|  |
| --- |
| Máster en Big Data |
| Tecnologías de Almacenamiento |
| 9. Hands-On: Spark |

Índice

[1. Introducción 3](#__RefHeading___Toc453_1462167123)

[2. Entorno 3](#__RefHeading___Toc455_1462167123)

[3. Inspección de los datos locales 3](#__RefHeading___Toc457_1462167123)

# Introducción

El objetivo de este Hands-On es el de familiarizarse con una de las librerías más populares del framework de Spark como es SparkSQL

# Entorno

Para la realización de los ejercicios se va a utilizar *spark-shell* en scala o python ya que nos proporciona un entorno muy dinámico para la introducción de funciones y nos permite recibir una respuesta inmediata.

Para ello, nos descargaremos Spark y levantaremos una *spark-shell* en local.

Sigue las siguientes instrucciones si es necesario:

https://sparkbyexamples.com/spark/install-apache-spark-on-mac/

https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/

https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/

Abre un Terminal (en windows preferiblemente PowerShell) apuntando al directorio *../spark-x.y.z-bin-hadoop/bin/*

El dataset que utilizaremos se llama *Summer-Olympic-medals-1976-to-2008.csv*.

# Inspección de los datos locales

Carga el dataset de la siguiente manera (con la ruta del dataset que aplique):

*val dataset = spark.read.option("header","true").option("inferSchema", "true").csv("/path/to/Summer-Olympic-medals-1976-to-2008.csv")*

1. Explora el dataset
2. ¿Que han aportado las opciones “header” y “inferSchema”?
3. ¿Como harías para contar las medallas conseguidas por año y país?
4. Usando SparkSQL muestra alguna métrica interesante.
5. ¿Que muestra la SparkUI: [http://localhost:4040](http://localhost:4040/) ?