



Usando o Apache Iceberg em AWS

AWS Orientação prescritiva



AWS Orientação prescritiva: Usando o Apache Iceberg em AWS

Copyright © 2024 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 desprestige 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
Lagos de dados modernos	2
Casos de uso avançados em lagos de dados modernos	2
Introdução ao Apache Iceberg	3
AWS suporte para Apache Iceberg	4
Introdução às tabelas Iceberg no Athena SQL	6
Criação de uma tabela não particionada	6
Criar uma tabela particionada	7
Criando uma tabela e carregando dados com uma única instrução CTAS	7
Inserindo, atualizando e excluindo dados	8
Consultando tabelas Iceberg	9
Anatomia da mesa de iceberg	9
Trabalhando com a Iceberg no Amazon EMR	12
Compatibilidade de versões e recursos	12
Criação de um cluster do Amazon EMR com o Iceberg	12
Desenvolvendo aplicativos Iceberg no Amazon EMR	13
Usando notebooks Amazon EMR Studio	13
Executando trabalhos do Iceberg no Amazon EMR	14
Melhores práticas para o Amazon EMR	19
Trabalhando com a Iceberg em AWS Glue	21
Usando a integração nativa do Iceberg	21
Usando uma versão personalizada do Iceberg	22
Usando um conector personalizado	22
Trazendo seus próprios arquivos JAR	24
Configurações do Spark para Iceberg em AWS Glue	24
Melhores práticas para AWS Glue empregos	26
Trabalhando com tabelas Iceberg usando o Spark	27
Criando e escrevendo tabelas Iceberg	27
Usando o Spark SQL	27
Usando a DataFrames API	28
Atualização de dados em tabelas Iceberg	29
Atualizando dados em tabelas Iceberg	30
Excluindo dados nas tabelas do Iceberg	30
Leitura de dados	31

Usando a viagem no tempo	31
Usando consultas incrementais	32
Acessando metadados	33
Trabalhando com tabelas Iceberg usando o Athena SQL	34
Compatibilidade de versões e recursos	34
Suporte à especificação da tabela Iceberg	34
Suporte ao recurso Iceberg	34
Trabalhando com mesas Iceberg	35
Migração de tabelas existentes para o Iceberg	36
Migração local	36
Migração completa de dados	41
Selecionar uma estratégia de migração	42
Melhores práticas para otimizar as cargas de trabalho do Iceberg	44
Práticas recomendadas gerais	44
Otimizando o desempenho de leitura	45
Particionamento	46
Ajustando os tamanhos dos arquivos	48
Otimize as estatísticas da coluna	50
Escolha a estratégia de atualização correta	51
Use compressão ZSTD	51
Definir a ordem de classificação	51
Otimizando o desempenho de gravação	54
Definir o modo de distribuição da tabela	54
Escolha a estratégia de atualização correta	54
Escolha o formato de arquivo correto	55
Otimizando o armazenamento	56
Habilite o S3 Intelligent Tiering	56
Arquive ou exclua instantâneos históricos	57
Excluir arquivos órfãos	60
Manutenção de tabelas usando compactação	61
Compactação de iceberg	61
Ajustando o comportamento de compactação	63
Executando a compactação com o Spark no Amazon EMR ou AWS Glue	64
Executando a compactação com o Amazon Athena	65
Recomendações para executar a compactação	65
Usando cargas de trabalho do Iceberg no Amazon S3	67

Evite o particionamento a quente (erros HTTP 503)	67
Use as operações de manutenção do Iceberg para liberar dados não utilizados	67
Replique dados em Regiões da AWS	68
Monitorando cargas de trabalho do Iceberg	70
Monitoramento em nível de tabela	70
Monitoramento em nível de banco de dados	72
Manutenção preventiva	74
Governança e controle de acesso	75
Arquiteturas de referência	76
Ingestão noturna em lote	76
Lago de dados que combina ingestão em lote e quase em tempo real	77
Recursos	78
Colaboradores	79
Histórico do documento	81
Glossário	82
#	82
A	83
B	86
C	88
D	91
E	95
F	97
G	99
H	99
I	101
L	103
M	104
O	108
P	111
Q	114
R	114
S	117
T	121
U	122
V	123
W	123

Z	124
.....	cxxvi

Usando o Apache Iceberg na AWS

Amazon Web Services ([colaboradores](#))

Abril de 2024 ([histórico do documento](#))

O Apache Iceberg é um formato de tabela de código aberto que simplifica o gerenciamento de tabelas e melhora o desempenho. AWS serviços de análise, como Amazon EMR AWS Glue, Amazon Athena e Amazon Redshift, incluem suporte nativo para o Apache Iceberg, para que você possa criar facilmente lagos de dados transacionais com base no Amazon Simple Storage Service (Amazon S3). AWS

Este guia técnico fornece orientação sobre como começar a usar o Apache Iceberg em diferentes Serviços da AWS áreas e inclui as melhores práticas e recomendações para executar o Apache Iceberg em grande escala e, ao mesmo tempo, otimizar o AWS custo e o desempenho.

Este guia se aplica a qualquer pessoa que esteja usando o Apache Iceberg on AWS, desde usuários iniciantes que desejam começar a usar o Apache Iceberg rapidamente até usuários avançados que desejam otimizar e ajustar suas cargas de trabalho existentes do Apache Iceberg. AWS

Neste guia:

- [Lagos de dados modernos](#)
- [Introdução às tabelas Iceberg no Athena SQL](#)
- [Trabalhando com a Iceberg no Amazon EMR](#)
- [Trabalhando com a Iceberg em AWS Glue](#)
- [Trabalhando com tabelas Iceberg usando o Spark](#)
- [Trabalhando com tabelas Iceberg usando o Athena SQL](#)
- [Melhores práticas para otimizar as cargas de trabalho do Iceberg](#)
- [Monitorando cargas de trabalho do Iceberg](#)
- [Governança e controle de acesso](#)
- [Arquiteturas de referência](#)
- [Recursos](#)
- [Colaboradores](#)

Lagos de dados modernos

Casos de uso avançados em lagos de dados modernos

Os data lakes oferecem uma das melhores opções para armazenar dados em termos de custo, escalabilidade e flexibilidade. Você pode usar um data lake para reter grandes volumes de dados estruturados e não estruturados a um custo baixo e usar esses dados para diferentes tipos de cargas de trabalho de análise, desde relatórios de business intelligence até processamento de big data, análises em tempo real, aprendizado de máquina e inteligência artificial generativa (IA), para ajudar a orientar melhores decisões.

Apesar desses benefícios, os data lakes não foram projetados inicialmente com recursos semelhantes aos de bancos de dados. Um data lake não fornece suporte à semântica de processamento de atomicidade, consistência, isolamento e durabilidade (ACID), que você pode precisar para otimizar e gerenciar com eficiência seus dados em grande escala para centenas ou milhares de usuários usando uma infinidade de tecnologias diferentes. Os data lakes não fornecem suporte nativo para as seguintes funcionalidades:

- Realizar atualizações e exclusões eficientes em nível de registro à medida que os dados mudam em sua empresa
- Gerenciando o desempenho das consultas à medida que as tabelas crescem para milhões de arquivos e centenas de milhares de partições
- Garantindo a consistência dos dados entre vários escritores e leitores simultâneos
- Prevenir a corrupção de dados quando as operações de gravação falham no meio da operação
- Esquemas de tabela em evolução ao longo do tempo sem reescrever (parcialmente) conjuntos de dados

Esses desafios se tornaram particularmente predominantes em casos de uso, como lidar com captura de dados alterados (CDC) ou casos de uso relacionados à privacidade, exclusão de dados e ingestão de dados por streaming, o que pode resultar em tabelas abaixo do ideal.

Os data lakes que usam as tabelas tradicionais no formato Hive oferecem suporte a operações de gravação somente para arquivos inteiros. Isso torna as atualizações e exclusões difíceis de implementar, demoradas e caras. Além disso, controles e garantias de concorrência oferecidos em sistemas compatíveis com ACID são necessários para garantir a integridade e a consistência dos dados.

Para ajudar a superar esses desafios, o Apache Iceberg fornece uma funcionalidade adicional semelhante a um banco de dados que simplifica a sobrecarga de otimização e gerenciamento de lagos de dados, ao mesmo tempo em que oferece suporte ao armazenamento em sistemas econômicos, como o Amazon Simple Storage Service ([Amazon S3](#)).

Introdução ao Apache Iceberg

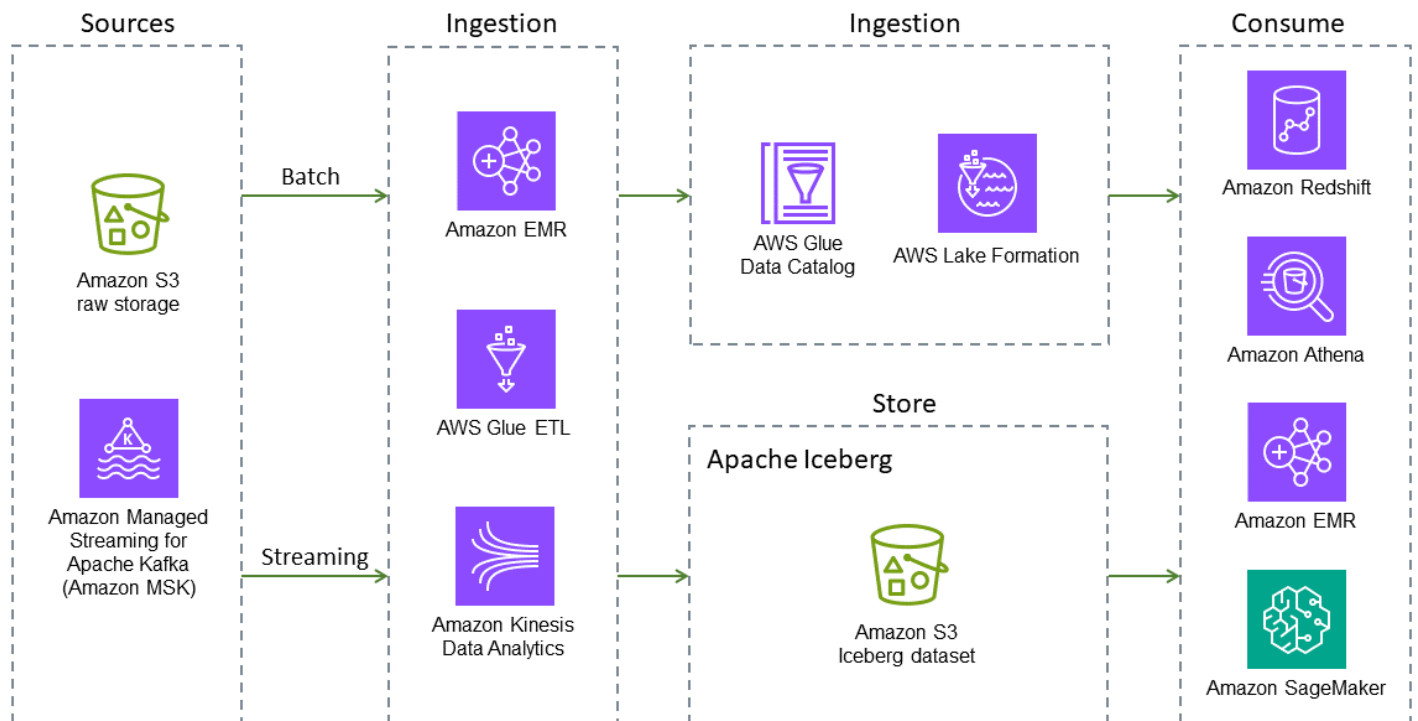
O Apache Iceberg é um formato de tabela de código aberto que fornece recursos em tabelas de data lake que, historicamente, estavam disponíveis apenas em bancos de dados ou data warehouses. Ele foi projetado para oferecer escala e desempenho e é adequado para gerenciar tabelas com mais de centenas de gigabytes. Algumas das principais características das mesas Iceberg são:

- Excluir, atualizar e mesclar. O Iceberg suporta comandos SQL padrão para armazenamento de dados para uso com tabelas de data lake.
- Planejamento rápido de escaneamento e filtragem avançada. O Iceberg armazena metadados, como estatísticas em nível de partição e coluna, que podem ser usados pelos mecanismos para acelerar o planejamento e a execução de consultas.
- Evolução completa do esquema. O Iceberg suporta adicionar, descartar, atualizar ou renomear colunas sem efeitos colaterais.
- Evolução da partição. Você pode atualizar o layout da partição de uma tabela à medida que o volume de dados ou os padrões de consulta mudam. O Iceberg suporta a alteração das colunas nas quais uma tabela é particionada, a adição ou remoção de colunas de partições compostas.
- Particionamento oculto. Esse recurso impede a leitura automática de partições desnecessárias. Isso elimina a necessidade de os usuários entenderem os detalhes de particionamento da tabela ou adicionarem filtros extras às consultas.
- Reversão da versão. Os usuários podem corrigir problemas rapidamente voltando ao estado anterior à transação.
- Viagem no tempo. Os usuários podem consultar uma versão anterior específica de uma tabela.
- Isolamento serializável. As alterações na tabela são atômicas, portanto, os leitores nunca veem alterações parciais ou não confirmadas.
- Escritores simultâneos. O Iceberg usa simultaneidade otimista para permitir que várias transações sejam bem-sucedidas. Em caso de conflito, um dos redatores precisa repetir a transação.
- Formatos de arquivo abertos. [O Iceberg suporta vários formatos de arquivo de código aberto, incluindo Apache Parquet, Apache Avro e Apache ORC.](#)

Em resumo, os data lakes que usam o formato Iceberg se beneficiam da consistência transacional, da velocidade, da escala e da evolução do esquema. Para obter mais informações sobre esses e outros recursos do Iceberg, consulte a documentação do [Apache Iceberg](#).

AWS suporte para Apache Iceberg

[O Apache Iceberg é suportado por estruturas populares de processamento de dados de código aberto, como Serviços da AWS Amazon EMR, Amazon Athena, Amazon Redshift e AWS Glue](#) O diagrama a seguir mostra uma arquitetura de referência simplificada de um data lake baseado no Iceberg.



Os itens a seguir Serviços da AWS fornecem integrações nativas do Iceberg. Existem outros Serviços da AWS que podem interagir com o Iceberg, indiretamente ou empacotando as bibliotecas do Iceberg.

- O Amazon S3 é o melhor lugar para criar lagos de dados devido à sua durabilidade, disponibilidade, escalabilidade, segurança, conformidade e recursos de auditoria. [O Iceberg foi projetado e construído para interagir perfeitamente com o Amazon S3 e fornece suporte para muitos recursos do Amazon S3, conforme listado na documentação do Iceberg.](#)
- O Amazon EMR é uma solução de big data para processamento de dados em escala de petabytes, análise interativa e aprendizado de máquina usando estruturas de código aberto,

como Apache Spark, Flink, Trino e Hive. O Amazon EMR pode ser executado em clusters personalizados do Amazon Elastic Compute Cloud (Amazon EC2), no Amazon Elastic Kubernetes Service (Amazon EKS) ou no Amazon EMR Serverless. AWS Outposts

- O Amazon Athena é um serviço de análise interativo e sem servidor baseado em estruturas de código aberto. Ele suporta formatos de tabela aberta e de arquivo e fornece uma maneira simplificada e flexível de analisar petabytes de dados onde eles estão. O Athena fornece suporte nativo para consultas de leitura, viagem no tempo, gravação e DDL para o Iceberg e usa o AWS Glue Data Catalog metastore for Iceberg.
- O Amazon Redshift é um data warehouse em nuvem na escala de petabytes que oferece suporte a opções de implantação baseadas em cluster e sem servidor. O Amazon Redshift Spectrum pode consultar tabelas externas registradas e armazenadas no Amazon S3. AWS Glue Data Catalog O Redshift Spectrum também fornece suporte para o formato de armazenamento Iceberg.
- AWS Glue é um serviço de integração de dados sem servidor que facilita a descoberta, a preparação, a movimentação e a integração de dados de várias fontes para análise, aprendizado de máquina (ML) e desenvolvimento de aplicativos. AWS Glue As versões 3.0 e posteriores oferecem suporte à estrutura Iceberg para lagos de dados. Você pode usar AWS Glue para realizar operações de leitura e gravação em tabelas Iceberg no Amazon S3 ou trabalhar com tabelas Iceberg usando o. AWS Glue Data Catalog Operações adicionais, como inserção, atualização, consultas do Spark e gravações do Spark, também são suportadas.
- AWS Glue Data Catalog fornece um serviço de catálogo de dados compatível com o Hive metastore que suporta tabelas Iceberg.
- Crawler do AWS Glue fornece automações para registrar tabelas Iceberg no. AWS Glue Data Catalog
- A Amazon SageMaker suporta o armazenamento de conjuntos de recursos na Amazon SageMaker Feature Store usando o formato Iceberg.
- AWS Lake Formation fornece permissões de controle de acesso gerais e refinadas para acessar dados, incluindo tabelas Iceberg consumidas pelo Athena ou pelo Amazon Redshift. Para saber mais sobre o suporte de permissões para tabelas Iceberg, consulte a [documentação do Lake Formation](#).

AWS tem uma ampla variedade de serviços que oferecem suporte ao Iceberg, mas cobrir todos esses serviços está além do escopo deste guia. As seções a seguir abordam o Spark (streaming estruturado e em lote) no Amazon EMR AWS Glue e também no Amazon Athena SQL. A seção a seguir fornece uma visão rápida do suporte do Iceberg no Athena SQL.

Introdução às tabelas Apache Iceberg no Amazon Athena SQL

O Amazon Athena fornece suporte integrado para o Apache Iceberg. Você pode usar o Iceberg sem nenhuma etapa ou configuração adicional, exceto para definir os pré-requisitos de serviço detalhados na seção [Introdução](#) da documentação do Athena. Esta seção fornece uma breve introdução à criação de tabelas no Athena. Para obter mais informações, consulte [Trabalhando com tabelas do Apache Iceberg usando o Athena SQL mais adiante neste guia](#).

Você pode criar tabelas Iceberg AWS usando mecanismos diferentes. Essas tabelas funcionam perfeitamente em todos os Serviços da AWS. Para criar suas primeiras tabelas Iceberg com o Athena SQL, você pode usar o seguinte código padronizado.

```
CREATE TABLE <table_name> (  
    col_1 string,  
    col_2 string,  
    col_3 bigint,  
    col_ts timestamp)  
PARTITIONED BY (col_1, <<<partition_transform>>(col_ts))  
LOCATION 's3://<bucket>/<folder>/<table_name>/'  
TBLPROPERTIES (  
    'table_type' = 'ICEBERG'  
)
```

As seções a seguir fornecem exemplos de criação de tabelas Iceberg particionadas e não particionadas no Athena. [Para obter mais informações, consulte a sintaxe do Iceberg detalhada na documentação do Athena](#).

Criação de uma tabela não particionada

O exemplo de instrução a seguir personaliza o código SQL padronizado para criar uma tabela Iceberg não particionada no Athena. Você pode adicionar essa instrução ao editor de consultas no [console do Athena para](#) criar a tabela.

```
CREATE TABLE athena_iceberg_table (  
    color string,  
    date string,  
    name string,
```

```
price bigint,  
product string,  
ts timestamp)  
LOCATION 's3://DOC_EXAMPLE_BUCKET/ice_warehouse/iceberg_db/athena_iceberg_table/'  
TBLPROPERTIES (  
  'table_type' = 'ICEBERG'  
)
```

Para step-by-step obter instruções sobre como usar o editor de consultas, consulte [Introdução](#) na documentação do Athena.

Criar uma tabela particionada

A declaração a seguir cria uma tabela particionada com base na data usando o conceito de particionamento [oculto](#) do Iceberg. Ele usa a `day()` transformação para derivar partições diárias, usando o `dd-mm-yyyy` formato, de uma coluna de carimbo de data/hora. O Iceberg não armazena esse valor como uma nova coluna no conjunto de dados. Em vez disso, o valor é derivado dinamicamente quando você grava ou consulta dados.

```
CREATE TABLE athena_iceberg_table_partitioned (  
  color string,  
  date string,  
  name string,  
  price bigint,  
  product string,  
  ts timestamp)  
PARTITIONED BY (day(ts))  
LOCATION 's3://DOC_EXAMPLE_BUCKET/ice_warehouse/iceberg_db/athena_iceberg_table/'  
TBLPROPERTIES (  
  'table_type' = 'ICEBERG'  
)
```

Criando uma tabela e carregando dados com uma única instrução CTAS

Nos exemplos particionados e não particionados das seções anteriores, as tabelas Iceberg são criadas como tabelas vazias. Você pode carregar dados nas tabelas usando a `MERGE` instrução `INSERT` or. Como alternativa, você pode usar uma `CREATE TABLE AS SELECT` (CTAS) instrução para criar e carregar dados em uma tabela Iceberg em uma única etapa.

O CTAS é a melhor maneira no Athena de criar uma tabela e carregar dados em uma única instrução. O exemplo a seguir ilustra como usar o CTAS para criar uma tabela Iceberg () a partir de uma tabela Hive/Parquet existente (iceberg_ctas_table) no Athena. hive_table

```
CREATE TABLE iceberg_ctas_table WITH (  
  table_type = 'ICEBERG',  
  is_external = false,  
  location = 's3://DOC_EXAMPLE_BUCKET/ice_warehouse/iceberg_db/iceberg_ctas_table/'  
) AS  
SELECT * FROM "iceberg_db"."hive_table" limit 20  
---  
SELECT * FROM "iceberg_db"."iceberg_ctas_table" limit 20
```

Para saber mais sobre o CTAS, consulte a documentação do [Athena CTAS](#).

