Semana 6

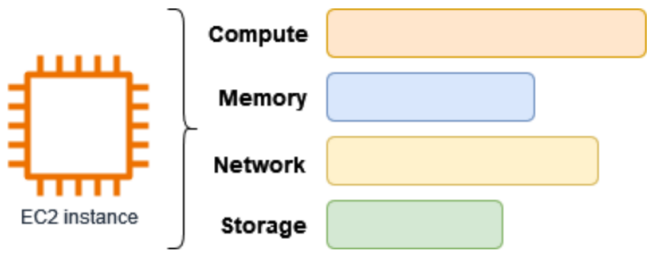
EC2

O que é o Amazon EC2?

PDF | RSS

O Amazon Elastic Compute Cloud (Amazon EC2) oferece uma capacidade de computação escalável sob demanda na Nuvem Amazon Web Services (AWS). O uso do Amazon EC2 reduz os custos de hardware para que você possa desenvolver e implantar aplicações com mais rapidez. É possível usar o Amazon EC2 para executar quantos servidores virtuais forem necessários, configurar a segurança e as redes e gerenciar o armazenamento. Você pode adicionar capacidade (aumentar a escala verticalmente) para lidar com tarefas de computação pesada, como processos mensais ou anuais ou picos no tráfego do site. Quando o uso diminui, você pode reduzir a capacidade (reduzir a escala verticalmente) de novo.

Uma instância do EC2 é um servidor virtual na Nuvem AWS. Quando executa uma instância do EC2, o tipo de instância que você especifica determina o hardware disponível para sua instância. Cada tipo de instância oferece um equilíbrio diferente entre recursos de computação, memória, armazenamento e rede. Para obter mais informações, consulte o Guia de tipos de instância do Amazon EC2.

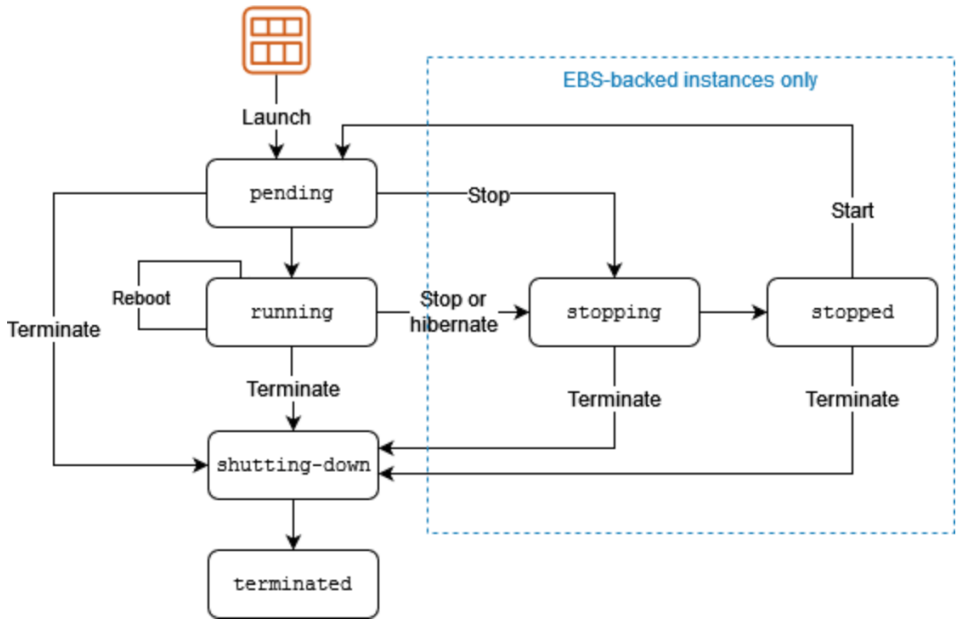


Ciclo de vida da instância

PDF | RSS

Uma instância do Amazon EC2 passa por diferentes estados do momento em que você a inicia até seu encerramento.

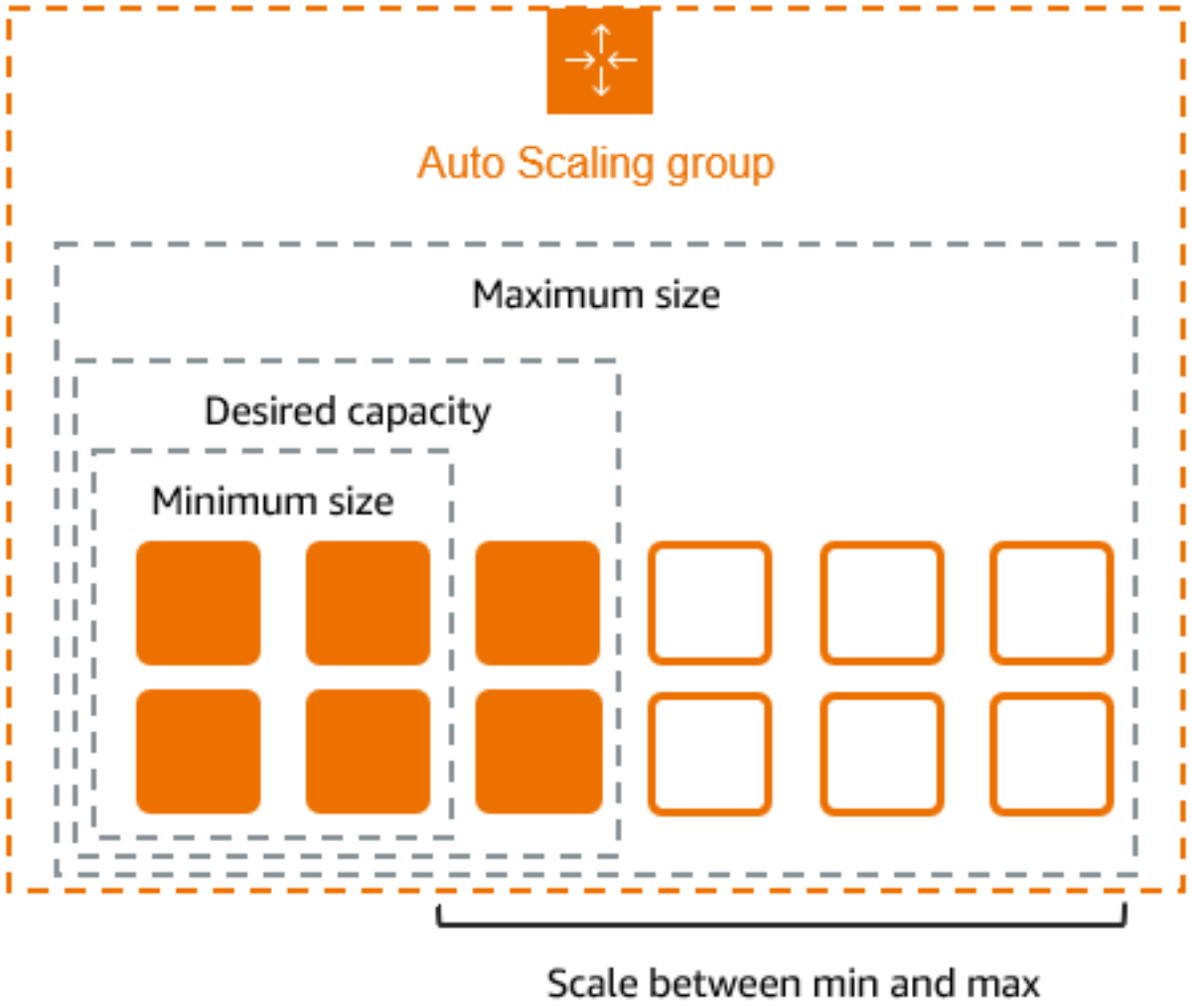
A ilustração a seguir representa as transições entre os estados da instância. Observe que você não pode parar e o iniciar uma instância com armazenamento de instâncias. Para obter mais informações sobre instâncias baseadas em armazenamento de instâncias, consulte Armazenamento para o dispositivo raiz.



Auto Scaling Group

Destaque para os conceitos:

- Desired capacity
- Minimum Size
- Maximum Size



Opções de compra

Opções de compra de instância

PDF | RSS

O Amazon EC2 fornece as seguintes opções de compra para permitir otimizar os custos com base em suas necessidades:

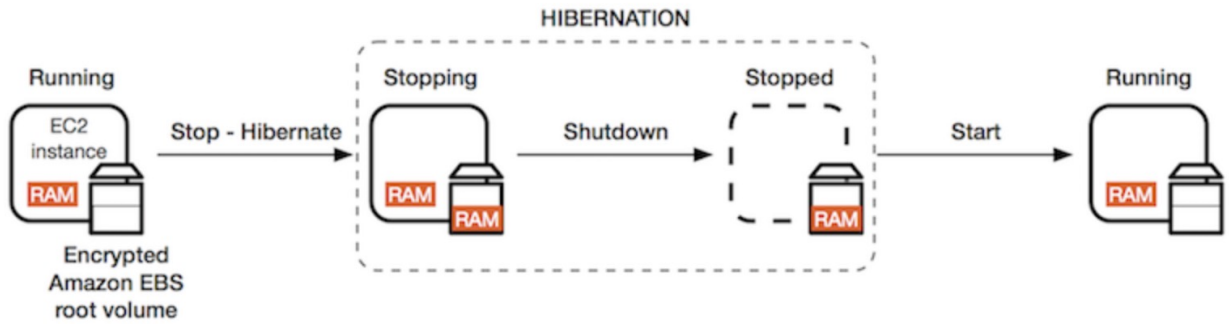
- **Instâncias sob demanda**: pague pelas instâncias que você iniciar
- **Savings Plans**: reduza os custos do Amazon EC2 se comprometendo com uma quantidade consistente de uso, em USD por hora, por um período de vigência de um ou de três anos.
- **Reserved Instances (Instâncias reservadas)**: reduza os custos do Amazon EC2 se comprometendo com uma configuração consistente de instância, incluindo o tipo de instância e a região, por um período de vigência de um ou de três anos.
- **Spot Instances (Instâncias spot)**: solicite instâncias do EC2 não utilizadas, o que pode reduzir os custos do Amazon EC2 significativamente.
- **Dedicated Hosts (Hosts dedicados)**: pague por um host físico que seja totalmente dedicado à execução de suas instâncias e traga suas licenças de software existentes por soquete, por núcleo ou por VM para reduzir custos.
- **Dedicated Instances (Instâncias dedicadas)**: pague por hora pelas instâncias que são executadas no hardware de um ocupante único.
- **Reservas de capacidade**: reserve capacidade para as instâncias do EC2 em uma zona de disponibilidade específica.

Hibernação EC2

Como a hibernação de instâncias do Amazon EC2 funciona

PDF | RSS

O diagrama a seguir mostra uma visão geral básica do processo de hibernação para instâncias do EC2.



- O instância entrará no estado `stopping`. O Amazon EC2 sinaliza o sistema operacional para realizar a hibernação (suspend-to-disk). A hibernação congela todos os processos, salva o conteúdo da RAM no volume raiz do EBS e, depois, executa um desligamento normal.
- Quando o desligamento é concluído, a instância muda para o estado `stopped`.
- Todos os volumes do EBS permanecem anexados à instância, e seus dados são mantidos, incluindo o conteúdo salvo da RAM.

Hibernação EC2

Pré-requisitos para a hibernação de instâncias do Amazon EC2

PDF | RSS

É possível habilitar o suporte à hibernação para uma instância sob demanda ou uma instância spot ao iniciá-la. Não é possível habilitar a hibernação em uma instância existente, esteja ela em execução ou parada. Para ter mais informações, consulte Habilitar hibernação da instância.

“Não é possível habilitar a hibernação em uma instância existente, esteja ela em execução ou parada.”

S3

O que é o Amazon S3?

PDF | RSS

O Amazon Simple Storage Service (Amazon S3) é um serviço de armazenamento de objetos que oferece escalabilidade líder do setor, disponibilidade de dados, segurança e performance. Clientes de todos os tamanhos e setores podem usar o Amazon S3 para armazenar e proteger qualquer volume de dados para uma variedade de casos de uso, como data lakes, sites, aplicações móveis, backup e restauração, arquivamento, aplicações corporativas, dispositivos IoT e análises de big data. O Amazon S3 fornece recursos de gerenciamento para que você possa otimizar, organizar e configurar o acesso aos seus dados para atender aos seus requisitos específicos de negócios, organizacionais e de compatibilidade.

