# Ejercicios de práctica Scala-Spark

Todos los apartados deben resolverse con métodos generales (siempre que sea posible) sin tener en cuenta los datos que se dan de ejemplo.

1. Dada una lista de números enteros (por ejemplo: [1, 4, 5, 6, 7]) se quiere realizar varias operaciones en paralelo sobre cada elemento. Los pasos a realizar son:
   1. Abstraer la lista a un conjunto *resilient distributed dataset* (RDD).
   2. Formar un conjunto de RDD que contenga una lista con el elemento junto con el valor “0” si el elemento es divisible entre 2 y “1” en caso contrario. La salida debería ser de la forma: [ [1,1], [4,0], [5,1], [6,0], [7,1] ]
   3. Descartar aquellos que sean pares (filtrar en función de si es 0 o 1). La salida debería ser: [ [1,1], [5,1], [7,1] ]
   4. Tomando la salida anterior, sumar el valor de los elementos originales. En el ejemplo: sumar 1 + 5 + 7. La salida deberá ser 13.
   5. Tomando la salida del apartado b, realizar la suma de los pares por un lado y de los impares por otro (usar reduce by key). La salida deberá ser de la forma [(0, 10), (1,13)]
2. Se quiere recoger los datos de una tabla y realizar transformaciones sobre esta. Todos los datos se van a recoger como tipo String. Un ejemplo es:

|  |  |  |  |
| --- | --- | --- | --- |
| **id producto** | **cantidad** | **categoría** | **precio** |
| 1 | 5 | tecnología | 21.5 |
| 2 | 7 | tecnologia | 44.365 |
| 3 | 2 | libros | 10.7 |
| 4 | 4 | libros | 7,61 |
| 5 | 4 | Oficina | 132.97 |
| 6 | 2 | tecnología | 23 |
| 7 | 1 | oficina | 11,52 |
| 8 | 1 | libros | NA |
| 9 | 6 | escalada | 13.23 |
| 10 | 3 | escalada | 35.11 |
| 11 | 2 | tecnología | 4 |

Además, se sabe que:

* Las categorías “tecnología” y “oficina” cuentan con ventajas Premium en sus productos.
* Las categorías “libros” y “electrodomésticos” no cuentan con ventajas Premium.

1. Leer la tabla principal y limpiarla para poder trabajar de manera cómoda con ella.
2. Calcular precio de los productos por categorías. ¿Cómo se controlarán los decimales?
3. Calcular el precio de los productos que cuentan con ventajas Premium. Hay que recordar que se debe realizar de forma general, es decir, hay que considerar que pueden cambiar las categorías que cuentan con ventajas Premium de un día para otro o añadir nuevas. Se aconseja realizar una tabla paramétrica para ello y apoyarse en esta:

|  |  |
| --- | --- |
| **Categoría** | **Premium** |
| tecnología | YES |
| libros | NO |
| oficina | YES |
| electrodomésticos | NO |

1. AVANZADO. A la tabla inicial se le añaden dos columnas adicionales. Una llamada “precio\_premium” y otra “campo\_precio”. La segunda indica qué columna del precio se debe tener en cuenta. Además, hay que considerar que en el futuro pueden añadir nuevas columnas para las distintas cuentas (Estándar, Gold, etc.) y estas pueden venir indicadas en campo\_precio. Se debe realizar el apartado b tomando el valor del precio de la columna indicada en “campo\_precio”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **id producto** | **cantidad** | **categoría** | **precio** | **Precio\_premium** | **Campo\_precio** |
| 1 | 5 | tecnología | 21.5 | 20.02 | precio |
| 2 | 7 | tecnologia | 44.365 | 25.3 | Precio\_premium |
| 3 | 2 | libros | 10.7 | 3.15 | precio |
| 4 | 4 | libros | 7,61 | 5.1 | precio |
| 5 | 4 | Oficina | 132.97 | 124.0 | Precio\_premium |
| 6 | 2 | tecnología | 23 | 23 | precio |
| 7 | 1 | oficina | 11,52 | 11.52 | precio |
| 8 | 1 | libros | NA | 4 | Precio\_premium |
| 9 | 6 | escalada | 13.23 | 12.23 | precio |
| 10 | 3 | escalada | 35.11 | 34.01 | precio |
| 11 | 2 | tecnología | 4 | 4 | precio |