ECOSISTEMA SPARK

Práctica 2: SparkSQL
TAREA 3

Vamos a trabajar con un nuevo conjunto de datos de BookCrossing (http://www.bookcrossing.com), una comunidad de amantes de los libros que intercambia libros por todo el mundo y comparte sus experiencias.

El primer paso es descargarse el CSV Dump de la página http://www2.informatik.uni-freiburg.de/~cziegler/BX/

1. El primer objetivo de esta tarea es utilizar las funciones de la API para resolver las siguientes consultas:
    a. Lista de usuarios junto con el número de libros que han valorado
    b. Rating máximo recibido por cada editorial
    c. Nombre del autor que ha recibido más ratings

2. El segundo objetivo de esta tarea es utilizar las Window Functions de Spark SQL para resolver las siguientes consultas:
    a. ¿Cuál es el título del libro con mayor número de ratings para cada editorial?
    b. ¿Cuál es la diferencia entre el número de ratings de cada libro y el número de ratings del libro con mayor número de ratings de la misma editorial?

Para esta tarea se utilizará un único notebook que formará parte del archivo .zip correspondiente a la Práctica 2. No se deben incluir los ficheros de datos. Las funciones deben estar documentadas.

FECHA DE ENTREGA: 9 de enero