UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Bancos de Dados em Colunas

Luís Guilherme Gaboardi Lins - 180022962

Brasília, DF
2023

Introdução

Os Bancos de Dados em Colunas são uma tecnologia inovadora de armazenamento e gerenciamento de dados que desempenham um papel fundamental no contexto atual de grandes volumes de informações. Essa abordagem diferenciada de estruturação dos dados tem se mostrado uma alternativa eficiente para otimizar consultas e análises, oferecendo vantagens significativas em relação aos modelos tradicionais de bancos de dados. Neste relatório, exploraremos os conceitos, objetivos e aplicações dos Bancos de Dados em Colunas, além de analisar exemplos reais de seu uso em empresas, projetos e instituições.

Essa tecnologia é uma alternativa ao modelo tradicional de bancos de dados relacionais, que armazenam informações em linhas. Nessa abordagem, os dados são organizados e armazenados em colunas, em vez de registros completos, permitindo um acesso mais eficiente às informações específicas de interesse em consultas e análises. Enquanto nos bancos de dados relacionais as consultas geralmente envolvem a recuperação de todos os atributos de uma linha, nos Bancos de Dados em Colunas é possível selecionar apenas as colunas relevantes para uma determinada consulta, resultando em um processamento mais rápido e eficiente.

Os Bancos de Dados em Colunas são amplamente utilizados em diferentes contextos e setores da indústria. Eles encontram aplicação em empresas que lidam com grandes volumes de dados, como instituições financeiras, empresas de telecomunicações, varejistas e provedores de serviços de Internet. Além disso, essa tecnologia é adotada em projetos de ciência de dados, análise de negócios, pesquisa científica e qualquer domínio que exija consultas e análises eficientes de dados.

Os seus fundamentos estão baseados na estruturação dos dados em colunas individuais, armazenando valores similares em conjunto e permitindo a compressão e otimização dos dados. Essa abordagem possibilita um armazenamento mais eficiente, reduzindo o espaço necessário e melhorando o desempenho das operações de leitura e gravação. Além disso, a estruturação dos dados em colunas também permite a execução de operações paralelas e a

distribuição dos dados em diversos nós de processamento, o que facilita a escalabilidade horizontal do banco de dados.

Objetivo Principal dos Bancos de Dados em Colunas

Os Bancos de Dados em Colunas têm como objetivo facilitar a realização de consultas analíticas complexas, fornecendo um desempenho aprimorado nesses cenários. A estrutura de colunas permite a busca seletiva de dados específicos, eliminando a necessidade de acessar colunas desnecessárias e melhorando a eficiência das análises. Conforme destacado por Abadi (2012), essa abordagem de armazenamento de dados em colunas permite uma recuperação seletiva das informações necessárias, reduzindo o tempo de resposta das consultas e tornando-se especialmente benéfica em cenários analíticos, nos quais são realizadas operações de agregação, filtragem e análise complexa dos dados.

Além disso, os Bancos de Dados em Colunas se destacam pela alta taxa de compressão de dados, resultante da natureza da estruturação em colunas. Isso proporciona economia de espaço de armazenamento e torna a transferência de informações mais eficiente. A capacidade de escalabilidade horizontal é outra característica importante desses bancos de dados, permitindo lidar com grandes volumes de dados e picos de carga de maneira eficiente. Conforme mencionado por Stonebraker et al. (2005) em seu estudo sobre o C-Store, um sistema de gerenciamento de bancos de dados em colunas, essa escalabilidade é fundamental para garantir a disponibilidade e o desempenho do sistema, à medida que a demanda cresce.

Em resumo, os Bancos de Dados em Colunas têm o objetivo de otimizar o processamento e a análise de dados, oferecendo maior velocidade, eficiência e flexibilidade em consultas analíticas complexas. A recuperação seletiva de dados, a compressão eficiente de informações e a capacidade de escalabilidade horizontal são alguns dos aspectos que contribuem para o alcance desses objetivos, conforme evidenciado pelas pesquisas de Abadi (2012) e Stonebraker et al. (2005).

Vantagens dos Bancos de Dados em Colunas

Os Bancos de Dados em Colunas oferecem uma série de vantagens significativas em relação a outros modelos de armazenamento. Algumas das principais vantagens incluem:

Desempenho aprimorado em consultas analíticas: Os Bancos de Dados em Colunas são especialmente projetados para consultas analíticas complexas, permitindo um desempenho superior. A estrutura de colunas facilita a recuperação seletiva de dados específicos, eliminando a necessidade de acessar colunas desnecessárias, resultando em tempos de resposta mais rápidos e análises mais eficientes (ABADI, 2012; STONEBRAKER et al., 2005).

