemplo de como configurar o limite de custo do paralelismo:

```
USE sampledb;
GO
EXEC sp_configure 'show advanced options', 1 ;
GO
RECONFIGURE
GO
EXEC sp_configure 'cost threshold for parallelism', 50 ;
GO
RECONFIGURE
GO
```

Otimize para workloads pontuais

Ative a opção otimizar para workloads pontuais para melhorar a eficiência do cache do plano para workloads que contenham muitos lotes ad hoc de uso único. Quando você executa inicialmente uma consulta pontual, o mecanismo do banco de dados armazena em cache um esboço do plano compilado em vez do plano de execução completo, economizando espaço no cache do plano. Se você executar o lote pontual novamente, o mecanismo do banco de dados reconhecerá que o lote já foi executado antes e substituirá o esboço do plano compilado pelo plano completo compilado no cache do plano.

Para verificar se essa opção está habilitada, use a consulta:

```
$ sp_configure 'optimize for ad hoc workloads'
```

Para obter mais informações sobre como otimizar workloads pontuais, consulte a documentação do Microsoft SQL Server.

Use sinalizadores de rastreamento para melhorar o desempenho

Considere o uso de sinalizadores de rastreamento do SQL Server que sejam aplicáveis ao seu ambiente para melhorar o desempenho. Por exemplo:

- 4199: Ativa as alterações do otimizador de consultas (QO) lançadas nas Atualizações Cumulativas () e nos Service Packs (CUs) do SQL Server. SPs
- 8048: Converte objetos de memória particionados por NUMA em objetos de memória particionados por CPU.
- 9024: Converte um objeto de memória do pool de logs global em um objeto de memória particionado por NUMA.

Os exemplos a seguir ilustram como ativar e desativar os sinalizadores de rastreamento para o SQL Server na Amazon. EC2 Se tiver problemas ao ativar o rastreamento, verifique se você tem as permissões apropriadas para a conta.

Para ativar o sinalizador de rastreamento 4199, execute:

```
dbcc traceon (4199, -1);
```

Para verificar o status do sinalizador de rastreamento, execute:

```
dbcc tracestatus (4199);
```

Para desativar o sinalizador de rastreamento 4199, execute:

```
dbcc traceoff (4199, -1);
dbcc tracestatus (4199);
```

Para obter uma lista completa dos sinalizadores de rastreamento, consulte a documentação do Microsoft SQL Server.

Instalar os patches mais recentes

A partir do SQL Server 2017, a Microsoft interrompeu o lançamento dos Service Packs () SPs. Ele lança somente atualizações cumulativas (CUs) e atualizações críticas (GDRs).

SPs inclua correções cruciais para o SQL Server, portanto, certifique-se de que a controladora mais recente tenha sido instalada. Além disso, se possível, instale o pacote de CUs mais recente.

Para obter informações sobre as atualizações mais recentes do SQL Server, consulte Atualizações mais recentes para o Microsoft SQL Server no site da Microsoft.

Limite a memória máxima do servidor para evitar pressão na memória

Para fins de desempenho, o SQL Server não libera a memória que já tenha alocada. Quando o SQL Server é iniciado, ele usa lentamente a memória especificada na opção min_server_memory e continua a crescer até atingir o valor especificado na opção max_server_memory. (Para obter mais informações sobre essas configurações, consulte as Opções de configuração da memória do servidor na documentação do SQL Server.)

A memória do SQL Server tem dois componentes: o buffer pool e o non-buffer pool (também chamado de memória a sair ou MTL). O valor da opção max_server_memory determina o tamanho do buffer pool do SQL Server, que consiste em cache de buffer, cache de procedimento, cache de planos, estruturas de buff e outros caches.

A partir do SQL Server 2012, min_server_memory e max_server_memory respondem por todas as alocações de memória para todos os caches, incluindo `SQLGENERAL`, `SQLBUFFERPOOL`, `SQLQUERYCOMPILE`, `SQLQUERYPLAN`, `SQLQUERYEXEC`, `SQLOPTIMIZER` e `SQLCLR`. Para obter uma lista completa dos gerenciadores de memória no max_server_memory, consulte sys.dm_os_memory_clerks na documentação do Microsoft SQL Server.

