# Actividad. Procesamiento de datos de Twitter con Apache Spark

| | |
|---|---|
| Unidad | 5 |
| Entrega | Documento en formato PDF |

## 1. Enunciado

En esta actividad se van a aplicar los conceptos vistos a lo largo de esta unidad. Los objetivos de aprendizaje a alcanzar en esta actividad son: disponer de las capacidades de procesar una gran cantidad de información mediante Spark, saber definir métricas y analíticas de los datos procesados, y ser capaces de reproducir y acelerar consultas de base de datos con Spark sobre datos estructurados.

El alcance principal de esta actividad consiste en procesar datos formateados en JSON provenientes de la red social Twitter. La fuente de datos está disponible para su procesamiento en el siguiente directorio de HDFS: hdfs:///Twitter/(esta carpeta contiene tweets extraídos del canal general).

La actividad consta de lo siguiente:

1. Paso 1. Elegir uno de los ficheros de esa carpeta y realizar los estudios que se consideren necesarios sobre los datos que contiene.
2. Paso 2. Preparar un informe que indique los siguientes aspectos:
   a. Número de ejecutores y núcleos elegidos.
   b. Cantidad de datos procesados en función de cada intervalo temporal (hora y día).
   c. Esquema JSON del contenido de la fuente de datos.
   d. Los 10 *hashtags* con mayor número de apariciones (*trending topic*).
   e. *Trending topic* en función de cuatro idiomas del perfil de usuario.
   f. El usuario con mayor número de seguidores que participó ese día.

   Se recomienda iniciar el análisis sobre un único fichero y, posteriormente, aumentar la cantidad de datos procesados.

## 2. Detalles de la entrega

- Escribir un documento que describa las soluciones propuestas.
- Subir de forma grupal el documento en formato PDF al campus virtual.

# Actividad. Procesamiento de datos de Twitter con Apache Spark

**Enlace**

Información práctica sobre la lectura de archivos en formato JSON:

https://spark.apache.org/docs/latest/sql-programming-guide.html