Assignment 2: Neo4j Project – A chess graph database

Natasa Farmaki - DS3517018

Stratos Gounidellis - DS3517005

Course: Data Mining Techniques

Professor: Y.Kotidis

Assistant: I.Filippidou

Athens, April 2018

Introduction

Graph databases constitute the most efficient and effective way of representing and utilizing interconnected data, i.e. data, whose interpretation and values depends mainly on the understanding of their interconnections. Neo4j is according to the DB-engines ranking the most popular database. More specifically, Neo4j is an open source and world's leading graph database management system. It excels at manipulating data with intriguing relationships and at exploiting those relationships to extract useful knowledge. Its developers describe it as an ACID-compliant transactional database with native graph storage and processing.

In that project, we aim to show the functionality of Neo4j by using it as a database for chess data. In this case, there are relationships between players, events, games and positions. Consequently, it is rational to use a graph database like Neo4j to represent those relationships and the respective entities.

The relative project files are available in the following link: https://dataminersns.github.io/miningRepo/

Input files' description

In order to create the graph database, two .csv files were used, one for the games and one for the moves. Those files were created with the python script "generateCSVs.py". More specifically, as far as the games file is concerned, the initial data file was used in order to extract the needed information i.e. the name of the white player, the name of the black player, the date of the game, number of moves and of half moves, the result, the white ELO and the black ELO, the number of the game, the event in the context of which the game took place, the site and the date of the event, the round, the ECO and the opening. For instance, the first game's data were transformed as shown below:

```
========================== Game =====
White: Zukertort  Johannes H
Black: Steinitz  Wilhelm
Date: 1886.01.11
HalfMoves: 92
Moves: 46
Result: Black
WhiteElo: 0
BlackElo: 0
GameNumber: 1
Event: World Championship 1st
Site: USA
EventDate: 1886.01.11
Round: 1
ECO: D11
Opening: Queen's Gambit Declined Slav
```

```
White,Black,Date,HalfMoves,Moves,Result,WhiteElo,BlackElo,GameNumber,Event,Site,EventDate,Round,ECO,Opening
Zukertort  Johannes H,Steinitz  Wilhelm,1886.01.11,92,46,Black,0,0,1,World Championship 1st,USA,1886.01.11,1,D11,Queen's Gambit Declined Slav
```

As far as the moves' file is concerned, again the initial file was used in order to extract the needed information i.e. the number of the move, the move itself, the side of the player, the FEN and the number of the game. For each move, we combined its information with the information of the next move. This is useful for the graph creation in Neo4j. For instance, the first two moves of the first game were transformed as shown below:

```
---------------------------------------------------------- Game Moves ----------------------------------
MoveNumber: 1,  Side: white,  Move: d4,  FEN: rnbqkbnrpppppppp883P48PPP1PPPPRNBQKBNR,  GameNumber: 1
MoveNumber: 2,  Side: black,  Move: d5,  FEN: rnbqkbnrppp1pppp83p43P48PPP1PPPPRNBQKBNR,  GameNumber: 1
```

```
MoveNumber,Side,Move,FEN,GameNumber,MoveNumber1,Side1,Move1,FEN1
1,white,d4,rnbqkbnrpppppppp883P48PPP1PPPPRNBQKBNR,1,2,black,d5,rnbqkbnrppp1pppp83p43P48PPP1PPPPRNBQKBNR
2,black,d5,rnbqkbnrppp1pppp83p43P48PPP1PPPPRNBQKBNR,1,3,white,c4,rnbqkbnrppp1pppp83p42PP48PP2PPPPRNBQKBNF
```

Graph design

The decision on the design of the graph was based on the principle of not storing redundant information. Each entity is stored only once and is connected to the fewest possible entities. In our design, we have used Game as a central node that connects all the other entities, i.e. Player, Position and Event. The rest of the entities are not connected to each other except for the position entity that has a self-referential relationship. The entities contain the following fields:

Event

```
{
  "name": "World Championship 1st",
  "site": "USA",
  "month": 1,
  "year": 1886,
  "day": 11
}
```

