VINÍCIUS HENRIQUE DOS SANTOS

MACHINE LEARNING ENGINEERING

BIG DATA STORAGE STRUCTURES

AULA 02

SUMÁRIO

O QUE VEM POR AÍ?

Todos nós já estamos cansados(as) de saber que o volume de dados gerados nos dias de hoje cresce em uma escala assustadora. Em 2010, o volume de dados produzidos globalmente era de 1,2 zettabytes, passando para 8 zettabytes em 2015... Entretanto, as coisas não pararam por aí e continuam crescendo, estimando-se atingir 175 zettabytes em 2025.

Por conta desse cenário, você tem o papel fundamental de entender como manipular, armazenar e usar parte desse volume gigantesco de dados e aprenderá como fazer isso usando o Amazon Simple Storage Service, ou Amazon S3. Com o S3, você entenderá como armazenar dados estruturados, desestruturados ou semi-estruturados em seu Data Lake de forma segura, resiliente e distribuída.

HANDS ON

Nessa aula você entenderá como usar um dos principais serviços de armazenamento do mercado como um Data Lake através do Amazon S3. Assim, temos:

1. Definir os componentes fundamentais do S3.
2. Criar buckets e realizar upload de objetos dentro do S3.
3. Interagir com o data lake através de 3 formas:
    a) AWS CLI.
    b) Console de Gerenciamento.
    c) SDK Python.
4. Habilitar e testar o versionamento de objetos dentro do Data Lake.

SAIBA MAIS

Amazon s3 e o papel de Data Lake

Em abril de 2023, as estimativas indicavam que o volume global de dados criados, capturados, copiados e consumidos no mundo estava projetado para crescer de 64,2 zettabytes em 2020 para mais de 180 zettabytes até 2025.

No entanto, é importante notar que esses números são projeções baseadas em tendências de crescimento e podem variar com o desenvolvimento de novas tecnologias, mudanças nos padrões de consumo digital e outros fatores imprevisíveis que afetam a geração de dados.

A quantidade de dados gerados anualmente tem visto um aumento significativo a cada ano, impulsionada por fatores como a digitalização crescente dos negócios e da vida cotidiana, o aumento do uso de dispositivos inteligentes e conectados à internet e a popularidade das plataformas de mídia social e serviços de streaming, além de avanços em tecnologias como Internet das Coisas (IoT), Inteligência Artificial (IA) e aprendizado de máquina.

Para dados mais atuais sobre o volume específico de dados gerados em um determinado ano após 2023, seria necessário consultar as últimas pesquisas e relatórios do setor, já que as estimativas e realidades podem mudar rapidamente devido à natureza dinâmica da tecnologia e dos padrões de consumo de dados.

Isso significa que empresas que trabalham com silos de dados tendem a ter cada vez mais dificuldades para transformar seus dados em insights de negócio, além de se tornar cada vez mais inviável pensar em gerenciamento de grandes volumes de dados em um ambiente on-premises tanto no âmbito de custos quanto performance ou dificuldade de uso.

Pensando nisso, uma arquitetura de Data Lake como o Amazon S3 permite agilidade e a capacidade de obter mais insights e valor de seus dados, além de ser possível adotar ferramentas e processos analíticos mais sofisticados à medida que suas necessidades evoluem.

Pense em um Data Lake como um repositório de dados centralizados que permite armazenar e migrar todos os dados estruturados e não estruturados sem

limitações. Uma vez que os dados estão dentro desse repositório central, você pode utilizar diversas ferramentas para tirar insights da forma que for mais viável para sua necessidade.

Empresas que implementam boas soluções de Data Lake tornam os dados mais disponíveis e acessíveis para usuários gerarem valor através de ferramentas e funções analíticas de Machine Learning, bancos de dados, indicadores, KPIs, automações ou os mais variados tipos de serviços.