Para verificar o valor atual da max_server_memory, use o comando:

```
$ sp_configure 'max_server_memory'
```

Recomendamos que você limite a max_server_memory a um valor que não cause pressão na memória em todo o sistema. Não há uma fórmula universal que se aplique a todos os ambientes, mas fornecemos algumas diretrizes nesta seção. max_server_memory é uma opção dinâmica e, portanto, pode ser alterada no tempo de execução.

Como ponto de partida, você pode determinar a max_server_memory da seguinte forma:

```
max_server_memory = total_RAM - (memory_for_the_OS + MTL)
```

onde:

- A memória do sistema operacional é de 1 a 4 GB.
- A MTL (memória para sair) inclui o tamanho da pilha, que é de 2 MB em máquinas de 64 bits por thread de trabalho e pode ser calculada da seguinte forma: `MTL = stack_size * max_worker_threads`

Você também pode usar:

```
max_server_memory = total_RAM - (1 GB for the OS
 + memory_basis_amount_of_RAM_on_the_server)
```

onde a quantidade básica de memória RAM é determinada da seguinte forma:

- Se a RAM no servidor estiver entre 4 GB e 16 GB, deixe 1 GB a cada 4 GB de RAM. Por exemplo, para um servidor com 16 GB, deixe 4 GB.
- Se a RAM no servidor for superior a 16 GB, deixe 1 GB a cada 4 GB de RAM até 16 GB, e 1 GB por 8 GB de RAM acima de 16 GB.

Por exemplo, se um servidor tiver 256 GB de RAM, o cálculo será:

- 1 GB para o sistema operacional
- Até 16 GB de RAM: 16/4 = 4 GB
- RAM restante acima de 16 GB: (256-16)/8 = 30
- Total de RAM a ser deixado: 1 + 4 + 30 = 35 GB
- max_server_memory: 256 - 35 = 221 GB

Após a configuração inicial, monitore a memória que você pode liberar durante um período normal de workload para determinar se você precisa aumentar ou diminuir a memória alocada para o SQL Server.

> **Note**
>
> O Windows sinaliza a notificação de poucos recursos de memória aos 96 MB, então você quer um buffer, mas pode definir que os Mbytes disponíveis estejam acima de 1 GB em servidores maiores com 256 GB ou mais de RAM.

Para obter informações adicionais, consulte o Guia de Arquitetura de Gerenciamento de Memória na documentação do Microsoft SQL Server.

Use o nível mais alto de compatibilidade do banco de dados

Verifique se você está usando o nível atual de compatibilidade do banco de dados para aproveitar as melhorias mais recentes no SQL Server. É importante verificar isso porque, ao restaurar um banco de dados de um versionamento inferior para um versionamento superior, ele manterá o nível de compatibilidade do versionamento inferior. Alguns dos aprimoramentos mais recentes do banco de dados funcionam somente quando você define a compatibilidade do banco de dados no nível mais recente disponível para o versionamento do mecanismo que você instalou.

Para verificar a compatibilidade atual do banco de dados, use:

```
$ select name, compatibility_level from sys.databases
```

Para obter mais informações sobre os níveis de compatibilidade do banco de dados, consulte a documentação do Microsoft SQL Server.

Controle o número de VLFs

Defina previamente o tamanho máximo dos dados e arquivos de log. Para melhorar o desempenho, controle o número de arquivos de log virtuais (VLFs) pré-alocando o espaço e corrigindo as configurações de crescimento automático (crescimento automático) dos arquivos de log.

Normalmente, um fator de crescimento automático de 8 GB funciona bem na maioria dos ambientes de produção. Considere aumentar os arquivos de log de transações em blocos de 8 GB. Um número maior de VLFs pode estender o tempo de backup e recuperação do banco de dados e causar problemas de desempenho em qualquer operação (por exemplo, replicação) que exija a análise dos arquivos de log.

Para obter mais informações sobre o algoritmo de criação e crescimento do VLF, consulte o SQLskills blog.

Verifique as configurações de crescimento automático do banco de dados

Qualquer transação que precise dos dados ou do arquivo de log para crescer incluirá o tempo gasto pela operação de crescimento do arquivo. O arquivo cresce de acordo com o tamanho do incremento definido pela opção FILEGROWTH. Você pode procurar eventos de crescimento de arquivos nos rastreamentos do SQL Server Profiler. Se o crescimento do arquivo demorar muito, você poderá ver tipos de espera como `ASYNC_IO_COMPLETION`, que ocorre quando o processamento de dados é muito lento. Esses tipos de espera não afetam apenas o desempenho, mas também podem resultar em timeouts de transação. Se essa transação bloquear os recursos buscados por outras transações, o timeout causaria graves problemas de bloqueio do servidor.