S3


Analytics data
Log files
Application data
Videos and pictures
Backup and archival
Move data
Move your data to Amazon S3 from wherever it lives – in the cloud, in applications, or on-premises
Amazon S3
Object storage built to store and retrieve any amount of data from anywhere
Store data
Create bucket, specify the Region, access controls, and management options. Upload any amount of data
Control access to data
Optimize cost with storage classes
Replicate data to any Region
Access from on-premises or VPC
Protect and secure your data
Gain visibility into your storage
Artificial Intelligence (AI)
Advanced analytics
Machine Learning (ML)
Analyze data
Use AWS and 3rd party services to analyze your data to gain insights

S3

Parallel processing with "Fan Out" architecture

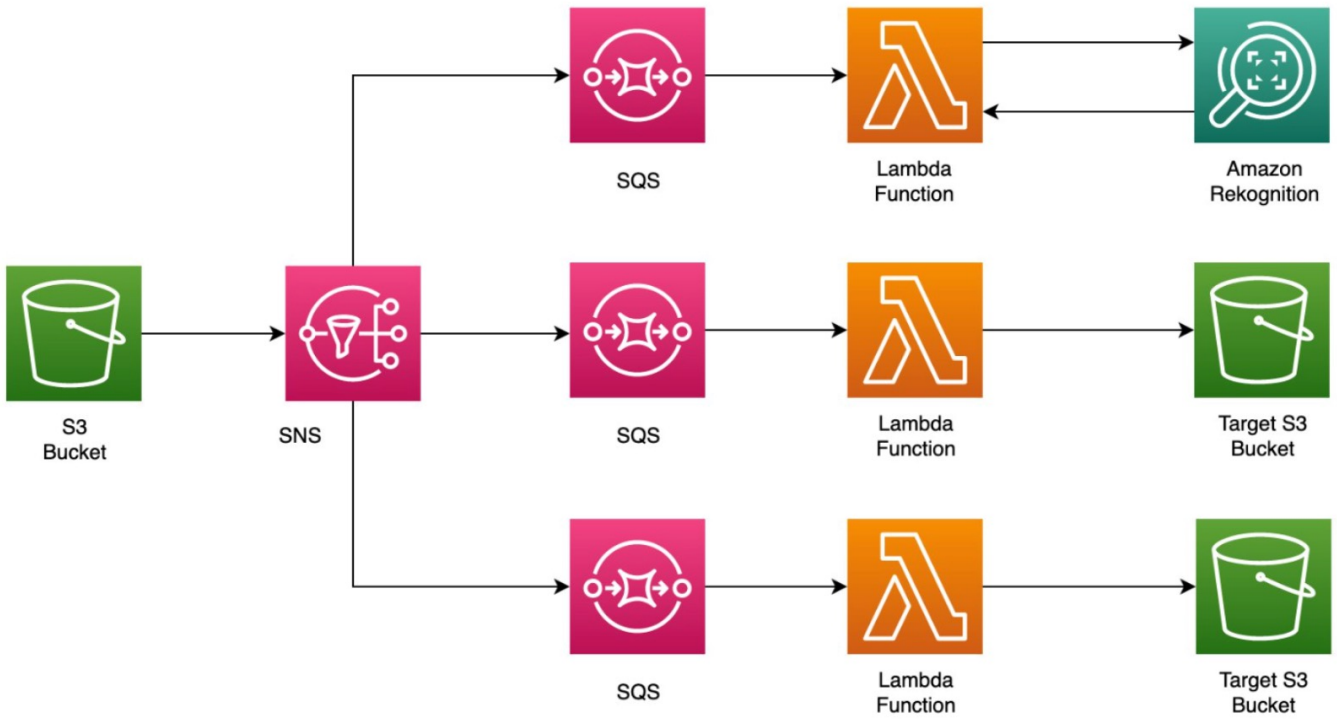
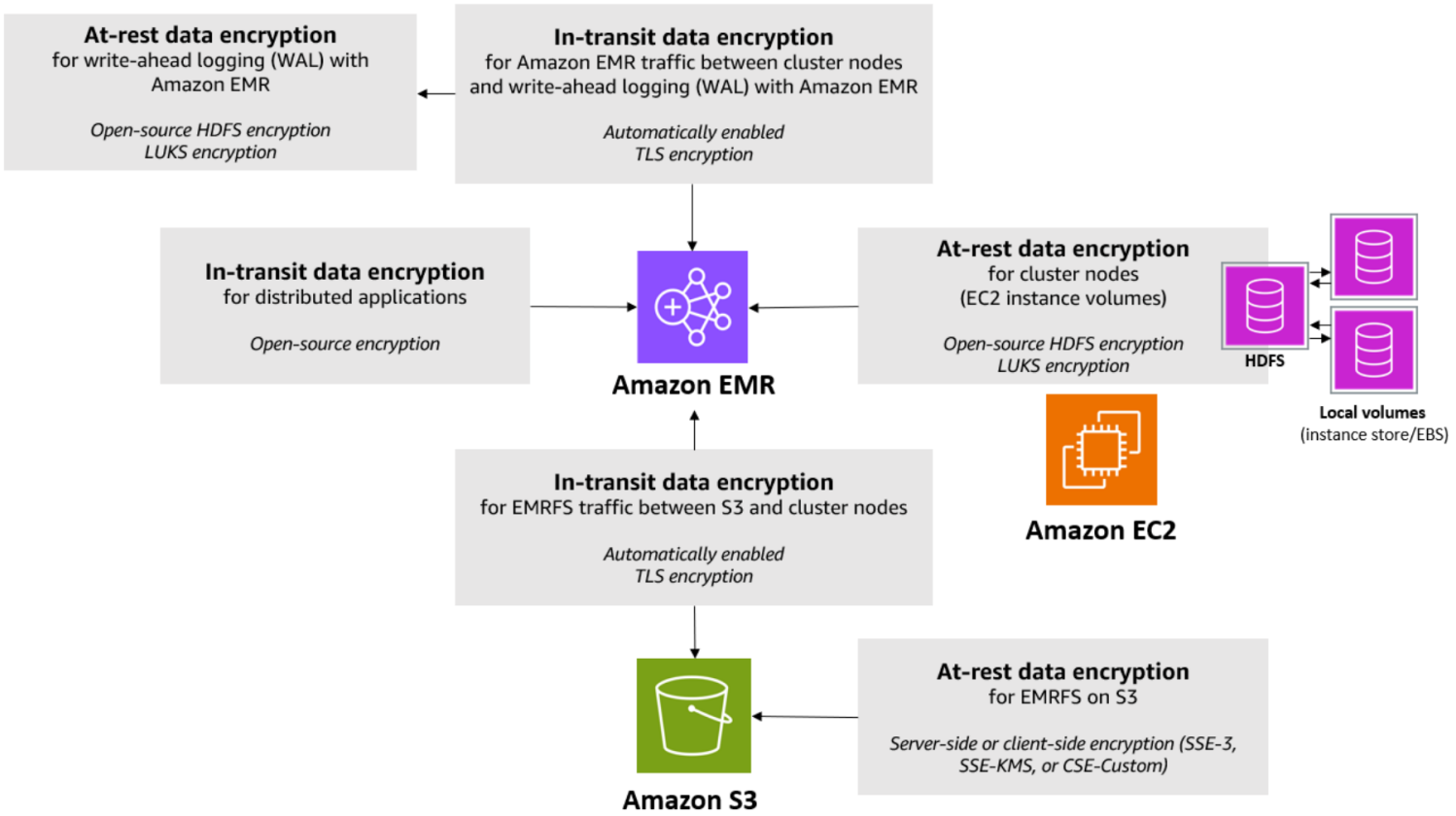


Figure 6. Fan out design pattern with S3, SNS, and SQS before sending to a Lambda function

https://aws.amazon.com/blogs/architecture/get-started-with-amazon-s3-event-driven-design-patterns/

S3 – Criptografia em repouso

O diagrama a seguir mostra as diferentes opções de criptografia de dados disponíveis com as configurações de segurança.



https://docs.aws.amazon.com/pt_br/emr/latest/ManagementGuide/emr-data-encryption-options.html

S3 – Criptografia em repouso

Proteger dados com criptografia

PDF | RSS

Importante

O Amazon S3 agora aplica criptografia do lado do servidor com chaves gerenciadas do Amazon S3 (SSE-S3) como nível básico de criptografia para cada bucket no Amazon S3. Desde 5 de janeiro de 2023, todos os novos uploads de objetos para o Amazon S3 são automaticamente criptografados sem custo adicional e sem impacto na performance. O status de criptografia automática para a configuração de criptografia padrão do bucket do S3 e para novos uploads de objetos está disponível em logs do AWS CloudTrail, no Inventário do S3, na Lente de Armazenamento do S3, no console do Amazon S3 e como cabeçalho adicional de resposta da API do Amazon S3 na AWS Command Line Interface e em AWS SDKs. Para obter mais informações, consulte Perguntas frequentes sobre criptografia padrão.

S3 – Criptografia em repouso

Se você precisar que os uploads de dados sejam criptografados usando somente chaves gerenciadas pelo Amazon S3, poderá usar a política de bucket a seguir. Por exemplo, a política de bucket a seguir negará permissões para fazer upload de um objeto, a menos que a solicitação não inclua o cabeçalho `x-amz-server-side-encryption` a fim de solicitar criptografia no lado do servidor:

```
{
  "Version": "2012-10-17",
  "Id": "PutObjectPolicy",
  "Statement": [
    {
      "Sid": "DenyObjectsThatAreNotSSES3",
      "Effect": "Deny",
      "Principal": "*",
      "Action": "s3:PutObject",
      "Resource": "arn:aws:s3:::amzn-s3-demo-bucket/*",
      "Condition": {
        "StringNotEquals": {
          "s3:x-amz-server-side-encryption": "AES256"
        }
      }
    }
  ]
}
```

https://docs.aws.amazon.com/pt_br/AmazonS3/latest/userguide/UsingServerSideEncryption.html

S3 – Criptografia em trânsito

In the following Amazon S3 bucket policy example, you can see how to enforce a bucket to only accept requests of GetObject if the connection is using TLS version 1.2 or higher.

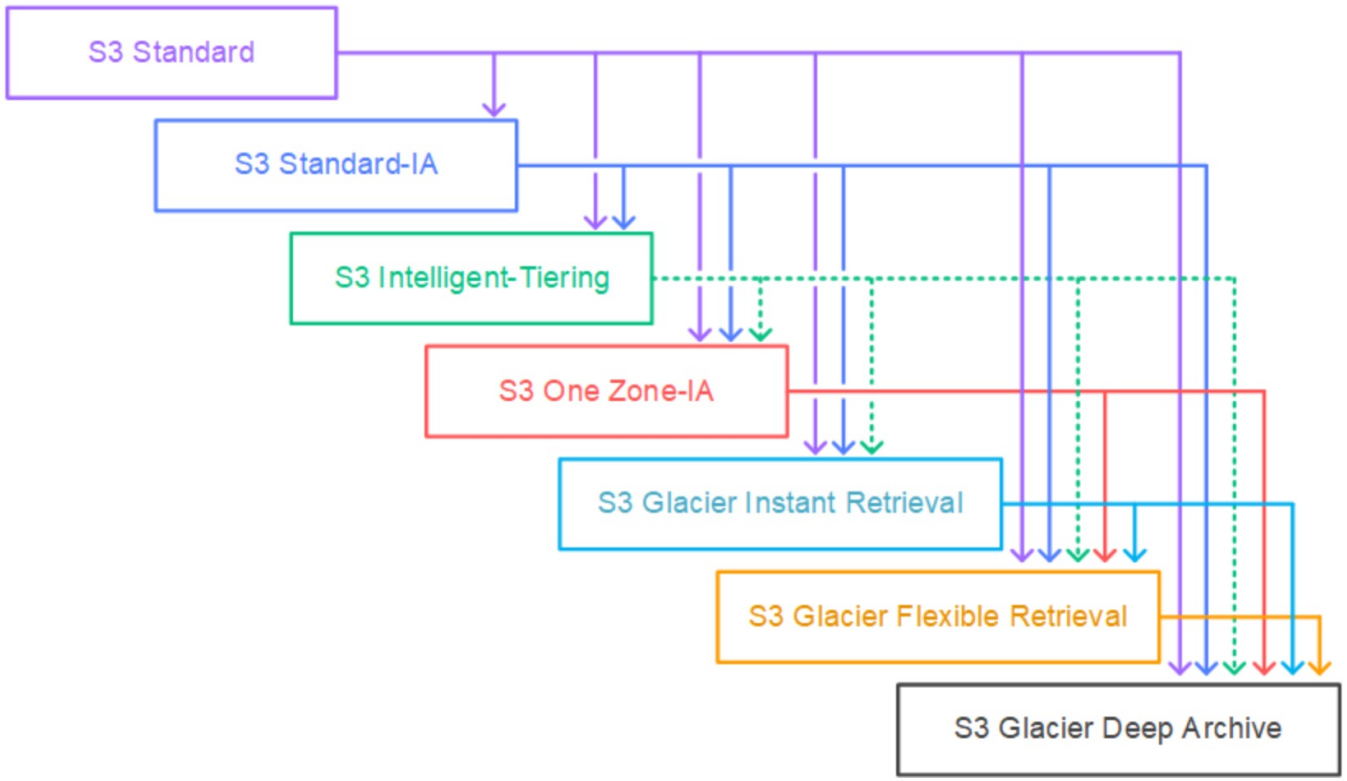
JSON

```
 {
      "Version": "2012-10-17",
      "Statement": [
          {
               "Effect": "Deny",
               "Principal": "*",
               "Action": "s3:*",
               "Resource": "arn:aws:s3::awsexamplebucket1/*",
               "Condition": {
                   "Bool": {
                       "aws:SecureTransport": "true"
                    },
                   "NumericLessThan": {
                       "s3:TlsVersion": [
                           "1.2"
                       ]
                   }
               }
          }
```

https://aws.amazon.com/blogs/storage/enforcing-encryption-in-transit-with-tls1-2-or-higher-with-amazon-s3/

