**Relatório Trabalho Prático**

Footbal Data from Transfermarkt

14829, José Pedro Gomes

15708, Pedro Carvalho

15709, José Carreira

Barcelos, dezembro 2023

Índice

[1. Introdução 3](#_Toc151250619)

[1.1. Objetivos do Projeto 3](#_Toc151250620)

[1.2. Recursos 4](#_Toc151250621)

[1.3. Carga de trabalho 5](#_Toc151250622)

[2. Modelo de dados relacional — SQL Server 5](#_Toc151250623)

[2.1. Data model design 6](#_Toc151250624)

[2.2. Data model tuning and testing 6](#_Toc151250625)

[3. Document-based data model — MongoDB 6](#_Toc151250626)

[3.1. Data model design 7](#_Toc151250627)

[3.2. Data model tuning and testing 7](#_Toc151250628)

[4. Graph-based data model — Neo4J 7](#_Toc151250629)

[4.1. Data model design 7](#_Toc151250630)

[4.2. Data model tuning and testing 7](#_Toc151250631)

[5. Discussion and conclusion 7](#_Toc151250632)

[References 7](#_Toc151250633)

[Contributions 8](#_Toc151250634)

[Annexes 9](#_Toc151250635)

[Annex 01 – <<Nome do Annex 01>> 9](#_Toc151250636)

[Annex 02 – <<Nome do Annex 02>> 9](#_Toc151250637)

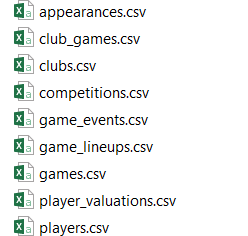
# Introdução

Este relatório é apresentado no âmbito do trabalho prático da unidade curricular de Bases de Dados Avançadas, do Mestrado de Engenharia Informática, lecionada pelo Professor Joaquim Silva. Esta UC tem como objetivo criar competências para o desenvolvimento e otimização de sistemas de armazenamento de dados relacionais e não relacionais, assim com análise de *performance* e persistência de dados. Também se pretende aumentar o conhecimento em relação a modelos e sistemas de dados orientados a documentos, *key/value* egrafos.

O desafio consiste em criar bases de dados de tipologias diferentes: relacional, não relacional e grafos. Modelar as bases de dados, fazer o *design*, implementar e monitorizar a sua *performance*. Pretende-se descrever o processo de cada base de dados, fazer comparações de resultados, conseguindo perceber as diferenças entre as BDs e quais seriam as melhores formas de otimizar cada uma delas, conseguindo identificar qual seria a melhor escolha para responder a determinadas perguntas.

## Objetivos do Projeto

O tema escolhido pelo grupo foi um *dataset* (constituído por vários ficheiros *CSV*) de estatísticas de futebol do Transfermarkt obtido no kaggle ([*https://www.kaggle.com/datasets/davidcariboo/player-scores*](https://www.kaggle.com/datasets/davidcariboo/player-scores)) com o nome “Footbal data from Transfermarkt”, que contém informação de várias competições nacionais e europeias desde 2012.

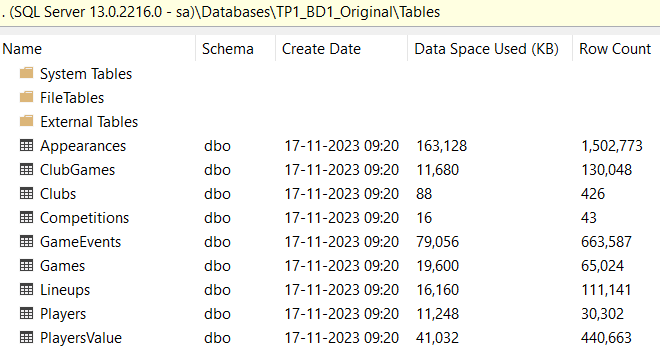


O objetivo passa por integrar todos os ficheiros *CSV,* pois cada ficheiro corresponde a determinada parte da informação, analisar o seu conteúdo e de seguida importar a informação para uma base de dados em *SQL*, para se conseguir ter melhor perceção do potencial, ou eventuais falhas, dos dados. Depois da análise de toda a informação disponível será normalizada a BD, serão validados e importados os dados e será testada a *performance* da BD, este trabalho de seguida também será efetuado em *MongoDB* e *Neo4j*.

