UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Data Lake

Pedro Lucas Cassiano Martins - 190036567

Brasília, DF

2022

1 Introdução

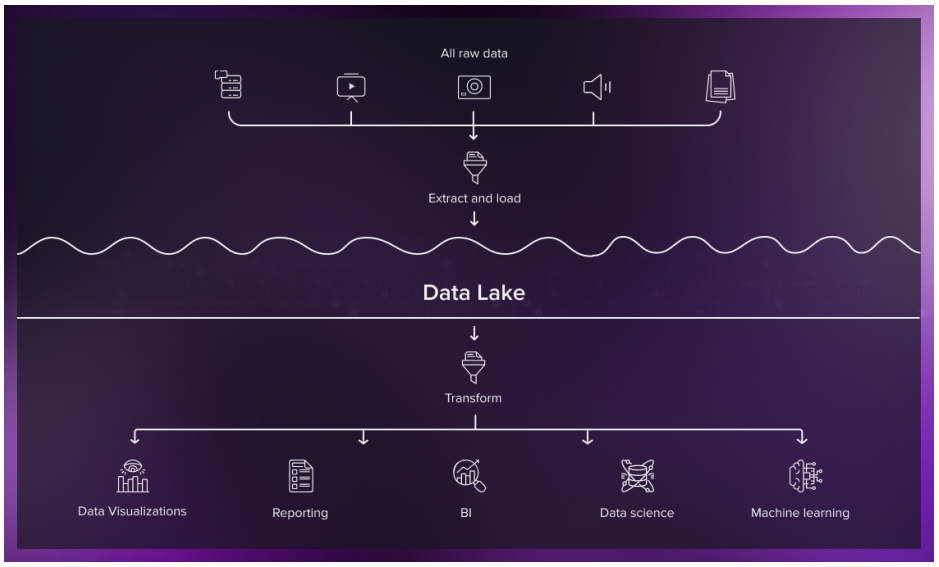
Com o avanço da tecnologia e com a digitalização de arquivos foi-se desenvolvendo uma grande importância para o armazenamento de arquivos, de ter um lugar para guardar os seus arquivos antigos e arquivos importantes. Com isso foi-se criando repositórios na internet para fazer esse armazenamento. O tempo foi passando e grandes empresas foram se preocupando mais com o tamanho desses armazéns pois foram se enchendo mais e mais com os dados da empresa e elas não queriam perder esses dados para liberar mais espaço para novos pois os antigos também eram importantes. Assim criou-se então grandes repositórios como o *Data Warehouse* e um tempo depois o *Data Lake*, cada um com suas diferenças.

2 Data Lake

O nome "Data Lake" (ou lago de dados, em português) se refere a um lago por, além de ser um grande repositório que guarda seus dados de uma forma natural como um grande corpo d'água natural, ser abastecido por diferentes base de dados como um lago real que é abastecido por rios e afluentes (ARAUJO, 2022).

Um Data Lake é um sistema ou repositório que permite guardar/armazenar vários tipos de dados como dados estruturados (dados de bancos de dados relacionais), semi-estruturados (planilhas, arquivos .csv, .xml, .json), não-estruturados (emails, .txt, PDFs) e dados binários (imagens, sons, etc). Esses dados armazenados podem ser guardados no formato nativo/natural ou *raw,* em inglês, e sem limitação de tamanho, o que é um destaque do Data Lake.

Figura 1 - Representação Visual de um Data Lake



Fonte: Logunova, 2022

É muito utilizado em ferramentas de analytics e Business Inteligence(BI) por causa de seu grande armazenamento e por poder guardar arquivos nativos, pode-se “executar diferentes tipos de análise, desde painéis e visualizações até processamento de big data, análise em tempo real e machine learning para orientar melhores decisões” (Amazon, 2022).

