Project Assignment - Part 1

Roberto Pellungrini, Anna Monreale

October 14, 2021

Introduction

In **Part 1** of the project you are required to create and populate a database starting from .csv files and perform different operations on it. In the following you can find a set of incremental assignments, each one with a brief description of what you are required to produce and what tools you can use for the task.

Build the datawarehouse

tennis.csv contains the main body of data: a fact table with tennis match data. For each match we have information about the tournament, the players involved (winner and loser) and several other metrics.

Files **male_players.csv** and **female_players.csv** contain the list of male players and female players respectively, while **geography.csv** contain a list of IOC codes with country names and continents.

In these four files you will find all the attributes to reproduce the schema shown in 1. The file **tennis.csv** will have to be split appropriately and combined with the other files to achieve this goal.

The goal of the following assignments is to build the schema and deploy it on server lds.di.unipi.it. Beware that, just as in real-life scenario, files may contain missing values and/or slight mistakes.

Assignment 0

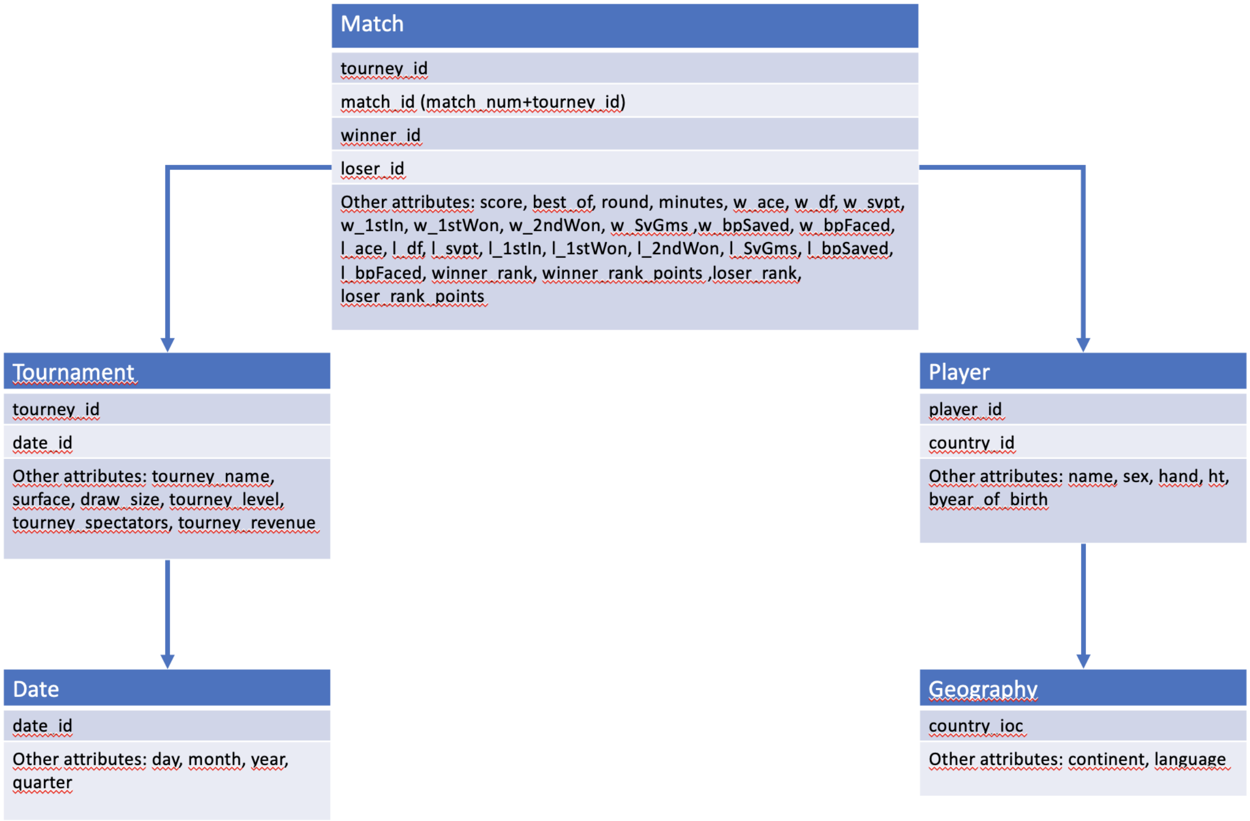
Create the database schema in Figure 1 using SQL Server Management Studio in server lds.di.unipi.it. The name of the database must be *GroupIDHWMart* (example: Group01HWMart).

Assignment 1

Write a python program that splits the content of **tennis.csv** into four separate tables: match, tournament, date and player. Use the files **male_players.csv** and **female_plauyers.csv** to create the attribute ”sex” for the player table. The use of the pandas library is forbidden for this assingment.

Assignment 2

Write a Python program that populates the database *GroupIDHWMart* with the various tables from the .csv files, establishing schema relations as necessary.



Match
tourney_id
match_id (match_num+tourney_id)
winner_id
loser_id
Other attributes: score, best_of, round, minutes, w_ace, w_df, w_svpt, w_1stIn, w_1stWon, w_2ndWon, w_SvGms ,w_bpSaved, w_bpFaced, l_ace, l_df, l_svpt, l_1stIn, l_1stWon, l_2ndWon, l_SvGms, l_bpSaved, l_bpFaced, winner_rank, winner_rank_points ,loser_rank, loser_rank_points

Tournament
tourney_id
date_id
Other attributes: tourney_name, surface, draw_size, tourney_level, tourney_spectators, tourney_revenue

Date
date_id
Other attributes: day, month, year, quarter

Player
player_id
country_id
Other attributes: name, sex, hand, ht, byear_of_birth

Geography
country_ioc
Other attributes: continent, language

Figure 1: Datawarehouse schema of reference.