UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Data Warehouse e Data Marts (DW)

Ailton Aires Amado - 180011600

Brasília, DF

2023

Definição

Segundo Vida (2021), **“**todos os sistemas de informação atuam com o mesmo propósito: fornecer informações com precisão, qualidade, abrangência e em tempo hábil”. E para que isso continue ocorrendo, passaram a ser criados sistemas para apoiar as empresas, e que fornecem tipos especializados de informações, como os Data Warehouses (DW) ou Armazém de Dados.

Um Data Warehouse é um repositório centralizado de informações integradas, variáveis, acessíveis, compartilháveis e não voláteis, que visa apoiar as decisões importantes de uma organização (VIDA et. al., 2021; FERREIRA et. al., 2010; MATTIODA e FAVARETTO, 2010). Segundo a Amazon ([20--?]), esses dados são extraídos de sistemas transacionais, bancos de dados relacionais e de outras fontes, para que sejam carregados no Data Warehouse de forma periódica e consistente. Vale lembrar também que no DW os dados armazenados já estão tratados, e não há redundância de quaisquer tipos de informações. Esse processo é conhecido como ETL (Extração, Transformação e Carga) e será discutido mais à frente.

De forma bem direta e informal para que se torne fácil a compreensão da definição de um Data Warehouse, essa estrutura pode ser vista como um grande banco de dados com muitas informações históricas e que nunca são apagadas. Essa é uma forma de permitir que essas informações sejam acessadas de forma fácil para a aquisição de conhecimento.

Enquanto um Data Warehouse é um repositório centralizado de dados integrados e organizados, um Data Mart é um subconjunto desse mesmo repositório que atende às necessidades de uma equipe ou unidade de negócios específica (AMAZON, [20--?]), desempenhando “o papel de um DW (departamental, regional ou funcional), podendo-se construir uma série deles ao longo do tempo e eventualmente vinculá-los através de um DW lógico empresa-inteira” (SINGH, 2001). Construir vários Data Marts ao longo do tempo é uma abordagem comum e que pode expandir e refinar as capacidades de análise em uma organização.

Dentro dos objetivos de um Data Warehouse, temos a acessibilidade das informações, consistência, qualidade e muito mais, que serão abordados

posteriormente.

Na tabela 1 é apresentado uma comparação entre Data Warehouse e Data Mart, levando em consideração suas características.

Tabela 1 - Características dos Data Warehouse e Data Mart

Características	Data warehouse	Data mart
Escopo	Várias áreas centralizadas e integradas	Uma área específica e descentralizada
Usuários	De toda a organização	Uma única comunidade ou departamento
Fonte de dados	Muitas fontes	Uma ou poucas fontes, ou uma parte dos dados já coletados em um data warehouse
Tamanho	Grande, pode variar de centenas de gigabytes a petabytes	Pequeno, normalmente até algumas dezenas de gigabytes
Projeto	De cima para baixo	De baixo para cima
Detalhes dos dados	Dados completos e detalhados	Pode manter dados resumidos

Fonte: AMAZON ([20??])

Na tabela acima temos as principais características entre Data Warehouse e Data Mart, levando em consideração que os Data Mart são apenas estruturas menores de um Data Warehouse.

Objetivo(s) principal(is) da Tecnologia Pesquisada

Segundo Kimball e Ross (2002), podemos elencar alguns dos principais objetos com relação aos Data Warehouse. São eles:

- O Data Warehouse deve tornar as informações de uma organização facilmente acessíveis: O conteúdo do Data Warehouse deve ser facilmente compreendido. Os dados devem ser intuitivos e óbvios para todos os usuários, e não somente para o desenvolvedor. Vale lembrar que informações de fácil compreensão resultam em uma facilidade na legibilidade delas. É importante que o conteúdo do Data Warehouse seja rotulado de maneira significativa. Os *stakeholders* desejam separar e combinar os dados no warehouse de diversas formas. Para isso, um processo comumente conhecido e utilizado é o de dividir e conquistar. Essa abordagem envolve a criação de estruturas de dados

flexíveis que permitem a combinação de diferentes elementos para atender às necessidades específicas de análise. As ferramentas que acessam o data warehouse devem ser simples e fáceis de usar. Eles também devem retornar os resultados da consulta ao usuário de forma rápida com o menor tempo de espera possível.