Inserindo, atualizando e excluindo dados

O Athena oferece suporte a diferentes formas de gravar dados em uma tabela Iceberg usando as INSERT INTO instruções, UPDATEMERGE INTO, e M. DELETE FRO

Nota:UPDATE,MERGE INTO, e DELETE FROM use a merge-on-read abordagem com exclusões posicionais. Atualmente, a copy-on-write abordagem não é compatível com o Athena SQL.

Por exemplo, a declaração a seguir é usada INSERT INTO para adicionar dados a uma tabela Iceberg:

```
INSERT INTO "iceberg_db"."ice_table" VALUES (  
  'red', '222022-07-19T03:47:29', 'PersonNew', 178, 'Tuna', now()  
)  
  
SELECT * FROM "iceberg_db"."ice_table"  
where color = 'red' limit 10;
```

Exemplo de resultado:

#	color	date	name	price	product	ts
1	red	222022-07-19T03:47:29	PersonNew	178	Tuna	2023-10-11 11:35:01.298000 UTC

Para obter mais informações, consulte a documentação do [Athena](#).

Consultando tabelas Iceberg

Você pode executar consultas SQL regulares em suas tabelas do Iceberg usando o Athena SQL, conforme ilustrado no exemplo anterior.

Além das consultas usuais, o Athena também oferece suporte a consultas de viagem no tempo para tabelas Iceberg. Conforme discutido anteriormente, você pode alterar os registros existentes por meio de atualizações ou exclusões em uma tabela do Iceberg, portanto, é conveniente usar consultas de viagem no tempo para examinar as versões mais antigas da sua tabela com base em um carimbo de data/hora ou em um ID de instantâneo.

Por exemplo, a instrução a seguir atualiza um valor de cor para ePerson5, em seguida, exibe um valor anterior de 4 de janeiro de 2023:

```
UPDATE ice_table SET color='new_color' WHERE name='Person5'  
  
SELECT * FROM "iceberg_db"."ice_table" FOR TIMESTAMP AS OF TIMESTAMP '2023-01-04  
12:00:00 UTC'
```

Exemplo de resultado:

#	color	date	name	price	product	ts
1	cyan	222022-07-19T03:47:29	Person5	353	Keyboard	2023-01-03 10:15:52.268000 UTC
2	lime	222022-07-19T03:47:29	Person1	833	Towels	2023-01-03 10:15:52.268000 UTC
3	turquoise	222022-07-19T03:47:29	Person1	1319	Shirt	2023-01-03 10:15:52.268000 UTC
4	blue	222022-07-19T03:47:29	Person3	163	Sausages	2023-01-03 10:15:52.268000 UTC

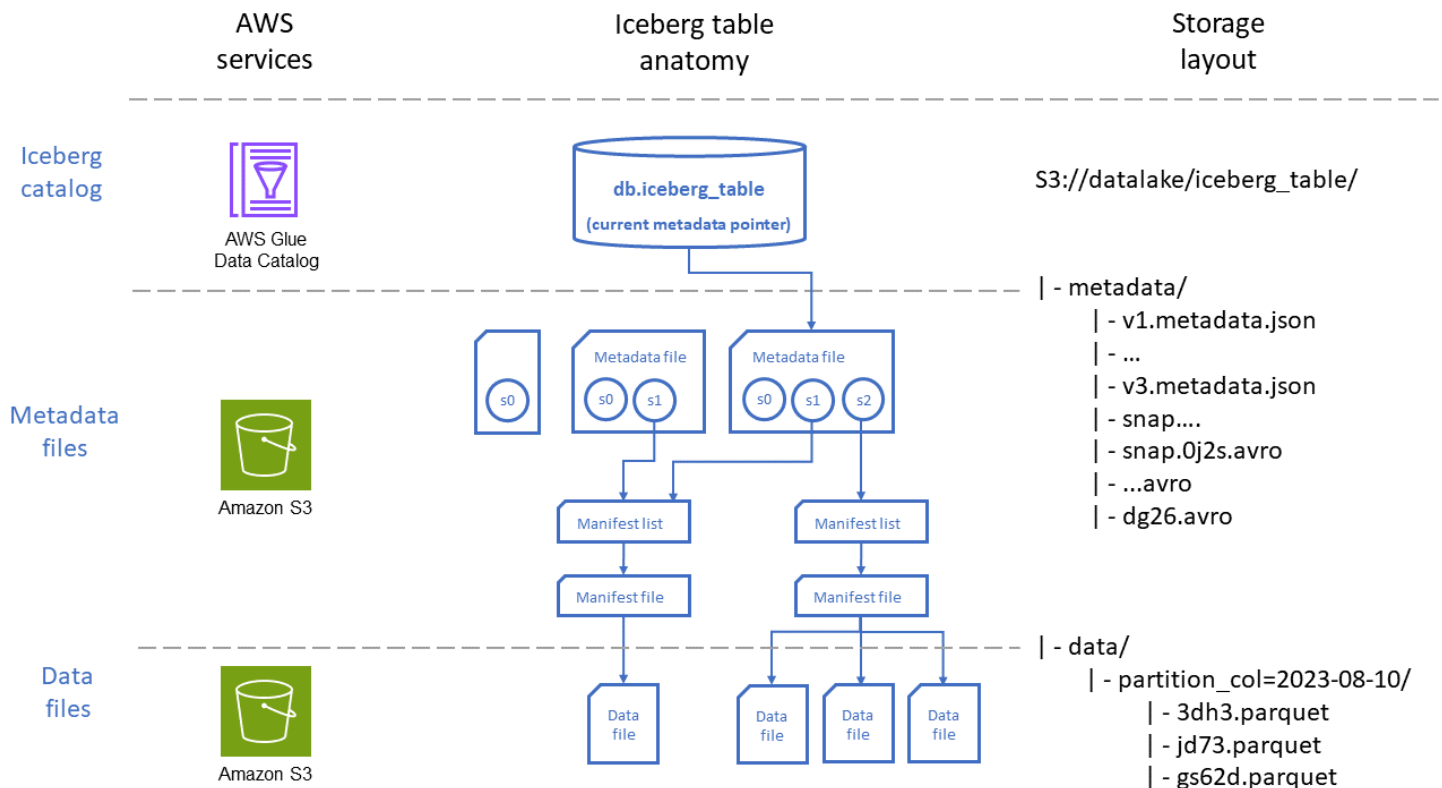
[Para ver exemplos adicionais e de sintaxe de viagens no tempo, consulte a documentação do Athena.](#)

Anatomia da mesa de iceberg

Agora que abordamos as etapas básicas do trabalho com mesas Iceberg, vamos nos aprofundar nos detalhes intrincados e no design de uma mesa Iceberg.

Para habilitar os recursos [descritos anteriormente](#) neste guia, o Iceberg foi projetado com camadas hierárquicas de dados e arquivos de metadados. Essas camadas gerenciam os metadados de forma inteligente para otimizar o planejamento e a execução de consultas.

O diagrama a seguir retrata a organização de uma tabela Iceberg por meio de duas perspectivas: a Serviços da AWS usada para armazenar a tabela e a colocação do arquivo no Amazon S3.



Conforme mostrado no diagrama, uma tabela Iceberg consiste em três camadas principais:

- **Catálogo Iceberg:** AWS Glue Data Catalog integra-se nativamente ao Iceberg e é, para a maioria dos casos de uso, a melhor opção para cargas de trabalho executadas em AWS. Os serviços que interagem com as tabelas do Iceberg (por exemplo, Athena) usam o catálogo para encontrar a versão instantânea atual da tabela, seja para ler ou gravar dados.
- **Camada de metadados:** os arquivos de metadados, ou seja, os arquivos de manifesto e os arquivos de lista de manifestos, controlam informações como o esquema das tabelas, a estratégia de partição e a localização dos arquivos de dados, bem como estatísticas em nível de coluna, como intervalos mínimos e máximos dos registros armazenados em cada arquivo de dados. Esses arquivos de metadados são armazenados no Amazon S3 dentro do caminho da tabela.
 - Os arquivos de manifesto contêm um registro para cada arquivo de dados, incluindo sua localização, formato, tamanho, soma de verificação e outras informações relevantes.
 - As listas de manifestos fornecem um índice dos arquivos de manifesto. À medida que o número de arquivos de manifesto aumenta em uma tabela, dividir essas informações em subseções

menores ajuda a reduzir o número de arquivos de manifesto que precisam ser verificados por consultas.

- Os arquivos de metadados contêm informações sobre toda a tabela Iceberg, incluindo as listas de manifestos, os esquemas, os metadados da partição, os arquivos de instantâneos e outros arquivos usados para gerenciar os metadados da tabela.
- Camada de dados: essa camada contém os arquivos que têm os registros de dados nos quais as consultas serão executadas. [Esses arquivos podem ser armazenados em diferentes formatos, incluindo Apache Parquet, ApacheAvro e Apache ORC.](#)
- Os arquivos de dados contêm os registros de dados de uma tabela.
- Os arquivos de exclusão codificam operações de exclusão e atualização em nível de linha em uma tabela Iceberg. O Iceberg tem dois tipos de arquivos de exclusão, conforme descrito na documentação do [Iceberg](#). Esses arquivos são criados por operações usando o merge-on-read modo.

Trabalhando com o Apache Iceberg no Amazon EMR

O Amazon EMR fornece processamento de dados em escala de petabytes, análises interativas e aprendizado de máquina na nuvem usando estruturas de código aberto, como Apache Spark, Apache Hive, Flink e Trino.

Note

Este guia usa o Apache Spark como exemplos.

O Amazon EMR oferece suporte a várias opções de implantação: Amazon EMR no Amazon EC2, Amazon EMR no Amazon EKS, Amazon EMR Serverless e Amazon EMR on. AWS Outposts Para escolher uma opção de implantação para sua carga de trabalho, consulte as perguntas frequentes do [Amazon EMR](#).

Compatibilidade de versões e recursos

A versão 6.5.0 e versões posteriores do Amazon EMR oferecem suporte nativo ao Apache Iceberg. Para obter uma lista das versões compatíveis do Iceberg para cada versão do Amazon EMR, [consulte o histórico de lançamentos do Iceberg na documentação](#) do Amazon EMR. Analise também [as considerações e limitações do uso do Iceberg no Amazon EMR](#) para ver quais recursos do Iceberg são compatíveis com o Amazon EMR em diferentes estruturas.

Recomendamos que você use a versão mais recente do Amazon EMR para se beneficiar da versão mais recente compatível do Iceberg. Os exemplos de código e as configurações nesta seção pressupõem que você esteja usando a versão emr-6.9.0 do Amazon EMR.

Criação de um cluster do Amazon EMR com o Iceberg

[Para criar um cluster do Amazon EMR no Amazon EC2 com o Iceberg instalado, siga as instruções na documentação do Amazon EMR.](#)

Especificamente, seu cluster deve ser configurado com a seguinte classificação:

```
[{
  "Classification": "iceberg-defaults",
  "Properties": {
```

```
        "iceberg.enabled": "true"
    }
}]
```

Você também pode optar por usar o Amazon EMR Serverless ou o Amazon EMR no Amazon EKS como opções de implantação para suas cargas de trabalho do Iceberg, a partir do Amazon EMR 6.6.0.

Desenvolvendo aplicativos Iceberg no Amazon EMR

Para desenvolver o código Spark para seus aplicativos Iceberg, você pode usar o Amazon [EMR Studio](#), que é um ambiente de desenvolvimento integrado (IDE) baseado na web para notebooks Jupyter totalmente gerenciados que são executados em clusters do Amazon EMR.

Usando notebooks Amazon EMR Studio

Você pode desenvolver aplicativos Spark de forma interativa nos notebooks Amazon EMR Studio Workspace e conectar esses notebooks ao seu Amazon EMR nos clusters do Amazon EC2 ou ao Amazon EMR nos endpoints gerenciados do Amazon EKS. Consulte a [Serviço da AWS documentação](#) para obter instruções sobre como configurar um EMR Studio para o Amazon [EMR no Amazon EC2 e o Amazon EMR](#) no Amazon [EKS](#).

Para usar o Iceberg no EMR Studio, siga estas etapas:

1. Inicie um cluster do Amazon EMR com o Iceberg ativado, conforme as instruções em [Use um cluster](#) com o Iceberg instalado.
2. Configure um EMR Studio. Para obter instruções, consulte [Configurar um Amazon EMR Studio](#).
3. Abra um notebook do EMR Studio Workspace e execute o código a seguir como a primeira célula no notebook para configurar sua sessão do Spark para usar o Iceberg:

```
%%configure -f
{
  "conf": {
    "spark.sql.catalog.<catalog_name>": "org.apache.iceberg.spark.SparkCatalog",
    "spark.sql.catalog.<catalog_name>.warehouse": "s3://YOUR-BUCKET-NAME/YOUR-
FOLDER-NAME/",
    "spark.sql.catalog.<catalog_name>.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
    "spark.sql.catalog.<catalog_name>.io-impl":
"org.apache.iceberg.aws.s3.S3FileIO",
```

```
"spark.sql.extensions":  
  "org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions"  
  }  
}
```

onde:

- `<catalog_name>` é o nome do catálogo de sessões do Iceberg Spark. Substitua-o pelo nome do seu catálogo e lembre-se de alterar as referências em todas as configurações associadas a esse catálogo. Em seu código, você deve então se referir às suas tabelas do Iceberg com o nome da tabela totalmente qualificado, incluindo o nome do catálogo da sessão do Spark, da seguinte forma:

```
<catalog_name>.<database_name>.<table_name>
```

- `<catalog_name>.warehouse` aponta para o caminho do Amazon S3 onde você deseja armazenar seus dados e metadados.
 - Para tornar o catálogo um AWS Glue Data Catalog, `<catalog_name>.catalog-impl` defina como `org.apache.iceberg.aws.glue.GlueCatalog`. Essa chave é necessária para apontar para uma classe de implementação para qualquer implementação de catálogo personalizado. A seção de [melhores práticas gerais](#), mais adiante neste guia, descreve os diferentes catálogos compatíveis com o Iceberg.
 - Use `org.apache.iceberg.aws.s3.S3FileIO` como para aproveitar `<catalog_name>.io-impl` as vantagens do upload de várias partes do Amazon S3 para obter alto paralelismo.
4. Agora você pode começar a desenvolver interativamente seu aplicativo Spark para o Iceberg no notebook, como faria com qualquer outro aplicativo Spark.

Para obter mais informações sobre como configurar o Spark para o Apache Iceberg usando o Amazon EMR Studio, consulte a postagem do blog [Crie um data lake evolutivo de alto desempenho, compatível com ACID](#), usando o Apache Iceberg no Amazon EMR.

Executando trabalhos do Iceberg no Amazon EMR

[Depois de desenvolver o código do aplicativo Spark para sua carga de trabalho do Iceberg, você pode executá-lo em qualquer opção de implantação do Amazon EMR que suporte o Iceberg \(consulte as perguntas frequentes do Amazon EMR\).](#)

Assim como em outros trabalhos do Spark, você pode enviar trabalhos para um cluster do Amazon EMR no Amazon EC2 adicionando etapas ou enviando interativamente trabalhos do Spark para o nó principal. Para executar um trabalho do Spark, consulte as seguintes páginas de documentação do Amazon EMR:

- Para obter uma visão geral das diferentes opções de envio de trabalho para um cluster do Amazon EMR no Amazon EC2 e instruções detalhadas para cada opção, [consulte Enviar trabalho para um cluster](#).
- Para o Amazon EMR no Amazon EKS, consulte [Executando trabalhos do Spark](#) com StartJobRun
- [Para o Amazon EMR Serverless, consulte Executando trabalhos](#).

As seções a seguir fornecem um exemplo para cada opção de implantação do Amazon EMR.

Amazon EMR no Amazon EC2

Você pode usar essas etapas para enviar a tarefa do Iceberg Spark:

1. Crie o arquivo `emr_step_iceberg.json` com o seguinte conteúdo em sua estação de trabalho:

```
[{
  "Name": "iceberg-test-job",
  "Type": "spark",
  "ActionOnFailure": "CONTINUE",
  "Args": [
    "--deploy-mode",
    "client",
    "--conf",

    "spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
    "--conf",
    "spark.sql.catalog.<catalog_name>=org.apache.iceberg.spark.SparkCatalog",
    "--conf",
    "spark.sql.catalog.<catalog_name>.catalog-
impl=org.apache.iceberg.aws.glue.GlueCatalog",
    "--conf",
    "spark.sql.catalog.<catalog_name>.warehouse=s3://YOUR-BUCKET-NAME/YOUR-
FOLDER-NAME/",
    "--conf",
    "spark.sql.catalog.<catalog_name>.io-
impl=org.apache.iceberg.aws.s3.S3FileIO",
    "s3://YOUR-BUCKET-NAME/code/iceberg-job.py"
```

```
]
}]
```

2. Modifique o arquivo de configuração para sua tarefa específica do Spark personalizando as opções de configuração do Iceberg destacadas em negrito.
3. Envie a etapa usando o AWS Command Line Interface (AWS CLI). Execute o comando no diretório em que o `emr_step_iceberg.json` arquivo está localizado.

```
aws emr add-steps --cluster-id <cluster_id> --steps file://emr_step_iceberg.json
```

Amazon EMR Serverless

Para enviar uma tarefa do Iceberg Spark para o Amazon EMR Serverless usando: AWS CLI

1. Crie o arquivo `emr_serverless_iceberg.json` com o seguinte conteúdo em sua estação de trabalho:

```
{
  "applicationId": "<APPLICATION_ID>",
  "executionRoleArn": "<ROLE_ARN>",
  "jobDriver": {
    "sparkSubmit": {
      "entryPoint": "s3://YOUR-BUCKET-NAME/code/iceberg-job.py",
      "entryPointArguments": [],
      "sparkSubmitParameters": "--jars /usr/share/aws/iceberg/lib/iceberg-
spark3-runtime.jar"
    }
  },
  "configurationOverrides": {
    "applicationConfiguration": [{
      "classification": "spark-defaults",
      "properties": {
        "spark.sql.extensions":
"org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
        "spark.sql.catalog.<catalog_name>":
"org.apache.iceberg.spark.SparkCatalog",
        "spark.sql.catalog.<catalog_name>.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
        "spark.sql.catalog.<catalog_name>.warehouse": "s3://YOUR-BUCKET-NAME/
YOUR-FOLDER-NAME/",

```

```

    "spark.sql.catalog.<catalog_name>.io-impl":
    "org.apache.iceberg.aws.s3.S3FileIO",
    "spark.jars": "/usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar",

    "spark.hadoop.hive.metastore.client.factory.class": "com.amazonaws.glue.catalog.metastore.AWSGlueCatalogMetastoreClientFactory",
  }
}],
  "monitoringConfiguration": {
    "s3MonitoringConfiguration": {
      "logUri": "s3://YOUR-BUCKET-NAME/emr-serverless/logs/"
    }
  }
}
}

```

2. Modifique o arquivo de configuração para sua tarefa específica do Spark personalizando as opções de configuração do Iceberg destacadas em negrito.
3. Envie o trabalho usando AWS CLI o. Execute o comando no diretório em que o `emr_serverless_iceberg.json` arquivo está localizado:

```
aws emr-serverless start-job-run --cli-input-json file://emr_serverless_iceberg.json
```

Para enviar uma tarefa do Iceberg Spark para o Amazon EMR Serverless usando o console do EMR Studio:

1. Siga as instruções na documentação do [Amazon EMR Serverless](#).
2. Para a configuração do Job, use a configuração do Iceberg para o Spark fornecida para o AWS CLI e personalize os campos destacados do Iceberg. Para obter instruções detalhadas, consulte [Usando o Apache Iceberg com o EMR Serverless na documentação do Amazon EMR](#).

Amazon EMR no Amazon EKS

Para enviar uma tarefa do Iceberg Spark para o Amazon EMR no Amazon EKS usando: AWS CLI

1. Crie o arquivo `emr_eks_iceberg.json` com o seguinte conteúdo em sua estação de trabalho:

```

{
  "name": "iceberg-test-job",
  "virtualClusterId": "<VIRTUAL_CLUSTER_ID>",

```

```

"executionRoleArn": "<ROLE_ARN>",
"releaseLabel": "emr-6.9.0-latest",
"jobDriver": {
  "sparkSubmitJobDriver": {
    "entryPoint": "s3://YOUR-BUCKET-NAME/code/iceberg-job.py",
    "entryPointArguments": [],
    "sparkSubmitParameters": "--jars local:///usr/share/aws/iceberg/lib/
iceberg-spark3-runtime.jar"
  }
},
"configurationOverrides": {
  "applicationConfiguration": [{
    "classification": "spark-defaults",
    "properties": {
      "spark.sql.extensions":
"org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
      "spark.sql.catalog.<catalog_name>":
"org.apache.iceberg.spark.SparkCatalog",
      "spark.sql.catalog.<catalog_name>.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
      "spark.sql.catalog.<catalog_name>.warehouse": "s3://YOUR-BUCKET-NAME/
YOUR-FOLDER-NAME/",
      "spark.sql.catalog.<catalog_name>.io-impl":
"org.apache.iceberg.aws.s3.S3FileIO",
      "spark.hadoop.hive.metastore.client.factory.class":
      "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory"
    }
  }],
  "monitoringConfiguration": {
    "persistentAppUI": "ENABLED",
    "s3MonitoringConfiguration": {
      "logUri": "s3://YOUR-BUCKET-NAME/emr-serverless/logs/"
    }
  }
}
}

```

2. Modifique o arquivo de configuração da sua tarefa do Spark personalizando as opções de configuração do Iceberg destacadas em negrito.
3. Envie o trabalho usando AWS CLI o. Execute o comando a seguir no diretório em que o `emr_eks_iceberg.json` arquivo está localizado:


```
aws emr-containers start-job-run --cli-input-json file://emr_eks_iceberg.json
```

Para obter instruções detalhadas, consulte Como [usar o Apache Iceberg com o Amazon EMR no EKS na documentação do Amazon EMR no EKS](#).

