Analyzing Stock Market Values

Data Processing and Analytics (DPA) Course - International Master DISS UCBL 23/24

The objective of this project is to use Spark Streaming to analyze the value of different stocks through time.

The Dataset

The data is contained in a CSV file named **stock.csv.** For this project, we will consider only two columns, the last one that is the name of the stock, and the second one that is its value. You don't have to worry about parsing the file, the notebook "Kafka_Producer_for_Project" will take care of that.

Development environment

For this project, I suggest you use the same environment you used for the TP related to Spark and Spark Streaming.
Use the notebooks you find on Moodle, in particular

- The notebook **Kafka_Producer_for_Project** reads the file and ingests the data into Kafka with the schema < name, price, timestamp >. Consider that for the timestamp it uses the current one.
    - Put the csv with the data in the same folder.
- The notebook **Project Template** registers Spark to the stream and puts it in a manageable form. Be careful not to edit the cell already there unless you know what you are doing ;).

Tasks

Imagine you want to create a dashboard that shows statistics in real time about what's at the stock market. For instance - given a time window of your choice (it can vary between tasks, and motivate it in the report) - compute:

1. The N most valuable stocks in each windows
2. Select the stocks that lost value between windows
3. Find the stocks that gained the most between windows
4. Implement a control that checks if a stock does not lose too much value in a period of time (feel free to choose the value you prefer).
5. Imagine you own some stocks (stored in a data frame with the schema <name, amount of stocks owned>). Compute how your asset changes with the fluctuation of the market.

Expected Deliverable

- The Jupyter Notebook. The code should be working.
- A document (pdf) containing
    - An explanation of what you did
    - The results of the analysis for each task
        - Bonus point if you manage to use some visualizations (OPTIONAL)
    - A user guide on how to run the notebook (e.g., which cells do what, or any other information that makes it easy for us to understand how to make your code work)

Evaluation

- 3pts for task 1 and 5.
- 4 pts for task 2,3 and 4
- 2 pts for the report
- 1pt (bonus) for visualizations
- 21 points total