Game

```
{
  "eco": "D11",
  "result": "Black",
  "number": 1,
  "month": 1,
  "year": 1886,
  "blackElo": "0",
  "moves": 46,
  "opening": "Queen's Gambit Declined
Slav",
  "whiteElo": "0",
  "halfmoves": 92,
  "day": 11
}
```

Player

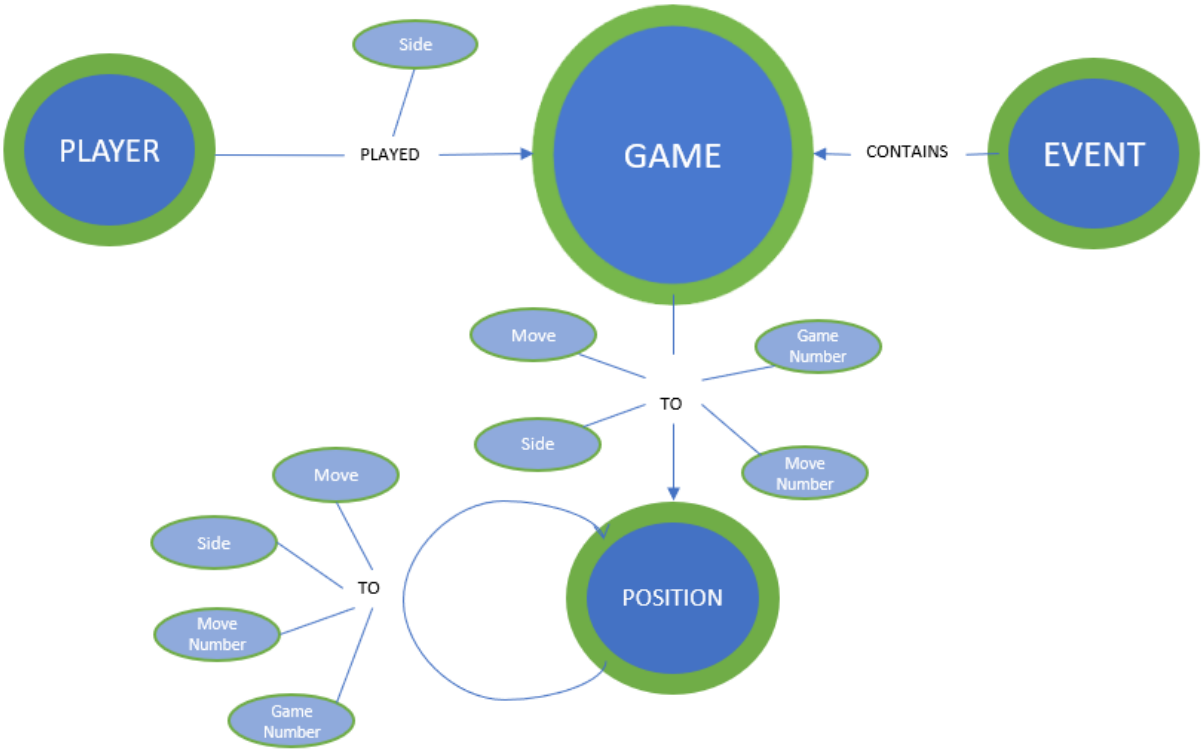
```
{
  "name": "Zukertort  Johannes H"
}
```

Position

```
{
 "fen":
"N4r1kpp4pp3n1p22b1p3P71B3P1P1P3P23R1RK1"
}
```

In more detail a Player is connected to a Game and in the relationship, information about the side/color of this player in this game is stored. For the connection of Game entity to the Position the first position/move of each game (i.e. with MoveNumber=1) is connected to the respective game storing into the relationship information about the move itself, the number of the game, the side (Black, White) that made that move and the number of that move for that particular game. Same information is also stored on the self-referential relationship of a position to another. The Event is connected to the Game, but no additional information is stored.

The script to create the aforementioned graph database schema for the chess data using the two .csv files can be found in the “”create_graph.cy” file.



Queries' results

After creating, the graph and storing the data of the files mentioned above we executed the following queries. The script for the execution of those queries can be found in the file "chess_queries.cy".