- O Data Warehouse deve apresentar as informações da organização de forma consistente: Os dados no Data Warehouse devem ser confiáveis. Eles devem ser cuidadosamente organizados a partir de uma variedade de fontes de toda a organização. Devem ser tratados, concisos e liberados somente quando estiverem adequados para que o usuário possa utilizá-los. As informações de um processo de negócios devem corresponder às informações de outro, para garantir a consistência e a integridade dos dados. Isso é fundamental para que os usuários confiem nas informações e realizem análises confiáveis e precisas dentro da organização. Isso ocorre por meio do mapeamento, padronização e integração dos dados, juntamente com os metadados. A exemplo, temos que, se duas medidas de desempenho tiverem o mesmo nome, elas devem significar a mesma coisa. Porém, se duas medidas não significam a mesma coisa, elas devem ser rotuladas de forma diferente. Consistência significa qualidade dentro das informações trabalhadas, ou seja, é uma garantia de que todos os dados necessários foram contabilizados e estão completos, livre de duplicações ou discrepâncias.

- O Data Warehouse deve ser adaptável e resiliente a mudanças: mudança é um processo inevitável. Necessidades de usuário, condições de negócios, dados e a tecnologia estão sujeitos às mudanças ao longo do tempo, por diversos motivos. O Data Warehouse deve ser projetado para lidar com esses tipos de mudanças. Porém, as alterações devem ser simples, o que significa que não invalidam dados ou aplicativos existentes. Os dados e aplicativos existentes não devem ser alterados ou interrompidos quando novos dados forem adicionados ao Data Warehouse.

- O Data Warehouse deve servir como base para uma melhor tomada de decisão: O Data Warehouse deve garantir a presença de dados precisos e confiáveis para auxiliar no processo de tomada de decisões da organização. O resultado final de um Data Warehouse é representado pelas decisões que são tomadas após a apresentação das evidências fornecidas. Essas decisões possuem um grande impacto comercial e agregam valor ao Data Warehouse. O termo que melhor se adequa à descrição do que é um Data Warehouse é o de um sistema que oferece suporte à tomada de decisões com base nas informações armazenadas.

- A organização deve aceitar o Data Warehouse para que seja considerado bem-sucedido: Não importa o quanto construímos uma solução elegante usando os melhores serviços ou ferramentas. Se a organização não adotou o Data Warehouse então o teste de aceitação falhou. Ao contrário de uma reescrita do sistema operacional, onde os usuários de negócios não têm escolha a não ser usar o novo sistema, o uso do data warehouse às vezes é opcional. A aceitação da organização tem mais a ver com simplicidade do que qualquer outra coisa.

Vantagens da Tecnologia Pesquisada

Segundo a Amazon ([20--?]), podemos elencar 5 vantagens principais de um Data Warehouse. São elas:

- Tomada de decisão adequada: Um Data Warehouse fornece uma visão consolidada dos dados da organização, permitindo uma melhor tomada de decisão. Os dados são organizados e estruturados de forma com que facilite a análise e a extração de insights que são relevantes no apoio da tomada de decisões estratégicas e operacionais.

- Dados consolidados de várias fontes: O Data Warehouse é projetado para suportar os dados de várias fontes, como sistemas transacionais, aplicativos de negócios e outros sistemas de dados, sejam eles internos ou externos. Isso

permite a consolidação e centralização dos dados de diferentes fontes em um único local, o que elimina a necessidade de consultar várias fontes de forma separada. Essa integração de dados é um ponto forte que facilita a análise abrangente e a obtenção de uma visão holística dos negócios.

