UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Data Lake

Eduardo Maia Rezende - 180119231

Brasília, DF

2023

1. Introdução

Com o progresso contínuo da tecnologia e a digitalização de documentos, a necessidade de armazenamento de arquivos tem se tornado cada vez mais relevante. É imprescindível dispor de um local adequado para preservar arquivos antigos e informações de importância crítica. Como resposta a essa demanda, repositórios digitais foram desenvolvidos na internet para atender às necessidades de armazenamento. Ao longo do tempo, empresas de grande porte passaram a demonstrar maior preocupação com a capacidade desses repositórios, uma vez que eles se tornaram cada vez mais volumosos devido à crescente quantidade de dados corporativos. Essas organizações manifestaram a intenção de evitar a perda de dados relevantes para abrir espaço para novas informações, dado que os dados antigos também possuíam valor significativo. Dessa maneira, foram criados grandes repositórios, como o Data Warehouse, e posteriormente o Data Lake, cada um com suas particularidades distintas.

2. Data Lake

O Data Lake é um sistema ou repositório projetado para armazenar uma ampla variedade de dados, incluindo dados estruturados provenientes de bancos de dados relacionais, dados semi-estruturados como planilhas, arquivos .csv, .xml e .json, dados não estruturados como e-mails, arquivos de texto .txt e PDFs, e também dados binários, tais como imagens e sons.

O conceito de "Data Lake" ou "Lago de Dados" se refere a um repositório de informações que utiliza a metáfora de um lago para ilustrar a maneira como os dados são armazenados. Assim como um lago natural é alimentado por diversos rios e afluentes, um Data Lake é abastecido por diferentes fontes de dados, permitindo a sua reunião em um único ambiente (ARAUJO, 2022).

Uma das características distintivas do Data Lake é a capacidade de armazenar os dados no formato nativo ou "raw", sem limitações quanto ao tamanho. Essa flexibilidade permite uma maior versatilidade na análise dos dados, possibilitando a execução de diferentes tipos de análises, desde a criação de painéis e visualizações até o processamento de grandes volumes de dados, análises em tempo real e a aplicação de técnicas de aprendizado de máquina para obter insights valiosos e embasar tomadas de decisão.

2.1. Início do Data Lake

Supostamente, James Dixon, que na época era diretor de tecnologia da Pentaho, teria introduzido o termo "data lake" como uma contraposição ao conceito de "data mart", que representa um repositório de menor escala contendo atributos de interesse derivados de dados brutos. Dixon argumentou que os data marts apresentam diversos problemas inerentes, como a fragmentação das informações. A PricewaterhouseCoopers, por sua vez, afirmou que os data lakes têm a capacidade de "eliminar os silos de dados". Em seu estudo sobre data lakes, a empresa observou que as organizações estavam começando a extrair e consolidar dados em um único repositório baseado na tecnologia Hadoop. Atualmente, diversas empresas, como Hortonworks, Google, Oracle, Microsoft, Zaloni, Teradata, Impetus Technologies, Cloudera e Amazon, oferecem soluções relacionadas aos data lakes.(Wikipédia,2023)

Levando em consideração os problemas inerentes aos data marts, Dixon desenvolveu uma solução denominada Data Lake, que consiste em um reservatório onde diversos tipos de dados provenientes de várias fontes são armazenados. Nesse ambiente, os usuários têm a possibilidade de acessar, explorar ou obter dados de forma flexível, fazendo analogia com a atividade de mergulho ou pesca.

3. Data Lake e Outras Tecnologias

3.1 Data Warehouse

Dependendo dos requisitos específicos, uma organização típica pode necessitar tanto de um data warehouse quanto de um data lake, pois esses dois recursos atendem a diferentes necessidades e cenários de uso.

Um data warehouse consiste em um banco de dados otimizado para a análise de dados relacionais provenientes de sistemas transacionais e aplicações de negócios. A estrutura e o esquema dos dados são definidos previamente para garantir consultas SQL rápidas, cujos resultados são geralmente utilizados em relatórios e análises operacionais. Os dados passam por processos de limpeza, enriquecimento e transformação, a fim de atuarem como a "fonte única da verdade" na qual os usuários podem confiar.

Por outro lado, um data lake difere do data warehouse, pois armazena tanto dados relacionais provenientes de aplicações de negócios quanto dados não relacionais provenientes de aplicativos móveis, dispositivos IoT e mídias sociais. A estrutura dos dados ou o esquema não é definido no momento da captura dos dados. Isso permite que todos os dados sejam armazenados sem a necessidade de um design prévio detalhado ou o conhecimento prévio sobre quais perguntas precisarão ser respondidas no futuro. Diversas formas de

análise, como consultas SQL, análise de big data, pesquisa de texto completo, análise em tempo real e aprendizado de máquina, podem ser aplicadas aos dados para descobrir insights relevantes.

