Instructions for Exercise 1: Data-Driven Computing Architectures

Deadline: Thursday 23.01.2024 23:59

Submission: Individual submissions only. Each student must submit their own completed notebook.

1. **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.

2. **Copy the Notebook:**
   - Locate the notebook titled **Exercise1** in the shared folder.
   - Copy the notebook to your own workspace in Noppe.
   - Open the copied notebook in your personal workspace.

3. **Dataset Information:**
   - We will be using a dataset called **air_quality_data.csv** in this exercise.
   - The dataset is stored in the shared directory. Do not move or modify this dataset; it must remain in the shared directory.

4. **Complete the Notebook Exercises:**
   - Open your copied notebook.
   - Follow the instructions and code examples provided in the notebook. This will include:
     - Configuring Spark with Delta Lake.
     - Performing operations on Delta tables, such as updates, deletes, and appends.
     - Exploring Delta features like time travel and vacuuming.
   - Answer the questions embedded in the notebook. These are designed to test your understanding of the operations.

5. **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., Exercise1_YourName.pdf) and submit it in Moodle.

6. **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 with the exercise!