

UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Data warehouse e Data marts

Kevin Luis Apolinario Batista - 18/0042386

Brasília, DF

2023

A) Definição da tecnologia Data Warehouse e Data Marts

Data Warehouse é um sistema de armazenamento de dados projetado para facilitar a análise e o relatório de informações em uma organização. É uma estrutura centralizada que integra dados de várias fontes, como bancos de dados operacionais, arquivos, feeds de dados externos, etc. O objetivo de um Data Warehouse é fornecer uma visão ampla e consolidada dos dados, de modo que os usuários possam tomar decisões estratégicas com base em informações confiáveis e consistentes (Elmasri e Navathe, p. 720). Ele é otimizado para consultas complexas e análises de longo prazo, geralmente armazenando grandes quantidades de dados históricos. Diferentes tipos de aplicações, como OLAP (Online Analytical Processing), DSS (Decision Support Systems) e aplicações de mineração de dados, são amplamente utilizados nesse contexto. (Elmasri e Navathe, p. 721)

Data Marts, de acordo com Wylton (2007), são subconjuntos de um Data Warehouse. Eles são projetados para atender a necessidades específicas de um departamento ou equipe dentro da organização. Um Data Mart é uma versão segmentada e simplificada do Data Warehouse, focada em fornecer informações relevantes para usuários específicos (Oracle). Geralmente, cada Data Mart é construído em torno de uma área de negócio ou função específica, como vendas, marketing, finanças e recursos humanos. Eles contêm um subconjunto de dados do Data Warehouse e são dimensionados para oferecer um desempenho otimizado para as consultas relacionadas à área específica.

A principal diferença entre um Data Warehouse e Data Marts está na escala e no escopo dos dados armazenados e no público-alvo de usuários. Um Data Warehouse é uma solução abrangente que engloba toda a organização, integrando uma ampla variedade de fontes de dados e fornecendo uma visão holística dos negócios. Por outro lado, um Data Mart é mais focado e limitado em termos de dados e usuários. Ele serve a uma função específica ou departamento, oferecendo informações mais detalhadas e personalizadas para atender às necessidades específicas dos usuários dentro desse contexto.

B) Principais objetivos da tecnologia

O principal objetivo de um Data Warehouse é fornecer uma fonte única de verdade para os dados de uma organização. Ele visa integrar e consolidar dados de várias fontes, eliminando inconsistências e duplicações, para criar um repositório confiável e consistente. Ao centralizar os dados em um Data Warehouse, as empresas podem obter uma visão completa e abrangente de suas operações, permitindo uma análise mais precisa e uma tomada de decisão mais informada (Oracle). Além disso, o Data Warehouse é projetado para facilitar consultas complexas e análises de longo prazo, permitindo aos usuários explorar tendências, padrões e insights valiosos.

Os Data Marts têm como objetivo principal fornecer informações direcionadas e especializadas para usuários específicos ou departamentos dentro da organização (Elmasri e Navathe, p. 722). Eles são projetados para atender às necessidades analíticas específicas de cada área de negócio, fornecendo um subconjunto de dados relevantes e adaptados às suas operações. Ao criar Data Marts, as empresas podem atender às demandas individuais de cada departamento, oferecendo informações personalizadas que são relevantes para suas necessidades de análise e tomada de decisão. Isso permite uma visão mais detalhada e direcionada das operações em nível departamental, promovendo uma melhor compreensão dos negócios e aprimorando a eficiência das atividades específicas de cada área (Oracle).

Data Warehouse e Data Marts desempenham papéis complementares no cenário da análise de dados em uma organização. Um Data Warehouse é construído para ser a principal fonte de dados, consolidando informações de várias fontes em um único local. Ele realiza processos de extração, transformação e carga (ETL) para integrar os dados e garantir a consistência e a qualidade das informações armazenadas. Além disso, o Data Warehouse geralmente armazena grandes volumes de dados históricos, permitindo análises de tendências, padrões sazonais e comportamentos ao longo do tempo. Esse repositório centralizado fornece uma base sólida para a tomada de decisões estratégicas em toda a organização.

