Data Engineering Slot 2: Transform and Load

For the first slot your task consisted of implementing a tool that will allow us to automatically retrieve information from the job web-sites.

You will now have a repository of HTML pages that contain the text of job offers.

Slot 2 of the data engineering track focuses on transforming and loading (as in ETL) this data into a database.

Transformation Step:

For the transformation step we have to achieve two things:

1. **Identify extraction targets in the underlying data**: In order to be able to later on successfully run queries that allow us to gain some insights on the Swiss job market we will have to identify individual data points and extract those from the job descriptions.
    - A starting data point would be the 'job title'.
    - What you extract is up to you. You should however make sure to extract 4-5 different data points
    - The more data point the better, as it will offer more possibilities for data analysis for the task in slot3.
2. **Extract targets:** For the actual extraction you can use tooling such as beautifulSoup (see for example here for a tutorial https://realpython.com/beautiful-soup-web-scraper-python/#step-3-parse-html-code-with-beautiful-soup), but other tools such as jq (https://pypi.org/project/jq/) are also okay.

At the end of the transformation step you should have extracted a set of data points in a form that is easily loadable into a database such as duckdb.

Loading Step:

The target database for this part is duckdb.
The part consists of the following steps:

1. Create table structure in duckdb to hold your extractions
2. Load a sizable amount of data (based on at least 1000 job descriptions, the more the better) into duckdb.

Deliverables:

The deliverables for this step consist of

- Your code for the transformation as a Jupyter Notebook which displays the results of your last execution including the and loading step into the database.
- Explain your thoughts in Markdown cells in the notebook.
- A short demonstration of the contents of your database to show in detail, how your database looks like (e.g. screenshots in code review style or short screencast (mp4) < 5 min with spoken comments, or other suited forms).
- Upload to your repository on BFH Gitlab and provide clickable links to these resources in Moodle.

If you have questions or require support please use the following link to book appointments:

https://outlook.office365.com/book/BFHBookingsbsc_data_engineering@bfh.ch/?ae=true