UNIVERSIDAD DE JAÉN

Computación Distribuida para la gestión de datos a gran escala

Práctica 3. sample_ncdc.text

Sergio Perea De La Casa

Máster Universitario en Ingeniería Informática

Ejercicio 1

Calcular la máxima temperatura registrada para cada año.

Unset

```
//Leemos el archivo

val practica3 = sc.textFile("../sample_ncdc.txt")


//Split para separar las líneas donde tiene + o -

val test = practica3.map(x => x.split("[+-]"))


//RDD pair (clave, valor) donde la clave es el año y valor es
la temperatura, usando el método de substring para coger una
subcadena en concreto y casteo a entero.

val pair = test.map(x=> (x(0).substring(15,19).toInt,
x(5).substring(1,4).toInt))


//Aplico el máximo
val max = pair.reduceByKey((a,b) => Math.max(a,b))
```

```
scala> val practica3 = sc.textFile("../sample_ncdc.txt")
practica3: org.apache.spark.rdd.RDD[String] = ../sample_ncdc.txt MapPartitionsRDD[1] at textFile at <console>:23

scala> val test = practica3.map(x => x.split("[+-]"))
test: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[2] at map at <console>:23

scala> test.collect()
res0: Array[Array[String]] = Array(Array(0067011990999992016051507004, 68750, 023550FM, 12, 038299999V0203301N00671220001CN9999999N9,
 02300, 99999999999), Array(0043011990999992017051512004, 68750, 023550FM, 12, 038299999V0203201N00671220001CN9999999N9, 01221, 99999
999999), Array(0043011990999992017051518004, 68750, 023550FM, 12, 038299999V0203201N00261220001CN9999999N9, 00111, 99999999999), Arra
y(0043012650999992018032412004, 62300, 010750FM, 12, 048599999V0202701N00461220001CN0500001N9, 01111, 99999999999), Array(00430126509
99992018032418004, 62300, 010750FM, 12, 048599999V0202701N00461220001CN0500001N9, 00781, 99999999999))

scala> val pair = test.map(x=> (x(0).substring(15,19).toInt, x(5).substring(1,4).toInt))
pair: org.apache.spark.rdd.RDD[(Int, Int)] = MapPartitionsRDD[3] at map at <console>:23

scala> pair.collect()
res1: Array[(Int, Int)] = Array((2016,230), (2017,122), (2017,11), (2018,111), (2018,78))

scala> val max = pair.reduceByKey((a,b) => Math.max(a,b))
max: org.apache.spark.rdd.RDD[(Int, Int)] = ShuffledRDD[4] at reduceByKey at <console>:23

scala> max.collect()
res2: Array[(Int, Int)] = Array((2018,111), (2016,230), (2017,122))
```

De esta forma obtenemos como resultado:

- 2018 → 11.1
- 2016 → 23.0
- 2017 → 12.2

Ejercicio 2

Calcula la temperatura media para cada año.

```
Unset
//Uso el groupBy para agrupar por el primer elemento de cada tupla (años), ahora se usa el mapValues para aplicar una función que calcula la media de los valores de cada grupo,donde lo divido entre 10 para ajustar a los grados a nuestro estandar.

val media = pair.groupBy(_._1).mapValues(values => values.map(_._2 / 10.0).sum / values.size.toFloat)
```

```
scala> val media = pair.groupBy(_._1).mapValues(values => values.map(_._2 / 10.0).sum / values.size.toFloat)
media: org.apache.spark.rdd.RDD[(Int, Double)] = MapPartitionsRDD[21] at mapValues at <console>:23

scala> media.collect()
res10: Array[(Int, Double)] = Array((2018,9.45), (2016,23.0), (2017,6.6499999999999995))
```

De esta forma obtenemos como resultado:

- 2018 → 9.45
- 2016 → 23.0
- 2017 → 6.65

Ejercicio 3

Calcular el día, mes y año en el que se alcanza la temperatura máxima.

```
Unset
//Creamos otro rdd par (clave, valor) donde la clave es el
año, dia y el mes y el valor es la temperatura. Usamos para
eso el método substring que coge la cantidad de string que
quieres de acuerdo con la posición del index de los caracteres


scala> val pair_2 = test.map(x=> (x(0).substring(15,23),
x(5).substring(1,4).toInt))
pair_2: org.apache.spark.rdd.RDD[(String, Int)] =
MapPartitionsRDD[24] at map at <console>:23


scala> val max_temp = pair_2.fold("test",0)((x, y) => {if(x._2
< y._2) y else x})
max_temp: (String, Int) = (20160515,230)
```

```
scala> val pair_2 = test.map(x=> (x(0).substring(15,23), x(5).substring(1,4).toInt))
pair_2: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[24] at map at <console>:23
```

```
scala> val max_temp = pair_2.fold("test",0)((x, y) => {if(x._2 < y._2) y else x})
max_temp: (String, Int) = (20160515,230)
```

Tenemos como resultado que en 2016, el mes de mayo, el día 15 fue la temperatura máxima con 23 grados.