Por esse motivo, recomendamos que você defina as configurações de crescimento automático com muito cuidado. Lembre-se também de que:

- O crescimento de arquivos é uma das operações mais onerosas no SQL Server.
- O crescimento automático frequente em pequenos pedaços pode levar à fragmentação do disco.
- O crescimento automático frequente dos arquivos de log resulta em um grande número de arquivos de log virtuais (VLFs) e afeta o desempenho, conforme discutido na seção anterior.

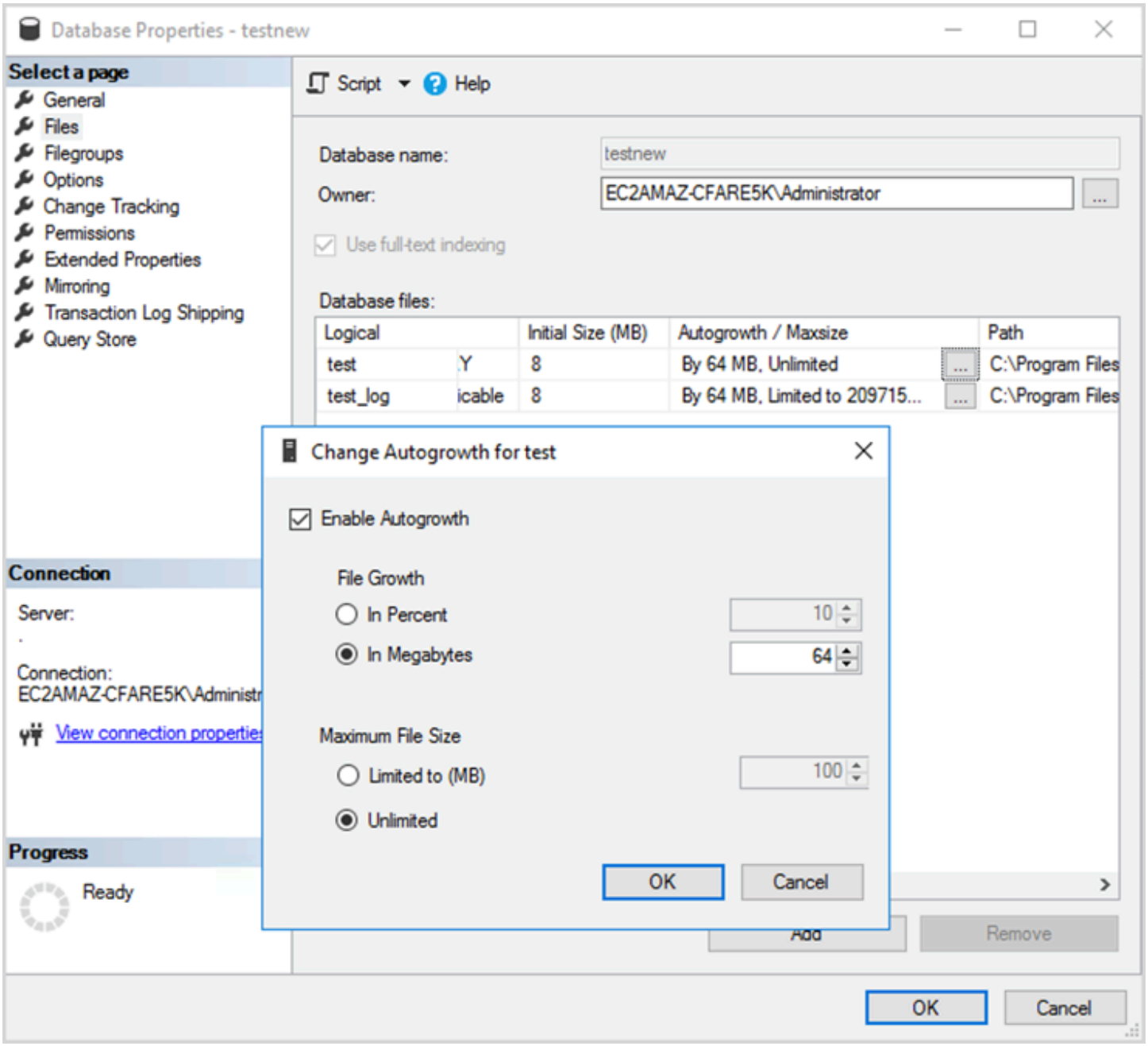
Todos esses motivos podem levar à lentidão no startup do banco de dados e ao aumento do tempo de backup e recuperação.