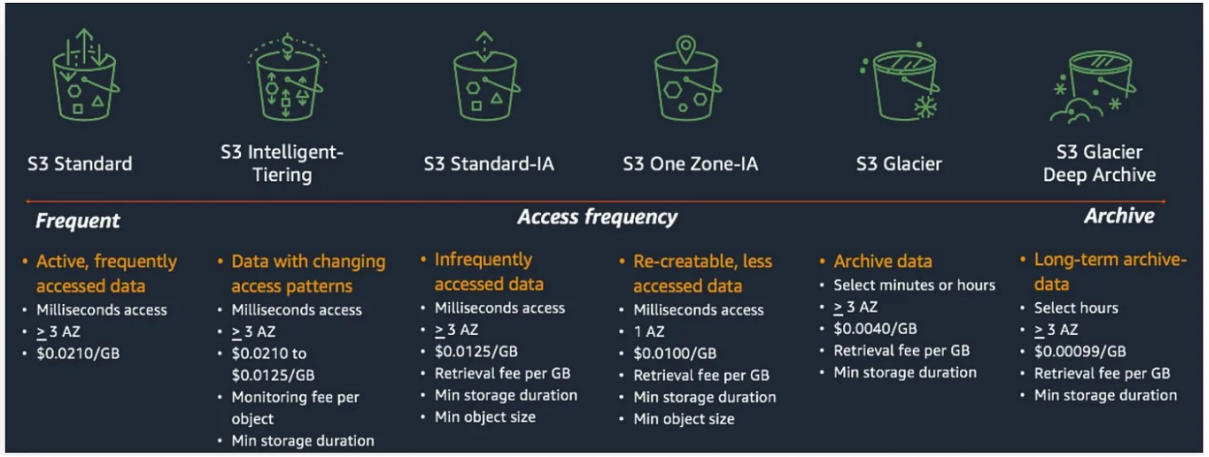
S3 – Lifecycle

O Amazon S3 da suporte a um modelo de cachoeira para fazer a transição entre classes de armazenamento, conforme mostrado no diagrama a seguir.

