Conduct a Data Science project

Topic: the exact topic can be chosen by yourself

Procedure and general conditions

Teamwork

Each project is worked on by a **team of 3 students**. you will be evaluated as a group (of course there are exceptions for exceptional cases).

Big Picture

This project is a continuation of the topics from the previous Data Science courses with special focus on Big Data Engineering. It is allowed to draw on results from the previous courses and incorporate them here if appropriate.

In this project, you will learn how to independently carry out data science projects with a focus on data engineering. It is not mandatory to use exactly the tools used in the course, unless they are explicitly required. Most of the time there are many ways to reach the goal. It is your task to find a suitable way and to organize yourself and the necessary infrastructure.

1. Find a topic you are interested in (might be the same from previous courses)
2. Get familiar with the provided infrastructure
3. Obtain data for your project that can be processed
4. Analyse your data ((analysis can be very simple from the algorithmic side, here it is more about the data engineering setup)
5. Present and visualize your results

MUST HAVE criteria for the project:

- **At least 3 different data sources** of different data:
  - 1 from a file (csv, json, parquet, ...) or a database (RDBMS, NoSQL).
  - 1 obtained by web scraping.
  - 1 obtained by using a REST API.
- Use **Kafka** to make the data (for at least one of the data sources) available on a broker. Write Kafka producers that push data to one or more Kafka topics. It is your job to organize the data in a way so that Kafka can be used as central data broker for the data you have.
- Use **Spark** to read data from a Kafka topic and process the data.
- **Store** your analyzed and transformed data or results (some kind of ETL/ELT) to a flat file or database of your choice and preserve them for later use. It is your choice to determine which data is stored.
- **Show** your results, tell a „story". There are a large number of examples on Kaggle. Storytelling is not a main aspect of the project but helps for a coherent, easy to understand presentation. It's your task to find a story behind your data that can be presented.
- **Visualize** the data flow of your project. List all the data sources, the transformations and the results.
- **Document** each step in a Jupyter notebook (even if not all steps need to be performed in a notebook).

Hint: in most cases, the project will not actually fall into the Big Data category due to the relatively small amount of data that will be used, but it is still necessary to apply similar procedures.

Delivering results

For a Data Science project it is important to intensively engage in the topic. Both topic and project goal are not clearly defined in the final project; give full scope to your imagination and make something "vivid" out of the data. Primarily, however, the final project is about various technologies, only secondarily about the "story that is told".

To implement the individual steps, you have to:

- Conduct supplementary research on the topic.
- Establish the technological bases and understand their functioning (through manual study in addition to the course).
- Implement the self-imposed task as well as possible
- In a final presentation, present the results, the chosen paths and methods to the class.
- In addition, create a documentation in form of Jupyter Notebooks.
- Create a public GitHub repository to share your notebooks.

Last modified: Monday, 13 January 2025, 4:38 PM