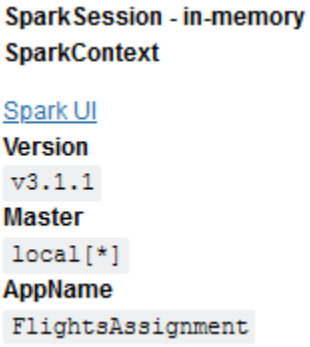
Big Data Systems and Architectures - Spark Assignment 2021

Exploring International Flights in 2017 Data – Task 1

The main scope of this assignment is to analyze a dataset about international flights in 2017, using Apache Spark to reveal insights about these data. The aforementioned data could be found here: http://andrea.imis.athena-innovation.gr/aueb-master/flights.csv.zip

The requested task is to calculate the average flight delays in our dataset. Before starting the implementation, it is required to install numpy, spark, findspark and pyspark. Some packages such as SparkContext, SparkSession, SQLContext, DataFrameReader, StringType should also be imported. It is important to run findspark.init(), so as to initialize the findspark.

A new entry point to Spark SQL is created by the use of SparkSession.builder with appName("FlightsAssignment"), as shown below:



After creating the temporary view, the data are loaded to a variable named `flights_data` by the use of spark.read() function with different options, such as that the dataset contains header and it is in comma delimited format.

Different functions, like show() and printSchema() are used to better understand the data, by showing the first rows of the dataset and the type of each variable.

To calculate the average delay of flights, the columns DEP_DELAY and ARR_DELAY are used as input for the departure and arrival delay respectively. The agg() function is used to aggregate the two different averages to one data frame, by implementing the avg() functions over the

columns DEP_DELAY and ARR_DELAY and renaming these averages to the aliases `AverageDepartureDelay` and `AverageArrivalDelay` respectively. In order the results to be more recognizable they are rounded with two decimal digits, with the use of round() function. The final result is assigned to a variable named `AvgDelay` and it is shown.