UNIVERSIDAD AUTONOMA DE MADRID

Laboratorio 6: SparkSQL
TAREA 2

Vamos a seguir trabajando con el conjunto de datos del Heterogeneity Dataset for Human Activity Recognition (HHAR) que contiene información de los sensores de movimientos de teléfonos y relojes. Utilizaremos los datos completos publicados en la página de la asignatura (no los datos de muestra).

1. El primer objetivo de esta tarea es comprobar que el formato parquet reduce los tamaños de los ficheros de forma considerable con respecto a ficheros de texto en formato csv, por ejemplo. Para ello, debes generar un fichero parquet para cada uno de los ficheros csv y entregar una tabla en la que aparezcan los tamaños de cada uno de los ficheros csv y de sus correspondientes ficheros parquet.

2. El segundo objetivo de esta tarea es medir el tiempo de ejecución de la tarea 1 cuando ésta se realiza de distintas formas. Consideraremos los siguientes casos:

   Caso 1. Se crean RDDs para cada uno de los ficheros csv (esto corresponde al notebook realizado en el Laboratorio 5)
   Caso 2. Se crean DataFrames a partir de los RDDs (esto corresponde a la Tarea 1 del Laboratorio 6)
   Caso 3. Se crean DataFrames a partir de los ficheros parquet generados en el apartado 1
   Caso 4. Se crean DataFrames a partir de los ficheros csv originales. Para ello, puedes utilizar la función spark.read.csv *http://spark.apache.org/docs/latest/api/python/pyspark.sql.html#pyspark.sql.DataFrame*

Para esta tarea se utilizará un único notebook que formará parte del archivo .zip correspondiente al Laboratorio 6. No se deben incluir los ficheros de datos. Las funciones deben estar documentadas.

Se entregarán también las dos tablas correspondientes a los apartados 1 y 2.

FECHA DE ENTREGA: 9 de enero