Melhores práticas para o Amazon EMR

Esta seção fornece diretrizes gerais para ajustar trabalhos do Spark no Amazon EMR para otimizar a leitura e gravação de dados nas tabelas do Iceberg. Para conhecer as melhores práticas específicas do Iceberg, consulte a seção [Melhores práticas](#) mais adiante neste guia.

- Use a versão mais recente do Amazon EMR — O Amazon EMR fornece otimizações do Spark prontas para usar com o tempo de execução do Amazon EMR Spark. AWS melhora o desempenho do mecanismo de tempo de execução do Spark a cada nova versão.
- Determine a infraestrutura ideal para suas cargas de trabalho do Spark — As cargas de trabalho do Spark podem exigir diferentes tipos de hardware para diferentes características de trabalho para garantir o desempenho ideal. O Amazon EMR [oferece suporte a vários tipos de instância](#) (como otimizado para computação, otimizado para memória, uso geral e otimizado para armazenamento) para cobrir todos os tipos de requisitos de processamento. Ao integrar novas cargas de trabalho, recomendamos que você faça o benchmark com tipos gerais de instância, como M5 ou M6g. Monitore o sistema operacional (OS) e as métricas do YARN da Ganglia e da Amazon CloudWatch para determinar os gargalos do sistema (CPU, memória, armazenamento e E/S) no pico de carga e escolha o hardware apropriado.
- Ajuste `spark.sql.shuffle.partitions` — defina a `spark.sql.shuffle.partitions` propriedade como o número total de núcleos virtuais (vCores) em seu cluster ou como um múltiplo desse valor (normalmente, 1 a 2 vezes o número total de vCores). Essa configuração afeta o paralelismo do Spark quando você usa particionamento de hash e intervalo como modo de distribuição de gravação. Ele solicita uma reprodução aleatória antes de gravar para organizar os dados, o que garante o alinhamento das partições.
- Ative a escalabilidade gerenciada — Para quase todos os casos de uso, recomendamos que você habilite a escalabilidade gerenciada e a alocação dinâmica. No entanto, se você tiver uma carga de trabalho com um padrão previsível, sugerimos que você desative o escalonamento automático e a alocação dinâmica. Quando a escalabilidade gerenciada estiver habilitada, recomendamos que você use instâncias spot para reduzir custos. Use instâncias spot para nós de tarefas em vez de

nós principais ou principais. Ao usar instâncias spot, use frotas de instâncias com vários tipos de instância por frota para garantir a disponibilidade spot.

- Use a junção de transmissão quando possível — A junção de transmissão (no lado do mapa) é a melhor junção, desde que uma de suas tabelas seja pequena o suficiente para caber na memória do seu menor nó (na ordem de MBs) e você esteja realizando uma junção equi (=). Todos os tipos de junção, exceto as junções externas completas, são suportados. Uma junção de transmissão transmite a tabela menor como uma tabela de hash em todos os nós de trabalho na memória. Depois que a pequena tabela for transmitida, você não poderá fazer alterações nela. Como a tabela de hash está localmente na máquina virtual Java (JVM), ela pode ser mesclada facilmente com a tabela grande com base na condição de junção usando uma junção de hash. As junções de transmissão oferecem alto desempenho devido à sobrecarga mínima de reprodução aleatória.
- Ajuste o coletor de lixo — Se os ciclos de coleta de lixo (GC) forem lentos, considere mudar do coletor de lixo paralelo padrão para o G1GC para melhorar o desempenho. Para otimizar o desempenho do GC, você pode ajustar os parâmetros do GC. Para monitorar o desempenho do GC, você pode monitorá-lo usando a interface do usuário do Spark. Idealmente, o tempo de GC deve ser menor ou igual a 1% do tempo de execução total da tarefa.

Trabalhando com o Apache Iceberg em AWS Glue

[AWS Glue](#) é um serviço de integração de dados sem servidor que facilita a descoberta, a preparação, a movimentação e a integração de dados de várias fontes para análise, aprendizado de máquina (ML) e desenvolvimento de aplicativos. Um dos principais recursos do AWS Glue é a capacidade de realizar operações de extração, transformação e carregamento (ETL) de maneira simples e econômica. Isso ajuda a categorizar seus dados, limpá-los, enriquecê-los e movê-los de forma confiável entre vários armazenamentos de dados e fluxos de dados.

[AWS Glue trabalhos](#) encapsulam scripts que definem a lógica de transformação usando um tempo de execução [Apache Spark ou Python](#). AWS Glue os trabalhos podem ser executados tanto no modo batch quanto no modo de streaming.

Ao criar trabalhos do Iceberg em AWS Glue, dependendo da versão do AWS Glue, você pode usar a integração nativa do Iceberg ou uma versão personalizada do Iceberg para anexar dependências do Iceberg ao trabalho.

Usando a integração nativa do Iceberg

AWS Glue as versões 3.0 e 4.0 oferecem suporte nativo a formatos de data lake transacionais, como Apache Iceberg, Apache Hudi e Linux Foundation Delta Lake in for Spark. AWS Glue Esse recurso de integração simplifica as etapas de configuração necessárias para começar a usar essas estruturas no. AWS Glue

Para ativar o suporte do Iceberg para seu AWS Glue trabalho, defina o trabalho: escolha a guia Detalhes do trabalho para seu AWS Glue trabalho, vá até Parâmetros do trabalho em Propriedades avançadas e defina a chave `--datalake-formats` e seu valor como `iceberg`.

Se você estiver criando um trabalho usando um notebook, você pode configurar o parâmetro na primeira célula do notebook usando a `%%configure` mágica da seguinte forma:

```
%%configure
{
  "--conf" : <job-specific Spark configuration discussed later>,
  "--datalake-formats" : "iceberg"
}
```

Usando uma versão personalizada do Iceberg

Em algumas situações, talvez você queira manter o controle sobre a versão do Iceberg para o trabalho e atualizá-la no seu próprio ritmo. Por exemplo, a atualização para uma versão posterior pode desbloquear o acesso a novos recursos e aprimoramentos de desempenho. Para usar uma versão específica do Iceberg com AWS Glue, você pode usar um conector personalizado ou seu próprio arquivo JAR.

Usando um conector personalizado

AWS Glue suporta conectores, que são pacotes de código opcionais que auxiliam no acesso aos armazenamentos de dados em AWS Glue Studio. Você pode [assinar um conector](#) em AWS Marketplace, ou você pode criar um conector personalizado.

Note

AWS Marketplace oferece o [conector Apache Iceberg](#) para. AWS Glue No entanto, recomendamos que você use um conector personalizado para manter o controle sobre as versões do Iceberg.

Por exemplo, para criar um conector de cliente para a versão 0.13.1 do Iceberg, siga estas etapas:

1. Faça o upload dos arquivos `iceberg-spark-runtime-3.1_2.12-0.13.1.jar`, `bundle-2.17.161.jar` e `url-connection-client-2.17.161.jar` para um bucket do Amazon S3. Você pode baixar esses arquivos de seus respectivos repositórios Apache Maven.
2. No [AWS Glue Studio console](#), crie um conector Spark personalizado:
 - a. No painel de navegação, escolha Conexões de dados. (Se você estiver usando a navegação mais antiga, escolha Conectores, Criar conector personalizado.)
 - b. Na caixa Conectores, escolha Criar conector personalizado.
 - c. Na página Criar conector personalizado:
 - Especifique o caminho para os arquivos JAR no Amazon S3.
 - Insira um nome para o conector.
 - Escolha Spark como o tipo de conector.
 - Em Nome da classe, especifique o nome da classe da fonte de dados totalmente qualificado (ou seu alias) que você usa ao carregar a fonte de dados do Spark com o `format` operador.

- (Opcional) Forneça uma descrição do conector.

3. Escolha Criar conector.

Ao trabalhar com conectores AWS Glue, você deve criar uma conexão para o conector. Uma conexão contém as propriedades necessárias para se conectar a um armazenamento de dados específico. Você usa a conexão com suas origens e destinos de dados no trabalho de ETL. Conectores e conexões trabalham juntos para facilitar o acesso aos armazenamentos de dados.

Para criar uma conexão usando o conector Iceberg personalizado que você criou:

1. No [AWS Glue Studio console](#), selecione seu conector Iceberg personalizado.
2. Siga as instruções para fornecer os detalhes, como sua VPC e outras configurações de rede exigidas pelo trabalho, e escolha Criar conexão.

Agora você pode usar a conexão em sua tarefa de AWS Glue ETL. Dependendo de como você cria o trabalho, há diferentes maneiras de vincular a conexão ao seu trabalho:

- Se você criar uma tarefa visual usando AWS Glue Studio, poderá selecionar a conexão na lista Conexão na guia Propriedades da fonte de dados — Conector.
- Se você desenvolver o trabalho em um notebook, use a `%connections` mágica para definir o nome da conexão:

```
%glue_version 3.0

%connections <name-of-the iceberg-connection>

%%configure
{
  "--conf" : "job-specific Spark configurations, to be discussed later",
  "--datalake-formats" : "iceberg"
}
```

- Se você criar o trabalho usando o editor de scripts, especifique a conexão na guia Detalhes do trabalho, em Propriedades avançadas, Conexões de rede adicionais.

Para obter mais informações sobre os procedimentos desta seção, consulte [Usando conectores e conexões AWS Glue Studio](#) na AWS Glue documentação.

Trazendo seus próprios arquivos JAR

Dentro AWS Glue, você também pode trabalhar com o Iceberg sem precisar usar um conector. Essa abordagem é útil quando você deseja manter o controle sobre a versão do Iceberg e atualizá-la rapidamente. Para usar essa opção, faça o upload dos arquivos JAR do Iceberg necessários em um bucket do S3 de sua escolha e faça referência aos arquivos em seu AWS Glue trabalho. Por exemplo, se você estiver trabalhando com o Iceberg 1.0.0, os arquivos JAR necessários são `iceberg-spark-runtime-3.0_2.12-1.0.0.jar`, `url-connection-client-2.15.40.jar`, e `bundle-2.15.40.jar`. Você também pode priorizar os arquivos JAR adicionais no caminho da classe definindo o `--user-jars-first` parâmetro como `true` para o trabalho.

Configurações do Spark para Iceberg em AWS Glue

Esta seção discute as configurações do Spark necessárias para criar uma tarefa de AWS Glue ETL para um conjunto de dados do Iceberg. Você pode definir essas configurações usando a chave `--conf` Spark com uma lista separada por vírgulas de todas as chaves e valores de configuração do Spark. Você pode usar a `%%configure` mágica em um notebook ou na seção Job parameters do AWS Glue Studio console.

```
%glue_version 3.0

%connections <name-of-the iceberg-connection>

%%configure
{
  "--conf" : "spark.sql.extensions=org.apache.iceberg.spark.extensions...",
  "--datalake-formats" : "iceberg"
}
```

Configure a sessão do Spark com as seguintes propriedades:

- `<catalog_name>` é o nome do catálogo de sessões do Iceberg Spark. Substitua-o pelo nome do seu catálogo e lembre-se de alterar as referências em todas as configurações associadas a esse catálogo. No seu código, você deve então se referir às suas tabelas Iceberg com o nome da tabela totalmente qualificado, incluindo o nome do catálogo da sessão do Spark, do seguinte modo:
`<catalog_name>.<database_name>.<table_name>`
- `<catalog_name>.<warehouse>` aponta para o caminho do Amazon S3 onde você deseja armazenar seus dados e metadados.

- Para tornar o catálogo um AWS Glue Data Catalog, `<catalog_name>.catalog-impl` defina como `org.apache.iceberg.aws.glue.GlueCatalog`. Essa chave é necessária para apontar para uma classe de implementação para qualquer implementação de catálogo personalizado. Para catálogos compatíveis com o Iceberg, consulte a [???](#) seção de [melhores práticas gerais](#) mais adiante neste guia.
- Use `org.apache.iceberg.aws.s3.S3FileIO` como para aproveitar `<catalog_name>.io-impl` as vantagens do upload de várias partes do Amazon S3 para obter alto paralelismo.

Por exemplo, se você tiver um catálogo chamado `glue_iceberg`, poderá configurar seu trabalho usando várias `--conf` chaves da seguinte forma:

```
%%configure
{
  "--datalake-formats" : "iceberg",
  "--conf" :
  "spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
  "--conf" : "spark.sql.catalog.glue_iceberg=org.apache.iceberg.spark.SparkCatalog",
  "--conf" : "spark.sql.catalog.glue_iceberg.warehouse=s3://<your-warehouse-dir>/",
  "--conf" : " spark.sql.catalog.glue_iceberg.catalog-
impl=org.apache.iceberg.aws.glue.GlueCatalog ",
  "--conf" : " spark.sql.catalog.glue_iceberg.io-
impl=org.apache.iceberg.aws.s3.S3FileIO
}
```

Como alternativa, você pode usar o código para adicionar as configurações acima ao seu script do Spark da seguinte forma:

```
spark = SparkSession.builder\

.config("spark.sql.extensions", "org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensi
    .config("spark.sql.catalog.glue_iceberg",
"org.apache.iceberg.spark.SparkCatalog")\
    .config("spark.sql.catalog.glue_iceberg.warehouse", "s3://<your-
warehouse-dir>/")\
    .config("spark.sql.catalog.glue_iceberg.catalog-impl",
"org.apache.iceberg.aws.glue.GlueCatalog") \
    .config("spark.sql.catalog.glue_iceberg.io-impl",
"org.apache.iceberg.aws.s3.S3FileIO") \
    .getOrCreate()
```

Melhores práticas para AWS Glue empregos

Esta seção fornece diretrizes gerais para ajustar as tarefas do Spark para otimizar AWS Glue a leitura e gravação de dados nas tabelas do Iceberg. Para conhecer as melhores práticas específicas do Iceberg, consulte a seção [Melhores práticas](#) mais adiante neste guia.

- Use a versão mais recente do AWS Glue e atualize sempre que possível — novas versões AWS Glue oferecem melhorias de desempenho, tempos de inicialização reduzidos e novos recursos. Eles também oferecem suporte às versões mais recentes do Spark, que podem ser necessárias para as versões mais recentes do Iceberg. Para ver uma lista das AWS Glue versões disponíveis e das versões do Spark compatíveis com elas, consulte a [AWS Glue documentação](#).
- Otimize a memória do AWS Glue trabalho — siga as recomendações na postagem do AWS blog [Otimize o gerenciamento de memória em AWS Glue](#).
- Use o AWS Glue Auto Scaling — Quando você ativa o Auto Scaling AWS Glue, ajusta automaticamente o número de trabalhadores dinamicamente com base na sua AWS Glue carga de trabalho. Isso ajuda a reduzir o custo do seu AWS Glue trabalho durante picos de carga, porque AWS Glue reduz o número de trabalhadores quando a carga de trabalho é pequena e os trabalhadores estão ociosos. Para usar o AWS Glue Auto Scaling, você especifica um número máximo de trabalhadores para os quais seu AWS Glue trabalho pode ser escalado. Para obter mais informações, consulte [Como usar o escalonamento automático](#) AWS Glue na AWS Glue documentação.
- Use conectores personalizados ou adicione dependências de biblioteca - a integração AWS Glue nativa com o Iceberg é melhor para começar a usar o Iceberg. No entanto, para cargas de trabalho de produção, recomendamos que você use contêineres personalizados ou adicione dependências de biblioteca (conforme discutido [anteriormente neste guia](#)) para ter controle total sobre a versão do Iceberg. Essa abordagem ajuda você a se beneficiar dos recursos mais recentes do Iceberg e das melhorias de desempenho em seus AWS Glue trabalhos.
- Ative a interface do Spark para monitoramento e depuração — Você também pode usar a [interface do Spark AWS Glue para inspecionar sua tarefa do Iceberg visualizando os diferentes estágios de uma tarefa do Spark em](#) um gráfico acíclico direcionado (DAG) e monitorando as tarefas em detalhes. A interface do usuário do Spark fornece uma maneira eficaz de solucionar problemas e otimizar tarefas do Iceberg. Por exemplo, você pode identificar estágios de gargalo que apresentam grandes confusões ou vazamento de disco para identificar oportunidades de ajuste. Para obter mais informações, consulte [Monitoramento de trabalhos usando a interface web do Apache Spark](#) na AWS Glue documentação.

Trabalhando com tabelas do Apache Iceberg usando o Apache Spark

Esta seção fornece uma visão geral do uso do Apache Spark para interagir com tabelas Iceberg. Os exemplos são códigos padronizados que podem ser executados no Amazon EMR ou. AWS Glue

Observação: a interface principal para interagir com as tabelas do Iceberg é o SQL, então a maioria dos exemplos combinará o Spark SQL com a API. DataFrames

Criando e escrevendo tabelas Iceberg

Você pode usar o Spark SQL e o Spark DataFrames para criar e adicionar dados às tabelas do Iceberg.

Usando o Spark SQL

Para escrever um conjunto de dados do Iceberg, use instruções SQL padrão do Spark, como e. CREATE TABLE INSERT INTO

Tabelas não particionadas

Aqui está um exemplo de criação de uma tabela Iceberg não particionada com o Spark SQL:

```
spark.sql(f"""
    CREATE TABLE IF NOT EXISTS {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions (
        c_customer_sk          int,
        c_customer_id          string,
        c_first_name           string,
        c_last_name            string,
        c_birth_country        string,
        c_email_address        string)
    USING iceberg
    OPTIONS ('format-version'='2')
    """)
```

Para inserir dados em uma tabela não particionada, use uma instrução padrão INSERT INTO:

```
spark.sql(f"""
INSERT INTO {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions
```

```
SELECT c_customer_sk, c_customer_id, c_first_name, c_last_name, c_birth_country,  
       c_email_address  
FROM another_table  
""")
```

Tabelas particionadas

Aqui está um exemplo de criação de uma tabela Iceberg particionada com o Spark SQL:

```
spark.sql(f"""  
    CREATE TABLE IF NOT EXISTS {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions (  
        c_customer_sk          int,  
        c_customer_id          string,  
        c_first_name           string,  
        c_last_name            string,  
        c_birth_country        string,  
        c_email_address        string)  
    USING iceberg  
    PARTITIONED BY (c_birth_country)  
    OPTIONS ('format-version'='2')  
""")
```

Para inserir dados em uma tabela Iceberg particionada com o Spark SQL, você executa uma classificação global e, em seguida, grava os dados:

```
spark.sql(f"""  
INSERT INTO {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions  
SELECT c_customer_sk, c_customer_id, c_first_name, c_last_name, c_birth_country,  
       c_email_address  
FROM another_table  
ORDER BY c_birth_country  
""")
```

Usando a DataFrames API

Para escrever um conjunto de dados do Iceberg, você pode usar a `DataFrameWriterV2` API.

Para criar uma tabela Iceberg e gravar dados nela, use a função `df.writeTo(t)`. Se a tabela existir, use a `.append()` função. Se isso não acontecer, use `.create()`. Os exemplos a seguir usam `.createOrReplace()`, que é uma variação do `.create()` que é equivalente `CREATE OR REPLACE TABLE AS SELECT a`.

Tabelas não particionadas

Para criar e preencher uma tabela Iceberg não particionada usando a API: `DataFrameWriterV2`

```
input_data.writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions") \  
    .tableProperty("format-version", "2") \  
    .createOrReplace()
```

Para inserir dados em uma tabela Iceberg não particionada existente usando a API: `DataFrameWriterV2`

```
input_data.writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions") \  
    .append()
```

Tabelas particionadas

Para criar e preencher uma tabela Iceberg particionada usando a `DataFrameWriterV2` API, você pode usar uma classificação local para ingerir dados:

```
input_data.sortWithinPartitions("c_birth_country") \  
    .writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions") \  
    .tableProperty("format-version", "2") \  
    .partitionedBy("c_birth_country") \  
    .createOrReplace()
```

Para inserir dados em uma tabela Iceberg particionada usando a `DataFrameWriterV2` API, você pode usar uma classificação global para ingerir dados:

```
input_data.orderBy("c_birth_country") \  
    .writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions") \  
    .append()
```

Atualização de dados em tabelas Iceberg

O exemplo a seguir mostra como atualizar dados em uma tabela Iceberg. Este exemplo modifica todas as linhas que têm um número par na `c_customer_sk` coluna.

```
spark.sql(f"""  
UPDATE {CATALOG_NAME}.{db.name}.{table.name}  
SET c_email_address = 'even_row'
```

```
WHERE c_customer_sk % 2 == 0
""")
```

Essa operação usa a copy-on-write estratégia padrão e, portanto, reescreve todos os arquivos de dados afetados.