Por outro lado, os Data Marts são projetados para serem mais ágeis, focados em um conjunto particular de dados relevantes para a área em questão (Oracle). Os Data Marts podem ser construídos a partir do Data Warehouse, selecionando e resumindo os dados necessários para análises específicas, a fim de fornecer informações específicas e detalhadas para os usuários em suas operações diárias.

Ao utilizar tanto um Data Warehouse quanto Data Marts, as organizações podem obter uma visão completa e detalhada dos dados em todos os níveis da empresa. O Data Warehouse oferece uma visão global, permitindo a análise e a compreensão de tendências e padrões corporativos em larga escala, enquanto os Data Marts fornecem insights granulares e especializados para áreas específicas. Essa combinação de recursos permite que os usuários acessem informações relevantes e precisas, adaptadas às suas necessidades individuais, impulsionando a eficiência operacional, a tomada de decisões baseada em dados e o desempenho organizacional como um todo.

C) Vantagens da Tecnologia

Data Warehouse possui várias características que o tornam uma peça fundamental para a análise de dados em uma organização. Algumas dessas características incluem:

1. Qualidade dos dados

Com a implementação adequada de processos de controle de qualidade, é possível garantir a precisão, integridade e consistência dos dados armazenados. Isso resulta em informações confiáveis e de alta qualidade para análise e tomada de decisões. Ao eliminar inconsistências e duplicidades nos dados, um data warehouse oferece aos usuários uma fonte confiável e precisa de informações, permitindo uma análise mais precisa e uma compreensão aprofundada dos negócios.

2. Acesso rápido aos dados

Data warehouses são projetados para otimizar o desempenho de consultas e análises complexas. Com estratégias de indexação, particionamento e otimização de consultas, os data warehouses oferecem tempos de resposta rápidos, mesmo ao lidar com grandes volumes de dados. Isso permite que os usuários realizem análises exploratórias, identifiquem tendências emergentes e obtenham insights valiosos de maneira eficiente. O acesso rápido aos dados é crucial para uma análise ágil e eficaz, apoiando a tomada de decisões informadas e oportunas.

3. Análise de dados históricos

Ao armazenar e manter o histórico dos dados ao longo do tempo, o Data Warehouse possibilita a análise retrospectiva e a identificação de tendências, padrões e comportamentos relevantes. Através da análise dos dados históricos, os interessados podem compreender o desempenho passado, avaliar a eficácia das estratégias adotadas e embasar suas decisões em insights baseados em dados concretos. Essa perspectiva temporal dos dados proporciona uma visão mais abrangente e permite uma análise mais aprofundada, proporcionando uma valiosa compreensão do contexto histórico para apoiar a tomada de decisões futuras.

Além dessas vantagens citadas, a facilidade de uso é outra vantagem significativa dos data warehouses. Eles são projetados para fornecer uma interface amigável e intuitiva, facilitando o acesso e a análise dos dados por usuários de diferentes níveis de habilidade. As ferramentas de visualização e consulta permitem que os usuários explorem os dados de maneira interativa, realizem análises ad-hoc e gerem relatórios personalizados. A facilidade de uso dos data warehouses aumenta a adoção e a utilização efetiva das informações pelos usuários, promovendo uma cultura orientada por dados dentro da organização.

D) Desvantagens do uso da tecnologia

Embora o data warehouse seja uma ferramenta valiosa para análise de dados, é importante reconhecer que sua implementação também pode trazer algumas desvantagens em termos de questões operacionais. Uma das desvantagens está relacionada à construção do data warehouse. Esse processo é complexo e requer um planejamento cuidadoso, envolvendo a identificação das fontes de dados, o projeto do esquema de dados e a extração, transformação e carga dos dados. A construção do data warehouse pode ser demorada e exigir recursos significativos, tanto em termos de tempo quanto de investimento financeiro.

