Práctica SDPD (Parte I)

- Datos: Fichero con tweets del GP MotoGP de Qatar 2014 (ALTO DATABASE).
- Tecnologías: Spark SQL, Spark Streaming, Apache Kafka.
- Versión Spark: 2.0.2 o superior.

En esta primera parte de la práctica, se pide resolver los siguientes ejercicios utilizando las tecnologías vistas en clase hasta este momento en la asignatura.

SPARK SQL

1. Cargar en Spark el fichero de datos "DATASET-Twitter-23-26-Mar-2014-MotoGP-Qatar.csv", definiendo para ello el esquema adecuado en Spark SQL.
2. Realizar las siguientes tareas:
   - a) Contabilizar el número total de menciones a los pilotos Marc Márquez, Valentino Rossi y Dani Pedrosa.
   - b) Contabilizar los 5 países que más tweets han publicado (considerando los tweets que contengan dicha información).
   - c) Contabilizar los 3 hashtags más utilizados (que aparezcan el mayor número de veces) en el cuerpo de los tweets (campo "body").

SPARK STREAMING + KAFKA

- Utilizando como base las herramientas presentadas en clase (productor y consumidor de Kafka genéricos en Python), crear una aplicación local de Spark Streaming que lea progresivamente los tweets insertados en una cola de Kafka identificada por el topic "Quatar_GP_2014", defina un intervalo de procesamiento de datos de 5 segundos y realice las siguientes tareas:

- a) Calcular el número total de menciones recibidas por cada cuenta de usuario durante el intervalo de 5 segundos.
- b) Calcular la frecuencia total acumulada de apariciones de cada hashtag en el campo body, actualizando un ranking con los 5 hashtags con mayor frecuencia de aparición.
- c) Calcular en una ventana temporal 20 segundos con offset de 10 segundos la frecuencia de aparición de cada uno de los 3 posibles tipos de tweets (TW-RT-MT).