- Análise de dados históricos: O Data Warehouse armazena e mantém o histórico dos dados ao longo do tempo. Isso permite que sejam analisados os dados históricos e a identificação de tendências, padrões e comportamentos ao longo do tempo pelos interessados. A capacidade de analisar dados históricos é crucial para compreender o desempenho passado, avaliar a eficácia das estratégias e tomar decisões embasadas em insights baseados em dados.

- Qualidade, consistência e precisão de dados: O Data Warehouse foi desenvolvido para garantir a qualidade, consistência e precisão dos dados que são armazenados. Isso envolve a aplicação de processos de limpeza, transformação e padronização dos dados para remover duplicações, erros e inconsistências. A garantia de qualidade dos dados é essencial para fornecer informações confiáveis e precisas para análise e tomada de decisão, conforme apresentado anteriormente dentro dos objetivos.

- Separação do processamento analítico dos bancos de dados transacionais, o que melhora o desempenho dos dois sistemas: O Data Warehouse separa o processamento analítico dos bancos de dados transacionais, utilizados para operações do dia a dia da organização. Essa separação melhora o desempenho dos dois sistemas, pois cada um pode ser otimizado para suas respectivas finalidades. O Data Warehouse é projetado para consultas analíticas complexas e de alto desempenho, permitindo uma análise eficiente e ágil dos dados, sem impactar negativamente as transações do banco de dados transacional.

Desvantagens da Tecnologia Pesquisada

Estruturar um Data Warehouse não é uma tarefa fácil e deve possuir

atenção em alguns detalhes antes que isso ocorra. Algumas das desvantagens dessa tecnologia, segundo Five Acts (2021) são:

- Dificuldade em integrar com sistemas e softwares legados: A integração do Data Warehouse com sistemas e softwares legados pode ser um desafio devido às enormes diferenças entre possíveis estruturas de dados distintas, formatos de dados e até mesmo as tecnologias utilizadas. A falta de uma padronização acaba dificultando todo o processo, havendo a necessidade de um esforço maior para tratar esses dados.

- Problemas no controle de acesso aos dados: Garantir a segurança e o controle de acesso adequados aos dados do Data Warehouse pode ser uma tarefa complicada, já que fazer uma gestão dos usuários quanto ao acesso aos dados não é uma tarefa fácil, principalmente se houver dados sensíveis e que demande uma confidencialidade maior.

- Complicações ao estruturar dados e para agregar valor a eles: Conforme já foi abordado, a filtragem e preparação dos dados são extremamente importantes para uma maior confiabilidade. No entanto, a definição e o projeto desses esquemas podem ser complexos, requerendo conhecimento especializado e análise cuidadosa das regras de negócio envolvidas. Além disso, agregar valor aos dados para que se tornem informações úteis e significativas pode exigir a criação de medidas, métricas e cálculos personalizados, o que também pode ser um desafio.

- Estruturação trabalhosa: A construção e manutenção de um Data Warehouse podem exigir esforço e tempo significativos. Isso inclui a extração, transformação e carga (ETL) de dados de várias fontes. A complexidade dessa atividade é mais uma das dificuldades encontradas dentro do uso dos Data Warehouse.

- Rápida obsolescência: Conforme já foi falado, mudanças são inevitáveis em

qualquer negócio. Dentro dos Data Warehouse, novos requisitos de dados, mudanças nas fontes de dados ou avanços tecnológicos podem exigir a atualização ou reestruturação constante para mantê-lo com as informações ainda adequadas e úteis. Por isso a importância da flexibilidade e adaptabilidade do Data Warehouse com relação às possíveis mudanças durante seu ciclo de vida.

- Imprevisibilidade em relação aos problemas: O Data Warehouse pode enfrentar problemas imprevisíveis, como falhas de hardware, erros de integração de dados, inconsistências nos dados, entre outros. Esses problemas podem afetar a disponibilidade, a integridade e a qualidade dos dados armazenados no Data Warehouse. A detecção e a resolução desses problemas podem exigir monitoramento constante, manutenção regular e ação rápida para evitar impactos negativos nas operações de negócios.

