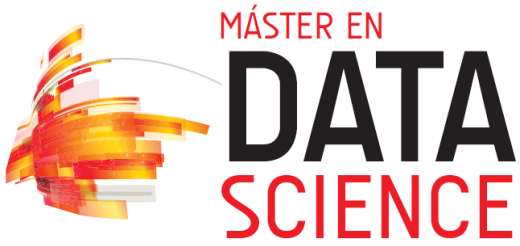

Sistemas Distribuidos de Procesamiento de Datos

Profesor: Felipe Ortega.

Práctica final T2 (Parte I)

- **Datos**: Fichero con tweets del GP MotoGP de Qatar 2014 (ALTO DATABASE).
- **Tecnologías**: Spark SQL, DataFrames, Jupyter notebook.
- **Versión Spark**: 2.2.1 o superior.
- **Fecha de entrega**: Domingo, 6 de mayo de 2018, a las 23:55.

En esta primera parte de la práctica, se pide resolver los siguientes ejercicios utilizando un notebook de Jupyter con Spark SQL .

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) Calcular el **número total de *retweets* por usuario** para los 50 usuarios con más mensajes en la muestra de tweets analizados. Calcular, para cada uno de estos usuarios la media de enlaces (URLs) enviados por mensaje. **(2.5 puntos)**.
   - b) Calcular el número total de mensajes que contienen información de geolocalización en el campo **LOCATION**.**(2.5 puntos)**.
   - c) Calcular las 10 cuentas de Twitter que más han sido mencionadas en todo el conjunto de datos analizados. **(2.5 puntos)**.
   - d)Calcular los 10 mensajes más retweeteados y los 10 mensajes que han acumulado más respuestas en la muestra de datos analizados. Ahora, restringe la búsqueda a los mensajes en el intervalo **2014-03-24 04:00 - 2014-03-24 10:00**. **(2.5 puntos)**.