Ejercicios de RDDs

Lectura de archivos

1. Lee todos los archivos de texto de un directorio en un único RDD.

- En Scala:

sparkContext porque es lo que maneja la funcionalidad Spark
textFile lee todos los archivos de la ubicación.

- En Python:

importa pyspark porque es lo que maneja la funcionalidad Spark
se crea una instancia de SparkSession con el nombre de “LecturaArchivos”
esta instancia lee mediante textFile todos los archivos de la ubicación que se le ha pasado

2. Lee varios archivos de texto en un solo RDD.

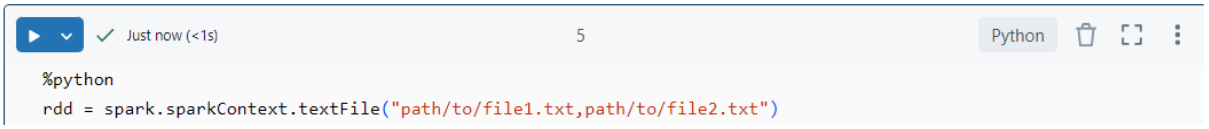
Se leen los archivos de texto y se combinan en un solo RDD.

- En Scala

  Se leen los archivos de texto y se combinan en un solo RDD.

- En Python

  No hace falta importar from pyspark.sql porque ya se ha importado antes



3. Lee todos los archivos de texto que coinciden con un patrón.

Leo todos los archivos de texto que acaben en .txt

- En Scala

```
Just now (1s)                                                    6
val rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/*.txt")

rdd: org.apache.spark.rdd.RDD[String] = C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/*.txt MapPartitionsRDD[7] at textF
ile at command-4335281140898157:1
```

- En Python
  No hace falta importar from pyspark.sql porque ya se ha importado antes

```
Just now (<1s)                                                   7
%python
rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/*.txt")
```

4. Lee archivos de varios directorios en un solo RDD.

- En Scala. Con * me aseguro de leer todos los archivos de los directorios

```
val rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/*,C:/Users/santiago.gil/Desktop/
Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Textfiles_Nuevos/*")

rdd: org.apache.spark.rdd.RDD[String] = C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/*,C:/Users/santiago.gil/Desktop/Bi
g Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Textfiles_Nuevos/* MapPartitionsRDD[13] at textFile at command-4335281140898159:1
```

- En Python
  No hace falta importar from pyspark.sql porque ya se ha importado antes

```
%python
rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/*,C:/Users/santiago.gil/Desktop/Big
Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Textfiles_Nuevos/*")
```

5. Lectura de archivos de texto de directorios anidados en un único RDD.

Leo todos los archivos de texto de directorios anidados en un único RDD.

- En Scala
  Leo todos los archivos de texto en el directorio especifico y sus subdirectorios. “ **/* “ me asegura que se incluyan todos los archivos en los subdirectorios.

```
val rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/**/*")

rdd: org.apache.spark.rdd.RDD[String] = C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/**/* MapPartitionsRDD[19] at textFile at comm
and-4335281140898161:1
```

- En Python

```
%python
rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/**/*")
```

6. Lectura de varios archivos csv en RDD.

Leo todos los archivos de texto de directorios anidados en un único RDD.

- En Scala
  Leo todos los archivos de texto en el directorio especifico y sus subdirectorios. “ **/* “ me asegura que se incluyan todos los archivos en los subdirectorios.

```
val rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Population.csv,C:/Users/santiago.
gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Education.csv")

rdd: org.apache.spark.rdd.RDD[String] = C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Population.csv,C:/Users/santiago.g
il/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Education.csv MapPartitionsRDD[23] at textFile at command-4335281140898163:1
```

- En Python

```
%python
rdd = spark.sparkContext.textFile("C:/Users/santiago.gil/Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Population.csv,C:/Users/santiago.gil/
Desktop/Big Data - Formacion/4.ApacheSparkBasico/3.RDDs/Ejercicios/Education.csv")
```