Data Analytics with Spark

The evaluation project of the course consists of two exercises.

Organization

- You can work in teams 2 or 3 maximum. Each team shall submit one report.
- The report is expected in PDF or ipynb format. It should contain the pyspark code, results of execution, and explanation of the solution. You can separate the solution of each exercise in a different file.
- Don't forget to indicate the names of the team members in the report
- The deadline for submission is **22/01/2023** midnight.
- The report is to be submitted on the moodle page of the course.

Exercise 1

- The revenues of a souvenir **shop** that is present in different cities in France is given in text files. Each file contains the monthly income of a shop's branch over a one-year period (each line contains a month and the corresponding revenue). The data is available in the zip file input.zip
- Each **store** (shop's branch) is identified by a name as follows:
    - city_i; i=1, 2, ... Depending on the number of stores in the city
    - If only one store is present in a city X, it is identified by the name of the city

You need to write a pyspark script that allows to display different statistics on the shop's performance:

- Average monthly income of the shop in France
- Average monthly income of the shop in each city
- Total revenue per city per year
- Total revenue per store per year
- The store that achieves the best performance in each month

You can solve the exercise using RDD <u>or</u> DataFrames transformations.

Using *wholeTextFiles* method (https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.SparkContext.wholeTextFiles.html) could be useful when loading data.

Exercise 2

The objective is to build a Spark Streaming application that shows popular hashtags on Twitter.

- You will need first to write a python application that can stream live tweets from Twitter API. You will need to create a Twitter Developer account and generate tokens to get access the twitter API. You can use existing python libraries (such as Tweepy) for this part.
- Then you can write a Spark Streaming application that connects to the first part, extracts hashtags, and displays the 10 most popular among them in the last 10 minutes.

It is sufficient to print out the results. Plotting the results in a graph is a plus.