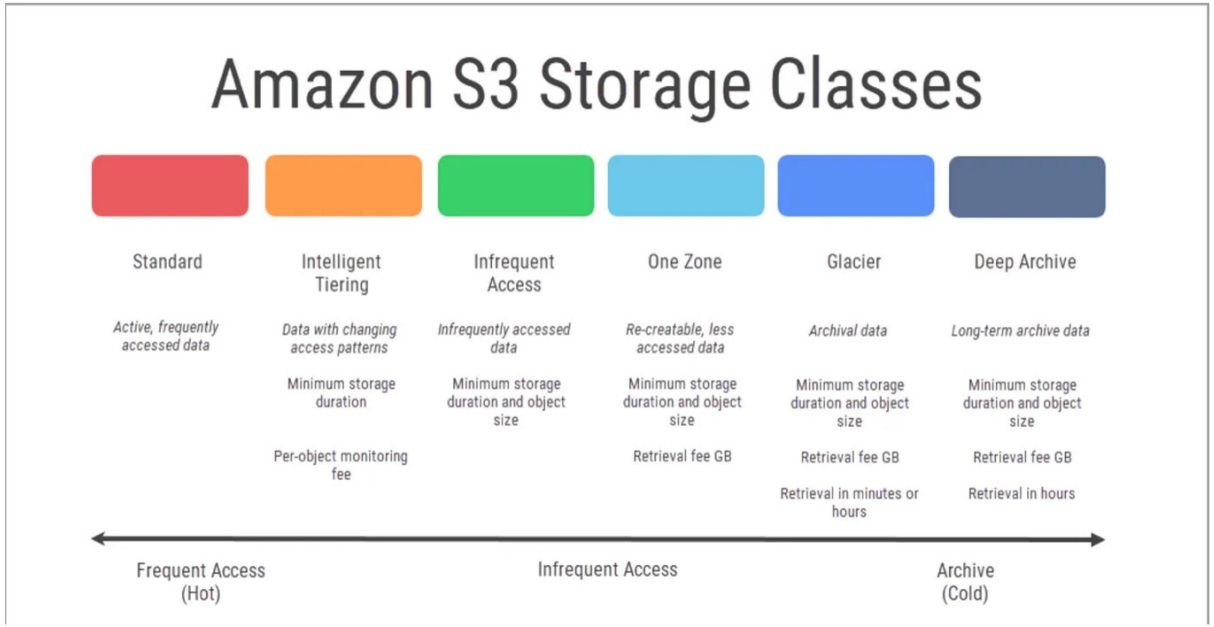


Quadro comparativo

Referência:
https://medium.com/@suryateja233/s3-storage-classes-4cc65b5f55c4

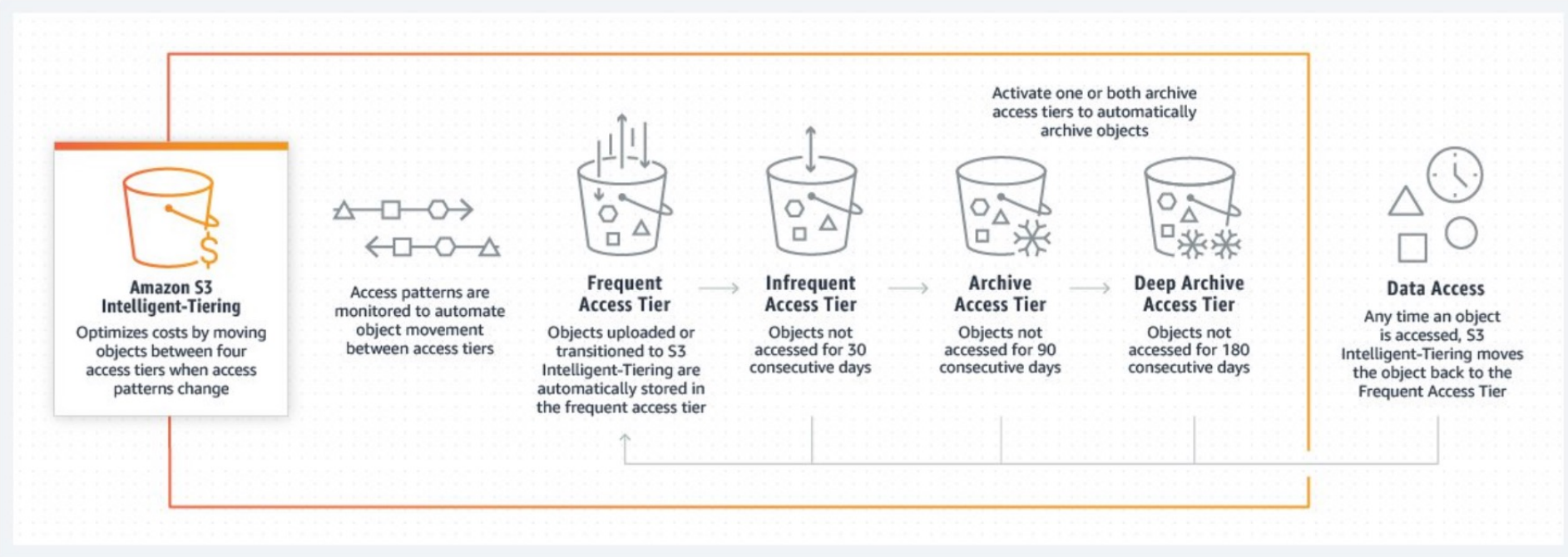


Source: Amazon AWS

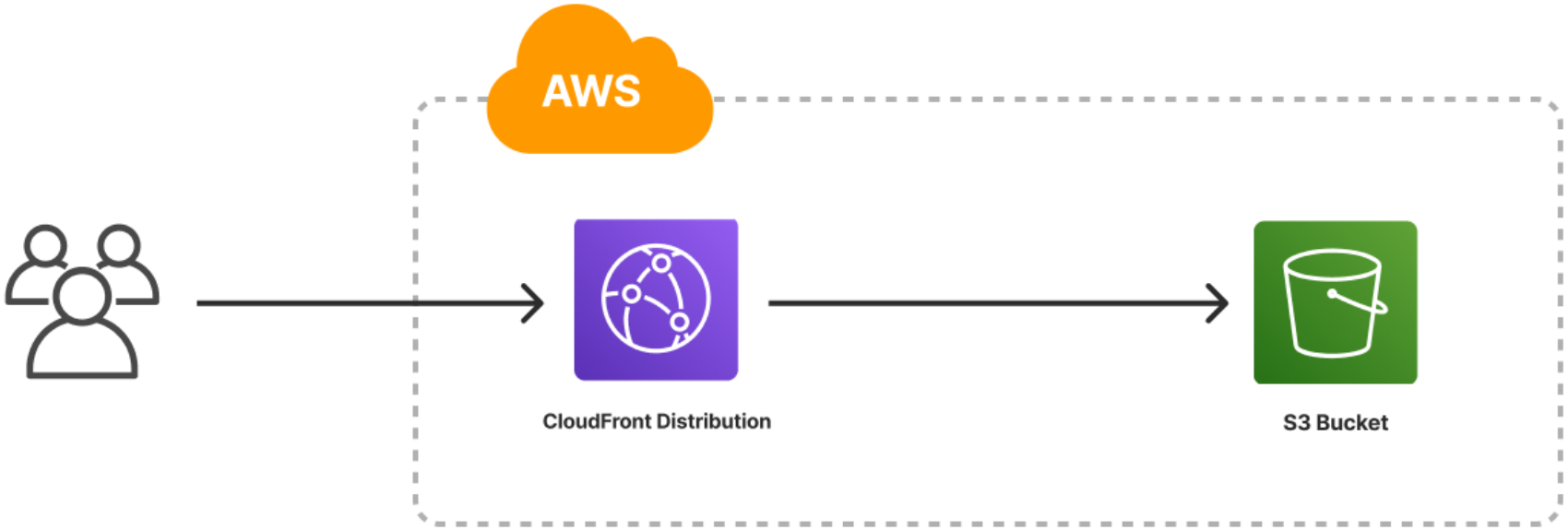


Source: Google

S3 – Intelligent-Tiering



S3 – Hospedando sites estáticos



https://www.pulumi.com/templates/static-website/aws/

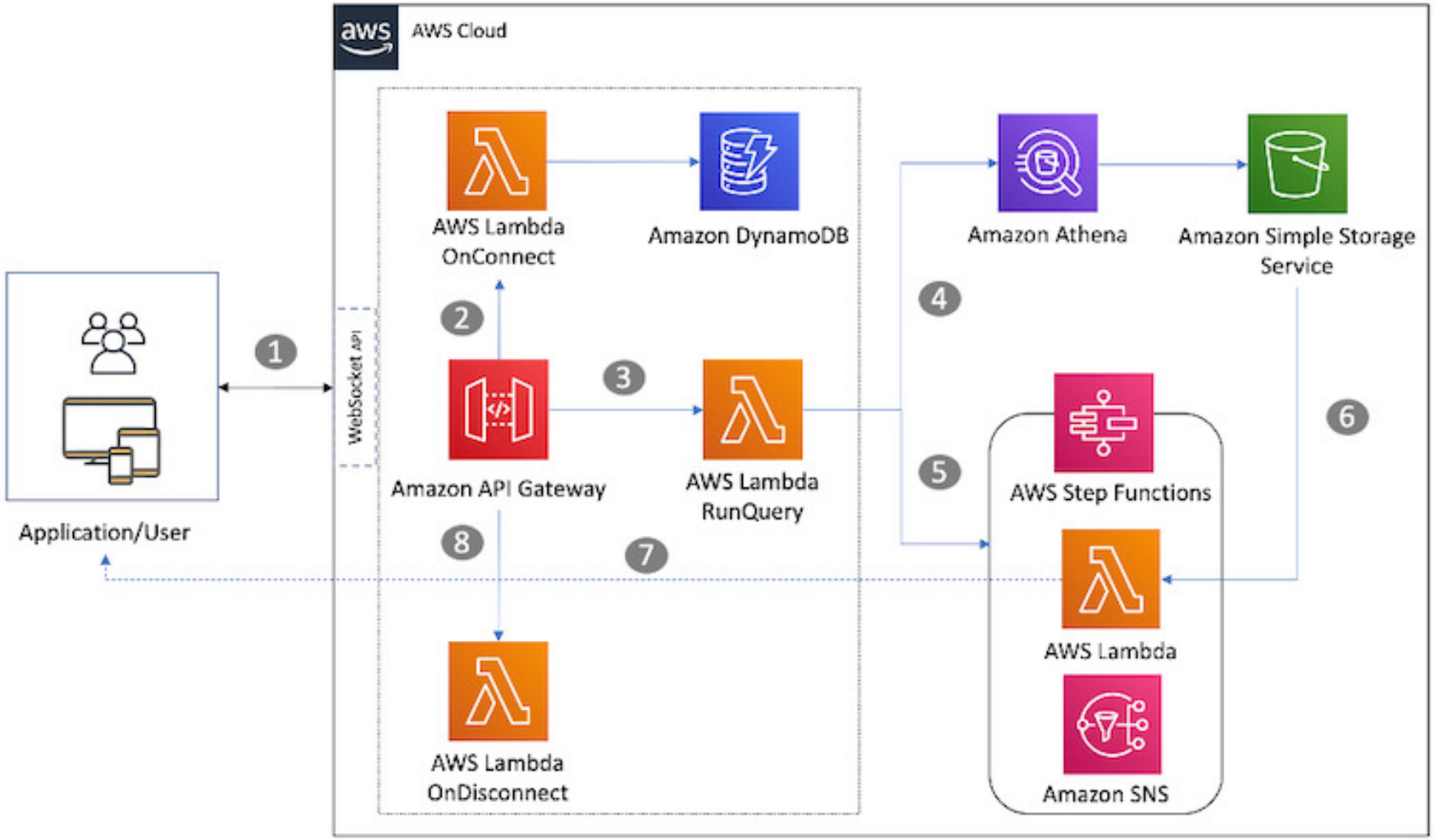
S3 – Multipart Upload

Carregar e copiar objetos usando multipart upload

PDF | RSS

O multipart upload permite que você faça upload de um único objeto como um conjunto de partes. Cada parte é uma parte contígua de dados do objeto. O upload dessas partes de objetos pode ser feito de maneira independente e em qualquer ordem. Se a transmissão de alguma parte falhar, você poderá retransmitir essa parte sem afetar outras partes. Depois que todas as partes do objeto forem carregadas, o Amazon S3 montará essas partes e criará o objeto. Geralmente, quando seu objeto alcança 100 MB de tamanho, você deve considerar o uso de multipart uploads em vez de fazer upload do objeto em uma única operação.

Athena


aws
AWS Cloud
AWS Lambda
OnConnect
Amazon DynamoDB
Amazon Athena
Amazon Simple Storage
Service
WebSocket API
1
2
3
4
5
6
7
8
Amazon API Gateway
AWS Lambda
RunQuery
AWS Step Functions
Application/User
AWS Lambda
Amazon SNS
AWS Lambda
OnDisconnect

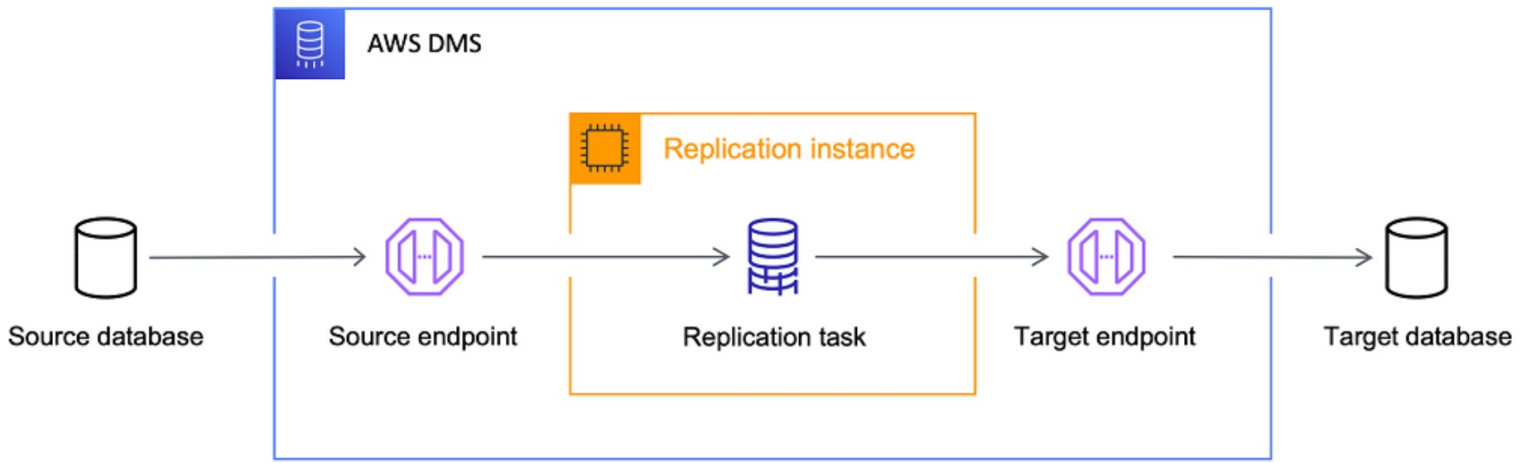
Athena

Cuidado para não confundir com o RedShift. O Redshift é um banco de dados que permite utilizar sintaxe SQL para executar consultas. O Athena é uma ferramenta de consulta ao S3 utilizando sintaxe SQL (**não é banco de dados**).

DMS

Em um nível básico, o AWS DMS é um servidor na Nuvem AWS que executa software de replicação. Você cria uma conexão de origem e de destino para informar ao AWS DMS de onde extrair e para onde carregar. E programa uma tarefa que é executada nesse servidor para mover os dados. O AWS DMS criará as tabelas e as chaves primárias associadas se ainda não existirem no destino. É possível criar as tabelas de destino manualmente, se preferir. Ou utilizar o AWS Schema Conversion Tool (AWS SCT) para criar algumas ou todas as tabelas, índices, visualizações, acionadores e assim por diante de destino.

O diagrama a seguir ilustra o processo de replicação do AWS DMS.



DynamoDB

O que é o Amazon DynamoDB?

PDF | RSS

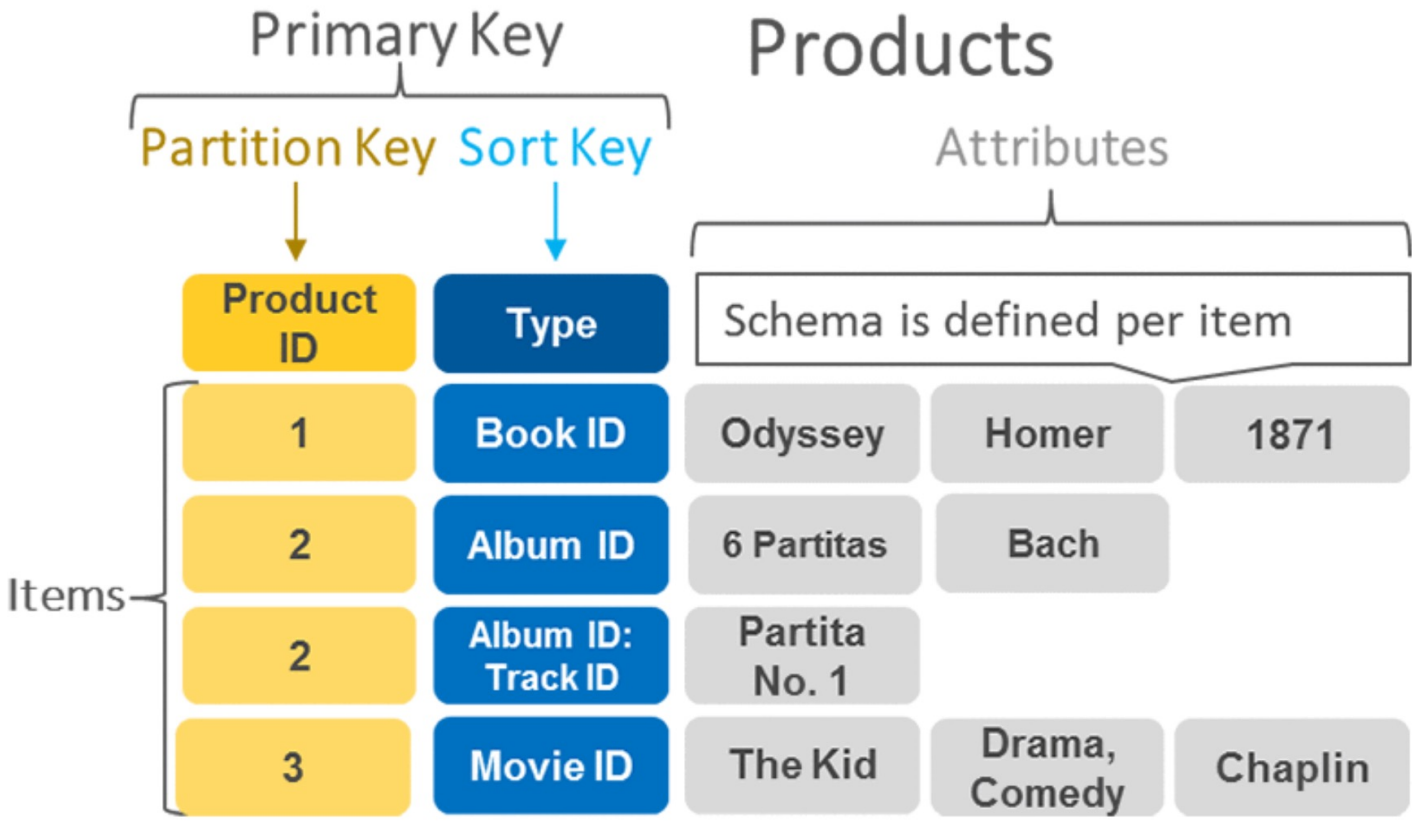
O Amazon DynamoDB é um banco de dados sem servidor, NoSQL e totalmente gerenciado com desempenho de latência inferior a dez milissegundos em qualquer escala.

O DynamoDB atende às suas necessidades para superar as complexidades operacionais e de escalabilidade dos bancos de dados relacionais. O DynamoDB tem propósito específico e é otimizado para workloads operacionais que exigem desempenho consistente em qualquer escala. Por exemplo, o DynamoDB oferece desempenho consistente com latência inferior a dez milissegundos para um caso de uso de carrinho de compras, independentemente de você ter dez ou cem milhões de usuários. Lançado em 2012, o DynamoDB continua ajudando você a abandonar os bancos de dados relacionais e, ao mesmo tempo, reduzir os custos e melhorar o desempenho em grande escala.

