Hands-On: Queries with SQL, pandas, Unix commands

1. Retrieve the three CSV files for this practical and study their content; they provide information, respectively, on country codes, country population, and Covid cases and deaths per country (a country being identified by its ISO 3166-1 alpha-3 code).
2. We will be interested in the following four queries on this dataset:
    - Q1. What are the total number of Covid deaths?
    - Q2. What are the ten countries with the highest number of deaths, with the corresponding number?
    - Q3. What are the ten countries with the highest number of deaths relative to their population, with the corresponding number in ‰?
    - Q4. What are the ten countries with the highest number of deaths relative to their population within Europe, with the corresponding number in ‰?
3. Use `sqlite3` to load this dataset into a new database (using `.open`, `.mode csv` and then `.import`) and write SQL queries to answer these queries. Note the time needed for each query (using `.timer on`).
4. Add indexes (using `CREATE INDEX`) on relevant columns. Rerun the queries and check the timings.
5. Use `pandas` to load the dataset in main memory and run the same queries. Using Python's standard `time` module, note the time taken by each query.
6. Use Unix commands to process the dataset on disk and run the same queries. Use the `time` command as a prefix to your other commands to note the time taken by each query.

Modifié le: vendredi 13 octobre 2023, 10:13