Compressão de dados eficiente: Devido à natureza dos dados organizados em colunas, os Bancos de Dados em Colunas têm uma alta taxa de compressão de dados. Valores repetidos em uma coluna podem ser armazenados de forma compacta, reduzindo o espaço necessário para armazenar as informações. Isso resulta em economia de armazenamento e transferência de dados mais eficiente (MANEGOLD, 2011).

Escalabilidade horizontal: Os Bancos de Dados em Colunas são altamente escaláveis horizontalmente, o que significa que é possível adicionar mais nós de processamento e armazenamento conforme a demanda cresce. Essa abordagem permite lidar com grandes volumes de dados e picos de carga de maneira eficiente, garantindo a disponibilidade e o desempenho do sistema (ABOUZEID et al., 2014).

Flexibilidade na adição de colunas: Uma das vantagens dos Bancos de Dados em Colunas é a capacidade de adicionar novas colunas sem afetar a estrutura das tabelas existentes. Isso permite uma maior flexibilidade e adaptabilidade aos requisitos em constante mudança do sistema, facilitando a incorporação de novos campos de dados sem a necessidade de modificações extensas no esquema (ABADI, 2012).

Os Bancos de Dados em Colunas se destacam em relação aos Bancos de Dados Relacionais por diversas razões. Em comparação aos bancos de dados relacionais, os Bancos de Dados em Colunas oferecem desempenho aprimorado em consultas analíticas, devido à recuperação seletiva de dados e

à eliminação do acesso a colunas desnecessárias. Além disso, a compressão eficiente de dados nesse modelo resulta em economia de espaço de armazenamento e transferência de dados mais eficiente.

Essas características destacam a superioridade dos Bancos de Dados em Colunas em relação aos Bancos de Dados Relacionais, especialmente em termos de desempenho analítico, compressão eficiente de dados, escalabilidade horizontal e flexibilidade na adição de colunas.

Desvantagens dos Bancos de Dados em Colunas

Apesar de suas vantagens, os Bancos de Dados em Colunas também possuem algumas desvantagens que devem ser consideradas:

Desempenho inferior em consultas transacionais e atualizações frequentes: Os Bancos de Dados em Colunas são projetados principalmente para consultas analíticas e não são ideais para operações de escrita intensiva ou atualizações frequentes de dados. Isso ocorre porque a estrutura de colunas exige alterações em várias colunas para atualizar um único registro, o que pode resultar em desempenho inferior em cenários transacionais.

Complexidade de modelagem: A modelagem de dados em um Banco de Dados em Colunas pode ser mais complexa do que em outros modelos. A necessidade de entender as consultas e os padrões de acesso aos dados é essencial para projetar uma estrutura eficiente. A segmentação adequada dos dados em colunas é crucial para garantir consultas otimizadas e eficazes.

Custo inicial mais elevado: A implementação de um Banco de Dados em Colunas pode envolver custos iniciais mais elevados em termos de hardware especializado e infraestrutura. Os requisitos de armazenamento e processamento podem ser maiores em comparação com outros modelos, o que pode exigir investimentos adicionais para atender às demandas de desempenho e escalabilidade.

Em operações transacionais e atualizações frequentes, seu desempenho pode ser inferior devido à necessidade de alterar várias colunas para atualizar um único registro. Isso pode resultar em um desempenho inferior em cenários transacionais, nos quais os Bancos de Dados Relacionais tendem

a ser mais adequados.

Além disso, a modelagem de dados em um Banco de Dados em Colunas pode ser mais complexa em comparação com outros modelos, exigindo um planejamento cuidadoso e uma compreensão profunda das consultas e padrões de acesso aos dados (ABADI, 2012; STONEBRAKER et al., 2005). Outra desvantagem a ser considerada é o custo inicial mais elevado da implementação de um Banco de Dados em Colunas. Essa tecnologia pode requerer hardware especializado e infraestrutura adequada para atender às demandas de desempenho e escalabilidade. Os requisitos de armazenamento e processamento podem ser maiores em comparação com outros modelos de banco de dados, o que pode resultar em investimentos adicionais (MANEGOLD, 2011; ABOUZEID et al., 2014).

