Project spezification to the course „Big Data Infrastructure“

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 3 students** (teamwork)

You will be evaluated as a group.

2. Big Picture

In this project, you will learn how to independently carry out Data Science projects with a focus on methods as applied to Big Data. Infrastructural issues in particular will be considered. 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

Step 2: obtain data for this purpose, which must be processed afterwards (ready-made data sets, access to data via APIs, ….)

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

Step 4: present / visualize your results (no complex visualizations necessary, because this is not part of the course. Simple visualizations are necessary to tell "your story", however)

MUST criteria for the project:

a) At least 1 data source, preferably 2 or more, which must be connected in some way.

b) Store and/or read and/or process the data using a database (must contain some form of NoSQL aspect, ...). The type of database will depend on the type of data. Argue your choice.

c) Use Big Data technologies to process the data. In our case, the use of at least one MapReduce calculation is mandatory.

d) Consider "your project" based on the Big Data criteria (5 Vs – like time-critical aspects, where must special attention be paid to data volume, ...) and the 4 levels of data processing (data source, data storage, …). Even if your project will probably not be Big Data relevant in practice, it is important to consider these points theoretically. The exercise example with the Connected Car is a starting point for this.

e) 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.

f) Document each step in a Jupyter notebook (even if not all steps need to be performed in a notebook).

g) Save (intermediate) results (at least the notebook) in a Git repository.

h) You can (but don't have to) use Docker to provision infrastructure.

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 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 • examples: ○ use ready datasets (e.g. Open Data Austria, Kaggle) ○ use data from Web-APIs (e.g. OpenWeatherMap) ○ …	5
Data Storage	• use one or more databases (RDBMS and / or NoSQL) • the use of a NoSQL aspect is mandatory • communicate withe the DB (Import / Export / Python Scripts) • provide a suitable interface • Exploitation of specific properties of the database used	5
Analysis of Big Data criteria	• Similar to the analysis of the Connected Car video according to Big Data criteria. • Consider your project according to the Big Data Vs (Volume, Velocity, Variety, Veracity, Value). Argue for each point the implications to your project idea. • Consider your project according to the 4 Levels of Data Handling in Data Science (Data Source, Data Storage, Data Analysis, Data output). Argue for each point the implications to your project idea.	5
MapReduce	• there should be a calculation according to the MapReduce algorithm in any form • generally design your calculations so that they can be performed even with large amounts of data • example: multiplication of a matrix with a vector, so that data can also be distributed over several nodes.	5
Visualization	• presentation of results in any form • should contain a least simple forms of diagrams • tell a story with your project	5

Documentation	- in the form of a Jupyter notebook	5
Provision and use of infrastructure (Git, Docker, …)	- must be visible (documented) in some form - Git is mandatory	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:	50