Final Examination Assignment

1. This exam is about a group project. You are organized in groups of three-four students.
2. Find a suitable dataset (also different datasets joined together). During this phase you can get in touch with me to double check the quality of your choice.
3. **ESSENTIAL POINT:** You have to think also to data scalability and you should IMAGINE that your data source could be a **Big Data Source**. Therefore, you have to **design** a solution (provide an architecture) that scales in such a Big Data context. Do not make assumptions about input data dimension unless you are filtering them - meaning that you control the final size (using your Big Data tools).
4. Your support NOSQL dataset is MongoDB. You should use it for long-term storage and data selection support.
5. Perform a *preliminary* data exploration analysis to understand how your data look like and to formulate initial hypotheses (choose your analysis direction and strategy).
6. Consolidate your hypotheses and start to prepare your data for the analysis.
7. Design a data analysis campaign, provide a proof to your hypotheses.
8. Prepare a **10 minutes** (please respect this timeline) PowerPoint (or similar tool) presentation for discussing your project during the exam.

    1. During the presentation you have to convince me about the quality of your technical solution and the utility of your findings.

9. [OPTIONAL] - prepare a document reporting your project (no more than 5 pages).
10. The exam will be about an oral presentation of your work using your Powerpoint presentation. Organize your talk so that each member of the group presents a piece of the work. During the exam I will ask details about your technical solution, I will ask you notions studied during the course (also and especially if they are not used in the project), I could ask you to write little pieces of new code to prove that you are able to do so.

Points to be covered in the project

1. You must write at least one batch MapReduce Job on Hadoop (you can also use Hive - in this case, you may expect questions on the form of your Mapper and Reducer functions).
2. You must use PySpark facilities to use Spark (DataFrames or MLib) for designing a

scalable solution so that your analysis can be reproduced also if your data source is a Big Data one.

3. You must use MongoDB as a project repository.
    1. Write at list one complex MongoDB query to filter your data.

4. You must use Python libraries for data analysis (depending on your data and on your analysis direction: Pandas, Scikit-learn, and/or Networkx, Seaborn, Matplotlib, and so forth).

Final Comment

All the technologies and tools adopted (Hadoop, Spark, python, etc.) must be used in your local machine (in a local Virtualbox virtual machine - or similar virtualization environment - as seen in the practical lessons). Members of each group should cooperate (also using Skype/Meet/Zoom/Teams free video call) to achieve project results.