Prática sesión 2 Spark -

Ejercicios

El objetivo de la siguiente práctica es que el alumno ponga en practica las operaciones que sen visto en clase sobre pair RDDs, persistencia, acumuladores, variables...

**Índice**

[Ejercicio propuesto 1 – Pair RDD 2](#__RefHeading___Toc242_3008056549)

[Ejercicio propuesto 2 – Pair RDD 3](#__RefHeading___Toc256_3008056549)

[Ejercicio propuesto 3 – Pair RDD 4](#__RefHeading___Toc262_3008056549)

[Ejercicio propuesto 4 - Pair RDD 5](#__RefHeading___Toc264_3008056549)

[Ejercicio propuesto 5 – Pair RDD 7](#__RefHeading___Toc266_3008056549)

[Ejercicio propuesto 6 - Pair RDD 8](#__RefHeading___Toc360_3008056549)

[Ejercicio propuesto 7 - Persistencia 10](#__RefHeading___Toc362_3008056549)

[Ejercicio propuesto 8 – Check Point 10](#__RefHeading___Toc364_3008056549)

[Ejercicio propuesto 9 – Acumuladores 11](#__RefHeading___Toc366_3008056549)

[Ejercicio propuesto 10 – Variables de transmisión 11](#__RefHeading___Toc368_3008056549)

## Ejercicio propuesto 1 – Pair RDD

A partir de la lista:

animales = Perro, Gato, Loro, Pez, León, Tortuga, Gallina

a) Crea un RDD a partir de esta lista, o desde un fichero con esta información

|  |  |
| --- | --- |
| SOLUCIÓN |  |

b)Convierte el RDD normal en un RDD de pares cuya clave sea la primera letra de animal

|  |  |
| --- | --- |
| SOLUCIÓN |  |

c) Crea otro pair rdd asignando a cada elemento un número no secuencial

|  |  |
| --- | --- |
| SOLUCIÓN |  |

d) ¿Y un índice incremental?

|  |  |
| --- | --- |
| SOLUCIÓN |  |

e) ¿Y un índice incremental que empiece en 100?

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 2 – Pair RDD

Dada una lista de compra:

|  |  |
| --- | --- |
| **Compra** | **Precio** |
| Pan | 3€ |
| Agua | 2€ |
| Azúcar | 1€ |
| Leche | 2€ |
| Pan | 1€ |
| Cereales | 3€ |
| Agua | 0.5€ |
| Leche | 2€ |
| Filetes | 5€ |

**Calcula:**

a) El total que se ha gastado por cada producto

|  |  |
| --- | --- |
| SOLUCIÓN |  |

b) Cuánto es lo máximo que ha pagado por cada producto

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 3 – Pair RDD

Después de ver las agregaciones groupbyKey y combineByKey…

Tenemos las cuentas de las compras de 3 días:

|  |  |  |
| --- | --- | --- |
| **COMPRA DÍA 1** | **COMPRA DÍA 2** | **COMPRA DÍA 3** |
| pan 3€ | pan 1€ | filetes 2€ |
| agua 2€ | cereales 3€ | cereales 1€ |
| azúcar 1€ | agua 0.5€ |  |
| leche 2€ | leche 2€ |  |
| leche 2€ | filetes 5€ |  |
| pan 4€ |  |  |

¿Cómo haríamos ahora para calcular cuanto ha sido lo máximo que hemos pagado por cada producto?

