|  |
| --- |
| Máster en Big Data |
| Tecnologías de Almacenamiento |
| 1. Hands-On: Desarrollo Apache Spark SQL |

Índice

[1. Introducción 3](#_Toc96244381)

[2. Entorno 3](#_Toc96244382)

[3. Creación del DataFrame 3](#_Toc96244383)

[4. Inspección de datos 4](#_Toc96244384)

# 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 ya que nos proporciona un entorno muy dinámico para la introducción de funciones y nos permite recibir una respuesta inmediata.

Para ello, utilizaremos la máquina virtual desplegada en Hands-On anteriores llamada Developer\_Hadoop y ejecutaremos el Spark Shell ubicado en /home/training/spark-1.3.1/bin

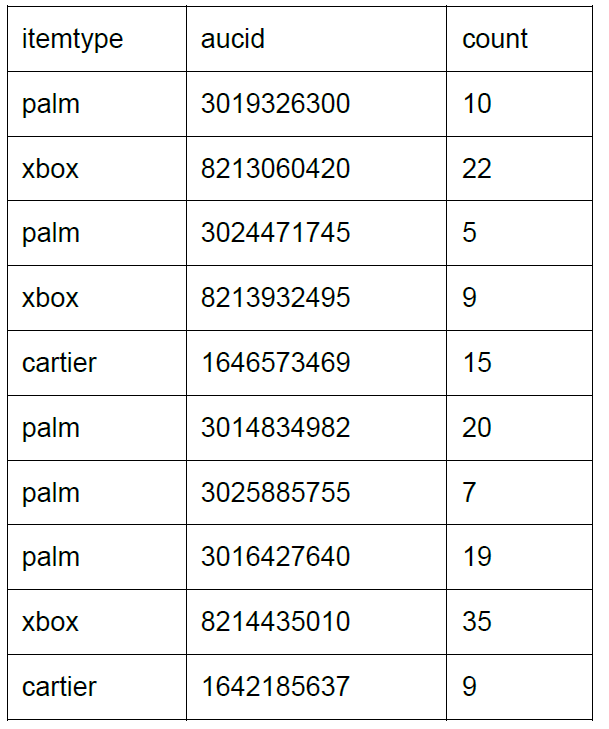
El dataset que utilizaremos se llama auctiondata.csv y está ubicado en */home/training/training\_materials/developer/data/auction.csv*

# Creación del DataFrame

1. Crear el SQL Context
2. Realiza el import correspondiente para convertir un RDD en un DataFrame
3. Definición del esquema usando una *case class*
4. Crea un RDD llamado *inputRDD* para cargar el *auctiondata.csv*. Asegurate de dividir el archivo de entrada con el separador “,”
5. Ahora Mapea el *inputRDD* a *case class*
6. Convierte el RDD resultante del punto anterior en un *DataFrame* llamado *auctionsDF* y regístralo como una tabla (Registrandolo de esta manera, podremos ejecutar consultas SQL usando los métodos proporcionados por sqlContext)
7. ¿Que acción puedes hacer para comprobar los datos en el DataFrame?
8. ¿Que función del DataFrame puedes utilizar para ver el esquema del mismo?

# Inspección de datos

1. ¿Cuál es el número total de pujas?
2. ¿Cuál es el número total de subastas distintas?
3. ¿Cuál es el número de objetos distintos?
4. Queremos contar el número de pujas por subasta y tipo de objeto (como se muestra a continuación). ¿Como se podría hacer? *(pista: usar groupBy)*

****

1. Por cada elemento subastada y tipo, calcular el máximo, mínimo y la media de pujas
2. Por cada elemento subastada y tipo, queremos la siguiente información (pista: utilizar *groupBy* y *agg*)
   1. Minima puja
   2. Maxima puja
   3. Puja media
3. ¿Cuál es el número de subastas cuyo precio es superior a 200?
4. Queremos ejecutar algunos comandos básicos sobre todas las subastas que tienen un elemento del tipo “xbox”. Es decir, queremos asilar en un nuevo dataframe con datos de los artículos “xbox” ¿Qué manera tenemos de hacer esto?

*(pista: al tener registrado el DataFrame como una tabla podemos usar sentencias sql, el resultado será un nuevo DataFrame donde podremos aplicar acciones sobre el)*

1. Mostrar el precio de todas las subastas con el elemento Xbox implicado (puedes utilitzar el dataframe anterior)