## Recursos

De seguida será explicada a estrutura de informação disponível, que deverá ser capaz de dar respostas às necessidades de uma plataforma de resultados, com informação detalhada dos jogos, jogadores e equipas.

A estrutura do *dataset* original foi importada para uma base de dados em SQL (“TP1\_BD1\_Original”), tendo sido criadas tabelas diretamente a partir de cada ficheiro *CSV* disponível. Como é visível na imagem abaixo, o *dataset* abrange informação variada em relação aos jogos, competições e intervenientes, contém alguma complexidade a relacionar as várias informações disponíveis e possui um volume de registos relativamente elevado, o que deverá permitir as análises pretendidas.



* *Appearances*, detalhe de cada jogador presente no jogo (golos, cartões, etc);
* *ClubGames,* informação de cada jogo de determinada equipa;
* *Clubs,* informação de cada equipa;
* *Competitions,* competições disponíveis (campeonatos, taças nacionais, competições europeias, etc.);
* *GameEvents,* detalhe deeventos do jogo (golos, substituições, cartões, etc.);
* *Games,* informação de cada jogo;
* *Lineups,* onze inicial e substitutos que entraram no jogo;
* *Players,* informação relacionada com os jogadores;
* *PlayersValue,* detalhe das avaliações de cada jogador.

A partir desta informação pretende-se ser capaz efetuar análises detalhadas de competições, jogos, equipas, jogadores, treinadores, árbitros e estádios. O objetivo do grupo passará por estruturar de forma correta toda esta informação, para se conseguir desenvolver uma base de dados capaz de dar as respostas necessárias, de forma rápida e assertiva.

## Carga de trabalho

O dataset depois de ser importado para SQL na BD “TP1\_BD1\_Original” permite ter a noção da dimensão de cada tabela, depois ser de analisar a informação chegou-se a algumas conclusões em relação aos totais por época (tendo por base os registos disponíveis desde 2012):

* Cada época tem entre 5.5k e 6k registos de jogos;
* Os registos de eventos por jogo variam entre 46k e 73k registos;
* A informação relacionada com os detalhes do jogador em cada jogo varia entre 120k e 145k.

As operações mais frequentes deverão estar relacionadas com informação dos jogos, equipas, jogadores e eventos do jogo. Pelo que uma correta modelação da BD, armazenamento dos dados e a otimização no acesso a estes dados seja fundamental.