Outra desvantagem está na administração do data warehouse. A manutenção do sistema, o gerenciamento de usuários e permissões de acesso, o monitoramento de desempenho e a garantia de disponibilidade contínua demandam esforço e habilidades técnicas especializadas. Essa administração pode exigir recursos adicionais da equipe de TI e pode ser um desafio em termos de complexidade e necessidade de atualizações constantes para atender às demandas em evolução da organização.

O controle de qualidade dos dados é outra questão operacional desafiadora para o data warehouse. Garantir a precisão, integridade e consistência dos dados armazenados requer a implementação de processos de limpeza, validação e padronização dos dados. Além disso, é necessário realizar auditorias regulares para identificar e corrigir possíveis problemas. Essa atividade de controle de qualidade exige recursos adicionais e pode exigir um esforço contínuo para manter a confiabilidade dos dados no data warehouse.

Por fim, vale ser ressaltado o alto custo envolvido na formação e manutenção de uma equipe especializada. Para implementar e gerenciar efetivamente um data warehouse, é necessário contar com profissionais capacitados em áreas como arquitetura de dados, administração de bancos de dados, desenvolvimento de ETL e análise de dados. Esses especialistas

demandam salários competitivos, além de investimentos contínuos em treinamento e capacitação para acompanhar as mudanças tecnológicas e as melhores práticas da área. Além disso, é necessário considerar os custos associados à contratação de consultores externos, caso a equipe interna não possua todos os conhecimentos necessários. Portanto, é essencial ter em mente que os custos relacionados à formação e manutenção de uma equipe especializada podem ser significativos ao se implementar e manter um data warehouse.

E) Exemplos de uso interessantes em empresas, projetos ou instituições

Para citarmos exemplos de uso da tecnologia de data warehouse podemos citar algumas ferramentas que são comumente utilizadas no mercado mundial, entre elas estão:

- Google BigQuery: É um data warehouse corporativo sem servidor e econômico que funciona em nuvens e pode ser escalonado de acordo com seus dados. Use ML/IA e BI integrados para insights em grande escala.
- Amazon Redshift: Solução de data warehousing baseada em nuvem. É altamente escalável e permite o processamento de grandes volumes de dados. Ele também é integrado com outras ferramentas e serviços da Amazon Web Services (AWS).

Caso de sucesso: **Mediexplain* utilizando cervinodata com google bigquery**

Cervinodata coleta todas as publicidades digitais e permite criar conexões entre plataformas, usando grupos de campanhas. Antes do Cervinodata, essas conexões eram feitas em planilhas do Google, usando aplicativos de terceiros para obter os dados. Essas conexões eram pouco

confiáveis e precisavam de verificações regulares e correções manuais. Agora essa tarefa é automatizada, eliminando a necessidade de trabalho manual e reduzindo erros. Os dados de todas as plataformas de anúncios são coletados e integrados de forma eficiente em um único local, o que economiza tempo e aumenta a confiabilidade dos dados. Além disso, a criação de conexões entre campanhas de várias plataformas agora é simplificada e mais fácil de gerenciar. (DORP)

Segundo Dorp (20??), depois de configurar o Cervinodata para o BigQuery, agora o Mediexplain* também traz dados de mídia off-line (TV, rádio) para o BigQuery. Isso permite que a Mediexplain* ofereça a cada cliente um painel atualizado com informações sobre campanhas online e offline. Devido a essa automação, o tempo é economizado estruturalmente, cerca de 4 horas toda semana, permitindo mais tempo para análises aprofundadas e contato com o cliente.

Caso de sucesso: **Financial Times utilizando amazon redshift**

O diário internacional Financial Times precisava de uma maneira de aumentar a velocidade, o desempenho e a flexibilidade de sua análise de dados. Após a análise do mercado, a plataforma passou a usar o AWS Redshift, permitindo assim serem executadas 450.000 consultas on-line 98% mais rápido do que seu datacenter tradicional anterior, ao mesmo tempo em que reduz os custos de infraestrutura em 80%.

