Task 1

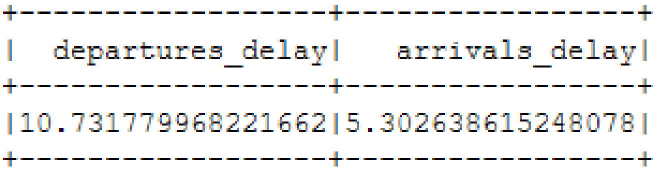
Your first task is to calculate the average flight delays in the dataset.

For answering this task we have to first create a data-frame containing all the values of the given csv file. Then it was necessary this data-frame to be converted into view in order to make use the SparkSQL.

After that, we created an SQL query and we have as a result the two averages from the columns “DEP_DELAY” and “ARR_DELAY”, as the below picture indicates. We have to clarify that the records were used for this query had both negative and positive values. The negative values in our case, probably indicate the earlier time of arrival or even departure.

Also, the dataset contained many null values, which they probably mean no delay. So, we had to replace these values with 0 number.

Finally, we have to mention the internal process of the query from the system, which is the splitting of the dataset to partitions and the editing on them in parallel in order to be as fast as possible.



```
+------------------+-----------------+
|  departures_delay|   arrivals_delay|
+------------------+-----------------+
|10.731779968221662|5.302638615248078|
+------------------+-----------------+
```