Instructions for Exercise 3: Data-Driven Computing Architectures

Deadline: Thursday 20.02.2025 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.

- **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 **Exercise3** 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:**
  - Build a medallion architecture pipeline using Delta Lake to clean, structure, and analyze the provided datasets.
  - Include comments where you describe what you have done and why
  - Follow the tasks outlined in the notebook.
  - 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., Exercise3_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!