E ainda de acordo com o CTO do diário, é possível analisar tendências e correlacionar essas informações com outros dados relevantes, permitindo uma sobreposição de informações de maneira poderosa. Essa análise possibilita a obtenção de insights valiosos sobre o comportamento dos usuários e a identificação de padrões de interesse. Essas informações podem ser usadas para tomar decisões estratégicas e otimizar a experiência do usuário no site, melhorando assim o desempenho e os resultados.

Caso de insucesso: **North American Tax Agency**

A alta direção do Departamento de Compliance de uma agência tributária norte-americana reconheceu os potenciais benefícios de um data warehouse. Eles buscavam aumentar o cumprimento voluntário e o grupo de TI dessa agência estava interessado em auxiliar a unidade na construção do data warehouse. Uma proposta formal foi elaborada, identificando as aplicações específicas do data warehouse, os benefícios a serem obtidos por meio dessas aplicações e o custo geral envolvido. O projeto seguiu as etapas adequadas para documentar os benefícios do projeto, especialmente os entregáveis de curto prazo. Estimava-se que essa empreitada ambiciosa levaria de três a cinco anos para ser concluída, a um custo de 25 a 30 milhões de dólares.

Após a aprovação da proposta, o interesse da unidade de negócios começou a diminuir, apesar do apoio inicial do patrocinador executivo. Uma das razões apresentadas foi que o patrocinador executivo não compreendia completamente o tempo necessário para concluir o projeto, o montante de financiamento necessário ou os benefícios intermediários que seriam alcançados (ou seja, a cada estágio de desenvolvimento, valor estava sendo adicionado ao Departamento de Compliance). Outra razão para a perda de interesse pode ter sido que o Departamento de Compliance foi solicitado a liderar do ponto de vista corporativo (ou seja, tomar decisões sobre arquitetura de dados e questões relacionadas aos dados). No entanto, o departamento relutou em assumir esse papel e expressou preocupação de que talvez fosse uma responsabilidade do grupo de TI. Como resultado, eles hesitaram em fornecer recursos adicionais à equipe do projeto.

Enquanto as razões para a perda de patrocínio não são claras, é provável que o custo total e a duração do projeto tenham contribuído para o enfraquecimento do interesse, desvantagens citadas neste texto anteriormente.

F) Referências Bibliográficas

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Banco de Dados**. 6. ed. São Paulo: Pearson, 2011.

WYLTON. **Data Warehouse ou Data Mart, por onde começar?** Disponível em: [link para o site](#). Acesso em: 10 jun. 2023.

ORACLE. **Oracle Autonomous Database**. Disponível em: [link para o site](#). Acesso em: 10 jun. 2023.

ORACLE BRASIL. **O que é um Data Warehouse?** Disponível em: <https://www.oracle.com/br/database/what-is-a-data-warehouse/>. Acesso em: 10 jun. 2023.

CETAX. **Data Warehouse: Tudo o que você precisa saber!** Disponível em: [link para o site](#). Acesso em 10 de junho de 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.

OLAVSRUD, THOR. **7 Amazon Redshift Success Stories**. [S. l.], 17 dez. 2014. Disponível em: [link para o site](#). Acesso em: 11 jun. 2023.

DORP, ERIK VAN. **Google BigQuery case study Mediexplain***. Disponível em: [link para o site](#). Acesso em: 11 jun. 2023.

