

1. Capítulo 2

- a. Descargar el Quijote

<https://gist.github.com/jsdario/6d6c69398cb0c73111e49f1218960f79>

Aplicar no solo count (para obtener el número de líneas) y show sino probar distintas sobrecargas del método show (con/sin truncate, indicando/sin indicar num de filas, etc) así como también los métodos, head, take, first (diferencias entre estos 3?)

- b. Del ejercicio de M&M aplicar:

- i. Otras operaciones de agregación como el Max con otro tipo de ordenamiento (descendiente).
- ii. hacer un ejercicio como el “where” de CA que aparece en el libro pero indicando más opciones de estados (p.e. NV, TX, CA, CO).
- iii. Hacer un ejercicio donde se calculen en una misma operación el Max, Min, Avg, Count. Revisar el API (documentación) donde encontrarán este ejemplo:

```
ds.agg(max($"age"), avg($"salary"))
```

```
ds.groupBy().agg(max($"age"), avg($"salary"))
```

NOTA: \$ es un alias de col()

- iv. Hacer también ejercicios en SQL creando tmpView

2. Capítulo 3

- a. Realizar todos los ejercicios propuestos de libro
- b. Leer el CSV del ejemplo del cap2 y obtener la estructura del schema dado por defecto.
- c. Cuando se define un schema al definir un campo por ejemplo `StructField('Delay', FloatType(), True)` ¿qué significa el último parámetro Boolean?
- d. Dataset vs DataFrame (Scala). ¿En qué se diferencian a nivel de código?
- e. Utilizando el mismo ejemplo utilizado en el capítulo para guardar en parquet y guardar los datos en los formatos:
 - i. JSON
 - ii. CSV (dándole otro nombre para evitar sobrescribir el fichero origen)
 - iii. AVRO
- f. Revisar al guardar los ficheros (p.e. json, csv, etc) el número de ficheros creados, revisar su contenido para comprender (constatar) como se guardan.
 - i. ¿A qué se debe que hayan más de un fichero?
 - ii. ¿Cómo obtener el número de particiones de un DataFrame?
 - iii. ¿Qué formas existen para modificar el número de particiones de un DataFrame?
 - iv. Llevar a cabo el ejemplo modificando el número de particiones a 1 y revisar de nuevo el/los ficheros guardados.

3. Capítulo 4

- a. Realizar todos los ejercicios propuestos de libro
- b. GlobalTempView vs TempView
- c. Leer los AVRO, Parquet, JSON y CSV escritos en el cap3

4. Capítulo 5

- a. Realizar todos los ejercicios propuestos de libro (excepto los de hive en caso de utilizar spark instalado en local y en el caso de RDBMS hacer únicamente ejemplo especificado más adelante)
- b. Pros y Cons utilizar UDFs
- c. Instalar MySQL, descargar driver y cargar datos de BBDD de empleados <https://dev.mysql.com/doc/employee/en/>
 - i. Cargar con spark datos de empleados y departamentos
 - ii. Mediante Joins mostrar toda la información de los empleados además de su título y salario.
 - iii. Diferencia entre Rank y dense_rank (operaciones de ventana)