Big Data
PySpark application

In this lab, we will go through the entire process of analyzing a set of Nasdaq tech stocks using the spark framework. The main idea of this lab is to use spark to build a python application that will deliver interesting insights from this data in order to bring value to someone who is wishing to invest in those stocks.

The objectives of the lab are :

Apply the concepts we have seen during the lectures and the precedent introduction tutorial in a real python program : it means you have to structure your program into functions (and classes if you can)

Remember to always check for the documentation :
https://spark.apache.org/docs/latest/api/python/index.html

Be innovative and extract useful information from the stock data

Build an application that can be used by a trader to get insights and take decisions

Think genericity by a trader t

Optional : Use visualization libraries like Matplotlib, Plotly, Seaborn or Ggplot to present your findings

Before start working on your solution, create a notebook with your full name and student number in **the first markdown** cell. You should write the full solution and explanations in this notebook; it will be your rendering for this lab.

For each step, I encourage you to have two cells, one markdown cell to explain your code and approach, and the other one for the code functions and the code instructions and functions calls

Project main steps :

In order to build the application that will deliver insights from the csv files. We have to implement the following steps :

1. Exploration
2. pre-processing
3. analysis and visualizations

Use spark to read the files, build a dataframe and check (eventually set-up) the correct columns types (datetime type…) . The read function should be generic and take the path as argument, you can also program the schema.

- Explore the different stocks historical prices and try to code **functions** that will look at the dataframe and extract some useful information about the input data :

    - Show the first and last 40 rows of each stock price
    - Get the number of observations
    - Deduce programmatically what is the period you have between the data points : for example, if you have data point with the following date [01/01, 02/01, …..], you shoud have a function that will analyse the difference between the dates automatically and deduce it is a day period
    - Descriptive statistics for each dataframe and each column (min, max, standard deviation)
    - Number of missing values for each dataframe and column
    - Correlation between values
- The exploration process involves answering some questions about the data you have in order to gain a good understanding of it, the questions should be technical and business in the same time. Some examples of the question you could try to answer using spark and koalas are :
    - What is the average of the opening and closing prices for each stock price and for different time periods (week, month, year)
    - How do the stock prices change day to day and month to month (may be you can create new columns to save those calculations)
    - Based on the opening and closing price, calculate the daily return of each stock
    - What are the stocks with the highest daily return
    - Calculate the average daily return for different periods (week, month, and year)

- **Moving average :** The moving average is calculated by adding a stock's prices over a certain period and dividing the sum by the total number of periods. For example, if you want to calculate the moving average for the opening price of the stock ABC. You look at the opening price over five periods and calculate the average. For example if the opening price over the past five days were 25.40, 25.90. 26.50, 26.30 and 27.90. Then, the moving average of the opening price of the last day is 26.40. **Code a function** that take as input a dataframe, a column name, the number of points to consider for the moving average (5 in the example) and add a new column to the dataframe with the values of calculated moving average

- Correlation : Is there any correlation between the different stocks you have ? **Code a function** that takes as input the values of two stocks (you should decide what is the data type that will handle the values) and calculate the correlation between them

- When investing in stocks, the return rate is very important. **Code a function** that calculates the return rate of the stock in different periods (week, month and year)

- Given a specific month, what is the stock with the best return rate. **Code a function** that takes as input a start date and a period (month, year), calculate the return rate for each stock and return the one with the best return rate.

- The exploration process is the opportunity for you to gain a good understanding of the data you are working with. Work on your imagination (helped with some google research) and think about **8 insights** that can be helpful for our use case.