Instructions for Exercise 2: Data-Driven Computing Architectures

Deadline: Thursday 06.02.2025 23:59

Submission: Submissions can be made individually or in groups of two. If you are submitting as a group, remember to register your group in Moodle.

In this exercise, we will use Spark and Apache Hive to perform ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) tasks on the same datasets, provided in different formats. For the Data Warehouse approach, you will define schemas, transform the data, and load it into structured tables (schema-on-write). For the Data Lake approach, you will load raw data first and apply transformations during analysis (schema-on-read).

- **Access the Environment:**
  - Open the Noppe PySpark Environment. (https://noppe.2.rahtiapp.fi/main/catalog)
  - Login to the environment using your Haka credentials.
  - Click "Join workspace" and use the joining code: **dat-vke5xyaw**.
  - After joining, go to "My Workspaces" and enter the PySpark environment. This will direct you to the Jupyter notebook interface.

- **Copy the Notebook:**
  - Locate the notebook titled **Exercise2** in the **shared** folder.
  - Copy the notebook to your personal workspace in Noppe, in the **my-work folder.**
  - Open the copied notebook from your workspace.

- **Complete the Notebook:**
  - Follow the tasks outlined in the notebook.
  - Create two pipelines: ETL and ELT and answer the questions.
  - We will be using datasets located in the **shared** folder.

- **Export Your Completed Notebook:**
  - Once you have completed all tasks and questions, export your notebook as a **.pdf file**
  - Save the file with your name included in the filename (e.g., Exercise2_YourNameOrGroup.pdf) and submit it in Moodle.

- **Additional Notes:**
  - Double-check that all cells in your notebook have been executed, and the outputs are visible.
  - If you encounter issues with the notebook or dataset, contact the teaching team for assistance.

Good luck!