Casos de sucesso:

Um dos grandes produtos oferecidos de Data Warehouse é o Amazon Redshift, serviço hospedado em nuvem que utiliza SQL para analisar dados estruturados e semiestruturados, utilizando hardware e machine learning para oferecer uma melhor performance (AMAZON, [20??]). Abaixo algumas empresas com casos de sucessos na utilização de Data Warehouse.

- Nokia: Conforme relatado por Olavsrud (2014), em 2012, os volumes de dados da Nokia literalmente quebraram o banco de dados da empresa, baseado em uma fala de um dos chefes da Nokia. A plataforma da Nokia afetada foi a Xpress Internet Services, que fornece serviço de internet móvel em alguns países. Antes da migração para o Redshift da Amazon, com o banco de dados atual não era possível fazer mais tarefas simples de forma econômica ou qualquer coisa útil a nível de consultas. Após a migração, em apenas 2 meses, a maioria das informações foram migradas e as consultas ocorriam duas vezes mais rápido, sem contar a possibilidade de analisar os dados de forma mais segura, com uma economia de custo de até 50%. Isso tudo só é

possível graças a consultas analíticas complexas e de alto desempenho, permitindo uma análise eficiente e ágil dos dados.

- Nasdaq: “A Nasdaq é uma empresa multinacional de tecnologia e serviços financeiros que detém e opera a Bolsa de Valores Nasdaq” (AMAZON, 2020). É ela que opera a maior quantidade de dados na Bolsa de Valores no mundo, gerenciando a conciliação de compradores e vendedores em alto volume e velocidade. São bilhões de registros que precisam ser armazenados.

Com o aumento das transações realizadas à medida que as plataformas de negociação automatizadas entraram no mercado, fez-se necessário a migração para um Data Warehouse, onde, segundo o Vice-presidente de Engenharia de software da Nasdaq, Robert Hunt, pôde-se suportar um salto de 30 para 70 bilhões de registros por dia devido a melhor flexibilidade e escalabilidade oferecida pelo Data Warehouse da Amazon. Também foi reduzido em 5 horas o tempo de carregamento dos dados e acelerou em 32% a execução de consultas. como a Nasdaq busca que a equipe de pesquisa econômica realize a análise de dados e execute consultas complexas nas informações presentes nos Data Warehouse, conforme apresentado no caso de sucesso anterior, isso só é permitido graças às consultas complexas e eficientes que os Data Warehouse são capazes de realizar.

Casos de insucesso

- High-Tech Company: Um caso de insucesso que pode ser apresentado é o de uma empresa de tecnologia americana, chamada High-Tech Company. Nesse caso, a equipe de marketing e finanças precisava de algumas informações sobre análises de tendências para tomar decisões estratégicas (um dos principais objetivos dos Data Warehouse). Para tal, foi então sugerido pela equipe de TI uma solução envolvendo Data Warehouse. A abordagem era construir data marts compatíveis e independentes.

Segundo Fenton (1999), “nesta empresa, três projetos diferentes de armazéns falharam. A primeira foi devido a software inadequado; a segunda à

falta de comprometimento formal da gestão e rotatividade organizacional; e o terceiro a fontes de dados inadequadas e, novamente, rotatividade organizacional". Levando em consideração o primeiro e o terceiro motivo para que a empresa viesse a falhar na implementação do Data Warehouse, podemos concluir que esses motivos podem estar ligados, respectivamente, as seguintes desvantagens supracitadas: dificuldade em integrar com sistemas e softwares legados e complicações ao estruturar dados e para agregar valor a eles.

Descrição: A base de dados reúne centenas de milhares de avaliações de alimentos da Amazon, mais precisamente 568.454 avaliações. As avaliações estão dentro do período de outubro de 1999 até outubro de 2012. Também existem 256.454 usuários e 74.258 produtos. As análises incluem informações sobre produtos e usuários, classificações e uma análise em texto simples. Essa base possui um arquivo csv e um sqlite, que são arquivos de fácil manipulação.

