UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Data Lake

Kess Jhones Gomes Tavares - 180124498

Brasília, DF

2023

Data Lake

Atualmente, dispositivos conectados, mídias sociais, internet das coisas, entre outras fontes de dados, geram enormes quantidades de dados, incluindo dados estruturados, semiestruturados e não estruturados. Esses dados podem criar valor para sistemas corporativos e de suporte à decisão, muitas vezes apresentados como *data warehouses* (MACIEL, 2022).

Assim, a utilização de um *data warehouses* é explicada por Vida *et al* (2021, p. 14):

> "O data warehouse permite que sejam realizadas consultas a fim de extrair informações para a tomada de decisão. Nele, os dados armazenados já foram tratados, e não há redundância de informações. Ao analisar os dados de um data warehouse, eles estarão relacionados a determinado período, mostrando as métricas referentes a esse intervalo de tempo medido".

De acordo com Maciel (2022), “No entanto, tais sistemas lidam constantemente com o desafio de processar dados heterogêneos e volumosos, e disponibilizá-los para análises em tempo hábil com performance”. Assim, os *Data Lakes* vem ganhando popularidade por serem uma alternativa viável e que contorna alguns desses problemas.

Os data lakes fornecem armazenamento de dados em formatos de dados brutos. Ele permite análises em um repositório volumoso, diversificado e em evolução, possibilitando a extração de relatórios, painéis de visualizações, processamento de big data, análise em tempo real e aprendizado de máquina para uma melhor tomada de decisão (NOGUEIRA; ROMDHANE, 2018).

Um *Data Lake* não realiza pré-processamento dos dados, facilitando a integração e armazenamento de dados oriundos de diversas fontes assim reduzindo tempo, esforço e custo necessários para sua implementação (Betrybe, 2022). Entretanto, deve ser bem projetado. Caso contrário, poderá rapidamente se tornar um pântano de dados inoperante. Ou seja, um data lake deve permitir a consulta de dados (seleção/restrição) com bons tempos de resposta, e não apenas armazenando dados (NOGUEIRA; ROMDHANE, 2018).

O termo, *Data Lake* tem sua criação vinculado a James Dixon que enquanto diretor de tecnologia da Pentaho, em 2011, teria criado o termo para se desvincular do termo *data mart*, que é a subdivisão por setores de um *data warehouse* (Vida *et al*, 2021). Desta forma vale a ressalva de que *data lake* não é uma tecnologia, mas sim o conceito de negócio (Betrybe, 2022).

Segundo Morais (2022), o Apache Hadoop é amplamente utilizado para a implementação local de *data lakes*, porém não é a única ferramenta capaz de realizar a tarefa, visto que que a composição básica é:

- **Ingestão de dados:** É a etapa que consiste em transportar os data para dentro do data lake, algumas ferramentas que podem ser usadas são: Flink, Samza, Flume, Kafka e Sqoop
- **Armazenamento:** Como Morais (2022) trás:

  > "… de acordo com Sawadogo e Darmont (2021), existem duas abordagens para armazenamento de dados em Data Lakes. A primeira é por meio da utilização de SGBDs tradicionais, como MySQL, PostgreSQL e Oracle. Geralmente, estes SGBDs são utilizados para armazenamento de dados estruturados e possuem também alguns recursos para armazenamento de dados não estruturados. Mas, geralmente, bancos de dados NoSQL são utilizados para o armazenamento de dados estruturados e não estruturados. A segunda forma é por meio do HDFS, utilizado em torno de 75% das vezes para armazenamento de dados em Data Lakes. No entanto, de acordo com Sawadogo e Darmont (2021), o HDFS sozinho é geralmente insuficiente para lidar com todos os formatos de dados, especialmente dados estruturados. Deste modo, segundo os autores, o ideal é que seja feita uma combinação do HDFS com bancos de dados relacionais e/ou NoSQL."
- **Processamento de dados:** Para o processamento dos dados MapReduce e Apache Spark são comumente utilizados, a diferença entre os dois é que o Spark realiza processamento 100% em memória o que o torna a melhor escolha para processamento de dados em tempo real (Moraes, 2022).
- **Acesso aos dados:** O acesso aos dados tem que ser feito de acordo com os tipos de dados presentes e as ferramentas utilizadas para seu

armazenamento. Em banco de dados relacionais utiliza-se SQL, ou JSONiq para bancos de dados XML. Em casos em que temos dados heterogêneos a utilização Spark SQL ou SQL++ se torna mais interessante, já que recuperam dados tanto de bancos relacionais como de dados semi estruturados em formato JSON.

