livelessons
video instruction from technology experts

Lesson 4: Spark Internals

4.5 A Day in the Life of a Spark Application

A Day in the Life of a Spark Application

1. Determine RDD lineage: construct DAG

2. Create execution plan for DAG: determine stages

3. Schedule and execute tasks

```
file_rdd.map(line => line.split(""))
      .map(split => (split(0), split(1).toInt))
      .groupByKey()
      .mapValues(iter => iter.reduce(_ + _)).collect()
```

dummy.txt

```
jon 2
mary 3
anna 1
jon 1
jesse 3
mary 5
```

What's in a Task?

```
sc.textFile('file:///dummy.txt')
```

```
map(line => line.split(" "))
```

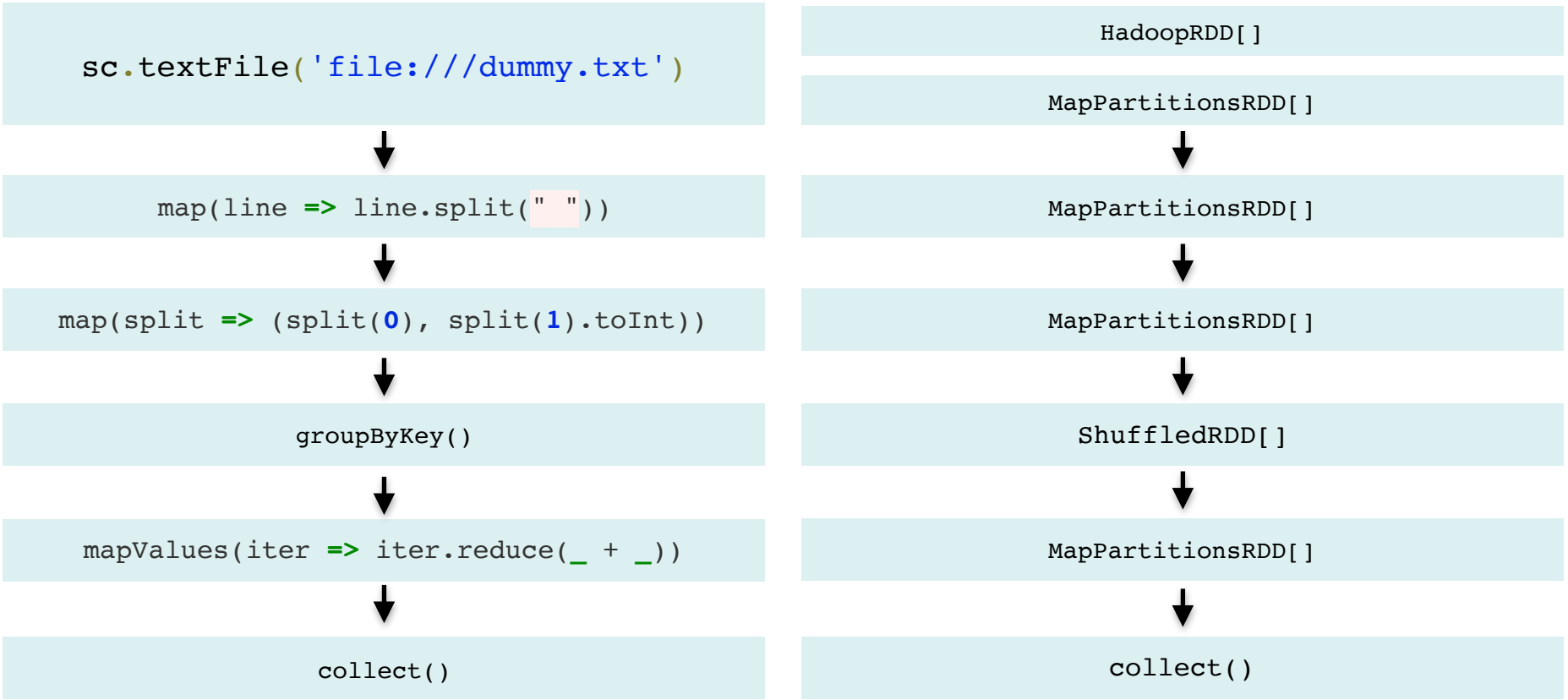
```
map(split => (split(0), split(1).toInt))
```

```
groupByKey()
```

```
mapValues(iter => iter.reduce(_ + _))
```

```
collect()
```

1. Create RDDs



```
scala> file_rdd.partitions
res51: Array[org.apache.spark.Partition] = Array(org.apache.spark.rdd.HadoopPartition@a17, org.apache.spark.rdd.HadoopPartition@a18)

scala> file_rdd.partitions.size
res52: Int = 2

scala> file_rdd.map(line => line.split(" ")).toDebugString
res53: String =
(2) MapPartitionsRDD[65] at map at <console>:24 []
 |  MapPartitionsRDD[23] at textFile at <console>:21 []
 |  file:///Users/jonathandinu/spark-ds-applications/dummy.txt HadoopRDD[22] at textFile at <console>:21 []

scala> file_rdd.map(line => line.split(" ")).map(split => (split(0), split(1).toInt)).toDebugString
res54: String =
(2) MapPartitionsRDD[67] at map at <console>:24 []
 |  MapPartitionsRDD[66] at map at <console>:24 []
 |  MapPartitionsRDD[23] at textFile at <console>:21 []
 |  file:///Users/jonathandinu/spark-ds-applications/dummy.txt HadoopRDD[22] at textFile at <console>:21 []

scala> file_rdd.map(line => line.split(" ")).map(split => (split(0), split(1).toInt)).groupByKey().toDebugString
res55: String =
(2) ShuffledRDD[70] at groupByKey at <console>:24 []
 +-(2) MapPartitionsRDD[69] at map at <console>:24 []
    |  MapPartitionsRDD[68] at map at <console>:24 []
    |  MapPartitionsRDD[23] at textFile at <console>:21 []
    |  file:///Users/jonathandinu/spark-ds-applications/dummy.txt HadoopRDD[22] at textFile at <console>:21 []

scala> file_rdd.map(line => line.split(" ")).map(split => (split(0), split(1).toInt)).groupByKey().mapValues(iter => iter.reduce(_ + _)).toDebugString
res56: String =
(2) MapPartitionsRDD[74] at mapValues at <console>:26 []
 |  ShuffledRDD[73] at groupByKey at <console>:26 []
 +-(2) MapPartitionsRDD[72] at map at <console>:26 []
    |  MapPartitionsRDD[71] at map at <console>:26 []
    |  MapPartitionsRDD[23] at textFile at <console>:21 []
    |  file:///Users/jonathandinu/spark-ds-applications/dummy.txt HadoopRDD[22] at textFile at <console>:21 []

scala> file_rdd.map(line => line.split(" ")).map(split => (split(0), split(1).toInt)).groupByKey().mapValues(iter => iter.reduce(_ + _)).collect()
res57: Array[(String, Int)] = Array((anna,1), (jesse,3), (jon,3), (mary,8))
```