Project specification to the course „Big Data Infrastructure“

Main focus is to provide the infrastructure and execute a simple data science project, according to the main contents of the course.

Topic: the exact topic can be chosen by yourself

Procedure & general conditions

1. Teamwork

Each project is worked on **by up to 4 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

Step 4: present / visualize your results

Note: the upcoming course "Big Data Engineering" will take a closer look at steps 1 to 4. In this project analysis can be very simple from the algorithmic side, no complex visualizations are necessary, because this is not part of the course, it is more about the infrastructural setup. Nevertheless, simple data processing and visualizations are necessary to tell "your story", however.

MUST criteria of the project:

a) At least 2 data sources, preferably more, which must be connected in some way. Ready-made datasets (e.g. csv-datasets) or access to data via APIs (e.g. REST-API) is fine.
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. Give arguments for 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 presented on the Big Data slides (5 Vs, 4 levels of data processing). Even if your project will probably not be Big Data relevant in practice, it is important to consider these points theoretically. The exercise 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) Organize yourself by means of a working infrastructure and document your infrastructure:
    a. Show the architecture in a diagram.
    b. Consider your setup in terms of BigData criteria.
    c. You work in a team on the project, therefore the setup must be multiuser capable. This means that all team members must have access to code, data and infrastructure in general.
        i. use Git for sharing (intermediate) results – at least the notebook must be available on Git.
        ii. your NoSQL database must be available for all persons working on the project.
        iii. 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. 20 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 AND Git
  - presentation of the results (in a team, 20 min)

4. Assessment

The following list gives an impression of the grading criteria and the points you can achieve:

Part	Description	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 does 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, NoSQL) • The use of a NoSQL aspect is mandatory • Communicate with the DB (Import / Export / Python Scripts) • Exploitation of specific properties of the database used	7
Data Analysis MapReduce	• At least simple data analysis should be visible (e.g. with Pandas) • In addition, 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	8
Visualization	• Present the results in form if at least simple diagrams • Tell a story with your project	5
Big Data criteria	• Describe your setup in terms of Big Data criteria.	5

	○ 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. • Should be similar to the exercise with the analysis of the Connected Car video.	
Documentation Architecture	• Documentation in the form of a Jupyter notebook (code inline comments, markdown cells, …) • Describe the architecture of your setup / infrasctructure (graphical representation, diagram) • Name the components and versions used (e.g. Python 3.9, Pandas 1.5.3, …) • Project setup must be multiuser capable (all users must have access to all parts and data of the project) • Git is mandatory and must be visible (documented)	10
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