02.Volume: Armazenamento de dados AWS

segunda-feira, 4 de setembro de 2023 15:00

- *Quando as empresas têm mais dados do que conseguem processar e analisar, elas têm um problema de volume.*

- *Dispositivos geram uma quantidade de dados enorme, onde você recebe informações como resultado de dados dos dispositivos (IoT).*

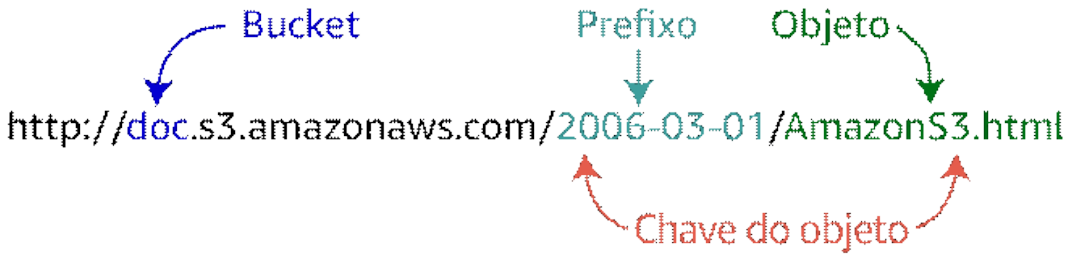
- ***Dados estruturados** são organizados e armazenados na forma de valores que são agrupados em linhas e colunas de uma tabela.*
- ***Dados semiestruturados** muitas vezes são armazenados em conjuntos de pares de chave-valor que são agrupados em elementos em um arquivo.*
- ***Dados não estruturados** não são estruturados de forma consistente. Alguns dados podem ter uma estrutura semelhante a dados semiestruturados, mas outros podem conter apenas metadados.*

- *Dados não estruturados precisam de Tags e precisam serem catalogados para serem analisados.*

Amazon S3 (Simple Storage Service)

- *Armazenamento de objetos*
- *Escalável*
- *Durável (99,999999999%)*
- *Uso para sites, aplicativos para dispositivos móveis e Data analytics*
- *Exabytes*
- *Desacoplar o armazenamento do processamento (Separar a maneira de armazenar os dados da maneira que processa os dados)*
- *Paralelização (Pode acessar os arquivos sem prejudicar os processos)*

Veja abaixo um exemplo de URL para um único objeto em um bucket chamado **doc**, com uma chave de objeto composta pelo prefixo **2006-03-01** e o arquivo nomeado **AmazonS3.html**.

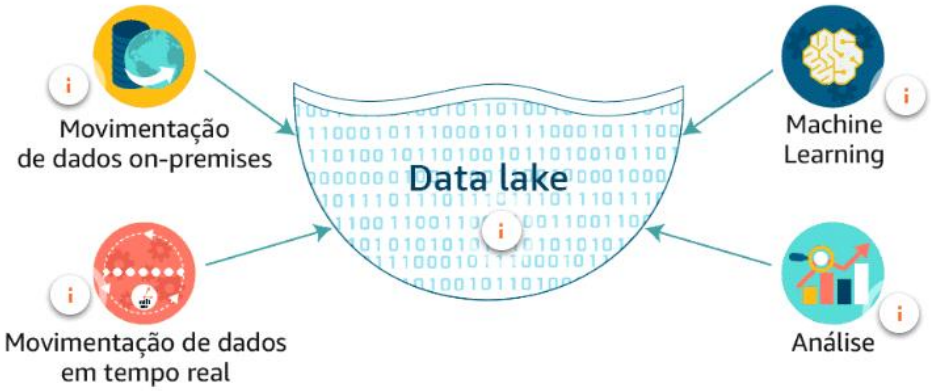


O Simple Storage Service pode armazenar dados organizados ou simplesmente jogar dentro do S3

- ***Data Lake** - Um data lake é um repositório centralizado que permite armazenar dados estruturados, semiestruturados e não estruturados em qualquer escala.*

Estruturados ou não estruturados

- *Pode usar as buckets do S3 e organizar os dados dentro do próprio S3*
- *A AWS tem um conjunto de ferramentas para gerenciar todo Data Lake sem tratar de cada bucket como objetos separados e não associados*
- ***Silos** - Quando as empresas armazenam os dados em diferentes tipos separados de armazenamentos, são raramente gerenciados e tratados pela mesma equipe.*
  - *Pode-se dividir os silos de dados e colocar em 1 repositório central gerenciado por uma única equipe*
  - *Não precisa transformar os dados para torna-los utilizáveis*
- *Precisão = Confiabilidade dos dados (Criar uma fonte confiável de dados = Data Lake)*
- *Armazenar o Data Lake no S3 Centraliza os dados.*



Benefícios de um data lake na AWS

- *São uma solução **de armazenamento de dados econômica**. Você pode armazenar de forma durável uma quantidade quase ilimitada de dados usando o Amazon S3.*
- *Implemente a **segurança e a conformidade** líderes do setor. A AWS usa rigorosos mecanismos de segurança, conformidade, privacidade e proteção de dados.*
- *Permite que você aproveite **muitas ferramentas diferentes de coleta e ingestão** de dados para ingerir dados em seu data lake. Esses serviços incluem o Amazon Kinesis para dados de streaming e dispositivos AWS Snowball para grandes volumes de dados locais.*
- *Ajudam você a **categorizar e gerenciar seus dados** de forma simples e eficiente. Use o AWS Glue para entender os dados dentro do seu data lake, prepará-los e carregá-los de forma confiável em datastores. Depois que o AWS Glue cataloga seus dados, eles são imediatamente pesquisáveis, podem ser consultados e estão disponíveis para processamento de ETL.*
- *Ajuda você a transformar dados em **informações significativas**. Utilize o poder dos serviços analíticos criados para finalidades específicas em vários casos de uso, como avaliação interativa, processamento de dados usando o Apache Spark e o Apache Hadoop, data warehousing, análise em tempo real, análise operacional, painéis e visualizações.*

- ***AWS Lake Formation***

  ***O AWS Lake Formation** facilita a ingestão, limpeza, catalogação, transformação e proteção dos seus dados, além de disponibilizá-los para avaliação e machine learning. O **Lake Formation** oferece um console central no qual você pode descobrir origens dos dados, configurar trabalhos de transformação para mover*

dados para um data lake do Amazon S3, remover duplicações e combinar registros, catalogar dados para acesso por ferramentas analíticas, configurar políticas de segurança e acesso a dados e auditar e controlar o acesso dos serviços analíticos e de machine learning da AWS.
***O Lake Formation** configura automaticamente os serviços AWS básicos para garantir a conformidade com suas políticas definidas. Se você configurou trabalhos de transformação que abrangem os serviços AWS, o Lake Formation configura os fluxos, centraliza a orquestração e permite que você monitore a execução dos trabalhos.*

- *Serviço para organizar e fazer curadoria dados*
- *Serviço para proteger dados em todo o data lake*
- *Serviço para orquestrar trabalhos de transformação com outros serviços AWS*