



# **UNIVERSIDADE DE BRASÍLIA**

## **Faculdade do Gama**

### **Sistemas de Banco de Dados 2**

#### **Trabalho Final (TF)**

#### **Tema D – Data Warehouse**

**Rogério S. dos Santos Júnior**

**17/0021751**

**Brasília, DF**

**2020**

## A) Definição da Tecnologia

Um *data warehouse*, do inglês “armazém de dados”, é um sistema gerenciador de dados especial que tem o objetivo de servir de fonte de consulta para análises mais complexas. Esta tecnologia tem como objetivo consolidar um grande volume de dados de diversas fontes.

**Data warehouse** é uma coleção de dados orientada a assunto, integrada, não volátil, variável no tempo para apoio às decisões da gerência. (Inmon et al. (2008) apud Elmasri e Navathe (2018), p. 996)

Como principal distinção dos bancos de dados tradicionais, que são transacionais (orientado à objeto, relacionais etc.), os *data warehouses* servem, principalmente, para aplicações que visão sustentar decisões. Proporcionam dados para estudos complexos.

Normalmente, a operação de um *data warehouse* envolve os seguintes aspectos:

- Base SQL para armazenar os dados;
- Pré-processamento ETL sobre os dados;
- Ferramentas para visualização de dados;
- Ferramentas que se usam os dados para treinar algoritmos de predição (ML/AI);

Vale destacar que, pela alta demanda de consultas, são bases muito bem otimizadas e que suportam demandas de alto desempenho dos dados e informações de uma organização. Por não possuírem margem para inserções ou atualizações, esse tipo de base é ideal para dar consultas, processamento de dados ou *data mining*.

O *data warehouse* é constituído, como um SGBD, por diversas tecnologias que permitem a sua operacionalização para que seja o mais otimizado e automatizado possível. Por passa por um pré-processamento, os dados tem

redundância minimizada e fornecem uma fonte única para ambiente de BI ou até mesmo para o *data science* da empresa.

## **B) Objetivos Principais da Tecnologia**

Para servir para consultas em larga escala, os data warehouses não podem receber dados de qualquer maneira, é necessário um tratamento prévio para que os dados possam ser disponibilizados na base. Dessa forma, surge como solução o ETL.

O ETL, do inglês “extrair, transformação e carga”, é a atividade que consiste em passar o dado por um extenso pré-processamento para que ele seja o mais consistente possível. Um *data warehouse* possui diversas fontes e é necessário que eles sejam convertidos para um formato que possa ser analisado e armazenado.

De acordo com a Oracle Brasil, dentre as principais características do *data warehouse*, vale citar:

- Atualização automática regular da base;
- Armazena dados históricos de meses, ano e até décadas;
- Proporciona análises avançadas e consultas ad-hoc;
- Usa modelagem parcialmente normalizada para otimizar o desempenho;
- Acessa milhões de linhas ao mesmo tempo.

O principal objetivo do *data warehouse* é servir como fonte de consulta para atividades de *business intelligence* (BI) e análise avançada. Esta tecnologia está intrinsicamente ligada a área de *data science* e é usando amplamente usando em atividades como *machine learning*, *artificial intelligence* e *data mining*.

Por ser uma fonte segura, rápida, convergente, homogênea e atualizada de dados é a opção mais utilizada ao se pensar em tecnologia de armazenamento de grandes fluxos de informações. Possuem uma gama variada de fontes, tais como: dados de uma empresa, log de aplicativos ou aplicativos de transações.

É o mais indicado para análise de dados volumosos ao longo de grandes períodos. Seus recursos permitem às empresas ter insights a partir dos dados armazenados na base e assim tomar as melhores decisões para o seu negócio

### **C) Vantagens da Tecnologia Utilizada**

O fator mais vantajoso para se manter uma base desse porte é que ela permite que dados que, anteriormente só ocupavam espaço, tenham algum valor. Extrai valor de grandes quantidades de dados históricos e fornece um registro consistente e organizado dos dados de uma empresa

Inmon, grande cientista da área, descreve as quatro características que fazem os *data warehouses* sejam a melhor opção para àqueles que se inserem dentro do contexto:

- **Integração:** por serem usados dados de diversas fontes é necessário padronizar previamente os dados antes de serem inseridos na base;
- **Não volátil:** em um *data warehouse*, as únicas operações realizadas são de consulta e exclusão e isso faz com que os dados não sofram alteração. De acordo com as boas práticas, a única forma de modificação presente na base é a inserção, que consiste na adição de novos dados, e a exclusão;
- **Variável com o tempo:** a sua manutenção consiste apenas na inserção de novos dados de um novo período. Não existem problemas de concorrência, como nos bancos transacionais, pois os dados não sofrem atualização e já são processados previamente;
- **Orientado para o assunto:** faz alusão que os *data warehouses* são feitos dentro de um contexto bem determinado que permite que a sua construção seja otimizada e planejada para uma determinada finalidade.