O ideal é que você aumente previamente os arquivos de forma proativa e com base no monitoramento frequente. Escolha cuidadosamente entre definir o crescimento automático como uma porcentagem ou como um valor estático (em MB). Normalmente, definir o crescimento automático para um oitavo do tamanho do arquivo é um bom ponto de partida, mas essa pode não ser a escolha certa. (Por exemplo, essa porcentagem seria muito alta se seu arquivo de dados tivesse vários TBs tamanhos.)

Na maioria dos casos, um valor de crescimento automático de 1024 MB funciona bem para arquivos de dados na maioria dos grandes bancos de dados. Para arquivos de log, 512 MB é um bom ponto de partida. Para medidas de contingência, é altamente recomendável que você defina o valor do crescimento automático, mas aumente os arquivos manualmente por alguns meses com base nas tendências passadas.

> **Note**
>
> Definir o crescimento automático deve ser uma medida de contingência e, portanto, você deve defini-la depois de pré-alocar o armazenamento em um arquivo.

Você pode alterar as configurações de crescimento automático usando o SQL Server Management Studio (SSMS) ou o Transact-SQL. A ilustração de tela a seguir mostra as configurações de crescimento automático no SSMS.



Ao usar a opção FILEGROWTH para arquivos de dados e log, escolha cuidadosamente entre configurá-la como uma porcentagem ou como um valor estático (em MB). Definir uma porcentagem resulta em um crescimento cada vez maior dos arquivos, então talvez você prefira usar um tamanho estático para controlar melhor a taxa de crescimento.

- Nas versões anteriores ao SQL Server 2022 (16.x), os registros de transações não podem usar a inicialização instantânea de arquivos. Portanto, tempos de crescimento de logs prolongados são especialmente críticos.

- A partir do SQL Server 2022 (16.x, todas as edições), a inicialização instantânea de arquivos pode beneficiar eventos de crescimento do log de transações em até 64 MB. O incremento padrão do tamanho do crescimento automático para novos bancos de dados é de 64 MB. Eventos de crescimento automático do arquivo de log de transações maiores que 64 MB não podem se beneficiar da inicialização instantânea do arquivo.

Configurando grupos de disponibilidade do Always On

Se você estiver usando bibliotecas de cliente nativas para o SQL Server versão 2012 e posterior e bibliotecas.NET Framework 4.5, você pode usar o MultiSubnetFailoverparâmetro para alterar o comportamento da conexão. Recomendamos que você defina o parâmetro como TRUE. Isso permitirá um failover mais rápido com grupos de disponibilidade Always On.

> **Note**
>
> Se você tiver aplicativos herdados que não podem usar o MultiSubnetFailoverparâmetro, você pode colocar um Network Load Balancer na frente de suas instâncias do SQL Server. O balanceador usa uma verificação de integridade que determina qual banco de dados do SQL Server está ativo e envia tráfego para a instância que atualmente hospeda esse banco de dados. O balanceador de carga abrange uma ou várias zonas de disponibilidade. Você pode usar uma porta dedicada, como 59999, para a verificação de integridade e, em seguida, modificar o parâmetro do grupo de clusters para responder a essa porta. Isso permite reduzir o tempo de failover do SQL Server para aproximadamente um minuto sem usar o MultiSubnetFailoverparâmetro. Para obter instruções detalhadas, consulte a postagem do blog Reduza os tempos de failover do SQL Server na EC2 instância da Amazon usando o Network Load Balancer.

Duas configurações afetam a forma como o ouvinte do grupo de disponibilidade é registrado com o DNS: RegisterAllProvidersIP e HostRecord TTL.

Defina RegisterAllProviders o IP como verdadeiro ao usar grupos de disponibilidade Always On

Recomendamos que você defina RegisterAllProviderso IP como 1 (verdadeiro). Quando o ouvinte do grupo de disponibilidade é criado com o RegisterAllProvidersIP definido como1, todos os endereços IP desse ouvinte são registrados no DNS. Quando RegisterAllProviderso IP é definido como `0` (falso), somente um IP ativo é registrado.