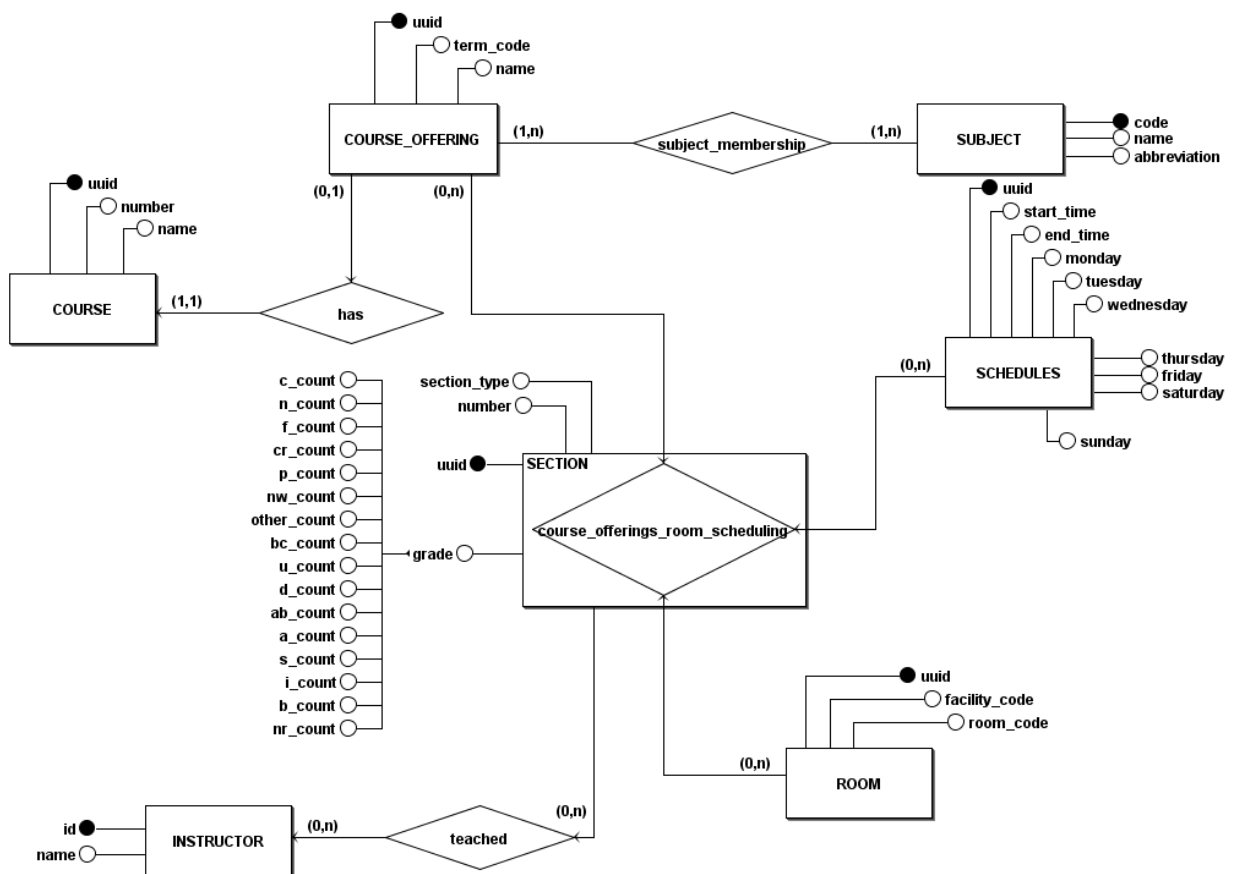
G) Documentação base de dados

A University of Wisconsin - Madison publica relatórios para todos os cursos (e seções desses cursos), instrutores, disciplinas e relatórios de notas para cada seção para cada semestre de outono e primavera desde 2006.

