# RDD编程

## Rdd创建

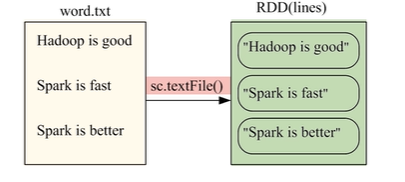
Spark core的核心，就在于rdd编程

1. 文件系统加载去创建rdd

.textfile()方法，可以从本地，也可以从云端如亚马逊amazon，也可以从hdfs中去加载

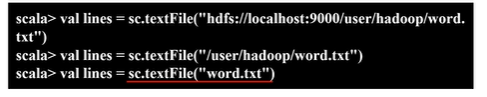




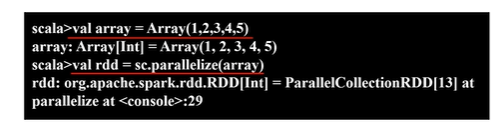




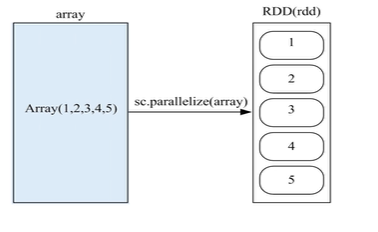
与下面的等价，不过是相对路径（用户名是hadoop，也即是在当前的用户的相对目录下）



1. 通过集合或者数组创建rdd



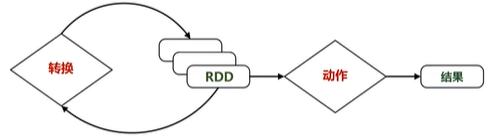
Parallslize（）方法



## Rdd操作(转换)

Rdd操作分类两类，一类是action 一类是transformation

注意这里有一个惰性机制

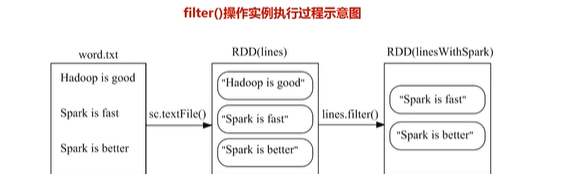


转换操作



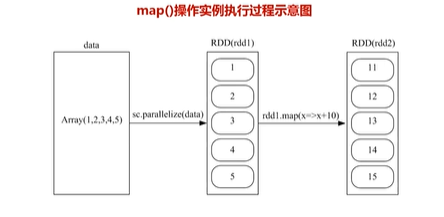
Filter（func）



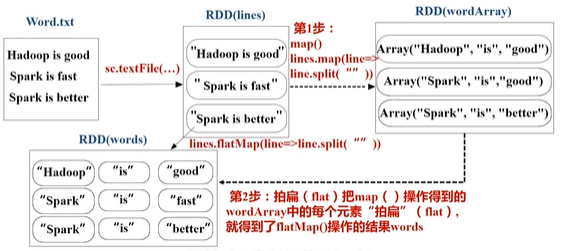


过滤

Map(func) 映射

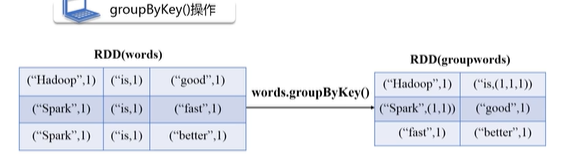


Flatmap（function）

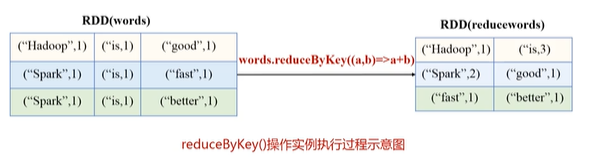


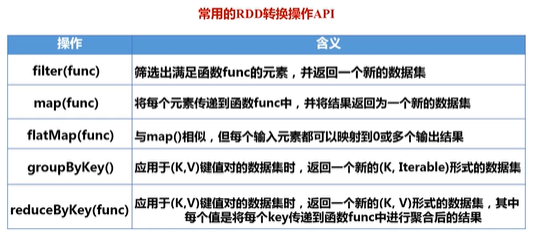
文本文件词频统计的常用方法

Groupby



Reducebykey(func)





## Rdd操作（动作）

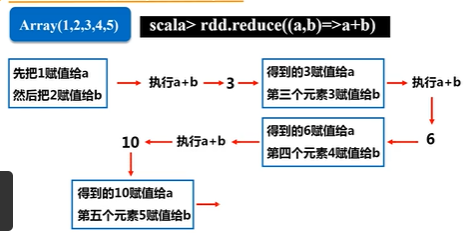
Count()

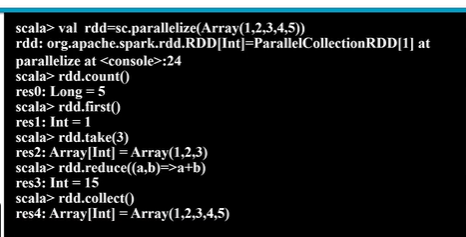
Collect()

First()

Take







惰性机制：

转换类的操作都不会去执行的，只有遇见动作类型的操作，才会去执行。