
UNIVERSIDAD DE GRANADA

MASTER PROFESIONAL EN INGENIERÍA INFORMÁTICA

PRÁCTICA 4

Hadoop

Autor:

Manuel Jesús García Manday
(nickter@correo.ugr.es)

Master en Ingeniería Informática

2 de junio de 2017

Índice

1. Objetivo.	3
2. Introducción.	3
3. Tareas.	4
3.1. Ejecutar el algoritmo "Random Forest" sobre el conjunto de datos BNG_heart y comprobar el rendimiento alcanzado de acuerdo a los siguientes casos	4

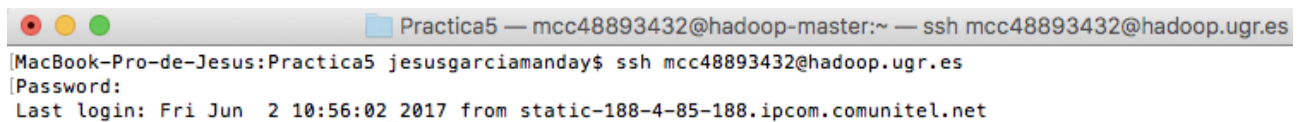
1. Objetivo.

El objetivo de esta práctica es conocer las alternativas para realizar experimentaciones de Ciencia de Datos. Para ello haremos uso del entorno **Hadoop**, utilizando **HDFS** como sistema de archivos distribuido y **MapReduce** como mecanismo de ejecución. Por último, aplicaremos la biblioteca **Mahout** para lanzar algoritmos de clasificación sobre conjuntos tipo **Big Data**.

2. Introducción.

Para comenzar a realizar las tareas que se piden en esta práctica, es necesario en primer lugar realizar una serie de pasos iniciales que se describen a continuación.

Realizamos una conexión remota hacia el servidor **hadoop.ugr.es** y una vez dentro comprobamos la existencia del conjunto de datos en el directorio indicado.

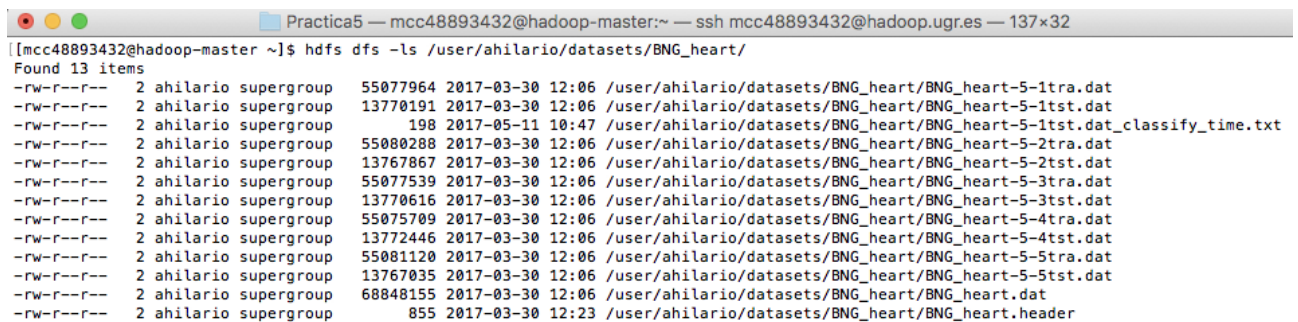


```

Practica5 — mcc48893432@hadoop-master:~ — ssh mcc48893432@hadoop.ugr.es
[MacBook-Pro-de-Jesus:Practica5 jesusgarciamanday$ ssh mcc48893432@hadoop.ugr.es
[Password:
Last login: Fri Jun  2 10:56:02 2017 from static-188-4-85-188.ipcom.comunitel.net

```

Figura 1: Conexión remota a **hadoop.ugr.es**.



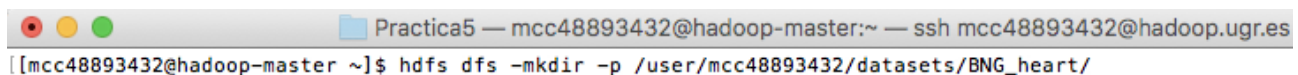
```

Practica5 — mcc48893432@hadoop-master:~ — ssh mcc48893432@hadoop.ugr.es — 137x32
[mcc48893432@hadoop-master ~]$ hdfs dfs -ls /user/ahilario/datasets/BNG_heart/
Found 13 items
-rw-r--r-- 2 ahilario supergroup 55077964 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-1tra.dat
-rw-r--r-- 2 ahilario supergroup 13770191 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-1tst.dat
-rw-r--r-- 2 ahilario supergroup 198 2017-05-11 10:47 /user/ahilario/datasets/BNG_heart/BNG_heart-5-1tst.dat_classify_time.txt
-rw-r--r-- 2 ahilario supergroup 55080288 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-2tra.dat
-rw-r--r-- 2 ahilario supergroup 13767867 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-2tst.dat
-rw-r--r-- 2 ahilario supergroup 55077539 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-3tra.dat
-rw-r--r-- 2 ahilario supergroup 13770616 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-3tst.dat
-rw-r--r-- 2 ahilario supergroup 55075709 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-4tra.dat
-rw-r--r-- 2 ahilario supergroup 13772446 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-4tst.dat
-rw-r--r-- 2 ahilario supergroup 55081120 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-5tra.dat
-rw-r--r-- 2 ahilario supergroup 13767035 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart-5-5tst.dat
-rw-r--r-- 2 ahilario supergroup 68848155 2017-03-30 12:06 /user/ahilario/datasets/BNG_heart/BNG_heart.dat
-rw-r--r-- 2 ahilario supergroup 855 2017-03-30 12:23 /user/ahilario/datasets/BNG_heart/BNG_heart.header

```

Figura 2: Comprobando los datasets.

