Advanced Data Management for Data Analysis

Assignment 4 *(individually!)*

18.11.2022

Due: ***Monday, 28 November 2022, 09:00 CET***

Notes:

- Using Google colab [4] (as with the DuckDB hands-on during class), from the given Python notebook template [6], produce a Python notebook (called "**code.ipynb**") that contains your solutions to all questions.
- Accompany the notebook with a PDF document (called "**report.pdf**"; preferably max. two A4 pages) that describes your solutions, lists the execution times for each question for all three systems (Pandas, SQLlite, DuckDB), and discusses the (potential) reasons for the observed differences in execution time between Pandas, SQLite & DuckDB for each question to the best of your knowledge.
- Create a (compressed) archive containing your "**code.ipynb**" and "**report.pdf"**.
- Name your submission file "**ADM2022_A4_<student_id>.zip**"
- Submit via BrightSpace

Points:

This assignment is worth a total of 100 points as specified in the Python notebook [6]: Data Loading: 5 points, Query 1: 10 points, Query 2: 15 points, Machine Learning: 40 points, PDF document: 30 points. The final score (grade) will be points divided by 10 to fit in the 0-10 grade system.

Please read the entire assignment instructions carefully and make sure you follow them properly.

In this assignment, you will experiment with the NYC Cab dataset from 2016 [5]. This dataset provides information (e.g., pickup/dropoff time, # of passengers, trip distance, fare) about cab trips done in New York City during 2016.
You will load this dataset into Pandas [1], SQLite [2], and DuckDB [3]. You will compare the performance of multiple data-science like queries, including performing a fare estimation (i.e., predicting how much a ride will cost depending on distance) using machine learning.
In the first section you will implement the loader in DuckDB [5 points].
The second section has two data-science like queries, the implementation in pandas is already given, and you should use it as a logical/correctness reference to write the queries for SQLite and DuckDB, remember to compare the performance of the three different systems [25 points].
Finally, in the third section you will implement a simple machine learning algorithm in DuckDB to predict fare costs. A full implementation of Pandas is given and a partial of SQLite. Again, use them as a logical/correctness reference and compare the performance of the three different systems. [40 points]
Remember to submit your notebook with the answers to all sections as well as a PDF document (max two pages) listing all experienced execution times and reasoning about the performance difference in these systems.

Given a Python notebook template [6], your task is to fill-in the missing code (mostly SQL queries) to do the requested tasks / answer the given questions. The code template already contains code to measure the execution times. Please collect the execution times you experience into a PDF documents and discuss the (potential) reasons for the observed differences in execution time between Pandas, SQLite & DuckDB for each question to the best of your knowledge.

You start with downloading the Python notebook template [6] to your computer, e.g., by clicking on “Raw” to get the raw text/code in your browser --- or directly go to [7] --- and then right-click and choose, say, “Save page as ...” --- on Unix-like systems you can also simply run

wget https://raw.githubusercontent.com/pdet/duckdb-tutorial/master/Project/NYC_Cab_DuckDB_Assignment.ipynb

in a shell ...
Then you login to Google colab [4] with your google account and upload (“File” → “Upload notebook” → “Upload”) the just saved Python notebook as Python 3 notebook.
Alternatively, instead of first downloading the notebook from github to your computer, you can also directly load it into Google colab: “File” → “Upload notbook” → “GitHub” and then specify the URL given as [6] below.

After that, you can complete the code as asked and run each part as you did during class with the first DuckDB hands-on. While working, recall to save your changes regularly (“File” → “Save” or Crtl-S), though colab should also do that automatically.

Once done, download your modified Python notebook for submission (“File” → “Download” → “Download .ipynb”) and (re)name it “**code.ipynb**”.

[1] https://pandas.pydata.org/
[2] https://www.sqlite.org/index.html
[3] https://duckdb.org/
[4] https://colab.research.google.com/
[5] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
[6] https://github.com/pdet/duckdb-tutorial/blob/master/Project/NYC_Cab_DuckDB_Assignment.ipynb
[7] https://raw.githubusercontent.com/pdet/duckdb-tutorial/master/Project/NYC_Cab_DuckDB_Assignment.ipynb