

UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Base de Dados 2

Trabalho Final - TF

Bancos de Dados Semiestruturados

Artur Vinicius Dias Nunes – 190142421

Brasília, DF
2023

Bancos de Dados Semiestruturados

Banco de dados semi-estruturado é um tipo de banco de dados que não possui uma estrutura rígida como os bancos de dados relacionais tradicionais, mas também não é completamente desestruturado. Em vez disso, ele permite o armazenamento de dados com uma estrutura flexível, onde diferentes registros podem ter diferentes campos e tipos de dados chamados de dados semiestruturados.

Dados semi-estruturados apresentam uma representação heterogênea, não sendo nem completamente não-estruturados nem estritamente tipados. Os dados semi-estruturados geralmente são representados em formatos como JSON (JavaScript Object Notation), XML (eXtensible Markup Language) ou YAML (YAML Ain't Markup Language). Esses formatos permitem que os dados sejam organizados em uma estrutura hierárquica usando pares chave-valor ou tags. Uma característica importante dos dados semi-estruturados é que eles podem conter anotações ou metadados adicionais que fornecem informações sobre o próprio dado. Por exemplo, um documento XML pode conter tags que descrevem o tipo de conteúdo, a data de criação, o autor, entre outros detalhes relevantes.

Os dados semi-estruturados são comumente encontrados em várias fontes, como páginas da web, documentos HTML, feeds de mídia social, logs de eventos, documentos XML de fontes diferentes, entre outros. Eles são especialmente úteis quando se lida com dados que possuem uma estrutura variável ou desconhecida, pois permitem acomodar mudanças na estrutura dos dados sem a necessidade de alterar um esquema rígido.

Vantagens Banco de Dados Semiestruturado:

- **Flexibilidade:** Um banco de dados semi-estruturado permite que você armazene dados de diferentes tipos e estruturas. Isso é particularmente útil quando você lida com dados heterogêneos, como documentos HTML, XML, JSON ou outros formatos não relacionais. A flexibilidade inerente ao modelo semi-estruturado permite que você adicione, modifique ou remova campos facilmente, sem a necessidade de redefinir a estrutura do banco de

dados.

- **Escalabilidade:** O modelo semi-estruturado é altamente escalável, o que significa que pode lidar com grandes volumes de dados sem comprometer o desempenho. Isso é especialmente importante quando você precisa lidar com dados não estruturados, como logs de servidor, dados de mídia social ou dados da web, que podem crescer rapidamente ao longo do tempo.
- **Integração de dados:** O modelo semi-estruturado facilita a integração de dados de várias fontes. Como não exige um esquema rígido, você pode combinar e agregar dados de diferentes formatos e estruturas em um único banco de dados. Isso é especialmente útil em ambientes onde você precisa extrair informações de várias fontes de dados, como em empresas que têm dados armazenados em sistemas legados, bancos de dados relacionais e APIs externas.
- **Consultas flexíveis:** Bancos de dados semi-estruturados oferecem flexibilidade nas consultas. Você pode realizar consultas ad hoc sem a necessidade de estruturar os dados de antemão. Além disso, a linguagem de consulta utilizada (como XQuery ou XPath) permite realizar consultas complexas, como pesquisas por padrões, navegação hierárquica e filtragem de dados.
- **Suporte a dados não estruturados:** Os bancos de dados semi-estruturados são projetados para lidar com dados não estruturados ou parcialmente estruturados, que não se encaixam bem em um modelo de banco de dados relacional. Isso inclui documentos, arquivos de texto, dados de mídia e muito mais. Ao usar um banco de dados semi-estruturado, você pode armazenar, pesquisar e analisar facilmente esses tipos de dados.

Desvantagens Banco de Dados Semiestruturado:

- **Falta de integridade de dados:** Ao contrário dos bancos de dados relacionais, os bancos de dados semi-estruturados não impõem um

esquema rígido ou restrições de integridade. Isso significa que não há garantia de consistência dos dados armazenados. Dados inconsistentes ou com qualidade inferior podem ser armazenados sem um controle rigoroso, o que pode levar a problemas de integridade e confiabilidade dos dados.