Objetivo Principal

Nos últimos anos, os data lakes evoluíram como plataformas para gerenciar e analisar grandes quantidades de dados. Eles são usados nas mais diversas áreas, como saúde, aviação, educação entre outras, e permitem que as empresas explorem o valor de seus dados usando análises avançadas, como aprendizado de máquina. Para isso, os dados heterogêneos são armazenados em sua forma bruta, possibilitando sua análise sem casos de uso pré-definidos (GIEBLER *et al.*, 2021).

Dessa forma podemos afirmar que os Data Lakes tem como objetivo principal disponibilizar a maior quantidade de dados gerado pela empresa em seu estado bruto, para que dessa forma profissionais qualificados possam extrair os dados necessários para a realização das mais diversas análises, assim auxiliando nas tomadas de decisões da empresa.

Além disso, podem-se tornar um ambiente de teste e aprendizado, tornando-se uma plataforma para experimentos. Ciências de dados podem realizar análises sobre dados inalterados e criarem protótipos para programas de análises.

Vale mencionar a pesquisa realizada pela empresa Aberdeen, que aponta que esse tipo de repositório, além de ajudar as empresas a explorarem o potencial dos mais diversos tipos de dados, também pode ajudar a tornar os sistemas legados mais eficientes, transferindo a capacidade para infraestrutura mais novas e flexíveis (ABERDEEN,2017).

Vantagens do Data Lake

A utilização traz diversas vantagens, abaixo detalhamento de algumas das principais:

- **Dados em estado bruto:** Os dados são salvos da forma que são gerados, pois a formatação dos dados é realizada apenas quando se está utilizando os dados para análise. Assim oferece uma maleabilidade para as regras de negócio que serão aplicadas sobre os dados (Betrybe, 2022).
- **Escalabilidade:** O Hadoop é uma plataforma bastante utilizada para implementação de *data lakes,* visto que permite uma alta escalabilidade utilizando computação distribuída com foco em clusters e conseguindo processar um grande volume de dados mantendo a tolerância a falhas. Desse modo a escalabilidade é um ponto importante para esse sistema (MACIEL, 2022).
- **Versatilidade:** Como já informado antes nesse texto, atualmente é gerado dados em diversos tipos de formato, e um dos pontos fortes dos data lake é justamente a capacidade de armazenar de forma rápida todos os tipos de dados por não fazer pré-processamento dos mesmos.
- **Análise avançada:** Um data lake se destaca por utilizar a disponibilidade de grandes quantidades de dados coerentes junto com algoritmos de aprendizado profundo. Ele ajuda na análise de decisões em tempo real, ao contrário de outros métodos utilizados no mercado (Betrybe, 2022).
- **Multilinguagens:** Por padrão o suporte é apenas para SQL, porém em casos específicos de uso avançado, é possível implementar suporte a outras linguagens, como por exemplo a utilização de Spark MLlib para aprendizado de máquina que traz suporte a uma variedade de linguagens.

Assim os data lakes se destacam por sua alta escalabilidade, custo reduzido em comparação com outras abordagens, como por exemplo *data warehouses*, além de permitir análises mais profundas e completas por armazenar diversos tipos de dados.