2.1 Origem do Data Lake

O, então Diretor de Tecnologia (CTO) da Pentaho James Dixon, patenteou o termo *Data Lake* em 2011 como contraste ao termo *Data Mart*, um repositório menor que serve para armazenar arquivos nativos específico de ambientes com *Data Warehouse* (um repositório central muito usado em Business Intelligence). Dixon argumentava que os *data marts* tinham vários problemas próprios como por exemplo silo de informação (information silo) que é um sistema isolado de gerenciamento onde as informações de um sistema ou subsistemas não conseguem se comunicar entre si e fazer operações reciprocar com outros subsistemas que são ou deveriam ser relacionadas (Wikipedia, 2023).

De acordo com seu blog “James Dixon’s Blog” escrito e publicado em 2010, o CTO James Dixon, depois da empresa em que trabalha Pentaho anunciar a disponibilidade do primeiro release utilizando Hadoop (uma plataforma de software que utiliza Java), conversou com outras empresas que usavam o mesmo software e percebeu alguns temas e assuntos em comuns dessas conversas:

- 80-90% das empresas estão lidando com dados estruturados ou semi-estruturados (mas não os não estruturados);
- A fonte dos dados é tipicamente uma única aplicação ou sistema;
- Os dados é tipicamente sub-transacional ou não-transacional;
- Existem algumas questões conhecidas para perguntar sobre os dados;
- Existem muitas questões desconhecidas que irão ser levantadas no futuro;
- Existem várias comunidades que possuem perguntas sobre os dados;
- Os dados são de uma escala ou volume diário tão grande que não irão caber tecnicamente e/ou economicamente em um RDBMS (Banco de Dados Relacional).

Ele também diz em seu blog que existem vários problemas em reportar e analisar os dados utilizando o Data Mart:

- Apenas um subconjunto dos atributos são examinados então apenas algumas questões pré-determinadas podem ser respondidas;
- Os dados são agregados, então a visibilidade de camadas mais profundas são perdidas.

Com esses problemas em mente, Dixon criou uma solução e chamou o

conceito de Data Lake, um lago onde vários tipos de dados de vários lugares são armazenados e usuários podem entrar, mergulhar ou pescar dados (Dixon, 2010).

3 Big Data

Dados que possuem maior **variedade** e que podem chegar em **volumes** crescentes e com mais **velocidades** são chamados de *Big Data*. Variedade, volume e velocidade são conhecidos como os três Vs:

Variedade: refere-se aos vários tipos de dados estruturados, semi-estruturados e não-estruturados;

Volume: é o volume do dado podendo chegar em grandes volumes como o petabyte ou o zettabyte;

Velocidade: trata-se da taxa mais rápida na qual os dados são recebidos. Alguns dados são operados em tempo real ou em quase tempo real e exigem avaliação e ação em tempo real (Oracle, 2022).

Big Data é um termo já utilizado há muito tempo desde quando os primeiros dados começaram a ser armazenados, porém o conceito de Big Data é relativamente novo. Foi por volta de 2005 que as pessoas começaram a perceber a quantidade enorme de dados gerados por usuários em redes sociais e na internet, então foi-se criando vários software e produtos para trabalhar com essa grande quantidade de dados e com o tempo o volume de Big Data foi crescendo disparadamente. O Big Data então foi criado para trabalhar com esses dados de usuários na internet, principalmente para trabalhar com Internet of Things (IoTs) ou Internet das Coisas, em português.

4 Data Science

Data Science, ou ciência dos dados, é o estudo da organização dos dados e informações inerentes sobre um negócio envolvendo todas as visões que cercam um determinado assunto desse negócio.

A ciência dos dados é uma ciência que estuda dados e informações, sua organização, processo de captura, transformação , geração e análises aplicadas. É uma ciência que abrange várias áreas de estudo como matemática, computação, estatística, organização e negócios.

O Data Science surgiu da necessidade de armazenamento, organização, processamento e análise do *Big Data*. O Data Science, diferentemente do BI, tenta procurar prever ao analisar os dados, o que nos leva ao Data Lake, usado muito pelos cientistas de dados como ferramenta para essas análises.

5 Vantagens e Desvantagens

Tudo possui pelo menos uma vantagem e uma desvantagem, nada é perfeito e por isso trago algumas vantagens e desvantagens do Data Lake.

