University of Basel

Prof. Dr. Heiko Schuldt
David Lengweiler, M. Sc.
Martin Vahlensieck, M. Sc.

Introduction to Databases — Autumn 2023

Exercise 1

(Practical) Hand-in: 06.10.2023 (during Exercise)
(Theory) Hand-in: 08.10.2023 (ADAM, 23:59)

Solving the Exercises: The exercises can be solved in small groups of a maximum of two people. Use the notations introduced in the lecture. The DMI plagiarism guidelines apply for this lecture.

Submission Information: Please upload all deliverables BEFORE the deadline to ADAM using the team hand-in feature. Solutions that are handed in too late cannot be considered. For practical exercises upload the deliverables to ADAM and present them to one of the assistants/tutors during the exercises, both is required to recieve the points!

Task (Practical) 1: Setup Process (2 points)

To solve the practical exercises, you need access to a running DBMS. In this lecture we use PostgreSQL, a popular open-source DBMS. To run PostgreSQL, we encourage you to use Docker[1]. Docker is an container framework, which allows you to easily start virtual systems with only limited setup necessary. Follow the instructions on `https://www.docker.com` (Mac users make sure you select the proper chip), then open a terminal and run the following command:

```
docker run --name introdb-postgres -e POSTGRES_USER=demo \
      -e POSTGRES_PASSWORD=demo -p 127.0.0.1:54321:5432 -d postgres
```

Now you need a way to interact with the database. Here you are free to choose. One way is the open-source community edition of dbeaver. Other options are JetBrains DataGrip, or pgsql if you prefer a command line tool. To verify that you are correctly connected, run the following SQL command:

```
SELECT version();
```

Hand-In: Show the working database to one of the assistants/tutors.

[1] You are of course free to use any other method as well, but expect less support in these cases.

Task (Practical) 2: Interaction (5 points)

We will now perform a little toy analysis. This consists of loading data into the database and executing a few queries. For now you can use the same tool as you have used in the previous task, in later exercises we will learn on how to run these queries from within a Python program.

The data consists of a subset of the games played on `lichess.org` in the year 2016. We only look at the standard games with a "normal" termination and where either white or black wins (no draws).

The data is provided as a SQL file and can be download from the following source: `https://drive.switch.ch/index.php/s/rspmfLq0g5xWzfv/download`. This file can be run, just as any other SQL scripts and recreates the data in your database. Be careful: The DBMS differ enough that the produced SQL file is generally not interchangeable between different DBMSs.

To insert the SQL file into the database, you will need to use the command line. Open a terminal in the directory with the file and execute the following command (if you are using Windows, use Command Prompt):

```
docker exec -i introdb-postgres psql -U demo < ex1.sql
```

Now have a look at the data. You should see a table named `games_simple` with four columns: `game_id`, `game_month`, `game_day` and `game_white_wins`. `game_id` is the eight character id of the game. `game_month` and `game_day` are are the month and day on which the game started. `game_white_wins` is true if white wins and false if black wins (remember, this dataset does not contain draws).

You can use the `game_id` to see the game on the website. Just enter `lichess.org/ID` into your browser and replace ID with the eight characters. Check that the other values in the row are actually correct. How many games are in the database?

Now we want to know, if there is a correlation between the month of the year and the number of times white wins. To do that, we need to aggregate the wins of black and the wins of white by the month the game took place. The following query accomplishes that:

```
SELECT game_month,
       SUM(game_white_wins::integer) AS white_wins,
       SUM((NOT game_white_wins)::integer) AS black_wins
  FROM games_simple
GROUP BY game_month
```

Now copy the data over to a plotting program of your choice and plot it.

Hand-In: Show the total amount of games, that you can execute the SQL statement and the plot to an assistant.

Task (Theory) 3: Questions (3 points)

Answer the following questions and explain your answers:

- What are the advantages and disadvantages of using databases to manage data compared to file-based approaches? *(2 advantages and 2 disadvantages as bullet points)*
- What is the link to the official documentation for the `version()` function used in the first task and to which category of functions does it belong?