Desvantagens do Data Lake

Depende da organização que está implementando o *data lakes* e como são seus dados, pode haver alguma desvantagem que impossibilite seu uso, abaixo está listado algumas delas:

- **Implantação complexa:** Criar um data lake na nuvem é uma tarefa relativamente fácil, embora existam soluções locais, como Hadoop e Splunk, essas podem ser consideravelmente mais complexas (Betrybe, 2022).
- **Curva de aprendizado:** Há uma curva de aprendizado ao se lidar com *data lake,* pois como os dados não estão estruturados e apresentados de forma clara, será necessário o treinamento da equipe para utilizar a ferramenta de forma adequada, ou realizar a contratação de novos colaboradores que tenham experiência com as tecnologias utilizadas.
- **Acúmulo de informações:** Como o data lake retém todos os dados da empresa em um único repositório sem tratamento, é importante que se tenha cuidado com quais dados estão sendo salvos, para não acabar criando um pântano que dificulta o acesso a dados relevantes (NOGUEIRA; ROMDHANE, 2018).
- **Capacidade de processamento:** A capacidade de processamento do sistema tem que ser analisada, como os dados estão armazenados de forma bruta, o que agiliza e economiza recursos na inserção de dados no repositório, será necessário um maior esforço computacional para extrair informações relevantes (Betrybe, 2022).
- **Metadados:** Os metadados facilitam a catalogação e a indexação dos dados no *data lake*. Ao registrar informações sobre os atributos dos dados, como nomes de colunas, tipos de dados e relacionamentos, os metadados permitem que os dados sejam pesquisados e localizados de forma eficiente. Isso melhora a capacidade de descoberta e recuperação de dados, economizando tempo e esforço ao procurar por informações específicas. Dessa forma, metadados ausentes ou incompletos podem

causar grandes problemas na utilização do *data lake* (NOGUEIRA; ROMDHANE, 2018).

Desta forma, para contornar esses problemas é necessário que o data lake tenha um controle de acesso, consistência semântica e uma supervisão periódica acerca das entradas, para garantir que esteja sendo bem utilizado e suprindo as demandas do público alvo.

Exemplos de uso

Segundo o caso de estudo realizado pela Amazon Web Services (2021), a Coca-Cola Andina, uma empresa chilena que produz e comercializa produtos licenciados da The Coca-Cola Company na américa latina, obteve um aumento de produtividade de de 80% em suas análises após a implementação de um *data lake* utilizando produtos AWS da Amazon.

Foi utilizado o Amazon S3 para o armazenamento dos dados em seu estado bruto, que retém 95% dos dados de interesse da empresa que possui cerca de 17.500 funcionários e aproximadamente 267.000 clientes. A união desses dados possibilitou que a empresa utilizasse dados coerentes e seguros (Amazon Web Services, 2021).

Com esses dados em mãos foi possível aplicar análises avançadas, gerar relatórios de qualidade por meio de machine learning e outras ferramentas de análise. Assim permitindo que a empresa tivesse uma melhora na sua receita, já que foi possível melhorar a eficiência das promoções e evitar falta de produto nos estoques. Tudo isso graças a melhora da equipe de análise que contou com cerca de 300 horas de treinamento realizado pela equipe da AWS (Amazon Web Services, 2021) .

Base de Dados

A base de dados escolhida foi obtida no site kaggle e contém 3 tabelas, sendo elas Books, Ratings e Users. Os dados foram retirados do site

Book-Crossing em 2004. Nela estão presentes dados sobre os livros, os usuários identificados apenas por ID, idade e localização e a pontuação que o usuários atribuiu a determinados livros. A tabela de ratings conta com 1.148.780 tuplas.

