Tratamiendo de datos masivos – Máster IoT

SparkSQL

Vamos a utilizar parte del fichero con los datos de pasajeros del Titanic (titanic.csv). Los campos en este fichero .csv son:

survived: 0 si no sobrevivió, 1 en otro caso

pclass: clase dentro del barco: 1,2,3

name:nombre del pasajero

sex:male,female

age: edad

sibsp: Número de parientes entre hermanos y pareja embarcados

parch: Número de padres e hijos embarcados

ticket: número de ticket (string, puede contener letras)

fare: precio pagado por el billete

cabin: camarote (string)

embarked: puerto de embarque C = Cherbourg, Q = Queenstown, S = Southampton

A partir de este fichero csv convertido en Dataset de Spark queremos:

1) [6pt] Queremos saber si el ratio de supervivientes fue mayor en el caso de pasajeros de clase 1 que entre los de clase 3. Utilizando SparkSQL escribir un código que
    a. Obtenga los pasajeros de clase 1 y de clase 3. Ver el ejemplo "where". Esto generará 2 nuevos Datasets, `first` y `third`. Mostrar el número de elementos de cada dataset. (debe salir 216 y 491).
    b. A partir de los dataset anteriores, obtenga los pasajeros de clase 1 y de clase 3 que sobrevivieron (usando "where"). Esto generará 2 nuevos Datasets firstSurvived y thirdSurvived. Mostrar el número de elementos de cada dataset.
    c. Mostrar los ratios de supervivencia por cada clase. Ayuda: hacer la división convirtiendo alguno de los valores a float (escribiendo (float) antes del valor). Si no hará la división entera que dará 0.

2) [2] Escribir una expresión en Spark que elimine todas las filas en las que *fare* valga null generando un nuevo RDD.
3) [2] Sobre el RDD anterior aplicar un 16% de IVA a los pasajeros de clase 3 y un 21% a los de clase 3, e indique el nombre de los pasajeros que han pagado más en total.