Em caso de failover, quando a réplica primária se move de uma sub-rede para outra, o endereço IP antigo não é registrado e o novo endereço IP é registrado. O DNS é atualizado com o novo IP quando o receptor do grupo de disponibilidade fica on-line. No entanto, os sistemas cliente não

resolverão o nome do receptor para o novo endereço IP até que a entrada atualmente em cache expire.

Defina HostRecord TTL como 60 ou menos ao usar grupos de disponibilidade Always On

A configuração HostRecordTTL controla o Time to Live (TTL) para entradas de DNS em cache. O valor padrão é de 1200 segundos. Recomendamos que você altere o HostRecordTTL para uma configuração muito mais baixa (60 segundos ou menos). Isso faz com que o valor em cache expire mais cedo, portanto, em caso de failover, os sistemas clientes podem resolver o novo IP mais rapidamente.

Desativar o failback automático para o grupo de clusters Always On

Verifique se o failback automático está desativado para os grupos de disponibilidade Always On no Gerenciador de Cluster do Windows.

Configuração dos backups

Conforme discutido na seção Otimizar o layout do disco ou a distribuição de arquivos, recomendamos que você envie seus backups nativos do SQL Server para uma unidade separada. Considere também fazer um snapshot programado do volume do EBS em que os arquivos de backup residem.

Melhorias de otimização do banco de dados

Esta seção fornece as melhores práticas para melhorar o desempenho ao trabalhar com o otimizador de consultas do SQL Server. Ela discute como a reconstrução de índices e a atualização regular das estatísticas da tabela podem ajudar a otimizar os planos de execução. As seções a seguir fornecem dicas de configuração e melhores práticas.

Tópicos

- Reconstruir índices
- Atualizar estatísticas

Reconstruir índices

Para que o otimizador de consultas gere os melhores planos de consulta possíveis e use os índices corretos, os índices não devem ser fragmentados. Os índices ficam fragmentados ao longo do tempo com base na taxa de atualização, inserção ou exclusão. Certifique-se de que as tabelas sejam reindexadas regularmente. A frequência de reconstrução depende da taxa na qual o banco de dados manipula as operações da linguagem de manipulação de dados (DML).

Um bom ponto de partida seria reconstruir índices fragmentados em mais de 30% e reorganizar índices fragmentados em menos de 30%. O valor de 30% funciona na maioria dos casos de uso, mas se você ainda vê planos de consulta insatisfatórios devido a índices não utilizados, talvez seja necessário revisitar essa porcentagem.

Use uma consulta como a seguinte para verificar a fragmentação:

```
SELECT OBJECT_NAME(OBJECT_ID), index_id,index_type_desc,index_level,
avg_fragmentation_in_percent,avg_page_space_used_in_percent,page_count
FROM sys.dm_db_index_physical_stats
(DB_ID(N'<your_database>'), NULL, NULL, NULL , 'SAMPLED')
ORDER BY avg_fragmentation_in_percent DESC
```

Recomendamos que você crie um trabalho de manutenção para reconstruir os índices regularmente.

Atualizar estatísticas

Assim como nos índices fragmentados, se o otimizador não tiver up-to-date informações sobre a distribuição dos valores-chave (estatísticas) das colunas da tabela, ele não poderá gerar planos de execução ideais. Recomendamos que você atualize as estatísticas de todas as tabelas regularmente. A frequência das atualizações depende da taxa na qual o banco de dados processa as operações de DML, mas normalmente é executada duas vezes por semana fora do horário de pico. No entanto, evite atualizar as estatísticas nos dias em que você está recriando índices. Para ter mais informações sobre atulização de estatísticas, consulte documentação do SQL Server.

Para otimização do banco de dados, recomendamos o uso de um script de manutenção de índices e estatísticas. Para obter um exemplo, consulte o script de manutenção de índices e estatísticas do SQL Server fornecido no site da Solução de Manutenção do SQL Server.

Otimizando as implantações do SQL Server na Amazon EC2 com AWS o Launch Wizard

AWS O Launch Wizard é o principal método para implantações de instância única e alta disponibilidade (HA) do SQL Server na Amazon EC2. As implantações do Launch Wizard são baseadas no AWS Well-Architected Framework e são otimizadas para segurança, confiabilidade, eficiência de desempenho e economia de custos.

