Project spezification to the course „Big Data Engineering“

Conduct a Data Science project on major contents of the course.

Topic: the exact topic can be chosen by yourself

Procedure & general conditions

1. Teamwork

Each project is worked on **by 2 students** (teamwork)

You will be evaluated as a group.

2. Big Picture

This project is the direct continuation of the project from the course “Big Data Infrastructure” with special focus on the newly added topics. It is allowed to draw on results from “Big Data Infrastructure” and incorporate them here as appropriate.

In this project, you will learn how to independently carry out Data Science projects with a focus on Data Engineering issues. 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.

Step 1: find a topic you are interested in (might be the same as for “Big Data Infrastructure”)

Step 2: obtain data for this purpose, which must be processed afterwards (new criteria compared to “Big Data Infrastructure”)

Step 3: analyze your data (analysis can be very simple from the algorithmic side, here it is more about the data engineering setup)

Step 4: present / visualize your results (simple visualizations are ok here)

MUST criteria for the project:

a) At least 3 different data sources of different data types
   - one (or more) data sources MUST be obtained by use of WebScraping
   - one more CAN be obtained by using REST APIs
   - another one CAN be data from a file (csv, parquet, avro, hdf5, …) or from a database (RDBMS, NoSQL)

b) Use Kafka to collect all data you have. 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 a central data broker for the data you have.

c) Use Spark to read data from Kafka and analyse the data

d) Store your analyzed/transformed data/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.
e) The use of Nifi is optional (you can use it, nevertheless the other points have to be fulfilled)
f) Show your results, tell a „story“. There are a large number of examples in Kaggle, such as: https://www.kaggle.com/parulpandey/geek-girls-rising-myth-or-reality
Hint: 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.
g) Somehow visualize the architecture of your system (data workflow, components you use, …) in a simple diagram
h) Document each step in a Jupyter notebook (even if not all steps need to be performed in a notebook).
i) You can (but don't have to) use Docker and / or Git.

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.

3. Delivering results

For a DataScience 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 group. (approx. 25 minutes per topic)
- In addition, create HOW-TOs (=documentation) in form of Jupyter Notebooks.

Milestones:

- class 4: Submit your topic (short talk in the unit)
  - topic (title)
  - members (team)
  - planned data sources
  - planned data storage
  - planned procedure
  - expected output

- class 6: intermediate delivery:
  - brief discussion during the attendance phase on the status of your project
- class 7: ask your questions (if you have any)
- class 8: final delivery:
  - all documents in Moodle
  - presentation of the results (in a team, 25 min)

4. Assessment

Depending on the selected topic, the individual parts can be prioritized differently. Nevertheless, parts from each area must be visible in some form. The points are then assigned according to the following key:

Part	what can be included (examples)	points

Data Source	• data identified, documented (what data do you have, how is it structured and organized)
	• make data available
	• describe your data, which metadata do exist
	• use at least 3 different data sources
	• examples:
	  ○ use WebScraping (mandatory)
	  ○ use ready datasets (e.g. Open Data Austria, Kaggle)
	  ○ use data from REST-APIs (e.g. OpenWeatherMap)
	  ○ …
		5

Kafka	• use one or more Kafka topics • use one or more producers to push data to Kafka	5
Spark	• use Spark to read data from Kafka (Kafka consumer, Spark streaming)	5

Spark data analysis	• use Spark for analyzing your data
	• use a mixture of your choice:
	  ○ Spark RDDs
	  ○ Spark dataframes
	  ○ SparkSQL
	  ○ Windowing with streaming data
	  ○ …
		5

ETL/ELT/data output	• store data to a flat file (parque, avro, HDF5, ..,) ○ CSV doesn't count (because it is too simple to use) • or to a database of your choice (RDBMS, NoSQL) • the data to be stored should be (somehow) analyzed (i.e. not the original data)	5
Visualization	• presentation of results in any form • should contain at least simple forms of diagrams	5

	• tell a story with your project	
Documentation	• in the form of a Jupyter notebook	5
architecture, **p**rovision and use of infrastructure (Git, Docker, …)	• show the architecture of your project in form of a diagram (mandatory) • organize your project by providing the necessary infrastructure • must be visible (documented) in some form	5
Quality in general	• overall impression of the project • how do the individual points interlock (do they give the impression of an overall project or are they rather independent partial solutions?) • everything that doesn't fit to above points	10
Presentation	• presentation, talk, adherence to deadlines, … • everything that doesn't fit to above points	5
	sum:	55