UNIVERSIDAD DE GRANADA

DECSAI

# Cloud Computing: Servicios y Aplicaciones

**Curso 2016 – 2017**

# Guión de prácticas 3

## Computación Distribuida y Escalable con Hadoop

El **objetivo** de esta práctica es realizar programas escalables para mejorar la eficiencia en entornos Big Data. Para ello, haremos uso del entorno que se ha convertido en un estándar de facto como es **Hadoop**, utilizando *HDFS* como sistema de archivos distribuido y *Hadoop-MapReduce* como mecanismo de ejecución.

Para constatar el manejo de la herramienta anterior, el alumno deberá realizar las **tareas** que se describen a continuación y entregar documentación describiendo las tareas realizadas. Se recomienda seguir el tutorial asociado en la página https://github.com/manuparra/MasterDegreeCC_Practice/

## Tareas: Realizar los objetivos que aparecen a lo largo del documento de prácticas

## Documentación: Es necesario entregar un informe con la siguiente estructura:

1. Portada con:
    a. Nombre de la asignatura
    b. Nombre de la práctica
    c. Nombre del alumno
    d. Dirección de correo electrónico
2. Salida resumida para los diferentes cálculos realizados.

El documento se entregará en formato PDF. El código fuente Java asociado con las diferentes tareas MapReduce se adjuntará también en una carpeta a tal efecto.

Tanto el código fuente como el documento, se empaquetará en un único fichero .zip con el nombre "`practicaHadoop`" y se entregará a través de la plataforma docente de decsai.

## Fecha límite para la entrega:

22 de Mayo de 2016 a las 23:59h.

Utilizando como base el conjunto de datos ECBDL14 situado en la carpeta `/tmp/BDCC/datasets/ECBDL14/ECBDL14_10tst.data` obtener los siguientes datos estadísticos descriptivos:

1. Calcula el valor mínimo de la variable (columna) 5
2. Calcula el valor máximo de la variable (columna) 5
3. Calcula al mismo tiempo los valores máximo y mínimo de la variable 5
4. Calcula los valores máximo y mínimo de todas las variables (salvo la última, que es la etiqueta de clase)
5. Realizar la media de la variable 5
6. Obtener la media de todas las variables (salvo la clase)
7. Comprobar si el conjunto de datos ECBDL es balanceado o no balanceado, es decir, que el ratio entre las clases sea menor o mayor que 1.5 respectivamente.
8. Cálculo del coeficiente de correlación entre todas las parejas de variables