Exemplos de Uso

Os Bancos de Dados em Colunas têm sido amplamente adotados em diversas indústrias para impulsionar a análise de dados e obter insights valiosos. Alguns exemplos de uso interessantes incluem os abaixo.

Uma empresa que adotou a tecnologia de Bancos de Dados em Colunas é a empresa de comércio eletrônico Amazon. A Amazon implementou um sistema de análise de dados chamado Redshift, que utiliza um modelo de armazenamento em colunas para lidar com suas vastas quantidades de informações de vendas, produtos e clientes. A transição para o Redshift permitiu à Amazon realizar consultas analíticas mais rápidas e eficientes, especialmente ao lidar com consultas complexas que envolvem agregações e análises de grandes conjuntos de dados. A capacidade de recuperação seletiva de dados e a compressão eficiente proporcionaram uma melhoria significativa no desempenho das análises, permitindo à empresa obter insights valiosos para otimizar suas operações de negócios.

Outro exemplo é o projeto Sloan Digital Sky Survey (SDSS), que utiliza Bancos de Dados em Colunas para armazenar e analisar dados astronômicos. O SDSS é um dos maiores levantamentos astronômicos já realizados, coletando uma quantidade massiva de informações sobre objetos celestes. A

adoção de Bancos de Dados em Colunas no projeto permitiu uma recuperação rápida e eficiente dos dados astronômicos para pesquisas científicas. A estrutura de colunas facilitou a análise de atributos específicos dos objetos, como brilho, localização e características espectrais. Isso proporcionou uma capacidade avançada de pesquisa e descoberta astronômica, impulsionando a compreensão e o conhecimento do universo.

Esses exemplos demonstram que a adoção de Bancos de Dados em Colunas pode trazer benefícios significativos para as empresas e projetos. A capacidade de realizar consultas analíticas mais rápidas e eficientes, juntamente com a recuperação seletiva de dados e a compressão eficiente, possibilita insights valiosos e melhor tomada de decisões. No entanto, é importante ressaltar que a adoção de qualquer nova tecnologia também pode apresentar desafios.

A implementação de um Banco de Dados em Colunas requer planejamento adequado, infraestrutura adequada e conhecimento especializado para garantir o sucesso do projeto. Além disso, pode haver custos iniciais mais elevados e desafios de modelagem de dados. Portanto, é fundamental avaliar cuidadosamente os requisitos e considerar os prós e contras antes de adotar essa tecnologia em uma empresa ou projeto.

Bibliografias Pesquisadas

1. ABADI, Daniel J. Column-Oriented Database Systems. 2012.
2. STONEBRAKER, Michael, et al. C-Store: A Column-oriented DBMS. 2005.
3. MANEGOLD, Stefan. Columnar Databases. 2011.
4. ABOUZEID, Azza, et al. Big Data: A Column-Oriented DBMS Perspective. 2014.

Base de Dados (Documentação)

Link: https://github.com/datacharmer/test_db

Diagrama Conceitual (DE-R):

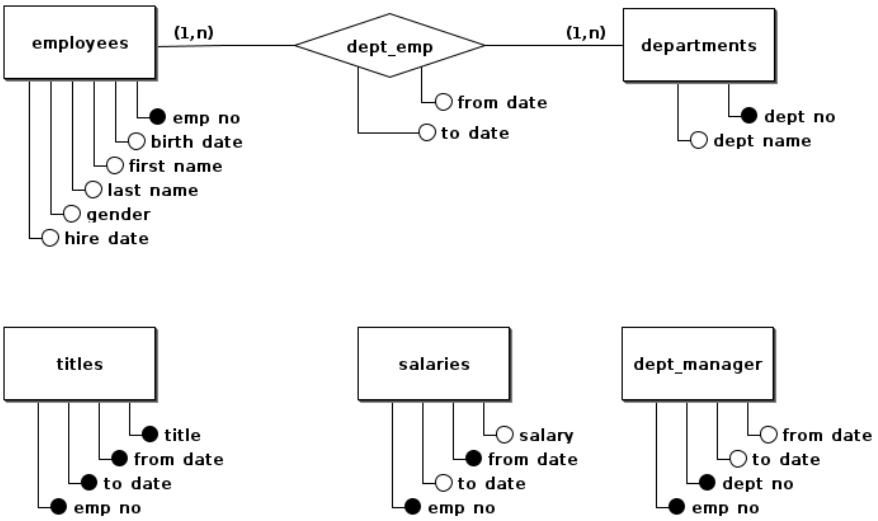

Diagrama Lógico (DLD)