# 课堂笔记

RDD可以分为：

单列元素RDD，和集合类Seq类似

双列元素（Pair）RDD，和集合类Map类似

多列元素RDD呢？

如下为3列元素RDD

1，2，3

1，3，4

2，3，5

转为PairRDD为：

1. （2，3）
2. （3，4）
3. （3，5）

通常多列RDD会转为PairRDD进行操作，这样就可以用PairRDD的Transform和Action操作。

## PairRDD操作

pair RDD如何创建？

1、可以通过sc.parallelize 创建

2、程序中其他RDD转的

pair RDD的元素不是Map，而是Tuple2。

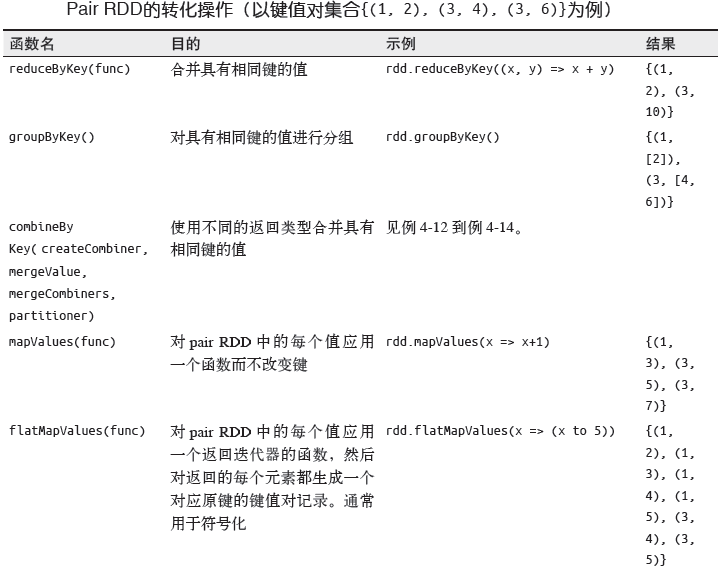
pair RDD 的Transform操作很多，是开发中最核心的部分。

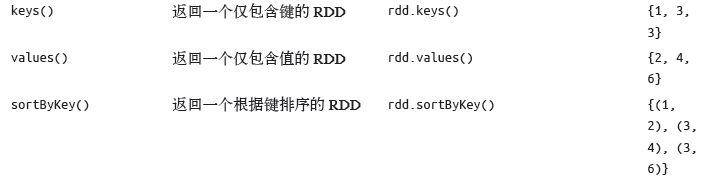
## Pair RDD的Transform操作函数

Pair RDD的Transform操作函数见下方，分为对单个RDD的操作，和2个RDD之间的操作。

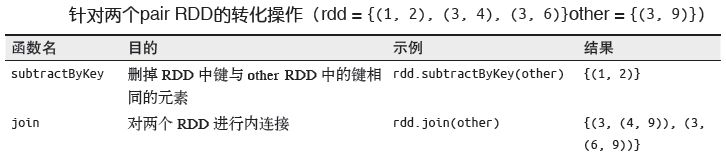
**对单个RDD的操作函数有：**

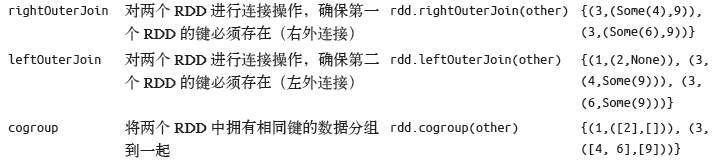
map() 、reduce() 这些和Mapreduce框架没联系。





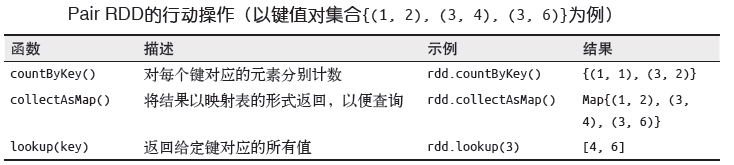
**两个RDD之间的Transform操作：**





后续将对每个重要函数进行细讲。

## Pair RDD的Action操作函数



countByKey : select key,count(value) group by key

## PairRDD的函数练习

pairRDD也是RDD，可以使用前边讲过的RDD的Transform和Action函数，只是用法不同了。

pairRDD如何使用map() , filter() ？

reduceByKey() 与reduce() 类似，做sum类

mapValues(func) 函数，功能类似于map{case (x, y): (x,func(y))}

## 案例，计算每个键对应的平均值

如果计算一个pairRDD每个**key对应的平均值**？

比如[(3,4),(2,4),(3,2),(2,6)] 最后得出[(3,3),(2,5)]

Select sum(value)/count(key)

From group by key

转换过程如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| pairRDD |  |  | mapValues |  |
|  |  |  |  |  |
| key | value |  | key | value |
| 3 | 4 |  | 3 | (4,1) |
| 2 | 4 |  | 2 | (4,1) |
| 3 | 2 |  | 3 | (2,1) |
| 2 | 6 |  | 2 | (6,1) |
|  |  |  |  | sum(\_1) 即为sum(value) sum(\_2) 即为count(key)  Avg=sum(\_1)/sum(\_2) |
| reduceByKey |  |  | mapValues |  |
| key |  |  | key |  |
| 3 | (6,2) |  | 3 | 3 |
| 2 | (10,2) |  | 2 | 5 |

## Join

join 对两个RDD 进行内连

leftOuterJoin 对两个RDD 进行左连接

rightOuterJoin 对两个RDD 进行右连接

效果同sql里，可以两个rdd想象成2张数据表(包含2个字段)。

第6课预告：

用电商企业真实数据，实现一个简单流量统计

第7课预告：

Spark Sql

1、读数据，注册临时表，执行sql

2、spark-sql 直接执行hive ql作业

spark-sql -e “select ... from ” hive on Spark

Hive -f -e -i

spark-sql -e -f -i

一天净增量100亿行，13-15T

200多个节点（相当于 苏宁易购 每天流量数据 ）