Atualizando dados em tabelas Iceberg

A atualização de dados se refere à inserção de novos registros de dados e à atualização dos registros de dados existentes em uma única transação. Para inserir dados em uma tabela Iceberg, você usa a declaração. SQL `MERGE INTO`

O exemplo a seguir altera o conteúdo da tabela `{UPSERT_TABLE_NAME}` dentro da tabela: `{TABLE_NAME}`

```
spark.sql(f"""
MERGE INTO {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME} t
USING {UPSERT_TABLE_NAME} s
  ON t.c_customer_id = s.c_customer_id
  WHEN MATCHED THEN UPDATE SET t.c_email_address = s.c_email_address
  WHEN NOT MATCHED THEN INSERT *
""")
```

- Se um registro de cliente que está em `{UPSERT_TABLE_NAME}` já existir `{TABLE_NAME}` com o mesmo `c_customer_id`, o `c_email_address` valor do `{UPSERT_TABLE_NAME}` registro substituirá o valor existente (operação de atualização).
- Se um registro de cliente que está em `{UPSERT_TABLE_NAME}` não existir em `{TABLE_NAME}`, o `{UPSERT_TABLE_NAME}` registro será adicionado à `{TABLE_NAME}` (operação de inserção).

Excluindo dados nas tabelas do Iceberg

Para excluir dados de uma tabela do Iceberg, use a `DELETE FROM` expressão e especifique um filtro que corresponda às linhas a serem excluídas.

```
spark.sql(f"""
DELETE FROM {CATALOG_NAME}.{db.name}.{table.name}
WHERE c_customer_sk % 2 != 0
""")
```

Se o filtro corresponder a uma partição inteira, o Iceberg executará uma exclusão somente de metadados e deixará os arquivos de dados no lugar. Caso contrário, ele reescreve somente os arquivos de dados afetados.

O método delete pega os arquivos de dados afetados pela WHERE cláusula e cria uma cópia deles sem os registros excluídos. Em seguida, ele cria um novo instantâneo da tabela que aponta para os novos arquivos de dados. Portanto, os registros excluídos ainda estão presentes nos instantâneos mais antigos da tabela. Por exemplo, se você recuperar o instantâneo anterior da tabela, verá os dados que acabou de excluir. Para obter informações sobre como remover instantâneos antigos desnecessários com os arquivos de dados relacionados para fins de limpeza, consulte a seção [Manutenção de arquivos usando compactação](#), mais adiante neste guia.

Leitura de dados

Você pode ler o status mais recente de suas tabelas Iceberg no Spark com o Spark SQL e DataFrames

Exemplo usando o Spark SQL:

```
spark.sql(f"""
SELECT * FROM {CATALOG_NAME}.{db.name}.{table.name} LIMIT 5
""")
```

Exemplo de uso da DataFrames API:

```
df = spark.table(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}").limit(5)
```

Usando a viagem no tempo

Cada operação de gravação (inserção, atualização, upsert, delete) em uma tabela do Iceberg cria um novo instantâneo. Em seguida, você pode usar esses instantâneos para viajar no tempo, para voltar no tempo e verificar o status de uma tabela no passado.

Para obter informações sobre como recuperar o histórico de instantâneos para tabelas usando snapshot-id e cronometrando valores, consulte a seção [Acessando metadados](#) mais adiante neste guia.

A consulta de viagem no tempo a seguir exibe o status de uma tabela com base em uma tabela específica snapshot-id.

Usando o Spark SQL:

```
spark.sql(f"""  
SELECT * FROM {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME} VERSION AS OF {snapshot_id}  
""")
```

Usando a DataFrames API:

```
df_1st_snapshot_id = spark.read.option("snapshot-id", snapshot_id) \  
    .format("iceberg") \  
    .load(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}") \  
    .limit(5)
```

A consulta de viagem no tempo a seguir exibe o status de uma tabela com base no último instantâneo criado antes de um timestamp específico, em milissegundos (). `as-of-timestamp`

Usando o Spark SQL:

```
spark.sql(f"""  
SELECT * FROM dev.{db.name}.{table.name} TIMESTAMP AS OF '{snapshot_ts}'  
""")
```

Usando a DataFrames API:

```
df_1st_snapshot_ts = spark.read.option("as-of-timestamp", snapshot_ts) \  
    .format("iceberg") \  
    .load(f"dev.{DB_NAME}.{TABLE_NAME}") \  
    .limit(5)
```

Usando consultas incrementais

Você também pode usar instantâneos do Iceberg para ler dados anexados de forma incremental.

Nota: Atualmente, essa operação oferece suporte à leitura de dados de `append` instantâneos. Ele não suporta a busca de dados de operações como `replaceoverwrite`, `oudelete`. Além disso, as operações de leitura incremental não são suportadas na sintaxe do Spark SQL.

O exemplo a seguir recupera todos os registros anexados a uma tabela Iceberg entre o instantâneo `start-snapshot-id` (exclusivo) e `end-snapshot-id` (inclusive).

```
df_incremental = (spark.read.format("iceberg")
    .option("start-snapshot-id", snapshot_id_start)
    .option("end-snapshot-id", snapshot_id_end)
    .load(f"glue_catalog.{DB_NAME}.{TABLE_NAME}")
)
```

Acessando metadados

O Iceberg fornece acesso aos seus metadados por meio de SQL. Você pode acessar os metadados de qualquer tabela (<table_name>) consultando o namespace. <table_name>.<metadata_table> Para obter uma lista completa das tabelas de metadados, consulte Como [inspecionar tabelas na documentação](#) do Iceberg.

O exemplo a seguir mostra como acessar a tabela de metadados do histórico do Iceberg, que mostra o histórico de confirmações (alterações) de uma tabela do Iceberg.

Usando o Spark SQL (com a %%sql mágica) de um notebook Amazon EMR Studio:

```
Spark.sql(f"""
SELECT * FROM {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}.history LIMIT 5
""")
```

Usando a DataFrames API:

```
spark.read.format("iceberg").load("{CATALOG_NAME}.{DB_NAME}.
{TABLE_NAME}.history").show(5, False)
```

Exemplo de resultado:

Type:	Table	Pie	Scatter	Line	Area	Bar
	made_current_at	snapshot_id	parent_id	is_current_ancestor		
	2023-01-09 02:50:17.547000+00:00	7501027970051178613	6598755163776233735			True
	2023-01-12 05:39:29.567000+00:00	7069175828427777019	7501027970051178613			True
	2023-01-12 05:39:58.807000+00:00	5173022175861138222	7069175828427777019			True
	2023-01-12 05:40:18.499000+00:00	3703414997660223390	5173022175861138222			True
	2023-01-12 05:40:41.827000+00:00	3807904412292252460	3703414997660223390			True

Trabalhar com tabelas do Apache Iceberg usando o SQL do Amazon Athena

O Amazon Athena fornece suporte integrado para o Apache Iceberg e não exige etapas ou configurações adicionais. Esta seção fornece uma visão geral detalhada dos recursos suportados e orientações de alto nível sobre o uso do Athena para interagir com as tabelas Iceberg.

Compatibilidade de versões e recursos

Note

As seções a seguir pressupõem que você esteja usando a [versão 3 do motor Athena](#).

Suporte à especificação da tabela Iceberg

A especificação da tabela Apache Iceberg especifica como as tabelas Iceberg devem se comportar. O Athena oferece suporte ao formato de tabela versão 2, portanto, qualquer tabela Iceberg que você criar com o console, a CLI ou o SDK usa essa versão inerentemente.

[Se você usa uma tabela Iceberg criada com outro mecanismo, como o Apache Spark no Amazon EMR ou AWS Glue, certifique-se de definir a versão do formato da tabela usando as propriedades da tabela.](#) Como referência, consulte a seção [Criando e escrevendo tabelas Iceberg](#) anteriormente neste guia.

Suporte ao recurso Iceberg

Você pode usar o Athena para ler e gravar em tabelas Iceberg. Quando você altera os dados usando as `DELETE FROM` instruções `UPDATE`, `EMERGE INTO`, o Athena suporta somente o merge-on-read modo. Essa propriedade não pode ser alterada. Para atualizar ou excluir dados com copy-on-write, você precisa usar outros mecanismos, como o Apache Spark no Amazon EMR ou. AWS Glue A tabela a seguir resume o suporte aos recursos do Iceberg no Athena.

		Suporte DDL		Suporte a DML		AWS Lake Formation para segurança (opcional)
	Formato da tabela	Create table	Evolução do esquema	Leitura de dados	Gravação de dados	Controle de acesso de linha/coluna
Amazon Athena	Versão 2	✓	✓	✓	XC opy-on-write	✓
					✓ Merge-on-read	✓

Note

O Athena não oferece suporte a consultas incrementais.

Trabalhando com mesas Iceberg

Para começar rapidamente a usar o Iceberg no Athena, consulte a [seção Introdução às tabelas do Iceberg no Athena](#) SQL, anteriormente neste guia.

A tabela a seguir lista as limitações e recomendações.

Cenário	Limitação	Recomendação
Tabela de geração de DDL	As tabelas de iceberg criadas com outros mecanismos podem ter propriedades que não estão expostas no Athena. Para essas tabelas, não é possível gerar o DDL.	Use a instrução equivalente no mecanismo que criou a tabela (por exemplo, a SHOW CREATE TABLE instrução do Spark).

Cenário	Limitação	Recomendação
Prefixos aleatórios do Amazon S3 em objetos que são gravados em uma tabela Iceberg	Por padrão, as tabelas Iceberg criadas com o Athena têm <code>write.object-storage.enabled</code> a propriedade ativada.	Para desativar esse comportamento e obter controle total sobre as propriedades da tabela Iceberg, crie uma tabela Iceberg com outro mecanismo, como o Spark no Amazon EMR ou. AWS Glue
Consultas incrementais	Atualmente, não há suporte no Athena.	Para usar consultas incrementais para permitir pipelines incrementais de ingestão de dados, use o Spark no Amazon EMR ou. AWS Glue

Migração de tabelas existentes para o Iceberg

Para migrar suas tabelas atuais do Athena AWS Glue (também conhecidas como tabelas do Hive) para o formato Iceberg, você pode usar a migração de dados local ou completa:

- A migração local é o processo de gerar os arquivos de metadados do Iceberg sobre os arquivos de dados existentes.
- A migração completa de dados cria a camada de metadados do Iceberg e também reescreve os arquivos de dados existentes da tabela original para a nova tabela do Iceberg.

As seções a seguir fornecem uma visão geral das APIs disponíveis para migrar tabelas e orientações para escolher uma estratégia de migração. Para obter mais informações sobre essas duas estratégias, consulte a seção [Migração de tabelas](#) na documentação do Iceberg.

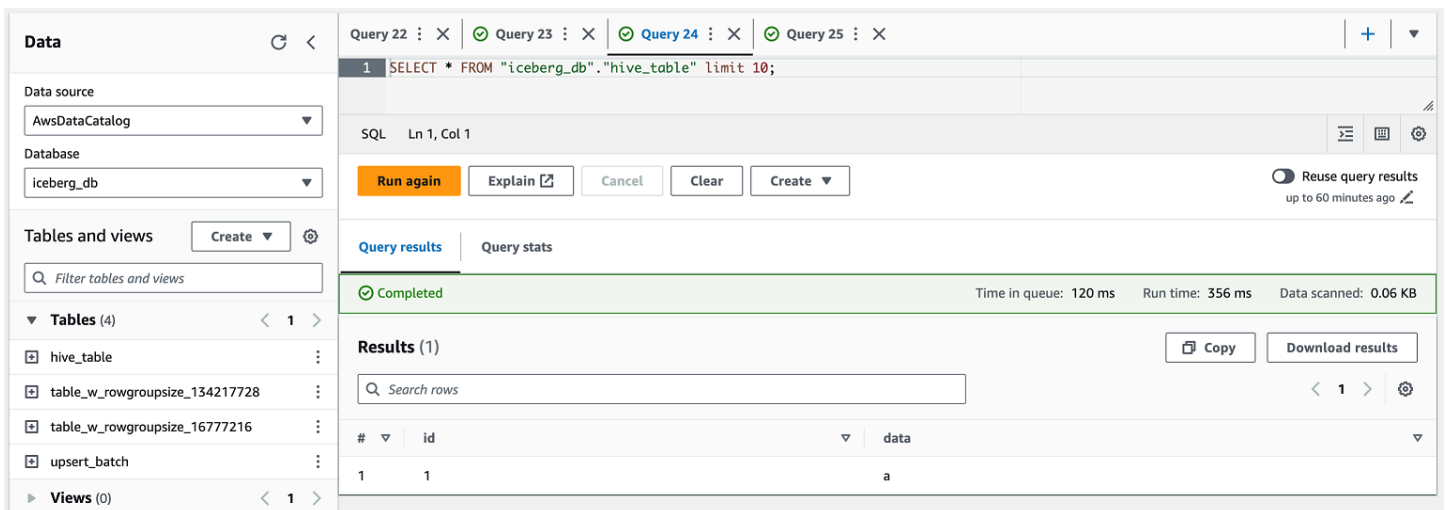
Migração local

A migração local elimina a necessidade de reescrever todos os arquivos de dados. Em vez disso, os arquivos de metadados do Iceberg são gerados e vinculados aos seus arquivos de dados existentes. O Iceberg oferece três opções para implementar a migração local:

- Usando o snapshot procedimento, conforme explicado nas seções [Tabela de instantâneos](#) e [Procedimento do Spark: instantâneo](#) na documentação do Iceberg.
- Usando o `add_files` procedimento, conforme explicado nas seções [Adicionar arquivos](#) e [Procedimento do Spark: add_files](#) na documentação do Iceberg.
- Usando o `migrate` procedimento, conforme explicado nas seções [Migrar tabela](#) e [Procedimento do Spark: Migrar](#) na documentação do Iceberg.

Atualmente, o procedimento de migração não funciona diretamente com o AWS Glue Data Catalog — ele funciona somente com o metastore Hive. Se você precisar usar o `migrate` procedimento em vez de `snapshot` ou `add_files`, você pode usar um cluster temporário do Amazon EMR com o metastore Hive (HMS). Essa abordagem requer a versão 1.2 ou posterior do Iceberg.

Digamos que você queira criar a seguinte tabela do Hive:



The screenshot shows the AWS Glue console interface. On the left, there's a sidebar with 'Data source' set to 'AwsDataCatalog' and 'Database' set to 'iceberg_db'. Below that, there's a 'Tables and views' section with a search filter and a list of tables including 'hive_table'. The main area shows a query editor with the SQL: `SELECT * FROM "iceberg_db"."hive_table" limit 10;`. Below the editor are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. The query execution status is 'Completed' with 'Time in queue: 120 ms', 'Run time: 356 ms', and 'Data scanned: 0.06 KB'. The results section shows one row with columns '#', 'id', and 'data', containing the value '1', '1', and 'a' respectively.

Você pode criar essa tabela do Hive executando este código no console do Athena:

```
CREATE EXTERNAL TABLE 'hive_table'(
  'id' bigint,
  'data' string)
USING parquet
LOCATION
  's3://datalake-xxxx/aws_workshop/iceberg_db/hive_table'

INSERT INTO iceberg_db.hive_table VALUES (1, 'a')
```

Se sua tabela do Hive estiver particionada, inclua a instrução de partição e adicione as partições de acordo com os requisitos do Hive.

```
ALTER TABLE default.placeholder_table_for_migration ADD
PARTITION (date = '2023-10-10')
```

Etapas:

1. Crie um cluster do Amazon EMR sem habilitar a AWS Glue Data Catalog integração, ou seja, não marque as caixas de seleção dos metadados das tabelas Hive ou Spark. Isso porque você usará o metastore nativo do Hive (HMS) que está disponível no cluster para essa solução alternativa.

AWS Glue Data Catalogue settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

- Use for Hive table metadata
- Use for Spark table metadata

2. Configure a sessão do Spark para usar a implementação do catálogo Iceberg Hive.

```
"spark.sql.extensions":"org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
"spark.sql.catalog.spark_catalog": "org.apache.iceberg.spark.SparkSessionCatalog",
"spark.sql.catalog.spark_catalog.type": "hive",
```

3. Valide se seu cluster do Amazon EMR não está conectado AWS Glue Data Catalog ao executando `show databases` ou `show tables`

```
[2]: %%sql
show databases
Last executed at 2023-07-05 12:24:26 in 35.03s
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
1	application_1686667730124_0002	pyspark	idle	Link	Link	None	✓

```
SparkSession available as 'spark'.

Type:  Table  Pie

namespace
default
```

4. Registre a tabela do Hive no metastore Hive do seu cluster do Amazon EMR e, em seguida, use o procedimento `Iceberg.migrate`

Register the Hive table in this local HMS catalog pointing to the S3 location where the files from the original Hive tables exist

```
%%sql -q
CREATE TABLE default.placeholder_hive_table (id bigint NOT NULL, data string)
USING parquet
LOCATION 's3://datalake-743490154766/aws_workshop/iceberg_db/hive_table/'
Last executed at 2023-07-05 12:55:19 in 3.25s
```

```
%%sql
select * from default.placeholder_hive_table limit 5
Last executed at 2023-07-05 12:57:13 in 7.43s
```

Type:

id	data
1	a

Once the Hive table is registered in this local HMS catalog, you can use Iceberg's migrate procedure.

```
spark.sql("CALL spark_catalog.system.migrate('default.placeholder_hive_table')")
Last executed at 2023-07-05 13:00:06 in 3.27s
```

► Spark Job Progress

DataFrame[migrated_files_count: bigint]

```
%%sql
show tables from default
Last executed at 2023-07-05 13:00:49 in 7.42s
```

Type:

namespace	tableName	isTemporary
default	placeholder_hive_table	False
default	placeholder_hive_table_backup_	False

Esse procedimento cria os arquivos de metadados do Iceberg no mesmo local da tabela do Hive.

5. Registre a tabela Iceberg migrada no. AWS Glue Data Catalog
6. Volte para um cluster do Amazon EMR que tenha a AWS Glue Data Catalog integração ativada.

AWS Glue Data Catalogue settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

- Use for Hive table metadata
- Use for Spark table metadata

7. Use a seguinte configuração do Iceberg na sessão do Spark.

```
"spark.sql.extensions": "org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
  "spark.sql.catalog.glue_catalog": "org.apache.iceberg.spark.SparkCatalog",
  "spark.sql.catalog.glue_catalog.warehouse": "s3://datalake-xxxx/
aws_workshop",
  "spark.sql.catalog.glue_catalog.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
  "spark.sql.catalog.glue_catalog.io-impl":
"org.apache.iceberg.aws.s3.S3FileIO",
```

Agora você pode consultar essa tabela no Amazon EMR ou no AWS Glue Athena.