Tome em mente que Data Lake não é apenas um repositório gigantesco de dados, mas sim serve também para várias outras funções: análise de dados, machine learning e organização de Big Data são alguns exemplos de utilização do Data Lake.

5.1 Vantagens

- **Machine Learning**

O Data Lake tem uma construção voltada para o aproveitamento dos dados em conjunto dos mais avançados sistemas de análise, onde os modelos são criados para prever resultados possíveis e sugerir um encadeamento de ações prescritas para alcançar o melhor resultado.

Isso permite vários avanços, especialmente o uso dos dados em aplicações com machine learning — como por exemplo, baseando-se no histórico de informações.

- **Análises**

Com tamanho potencial de armazenar dados (por longo período de

tempo), as empresas poderão aprimorar suas funções analíticas e aproveitar todo potencial de sistemas de BI e ciência de dados. Os data lakes permitem que você execute análises sem a necessidade de mover seus dados para um sistema de análise separado, podendo fazer essa análise localmente.

- **Os valores agregados por um Data Lake**

Ao possuir vários dados oriundos de um maior número de fontes, sua empresa impulsiona suas operações e agrega valor à tomada de decisão, melhorando áreas como: interações com os clientes, inovação e P&D, bem como sua eficiência operacional.

5.2 Desvantagens

- **Menor velocidade na análise de dados**

Porque o Data Lake armazena dados não-estruturados a velocidade de análise dos dados não é tão grande quanto aos de Data Warehouses e outros tipos de repositórios que não armazenam esses tipos de dados.

- **Suporte ruim de casos de BI**

A integração com ferramentas de business intelligence e análise pode ser difícil se os data lakes não forem bem gerenciados. Além disso, os resultados da consulta podem não ser precisos devido à falta de estruturas de dados consistentes (Logunova, 2022).

- **Sem supervisão**

O principal desafio com uma arquitetura de data lake é que os dados nativos/brutos são armazenados sem supervisão do conteúdo. Para que os dados sejam utilizados, um Data Lake precisa ter mecanismos definidos para catalogar e proteger os dados. Sem esses elementos, os dados não podem ser confiáveis e encontrados, resultando em uma bagunça, um “pântano de dados”. Atender às necessidades de públicos mais amplos exige que os data lakes tenham governança, consistência semântica e controles de acesso (Amazon, 2022).

6 História de Sucesso

Uma grande empresa que utiliza do sistema de Data Lake é o Google com a sua “suíte de computação” Google Cloud que oferece várias ferramentas para o usuário, algumas delas utilizando o Data Lake do Google para armazenamento e análise de dados como o Bigquery e Dataproc (análise), DataFlow (processamento) e o Cloud Storage (Armazenamento).

Outra empresa grande que cria um DL para o usuário é a AWS (Amazon Web Services, Inc.). Diretamente do website da empresa, fala que “A AWS oferece o portfólio de serviços mais seguro, escalável, abrangente e econômico para permitir que os clientes criem um data lake na nuvem, analisem todos os dados, incluindo dados de dispositivos de IoT, usando diversas abordagens analíticas, como machine learning. Como resultado, há mais organizações executando data lakes e análises na AWS do que em qualquer outro lugar, com clientes como **NETFLIX, Zillow, NASDAQ, Yelp, iRobot e FINRA** confiando na AWS para executar suas workloads de análise crítica de negócios”.

7 Referências Bibliográficas

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/big-data/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. **Data Lake**. **Wikipedia**, 2022. Disponível em: https://en.wikipedia.org/wiki/Data_lake. Acesso em 18 de jan., de 2023.

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 em: https://en.wikipedia.org/wiki/Data_mart. Acesso em 18 de jan., de 2023.

