OBRADOIRO: INICIACIÓN Ó BIG DATA CON R COA LIBRARÍA sparklyr

M. Aurora Baluja González[1], Javier López Cacheiro[2]

[1] Servizo de Anestesioloxía e Reanimación. Complexo Hospitalario Universitario de Santiago (CHUS)

[2] Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia (CESGA)

RESUMO

A tecnoloxía que permite a lectura e escritura paralela e eficiente de macrodatos ("Big Data") experimentou unha gran difusión nos últimos anos, coa aparición de plataformas open source - Hadoop, Spark-, e linguaxes como Scala.

O paquete sparklyr, lanzado en 2016 por RStudio, permite unha comunicación directa coa API de Spark para " dataframes2", utilizando a sintaxe de R e dplyr, de forma cómoda para o usuario final.

Neste obradoiro aprenderemos:

- Os conceptos máis importantes sobre o funcionamiento da tecnoloxía para Big Data.
- A posta en marcha dunha instancia de Spark no noso PC, ou (para os que teñan conta) no servidor Big Data do CESGA.
- As operacións máis frecuentes que permite o stack Hadoop-Spark-Sparklyr sobre os nosos datos.

Requírense coñecementos básicos de R.
Para o obradoiro requírese ter instalado e en funcionamento (apórtanse suxerencias de vencellos de axuda) :

1. Pasos comúns aos 3 sistemas operativos:
   - Instalar R: https://cran.r-project.org/.
   - Instalar R Studio: https://www.rstudio.com/products/rstudio/download/#download
   - Instalar algunha ferramenta que permita usar Git (recomendable, non imprescindible).
2. MS Windows:
   - Instalar Java (JDK): https://www.theserverside.com/tutorial/How-to-install-the-JDK-on-Windows-and-setup-JAVA_HOME.
   - Instalar Apache/Spark: https://hernandezpaul.wordpress.com/2016/01/24/apache-spark-installation-on-windows-10/.
   - Instalar paquete sparklyr desde R.

3. OSX (o proceso pode levar bastante tiempo no caso de se completar desde cero)

- Instalar xcode: https://developer.apple.com/xcode/ (a descarga e instalación poden levar tempo, ó instala-las command-line developer tools)
- Instalar Java e Apache/Spark: https://medium.freecodecamp.org/installing-scala-and-apache-spark-on-mac-os-837ae57d283f.
- Instalar paquete sparklyr desde R.

4. GNU/Linux (exemplo con Ubuntu 16.04):

- Instalar Java e Apache/Spark: https://www.tutorialkart.com/apache-spark/install-latest-apache-spark-on-ubuntu-16/.
- Instalar paquete sparklyr desde R.