Fundamentals of Data Science and Engineering

Databases • Practical Assignment • 06-11-2022

Each group will deliver its project inside a **zip** containing, at least, a README file. This file should contain a **brief** description of what was done and the contribution of each group member.

Words highlighted in **yellow** represent the name of the files expected in the delivered zip file.

1) Download Files.

a) **Download** the *all_races.csv* dataset from here. This dataset contains results from running events in Porto, Portugal from 2012 until 2016.

2) Design the Database.

a) **Draw** a UML diagram of a database capable of holding data about runners, events, distances, teams, … (uml.png).
b) **Convert** this diagram into the relational model (relational.txt).
c) **Write** a SQL script that creates the database (races.sql).
d) **Create** the corresponding tables in your PostgreSQL database (use the public schema).

Note: Think about what primary key you can use for each runner. Name is probably not the best candidate (you can have two runners with the same name). Maybe name and birthdate is a better choice. Try adding an artificial ID that can be used to represent a runner.

3) Load Data.

Create a Python script (load_races.py) that:

a) removes all data from the database (using the DELETE command),
b) reads the all_races.csv file, and
c) populates the database with new data (using the INSERT command).

4) User Interaction.

Create a Python script (races.py) that:

a) Prints a menu with several options (for example, search runner, show race, top runner for each distance, …) and that allows the user to select one of these options.
b) Allows the user to interact with the database in a friendly interface. For example, show the progress of a single runner on a single distance throughout the years.
c) Use your imagination so that you create an interesting application.

A small example of what is intended:

```
Menu:

[1] Search Runner
[2] Search Race
[3] …

What's your choice? 1

---

What runner do you want to search: André Restivo

---

Here are the races where André Restivo has run:

[1] …
[2] …
[3] …

To see more details, select the race number, or 0 to go back to the menu:
```

5) Ask Some Questions.

Ask the following questions using SQL (question.sql):

a) Who run the fastest 10K race ever (name, birthdate, time)?
b) What 10K race had the fastest average time (event, event date)?
c) What teams had more than 3 participants in the 2016 *maratona* (team)?
d) What are the 5 runners with more kilometers in total (name, birthdate, kms)?
e) What was the best time improvement in two consecutive *maratona* races (name, birthdate, improvement)?

Extra points: Think of other interesting questions to ask!

6) Extra.

Use other libraries (panda, matplotlib, sci-kit, …) to extract meaningful information from the database. For example, output charts that show the relationship between different variables in the datasets (e.g., age vs time, distance vs time), a single distance or event times, the evolution of a single runner…