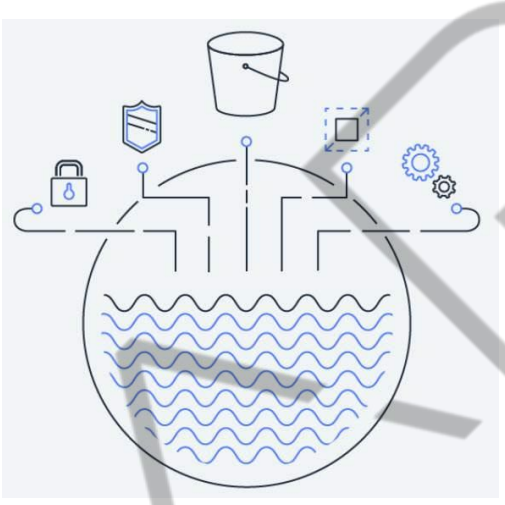

Figura 1 – Ilustração Datalake
Fonte: Google Imagens (2024)

Soluções de Data Lake baseadas no AWS S3 se beneficiam de um ambiente de performance com alta escalabilidade em uma arquitetura integrada com outros serviços, sendo estável e sempre disponível.

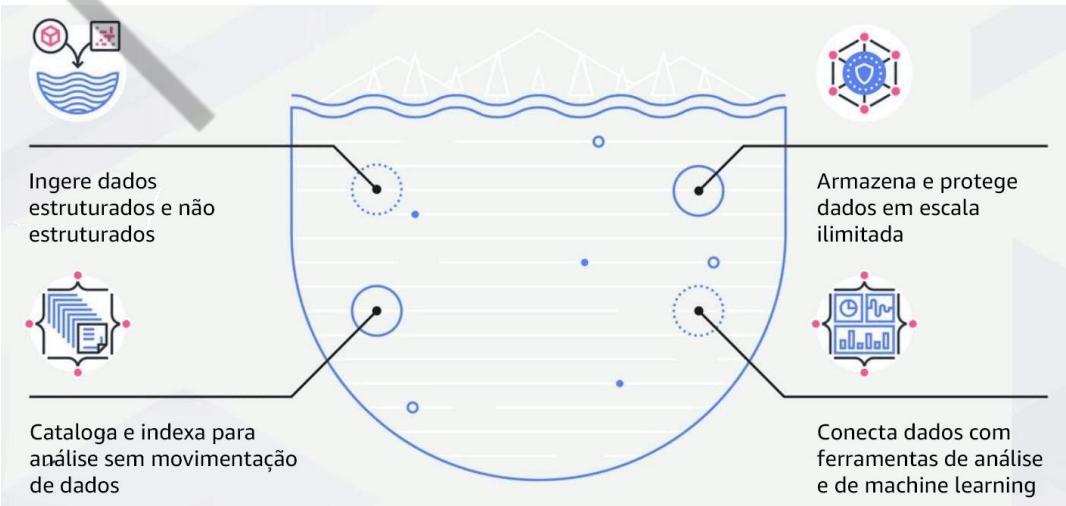

Figura 2 – Vantagens de um Datalake
Fonte: Google Imagens (2024)

Mais detalhes do Amazon s3

O Amazon Simple Storage Service é o serviço destinado a armazenamento de dados número um no setor de tecnologia e Big Data e oferece altíssima disponibilidade de dados, segurança e performance para leitura e consumo das informações.

Empresas de qualquer lugar do mundo, independentemente do setor ou do tamanho, podem usar o S3 para armazenar e garantir a segurança de qualquer volume de dados para uma infinidade de casos de uso, sendo a principal delas o Data Lake.

Dentro do S3, graças aos recursos oferecidos, é possível otimizar a utilização dos dados, organizar de acordo com a especificidade do negócio e configurar acessos que mantenham a segurança e a proteção dos dados envolvidos.

Classes de armazenamento

Dentro do S3 é ofertada uma grande variedade de classes de armazenamento dos dados que se adequa melhor a cada necessidade de uso. Em um cenário em que você possui dados de produção sendo consumidos frequentemente, faz sentido usar a classe Standard ou a S3 Express One Zone.

Para cenários em que seus dados são acessados com pouca frequência, você pode direcioná-los para um armazenamento de classe Standard-IA ou One Zone-IA. Dessa forma, você irá gerar economia de recursos e dinheiro.

Além disso, existem classes para dados em que você tem um custo ainda menor e cabem muito bem para cenários de acesso de pouquíssima ou nenhuma frequência, como as classes do tipo Glacier.

O tipo Amazon S3 Express One Zone é uma opção de armazenamento dentro do Amazon S3 criada especificamente para entregar acesso a dados com uma latência abaixo de dez milissegundos e é ideal para aplicações que demandam respostas em alta velocidade.