```

%%sql
show tables from iceberg_db
Last executed at 2023-07-05 13:10:50 in 7.44s

Type: Table Pie Scatter Line Area Bar

namespace      tableName      isTemporary
iceberg_db      hive_table      False
iceberg_db      table_w_rowgroupsize_134217728      False
iceberg_db      table_w_rowgroupsize_16777216      False
iceberg_db      upsert_batch      False
iceberg_db      ws_webpage_pk_partitioned_140gb_trino      False

%%bash
aws s3 ls s3://datalake-743490154766/aws_workshop/iceberg_db/hive_table/metadata/
Last executed at 2023-07-05 13:10:20 in 488ms
2023-07-05 12:00:07      2239 00000-12a20051-6a3f-4b46-bae1-985f6df254db.metadata.json
2023-07-05 12:00:07      5802 b3d40480-0cb9-4cea-a4af-94c40a123689-m0.avro
2023-07-05 12:00:07      3781 snap-6104693268717769849-1-b3d40480-0cb9-4cea-a4af-94c40a123689.avro

metadata_file = "s3://datalake-743490154766/aws_workshop/iceberg_db/hive_table/metadata/00000-12a20051-6a3f-4b46-bae1-985f6df254db.metadata.json"
Last executed at 2023-07-05 13:11:46 in 49ms

spark.sql(f"CALL glue_catalog.system.register_table('iceberg_db.migrated_iceberg_table',{metadata_file})")
Last executed at 2023-07-05 13:12:27 in 3.32s

▶ Spark Job Progress

DataFrame[current_snapshot_id: bigint, total_records_count: bigint, total_data_files_count: bigint]

%%sql -q
alter table glue_catalog.iceberg_db.migrated_iceberg_table SET TBLPROPERTIES('format-version'='2')
Last executed at 2023-07-05 13:12:33 in 2.24s

%%sql
select * from glue_catalog.iceberg_db.migrated_iceberg_table limit 5
Last executed at 2023-07-05 13:12:44 in 7.42s

Type: Table Pie Scatter Line Area Bar

id data
1 a

```

Migração completa de dados

A migração completa de dados recria os arquivos de dados e os metadados. Essa abordagem leva mais tempo e requer recursos computacionais adicionais em comparação com a migração local. No entanto, essa opção ajuda a melhorar a qualidade da tabela: você pode validar os dados, fazer alterações no esquema e na partição, recorrer aos dados e assim por diante. Para implementar a migração completa de dados, use uma das seguintes opções:

- Use a declaração `CREATE TABLE ... AS SELECT (CTAS)` no Spark no Amazon EMR ou no Athena. AWS Glue Você pode definir a especificação da partição e as propriedades da tabela para a nova tabela Iceberg usando as `TBLPROPERTIES` cláusulas `PARTITIONED BY` e. Você

pode ajustar o esquema e o particionamento da nova tabela de acordo com suas necessidades, em vez de simplesmente herdá-los da tabela de origem.

- Leia a tabela de origem e grave os dados como uma nova tabela Iceberg usando o Spark no Amazon EMR ou AWS Glue (consulte [Criação de uma tabela](#) na documentação do Iceberg).

Selecionar uma estratégia de migração

Para escolher a melhor estratégia de migração, considere as perguntas na tabela a seguir.

Pergunta	Recomendação
Qual é o formato do arquivo de dados (por exemplo, CSV ou Apache Parquet)?	<ul style="list-style-type: none">• Considere a migração local se o formato do arquivo de tabela for Parquet, ORC ou Avro.• Para outros formatos, como CSV, JSON e assim por diante, use a migração de dados completa.
Você quer atualizar ou consolidar o esquema da tabela?	<ul style="list-style-type: none">• Se você quiser desenvolver o esquema da tabela usando os recursos nativos do Iceberg, considere a migração local. Por exemplo, você pode renomear colunas após a migração. (O esquema pode ser alterado na camada de metadados do Iceberg.)• Se você quiser excluir colunas inteiras dos arquivos de dados, recomendamos usar a migração de dados completa.
A tabela se beneficiaria com a mudança da estratégia de partição?	<ul style="list-style-type: none">• Se a abordagem de particionamento do Iceberg atender aos seus requisitos (por exemplo, novos dados são armazenados usando o novo layout de partição enquanto as partições existentes permanecem como estão), considere a migração local.

Pergunta

A tabela se beneficiaria com a adição ou alteração da estratégia de ordem de classificação?

A tabela tem muitos arquivos pequenos?

Recomendação

- Se você quiser usar partições ocultas em sua tabela, considere a migração completa dos dados. Para obter mais informações sobre partições ocultas, consulte a seção [Práticas recomendadas](#).
- Adicionar ou alterar a ordem de classificação dos seus dados exige a reescrita do conjunto de dados. Nesse caso, considere usar a migração de dados completa.
- Para tabelas grandes em que é extremamente caro reescrever todas as partições da tabela, considere usar a migração local e executar a compactação (com a classificação ativada) para as partições acessadas com mais frequência.
- A mesclagem de arquivos pequenos em arquivos maiores exige a regravação do conjunto de dados. Nesse caso, considere usar a migração de dados completa.
- Para tabelas grandes em que é extremamente caro reescrever todas as partições da tabela, considere usar a migração local e executar a compactação (com a classificação ativada) para as partições acessadas com mais frequência.

Melhores práticas para otimizar as cargas de trabalho do Apache Iceberg

O Iceberg é um formato de tabela projetado para simplificar o gerenciamento de data lake e aprimorar o desempenho da carga de trabalho. Casos de uso diferentes podem priorizar aspectos diferentes, como custo, desempenho de leitura, desempenho de gravação ou retenção de dados, então o Iceberg oferece opções de configuração para gerenciar essas compensações. Esta seção fornece informações para otimizar e ajustar suas cargas de trabalho do Iceberg para atender às suas necessidades.

Tópicos

- [Práticas recomendadas gerais](#)
- [Otimizando o desempenho de leitura](#)
- [Otimizando o desempenho de gravação](#)
- [Otimizando o armazenamento](#)
- [Manutenção de tabelas usando compactação](#)
- [Usando cargas de trabalho do Iceberg no Amazon S3](#)

Práticas recomendadas gerais

Independentemente do seu caso de uso, ao usar o Apache Iceberg on AWS, recomendamos que você siga essas melhores práticas gerais.

- Use a versão 2 do formato Iceberg.

Athena usa a versão 2 do formato Iceberg por padrão.

[Ao usar o Spark no Amazon EMR AWS Glue ou para criar tabelas Iceberg, especifique a versão do formato conforme descrito na documentação do Iceberg.](#)

- Use o AWS Glue Data Catalog como seu catálogo de dados.

Athena usa o AWS Glue Data Catalog por padrão.

Ao usar o Spark no Amazon EMR AWS Glue ou para trabalhar com o Iceberg, adicione a seguinte configuração à sua sessão do Spark para usar o catálogo de dados do AWS Glue. Para obter mais

informações, consulte a seção [Configurações do Spark para Iceberg no AWS Glue](#), no início deste guia.

```
"spark.sql.catalog.<your_catalog_name>.catalog-impl":  
"org.apache.iceberg.aws.glue.GlueCatalog"
```

- Use o AWS Glue Data Catalog como gerenciador de bloqueio.

O Athena usa o AWS Glue Data Catalog como gerenciador de bloqueio por padrão para tabelas Iceberg.

Ao usar o Spark no Amazon EMR AWS Glue ou para trabalhar com o Iceberg, certifique-se de configurar sua sessão do Spark para AWS Glue Data Catalog usá-la como gerenciador de bloqueio. Para obter mais informações, consulte [Optimistic Locking na documentação](#) do Iceberg.

- Use a compressão Zstandard (ZSTD).

O codec de compactação padrão do Iceberg é o gzip, que pode ser modificado usando a propriedade `table.write.<file_type>.compression-codec`. O Athena já usa o ZSTD como codec de compactação padrão para tabelas Iceberg.

Em geral, recomendamos o uso do codec de compactação ZSTD porque ele equilibra o GZIP e o Snappy e oferece um bom desempenho de leitura/gravação sem comprometer a taxa de compactação. Além disso, os níveis de compressão podem ser ajustados para atender às suas necessidades. Para obter mais informações, consulte [Níveis de compressão ZSTD no Athena na documentação do Athena](#).

O Snappy pode fornecer o melhor desempenho geral de leitura e gravação, mas tem uma taxa de compactação menor do que GZIP e ZSTD. Se você priorizar o desempenho, mesmo que isso signifique armazenar grandes volumes de dados no Amazon S3, o Snappy pode ser a melhor escolha.

Otimizando o desempenho de leitura

Esta seção discute as propriedades da tabela que você pode ajustar para otimizar o desempenho de leitura, independentemente do mecanismo.

Particionamento

Assim como nas tabelas do Hive, o Iceberg usa partições como camada primária de indexação para evitar a leitura de arquivos de metadados e arquivos de dados desnecessários. As estatísticas das colunas também são consideradas como uma camada secundária de indexação para melhorar ainda mais o planejamento de consultas, o que leva a um melhor tempo geral de execução.

Particionar dados

Para reduzir a quantidade de dados digitalizados ao consultar tabelas do Iceberg, escolha uma estratégia de partição balanceada que se alinhe aos padrões de leitura esperados:

- Identifique as colunas que são usadas com frequência em consultas. Esses são candidatos ideais para particionamento. Por exemplo, se você normalmente consulta dados de um determinado dia, um exemplo natural de uma coluna de partição seria uma coluna de data.
- Escolha uma coluna de partição de baixa cardinalidade para evitar a criação de um número excessivo de partições. Muitas partições podem aumentar o número de arquivos na tabela, o que pode afetar negativamente o desempenho da consulta. Como regra geral, “muitas partições” podem ser definidas como um cenário em que o tamanho dos dados na maioria das partições é menor que 2 a 5 vezes o valor definido por `target-file-size-bytes`

Note

Se você costuma consultar usando filtros em uma coluna de alta cardinalidade (por exemplo, uma `id` coluna que pode ter milhares de valores), use o recurso de particionamento oculto do Iceberg com transformações de bucket, conforme explicado na próxima seção.

Use particionamento oculto

Se suas consultas geralmente são filtradas por um derivado de uma coluna de tabela, use partições ocultas em vez de criar explicitamente novas colunas para funcionarem como partições. Para obter mais informações sobre esse recurso, consulte a [documentação do Iceberg](#).

Por exemplo, em um conjunto de dados que tem uma coluna de carimbo de data/hora (por exemplo, `2023-01-01 09:00:00`), em vez de criar uma nova coluna com a data analisada (por exemplo, `2023-01-01`), use transformações de partição para extrair a parte da data do carimbo de data/hora e criar essas partições rapidamente.

Os casos de uso mais comuns para particionamento oculto são:

- Particionamento em data ou hora, quando os dados têm uma coluna de carimbo de data/hora. O Iceberg oferece várias transformações para extrair as partes de data ou hora de um carimbo de data/hora.
- Particionamento em uma função hash de uma coluna, quando a coluna de particionamento tem alta cardinalidade e resultaria em muitas partições. O bucket do Iceberg transforma grupos de vários valores de partição em menos partições ocultas (bucket) usando funções de hash na coluna de particionamento.

Consulte as [transformações de partição](#) na documentação do Iceberg para obter uma visão geral de todas as transformações de partição disponíveis.

As colunas usadas para particionamento oculto podem se tornar parte dos predicados de consulta por meio do uso de funções SQL regulares, como `e. year()` `month()` Os predicados também podem ser combinados com operadores como `BETWEEN` e `AND`

Note

O Iceberg não pode realizar a remoção de partições para funções que geram um tipo de dados diferente; por exemplo, `substring(event_time, 1, 10) = '2022-01-01'`

Use a evolução da partição

Use a [evolução de partições do Iceberg](#) quando a estratégia de partição existente não for a ideal. Por exemplo, se você escolher partições horárias que acabem sendo muito pequenas (apenas alguns megabytes cada), considere mudar para partições diárias ou mensais.

Você pode usar essa abordagem quando a melhor estratégia de particionamento para uma tabela inicialmente não está clara e você deseja refinar sua estratégia de particionamento à medida que obtém mais informações. Outro uso efetivo da evolução de partições é quando os volumes de dados mudam e a estratégia atual de particionamento se torna menos eficaz com o tempo.

Para obter instruções sobre como desenvolver partições, consulte as [extensões ALTER TABLE SQL](#) na documentação do Iceberg.

Ajustando os tamanhos dos arquivos

A otimização do desempenho da consulta envolve minimizar o número de arquivos pequenos em suas tabelas. Para um bom desempenho de consulta, geralmente recomendamos manter os arquivos Parquet e ORC maiores que 100 MB.

O tamanho do arquivo também afeta o planejamento de consultas para tabelas Iceberg. À medida que o número de arquivos em uma tabela aumenta, o mesmo acontece com o tamanho dos arquivos de metadados. Arquivos de metadados maiores podem resultar em um planejamento de consultas mais lento. Portanto, quando o tamanho da tabela aumentar, aumente o tamanho do arquivo para aliviar a expansão exponencial dos metadados.

Use as práticas recomendadas a seguir para criar arquivos de tamanho adequado nas tabelas do Iceberg.

Defina o tamanho do arquivo de destino e do grupo de linhas

O Iceberg oferece os seguintes parâmetros-chave de configuração para ajustar o layout do arquivo de dados. Recomendamos que você use esses parâmetros para definir o tamanho do arquivo de destino e o grupo de linhas ou tamanho do traçado.

Parâmetro	Valor padrão	Comentário
<code>write.target-file-size-bytes</code>	512 MB	Esse parâmetro especifica o tamanho máximo do arquivo que o Iceberg criará. No entanto, alguns arquivos podem ser gravados com um tamanho menor que esse limite.
<code>write.parquet.row-group-size-bytes</code>	128 MB	Tanto o Parquet quanto o ORC armazenam dados em partes para que os mecanismos possam evitar a leitura do arquivo inteiro em algumas operações.

Parâmetro	Valor padrão	Comentário
<code>write.orc.stripe-size-bytes</code>	64 MB	
<code>write.distribution-mode</code>	Nenhum, para a versão 1.1 e inferior do Iceberg Hash, começando com a versão 1.2 do Iceberg	O Iceberg solicita que o Spark classifique os dados entre suas tarefas antes de gravar no armazenamento.

- Com base no tamanho esperado da tabela, siga estas diretrizes gerais:
 - Tabelas pequenas (até alguns gigabytes) — Reduza o tamanho do arquivo de destino para 128 MB. Reduza também o tamanho do grupo de linhas ou da faixa (por exemplo, para 8 ou 16 MB).
 - Tabelas de médio a grande porte (de alguns gigabytes a centenas de gigabytes) — Os valores padrão são um bom ponto de partida para essas tabelas. Se suas consultas forem muito seletivas, ajuste o grupo de linhas ou o tamanho da faixa (por exemplo, para 16 MB).
 - Tabelas muito grandes (centenas de gigabytes ou terabytes) — Aumente o tamanho do arquivo de destino para 1024 MB ou mais e considere aumentar o tamanho do grupo de linhas ou da faixa se suas consultas geralmente extraírem grandes conjuntos de dados.
- Para garantir que os aplicativos Spark que gravam nas tabelas do Iceberg criem arquivos de tamanho adequado, defina a `write.distribution-mode` propriedade como `ou.hash.range`. Para obter uma explicação detalhada da diferença entre esses modos, consulte [Escrevendo modos de distribuição](#) na documentação do Iceberg.

Essas são diretrizes gerais. Recomendamos que você execute testes para identificar os valores mais adequados para suas tabelas e cargas de trabalho específicas.

Execute a compactação regular

As configurações na tabela anterior definem um tamanho máximo de arquivo que as tarefas de gravação podem criar, mas não garantem que os arquivos tenham esse tamanho. Para garantir tamanhos de arquivo adequados, execute a compactação regularmente para combinar arquivos pequenos em arquivos maiores. Para obter orientação detalhada sobre como executar a compactação, consulte a [compactação do Iceberg](#) mais adiante neste guia.

Otimizar as estatísticas da coluna

O Iceberg usa estatísticas de colunas para realizar a remoção de arquivos, o que melhora o desempenho das consultas ao reduzir a quantidade de dados que são digitalizados pelas consultas. Para se beneficiar das estatísticas de colunas, certifique-se de que o Iceberg colete estatísticas de todas as colunas que são usadas com frequência em filtros de consulta.

Por padrão, o Iceberg coleta estatísticas somente para as [primeiras 100 colunas em cada tabela](#), conforme definido pela propriedade da tabela. `write.metadata.metrics.max-inferred-column-defaults`. Se sua tabela tiver mais de 100 colunas e suas consultas frequentemente referenciam colunas fora das primeiras 100 colunas (por exemplo, você pode ter consultas que filtram na coluna 132), certifique-se de que o Iceberg colete estatísticas sobre essas colunas. Há duas opções para conseguir isso:

- Ao criar a tabela Iceberg, reordene as colunas para que as colunas necessárias para consultas estejam dentro do intervalo de colunas definido por `write.metadata.metrics.max-inferred-column-defaults` (o padrão é 100).

Observação: se você não precisar de estatísticas em 100 colunas, poderá ajustar a `write.metadata.metrics.max-inferred-column-defaults` configuração para um valor desejado (por exemplo, 20) e reordenar as colunas para que as colunas necessárias para ler e escrever consultas caiam nas primeiras 20 colunas no lado esquerdo do conjunto de dados.

- Se você usar apenas algumas colunas nos filtros de consulta, poderá desativar a propriedade geral da coleta de métricas e escolher seletivamente colunas individuais para as quais coletar estatísticas, conforme mostrado neste exemplo:

```
.tableProperty("write.metadata.metrics.default", "none")
.tableProperty("write.metadata.metrics.column.my_col_a", "full")
.tableProperty("write.metadata.metrics.column.my_col_b", "full")
```

Observação: as estatísticas das colunas são mais eficazes quando os dados são classificados nessas colunas. Para obter mais informações, consulte a seção [Definir a ordem de classificação](#) mais adiante neste guia.

Escolha a estratégia de atualização correta

Use uma copy-on-write estratégia para otimizar o desempenho de leitura, quando operações de gravação mais lentas forem aceitáveis para seu caso de uso. Essa é a estratégia padrão usada pelo Iceberg.

C copy-on-write resulta em melhor desempenho de leitura, porque os arquivos são gravados diretamente no armazenamento de forma otimizada para leitura. No entanto, em comparação com merge-on-read, cada operação de gravação leva mais tempo e consome mais recursos computacionais. Isso apresenta uma compensação clássica entre latência de leitura e gravação. Normalmente, copy-on-write é ideal para casos de uso em que a maioria das atualizações é colocada nas mesmas partições de tabela (por exemplo, para cargas diárias em lote).

copy-on-write As configurações C (`write.update.mode`, `write.delete.mode`, `write.merge.mode`) podem ser definidas no nível da tabela ou de forma independente no lado do aplicativo.

Use compressão ZSTD

Você pode modificar o codec de compactação usado pelo Iceberg usando a propriedade `table.write.<file_type>.compression-codec`. Recomendamos que você use o codec de compressão ZSTD para melhorar o desempenho geral nas tabelas.

Por padrão, as versões 1.3 e anteriores do Iceberg usam a compressão GZIP, que fornece um desempenho de leitura/gravação mais lento em comparação com o ZSTD.

Observação: alguns mecanismos podem usar valores padrão diferentes. Esse é o caso das [tabelas Iceberg criadas com o Athena](#) ou o Amazon EMR versão 7.x.

Definir a ordem de classificação

Para melhorar o desempenho de leitura nas tabelas do Iceberg, recomendamos que você classifique sua tabela com base em uma ou mais colunas usadas com frequência em filtros de consulta. A classificação, combinada com as estatísticas das colunas do Iceberg, pode tornar a remoção de arquivos significativamente mais eficiente, o que resulta em operações de leitura mais rápidas. A classificação também reduz o número de solicitações do Amazon S3 para consultas que usam as colunas de classificação nos filtros de consulta.

Você pode definir uma ordem de classificação hierárquica no nível da tabela executando uma instrução de linguagem de definição de dados (DDL) com o Spark. Para ver as opções disponíveis,

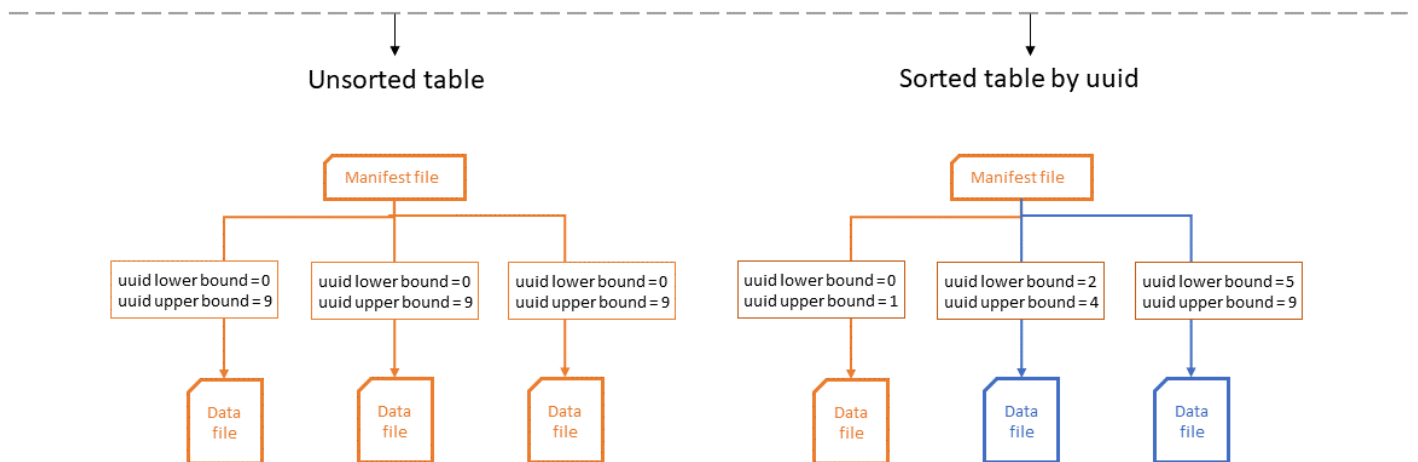
consulte a [documentação do Iceberg](#). Depois de definir a ordem de classificação, os gravadores aplicarão essa classificação às operações subsequentes de gravação de dados na tabela Iceberg.

Por exemplo, em tabelas particionadas por date (yyyy-mm-dd) em que a maioria das consultas é filtrada por uuid, você pode usar a opção DDL Write Distributed By Partition Locally Ordered para garantir que o Spark grave arquivos com intervalos não sobrepostos.

O diagrama a seguir ilustra como a eficiência das estatísticas de coluna melhora quando as tabelas são classificadas. No exemplo, a tabela classificada precisa abrir apenas um único arquivo e se beneficia ao máximo da partição e do arquivo do Iceberg. Na tabela não classificada, qualquer um uuid pode existir potencialmente em qualquer arquivo de dados, portanto, a consulta precisa abrir todos os arquivos de dados.

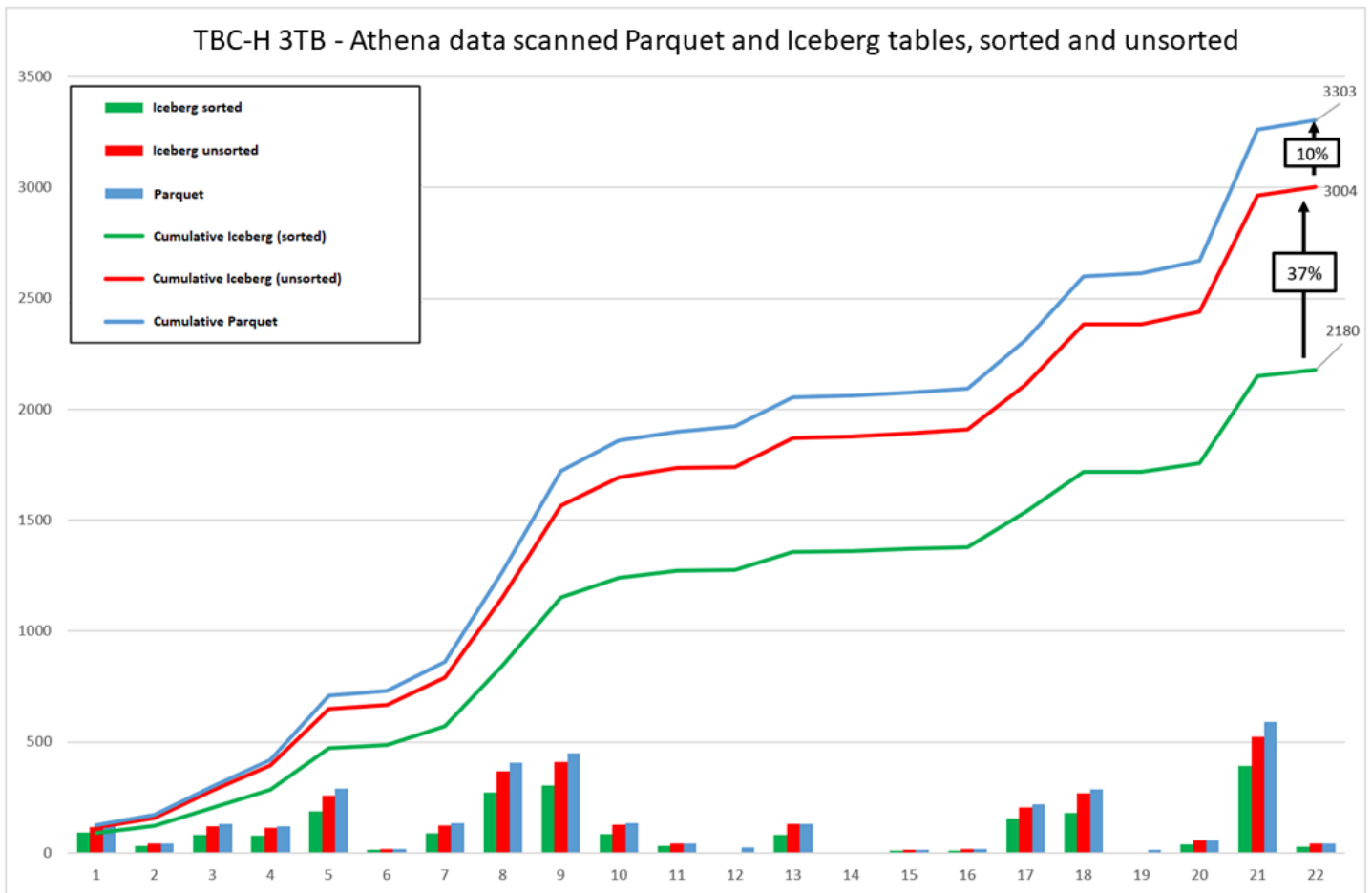
Query example:

```
SELECT * FROM Table
WHERE date > 2022-02-05 AND date < 2022-02-10 AND uuid = 1
```



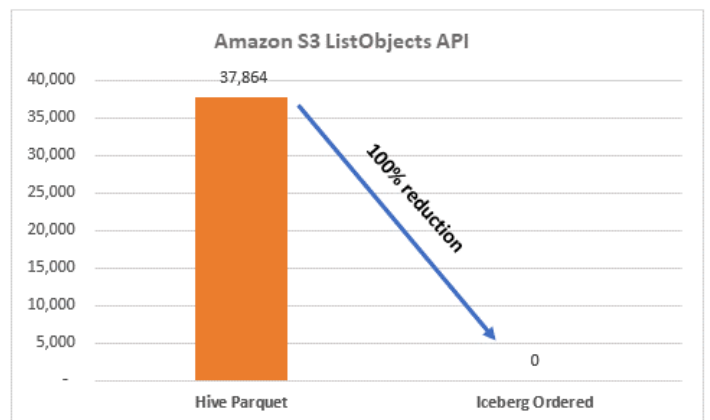
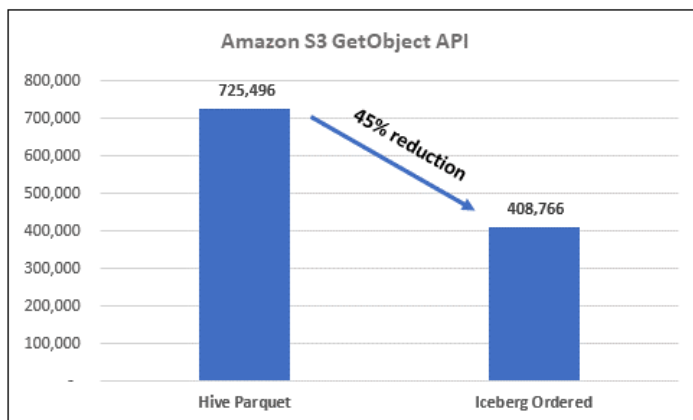
A alteração da ordem de classificação não afeta os arquivos de dados existentes. Você pode usar a compactação Iceberg para aplicar a ordem de classificação a eles.

Usar as tabelas classificadas do Iceberg pode diminuir os custos de sua carga de trabalho, conforme ilustrado no gráfico a seguir.



Esses gráficos resumem os resultados da execução do benchmark TPC-H para tabelas Hive (Parquet) em comparação com as tabelas classificadas do Iceberg. No entanto, os resultados podem ser diferentes para outros conjuntos de dados ou cargas de trabalho.

TPC-H 3TB - 22 queries



Otimizando o desempenho de gravação

Esta seção discute as propriedades da tabela que você pode ajustar para otimizar o desempenho de gravação nas tabelas do Iceberg, independentemente do mecanismo.

Definir o modo de distribuição da tabela

O Iceberg oferece vários modos de distribuição de gravação que definem como os dados de gravação são distribuídos nas tarefas do Spark. Para uma visão geral dos modos disponíveis, consulte [Escrevendo modos de distribuição](#) na documentação do Iceberg.

Para casos de uso que priorizam a velocidade de gravação, especialmente em cargas de trabalho de streaming, defina como `write.distribution-mode none`. Isso garante que o Iceberg não solicite embaralhamento adicional do Spark e que os dados sejam gravados assim que forem disponibilizados nas tarefas do Spark. Esse modo é particularmente adequado para aplicativos Spark Structured Streaming.

Note

Definir o modo de distribuição de gravação como `none` tende a produzir vários arquivos pequenos, o que degrada o desempenho de leitura. Recomendamos a compactação regular para consolidar esses arquivos pequenos em arquivos de tamanho adequado para o desempenho da consulta.

Escolha a estratégia de atualização correta

Use uma `merge-on-read` estratégia para otimizar o desempenho de gravação, quando operações de leitura mais lentas nos dados mais recentes forem aceitáveis para seu caso de uso.

Quando você usa `merge-on-read`, o Iceberg grava atualizações e exclui no armazenamento como pequenos arquivos separados. Quando a tabela é lida, o leitor precisa mesclar essas alterações com os arquivos base para retornar a visão mais recente dos dados. Isso resulta em uma penalidade de desempenho nas operações de leitura, mas acelera a gravação de atualizações e exclusões. Normalmente, `merge-on-read` é ideal para streaming de cargas de trabalho com atualizações ou trabalhos com poucas atualizações espalhadas por várias partições de tabela.

Você pode definir `merge-on-read` configurações (`write.update.modewrite.delete.mode`, `ewrite.merge.mode`) no nível da tabela ou de forma independente no lado do aplicativo.

O uso merge-on-read requer a execução de compactação regular para evitar que o desempenho de leitura se degrade com o tempo. A compactação reconcilia atualizações e exclusões com arquivos de dados existentes para criar um novo conjunto de arquivos de dados, eliminando assim a penalidade de desempenho incorrida no lado da leitura. Por padrão, a compactação do Iceberg não mescla arquivos excluídos, a menos que você altere o padrão da `delete-file-threshold` propriedade para um valor menor (consulte a documentação do [Iceberg](#)). Para saber mais sobre compactação, consulte a seção Compactação do [Iceberg](#) mais adiante neste guia.

Escolha o formato de arquivo correto

O Iceberg suporta a gravação de dados nos formatos Parquet, ORC e Avro. O parquet é o formato padrão. Parquet e ORC são formatos colunares que oferecem desempenho de leitura superior, mas geralmente são mais lentos para gravar. Isso representa a compensação típica entre desempenho de leitura e gravação.

Se a velocidade de gravação for importante para seu caso de uso, como em cargas de trabalho de streaming, considere escrever no formato Avro `write-format` configurando como Avro nas opções do gravador. Como o Avro é um formato baseado em linhas, ele fornece tempos de gravação mais rápidos ao custo de um desempenho de leitura mais lento.

Para melhorar o desempenho de leitura, execute a compactação regular para mesclar e transformar pequenos arquivos Avro em arquivos Parquet maiores. O resultado do processo de compactação é determinado pela configuração da `write.format.default` tabela. O formato padrão do Iceberg é o Parquet, então se você escrever em Avro e depois executar a compactação, o Iceberg transformará os arquivos Avro em arquivos Parquet. Veja um exemplo abaixo:

```
spark.sql(f"""
  CREATE TABLE IF NOT EXISTS glue_catalog.{DB_NAME}.{TABLE_NAME} (
    Col_1 float,
    <<<...other columns...>>
    ts timestamp)
  USING iceberg
  PARTITIONED BY (days(ts))
  OPTIONS (
    'format-version'='2',
    write.format.default='parquet')
  """)

query = df \
  .writeStream \
```

```

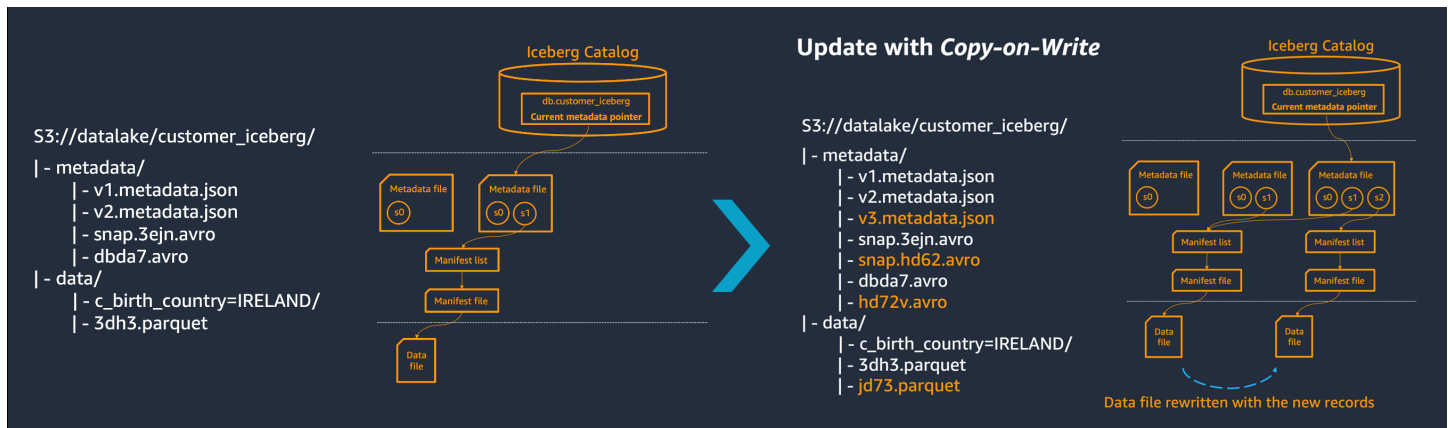
.format("iceberg") \
.option("write-format", "avro") \
.outputMode("append") \
.trigger(processingTime='60 seconds') \
.option("path", f"glue_catalog.{DB_NAME}.{TABLE_NAME}") \
.option("checkpointLocation", f"s3://{BUCKET_NAME}/checkpoints/iceberg/")

.start()

```

Otimizando o armazenamento

Atualizar ou excluir dados em uma tabela Iceberg aumenta o número de cópias de seus dados, conforme ilustrado no diagrama a seguir. O mesmo vale para a execução da compactação: ela aumenta o número de cópias de dados no Amazon S3. Isso porque o Iceberg trata os arquivos subjacentes a todas as tabelas como imutáveis.



Siga as práticas recomendadas nesta seção para gerenciar os custos de armazenamento.

Habilite o S3 Intelligent Tiering

Use a classe de armazenamento [Amazon S3 Intelligent-Tiering](#) para mover automaticamente os dados para o nível de acesso mais econômico quando os padrões de acesso mudarem. Essa opção não tem sobrecarga operacional nem impacto no desempenho.

Observação: não use os níveis opcionais (como Archive Access e Deep Archive Access) no S3 Intelligent-Tiering com tabelas Iceberg. Para arquivar dados, consulte as diretrizes na próxima seção.

[Você também pode usar as regras de ciclo de vida do Amazon S3 para definir suas próprias regras para mover objetos para outra classe de armazenamento do Amazon S3, como S3 Standard-IA ou](#)

[S3 One Zone-IA \(consulte Transições suportadas e restrições relacionadas na documentação do Amazon S3\).](#)

Arquive ou exclua instantâneos históricos

Para cada transação confirmada (inserção, atualização, mesclagem, compactação) em uma tabela Iceberg, uma nova versão ou instantâneo da tabela é criada. Com o tempo, o número de versões e o número de arquivos de metadados no Amazon S3 se acumulam.

É necessário manter instantâneos de uma tabela para recursos como isolamento de instantâneos, reversão de tabelas e consultas de viagem no tempo. No entanto, os custos de armazenamento aumentam com o número de versões que você retém.

A tabela a seguir descreve os padrões de design que você pode implementar para gerenciar custos com base em seus requisitos de retenção de dados.

Padrão de design	Solução	Casos de uso
Excluir instantâneos antigos	<ul style="list-style-type: none"> Use a instrução VACUUM no Athena para remover instantâneos antigos. Essa operação não incorre em nenhum custo computacional. Como alternativa, você pode usar o Spark no Amazon EMR AWS Glue ou remover instantâneos. Para obter mais informações, consulte <code>expire_snapshots</code> na documentação do Iceberg. 	<p>Essa abordagem exclui instantâneos que não são mais necessários para reduzir os custos de armazenamento. Você pode configurar quantos instantâneos devem ser retidos ou por quanto tempo, com base em seus requisitos de retenção de dados.</p> <p>Essa opção executa uma exclusão definitiva dos instantâneos. Você não pode reverter ou viajar no tempo para instantâneos expirados.</p>
Defina políticas de retenção para instantâneos específicos	<ol style="list-style-type: none"> Use tags para marcar instantâneos específicos e definir uma política de retenção no Iceberg. Para 	<p>Esse padrão é útil para a conformidade com requisitos comerciais ou legais que exigem que você mostre o</p>

Padrão de design

Solução

obter mais informações, consulte [Tags históricas](#) na documentação do Iceberg.

Por exemplo, você pode reter um snapshot por mês durante um ano usando a seguinte instrução SQL no Spark no Amazon EMR:

```
ALTER TABLE glue_catalog.db.table
CREATE TAG 'EOM-01' AS
OF VERSION 30 RETAIN
365 DAYS
```

2. Use o Spark no Amazon EMR AWS Glue ou remova os demais snapshots intermediários não marcados.

Casos de uso

estado de uma tabela em um determinado momento no passado. Ao colocar políticas de retenção em instantâneos marcados específicos, você pode remover outros instantâneos (não marcados) que foram criados. Dessa forma, você pode atender aos requisitos de retenção de dados sem reter cada snapshot criado.

Padrão de design	Solução	Casos de uso
Arquive instantâneos antigos	<ol style="list-style-type: none"><li data-bbox="535 210 1023 987">1. Use as tags do Amazon S3 para marcar objetos com o Spark. (As tags do Amazon S3 são diferentes das tags Iceberg; para obter mais informações, consulte a documentação do Iceberg.) Por exemplo: . <pre data-bbox="617 630 1023 987">spark.sql.catalog. my_catalog.s3.delete-enabled=false and \ spark.sql.catalog. my_catalog.s3.delete.tags.my_key=to_archive</pre><li data-bbox="535 987 1023 1428">2. Use o Spark no Amazon EMR AWS Glue ou para remover instantâneos. Quando você usa as configurações no exemplo, esse procedimento marca objetos e os separa dos metadados da tabela Iceberg em vez de excluí-los do Amazon S3.<li data-bbox="535 1428 1023 1810">3. Use as regras de ciclo de vida do S3 para fazer a transição de objetos marcados como <code>to_archive</code> uma das classes de armazenamento do S3 Glacier.	<p data-bbox="1023 210 1520 399">Esse padrão permite que você mantenha todas as versões e instantâneos da tabela a um custo menor.</p> <p data-bbox="1023 441 1520 777">Você não pode viajar no tempo ou reverter para instantâneos arquivados sem primeiro restaurar essas versões como novas tabelas. Isso geralmente é aceitável para fins de auditoria.</p> <p data-bbox="1023 819 1520 1029">Você pode combinar essa abordagem com o padrão de design anterior, definindo políticas de retenção para instantâneos específicos.</p>

Padrão de design

Solução

Casos de uso

4. Para consultar dados arquivados:

- [Restaure os objetos arquivados](#).
- Use o [procedimento `register_table`](#) no Iceberg para registrar o instantâneo como uma tabela no catálogo.

Para obter instruções detalhadas, consulte a postagem do AWS blog [Melhore a eficiência operacional das tabelas Apache Iceberg criadas nos data lakes do Amazon S3](#).

Excluir arquivos órfãos

Em determinadas situações, os aplicativos do Iceberg podem falhar antes que você confirme suas transações. Isso deixa os arquivos de dados no Amazon S3. Como não houve confirmação, esses arquivos não serão associados a nenhuma tabela, então talvez seja necessário limpá-los de forma assíncrona.

Para lidar com essas exclusões, você pode usar a [instrução `VACUUM`](#) no Amazon Athena. Essa instrução remove instantâneos e também exclui arquivos órfãos. Isso é muito econômico, porque o Athena não cobra pelo custo computacional dessa operação. Além disso, você não precisa agendar nenhuma operação adicional ao usar a `VACUUM` instrução.

Como alternativa, você pode usar o Spark no Amazon EMR AWS Glue ou executar `remove_orphan_files` o procedimento. Essa operação tem um custo computacional e precisa

ser programada de forma independente. Para obter mais informações, consulte a [documentação do Iceberg](#).

Manutenção de tabelas usando compactação

O Iceberg inclui recursos que permitem realizar [operações de manutenção da tabela](#) depois de gravar dados na tabela. Algumas operações de manutenção se concentram na simplificação dos arquivos de metadados, enquanto outras aprimoram a forma como os dados são agrupados nos arquivos para que os mecanismos de consulta possam localizar com eficiência as informações necessárias para responder às solicitações do usuário. Esta seção se concentra nas otimizações relacionadas à compactação.

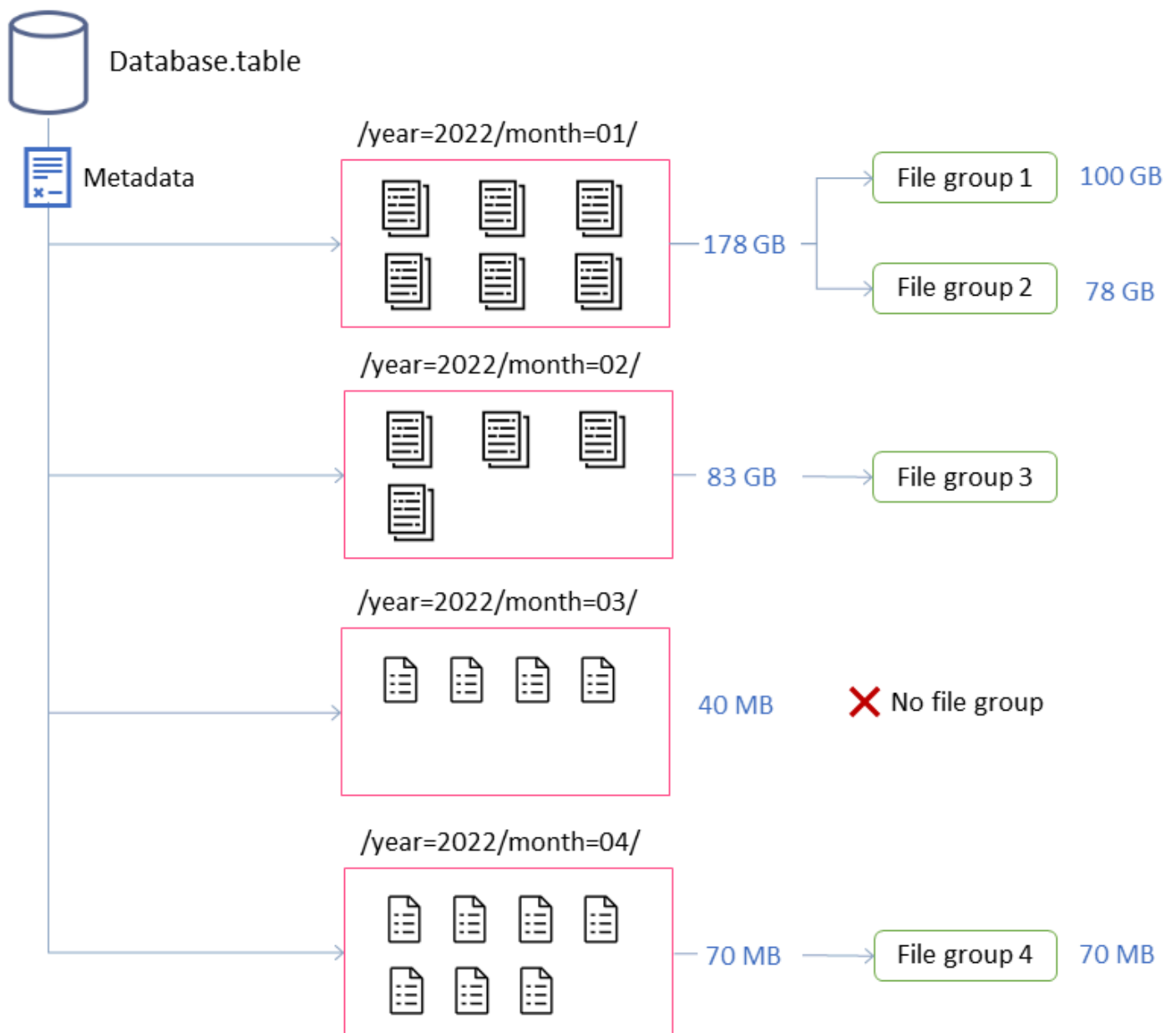
Compactação de iceberg

No Iceberg, você pode usar a compactação para realizar quatro tarefas:

- Combinação de arquivos pequenos em arquivos maiores que geralmente têm mais de 100 MB. Essa técnica é conhecida como embalagem de lixo.
- Mesclando arquivos excluídos com arquivos de dados. Os arquivos excluídos são gerados por atualizações ou exclusões que usam a merge-on-read abordagem.
- (Re) classificando os dados de acordo com os padrões de consulta. Os dados podem ser gravados sem qualquer ordem de classificação ou com uma ordem de classificação adequada para gravações e atualizações.
- Agrupar os dados usando curvas de preenchimento de espaço para otimizar padrões de consulta distintos, particularmente a classificação de ordem z.

Ativado AWS, você pode executar operações de compactação e manutenção de tabelas para o Iceberg por meio do Amazon Athena ou usando o Spark no Amazon EMR ou AWS Glue

Ao executar a compactação usando o procedimento [rewrite_data_files](#), você pode ajustar vários botões para controlar o comportamento da compactação. O diagrama a seguir mostra o comportamento padrão do empacotamento de lixo. Compreender a compactação de compartimentos é fundamental para entender as implementações de classificação hierárquica e classificação de ordem Z, porque elas são extensões da interface de embalagem de compartimentos e operam de maneira semelhante. A principal distinção é a etapa adicional necessária para classificar ou agrupar os dados.



Neste exemplo, a tabela Iceberg consiste em quatro partições. Cada partição tem um tamanho diferente e um número diferente de arquivos. Se você iniciar um aplicativo Spark para executar a compactação, o aplicativo criará um total de quatro grupos de arquivos para processar. Um grupo de arquivos é uma abstração do Iceberg que representa uma coleção de arquivos que serão processados por uma única tarefa do Spark. Ou seja, o aplicativo Spark que executa a compactação criará quatro trabalhos do Spark para processar os dados.

Ajustando o comportamento de compactação

As propriedades principais a seguir controlam como os arquivos de dados são selecionados para compactação:

- [MAX_FILE_GROUP_SIZE_BYTES](#) define o limite de dados para um único grupo de arquivos (tarefa do Spark) em 100 GB por padrão. Essa propriedade é especialmente importante para tabelas sem partições ou tabelas com partições que abrangem centenas de gigabytes. Ao definir esse limite, você pode dividir as operações para planejar o trabalho e progredir, evitando o esgotamento dos recursos no cluster.

Observação: cada grupo de arquivos é classificado separadamente. Portanto, se você quiser realizar uma classificação em nível de partição, deverá ajustar esse limite para corresponder ao tamanho da partição.

- [MIN_FILE_SIZE_BYTES](#) ou [MIN_FILE_SIZE_DEFAULT_RATIO](#) assumem como padrão 75 por cento do tamanho do arquivo de destino definido no nível da tabela. Por exemplo, se uma tabela tiver um tamanho alvo de 512 MB, qualquer arquivo menor que 384 MB será incluído no conjunto de arquivos que serão compactados.
- [MAX_FILE_SIZE_BYTES](#) ou [MAX_FILE_SIZE_DEFAULT_RATIO](#) assumem como padrão 180 por cento do tamanho do arquivo de destino. Assim como as duas propriedades que definem tamanhos mínimos de arquivo, essas propriedades são usadas para identificar arquivos candidatos para o trabalho de compactação.
- [MIN_INPUT_FILES](#) especifica o número mínimo de arquivos a serem compactados se o tamanho da partição da tabela for menor que o tamanho do arquivo de destino. O valor dessa propriedade é usado para determinar se vale a pena compactar os arquivos com base no número de arquivos (o padrão é 5).
- [DELETE_FILE_THRESHOLD](#) especifica o número mínimo de operações de exclusão de um arquivo antes de ser incluído na compactação. A menos que você especifique o contrário, a compactação não combina arquivos excluídos com arquivos de dados. Para habilitar essa funcionalidade, você deve definir um valor limite usando essa propriedade. Esse limite é específico para arquivos de dados individuais, portanto, se você defini-lo como 3, um arquivo de dados será regravado somente se houver três ou mais arquivos de exclusão que façam referência a ele.

Essas propriedades fornecem informações sobre a formação dos grupos de arquivos no diagrama anterior.

Por exemplo, a partição rotulada `month=01` inclui dois grupos de arquivos porque excede a restrição de tamanho máximo de 100 GB. Por outro lado, a `month=02` partição contém um único grupo de arquivos porque tem menos de 100 GB. A `month=03` partição não atende ao requisito mínimo padrão de arquivo de entrada de cinco arquivos. Como resultado, ele não será compactado. Por fim, embora a `month=04` partição não contenha dados suficientes para formar um único arquivo do tamanho desejado, os arquivos serão compactados porque a partição inclui mais de cinco arquivos pequenos.

Você pode definir esses parâmetros para o Spark executado no Amazon AWS Glue EMR ou. Para o Amazon Athena, você pode gerenciar propriedades semelhantes usando as propriedades da [tabela](#) (que começam com o prefixo `optimize_`).

Executando a compactação com o Spark no Amazon EMR ou AWS Glue

Esta seção descreve como dimensionar adequadamente um cluster Spark para executar o utilitário de compactação do Iceberg. O exemplo a seguir usa o Amazon EMR Serverless, mas você pode usar a mesma metodologia no Amazon EMR no Amazon EC2 ou no Amazon EKS, ou em AWS Glue.

Você pode aproveitar a correlação entre grupos de arquivos e trabalhos do Spark para planejar os recursos do cluster. Para processar os grupos de arquivos sequencialmente, considerando o tamanho máximo de 100 GB por grupo de arquivos, você pode definir as seguintes propriedades do [Spark](#):

- `spark.dynamicAllocation.enabled = FALSE`
- `spark.executor.memory = 20 GB`
- `spark.executor.instances = 5`

Se quiser acelerar a compactação, você pode escalar horizontalmente aumentando o número de grupos de arquivos que são compactados em paralelo. Você também pode escalar o Amazon EMR usando escalabilidade manual ou dinâmica.

- Dimensionamento manual (por exemplo, por um fator de 4)
 - `MAX_CONCURRENT_FILE_GROUP_REWRITES= 4` (nosso fator)
 - `spark.executor.instances= 5` (valor usado no exemplo) x 4 (nosso fator) = 20
 - `spark.dynamicAllocation.enabled = FALSE`
- Escalabilidade dinâmica
 - `spark.dynamicAllocation.enabled= TRUE` (padrão, nenhuma ação é necessária)

- [MAX_CONCURRENT_FILE_GROUP_REWRITES](#) = N (alinhe esse valor com `comspark.dynamicAllocation.maxExecutors`, que é 100 por padrão; com base nas configurações do executor no exemplo, você pode definir como 20) N

Essas são diretrizes para ajudar a dimensionar o cluster. No entanto, você também deve monitorar o desempenho de suas tarefas do Spark para encontrar as melhores configurações para suas cargas de trabalho.

Executando a compactação com o Amazon Athena

[O Athena oferece uma implementação do utilitário de compactação do Iceberg como um recurso gerenciado por meio da declaração OPTIMIZE](#). Você pode usar essa instrução para executar a compactação sem precisar avaliar a infraestrutura.

Essa instrução agrupa arquivos pequenos em arquivos maiores usando o algoritmo de embalagem de compartimentos e mescla arquivos excluídos com arquivos de dados existentes. Para agrupar os dados usando classificação hierárquica ou classificação de ordem z, use o Spark no Amazon EMR ou AWS Glue

Você pode alterar o comportamento padrão da OPTIMIZE instrução na criação da tabela passando as propriedades da tabela na CREATE TABLE instrução ou após a criação da tabela usando a ALTER TABLE instrução. Para valores padrão, consulte a documentação do [Athena](#).

Recomendações para executar a compactação

Caso de uso

Executando a compactação da embalagem de lixeiras com base em um cronograma

Recomendação

- Use a OPTIMIZE declaração no Athena se você não souber quantos arquivos pequenos sua tabela contém. O modelo de preços do Athena é baseado nos dados digitalizados, portanto, se não houver arquivos a serem compactados, não haverá custo associado a essas operações. Para evitar o tempo limite nas tabelas do Athena, execute OPTIMIZE com base. `per-table-partition`

Caso de uso

Executando a compactação de embalagens de lixo com base em eventos

Executando a compactação para classificar os dados

Executando a compactação para agrupar os dados usando a classificação de ordem z

Executando a compactação em partições que podem ser atualizadas por outros aplicativos devido à chegada tardia de dados

Recomendação

- Use o Amazon EMR ou AWS Glue com escalabilidade dinâmica quando você espera que grandes volumes de arquivos pequenos sejam compactados.
- Use o Amazon EMR ou AWS Glue com escalabilidade dinâmica quando você espera que grandes volumes de arquivos pequenos sejam compactados.
- Use o Amazon EMR ou AWS Glue, porque a classificação é uma operação cara e pode precisar transferir dados para o disco.
- Use o Amazon EMR ou AWS Glue, porque a classificação por ordem z é uma operação muito cara e pode precisar transferir dados para o disco.
- Use o Amazon EMR ou AWS Glue. Ative a propriedade Iceberg [PARTIAL_PROGRESS_ENABLED](#). Quando você usa essa opção, o Iceberg divide a saída de compactação em vários commits. Se houver uma colisão (ou seja, se o arquivo de dados for atualizado durante a execução da compactação), essa configuração reduz o custo da nova tentativa, limitando-a à confirmação que inclui o arquivo afetado. Caso contrário, talvez seja necessário recompactar todos os arquivos.

Usando cargas de trabalho do Iceberg no Amazon S3

Esta seção discute as propriedades do Iceberg que você pode usar para otimizar a interação do Iceberg com o Amazon S3.

Evite o particionamento a quente (erros HTTP 503)

Alguns aplicativos de data lake executados no Amazon S3 manipulam milhões ou bilhões de objetos e processam petabytes de dados. Isso pode levar a prefixos que recebem um alto volume de tráfego, que normalmente são detectados por meio de erros HTTP 503 (serviço indisponível). Para evitar esse problema, use as seguintes propriedades do Iceberg:

- `write.distribution-mode` Defina `hash` para `range` que o Iceberg grave arquivos grandes, o que resulta em menos solicitações do Amazon S3. Essa é a configuração preferida e deve abordar a maioria dos casos.
- Se você continuar enfrentando erros 503 devido a um grande volume de dados em suas cargas de trabalho, você pode `write.object-storage.enabled` configurá-los no Iceberg. `true` Isso instrui o Iceberg a fazer o hash dos nomes dos objetos e distribuir a carga em vários prefixos aleatórios do Amazon S3.

Para obter mais informações sobre essas propriedades, consulte [Write properties](#) na documentação do Iceberg.

Use as operações de manutenção do Iceberg para liberar dados não utilizados

Para gerenciar tabelas Iceberg, você pode usar a API principal do Iceberg, clientes Iceberg (como o Spark) ou serviços gerenciados, como o Amazon Athena. [Para excluir arquivos antigos ou não utilizados do Amazon S3, recomendamos que você use somente as APIs nativas do Iceberg para remover instantâneos, remover arquivos de metadados antigos e excluir arquivos órfãos.](#)

Usar as APIs do Amazon S3 por meio do Boto3, do Amazon S3 SDK ou do AWS Command Line Interface (AWS CLI) ou usar qualquer outro método que não seja do Iceberg para substituir ou remover arquivos do Amazon S3 de uma tabela do Iceberg leva à corrupção da tabela e a falhas de consulta.

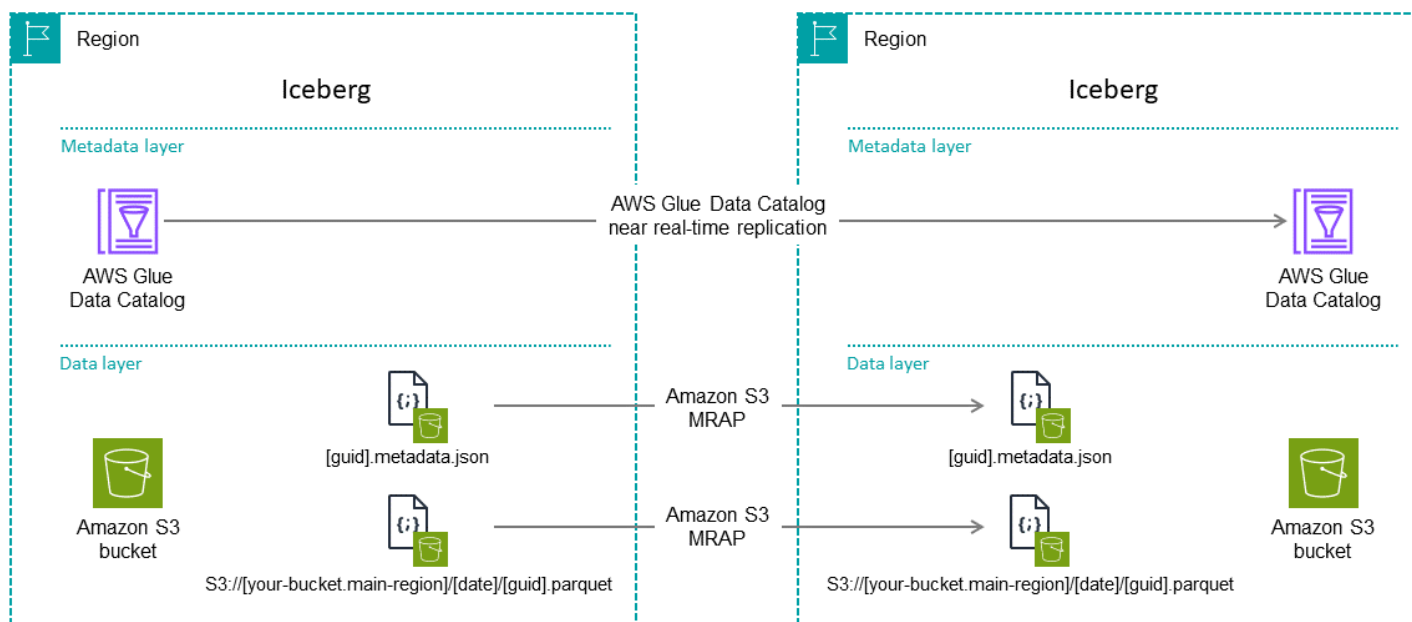
Replique dados em Regiões da AWS

Ao armazenar tabelas do Iceberg no Amazon S3, você pode usar os recursos integrados no Amazon S3, [como replicação entre regiões \(CRR\) e pontos de acesso multirregionais \(MRAP\), para replicar dados em várias regiões da AWS](#). O MRAP fornece um endpoint global para que os aplicativos acessem buckets S3 localizados em vários. Regiões da AWS O Iceberg não suporta caminhos relativos, mas você pode usar o MRAP para realizar operações do Amazon S3 mapeando buckets para pontos de acesso. O MRAP também se integra perfeitamente ao processo de replicação entre regiões do Amazon S3, o que introduz um atraso de até 15 minutos. Você precisa replicar os arquivos de dados e metadados.

⚠ Important

Atualmente, a integração do Iceberg com o MRAP funciona somente com o Apache Spark. Se você precisar fazer o failover para o secundário Região da AWS, planeje redirecionar as consultas do usuário para um ambiente Spark SQL (como o Amazon EMR) na região de failover.

Os recursos CRR e MRAP ajudam você a criar uma solução de replicação entre regiões para tabelas Iceberg, conforme ilustrado no diagrama a seguir.



Para configurar essa arquitetura de replicação entre regiões:

1. Crie tabelas usando a localização do MRAP. Isso garante que os arquivos de metadados do Iceberg apontem para o local do MRAP em vez do local físico do bucket.
2. Replique arquivos Iceberg usando o Amazon S3 MRAP. O MRAP oferece suporte à replicação de dados com um contrato de nível de serviço (SLA) de 15 minutos. O Iceberg impede que as operações de leitura introduzam inconsistências durante a replicação.
3. Disponibilize as tabelas AWS Glue Data Catalog na região secundária. Você pode escolher entre duas opções:
 - Configure um pipeline para replicar os metadados da tabela Iceberg usando a replicação. AWS Glue Data Catalog Esse utilitário está disponível no repositório de [replicação do GitHub Glue Catalog e do Lake Formation Permissions](#). Esse mecanismo orientado por eventos replica tabelas na região de destino com base nos registros de eventos.
 - Registre as tabelas na região secundária quando precisar fazer o failover. Para essa opção, você pode usar o utilitário anterior ou o [procedimento register_table](#) do Iceberg e apontá-lo para o arquivo mais recente. `metadata.json`

Monitorando cargas de trabalho do Apache Iceberg

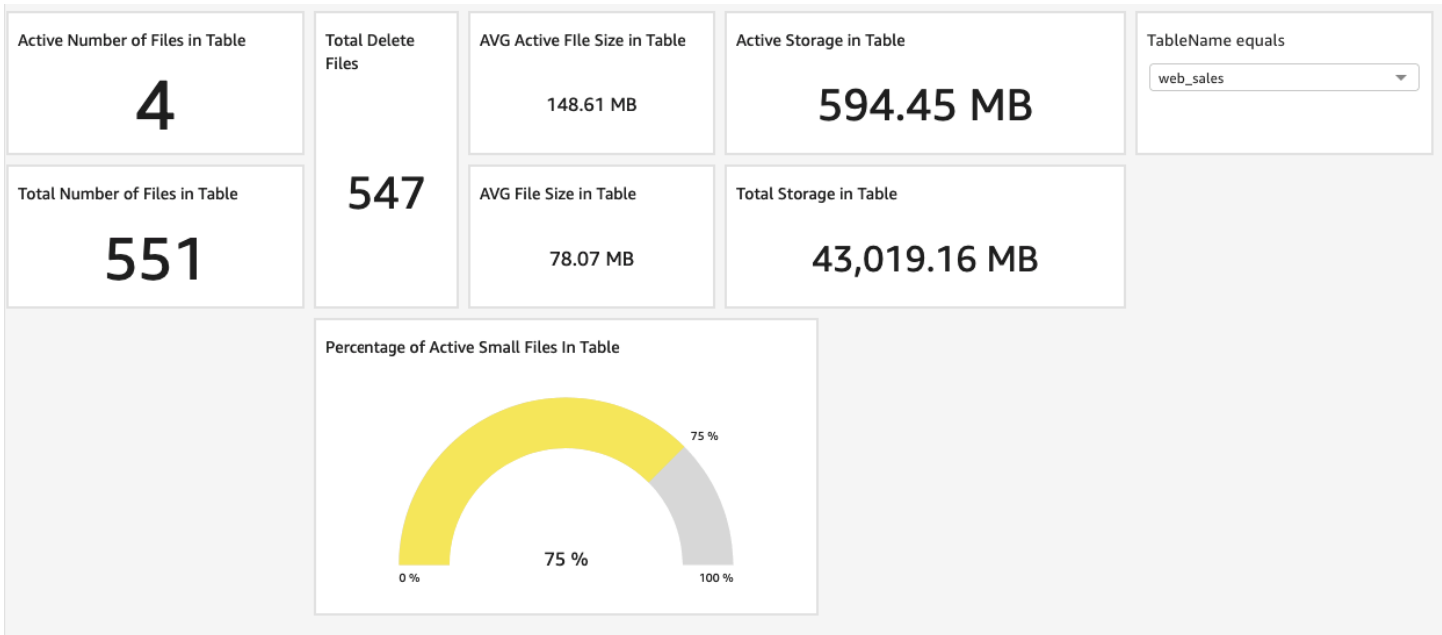
[Para monitorar as cargas de trabalho do Iceberg, você tem duas opções: analisar tabelas de metadados ou usar relatórios de métricas.](#) Os repórteres de métricas foram introduzidos na versão 1.2 do Iceberg e estão disponíveis somente para catálogos REST e JDBC.

Se você estiver usando AWS Glue Data Catalog, poderá obter informações sobre a integridade de suas tabelas do Iceberg configurando o monitoramento sobre as tabelas de metadados que o Iceberg expõe.