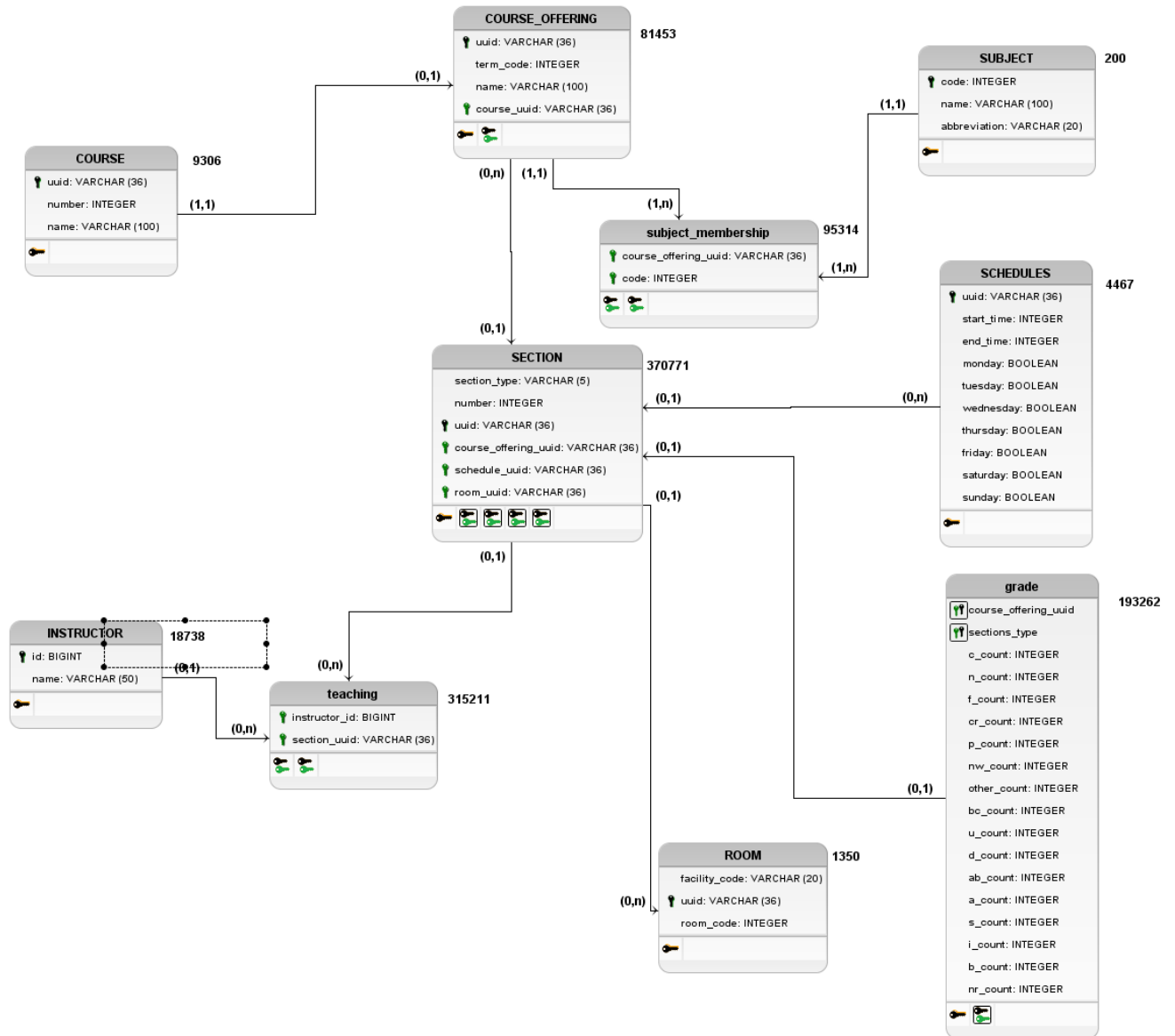
No contexto específico da disciplina de Sistema de Banco de Dados 2, oferecida também por uma Universidade, a escolha da base de dados se revela extremamente coerente com a realidade vivenciada pelos alunos. A importância dessa seleção reside no fato de que a base de dados reflete de maneira precisa e relevante os desafios e as demandas enfrentados no mundo real no campo dos sistemas de banco de dados.

Para início dos estudos, foi gerada uma documentação básica, documentada abaixo:

1. Diagrama Entidade-Relacionamento



2. Diagrama lógico de dados



A base de dados foi encontrada no Kaggle e está disponível no endereço virtual:

<https://www.kaggle.com/datasets/Madgrades/uw-madison-courses?resource=download&select=rooms.csv>