Project 5

**Overview**

The goal of the project is to build a time series analysis pipeline of a Covid19 dataset.

The system must provide distributed computing capabilities using Apache Spark middleware.

Being the problem a batch processing on a given dataset, we decided to use SparkSQL to develop the processing.

The overall computation is handled by one big Spark driver program which provides result for each of the three subqueries required, and each subquery uses the computations done by the previous queries.

To provide protection against crashes, each subquery is persisted, when computed, by Spark workers.

**Query 1**

Makes a windowing function

1. partitioning by country,
2. ordering by day,
3. aggregating in the moving average horizon.

Finally, an average of cases over the window is computed

**Query 2**

Makes a windowing function

1. partitioning by country,
2. ordering by day.

A new daily lagged column of previous query is then added to the data, replacing null border values with not lagged values.

Result is obtained by aggregating (dividing) previous query and previous lagged attribute.

**Query 3**

1. Filters out all first day rows, because previous query values are the same for all the countries.
2. Creates a new (second) dataframe of one ordered days column, then we add increasing ID column for later join.
3. Makes a windowing function for the first dataframe

* partitioning by day,
* ordering (descending) on the previous query,
* ranking each day group.

For each of the day ranking position we finally add to the second (result) dataframe the corresponding ranked country column, which is created by selecting original dataframe on given rank, adding increasing ID column on it and then joining with the day column of the second dataframe.