DynamoDB

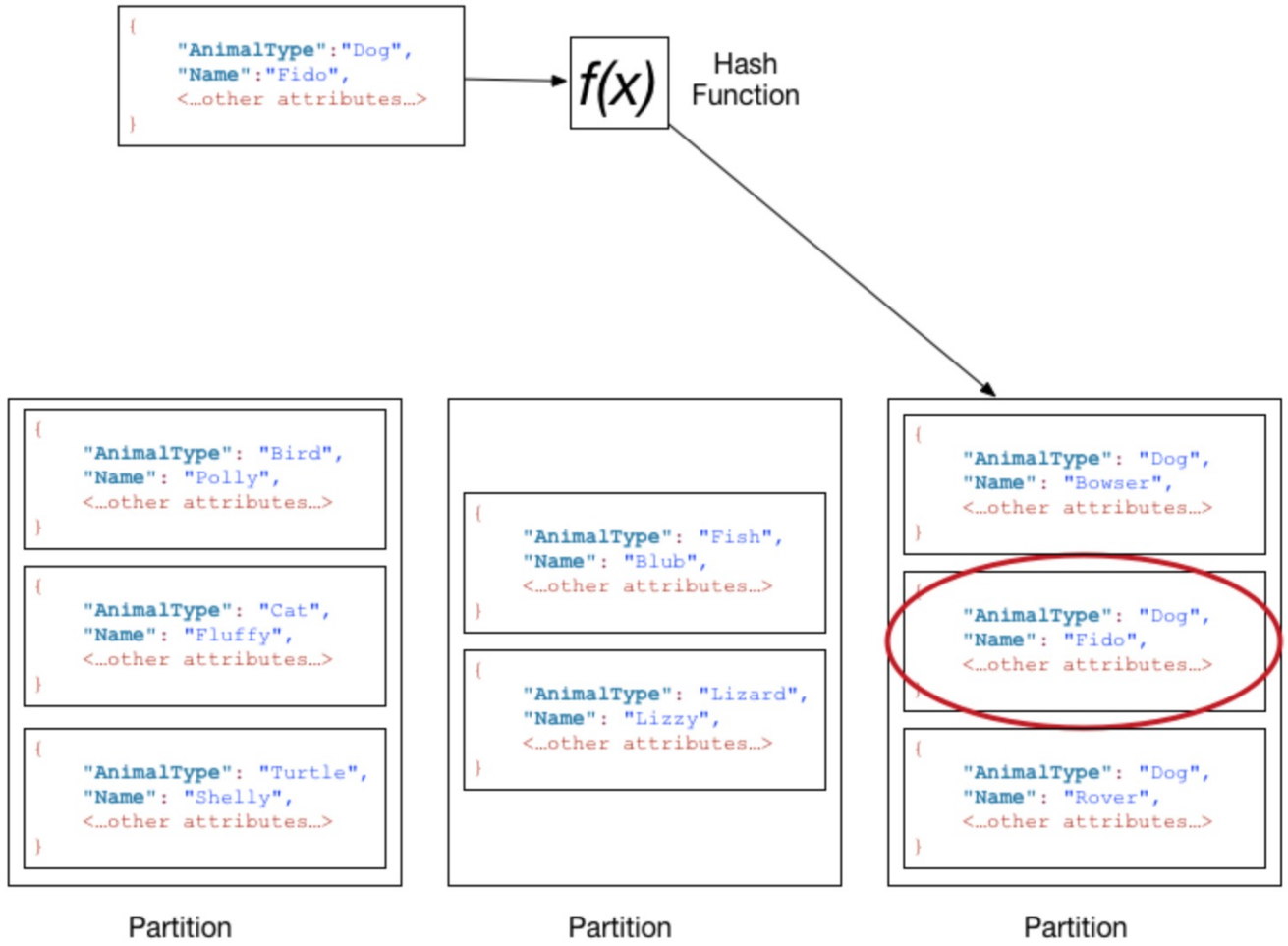
- TTL (expiração de itens) - time to live
- Partition Key (chave de partição)
- Sort Key (chave de ordenação)
- LSI (criação apenas no início) vs GSI (máximo 20, consistência eventual)
- Consistência eventual vs Consistência forte

DynamoDB


Primary Key
Products
Partition Key
Sort Key
Attributes
Product ID
Type
Schema is defined per item
Items
1
Book ID
Odyssey
Homer
1871
2
Album ID
6 Partitas
Bach
2
Album ID: Track ID
Partita No. 1
3
Movie ID
The Kid
Drama, Comedy
Chaplin

DynamoDB

Suponhamos que a tabela *Pets* possua uma chave primária composta que consiste em *AnimalType* (chave de partição) e *Name* (chave de classificação). O diagrama a seguir mostra o DynamoDB gravando um item com um valor de chave de partição *Dog* e um valor de chave de classificação *Fido*.



RDS - Autoscaling

IMPORTANTE: Cenários em que a capacidade de storage está chegando no limite

O Amazon RDS agora oferece suporte ao Storage Auto Scaling

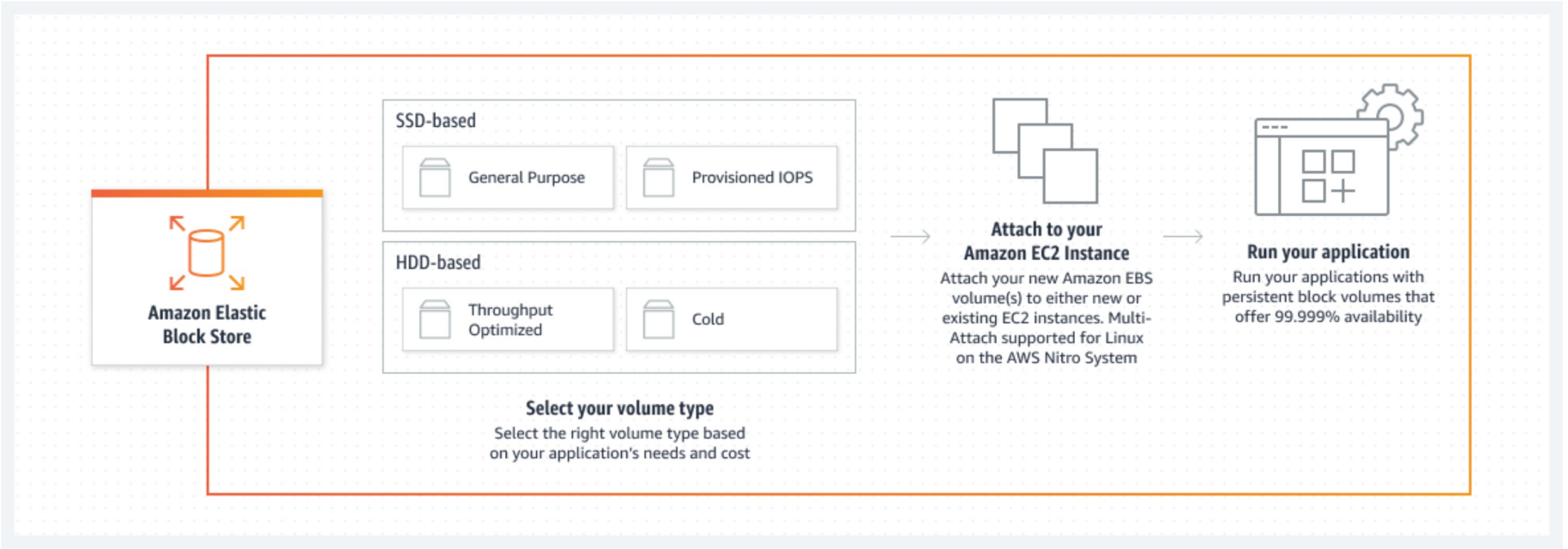
Publicado: Jun 20, 2019

A partir de hoje, o Amazon RDS for MariaDB, o Amazon RDS for MySQL, o Amazon RDS for PostgreSQL, o Amazon RDS for SQL Server e o Amazon RDS for Oracle oferecem suporte ao RDS Storage Auto Scaling. O RDS Storage Auto Scaling dimensiona automaticamente a capacidade de armazenamento em resposta às crescentes cargas de trabalho do banco de dados, sem tempo de inatividade.

EBS

Como funciona

O Amazon Elastic Block Store (Amazon EBS) é um serviço de armazenamento em blocos fácil de usar, escalável e de alta performance projetado para o Amazon Elastic Compute Cloud (Amazon EC2).



EBS

Tipo de EBS	Descrição	Uso Comum	Desempenho	Custo	Capacidad
General Purpose SSD (gp3)	Armazenamento SSD de propósito geral. Oferece um equilíbrio entre custo e desempenho, com a capacidade de ajustar IOPS e throughput independentemente.	- Servidores de aplicação - Bancos de dados pequenos e médios - Volume de boot de instâncias EC2	3.000 IOPS padrão (pode ser ajustado até 16.000 IOPS) Throughput de até 1.000 MB/s	Moderado	1 GiB a 16 TiB
General Purpose SSD (gp2)	Armazenamento SSD de propósito geral. Oferece uma boa relação custo-benefício com desempenho que escala com o tamanho do volume.	- Workloads de propósito geral - Volumes de boot - Bancos de dados de tamanho médio	3 IOPS por GiB, até 16.000 IOPS Throughput de até 250 MB/s	Moderado	1 GiB a 16 TiB
Provisioned IOPS SSD (io1/io2)	Armazenamento SSD de alto desempenho com IOPS provisionado para cargas de trabalho críticas que exigem desempenho consistente e baixa latência.	- Bancos de dados críticos (OLTP) - Workloads sensíveis à latência - Aplicações de missão crítica	Até 64.000 IOPS (io1) / Até 64.000 IOPS (io2) Throughput de até 1.000 MB/s (io2)	Alto	4 GiB a 16 TiB (io1) 4 GiB a 64 TiB (io2)

EBS

Throughput Optimized HDD (st1)	Armazenamento HDD otimizado para throughput, projetado para cargas de trabalho que requerem grandes volumes de leitura e escrita sequenciais.	- Big data - Processamento de log - Data warehouses	Até 500 IOPS Throughput de até 500 MB/s	Baixo	125 GiB a 16 TiB
Cold HDD (sc1)	Armazenamento HDD de baixo custo para dados acessados esporadicamente, onde o custo é mais importante do que o desempenho.	- Arquivos de grande volume - Arquivamento de dados - Backups de longa duração	Até 250 IOPS Throughput de até 250 MB/s	Muito baixo	125 GiB a 16 TiB
Magnetic (Standard)	Tipo de volume legada não recomendado para novos desenvolvimentos. Oferece armazenamento magnético com desempenho limitado e baixo custo.	- Workloads legadas	40-200 IOPS Throughput de até 90 MB/s	Baixo (não recomendado para novas cargas)	1 GiB a 1 TiB

EBS

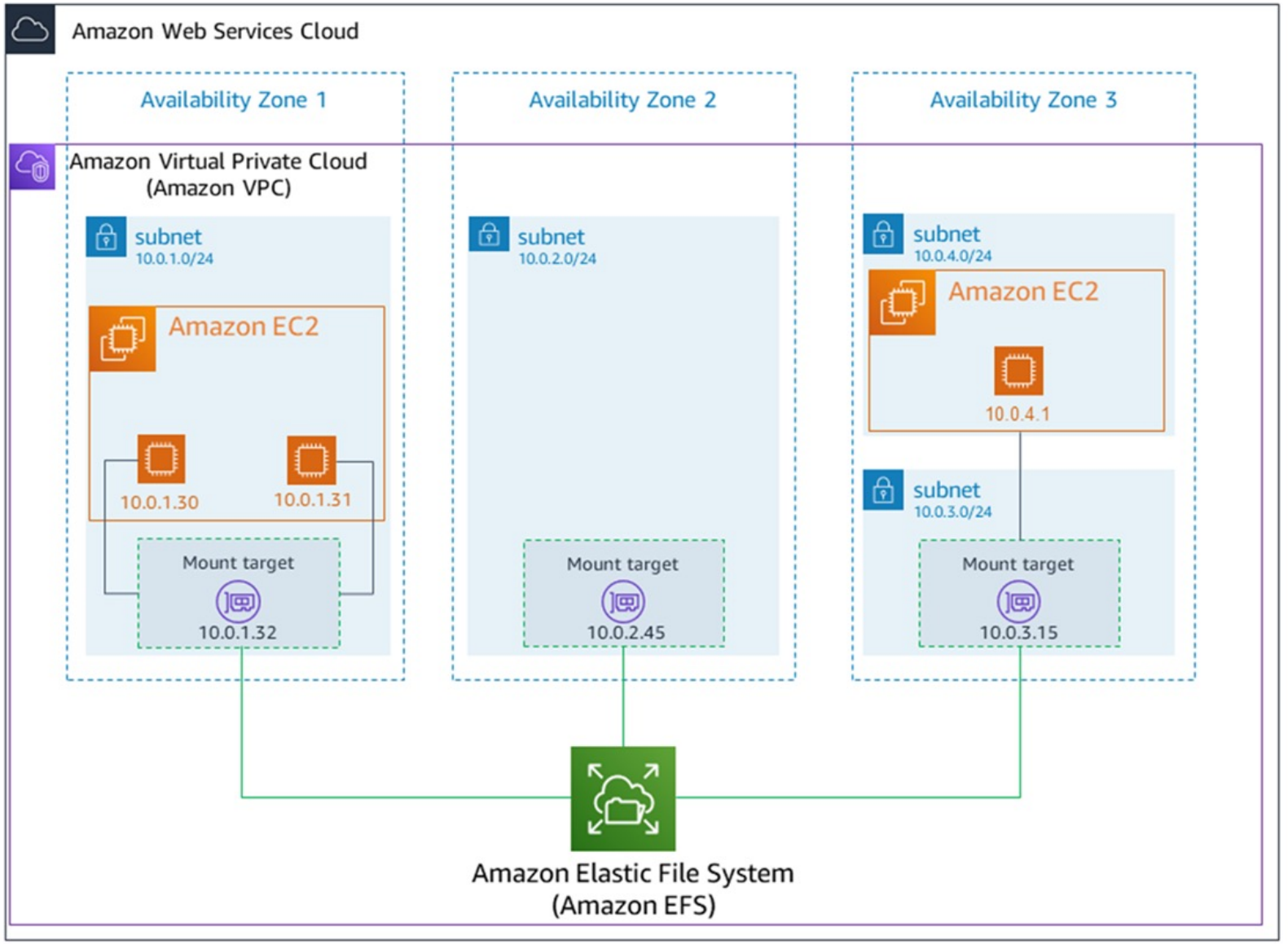
O EBS não é a melhor opção para usarmos para várias instâncias. Para isso, usaremos o EFS.

