UNIVERSITY OF CYPRUS

Computer Science Department

EPL 646 – Advanced Topics in Databases

Assignment 2 – Storage and DBs (Compare Querying Large Temporal Dataset in Pandas/Parquet, SQLite and InfluxDB)

Assignment Date: Friday, 07/02/2025
Due Date: Friday, 21/02/2025 (14 days)

Instructor: Demetris Zeinalipour

(Submit as a zip file through **Moodle**)

I. Exercise Goal

The aim of the exercise is to familiarize yourself with big data processing. In particular, you will be asked to use (in order to compare and bencmark) *Pandas/Parquet*, *SQLite* and *InfluxDB* to formulate queries for each of the questions listed below.

II. Preparation and groundwork

Download the CSV from https://github.com/Schlumberger/hackathon/blob/master/backend/dataset/data-large.csv to use as your dataset. As a first step you have to import the downloaded dataset using Python to i) Parquet files, ii) SQLite and iii) InfluxDB and compare the time needed for each case. Afterwords you have to benchmark and compare each case by running the following queries:

a) Retrieve all measurements
b) Retrieve all measurements for a given hour-long period (select a random period from the dataset)
c) Average the values of one measurement (one column) for an hour-long period
d) Average the values of one five measurements (five columns) for a day
e) Average the values of all measurements (all columns) for all time

For initial testing of your code and queries you can use the CSV from https://github.com/Schlumberger/hackathon/blob/master/backend/dataset/data-small.csv

You might find the following tutorials useful:

- Pandas with Parquet:
  - https://sql2pandas.pythonanywhere.com/articles/sql-aggregations-in-pandas
    - https://sql2pandas.pythonanywhere.com/cookbook/sql-avg-in-pandas
  - https://pandas.pydata.org/docs/reference/api/pandas.read_parquet.html
  - https://www.opendatablend.io/blog/querying-large-parquet-files-with-pandas/
- SQLite and Python:

    - https://www.sqlitetutorial.net/sqlite-python/sqlite-python-select/
- InfluxDB and Flux
    - https://docs.influxdata.com/influxdb/cloud/api-guide/client-libraries/python/
    - https://docs.influxdata.com/influxdb/cloud/query-data/flux/

You might find the following tools useful:

- Unix timestamp converter: https://www.epochconverter.com/
- SQLite Browser: https://sqlitebrowser.org/dl/

III. Exercise Requirements

Write the Python code needed to import and query your dataset for each mentioned case while timing each operation. Produce a report comparing the timing of each operation (do not forget the operation of data importation). In your report you must have graphs to show your results as well as a summary discussing the performance of each of the three cases (Pandas with Parquet files, SQLite and InfluxDB).

IV. Deliverables

Deliver **all** your source files (python source code and report document – in .docx or .pdf format) through Moodle in 1 compressed file (as1.zip):

Good luck!