

# INFO-H420 Management of Data Science and Business Workflows

## Assignment 4: Workflows with Apache Airflow (15 points)

The goal of this assignment is to write a simple workflow that analyzes a web server log file, extracts the required lines and fields, performs some transformations, and writes to another file. The log file is available on UV.

### Exercise 1 (2 points)

Install Docker on your system. Install Airflow via Docker. Describe in the report how you achieved this.

### Exercise 2 (10 points)

- Define the DAG. Create a DAG named “process\_web\_log” that runs daily.
- Create a task to scan for a log. Create a task named “scan\_for\_log” that scans a folder “the\_logs” for a “log.txt” file and triggers the rest of the workflow.
- Create a task to extract data. Create a task named “extract\_data”. This task should extract the ipaddress field from the web server log file and save it into a file named “extracted\_data.txt”.
- Create a task to transform data. Create a task named “transform\_data”. This task should filter out all the occurrences of ipaddress 198.46.149.143 from “extracted\_data.txt” and save the output to a file named “transformed\_data.txt”.
- Create a task to load the data. Create a task named “load\_data”. This task should archive the file “transformed\_data.txt” into a tar file named “weblog.tar”.
- Define the workflow that execute the aforementioned tasks in sequence.

Save the DAG you defined into a file named “process\_web\_log.py”. In the report, include code snippets describing how you achieved this.

### Exercise 3 (3 points)

Do a test run for each of the tasks you defined. Once all works as expected, do a test run for the workflow. Finally, trigger/run the workflow and monitor a few runs. In the report, document the test runs, and include any findings or observations that you may have from the runs.

## Instructions

The assignment contributes 15% to the overall grade.

This assignment is to be made in **groups** of two persons. You are asked to form the groups via "Groups for Assignment 4" on the Université Virtuelle (UV). If you cannot find a partner, please post a request in the "Discussion Forum" on UV.

You are asked to submit a short **report** presenting your solution to the exercises, including justifications for the choices and assumptions made.

The report has to be uploaded as a pdf file to "Assignment 4" on UV before January 9, 2023.