- Dificuldade de consultas complexas: Embora os bancos de dados semi-estruturados ofereçam flexibilidade nas consultas, a complexidade das consultas pode aumentar significativamente à medida que a estrutura dos dados se torna mais complexa. O modelo semi-estruturado não é adequado para consultas complexas e sofisticadas que exigem manipulação de várias relações e agregações complexas. Isso pode dificultar a obtenção de informações precisas e relevantes dos dados.
- Ausência de um esquema formal: Os bancos de dados semi-estruturados não possuem um esquema formal predefinido, o que pode dificultar a compreensão e interpretação dos dados armazenados. A falta de um esquema bem definido também pode dificultar a migração ou a integração dos dados com outros sistemas ou aplicativos.
- Desempenho e eficiência: Consultas em bancos de dados semi-estruturados podem ser mais lentas e menos eficientes em comparação com bancos de dados relacionais. Isso ocorre porque os bancos de dados semi-estruturados não têm otimizações específicas para consultas complexas e não possuem índices estruturados para acelerar a recuperação de dados. O desempenho pode ser comprometido à medida que o volume de dados aumenta.
- Ferramentas e habilidades específicas: O uso de bancos de dados semi-estruturados pode exigir ferramentas e habilidades específicas para gerenciar, consultar e analisar os dados. Isso pode limitar a disponibilidade de recursos e profissionais especializados nessa área, o que pode aumentar a complexidade e os custos de implantação e manutenção de um banco de dados semi-estruturado.

História de Sucesso utilizando Banco de Dados Semiestruturados

Uma história de sucesso no uso de banco de dados semi-estruturados é a da empresa Netflix. A Netflix é uma plataforma de streaming de vídeo que oferece uma ampla variedade de conteúdo aos seus assinantes. Para fornecer uma experiência personalizada e recomendada aos usuários, a Netflix precisa coletar e analisar grandes volumes de dados provenientes de diferentes fontes.

A empresa lida com uma enorme quantidade de dados semiestruturados, como informações de perfil do usuário, histórico de visualizações, avaliações, preferências de gênero, idioma e muitos outros atributos. Esses dados são altamente heterogêneos e variam de acordo com o tipo de conteúdo, região e preferências individuais dos usuários.

Para lidar com essa complexidade de dados, a Netflix adotou uma abordagem baseada em bancos de dados semi-estruturados.

Esses bancos de dados semi-estruturados permitem à Netflix armazenar, processar e analisar grandes volumes de dados em tempo real. Eles permitem a escalabilidade horizontal, o que significa que a Netflix pode adicionar mais servidores conforme a demanda de dados aumenta. Além disso, eles fornecem flexibilidade para lidar com diferentes tipos de dados, permitindo que a Netflix adicione ou modifique campos sem a necessidade de reestruturar todo o banco de dados.

Usando esses bancos de dados semi-estruturados, a Netflix pode executar análises sofisticadas e algoritmos de recomendação para fornecer aos usuários sugestões personalizadas de filmes e séries com base em seus hábitos de visualização, preferências e comportamentos.

A abordagem da Netflix no uso de bancos de dados semi-estruturados permitiu que eles enfrentassem os desafios de lidar com grandes volumes de dados heterogêneos, garantindo uma experiência personalizada e de alta qualidade para seus usuários.

História de Fracasso utilizando Banco de Dados Semiestruturados

Embora os bancos de dados semi-estruturados tenham muitos casos de

sucesso, também há exemplos de desafios e fracassos associados ao seu uso. Um exemplo de fracasso é o caso do banco de dados semi-estruturado da NASA (National Aeronautics and Space Administration) conhecido como NOSQL (Not Only SQL).

Em 2005, a NASA lançou o projeto NOSQL com o objetivo de criar um banco de dados semi-estruturado para armazenar e consultar dados provenientes de várias missões espaciais e experimentos científicos. O objetivo era lidar com a natureza heterogênea e não estruturada dos dados coletados, que incluíam informações sobre imagens, telemetria, metadados, entre outros.

No entanto, o projeto NOSQL enfrentou vários desafios e acabou sendo considerado um fracasso. Algumas das principais razões incluem:

Complexidade da modelagem de dados: O projeto NOSQL enfrentou dificuldades ao tentar modelar os dados de forma adequada para consulta e análise eficientes. A natureza semi-estruturada dos dados tornou difícil definir um esquema adequado e estabelecer relacionamentos entre os diferentes tipos de dados.

Desempenho insatisfatório: O NOSQL enfrentou problemas de desempenho ao lidar com consultas complexas e volumes significativos de dados. O modelo semi-estruturado utilizado no projeto não oferecia a mesma otimização de consultas que os bancos de dados relacionais, levando a tempos de resposta lentos e ineficiência na recuperação dos dados.