Ref.: https://stackoverflow.com/questions/841240/can-you-attach-amazon-ebs-to-multiple-instances

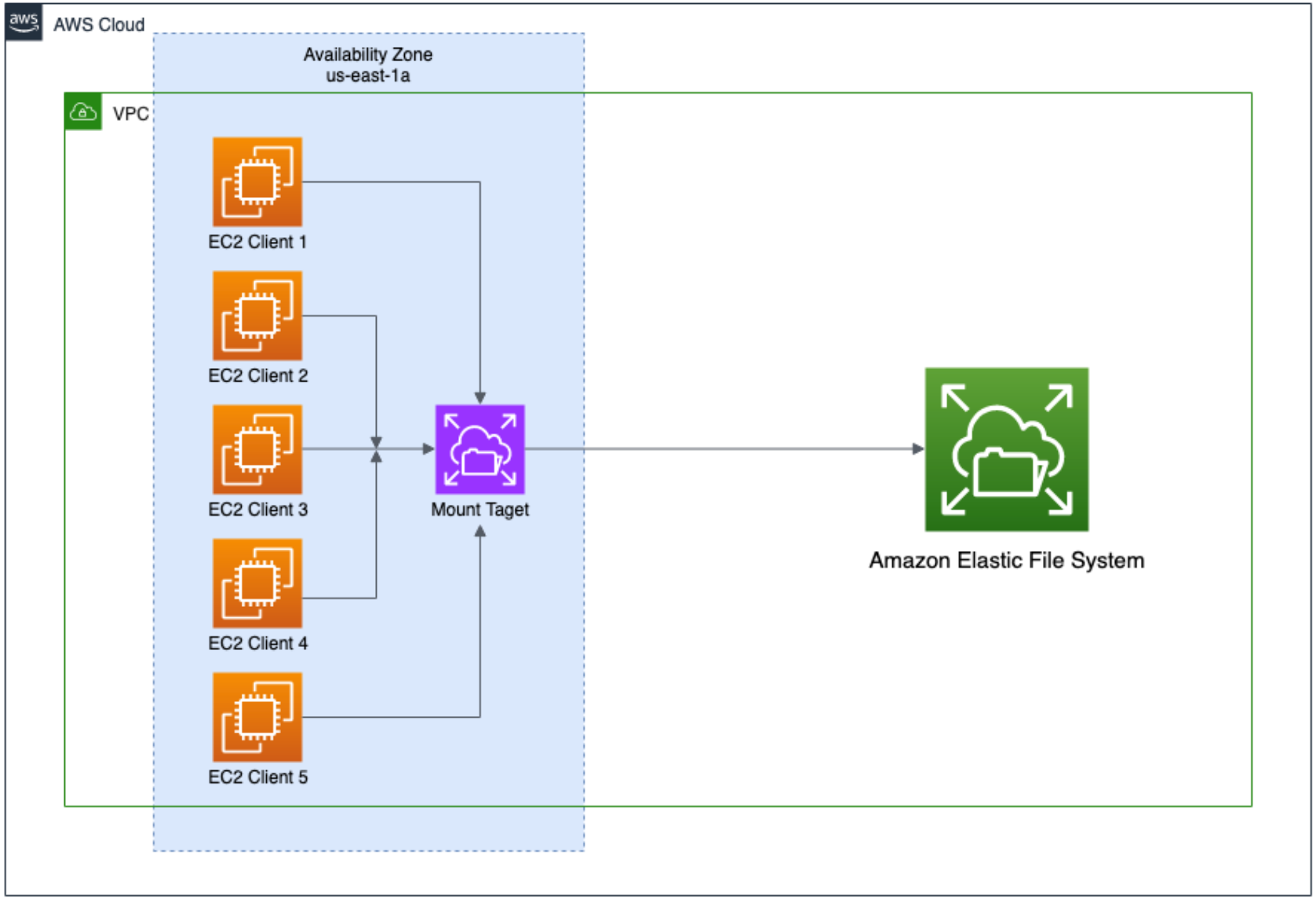
EFS

- Integra com o CloudTrail

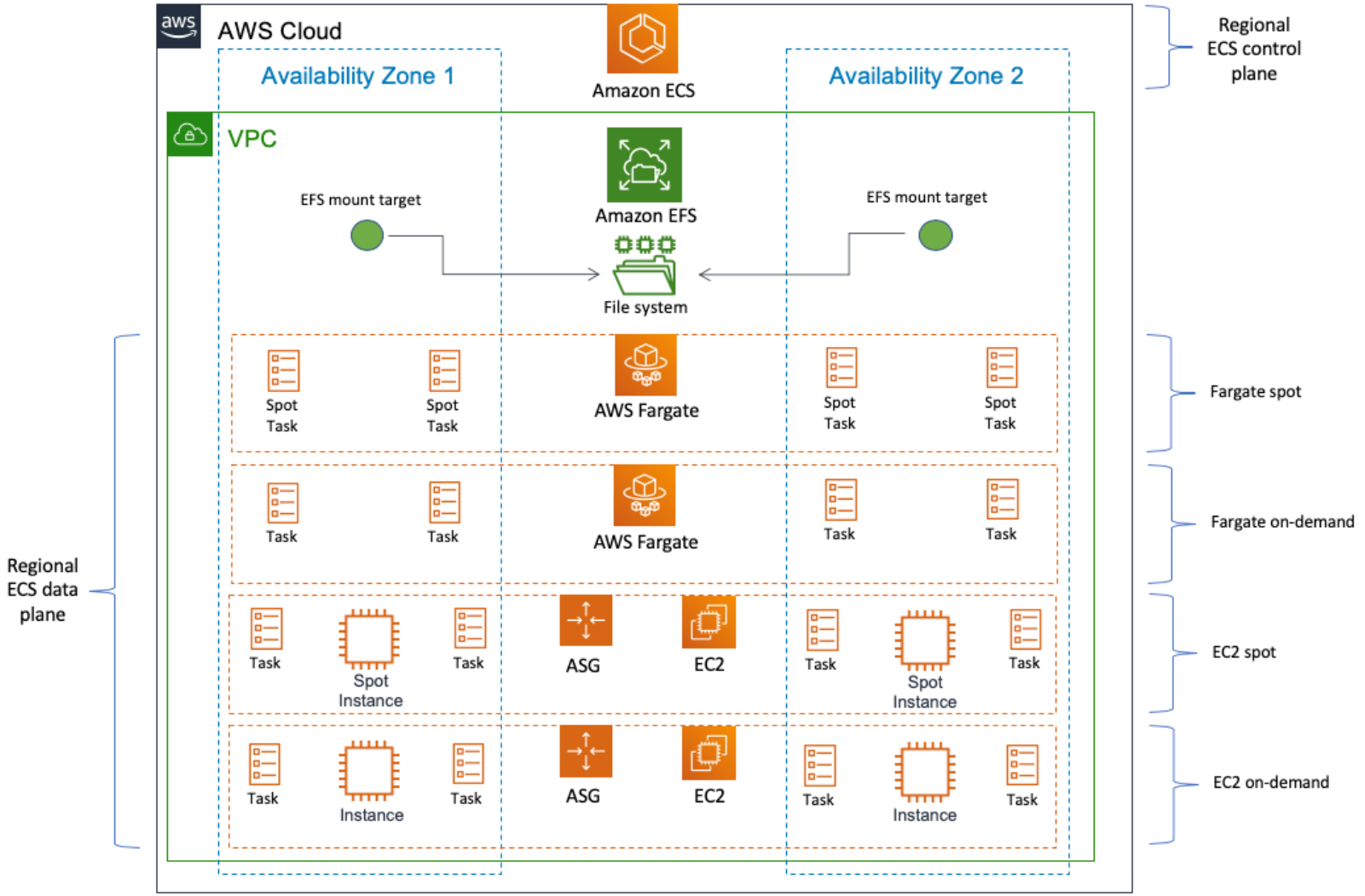
- Multi-AZ



EFS


AWS Cloud
Availability Zone
us-east-1a
VPC
EC2 Client 1
EC2 Client 2
EC2 Client 3
EC2 Client 4
EC2 Client 5
Mount Taget
Amazon Elastic File System

EFS + ECS: https://aws.amazon.com/blogs/containers/developers-guide-to-using-amazon-efs-with-amazon-ecs-and-aws-fargate-part-2/



EFS:

Modo de Throughput	Descrição	Casos de Uso Comuns	Escalabilidade	Custo
Bursting Throughput	Ajusta automaticamente o throughput com base no tamanho do sistema de arquivos.	- Workloads com padrões de acesso variáveis - Ambientes de desenvolvimento e teste	Escalável automaticamente com base no tamanho do sistema de arquivos	Custo baseado no armazenamento utilizado
Provisioned Throughput	Permite configurar manualmente o throughput independente do tamanho do sistema de arquivos.	- Aplicações críticas que requerem desempenho consistente - Workloads com requisitos de throughput elevados	Escalável conforme necessário, independente do tamanho do sistema de arquivos	Custo adicional baseado no throughput provisionado

Exercícios

Questão 1

Cenário:

Uma empresa de mídia digital armazena milhões de arquivos de imagem e vídeo no Amazon S3. Esses arquivos são acessados frequentemente por usuários ao redor do mundo, mas a empresa tem notado um aumento significativo nos custos de armazenamento e de recuperação de dados. A equipe de TI foi encarregada de otimizar os custos de armazenamento no S3 sem comprometer a disponibilidade dos arquivos para os usuários. Além disso, os arquivos mais antigos são acessados com menor frequência, mas ainda precisam estar disponíveis para visualização on-demand.

Pergunta:

Qual estratégia de armazenamento no Amazon S3 você recomendaria para reduzir os custos, mantendo a disponibilidade e o acesso on-demand para os arquivos mais antigos?

Questão 1

a) Transferir todos os arquivos para o S3 Standard-IA (Infrequent Access) para reduzir os custos de armazenamento, já que essa classe oferece menor custo de armazenamento para dados acessados com pouca frequência.

b) Implementar o S3 Intelligent-Tiering para mover automaticamente os arquivos entre as classes de armazenamento Standard e Infrequent Access com base nos padrões de acesso, otimizando custos automaticamente.

c) Mover todos os arquivos para o S3 Glacier para maximizar a economia de custos, pois oferece o menor custo de armazenamento, e configurar recuperações expeditas para os arquivos mais antigos.

d) Utilizar o S3 One Zone-IA para armazenar os arquivos antigos que são acessados com pouca frequência, reduzindo os custos de armazenamento, já que essa classe é mais barata que o S3 Standard-IA.

Resposta Correta: b) Implementar o S3 Intelligent-Tiering para mover automaticamente os arquivos entre as classes de armazenamento Standard e Infrequent Access com base nos padrões de acesso, otimizando custos automaticamente.

Explicação:

•**S3 Intelligent-Tiering** é a melhor opção neste cenário porque oferece uma maneira automática de otimizar os custos de armazenamento com base nos padrões de acesso aos dados. Ele move objetos automaticamente entre duas camadas de acesso (frequente e infrequente) sem impacto no desempenho e sem taxas adicionais de recuperação. Isso permite que a empresa mantenha todos os arquivos acessíveis on-demand, enquanto minimiza os custos de armazenamento com base no uso real.
•**S3 Standard-IA** é uma boa opção para dados que são acessados com menos frequência, mas podem incorrer em taxas de recuperação se os dados forem acessados mais frequentemente do que o esperado.
•**S3 Glacier** oferece o menor custo de armazenamento, mas é mais adequado para arquivos que podem tolerar tempos de recuperação mais longos, geralmente em cenários de arquivamento. Como os arquivos precisam estar disponíveis on-demand, o S3 Glacier não seria a melhor escolha.
•**S3 One Zone-IA** reduz custos armazenando dados em uma única zona de disponibilidade, mas isso compromete a disponibilidade e durabilidade em caso de falha na zona, o que pode não ser aceitável para a empresa que precisa de alta disponibilidade dos arquivos.
Referência na Documentação da AWS:
•S3 Intelligent-Tiering
•S3 Storage Classes
Essa pergunta ajuda a explorar as diferentes opções de armazenamento no Amazon S3 e como escolher a melhor estratégia para otimizar custos enquanto mantém a disponibilidade e o desempenho necessários para os usuários finais.