Esta classe se destaca como a opção de armazenamento de objetos na nuvem com a menor latência atualmente disponível, proporcionando velocidade de acesso até dez vezes superior e reduzindo os custos de solicitação em até 50% em comparação com a variante S3 Standard. Ela permite a escolha de uma única zona

de disponibilidade, possibilitando a co-localização do armazenamento de objetos e recursos computacionais para maximizar a velocidade de acesso.

Para aprimorar a rapidez no acesso e suportar um vasto número de solicitações por segundo, os dados são alocados em uma nova estrutura de armazenamento: um bucket de diretório Amazon S3.

Para dados cujo padrão de acesso pode variar ou é incerto o S3 Intelligent-Tiering é recomendado, uma vez que ajusta os custos de armazenamento ao mover automaticamente os dados entre quatro níveis de acesso com base na frequência de acesso. Estes níveis incluem dois para acesso frequente e infrequente com baixa latência e dois para arquivamento, destinados a dados raramente acessados, o que facilita o acesso assíncrono.

A organização dos dados é realizada em buckets dentro do S3, respeitando o armazenamento baseado em objetos. Cada objeto dentro do S3 é localizado pela composição de “nome do bucket” + “prefixo” + “nome do objeto”.

Buckets servem como recipientes estáveis que armazenam objetos e em sua conta AWS é possível estabelecer de 1 a 100 buckets. Para elevar esse número a até um máximo de 1.000, basta solicitar uma expansão no limite do serviço. Não há necessidade de definir um tamanho específico para os buckets, visto que eles suportam quantidades praticamente ilimitadas de dados, diferentemente do que ocorre com volumes de armazenamento ou partições.

Um bucket do Amazon S3 oferece uma solução de armazenamento flexível, permitindo funcionalidades como hospedagem de websites estáticos, armazenamento de informações de versões de objetos e implementação de estratégias de gerenciamento de ciclo de vida. Essas políticas ajudam a equilibrar a retenção de diferentes versões de objetos com o tamanho geral e os custos associados ao bucket.

A criação de um bucket deve respeitar algumas regras e restrições:

1. Uma vez que determinado bucket é criado em sua conta, ele jamais poderá ser transferido para uma outra conta AWS.
2. Os nomes dos buckets devem ser exclusivos a nível mundial, exigindo que você escolha uma combinação de nomes inexistente até o momento em toda a infraestrutura AWS.

3. Uma vez criado, nenhum bucket pode ser renomeado.
4. Os buckets são unidades de armazenamento permanentes, fazendo com que sua destruição seja possível apenas quando você já o esvaziou movendo ou deletando todos os arquivos dentro dele. Quando um bucket S3 é excluído, seu nome volta a ficar disponível globalmente para utilização.

O QUE VOCÊ VIU NESTA AULA?

Nessa aula você conseguiu compreender a importância de um Data Lake escalável, rápido e seguro para se trabalhar com o crescimento exponencial do volume de dados gerados e consumidos por todas as empresas.

Além disso, desenvolveu a habilidade de criar seus buckets dentro do AWS S3, realizou upload e deleção de objetos, compreendeu a arquitetura do serviço e definiu a política de gerenciamento dentro do Data Lake.

REFERÊNCIAS

AWS. **O que é o Amazon S3?** 2024. Disponível em: <https://docs.aws.amazon.com/pt_br/AmazonS3/latest/userguide/Welcome.html>. Acesso em: 24 abr. 2024.

AWS. **Getting started with Amazon S3**. 2024. Disponível em: <https://docs.aws.amazon.com/AmazonS3/latest/userguide/GetStartedWithS3.html>. Acesso em: 24 abr. 2024.

AWS. **AWS**. 2024. Disponível em: <https://awscli.amazonaws.com/v2/documentation/api/latest/reference/index.html>. Acesso em: 24 abr. 2024.

AWS. **Boto3 documentation**. Disponível em: <https://boto3.amazonaws.com/v1/documentation/api/latest/index.html>. Acesso em: 24 abr. 2024.

PALAVRAS-CHAVE

Palavras-chave: Data Lake. S3. Boto3. Bucket. Storage.