Exercise 01 [Matrix Multiplication]

For this exercise, a matrix is represented in the following way: (i, j, value) where "i" is the row number, "j" the column number and "value" the value of the matrix at the i^{th} row and j^{th} column.

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \end{pmatrix} \cdot \begin{pmatrix} 2 & 4 & 8 \\ 1 & 5 & 10 \\ 3 & 6 & 9 \end{pmatrix}$$

1 Create two RDDs to represent the following matrices.

2 Find the SQL query that allows you to compute matrix multiplication if you suppose that each matrix represents a table, i.e., Table A (column I, column J, column Value) represents matrix A, and Table B (column I, column J, column Value) represents matrix B.

3 Implement the SQL query you just designed in Spark and test out your implementation. Result should be:

$$\begin{pmatrix} 13 & 32 & 55 \\ 30 & 75 & 129 \end{pmatrix}$$

4 To test out the scalability of your algorithm generate two random matrices: A of dimensions (1 000 000 x 3) and B of dimensions (3 x 3) and apply your Spark code once again. Do you think you can improve the performance of your Spark code? If so how?
Note: the two random matrices generated must have the same structure as above, i.e. (row number, column number, value).

Thanks to AISSAOUI Ahlam for her contribution

Exercise 02 [Simple Moving Average (SMA)]

Suppose you have time series values Y1, Y2, … Yn. A value Yi is represented as such: (timestamp, value at given time). m is called the “window” of the Simple Moving Average we aim to compute, e.g.: 7 days.

For a given natural m>0, this operation consists of mapping each Yi to Yi’, where:

$$Yi' = avg(Y(i-m) + Y(i-(m-1))+\ldots +Y(i))$$

take Y(i-m) as NULL if Y(i-m) does not exist in the time series data Consider the following dataset:

Timestamp (i)	**Value at timestamp (Y(i))**
2018-03-10T15:27:18+00:00	17.00
2018-03-11T12:27:18+00:00	13.00
2018-03-12T11:27:18+00:00	25.00
2018-03-13T15:27:18+00:00	20.00
2018-03-14T12:27:18+00:00	56.00
2018-03-15T11:27:18+00:00	99.00
2018-03-22T11:27:18+00:00	156.00
2018-03-31T11:27:18+00:00	122.00
2018-04-15T11:27:18+00:00	7000.00
2018-04-16T11:27:18+00:00	9999.00

Given a window m=7 days, transform the dataset into an RDD, then design and implement the Spark algorithm that allows you to compute the simple moving average for each day of the dataset.

Note: Consider that the timestamps composing the dataset are not always sequential.