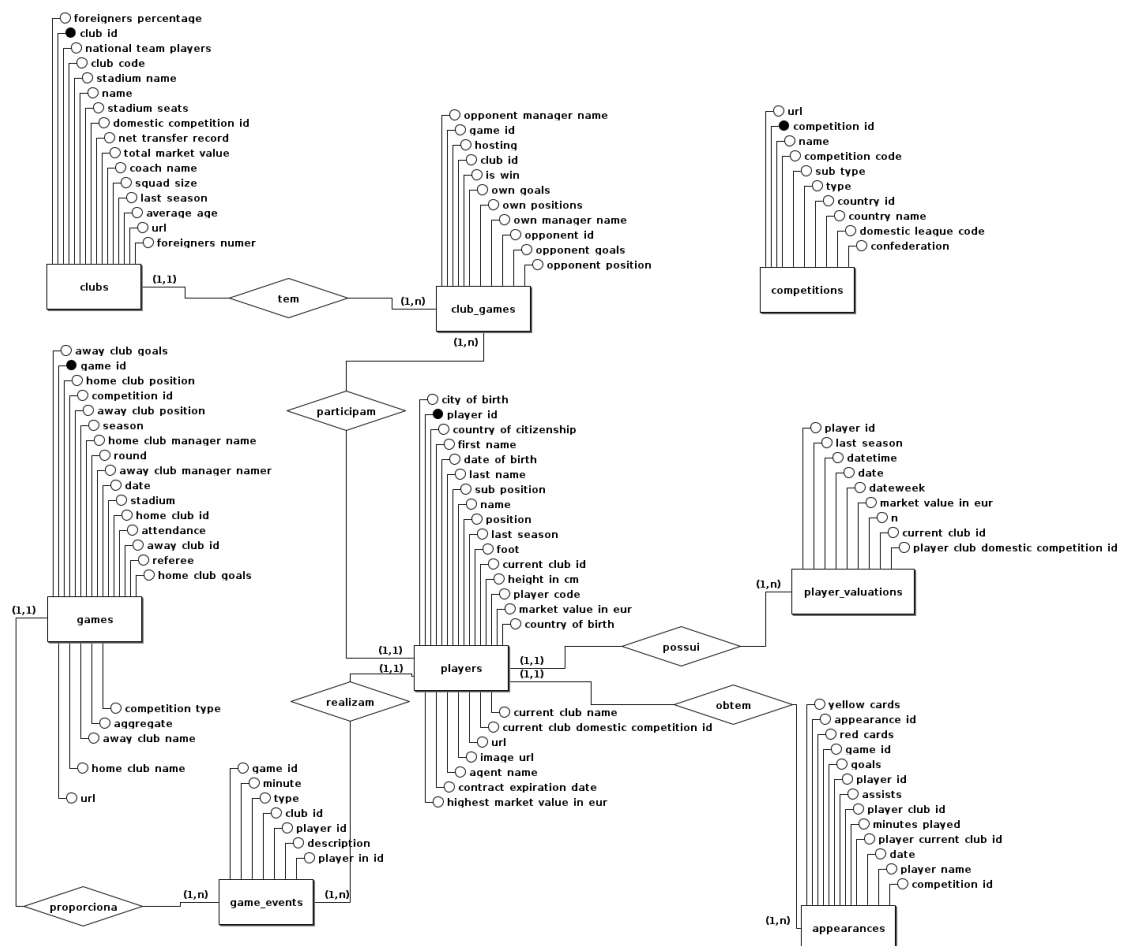
Limitações de ferramentas e habilidades: A falta de ferramentas adequadas e a escassez de profissionais com experiência em bancos de dados semi-estruturados contribuíram para o fracasso do projeto NOSQL. A equipe da NASA enfrentou dificuldades para gerenciar, consultar e analisar os dados de forma eficaz devido à falta de recursos especializados.

Como resultado, a NASA decidiu encerrar o projeto NOSQL e buscar soluções alternativas para lidar com a gestão de dados semiestruturados. Essa experiência destaca a importância de avaliar cuidadosamente os requisitos do projeto, a complexidade dos dados e a disponibilidade de recursos antes de optar por um banco de dados semi-estruturado. Além disso, a falta de experiência e ferramentas adequadas pode impactar negativamente o sucesso na implementação de um banco de dados semi-estruturado.