À medida que as organizações que possuem data warehouses percebem os benefícios dos data lakes, elas estão evoluindo suas estruturas para incorporar data lakes, habilitando assim recursos avançados de consulta, casos de uso de ciência de dados e capacidades avançadas para descoberta de novos modelos de informação. Essa evolução é chamada pela Gartner de "Solução de Gerenciamento de Dados para Análise" ou "DMSA" (Data Management Solution for Analytics).

Figura 1 – Data Warehouse versus Data Lake

Características	Data warehouse	Data lake
Dados	Relacionais de sistemas transacionais, bancos de dados operacionais e aplicações de linha de negócios	Não relacionais e relacionais de dispositivos de IoT, sites, aplicações móveis, mídia social e aplicações corporativas
Esquema	Definido antes da implementação do DW (esquema na gravação)	Gravado no momento da análise (esquema na leitura)
Preço/performance	Resultados de consulta mais rápidos, usando armazenamento de maior custo	Resultados de consulta ficando mais rápidos, usando armazenamento de menor custo
Qualidade dos dados	Dados altamente selecionados, que representam a versão central da verdade	Quaisquer dados, selecionados ou não (ou seja, dados brutos)
Usuários	Analistas de negócios	Cientistas de dados, desenvolvedores de dados e analistas de negócios (usando dados selecionados)
Análises	Geração de relatórios em lote, BI e visualizações	Machine learning, análises preditivas, descoberta de dados e criação de perfis

Fonte: AWS Amazon, 2023

3.2 Big Data

Devido à sua natureza centrada em dados, é comum haver confusão entre os termos "Data Lake" e "Big Data". Para esclarecer essa distinção, é importante compreender que o Data Lake é um conceito de negócio, enquanto o Big Data é um conceito relacionado à tecnologia.

O Data Lake é um ambiente que tem a capacidade de armazenar e gerenciar diversos tipos de dados, independentemente de sua estrutura ou formato. Ele serve como um repositório centralizado onde todas as informações relevantes da organização podem ser armazenadas. O conceito de Data Lake se baseia na ideia de preservar todos os dados brutos, sem a necessidade de definir previamente a estrutura ou o esquema dos dados. Isso permite que a empresa explore esses dados de maneira flexível e realize análises avançadas posteriormente, quando necessário.

Por outro lado, o Big Data é um conjunto de técnicas e ferramentas desenvolvidas para lidar com o processamento e a análise de grandes volumes

de dados, especialmente aqueles que não podem ser tratados de maneira eficiente com os métodos tradicionais. O objetivo principal do Big Data é extrair insights significativos e relevantes a partir desses dados massivos, por meio de técnicas como mineração de dados, aprendizado de máquina, análise preditiva e processamento em tempo real.

Assim, enquanto o Data Lake se concentra na coleta e armazenamento de dados de forma abrangente, o Big Data está relacionado às abordagens e técnicas para a análise desses dados em busca de conhecimento valioso para a organização. Portanto, embora relacionados, os conceitos de Data Lake e Big Data abrangem diferentes aspectos no ecossistema de dados de uma empresa.

4. Vantagens e Desvantagens

Cada fenômeno possui inerentemente vantagens e desvantagens, e o Data Lake não é uma exceção a essa regra. É essencial reconhecer que o Data Lake transcende a mera concepção de um repositório extenso de dados, uma vez que desempenha uma variedade de funções adicionais. Entre essas funções, destacam-se a análise de dados, o aprendizado de máquina e a organização de conjuntos de dados volumosos, comumente referidos como Big Data. Diante desse contexto, é pertinente apresentar algumas considerações acerca das vantagens e desvantagens inerentes ao Data Lake.

4.1 Vantagens

A utilização de um Data Lake oferece várias vantagens significativas para as organizações. Aqui estão algumas das principais vantagens associadas ao uso de um Data Lake:

- **Armazenamento centralizado de dados:** O Data Lake permite armazenar grandes volumes de dados brutos em um único local centralizado. Isso facilita o acesso e a busca por dados de várias fontes, eliminando a necessidade de procurar dados em diferentes sistemas ou silos. Os dados são armazenados em sua forma bruta, preservando seu contexto original e permitindo uma ampla gama de análises e explorações.