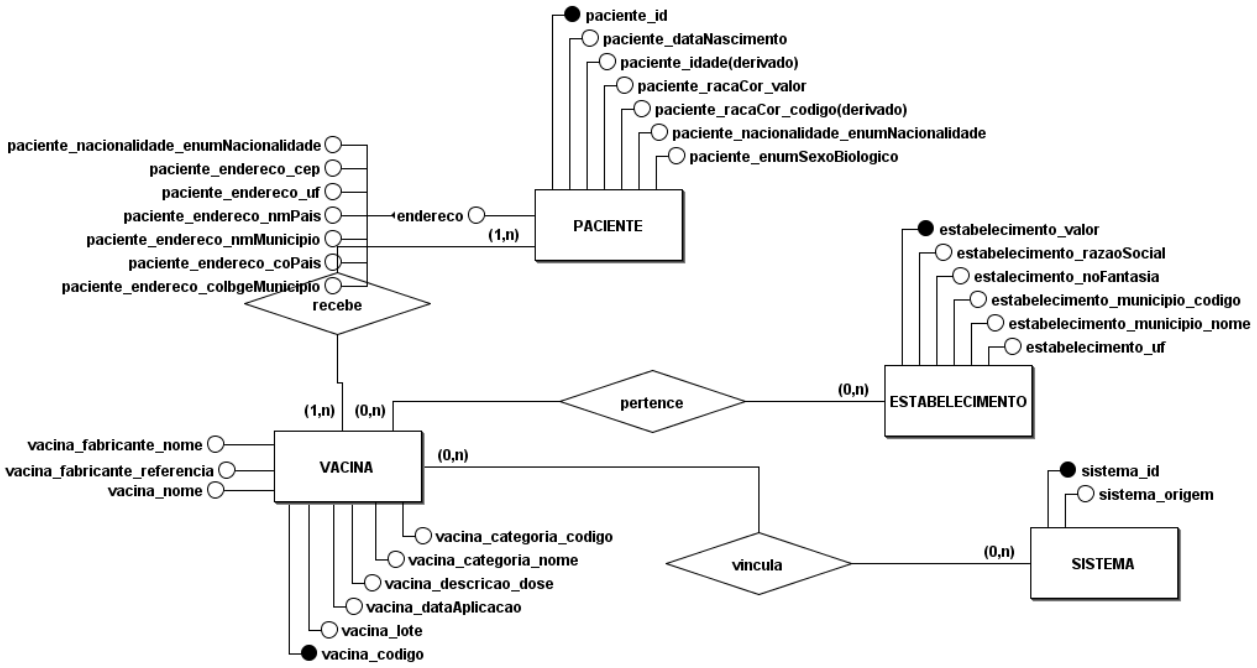
Base de Dados

A base de dados escolhida foi a de quantidade de vacinas aplicadas no Distrito Federal pega do site do governo dados.gov.br. Ela pode ser encontrada aqui: https://dados.gov.br/dados/conjuntos-dados/covid-19-vacinacao1.

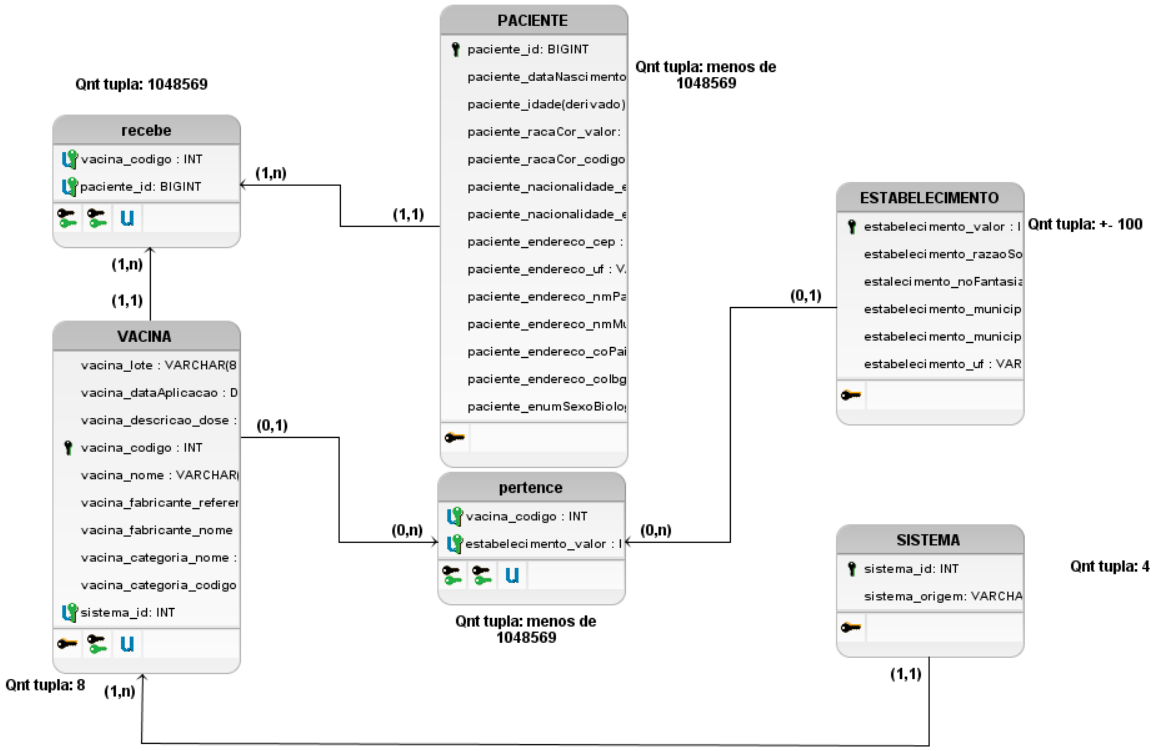
Nela possui Paciente, Vacina, Sistema e Estabelecimento como entidade, podendo possuir mais.