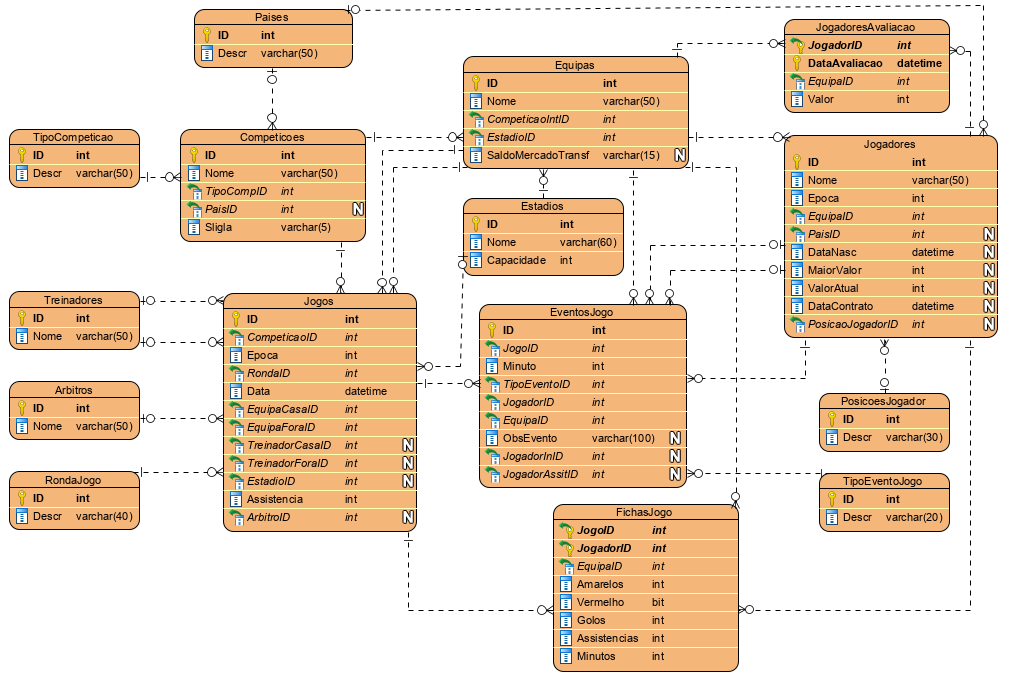
# Modelo de dados relacional — SQL Server

Depois de se analisar os dados chegou-se à conclusão que havia muita redundância de informação e que seria necessário partir a informação em várias entidades, de forma a garantir a integridade e a normalização dos dados.

Foram identificadas as principais tabelas:

* Competições;
* Equipas;
* Jogos;
* Eventos do Jogo;
* Ficha de Jogo;
* Jogadores;
* Avaliação dos Jogadores;

## Data model design

<< Identify the entities and their relationships. **Design and** **implement** the relational data model. **Populate the database** with meaningful data. The database should have enough size to draw conclusions about the performance and suitability of the data model.>> 

## Data model tuning and testing

<<Test and evaluate some scenarios for the database system. **Execute** operations previously described. **Implement solutions** to improve the operations performance, e. g., the creation of indexes. Results, performance metrics, and the conclusions **should be properly documented**.>>

# Document-based data model — MongoDB

<<Include here the tasks performed and its results. >>

## Data model design

<< Identify the entities and their relationships. **Design and** **implement** the data model – Collection-Relationship Diagram. **Populate the database** with meaningful data. The database should have enough size to draw conclusions about the performance and suitability of the data model.>>

## Data model tuning and testing

<<Test and evaluate some scenarios for the database system. **Execute** operations previously described. **Implement solutions** to improve the operations performance, e. g., the creation of indexes. Results, performance metrics, and the conclusions **should be properly documented**.>>

# Graph-based data model — Neo4J

<<Include here the tasks performed and its results. >>

## Data model design

<< Identify the entities and their relationships. **Design and** **implement** the Graph-based data model. **Populate the database** with meaningful data. The database should have enough size to draw conclusions about the performance and suitability of the data model.>>

## Data model tuning and testing

<<Test and evaluate some scenarios for the database system. **Execute** operations previously described. **Implement solutions** to improve the operations performance, e. g., the creation of indexes. Results, performance metrics, and the conclusions **should be properly documented**.>>

# Discussion and conclusion

<< Taking into account the results and conclusions, please provide a grounded recommendation for the database model that should be adopted. >>

# References

<< In this section, should be presented in APA format the list of bibliographic sources that were relevant the execution of the work. All the references must be cited on the report.>>

# Contributions

<<Here must be reported the **individual contributions** of each student. For each project task, identify the students that have actively participated.>>

# Annexes

<< The annexes should be used for the inclusion of additional information needed for a better understanding of the report or to complement it. The annexes created should be numbered, start at the top of a new page and be cited in the body text of the report.>>

## Annex 01 – <<Nome do Annex 01>>

## Annex 02 – <<Nome do Annex 02>>

Etc.