Questão 2

Cenário:

Uma startup de análise de dados armazena grandes volumes de dados de sensores em formato CSV no Amazon S3. Esses dados são usados para gerar relatórios e análises que precisam ser processados diariamente por um cluster de Amazon EMR (Elastic MapReduce). No entanto, a startup está enfrentando desafios com o tempo de processamento, pois o número de arquivos está crescendo rapidamente e o desempenho do processamento tem diminuído. A equipe de arquitetura precisa otimizar a estrutura de armazenamento no S3 para melhorar o desempenho do processamento de dados no EMR.

Pergunta:

Qual estratégia de armazenamento no Amazon S3 você recomendaria para melhorar o desempenho do processamento de dados no Amazon EMR?

Questão 2

a) Comprimir todos os arquivos CSV antes de armazená-los no S3 para reduzir o tamanho total dos dados e acelerar o tempo de transferência para o EMR.

b) Agrupar os arquivos CSV em arquivos maiores e usar o S3 Select para consultar apenas os dados necessários, reduzindo o volume de dados transferidos para o EMR.

c) Converter os arquivos CSV para o formato Apache Parquet antes de armazená-los no S3, aproveitando a compatibilidade com colunas e a eficiência de leitura do Parquet para melhorar o desempenho do processamento no EMR.

d) Mover os dados para o S3 Glacier para reduzir os custos de armazenamento e configurar processos de recuperação expeditos para carregar os dados no EMR diariamente.

Questão 2

Resposta Correta: c) Converter os arquivos CSV para o formato Apache Parquet antes de armazená-los no S3, aproveitando a compatibilidade com colunas e a eficiência de leitura do Parquet para melhorar o desempenho do processamento no EMR.

Explicação:

•**Converter os arquivos CSV para o formato Apache Parquet** é a melhor estratégia neste cenário porque o Parquet é um formato de armazenamento colunar otimizado para leitura eficiente de dados, especialmente em grandes conjuntos de dados. Ele reduz o tempo de processamento ao permitir que o EMR leia apenas as colunas necessárias, em vez de processar todas as linhas em arquivos CSV. Isso melhora significativamente o desempenho do processamento de dados.
•**Comprimir os arquivos CSV** (opção a) pode reduzir o tamanho dos dados e melhorar o tempo de transferência, mas não aborda diretamente os desafios de desempenho do processamento em um ambiente de grandes dados como o EMR.
•**Agrupar os arquivos CSV em arquivos maiores** (opção b) pode ajudar a reduzir a sobrecarga de manipular muitos arquivos pequenos, mas o S3 Select é mais adequado para consultas simples e não substitui a necessidade de processamento em um cluster EMR.
•**Mover os dados para o S3 Glacier** (opção d) não é apropriado, pois o Glacier é destinado ao arquivamento de longo prazo e não é otimizado para o processamento frequente de dados, como o exigido pelo EMR diariamente.

Referência na Documentação da AWS:
•Amazon S3 and EMR Best Practices
•Apache Parquet
Essa pergunta explora como otimizar o armazenamento no S3 para melhorar o desempenho do processamento de dados em um ambiente de big data, como o Amazon EMR, abordando questões reais de desempenho e eficiência.

Questão 3

Cenário:

Uma empresa de e-commerce está expandindo rapidamente e precisa garantir que sua aplicação web esteja sempre disponível e possa escalar automaticamente com base na demanda dos clientes. Atualmente, a aplicação está hospedada em instâncias Amazon EC2 em uma única região da AWS. Com o aumento do tráfego durante períodos de pico, como Black Friday, a empresa tem enfrentado desafios para garantir alta disponibilidade e escalabilidade, especialmente durante falhas em uma zona de disponibilidade.
A equipe de arquitetura foi encarregada de revisar a configuração atual e recomendar uma solução que melhore a disponibilidade, escalabilidade e resiliência da aplicação.

Pergunta:

Qual das seguintes abordagens você recomendaria para garantir alta disponibilidade e escalabilidade automática para a aplicação web da empresa?

Questão 3

a) Usar uma única instância EC2 em uma zona de disponibilidade maior (larger instance size) para lidar com o aumento do tráfego e configurar snapshots regulares para garantir a recuperação em caso de falhas.

b) Implementar um Auto Scaling Group com instâncias EC2 distribuídas em várias zonas de disponibilidade dentro da mesma região, e configurar um Elastic Load Balancer para distribuir o tráfego entre as instâncias.

c) Configurar um Auto Scaling Group em uma única zona de disponibilidade com uma política de escalabilidade agressiva para adicionar instâncias rapidamente durante períodos de pico.

d) Mover a aplicação para uma instância EC2 de capacidade reservada (Reserved Instance) para garantir capacidade dedicada durante períodos de pico e configurar backups manuais para recuperação.

Questão 3

Resposta Correta: b) Implementar um Auto Scaling Group com instâncias EC2 distribuídas em várias zonas de disponibilidade dentro da mesma região, e configurar um Elastic Load Balancer para distribuir o tráfego entre as instâncias.

Explicação:

•**Implementar um Auto Scaling Group com instâncias EC2 distribuídas em várias zonas de disponibilidade** e usar um **Elastic Load Balancer (ELB)** é a abordagem recomendada para garantir alta disponibilidade e escalabilidade. O Auto Scaling Group permite que a aplicação escale automaticamente com base na demanda, enquanto o ELB distribui o tráfego entre várias instâncias EC2. A distribuição em várias zonas de disponibilidade aumenta a resiliência contra falhas em uma única zona, garantindo que a aplicação permaneça disponível mesmo em caso de problemas de infraestrutura em uma das zonas.
•**Opção a**: Usar uma única instância EC2, mesmo que de tamanho maior, não garante alta disponibilidade, pois a falha na zona de disponibilidade ou na instância resultaria em indisponibilidade da aplicação.
•**Opção c**: Configurar um Auto Scaling Group em uma única zona de disponibilidade melhora a escalabilidade, mas não oferece alta disponibilidade em caso de falha na zona.
•**Opção d**: Mover para uma instância de capacidade reservada (Reserved Instance) oferece economia de custos em comparação com instâncias sob demanda, mas não melhora a disponibilidade ou a escalabilidade da aplicação.

Referência na Documentação da AWS:
•Auto Scaling Groups
•Elastic Load Balancing
Essa pergunta ajuda os alunos a entender como garantir alta disponibilidade e escalabilidade automática em uma aplicação web hospedada em EC2, abordando conceitos importantes como Auto Scaling, distribuição de carga, e o uso de múltiplas zonas de disponibilidade para resiliência.

Questão 4

Cenário:

Uma empresa de mídia social está lançando um novo recurso de compartilhamento de vídeos que permite aos usuários fazer upload e processar vídeos em sua plataforma. Esse recurso exige que os vídeos sejam processados em diferentes formatos e resoluções, dependendo do dispositivo do usuário. O processamento de vídeo é intensivo em termos de CPU e pode variar em demanda ao longo do tempo, especialmente com o crescimento do número de usuários. A empresa deseja implementar uma solução escalável e eficiente em termos de custos, utilizando instâncias EC2.

Pergunta:

Qual arquitetura você recomendaria para processar os vídeos de maneira escalável e otimizada em termos de custos?

Questão 4

a) Configurar instâncias EC2 reservadas (Reserved Instances) com alta capacidade de CPU para garantir que haja recursos suficientes para o processamento de vídeos em horários de pico.

b) Implementar um Auto Scaling Group com instâncias EC2 Spot, que são mais baratas, e configurar a aplicação para reprocessar vídeos automaticamente em caso de interrupções das instâncias Spot.

c) Usar instâncias EC2 On-Demand com Elastic Load Balancing (ELB) para distribuir o tráfego de processamento de vídeo, garantindo que as instâncias sejam lançadas apenas quando necessário.

d) Criar um Auto Scaling Group com instâncias EC2 dedicadas (Dedicated Instances) para garantir recursos exclusivos para o processamento de vídeos, evitando a interferência de outros workloads na AWS.

Questão 4

Resposta Correta: b) Implementar um Auto Scaling Group com instâncias EC2 Spot, que são mais baratas, e configurar a aplicação para reprocessar vídeos automaticamente em caso de interrupções das instâncias Spot.

Explicação:

•**Instâncias EC2 Spot** são ideais para cargas de trabalho que são tolerantes a interrupções e podem ser retomadas sem impacto significativo, como o processamento de vídeos. Elas oferecem um custo significativamente mais baixo em comparação com instâncias On-Demand ou Reservadas, o que ajuda a otimizar os custos. Usar um **Auto Scaling Group** com instâncias Spot permite que a empresa escale automaticamente o número de instâncias com base na demanda, aproveitando os preços mais baixos das instâncias Spot. Configurar a aplicação para lidar com interrupções e reprocessar vídeos garante que a solução seja resiliente e eficiente.

•**Opção a**: Instâncias EC2 reservadas garantem capacidade e previsibilidade de custos, mas não são tão escaláveis ou econômicas para cargas de trabalho variáveis e tolerantes a interrupções como o processamento de vídeos.

•**Opção c**: Instâncias On-Demand são boas para cargas de trabalho com requisitos imprevisíveis, mas são mais caras do que as Spot, e usar apenas ELB não aproveita totalmente a escalabilidade e economia de custos que as Spot oferecem.

•**Opção d**: Instâncias EC2 dedicadas são apropriadas para workloads que exigem hardware exclusivo por razões de conformidade ou desempenho, mas são mais caras e menos flexíveis em termos de escalabilidade e custos para o processamento de vídeos.

Referência na Documentação da AWS:

•Amazon EC2 Auto Scaling

•Amazon EC2 Spot Instances

•Best Practices for EC2 Spot Instances

Essa pergunta ajuda os alunos a entender como usar diferentes tipos de instâncias EC2 e arquiteturas de Auto Scaling para otimizar o processamento de cargas de trabalho intensivas, como o processamento de vídeos, equilibrando escalabilidade, resiliência e custos.

Questão 5

Cenário:

Uma empresa de análise de dados processa grandes volumes de dados diariamente e armazena esses dados em arquivos que precisam ser acessados simultaneamente por várias instâncias EC2 em uma arquitetura de processamento distribuído. A empresa deseja usar um sistema de arquivos compartilhado que permita a leitura e gravação simultâneas dos dados por várias instâncias e que seja escalável para lidar com o crescimento contínuo dos dados. A equipe de arquitetura decidiu utilizar o Amazon Elastic File System (EFS) como o sistema de arquivos compartilhado. Eles estão avaliando como configurar o EFS para garantir alta disponibilidade, desempenho adequado e otimização de custos para sua carga de trabalho.

Pergunta:

Qual abordagem você recomendaria para configurar o Amazon EFS de forma a otimizar o desempenho e o custo para essa carga de trabalho?

Questão 5

a) Configurar o EFS no modo de desempenho **Bursting Throughput** e associá-lo a instâncias EC2 em uma única zona de disponibilidade para maximizar a largura de banda.

b) Configurar o EFS no modo de desempenho **Provisioned Throughput** e armazenar os dados em camadas de armazenamento **Infrequent Access (IA)** para reduzir os custos de armazenamento sem sacrificar o desempenho.

c) Utilizar o modo de desempenho **General Purpose** com **Elastic Throughput** para permitir que o EFS escale automaticamente a capacidade de throughput com base nas necessidades da aplicação, garantindo alta disponibilidade em múltiplas zonas de disponibilidade.

d) Configurar o EFS no modo de desempenho **Max I/O** e replicar manualmente os dados entre diferentes sistemas de arquivos EFS em várias regiões para garantir alta disponibilidade e redundância.

Questão 5

Resposta Correta: c) Utilizar o modo de desempenho **General Purpose** com **Elastic Throughput** para permitir que o EFS escale automaticamente a capacidade de throughput com base nas necessidades da aplicação, garantindo alta disponibilidade em múltiplas zonas de disponibilidade.