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 4 - Pair RDD

Somos dueños de dos bibliotecas y tenemos en cada sala los siguientes ejemplares del Señor de los anillos:

|  |  |  |
| --- | --- | --- |
| **Sala** | **Libro** | **Ejemplares** |
| Sala A | La Comunidad del Anillo | 3 |
| Sala A | Las dos torres | 2 |
| Sala A | El retorno del Rey | 4 |
| Sala A | El hobbit | 8 |
| Sala B | Prólogo | 3 |
| Sala B | La Comunidad del Anillo | 1 |
| Sala B | Las dos torres | 5 |
| Sala B | El retorno del Rey | 2 |

Queremos saber...

a) Resumen de libros y cantidad por sala

|  |  |
| --- | --- |
| SOLUCIÓN |  |

b) Los libros que están disponibles en ambas salas

|  |  |
| --- | --- |
| SOLUCIÓN |  |

c) De los libros que tenemos en las dos salas, cuantos tendríamos en total de cada uno de ellos entre las 2 salas

|  |  |
| --- | --- |
| SOLUCIÓN |  |

d) Saca una lista de los libros que tenemos, nos da igual el numero de ejemplares

|  |  |
| --- | --- |
| SOLUCIÓN |  |

e) Total de ejemplares ordenado ascendentemente

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 5 – Pair RDD

Siguiendo con el ejercicio de la biblioteca…

Se solicita

a) ¿Cómo podríamos saber el total de ejemplares de 'El hobbit' de la sala\_A?

|  |  |
| --- | --- |
| SOLUCIÓN |  |

b) ¿Que libros no están en la sala A pero si en la B?

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 6 - Pair RDD

Dada la carpeta weblog, con un conjunto de ficheros de log de una web...

a) Carga los logs de la carpeta weblogs

|  |  |
| --- | --- |
| SOLUCIÓN |  |

b) Crea un RDD de pares a partir de estos ficheros con el usuario como clave y un valor 1.

userid1,1

userid2,1

userid3,1

|  |  |
| --- | --- |
| SOLUCIÓN |  |

c) Suma los valores de cada ID de usuario

useid1,4

userid2,9

|  |  |
| --- | --- |
| SOLUCIÓN |  |

d) Muestra los 10 usuarios con máximas apariciones

|  |  |
| --- | --- |
| SOLUCIÓN |  |

e) Crea un rdd de pares donde la clave sea el usuario y el valor una lista de Ips desde donde se ha conectado el usuario

ej:

usuario1,[ip1,ip2,ip3]

usuario2,[ip4,ip5,ip6]

En cvs de cuentas, tenemos información de los clientes, podrías añadir el nombre y apellido del usuario junto a las apariciones de ese usuario?

por ejemplo: Usuario1,(Nombre apellido,30)

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 7 - Persistencia

Vamos a leer un fichero más o menos, por ejemplo el del quijote que teníamos de la sesión anterior y hacerle alguna operación(en el caso de no ser suficiente, podemos ampliarlo o elegir un fichero más grande) , vamos a contar las palabras que son mayores que 10

Esto mismo lo haremos con persistencia y sin persistencia para comprobar el tiempo que tarda cada una, podemos sacar el tiempo antes y después y restarlo para calcular el tiempo que ha tardado

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 8 – Check Point

Vamos a crear un RDD simple por ejemplo:

rdd = 1,2,3,4,5

A este RDD vamos a iterar y cada iteración vamos a sumar 1 a cada elemento del RDD.

Vamos a ir subiendo el número de iteraciones hasta que la memoria ya no de más de sí y nos de error.

Para solucionar este error usaremos Check Points

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 9 – Acumuladores

Dados los ficheros de log de la carpeta weblogs, queremos contar cuantas veces se han enviado datos en formato png,jpg,html,css y txt.

Usa acumuladores para recorrer los ficheros y calcular cuantos hay de cada formato en total entre todos los logs

|  |  |
| --- | --- |
| SOLUCIÓN |  |

## Ejercicio propuesto 10 – Variables de transmisión

Queremos filtrar las peticiones web para solo ver las que aparezcan en una lista de modelos que tenemos (fichero: targetmodels.txt)

Necesitamos filtrar los logs del servidor para solo ver las peticiones de esta lista de modelos, como son muchos ficheros spark los va a repartir por muchos nodos y nos interesa crear una variable con estos datos en cada nodo.

Los logs estarán en una carpeta llamada weblogs de nuestro hdfs (o en local)

(Nota: Repasar funciones any y strip de python)

|  |  |
| --- | --- |
| SOLUCIÓN |  |