- **Escalabilidade e flexibilidade:** Os Data Lakes são projetados para lidar com grandes volumes de dados, permitindo escalabilidade horizontal à medida que a quantidade de dados aumenta. Eles são altamente flexíveis, permitindo que os dados sejam adicionados, transformados e processados conforme necessário. Isso torna mais fácil para as organizações lidar com a variedade e a complexidade crescentes dos dados, além de se adaptar a requisitos em evolução.

- **Diversidade de dados:** Os Data Lakes podem lidar com diversos tipos

de dados, sejam estruturados, não estruturados ou semi-estruturados. Isso inclui dados de várias fontes, como bancos de dados relacionais, logs de servidores, arquivos CSV, documentos de texto, fluxos de sensores IoT, dados de redes sociais, entre outros. A capacidade de armazenar e analisar uma ampla variedade de dados permite a obtenção de insights mais abrangentes e valiosos.

- **Análise avançada:** Com o Data Lake, é possível realizar análises avançadas e sofisticadas nos dados brutos. A natureza flexível do Data Lake permite que os usuários apliquem diferentes técnicas de análise, como mineração de dados, aprendizado de máquina e processamento de linguagem natural. Além disso, os Data Lakes podem integrar-se a ferramentas e ecossistemas de análise, como frameworks de big data, para processamento e extração de insights em larga escala.

- **Descoberta de insights e inovação:** O Data Lake fornece uma plataforma para a descoberta de insights e a geração de ideias inovadoras. Ao armazenar uma grande quantidade de dados em sua forma bruta, os Data Lakes permitem que as organizações explorem e experimentem diferentes abordagens analíticas. Isso pode levar à descoberta de padrões ocultos, identificação de tendências emergentes e insights valiosos para impulsionar a tomada de decisões e a inovação.

- **Baixo custo inicial:** Em comparação com as soluções tradicionais de data warehousing, os Data Lakes geralmente apresentam um custo inicial mais baixo. Isso ocorre porque eles aproveitam a tecnologia de armazenamento em nuvem e sistemas distribuídos de código aberto, como o Apache Hadoop. Além disso, os Data Lakes têm um modelo de armazenamento "pague à medida que cresce", o que significa que as organizações podem dimensionar os custos de armazenamento de acordo com suas necessidades.

Essas vantagens tornam os Data Lakes uma solução atraente para organizações que desejam aproveitar ao máximo seus dados, permitindo análises mais abrangentes, descoberta de insights e tomada de decisões fundamentadas.

4.2 Desvantagens

Embora os Data Lakes tenham muitas vantagens, também existem algumas desvantagens a serem consideradas ao utilizá-los. Aqui estão algumas das principais desvantagens associadas ao uso de um Data Lake:

- **Governança de dados:** Um desafio comum dos Data Lakes é a governança de dados. Como os Data Lakes armazenam dados brutos e não estruturados, pode ser difícil impor políticas e padrões de governança, como qualidade de dados, conformidade, segurança e privacidade. A falta de governança adequada pode levar a problemas como dados inconsistentes, duplicados ou incorretos.

- **Complexidade de dados:** Os Data Lakes permitem armazenar dados de várias fontes e formatos, o que pode resultar em uma grande variedade de dados complexos e não estruturados. Isso pode dificultar a descoberta, a organização e a compreensão dos dados, especialmente para usuários não técnicos. A complexidade dos dados pode levar a desafios na hora de realizar análises e obter insights significativos.

- **Necessidade de pré-processamento:** Embora o Data Lake armazene dados brutos, muitas vezes é necessário realizar um pré-processamento nos dados antes de usá-los para análises ou outras finalidades. Isso pode envolver a extração, transformação e carga (ETL) dos dados para estruturá-los adequadamente. O pré-processamento pode exigir tempo e recursos significativos, especialmente quando se lida com grandes volumes de dados.

- **Requisitos de habilidades técnicas:** Para aproveitar ao máximo um Data Lake, é necessário ter habilidades técnicas e conhecimento em ferramentas de processamento de dados, programação e análise. A equipe responsável pelo Data Lake precisa ter conhecimentos em áreas como engenharia de dados, ciência de dados e governança de dados. A falta de habilidades técnicas adequadas pode dificultar a utilização eficaz do Data Lake.

- **Custos de armazenamento e processamento:** Embora os custos de armazenamento tenham diminuído nos últimos anos, o armazenamento e o processamento de grandes volumes de dados em um Data Lake ainda podem representar um investimento significativo. Além do armazenamento, também é necessário considerar os custos de processamento e análise dos dados, especialmente quando se lida com cargas de trabalho intensivas em recursos.

