

# TODO titulo

Trabalho elaborado no âmbito da Unidade Curricular de Armazenamento para Big Data do 2º ano da Licenciatura de Ciência de Dados do Instituto Universitário de Lisboa ISCTE

André Plancha; 105289

Another Author; Num

07/16/2020

## Introdução

A Associação de Tenistas Profissionais (*ATP*) é um órgão de ténis profissional masculino, organizando torneios do desporto globalmente. A organização contém na sua base de dados um conjunto de jogos e jogadores que participaram em torneios pelo menos desde 1914, e incluem todos os grandes torneios do circuito masculino, incluindo os torneios de Grand Slam. O objetivo deste trabalho será limpar e preparar os dados de um modelo não-relacional para um modelo-relacional, para que possa ser utilizado em análises posteriores.

## Importação dos dados

Para o nosso projeto foi-nos provisionado o ficheiro *atpplayers.json*, que contém os jogos feitos pelos jogadores. Para importar este ficheiro, foi utilizado o comando *mongoimport*:

```
mongoimport `
  --db atp `
  --collection games `
  --file "$pwd\data\atpplayers.json"
```

Isto significa que foi criada uma base de dados chamada *atp*, e uma coleção chamada *games*, que contém os dados do ficheiro *atpplayers.json*.

```
use atp;
db.games.find({}, {_id:0}).limit(5);
```

<b>Born</b>	Belgrade, Serbia	Belgrade, Serbia	Belgrade, Serbia	Belgrade, Serbia	Belgrade, Serbia
<b>Date</b>	2022.02.21 - 2022.02.26	2021.08.30 - 2021.09.12	2021.11.15 - 2021.11.21	2021.08.30 - 2021.09.12	2021.11.01 - 2021.11.07
<b>GameRank</b>	26	145	5	121	
<b>GameRound</b>	Round of 16	Round of 128	Round Robin	Round of 64	Round of 64
<b>Ground</b>	Hard	Hard	Hard	Hard	Hard

<b>Hand</b>	Right- Handed, Two-Handed Backhand	Right- Handed, Two-Handed Backhand	Right- Handed, Two-Handed Backhand	Right- Handed, Two-Handed Backhand	Right- Handed, Two-Handed Backhand
<b>Height</b>	188	188	188	188	188
<b>Location</b>	Dubai, U.A.E.	New York, NY, U.S.A.	Turin, Italy	New York, NY, U.S.A.	Paris, France
<b>Oponent</b>	Karen Khachanov	Holger Rune	Andrey Rublev	Tallon Griekspoor	bye
<b>PlayerName</b>	Novak Djokovic	Novak Djokovic	Novak Djokovic	Novak Djokovic	Novak Djokovic
<b>Prize</b>	\$2,794,840	\$27,200,000	\$7,250,000	\$27,200,000	5,207,405
<b>Score</b>	63 76	61 67, 62 61	63 62	62 63 62	null
<b>Tournament</b>	Dubai	US Open	Nitto ATP Finals	US Open	ATP Masters 1000 Paris
<b>WL</b>	W	W	W	W	

nota: Foi retirado a coluna LinkPlayer para melhor visualização

A coleção contém 15 colunas: \* **PlayerName**: Nome do jogador do jogo \* **Born**: Onde este jogador nasceu (cidade, país) \* **Height**: Altura deste jogador (cm) \* **Hand**: A mão dominante do jogador, e o tipo de *backhand* que utiliza \* **LinkPlayer**: link para a página do jogador em atptour.com \* **Tournament**: nome do torneio do jogo \* **Location** : A cidade e país onde o torneio foi realizado \* **Date**: Período de tempo do torneio \* **GameRound**: fase do jogo no torneio \* **GameRank**: *ATP Rankings* do jogo \* **WL**: Vitória ou Derrota (W ou L) \* **Opponent**: Nome do Oponente \* **Score**: Sets do jogo

## Preparar dos dados

Para preparar os dados, nós planeamos transformar a nossa coleção em coleções diferentes, de forma a representar o modelo relacional, para facilitar a sua transição. Para isso, desenhámos o nosso diagrama do modelo relacional pretendido:

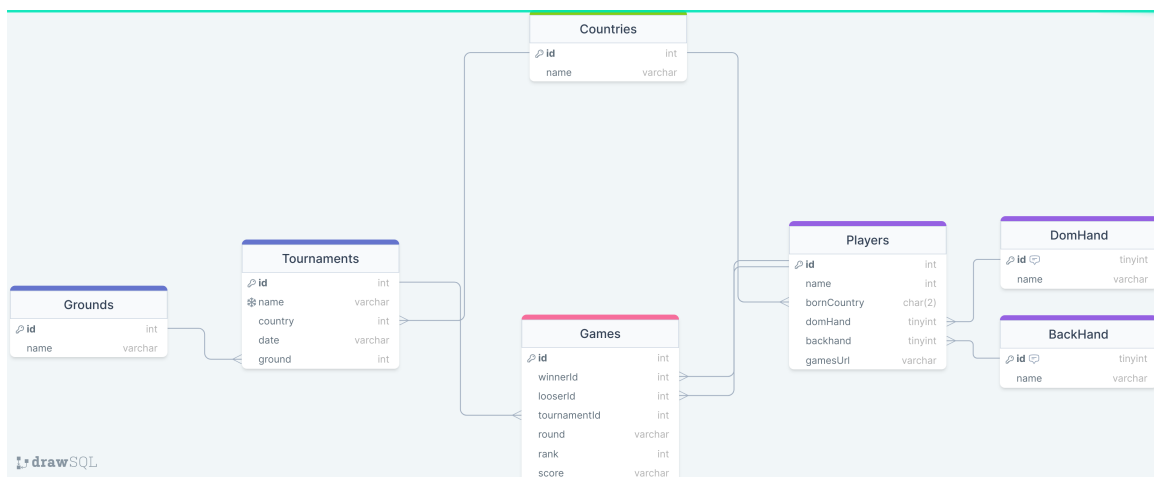


Figura 1: Diagrama do modelo relacional