03.Armazenamento de dados estruturados

segunda-feira, 4 de setembro de 2023 17:24

Data Warehouse

- Armazenado por banco de dados
- Usado Data Warehouse
- Um repositório central de dados estruturados de muitas origens de dados
    - Transformar
    - Agregar
    - Preparar
        - Os dados para armazenar no Data Warehouse
- É a espinha dorsal do Business Inteligence

Vantagens	Desvantagens
Rápida recuperação de dados	Custo elevado para implementar
Conjuntos de dados com curadoria	Manutenção pode ser desafiadora
Armazenamento centralizado	Preocupações com segurança
Melhor business intelligence	Dificuldade de escalar para atender à demanda

- Recomendações:

    - Ao armazenar objetos ou arquivos individuais, recomendamos o Amazon S3.
    - Ao armazenar volumes massivos de dados, semiestruturados e não estruturados, recomendamos a criação de um data lake no Amazon S3.
    - Ao armazenar grandes quantidades de dados estruturados para avaliações complexas, recomendamos armazenar seus dados no Amazon Redshift.

Data Marts

Um subconjunto de dados de um data warehouse é chamado de **data mart**. Os data marts **se concentram em apenas um assunto ou uma área funcional**. Um data warehouse pode conter todas as fontes relevantes para uma empresa, mas um data mart pode armazenar **apenas as fontes de um único departamento**. Como os data marts geralmente são uma cópia dos dados já contidos em um data warehouse, eles geralmente são **rápidos e simples de implementar.**

- **Amazon Redshift** - Implantar um Datawarehouse em minutos
    - Construido para armazenar e consultar dados de gigabytes até

petabytes de dados
    - Escalável

- **Amazon Redshift Spectrum** - Combina o Data Lake com o Data Warehouse como se fosse uma única fonte de dados.

    - Benefícios dos Data Warehouse
        - Recuperação de dados rápida e centralizada
        - Conjunto de dados agrupados
        - Auxilio em insights
        - Escalável
- **Amazon EMR** - Ferramenta para criar operações de ETL(Extract Transform Load)
    - Estrutura gerenciada do Hadoop (Framework auxilia a armazenar e processar dados de forma temporária)
    - Gerir grandes volumes de dados rápidos e produzir insights
    - O Amazon EMR é o serviço AWS que implementa frameworks Hadoop. O serviço fará a ingestão dos dados de praticamente qualquer tipo de origem a praticamente qualquer velocidade! O Amazon EMR consegue implementar dois sistemas de arquivos diferentes: HDFS ou Elastic MapReduce File System (EMRFS). Um sistema de arquivos é um conjunto de regras organizacionais que controlam como os arquivos são armazenados.

- ## Amazon EMRFS

    O Amazon EMR oferece uma alternativa ao HDFS: o EMR File System (EMRFS). O EMRFS pode ajudar a garantir que haja uma “fonte confiável” persistente para dados do HDFS armazenados no Amazon S3. Ao implementar o EMRFS, não é necessário copiar dados para o cluster antes de transformar e analisar os dados como no HDFS. O EMRFS pode catalogar dados em um data lake no Amazon S3. O tempo economizado eliminando a etapa de cópia pode melhorar drasticamente o desempenho do cluster.

Uso do Hadoop

O Hadoop facilita a navegação de dados, descoberta e avaliação única de dados. Com o Hadoop, você pode compensar acontecimentos inesperados analisando grandes quantidades de dados rapidamente para formar uma resposta.