- **Segurança e privacidade:** À medida que um Data Lake centraliza uma grande quantidade de dados em um único local, a segurança e a privacidade dos dados se tornam preocupações importantes. É necessário implementar medidas de segurança robustas para proteger os dados contra acessos não autorizados ou violações. Também é essencial garantir que os dados sensíveis estejam adequadamente protegidos e que sejam cumpridos os regulamentos de privacidade de dados aplicáveis.

É importante considerar essas desvantagens e avaliar cuidadosamente os requisitos e as necessidades da organização antes de adotar um Data Lake. Embora possa ser uma solução valiosa para muitos casos de uso, é fundamental abordar essas desvantagens de maneira adequada para garantir o sucesso da implementação do Data Lake.

5. Casos de Sucesso com Data Lake

Uma notável empresa que adota o sistema de Data Lake é o Google, por meio de sua "suíte de computação" chamada Google Cloud, que oferece uma variedade de ferramentas aos usuários. Dentre essas ferramentas, destacam-se o BigQuery e o Dataproc para armazenamento e análise de dados, o DataFlow para processamento e o Cloud Storage para armazenamento.

Outra empresa proeminente que proporciona um ambiente de Data Lake aos usuários é a AWS (Amazon Web Services, Inc.). Conforme declarado no próprio site da empresa, a AWS oferece o conjunto mais abrangente, seguro, escalável e econômico de serviços para permitir aos clientes criar um Data Lake na nuvem, analisar todos os tipos de dados, incluindo dados provenientes de dispositivos de IoT, por meio de diversas abordagens analíticas, como o aprendizado de máquina. Como resultado, um número significativo de organizações opta por executar seus Data Lakes e análises na AWS, incluindo clientes de renome como NETFLIX, Zillow, NASDAQ, Yelp, iRobot e FINRA, que confiam na AWS para a execução de suas cargas de trabalho de análise crítica de negócios.

6. Referências Bibliográficas

Data Lake e Big Data: entenda a diferença! Disponível em: https://blog.engdb.com.br/data-lake/#:~:text=Como%20est%C3%A3o%20relacionados%20a%20dados. Acesso em: 11 jun. 2023.

Amazon Web Services, Inc. ou suas afiliadas. Datalake and Analytics. 2022. Disponível em: https://aws.amazon.com/pt/big-data/datalakes-and-analytics/what-is-a-data-lake/. Acesso em: 18 de jan., de 2023.

ARAUJO, Juarez. Constância da Silva. Conheça a diferença entre Data Lake e Data Warehouse. 17 de maio de 2022. Disponível em: https://blog.dbacorp.com.br/2022/05/17/data-lake-data-warehouse/. Acesso em:18 de jan. de 2023.

DIXON, James. Pentaho, Hadoop, and Data Lakes, 14 de Outubro de 2010. Disponível em: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. Acesso em: 19 de jan. de 2023. RAU, Isabele Aurora Cândido Vitorino .

DATA LAKE: UMA NOVA ABORDAGEM PARA O ARMAZENAMENTO DE DADOS. Florianópolis: Universidade do Sul de Santa Catarina, 2021. Disponível em: https://repositorio.animaeducacao.com.br/bitstream/ANIMA/13790/1/versao-final-tcc%20%281%29.pdf. Acesso em: 20 de jan., de 2023.

GORELIK, Alex. The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science. 1ª edição. O'Reilly Media. 2019. LOGUNOVA, Inna. Data Warehouses vs. Data Lakes vs. Data Lakehouses, 4 de Outubro de 2022. Disponível em: https://serokell.io/blog/data-warehouse-vs-lake-vs-lakehouse. Acesso em: 21 de jan. de 2023.

Oracle and/or its affiliates. The Evolution of Big Data and the Future of the Data Lakehouse, 2022. Disponível em: https://www.oracle.com/br/a/ocom/docs/bigdata/big-data-evolution.pdf. Acesso em: 20 de jan., de 2023.

Oracle and/or its affiliates. What is Big Data, 2022. Disponivel em: https://www.oracle.com/br/big-data/what-is-big-data/#history. Acesso em: 20 de jan. de 2023.

WIKIPEDIA. DataLake. Wikipedia, 2022. Disponível https://en.wikipedia.org/wiki/Data_lake. Acesso em 18 de jan., de 2023. em:

WIKIPEDIA. Information Silo. Wikipedia, 2022. Disponível em: https://en.wikipedia.org/wiki/Information_silo. Acesso em 18 de jan., de 2023.

