Dear students,

Here follows **11 guidelines** for the project:

Guideline 1: Relevant Big Data Problem

According to the project assignment, your group should be able to identify a relevant Big Data problem. You will be evaluated by the shown creativity in the selected dataset and problem approach. This means, that if you select the iris dataset (for example) this would hardly scale in a true real-world environment. It is hardly considered a big data problem.

What is a big data problem? Theoretically and, by definition, a **big data problem** refers to a task that involves **data sets either large, sophisticated, or in real-time** that need to be treated in distributed environments. The most extreme cases include for example:

1. **Volume** – Large amounts of real-time data
   Example: Processing Twitter posts in real-time.
2. **Velocity** – High speed of data processing
   Example: Processing of stock market data in milliseconds.
3. **Variety** – Different formats (text, images, video, sensors, etc.)
   Example: Combining hospital EHRs, MRI scans, and voice notes.

Despite these being real-world scenarios, we don't expect your group to process stock market data in milliseconds or to process twitter posts in the form of terabytes. You are required to find a problem that is a big data problem and tackle it in small scale. This means to use Databricks Community Spark to process a subset or sampled set of data in a distributed manner (the code will not run distributed but will be prepared to do so).

Even if your dataset is constrained in terms of rows either by definition (the dataset is as it is) or by sampling (due to the limitations of databricks), please discuss how it still represents a big data problem. This will be evaluated across the report but especially in **problem definition**.

Guideline 2: Number of Rows and Columns?

The number of rows or columns in your dataset does not matter that much within reasonable limits (obviously 1000 rows would not be great). Do not to aim at a specific number, and more for complexity / general interest.

Guideline 3: Use of Numpy and Pandas

You can use numpy and pandas solely for visualization and presentation purposes and it should be shown inside Databricks. Refrain as much as possible from using Spark outside the Databricks environment. The idea is to give you some experience on how to

to handle Databricks. If you do use outside code, please find a very special reason to do it.

Guideline 4: Use of PySpark

Avoid the use of sklearn, pytorch, tensorflow. Sklearn and others are not distributed platforms. PySpark is preferred here and the one aligned with what we teach in classes. The purpose of this is that you understand the utilities of pyspark, MLLib (distributed machine learning library), Spark SQL (in the labs and allowed in the project), and get a theoretical sense of streaming and GraphX. Understand please what is vectorassembler and what is does in a distributed environment. If you use clustering, investigate how it works underneath (for the report it can be quite useful).

Guideline 5: What to do as modelling option(s)

The clustering - regression idea was an idea and proposal, not an enforcement. We expect you to be creative. Working ethics and effort is great but we would love to see the *Spark* in your work. I understand we are limited in terms of code, libraries, and platform but that can be seen as an opportunity to do our best within the limitations. Some of your colleagues have shown amazing ideas so far which is great.

Guideline 6: Use of ChatGPT

Avoid the use of ChatGPT, especially in the report. It is not interesting to read revolving text that does not say much about your work. You are free to use it in a smart way nevertheless.

Guideline 7: Appendixes and Plots

You can have appendixes but avoid tedious repetitive plots. It is not about how many plots you can do but more about the algorithms and the insights that result in the plots.

Guideline 8: Bonuses

2 extra values will be rewarded to outstanding writing performance to a single group (if that makes sense). NO extra values are expected for Spark Streaming of Spark GraphX. Unfortunately, it is not possible to implement these solutions in Community Edition. A simple solution of streaming that is just about copying or mirroring the code provided in the labs will be considered but not to the full extension of the streaming extra points.

Guideline 9: Oral Discussion

The project presentation referred in the report is replaced by a short 15 min discussion due to the large number of groups and the limited slot time. Please use your time effectively. Bring the code functioning and as (only) a last resort the printed code. We expect all group members to participate. If a group member shows little to no knowledge of the project, penalties apply.

Guideline 10: Oral Discussion Slot Allocation

The discussions will be held in the last 2 weeks of classes (13 - FIRST and 14th - SECOND weeks). The groups were allocated taking into consideration the members in the group and their practical allocations. Two teachers will be present at each discussion. We expect that recommendations for improvement can be reflected in the final report / code. We are here also to help, as usual.

Guideline 10: Evaluation Criteria

The grading criteria for the project is the following:

- Problem Definition - 10 points
- Data Preprocessing (Big Data & Model) - 45 points
- Analysis & Insights (Results) - 20 points
- Visualization & Presentation of Results (Visuals) - 15 points
- Expertise in Discussion - 10 points
- Award for Outstanding Writing - 10 points (extra)

The discussion is expected to differentiate (or not) the students. It can also change the grade in circumstances where the student shows less effort / contribution than his/her peers. We understand you partition the effort, but this is a coding exercise and not merely a writing contribution *for example*.

Guideline 11: What should be in the final moodle delivery

The delivery (zip file) will be expected at the 8 of June of 2025 (3 days of tolerance with penalty of 0.5 in 20) through moodle. We don't accept submissions by email.

The final submission must include, in the zip file: code (zip with Databricks Notebooks), data (zip with CSV files or text file with proper link), and report (PDF). **You may add other attachments to the zip (if they make sense, like a presentation).** Notebooks that do not run on Databricks will not be graded, even if they run Spark. Submit the actual notebook(s) in a zip, not a link to it. **Links will not be evaluated**.

These guidelines are available in the **Project Section on Moodle**.

I hope we have covered most of the questions,

Feel free to contact the teaching staff,

Best regards,

Márcia L. Baptista