O Launch Wizard simplifica suas implantações do SQL Server e também facilita a configuração do SQL Server. Os atributos incluem:

- Seleção automática de AWS recursos — O Launch Wizard pode recomendar o tipo de instância ideal com base nos requisitos de sua CPU virtual (vCPU), memória e rede. Ele também pode recomendar o tipo de volume com base na unidade de armazenamento e no throughput.
- Monitoramento com um clique — O Launch Wizard se integra ao Amazon CloudWatch Application Insights para configurar o monitoramento para implantações de HA do SQL Server em. AWS Quando você seleciona essa opção, o Application Insights configura automaticamente métricas, registros e alarmes relevantes e começa a monitorar cargas de trabalho recém-implantadas. CloudWatch
- Grupos de recursos de aplicativos para facilitar a descoberta — o Launch Wizard cria um grupo de AWS recursos para todos os recursos que foram criados para seu aplicativo SQL Server. Você pode gerenciar, corrigir e manter seus aplicativos do SQL Server a partir do AWS Systems Manager console.

O Launch Wizard fornece modelos de AWS CloudFormation código reutilizáveis. Esses modelos podem servir como uma linha de base para suas implantações subsequentes de aplicativos. Para saber mais, consulte a visão geral e o guia do usuário do AWS Launch Wizard.

Próximas etapas

Este guia abordou algumas das melhores práticas para configurar e executar cargas de trabalho do Microsoft SQL Server na Amazon. EC2 Seguir essas diretrizes nas fases de planejamento e implementação do processo de migração ajudará você a estabelecer um servidor estável no ambiente de produção.

Para obter mais informações sobre essas tarefas de configuração, consulte os links fornecidos em cada seção e visite as páginas da web listadas na seção Recursos adicionais.

Recursos adicionais

Estratégias, guias e padrões relacionados

- Estratégia de migração para bancos de dados relacionais
- Padrões SQL Server:
  - Todos os padrões
  - Padrões de rehospedagem (migrando o SQL Server para a Amazon) EC2
  - Padrões para redefinir a plataforma (migração do SQL Server para o Amazon RDS para SQL Server)
  - Reformule os padrões (migrando o SQL Server para bancos de dados de código aberto e AWS nativos da nuvem)
- AWS Site de orientação prescritiva

AWS recursos

- AWS documentação
- AWS referência geral
- Glossário da AWS

AWS serviços

- Amazon EBS
- Amazon EC2

Outros recursos

- Como mover o tempdb do Microsoft SQL Server para discos de instância/efêmeros na Amazon EC2
- Discos efêmeros do Stripe Windows no lançamento
- De quanta memória meu SQL Server realmente precisa?
- Estatísticas de espera do SQL Server (ou, me diga onde dói...)
- Tempos limite de conexão no grupo de disponibilidade de várias sub-redes

- Planejar o cache e otimizar para workloads ad hoc
- RAM, memória virtual, arquivo de paginação e gerenciamento de memória no Windows
- Como determinar o tamanho apropriado do arquivo de paginação para versões de 64 bits do Windows

Histórico do documento

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

Alteração	Descrição	Data
Informações corrigidas	Foram corrigidas as informaçõ es sobre o limite de custo da propriedade de paralelis mo. Seu valor é medido por unidades de custo, não por tempo.	4 de dezembro de 2023
Orientação atualizada	Foram atualizadas as seções sobre como definir o tamanho da unidade de alocação de NTFS, bloquear páginas na memória, usar atributos de descarregamento de tarefas, usar striping, alterar o limite de custo do paralelismo, usar sinalizadores de rastreame nto e usar configurações de crescimento automático do banco de dados.	8 de agosto de 2023
Orientação adicionada	Foram adicionadas informaçõ es sobre o uso do Network Load Balancer para aplicativ os legados que não podem usar o MultiSubnetFailove rparâmetro.	11 de novembro de 2022
Código fixo	Foi corrigido o PowerShel l código na seção sobre a	27 de junho de 2022

	inicialização do armazenamento de instâncias.	
Adição de uma nova seção	Foram adicionadas informações sobre AWS Assistente de inicialização para SQL Server.	18 de agosto de 2021
Publicação inicial	—	21 de julho de 2020

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.