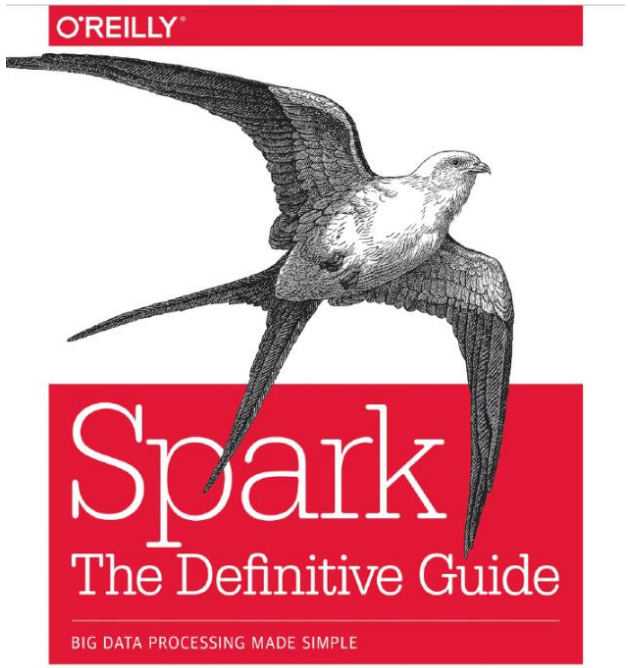
Advanced RDDs



Capítulo Abordado

13. Advanced RDDs

Copyright © 2019 BlueShift – All rights reserved.

<What is Key-Value (Pair) RDD?/>

Simple RDD vs Key-Value RDDs (ou Pair RDDs)

Simple RDDs

Os elementos não possuem chave.

```
val simple_string = "Databricks is awesome!"
val simple_rdd = sc.parallelize(List(simple_string))
```

simple_rdd.collect => **Array[String] = Array(Databricks is awesome!)**

Não possui métodos como agregações e joins.

Precisa ser criada uma chave para tais metodos!

```
val simple_string_withKey = simple_rdd.keyBy(<some_key_func>)
```

Key-Value RDDs

Todos os elementos possuem chave e valor

Se comportam como uma Tupla com 2 valores

```
val simple_string = (1, “First Element“)
val simple_rdd = sc.parallelize(List(simple_string))
val simple_rdd_grouped = simple_rdd.groupByKey()
```

simple_rdd_grouped().collect => **Array[(Int, Iterable[String])] = Array((1,CompactBuffer(Databricks is awesome!)))**

Copyright © 2019 BlueShift – All rights reserved.

<When to Use the Low-Level APIs?/>

Utilidades de Pair RDDs:

- **Agregações e unions**
- **Particionamento customizado (Principal motivo para descer para RDDs!)**

(ex: rdd.partitionBy(<num_partititons>, <custom_func>)

- **Métodos aplicados em partições!**

(ex: foreachPartitions, zipPartitions, glom, repartitionAndSortWithinPartitions)

- **Joins**

Copyright © 2019 BlueShift – All rights reserved.

<Manipulating and Saving RDDs/>

Manipulando RDDs:

- **RDD Programming Guide:**
- https://spark.apache.org/docs/latest/rdd-programming-guide.html
- **Scala Package RDD:**

  https://spark.apache.org/docs/2.3.0/api/scala/index.html#org.apache.spark.rdd.package
- **PySpark Class RDD:**

  https://spark.apache.org/docs/2.1.3/api/python/pyspark.html#pyspark.RDD

Salvando RDDs:

- **Salvando como arquivo texto: rdd.saveAsTextFile("<diretório_final>")**

Copyright © 2019 BlueShift – All rights reserved.