WIKIPEDIA. DataMart. Wikipedia, 2022. Disponível https://en.wikipedia.org/wiki/Data_mart. Acesso em 18 de jan., de 2023.

Base de Dados

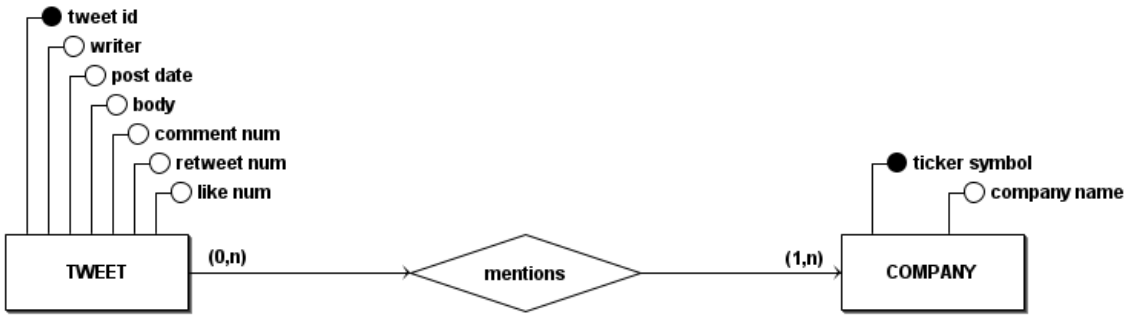
Tweets about the Top Companies from 2015 to 2020

Esta base de dados foi criada como parte de um artigo publicado na Conferência Internacional IEEE de Big Data em 2020, na 6ª Sessão Especial sobre Mineração de Dados Inteligente. Seu objetivo principal é determinar possíveis especuladores e influenciadores no mercado de ações. O conjunto de dados é composto por tweets relacionados às empresas Amazon, Apple, Google, Microsoft e Tesla, identificados pelos seus respectivos códigos de ações.

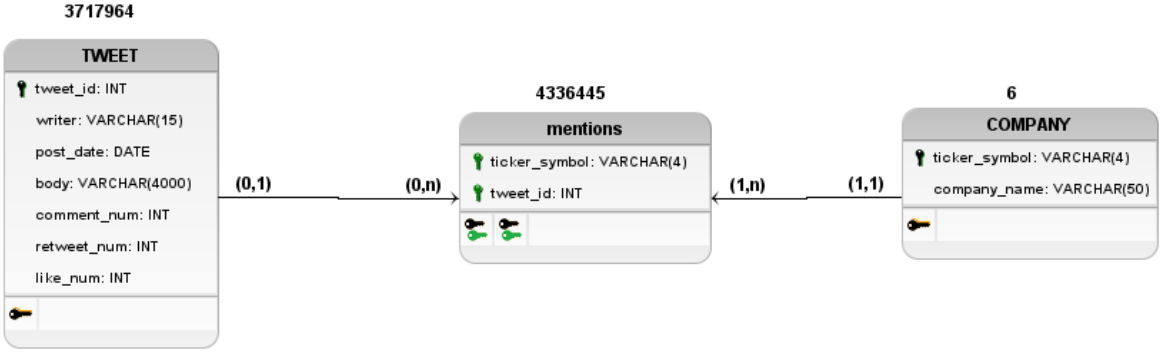
O conjunto de dados contém mais de 3 milhões de tweets exclusivos, cada um com informações como ID do tweet, autor, data de publicação, texto do tweet e o número de comentários, curtidas e retweets associados a cada tweet relacionado à empresa correspondente.

Essa base de dados é valiosa para aqueles interessados em analisar os tweets que foram escritos sobre as empresas mencionadas e sua relação com o desempenho do mercado de ações. Pode ser utilizada para pesquisas relacionadas à detecção de especuladores e influenciadores no mercado financeiro, bem como para análises de sentimento, estudos de comportamento de investidores e investigações de impacto das redes sociais nas tendências do mercado de ações.

DER



DLD


3717964
TWEET
tweet_id: INT
writer: VARCHAR(15)
post_date: DATE
body: VARCHAR(4000)
comment_num: INT
retweet_num: INT
like_num: INT
(0,1)
(0,n)
4336445
mentions
ticker_symbol: VARCHAR(4)
tweet_id: INT
(1,n)
(1,1)
6
COMPANY
ticker_symbol: VARCHAR(4)
company_name: VARCHAR(50)