Diagrama Entidade-Relacionamento

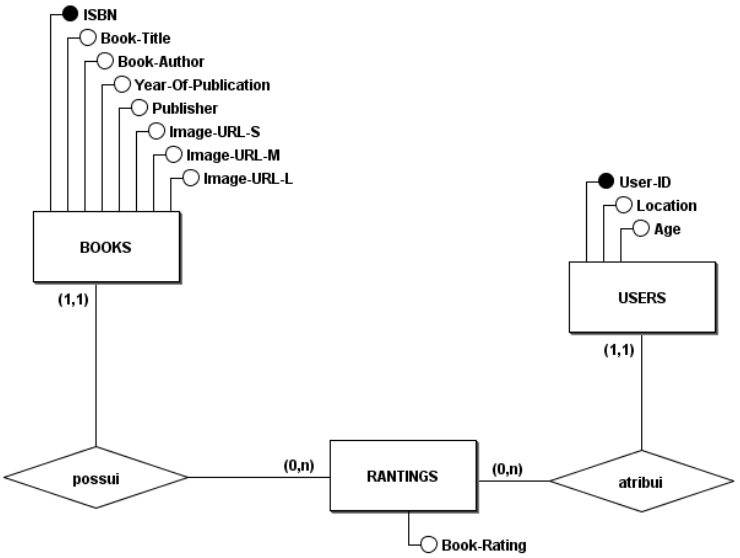
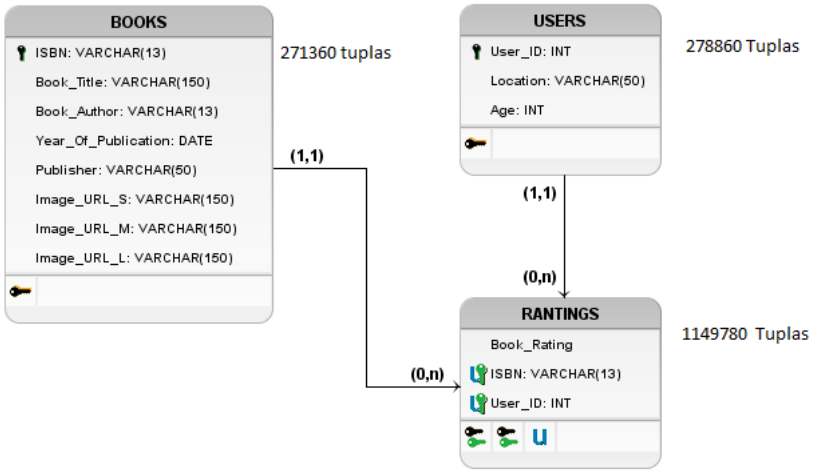


Diagrama Lógico de Dados



É possível encontrar a base de dados em: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset

Bibliografia

ABERDEEN. Angling for insight in today's data lake. 2017. Disponível em: https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/berdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf. Acesso em: 11 de jun. de 2023.

AMAZON WEB SERVICES. Coca-Cola Andina Case Study. 2021. Disponível em:https://aws.amazon.com/pt/solutions/case-studies/coca-cola-andina-case-study/. Acesso em: 12 jun. 2023.

GIEBLER, Corinna *et al*. The Data Lake Architecture Framework. BTW 2021, 2021.

MACIEL, Vitória Maria da Silva. Um modelo de suporte para conformidade de data lake com a LGPD. 2022. Dissertação de Mestrado. Universidade Federal de Pernambuco.

MARTINS, Vinicius. Data lake: conheça a fonte do Big data e saiba quais suas vantagens. 21 de setembro de 2022, Betrybe, Disponível em: <https://blog.betrybe.com/tecnologia/data-lake/#:~:text=Data%20lakes%20n%C3%A3o%20estruturados&text=Os%20dados%20n%C3%A3o%20estruturados%20tamb%C3%A9m,ser%C3%A3o%20cruciais%20para%20as%20empresas> Acesso em: 11 de jun. de 2023.

MORAIS, FLAVIO LOPES DE, Termos e Definições - Governança de Dados. 06 de dezembro de 2022, Base de Conhecimento - UFLA, Disponível em: <https://kb.ufla.br/books/termos-e-definicoes-governanca-de-dados>Acesso em: 12 jun. 2023.

NOGUEIRA, I. D.; ROMDHANE, M. Modeling Data Lake Metadata with a Data Vault. 22nd International Database Engineering & Applications Symposium, Villa San Giovanni, p. 253 – 261, Jun 2018. Disponível em:

<https://arxiv.org/abs/1807.04035>. Acesso em: 11 de jun. 2023.

VIDA, Edinilson da Silva *et al*. Data warehouse. Porto Alegre: Grupo A, 2021. E-book. ISBN 9786556901916. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786556901916/. Acesso em: 11 jun. 2023.