A maior quantidade de tupla encontrada foi de 1048576 d relacionamento com vacinas ou id_documento no csv.

DER



DLD



Dicionário de Dados

Ordem	Campo	Descrição	Categoria
1	document_id	Identificador do documento	
2	paciente id	Identificador do vacinado	
3	paciente_idade	Idade do vacinado	
4	paciente_data Nascimento	Data de nascimento do vacinado	
5	paciente_enumSexoBiologico	Sexo do vacinado	M = Masculino, F = Feminino

6	paciente_racaCor_codigo	Código da raça/cor do vacinado	1; 2; 3; 4; 99
7	paciente_racaCor_valor	Descrição da raça/cor do vacinado	1 = Branca; 2 = Preta; 3 = Parda; 4 = Amarela; 99 = Sem informação
8	paciente_endereco_coIbgeMunicipio	Código IBGE do município de endereço do vacinado	
9	paciente_endereco_coPais	Código do país de endereço do vacinado	
10	paciente_endereco_nmMunicipio	Nome do município de endereço do vacinado	
11	paciente_endereco_nmPais	Nome do país de endereço do vacinado	
12	paciente_endereco_uf	Sigla da UF de endereço do vacinado	
13	paciente_endereco_cep	5 dígitos para anonimizado e 7 dígitos para identificado	
14	paciente_nacionalidade_enumNacionalidade	Nacionalidade do vacinado	
15	estabelecimento_valor	Código do CNES do estabelecimento que realizou a vacinação	
16	estabelecimento_razaoSocial	Nome/Razão Social do estabelecimento	
17	estabelecimento_noFantasia	Nome fantasia do estabelecimento	
18	estabelecimento_municipio_codigo	Código do municipio do estabelecimento	
19	estabelecimento_municipio_nome	Nome do municipio do estabelecimento	
20	estabelecimento_uf	Sigla da UF do estabelecimento	

21	vacina_grupo_atendimento_code	Código do grupo de atendimento ao qual pertence o vacinado	
22	vacina_grupo_atendimento_nome	Nome do grupo de atendimento ao qual penence o vacinado	
23	vacina_categoria_code	Código da categoria	
24	vacina_categoria_nome	Descrição da categoria	
25	vacina_lote	Número do lote da vacina	
26	vacina_fabricante_nome	Nome do fabricante/fornecedor	
27	vacina_fabricante_referencia	CNPJ do fabricante/fornecedor	
28	vacina_dataAplicacao	Data de aplicação da vacina	
29	vacina_descricao_dose	Descrição da dose	
30	vacina_codigo	Código da vacina	
31	vacina_nome	Nome da vacina/produto	
32	sistema origem	Nome do sistema de origem	