Una vez corroborada su existencia copiamos dicha carpeta en un directorio local, para posteriormente copiarla en un directorio en **hdfs** que hayamos creado previamente.

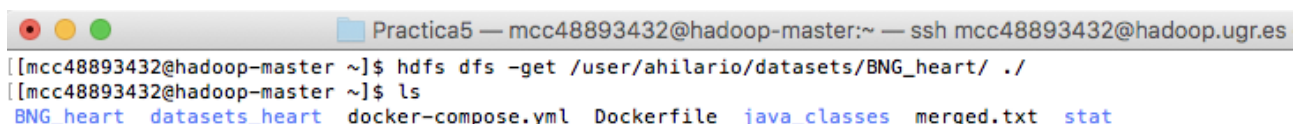


```

Practica5 — mcc48893432@hadoop-master:~ — ssh mcc48893432@hadoop.ugr.es
[mcc48893432@hadoop-master ~]$ hdfs dfs -mkdir -p /user/mcc48893432/datasets/BNG_heart/

```

Figura 3: Creando nuevo directorio en **hdfs**.

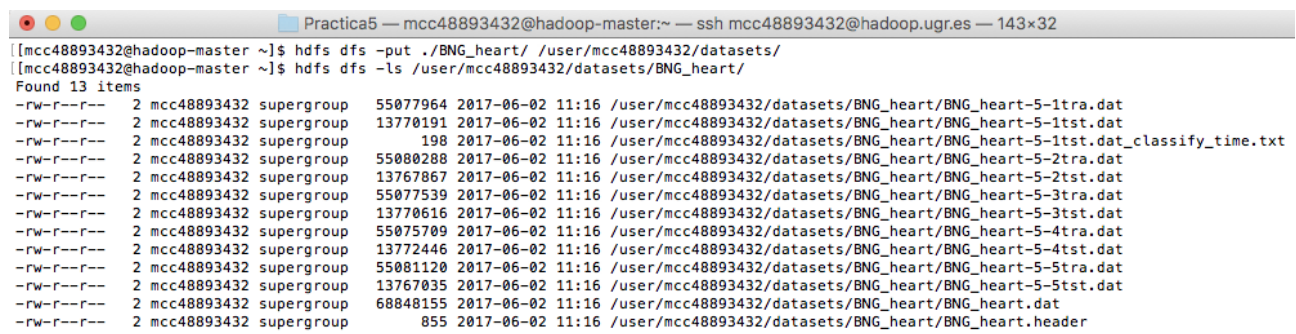


```

Practica5 — mcc48893432@hadoop-master:~ — ssh mcc48893432@hadoop.ugr.es
[mcc48893432@hadoop-master ~]$ hdfs dfs -get /user/ahilario/datasets/BNG_heart/ ./
[mcc48893432@hadoop-master ~]$ ls
BNG_heart  datasets_heart  docker-compose.yml  Dockerfile  java_classes  merged.txt  stat

```

Figura 4: Trayendo datasets en un directorio local.



```

Practica5 — mcc48893432@hadoop-master:~ — ssh mcc48893432@hadoop.ugr.es — 143x32
[mcc48893432@hadoop-master ~]$ hdfs dfs -put ./BNG_heart/ /user/mcc48893432/datasets/
[mcc48893432@hadoop-master ~]$ hdfs dfs -ls /user/mcc48893432/datasets/BNG_heart/
Found 13 items
-rw-r--r-- 2 mcc48893432 supergroup 55077964 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-1tra.dat
-rw-r--r-- 2 mcc48893432 supergroup 13770191 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-1tst.dat
-rw-r--r-- 2 mcc48893432 supergroup 198 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-1tst.dat_classify_time.txt
-rw-r--r-- 2 mcc48893432 supergroup 55080288 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-2tra.dat
-rw-r--r-- 2 mcc48893432 supergroup 13767867 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-2tst.dat
-rw-r--r-- 2 mcc48893432 supergroup 55077539 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-3tra.dat
-rw-r--r-- 2 mcc48893432 supergroup 13770616 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-3tst.dat
-rw-r--r-- 2 mcc48893432 supergroup 55075709 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-4tra.dat
-rw-r--r-- 2 mcc48893432 supergroup 13772446 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-4tst.dat
-rw-r--r-- 2 mcc48893432 supergroup 55081120 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-5tra.dat
-rw-r--r-- 2 mcc48893432 supergroup 13767035 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart-5-5tst.dat
-rw-r--r-- 2 mcc48893432 supergroup 68848155 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart.dat
-rw-r--r-- 2 mcc48893432 supergroup 855 2017-06-02 11:16 /user/mcc48893432/datasets/BNG_heart/BNG_heart.header

```

Figura 5: Importando datasets a un directorio **hdfs**.

3. Tareas.

- 3.1. Ejecutar el algoritmo "Random Forest" sobre el conjunto de datos **BNG_heart** y comprobar el rendimiento alcanzado de acuerdo a los siguientes casos