1. In how many games the position with FEN: r1bqkbnrpppp1ppp2n51B2p34P35N2PPPP1PPPRNBQK2R is found and in what percentage of these white won?

"total_games"	"white_percentage"
87	0.41379310344827586

2. In the games where the position with FEN: r1bqkbnrpppp1ppp2n51B2p34P35N2PPPP1PPPRNBQK2R is found, in how many the result was draw and in how many white won or black won?

"total_draws"	"total_whites"	"total_blacks"
34	36	17

3. What is the tournament with the most games played and in how many of these Karpov Anatoly had either white or black?

"eventName"	"frequency"
"World Championship 18th"	48
"World Championship 31th"	48

"eventName"	"frequency_KA"
"World Championship 31th"	48

For World Championship 18 no Karpov games were found.

4. Which player has most games with “Ruy Lopez” opening?

"p.name"	"frequency"
"Lasker Emanuel"	17

5. How many games have the moves “Nc6”, “Bb5”, “a6” and which players played these games?

"number_of_games"
52

"Player_name"
"Karpov Anatoly"
"Kasparov Gary"
"Korchnoi Viktor L"
"Spassky Boris V"
"Fischer Robert J"
"Petrosian Tigran V"
"Botvinnik Mikhail M"
"Smyslov Vassily V"
"Reshevsky Samuel H"
"Keres Paul"
"Euwe Max"

"Alekhine Alexander A"
"Bogoljubow Efim D"
"Schlechter Carl"
"Lasker Emanuel"
"Janowski Dawid M"
"Tarrasch Siegbert"
"Chigorin Mikhail I"
"Steinitz Wilhelm"

6. For GameNumber:636 show the game's information, the tournament where it was played, the players and all the moves played ordered.


```
{
  "eco": "C80",
  "result": "White",
  "number": 636,
  "month": 11,
  "year": 1981,
  "blackElo": "2695",
  "moves": 40,
  "opening": "Ruy Lopez, Open",
  "whiteElo": "2700",
  "halfmoves": 81,
  "day": 18
}
```

```
{
  "name": "World Championship 30th",
  "site": "Meran ITA",
  "month": 10,
  "year": 1981
}
```

```
{
  "name": "Korchnoi  Viktor L"
}
```

```
{
  "eco": "C80",
  "result": "White",
  "number": 636,
  "month": 11,
  "year": 1981,
  "blackElo": "2695",
  "moves": 40,
  "opening": "Ruy Lopez, Open",
  "whiteElo": "2700",
  "halfmoves": 81,
  "day": 18
}
```

```
{
  "name": "World Championship 30th",
  "site": "Meran ITA",
  "month": 10,
  "year": 1981
}
```

```
{
  "name": "Karpov  Anatoly"
}
```


As far as the moves of the game are concerned, we demonstrate here the first 16 moves. In total, this game had 81 moves.

r.MoveNumber	r.Move
1	"e4"
2	"e5"
3	"Nf3"
4	"Nc6"
5	"Bb5"
6	"a6"
7	"Ba4"
8	"Nf6"
9	"O-O"
10	"Nxe4"
11	"d4"
12	"b5"
13	"Bb3"
14	"d5"
15	"dxe5"
16	"Be6"

7. Show all the games with the position Fen: “1bqkbnrpppp1ppp2n51B2p34P35N2PPPP1PPPRNBQK2R where the next move was not “a6”. Also show the alternative moves that were played after this position and the result of the games.

   For the purposes of the report we demonstrate the first 16 results. In total, the results are 35.

gNumber	gameResult	altMove
4	"Black"	"Nf6"
6	"White"	"Nf6"
8	"Draw"	"Nf6"
10	"Draw"	"Nf6"
12	"White"	"Nf6"
14	"Draw"	"Nf6"
16	"White"	"Nf6"
18	"White"	"Nf6"
23	"White"	"d6"
39	"Black"	"d6"
58	"Draw"	"Nf6"
60	"White"	"Nf6"
67	"Black"	"d6"
70	"White"	"Nf6"
80	"White"	"d6"
81	"White"	"Nf6"