Explicação:

•**Modo de Desempenho General Purpose com Elastic Throughput** é a melhor abordagem para a maioria das cargas de trabalho de EFS que exigem alta disponibilidade e escalabilidade. O General Purpose oferece baixa latência e é adequado para sistemas de arquivos que precisam ser acessados por várias instâncias EC2 em várias zonas de disponibilidade, proporcionando alta disponibilidade e resiliência. O Elastic Throughput permite que o EFS ajuste automaticamente a capacidade de throughput com base nas demandas da aplicação, otimizando o desempenho sem a necessidade de provisionamento manual.

•**Opção a**: Configurar o EFS no modo Bursting Throughput pode ser benéfico para cargas de trabalho que têm picos de demanda, mas limitar o EFS a uma única zona de disponibilidade não oferece a alta disponibilidade desejada.

•**Opção b**: Provisioned Throughput oferece controle manual sobre o throughput, mas pode resultar em custos desnecessários se o throughput for provisionado além do necessário. O armazenamento em Infrequent Access (IA) é útil para economizar custos, mas pode não ser ideal se os dados forem acessados frequentemente.

•**Opção d**: O modo de desempenho Max I/O é adequado para cargas de trabalho que requerem alta taxa de transferência de dados e pode lidar com milhares de instâncias EC2 simultâneas, mas a replicação manual entre regiões é complexa e desnecessária para garantir alta disponibilidade, especialmente quando o EFS já fornece alta disponibilidade em múltiplas zonas de disponibilidade dentro de uma região.

Referência na Documentação da AWS:

•Amazon EFS Performance Modes

•Amazon EFS Throughput Modes

•Best Practices for Amazon EFS

Essa pergunta explora como configurar o Amazon EFS para suportar uma carga de trabalho de processamento de dados distribuída, abordando questões de desempenho, disponibilidade e otimização de custos.

Questão 6

Cenário:
Uma empresa de desenvolvimento de software utiliza instâncias Amazon EC2 para executar seus ambientes de desenvolvimento, teste e produção. Cada ambiente requer diferentes níveis de desempenho de I/O e resiliência de dados. O ambiente de produção lida com transações críticas de clientes, exigindo alta disponibilidade e consistência de dados, enquanto os ambientes de desenvolvimento e teste precisam de armazenamento rápido e econômico. A empresa precisa escolher o tipo de Amazon EBS (Elastic Block Store) mais adequado para cada ambiente, otimizando custos sem comprometer o desempenho ou a segurança dos dados.

Pergunta:
Qual abordagem você recomendaria para selecionar o tipo de volume EBS mais apropriado para cada ambiente, considerando os requisitos de desempenho e custos?

Questão 6

a) Usar **EBS General Purpose SSD (gp3)** para o ambiente de produção por seu equilíbrio entre desempenho e custo, e **EBS Cold HDD (sc1)** para os ambientes de desenvolvimento e teste, dado o seu baixo custo e armazenamento de alta capacidade.

b) Utilizar **EBS Provisioned IOPS SSD (io2)** para o ambiente de produção devido à sua alta durabilidade e desempenho consistente, e **EBS General Purpose SSD (gp3)** para os ambientes de desenvolvimento e teste, oferecendo uma boa relação custo-benefício.

c) Configurar **EBS Magnetic (standard)** para todos os ambientes para minimizar custos, enquanto confia no Auto Scaling para lidar com picos de demanda no ambiente de produção.

d) Implementar **EBS Throughput Optimized HDD (st1)** para o ambiente de produção, que requer altos níveis de I/O sequencial, e **EBS Provisioned IOPS SSD (io1)** para os ambientes de desenvolvimento e teste para maximizar a performance durante os testes.

Resposta Correta: b) Utilizar **EBS Provisioned IOPS SSD (io2)** para o ambiente de produção devido à sua alta durabilidade e desempenho consistente, e **EBS General Purpose SSD (gp3)** para os ambientes de desenvolvimento e teste, oferecendo uma boa relação custo-benefício.

Explicação:

•**EBS Provisioned IOPS SSD (io2)** é ideal para o ambiente de produção, pois oferece desempenho de IOPS altamente consistente e é projetado para cargas de trabalho críticas que exigem baixa latência, como bancos de dados transacionais. Além disso, o io2 oferece maior durabilidade (99.999% de durabilidade) em comparação com o io1, tornando-o adequado para ambientes de produção que exigem alta disponibilidade e consistência de dados.

•**EBS General Purpose SSD (gp3)** é uma excelente escolha para os ambientes de desenvolvimento e teste, pois oferece um bom equilíbrio entre desempenho e custo. O gp3 permite ajustar independentemente o IOPS e o throughput, oferecendo flexibilidade adicional para ajustar o desempenho conforme necessário durante o desenvolvimento e os testes.

•**Opção a**: EBS Cold HDD (sc1) é uma opção de baixo custo para cargas de trabalho que acessam dados esporadicamente, mas não oferece desempenho adequado para ambientes de desenvolvimento e teste que requerem tempos de resposta rápidos.

•**Opção c**: EBS Magnetic (standard) é uma opção legada e não é recomendada para novos desenvolvimentos, especialmente para ambientes de produção que exigem alta performance.

•**Opção d**: EBS Throughput Optimized HDD (st1) é adequado para cargas de trabalho que requerem I/O sequencial, como processamento de big data, mas não é a melhor escolha para um ambiente de produção com transações críticas que exigem IOPS consistentes. O uso de io1 para desenvolvimento e teste pode ser caro e desnecessário, considerando que o gp3 já oferece um bom desempenho com flexibilidade de custo.

Referência na Documentação da AWS:

•Amazon EBS Volume Types

•Amazon EBS Best Practices

Essa pergunta ajuda os alunos a entender como escolher o tipo de volume EBS apropriado com base nos requisitos específicos de desempenho, durabilidade e custo para diferentes ambientes, abordando as necessidades de um cenário real de arquitetura na AWS.

Questão 7

Cenário:

Uma empresa de marketing digital armazena grandes volumes de dados de log de cliques e interações de usuários em arquivos JSON no Amazon S3. Esses dados são utilizados para gerar relatórios de análise de comportamento dos usuários e para ajustar campanhas publicitárias em tempo real. A equipe de análise de dados precisa consultar esses dados frequentemente e gerar insights rápidos sem a necessidade de configurar e gerenciar uma infraestrutura complexa de banco de dados.
A empresa decidiu utilizar o Amazon Athena para consultar esses dados diretamente no S3, mas a equipe está buscando a melhor maneira de otimizar as consultas, reduzir os custos e melhorar o desempenho.

Pergunta:

Qual abordagem você recomendaria para otimizar as consultas no Amazon Athena e melhorar o desempenho e os custos ao analisar os dados armazenados no S3?

Questão 7

a) Manter os arquivos JSON no S3 como estão, pois o Amazon Athena pode consultar esses arquivos diretamente sem qualquer preparação adicional.

b) Converter os arquivos JSON para o formato CSV e armazená-los no S3, já que CSV é um formato de arquivo mais simples que pode acelerar o processamento de consultas no Athena.

c) Utilizar o AWS Glue para catalogar os dados no S3 e converter os arquivos JSON para o formato Parquet, que é colunar e compactado, melhorando o desempenho das consultas e reduzindo os custos.

d) Mover os dados do S3 para um banco de dados Amazon RDS e usar o Amazon Athena para consultar o banco de dados diretamente, aproveitando a integração com RDS para melhorar o desempenho.

Resposta Correta: c) Utilizar o AWS Glue para catalogar os dados no S3 e converter os arquivos JSON para o formato Parquet, que é colunar e compactado, melhorando o desempenho das consultas e reduzindo os custos.

Explicação:

•**Converter os arquivos JSON para o formato Parquet** e catalogá-los com o AWS Glue é a melhor abordagem para otimizar as consultas no Amazon Athena. O Parquet é um formato de arquivo colunar, o que significa que o Athena precisa ler apenas as colunas relevantes para uma consulta específica, em vez de todo o arquivo. Isso melhora significativamente o desempenho das consultas e reduz os custos, já que o Athena cobra com base na quantidade de dados escaneados.

•**Opção a**: Embora o Athena possa consultar arquivos JSON diretamente, JSON é um formato baseado em linhas, o que pode resultar em maior tempo de leitura e custos mais altos, pois o Athena terá que escanear mais dados.

•**Opção b**: Converter JSON para CSV não é ideal para consultas em grandes conjuntos de dados porque o CSV também é um formato baseado em linhas, e não aproveita as vantagens de desempenho e custo que um formato colunar como o Parquet oferece.

•**Opção d**: O Amazon Athena é projetado para consultar diretamente dados no S3, e mover os dados para o Amazon RDS adicionaria complexidade desnecessária e custos adicionais, sem melhorar significativamente o desempenho das consultas no Athena.

Referência na Documentação da AWS:

•Amazon Athena Best Practices

•Working with AWS Glue and Amazon Athena

•Columnar Formats like Parquet and ORC

Essa pergunta ajuda os alunos a entender como otimizar o uso do Amazon Athena para consultar grandes conjuntos de dados no S3, destacando a importância de escolher o formato de armazenamento correto para melhorar o desempenho e reduzir custos.

Questão 8

Cenário:

Uma empresa de jogos online está desenvolvendo um novo sistema de placares e conquistas para os jogadores. Esse sistema precisa armazenar dados como pontuações, níveis alcançados e conquistas desbloqueadas para milhões de usuários ao redor do mundo. O acesso a esses dados deve ser rápido e escalável, pois as consultas serão feitas em tempo real durante as sessões de jogo. Além disso, o sistema precisa lidar com picos de tráfego durante lançamentos de novos conteúdos ou eventos especiais dentro do jogo.
A equipe de arquitetura decidiu usar o Amazon DynamoDB para armazenar e consultar esses dados. Agora, eles precisam decidir sobre a melhor estratégia de modelagem de dados e configuração de DynamoDB para atender às necessidades da aplicação.

Pergunta:

Qual abordagem você recomendaria para modelar os dados e configurar o DynamoDB para garantir alta disponibilidade, desempenho e escalabilidade?

Questão 8

a) Criar uma tabela DynamoDB única para todos os usuários, utilizando a combinação de ID do usuário e tipo de dado (por exemplo, pontuação, conquistas) como chave composta, e configurar a tabela para capacidade provisionada com Auto Scaling habilitado.

b) Criar tabelas DynamoDB separadas para pontuações, níveis e conquistas, utilizando a ID do usuário como chave primária, e configurar cada tabela para capacidade sob demanda (on-demand) para lidar com picos de tráfego.

c) Usar uma única tabela DynamoDB para armazenar todas as informações, utilizando um esquema de chave-partição com base na ID do usuário, e configurar a tabela para capacidade sob demanda (on-demand) para escalar automaticamente com a demanda.

d) Configurar várias tabelas DynamoDB em diferentes regiões da AWS para cada tipo de dado (pontuações, níveis, conquistas) e utilizar Global Tables para replicar os dados entre as regiões, garantindo alta disponibilidade e desempenho.

Resposta Correta: c) Usar uma única tabela DynamoDB para armazenar todas as informações, utilizando um esquema de chave-partição com base na ID do usuário, e configurar a tabela para capacidade sob demanda (on-demand) para escalar automaticamente com a demanda.

Explicação:

•**Utilizar uma única tabela DynamoDB** com um esquema de chave-partição baseado na ID do usuário é a abordagem recomendada porque permite que todos os dados relacionados a um usuário específico (como pontuações, níveis e conquistas) sejam armazenados juntos e acessados eficientemente. Isso otimiza o desempenho de leitura e escrita, especialmente em um ambiente de alta escalabilidade como o de um jogo online. Configurar a tabela para capacidade sob demanda (on-demand) permite que o DynamoDB escale automaticamente para lidar com picos de tráfego sem a necessidade de provisionamento manual de capacidade.
•**Opção a**: Embora o uso de uma chave composta possa funcionar, a configuração para capacidade provisionada pode não ser tão flexível quanto a capacidade sob demanda, especialmente para um sistema com tráfego variável.
•**Opção b**: Criar tabelas separadas para diferentes tipos de dados pode levar a uma maior complexidade e potencialmente mais consultas de leitura, que podem aumentar os custos e a latência.
•**Opção d**: Usar Global Tables para replicação entre regiões é útil para garantir alta disponibilidade em um cenário de múltiplas regiões, mas a criação de várias tabelas em diferentes regiões aumenta a complexidade e pode não ser necessária se o foco principal for a escalabilidade dentro de uma única região.
Referência na Documentação da AWS:
•Amazon DynamoDB Best Practices
•Using On-Demand Capacity Mode
•Data Modeling in Amazon DynamoDB
Essa pergunta ajuda os alunos a entender como configurar e modelar dados no Amazon DynamoDB para uma aplicação que exige alta disponibilidade, desempenho e escalabilidade, abordando um cenário real de arquitetura de sistemas de jogos online.