Um projeto bem feito de *data warehouse* fornece a uma empresa a opção de proporcionar aos usuários mais flexibilidade na consulta aos dados, uma vez que pode proporcionar informações mais detalhadas ou não, e de possuir um maior volume de tráfego de dados.

## **D) Desvantagens da Tecnologia Utilizada**

Como toda implementação de nova funcionalidade é sempre importante analisar o contexto e a necessidade do escopo de implementação de um *data warehouse*, pois senão houver os profissionais, os recursos e a gestão necessária a operacionalização pode resultar em custos desnecessários e uma base que não consiga atingir os padrões necessários.

A montagem exige um extenso planejamento que pode levar até anos. Para que um *data warehouse* seja constituído é necessário um longo estudo dos dados da organização para ter propriedade e realizar a modelagem do banco de forma a prever como eles serão usados. O fato de as informações não serem alteradas faz com que toda a acomodação seja pensada com antecedência.

A escolha de tecnologias deve ser muito bem analisada para que a base tenha a melhor performance possível. O pré-processamento e o armazenamento dos dados são tarefas que vão lidar com volumes massivos de dado e por isso se faz necessário ferramentas que estejam de acordo com o escopo de atuação do *data warehouse*.

Com a evolução das tecnologias, tanto das fontes de dados quanto do *data warehouse*, o gerenciamento se torna um verdadeiro desafio. As bases de dados sofrem alterações e atualizações que devem ser refletidos no *data warehouse* para que a operação seja mantida. Dessa forma, é ele saiba lidar com esse tipo de demanda.

Outro ponto muito relevante é a qualidade e consistência dos dados. Por se tratar de um papel de decisão em muitas organizações, as informações devem possuir um elevado padrão de controle. A integração de dados é muito desafiadora, dado o domínio de informação de cada base, e o projetista do *data warehouse* deve ser capaz de lidar com as alterações das fontes de dados.

O sistema deve sempre refletir os melhores caminhos de modo que os padrões de utilização se modificam com o tempo para que o *data warehouse* esteja

sempre otimizado às demandas da organização. Isso é essencial para que o funcionamento pleno da base seja antigido.

A equipe que cuida de uma operação como essa deve possuir amplas habilidades. A multidisciplinaridade também é esperada, uma vez que não consiste somente em um banco, mas em conhecer todo o contexto da organização, regras, regulamentações, políticas e restrições, para realizar o planejamento de acordo.

Por fim, vale ressaltar os altos custos envolvidos no planejamento, execução e planejamento de uma operação como essa. Atualmente, os valores relacionados às tecnologias de armazenamento são bem onerosos, além da necessidade de manutenção. A capacitação dos profissionais necessários à atividade também é outro ponto bem relevante, pois geram mais gastos

## **E) Exemplo de Uso**

Para poder ilustrar as empresas que usam o data warehouse, estão elencados 3 casos de sucesso relacionadas ao data warehouse, o que ele proporcionou nas empresas, o que mudou e qual a nova realidade das organizações:

- **Vivo (Telefônica)**: para centralizar os dados de 6 empresas que a compõe seu conglomerado, a empresa possui um *data warehouse* que permite uma economia de US\$ 28 milhões todos os meses. Essa central reúne dados de uso de SMS, uso de rede, volume de voz trafegado, consumo financeiro do cliente etc., e usa esses dados para alimentar sua área de BI;
- **Ministério da Justiça do Brasil**: com um *data warehouse* que fornece dados para operações complexas, tais como a Lava Jato. Essa base central permite armazenar dados vitais ao funcionamento do órgão e possui bilhões de registros. O processamento conta poderoso computador da IBM que é capaz de coletar, agrupar e processar petabytes ( $10^6$  GB) em questão de milésimos de segundos. É capaz de cruzar informações de pessoas envolvidas nos mais diversos processos do judiciário.