O monitoramento é crucial para o gerenciamento de desempenho e a solução de problemas. Por exemplo, quando uma partição em uma tabela do Iceberg atinge uma certa porcentagem de arquivos pequenos, sua carga de trabalho pode iniciar um trabalho de compactação para consolidar os arquivos em arquivos maiores. Isso evita que as consultas fiquem mais lentas além de um nível aceitável.

Monitoramento em nível de tabela

A tela a seguir mostra um painel de monitoramento de tabelas criado na Amazon QuickSight. Esse painel consulta as tabelas de metadados do Iceberg usando o Spark SQL e captura métricas detalhadas, como o número de arquivos ativos e o armazenamento total. Essas informações são então armazenadas em AWS Glue tabelas para fins operacionais. Por fim, um QuickSight painel, conforme mostrado na ilustração a seguir, é criado usando o Amazon Athena. Essas informações ajudam você a identificar e resolver problemas específicos em seus sistemas.



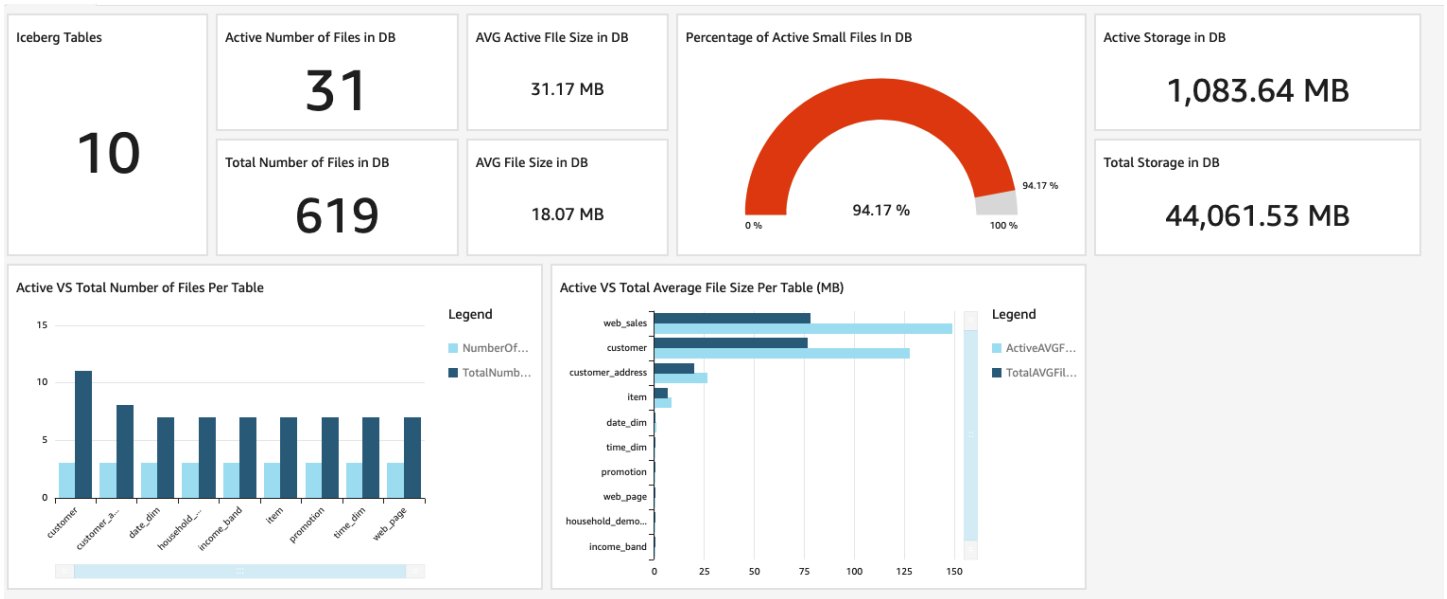
O QuickSight painel de exemplo coleta os seguintes indicadores-chave de desempenho (KPIs) para uma tabela Iceberg:

KPI	Descrição	Consulta
Número de arquivos	O número de arquivos na tabela Iceberg (para todos os instantâneos)	<pre>select count(*) from <catalog.database. table_name>.all_files</pre>
Número de arquivos ativos	O número de arquivos ativos no último instantâneo da tabela Iceberg	<pre>select count(*) from <catalog.database. table_name>.files</pre>
Tamanho médio do arquivo	O tamanho médio do arquivo, em megabytes, para todos os arquivos na tabela Iceberg	<pre>select avg(file_ size_in_bytes)/100 0000 from <catalog.database. table_name>.all_files</pre>
Tamanho médio do arquivo ativo	O tamanho médio do arquivo, em megabytes, dos arquivos ativos na tabela Iceberg	<pre>select avg(file_ size_in_bytes)/100 0000</pre>

KPI	Descrição	Consulta
Porcentagem de arquivos pequenos	A porcentagem de arquivos ativos menores que 100 MB	<pre>from <catalog.database. table_name>.files</pre>
Tamanho total do armazenamento	O tamanho total de todos os arquivos na tabela, excluindo arquivos órfãos e versões de objetos do Amazon S3 (se habilitado)	<pre>select cast(sum(case when file_size _in_bytes < 100000000 then 1 else 0 end)*100/ count(*) as decimal(1 0,2)) from <catalog.database. table_name>.files</pre>
Tamanho total do armazenamento ativo	O tamanho total de todos os arquivos nos instantâneos atuais de uma determinada tabela	<pre>select sum(file_ size_in_bytes)/100 0000 from <catalog.database. table_name>.all_files</pre>
Tamanho total do armazenamento ativo	O tamanho total de todos os arquivos nos instantâneos atuais de uma determinada tabela	<pre>select sum(file_ size_in_bytes)/100 0000 from <catalog.database. table_name>.files</pre>

Monitoramento em nível de banco de dados

O exemplo a seguir mostra um painel de monitoramento que foi criado QuickSight para fornecer uma visão geral dos KPIs em nível de banco de dados para uma coleção de tabelas Iceberg.



Esse painel coleta os seguintes KPIs:

KPI	Descrição	Consulta
Número de arquivos	O número de arquivos no banco de dados do Iceberg (para todos os instantâneos)	Esse painel usa as consultas em nível de tabela fornecidas na seção anterior e consolida os resultados.
Número de arquivos ativos	O número de arquivos ativos no banco de dados do Iceberg (com base nos últimos instantâneos das tabelas do Iceberg)	
Tamanho médio do arquivo	O tamanho médio do arquivo, em megabytes, de todos os arquivos no banco de dados Iceberg	
Tamanho médio do arquivo ativo	O tamanho médio do arquivo, em megabytes, de todos os arquivos ativos no banco de dados Iceberg	

KPI	Descrição	Consulta
Porcentagem de arquivos pequenos	A porcentagem de arquivos ativos menores que 100 MB no banco de dados do Iceberg	
Tamanho total do armazenamento	O tamanho total de todos os arquivos no banco de dados, excluindo arquivos órfãos e versões de objetos do Amazon S3 (se habilitado)	
Tamanho total do armazenamento ativo	O tamanho total de todos os arquivos nos instantâneos atuais de todas as tabelas no banco de dados	

Manutenção preventiva

Ao configurar os recursos de monitoramento discutidos nas seções anteriores, você pode abordar a manutenção da mesa de um ângulo preventivo em vez de reativo. Por exemplo, você pode usar as métricas em nível de tabela e em nível de banco de dados para programar ações como as seguintes:

- Use a compactação de compartimentos para agrupar arquivos pequenos quando uma tabela atingir N arquivos pequenos.
- Use a compactação de compartimentos para mesclar arquivos excluídos quando uma tabela atingir N arquivos excluídos em uma determinada partição.
- Remova arquivos pequenos que já foram compactados removendo instantâneos quando o armazenamento total for X vezes maior que o armazenamento ativo.

Governança e controle de acesso para o Apache Iceberg em AWS

O Apache Iceberg se integra AWS Lake Formation para simplificar a governança de dados. Essa integração permite que os administradores do data lake atribuam permissões de acesso em nível de célula às tabelas do Iceberg. Para ver um exemplo de consulta de tabelas Iceberg usando o Amazon Athena AWS Lake Formation e, consulte a [postagem AWS do blog Interaja com tabelas Iceberg do Apache usando o Amazon Athena e permissões refinadas entre contas](#) usando. AWS Lake Formation

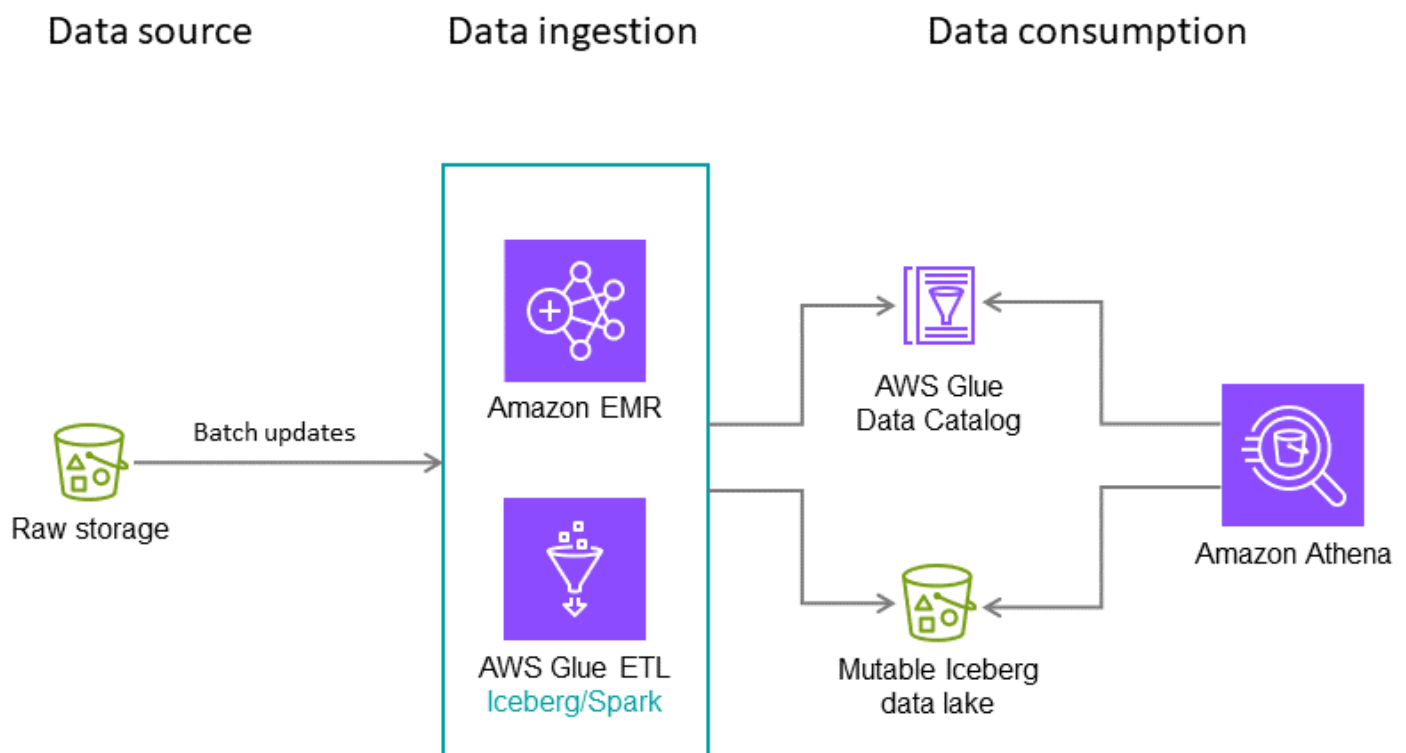
Arquiteturas de referência para o Apache Iceberg em AWS

Esta seção fornece exemplos de como aplicar as melhores práticas em diferentes casos de uso, como ingestão em lote e um data lake que combina ingestão de dados em lote e streaming.

Ingestão noturna em lote

Para esse caso de uso hipotético, digamos que sua mesa Iceberg consuma transações com cartão de crédito todas as noites. Cada lote contém somente atualizações incrementais, que devem ser mescladas na tabela de destino. Várias vezes por ano, dados históricos completos são recebidos. Para esse cenário, recomendamos a arquitetura e as configurações a seguir.

Nota: Este é apenas um exemplo. A configuração ideal depende de seus dados e requisitos.



Recomendações:

- Tamanho do arquivo: 128 MB, porque as tarefas do Apache Spark processam dados em blocos de 128 MB.
- Tipo de escrita: copy-on-write. Conforme detalhado anteriormente neste guia, essa abordagem ajuda a garantir que os dados sejam gravados de forma otimizada para leitura.

- Variáveis de partição: ano/mês/dia. Em nosso caso de uso hipotético, consultamos dados recentes com mais frequência, embora ocasionalmente executemos varreduras completas de dados nos últimos dois anos. O objetivo do particionamento é conduzir operações de leitura rápida com base nos requisitos do caso de uso.
- Ordem de classificação: carimbo de data/hora
- Catálogo de dados: AWS Glue Data Catalog

Lago de dados que combina ingestão em lote e quase em tempo real

Você pode provisionar um data lake no Amazon S3 que compartilha dados em lote e de streaming entre contas e regiões. Para obter um diagrama de arquitetura e detalhes, consulte a postagem do AWS blog [Crie um data lake transacional usando o Apache Iceberg e compartilhamentos de dados entre contas](#) usando o Amazon Athena. AWS Glue AWS Lake Formation

Recursos

- [Usando a estrutura Iceberg em AWS Glue](#) (AWS Glue documentação)
- [Iceberg](#) (documentação do Amazon EMR)
- [Usando tabelas do Apache Iceberg \(documentação\)](#) do Amazon Athena)
- [Documentação do Amazon S3](#)
- [QuickSight Documentação da Amazon](#)
- [Replicação do Glue Catalog e do Lake Formation Permissions](#) (GitHub repositório)
- [Documentação do Apache Iceberg](#)
- [Documentação do Apache Spark](#)

Colaboradores

As seguintes pessoas escreveram, foram AWS coautoras e revisaram este guia.

Colaboradores

- Carlos Rodrigues, arquiteto de soluções, Big Data
- Imtiaz (Taz) Sayed, arquiteto de soluções e líder técnico de análise
- Shana Schipers, arquiteta de soluções, Big Data
- Prashant Singh, engenheiro de desenvolvimento de software, Amazon EMR
- Stefano Sandona, arquiteto de soluções, Big Data
- Arun A K, arquiteto de soluções, Big Data e ETL
- Francisco Morillo, arquiteto de soluções, streaming
- Suthan Phillips, arquiteto de análise, Amazon EMR
- Sercan Karaoglu, arquiteto de soluções
- Yonatan Dolan, especialista em análise
- Guy Bachar, arquiteto de soluções
- Sofia Zilberman, arquiteta de soluções de streaming
- Ismail Makhoul, arquiteto de soluções, análise
- Dan Stair, arquiteto especializado em soluções
- Sakti Mishra, arquiteta de soluções

Revisores

- Rick Sears, gerente geral, Amazon EMR
- Linda OConnor, especialista em Amazon EMR
- Ian Meyers, diretor do Amazon EMR
- Vinita Ananth, diretora de gerenciamento de produtos do Amazon EMR
- Jason Berkowitz, gerente de produto, AWS Lake Formation
- Mahesh Mishra, gerente de produto, Amazon Redshift
- Vladimir Zlatkin, gerente de arquitetura de soluções, Big Data
- Karthik Prabhakar, arquiteto de análise, Amazon EMR

- Jack Ye, engenheiro de desenvolvimento de software, Amazon EMR
- Vijay Jain, gerente de produto
- Anupriti Warade, gerente de produto, Amazon S3
- Molly Brown, gerente geral, AWS Lake Formation
- Ajit Tandale, arquiteto de soluções, dados
- Gwen Chen, gerente de marketing de produto

Histórico do documento

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

Alteração	Descrição	Data
Publicação inicial	—	30 de abril de 2024

AWS Glossário de orientação prescritiva

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

Números

7 Rs

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

- Refatorar/rearquitetar: mova uma aplicação e modifique sua arquitetura aproveitando ao máximo os recursos nativos de nuvem para melhorar a agilidade, a performance e a escalabilidade. Isso normalmente envolve a portabilidade do sistema operacional e do banco de dados. Exemplo: migre seu banco de dados Oracle local para a edição compatível com o Amazon Aurora PostgreSQL.
- Redefinir a plataforma (mover e redefinir [mover e redefinir (lift-and-reshape)]): mova uma aplicação para a nuvem e introduza algum nível de otimização a fim de aproveitar os recursos da nuvem. Exemplo: Migre seu banco de dados Oracle local para o Amazon Relational Database Service (Amazon RDS) for Oracle no. Nuvem AWS
- Recomprar (drop and shop): mude para um produto diferente, normalmente migrando de uma licença tradicional para um modelo SaaS. Exemplo: migre seu sistema de gerenciamento de relacionamento com o cliente (CRM) para a Salesforce.com.
- Redefinir a hospedagem (mover sem alterações [lift-and-shift])mover uma aplicação para a nuvem sem fazer nenhuma alteração a fim de aproveitar os recursos da nuvem. Exemplo: Migre seu banco de dados Oracle local para o Oracle em uma instância do EC2 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 [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 as AIOps são usadas na estratégia de migração para a AWS, consulte o [guia de integração de operações](#).

criptografia assimétrica

Um algoritmo de criptografia que usa um par de chaves, uma chave pública para criptografia e uma chave privada para descryptografia. É possível compartilhar a chave pública porque ela não é usada na descryptografia, 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.

Availability Zone (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 containerizados 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 Serviço da AWS receba.

Centro de Excelência da Nuvem (CCoE)

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

computação em nuvem

A tecnologia de nuvem normalmente usada para armazenamento de dados remoto e gerenciamento de dispositivos de IoT. A computação em nuvem 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: realizar investimentos fundamentais para escalar sua adoção da nuvem (por exemplo, criar uma zona de pouso, definir um CCoE, estabelecer um modelo de operações)
- Migração: migrar aplicações individuais
- 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 ou AWS CodeCommit. 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, AWS Panorama oferece dispositivos que adicionam CV às redes de câmeras locais, e a Amazon SageMaker 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. O CI/CD é comumente descrito como um pipeline. O CI/CD pode ajudar você 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

Consulte [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](#).

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.

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).

environment (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

Consulte [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 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.

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.

G

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.

estratégia greenfield

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

barreira de proteção

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

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.

I

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 segurança da AWS](#) recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

migração incremental

Uma estratégia de substituição na qual você migra a aplicação em pequenas partes, em vez de realizar uma única substituição completa. Por exemplo, é possível mover inicialmente

I

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 das Coisas Industrial (IIoT)

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

VPC de inspeção

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

Internet das Coisas (IoT)

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

interpretabilidade

Uma característica de um modelo de machine learning que descreve o grau em que um ser humano pode entender como as previsões do modelo dependem de suas entradas. Para obter mais informações, consulte [Interpretabilidade do modelo de machine learning com a 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](#).

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](#).

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, vaziar 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-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 por meio de APIs bem definidas e normalmente pertence a equipes pequenas e autônomas. Por exemplo, um sistema de seguradora pode incluir microsserviços que mapeiam as capacidades comerciais, como vendas ou marketing, ou subdomínios, como compras, reclamações ou análises. Os benefícios dos

microserviços incluem agilidade, escalabilidade flexível, fácil implantação, código reutilizável e resiliência. Para obter mais informações, consulte [Integração de microserviços usando serviços sem AWS servidor](#).

arquitetura de microserviços

Uma abordagem à criação de aplicações com componentes independentes que executam cada processo de aplicação como um microserviço. Esses microserviços se comunicam por meio de uma interface bem definida usando APIs leves. Cada microserviço 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 microserviços 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: rehoste a migração para o Amazon EC2 AWS com o Application Migration Service.

Avaliação de Portfólio para Migração (MPA)

Uma ferramenta on-line que fornece informações para validar o caso de negócios para migrar para 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.

OU

Veja a [análise de prontidão operacional](#).

NÃO

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 segurança da AWS](#) recomenda configurar sua conta de rede com VPCs de entrada, saída e inspeção para proteger a interface bidirecional entre a aplicação e a Internet em geral.

P

limite de permissões

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

Informações de identificação pessoal (PII)

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

PII

Veja [informações de identificação pessoal](#).

manual

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

PLC

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

zonas hospedadas privadas

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

controle proativo

Um [controle de segurança](#) 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.

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.

publicar/assinar (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 \(RACI\)](#).

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](#).

rotação

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.

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 do Amazon EC2 ou a rotação de credenciais.

Criptografia do lado do servidor

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

política de controle de serviços (SCP)

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

service endpoint (endpoint de serviço)

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

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 pode ser usado para interconectar as VPCs e as redes on-premises. Para obter mais informações, consulte [O que é um gateway de trânsito](#) na AWS Transit Gateway documentação.

fluxo de trabalho baseado em troncos

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

Acesso confiável

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

tuning (ajustar)

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

equipe de duas pizzas

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

U

incerteza

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

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

Vulnerabilidade

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

W

cache quente

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

dados mornos

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

função de janela

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

workload

Uma coleção de códigos e recursos que geram valor empresarial, como uma aplicação voltada para o cliente ou um processo de 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 o [AWS Workload Qualification Framework](#).

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.

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.