Diagrama Entidade-Relacionamento (DE-R):

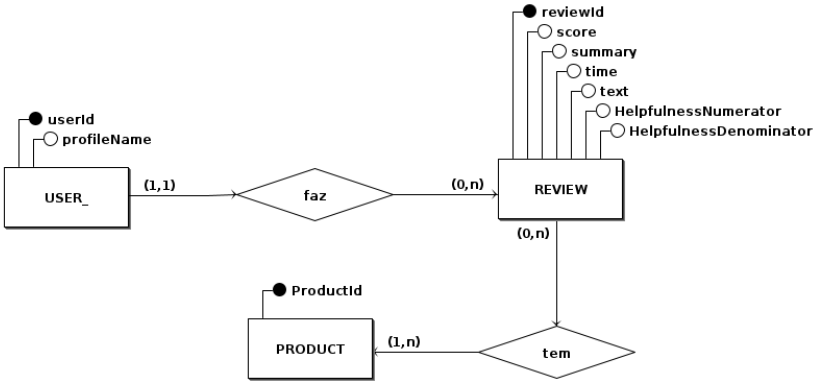
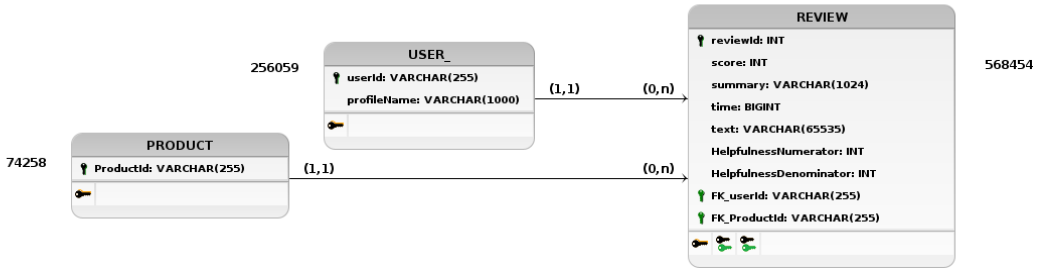


Diagrama Lógico de Dados (DLD):



Endereço virtual:

https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews?resource=download&select=database.sqlite

Referências Bibliográficas

AMAZON. **Conceitos de Data Warehouse**. [*S. l.*], [s.d]. Disponível em: link para o site. Acesso em: 9 jun. 2023.

AMAZON. **Nasdaq usa a AWS para tornar-se pioneira no armazenamento de dados de bolsas de valores na nuvem.** [*S. l.*], 2014. Disponível em: link para o site. Acesso em: 10 jun. 2023.

FENTON, M. D., HAYWOOD, M. E., GERARD, J. G., GONZALEZ, L. E., & WATSON, H. J. (1999). **Data Warehousing Failures: Case Studies and Findings.** Disponível em: link para o site. Acesso em: 10 jun. 2023.

FERREIRA, J. J. et al. **O processo ETL em sistemas data warehouse**. 1 jun. 2010. Disponível em: link para o site. Acesso em: 9 jun. 2023.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling** ; Second Edition. [s.l.] Wiley, 2002. Disponível em: link para o site. Acesso em: 9 jun. 2023.

MATTIODA, R. A.; FAVARETTO, F. Qualidade da informação em duas empresas que utilizam Data Warehouse na perspectiva do consumidor de informação: um estudo de caso. **Gestão & Produção**, v. 16, n. 4, p. 645–666, dez. 2009. Disponível em: link para o site. Acesso em: 10 jun. 2023

OLAVSRUD, THOR. **7 Amazon Redshift Success Stories.** [*S. l.*], 17 dez. 2014. Disponível em: link para o site. Acesso em: 9 jun. 2023.

SINGH, H. S. **Data warehouse conceitos, tecnologias, implementação e gerenciamento.** São Paulo: Makron Books, 2001. p. 14.

VIDA, Edinilson da S.; ALVES, Nicolli S R.; FERREIRA, Rafael G C.; et al. **Data warehouse**. Grupo A, 2021. *E-book.* ISBN 9786556901916. Disponível em: link para o site. Acesso em: 09 jun. 2023.