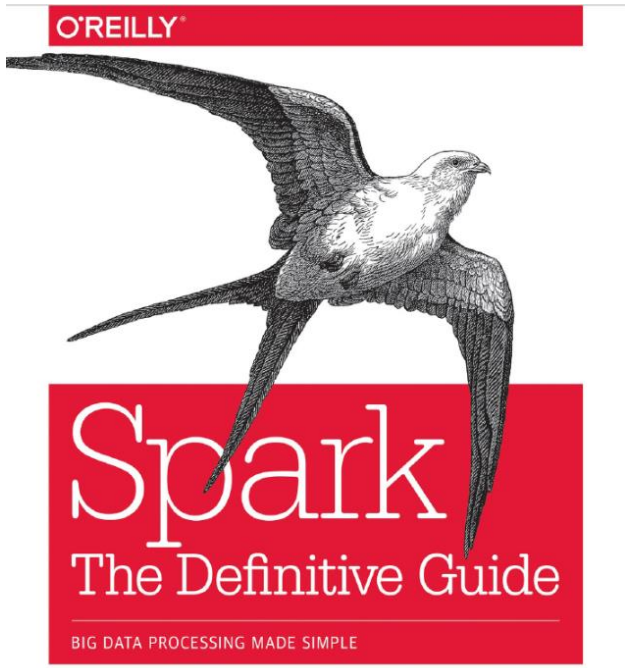
RDD Low-Level API



Capítulo Abordado

12. Resilient Distributed Datasets

(RDDs)

<What Are the Low-Level APIs?/>

Os dois tipos de Low-Level APIs

Manipulação de *dados* distribuídos

Resilient Distributed Datasets

(RDDs)

Distribuição e manipulação de *variáveis* distribuídas

Broadcast Variables

And

Accumulators

Coming Soon: BroadcastHashJoin!

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

Usar RDD quando você precisar:

- De alguma funcionalidade não presente nas APIs de alto nível (DataFrame/Dataset)
- Ter um controle melhor sobre o dado físico distribuído pelo cluster.
- Lidar com dados não estruturados
- Quando você não precisa de um *schema* definido para os seus dados.
- Sustentar um código legado escrito em RDD
- Utilizar broadcast variables ou accumulators (muito difícil!)

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

SparkContext: Ponto de entrada para utilizar Low-Level APIs.

sc = spark.sparkContext

Criando um RDD a partir de uma collection

```
// in Scala
val myCollection = "Spark The Definitive Guide : Big Data Processing Made Simple"
  .split(" ")
val words = spark.sparkContext.parallelize(myCollection, 2)

# in Python
myCollection = "Spark The Definitive Guide : Big Data Processing Made Simple"\
  .split(" ")
words = spark.sparkContext.parallelize(myCollection, 2)
```

Criando um RDD a partir de um DataFrame

```
new_rdd = df.rdd
```

Carregando arquivos de texto em um RDD

Esse método carrega cada linha do arquivo texto como um elemento do RDD:

```
spark.sparkContext.textFile("/some/path/withTextFiles")
```

Já esse outro método, carrega cada arquivo texto como um elemento de chave valor do RDD, onde a chave é o nome do arquivo e o valor o seu conteúdo inteiro.

```
spark.sparkContext.wholeTextFiles("/some/path/withTextFiles")
```

<Manipulating and Saving RDDs/>

Manipulando RDDs:

- **RDD Programming Guide:** https://spark.apache.org/docs/latest/rdd-programming-guide.html
- **Scala Class RDD:** https://spark.apache.org/docs/2.1.3/api/java/org/apache/spark/rdd/RDD.html
- **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>")**