# Assignment: Spark MLlib

Using the “Flights Stats” data for the year 2014, we will predict the likeliness of a flight delay. Flight delay causes a lot of inconvenience and predicting flight delays could enable the travellers to remove uncertainty and plan ahead. For the purpose of this case study, we will use .csv file with 441622 rows in 17 columns. The schema of the table is as follows:

**dayOfMonth: Int**

**dayOfWeek: Int**

**carrier: String**

**tailNumber: String**

**flightNumber: Int**

**originId: String**

**origin: String**

**destinationId: String**

**destination: String**

**crsDepartureTime: Double**

**actualDepartureTime: Double**

**departureDelayMinutes: Double**

**crsArrivalTime: Double**

**actualArrivalTime: Double**

**arrivalDelayMinutes: Double**

**crsElapsedTime: Double**

**distance: Int**

Import these packages:

*import org.apache.spark.mllib.linalg.Vectors*

*import org.apache.spark.mllib.regression.LabeledPoint*

*import org.apache.spark.mllib.tree.{DecisionTree, RandomForest}*

*import org.apache.spark.sql.SQLContext*

*import org.apache.spark.{SparkConf, SparkContext}*

**EXERCISE 1:** Define the schema and load data from "2014.csv"

**EXERCISE 2:** Create an index of carrier, origin and destination which contains distinct values for each column

**EXERCISE 3:** Create a dataframe "features" that contains dayOfMonth, dayOfWeek, crsDepartureTime, crsArrivalTime, carrierIndex, crsElapsedTime, origin, destination, departureDelayMinutes and map it to an array

**EXERCISE 4:** Prepare a test model for the data and train it based on Decision Tree method

**EXERCISE 5:** Compute the predictions and print the accuracy