Date: 23rd of May, 2025

The deadline to upload your solution to iCorsi is on the 29th of May, at 23:59. Late submission policy will apply; see the Intro slides for details. Upload a zip/compressed file with the Jupyter Notebook `hw07.ipynb` containing your solutions. Name the compressed file using the "HW7_FirstName_LastName" convention.

Description

Make suitable assumptions where necessary and state them explicitly in your answers.

In this assignment, you will familiarize yourself with Spark - the distributed computing framework. You will develop a python-based Spark application to execute queries on the TPCx-BB bigdata benchmark. Python will be used as an entrypoint to Spark Core and SparkSQL API. Hence the Spark code necessary to execute queries will be contained within a Python application. Queries executed inside Spark will be expressed using the SparkSQL API. Your task is to translate some queries in the TPCx-BB benchmark from SQL to SparkSQL using the guidance and templates provided. The queries must be developed and executed inside the jupyter lab environment provided in the starter kit.

The starter kit is available on iCorsi and on GitHub. You will need Jupyter Lab to run this exercise:

- Make sure you have python higher than v3.9 installed.
- Make sure you have exactly Java v17.
- Install Jupyter Lab with: `pip install jupyterlab`. If you have another operating system than Linux, follow the official Jupyter Lab website for installation instructions.
- Download the starter kit.
- Navigate to the starter kit folder and start Jupyter. Linux command: `cd dm-spark-tpcxbb && jupyter lab`

Implementation details and additional information are given in the `hw07.ipynb` Jupyter notebook file.

Tasks:

- Translate queries `07, 09, 20`. Execute queries on the given data to obtain final results which should be close enough to the results present in the corresponding folder (more info in the exercise intro).
- *Optional 1*: Translate 3 additional queries of your choice among the ones available in the `queries` folder.

Additional notes

- Make sure every query is in its own separate cell and executes when running all the cells above it.
- Solving the non-optional task will get you full points for this exercise. Solving the optional tasks will get you "extra" points: an additional delta on top of the overall homework grade.