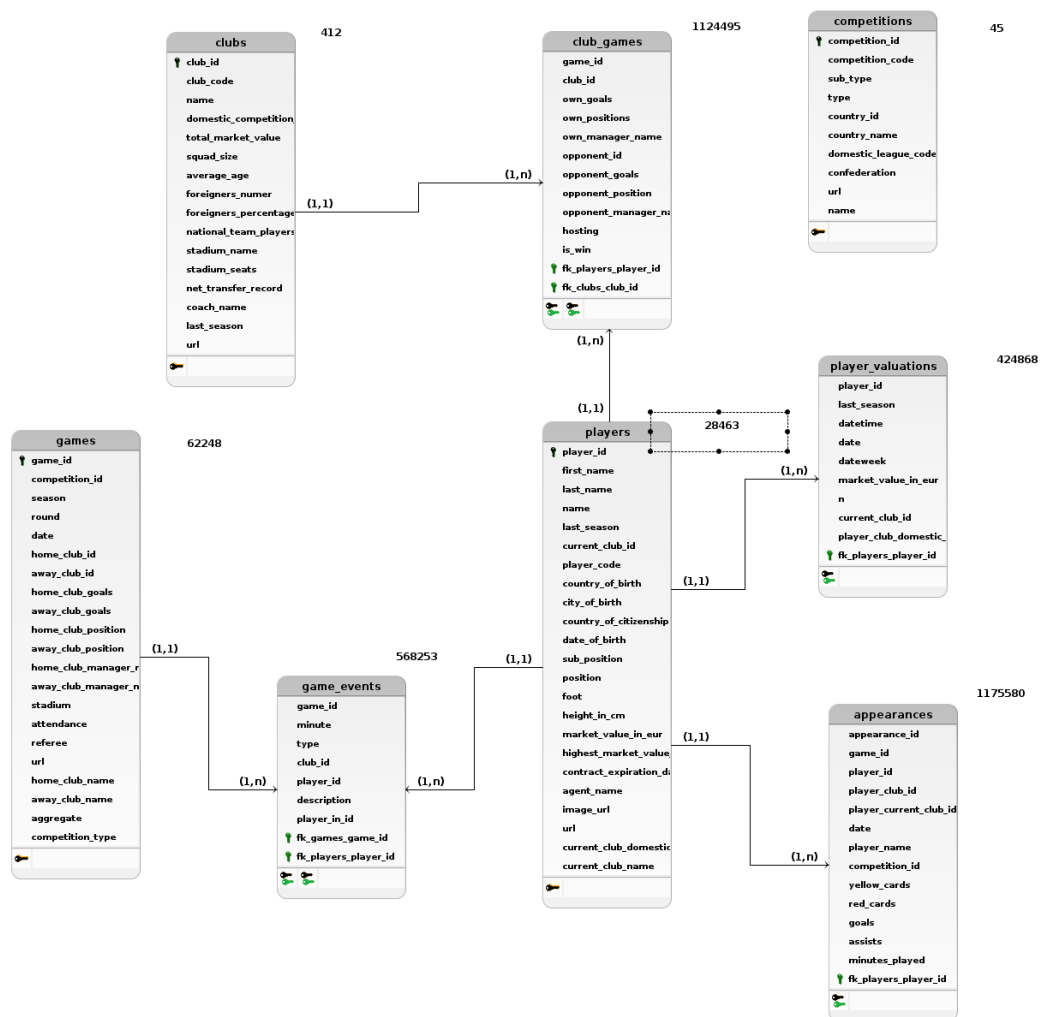
Base de Dados

O conjunto de dados é composto por informações sobre competições, jogos, clubes, jogadores e aparições que são atualizados automaticamente uma vez por semana. Cada tabela contém os atributos da entidade e os IDs que podem ser usados para juntá-los.

DER - Diagrama Entidade Relacionamento



DLD - Diagrama Lógico de Dados



Endereço Base de Dados :
<https://www.kaggle.com/datasets/davidcariboo/player-scores?select=appearances.csv>

Referência Bibliográfica

- Aitan Pontes. Disponível em:
<https://monografias.ufma.br/jspui/bitstream/123456789/3516/1/AITAN-PONTES.pdf>. Acesso em: 12 jun. 2023.
- Mello Ronaldo, Dados semi-estruturados. Disponível em:
<https://www.ime.usp.br/~jef/semi-estruturado.pdf>. Acesso em: 12 jun. 2023.
- Service WordPress IGTI. Dados semiestruturados. Disponível em:
<https://servicewordpressigti.azurewebsites.net/dados-semiestruturados/>. Acesso em: 12 jun. 2023.
- Tok Wang Ling, Semi-Structured Database Design. 1ªed.Springer; Softcover reprint of hardcover, 2005.