- 
- **Avon:** para substituir o formar de tomar decisões, que era tido por experiência e percepção dos gestores, a empresa optou por implementar um *data warehouse* para poder tomar decisões baseadas em dados mais concretos. Dessa forma, se pode monitorar as estratégias traçadas e identificar oportunidades de negócio;

Existem diversas ferramentas que implementam um *data warehouse* padrão e oferecem um ambiente em nuvem para que as empresas possam armazenar seus dados e recebam insights de forma automatizada. Uma das principais vantagens delas é o fato de não ser necessário realizar a manutenção física dos servidores, que gera muitos custos. As principais são:

- **Google BigQuery:** serviço oferecido pela gigante da tecnologia, este serviço pode ser contratado on-line e, assim, conceder acesso a uma suíte de ferramentas e armazenamento de grandes massas de dados. Conta com vários recursos que permitem as organizações ter insights de seus negócios. É uma plataforma que proporciona uma interface visual para seus usuários facilitando o seu uso. Possui vasta documentação e suporte e inúmeros clientes famosos (Twitter, Dow Jones, SBT, Mercado Livre etc.);
- **Oracle Autonomous Data Warehouse:** serviço em nuvem que promete eliminar a complexidade da operacionalização de um *data warehouse* e automatizar a gerência da base. Possui diversas ferramentas de predição e autoajuste para melhorar o desempenho;
- **Amazon Redshift:** serviço de alta performance oferecido pela amazon que tem fácil integração com outros serviços oferecido por essa outra gigante da tecnologia. Possui clientes bem famosos (Mc Donald's, Duolingo, Fox, Amazon).

---

## F) Bibliografias

AMAZON. **Amazon RedShift**. Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

DEVMEDIA. **Data Warehouse**. Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Banco de Dados**. 7. ed. São Paulo: Pearson, 2018. 1127 p. ISBN 978-85-430-2500-1.

GOOGLE. **Google BigQuery**. Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

GUERRA, Ana Rita. **Telefônica Vivo revoluciona sistema de oferta ao cliente com processamento de dados veloz**. 22 out. 2015. Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

MAGALHÃES, Paulo. **10 cases de sucesso de empresas que utilizaram o Big Data**. 26 abr. 2017. Disponível em: [link para o site](#).

ORACLE. **Oracle Autonomous Data Warehouse**. Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

ORACLE BRASIL. **O que É um Data Warehouse?** Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

SAS. **ETL: o que é e qual a sua importância?** Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

THINK CONSULTING. **3 empresas que utilizam business intelligence e obtiveram sucesso**. Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.

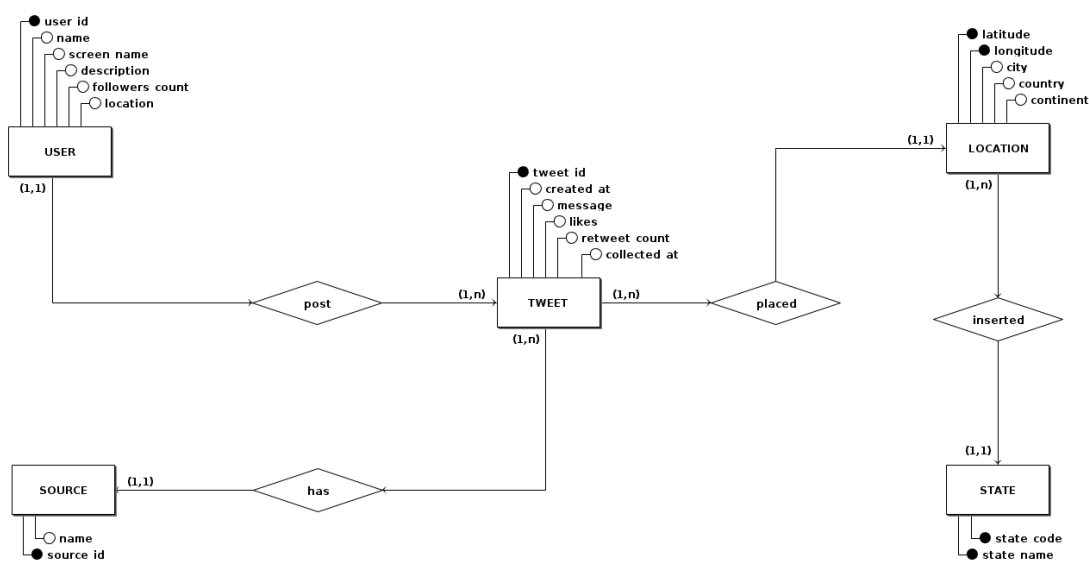
THINK CONSULTING. **Afinal, você sabe o que é data warehouse?** Disponível em: [link para o site](#). Acesso em: 16 nov. 2020.



## G) Documentação Base de Dados

Os dados foram extraídos do Kaggle dos [tweets das eleições dos Estados Unidos da América de 2020](#). A partir do arquivo CSV (hashtag\_joebidden.csv) foi feita a modelagem do banco MySQL de forma a normalizar os dados em suas respectivas tabelas. Em seguida, com o auxílio de um script Python os dados foram inseridos na base, totalizando 155 mil tuplas. Foi gerado o [dump](#) da base modelada.

### 1) Diagrama Entidade Relacionamento (DE-R)



### 2) Diagrama Lógico

