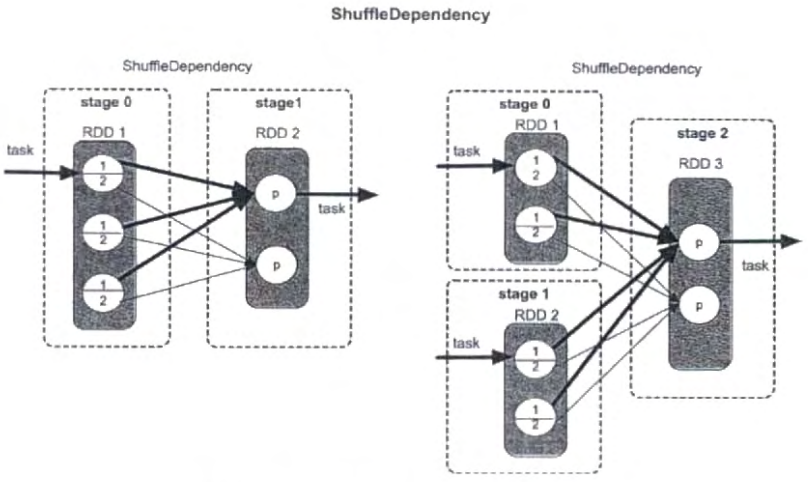

图 4.13　ShuffleDependency 的 stage 划分原则

OneToOneDependency 类型的操作

典型操作	语义特点
map(), mapValues(), filter(), filterByRange(), flatMap(), flatMapValues(), sample(), sampleByKey(), glom(), zipWithIndex(), zipWithUniqueId() 等	针对每个 record 执行 func 操作，输出一个或多个 record
mapPartitions(), mapPartitionsWithIndex() 等	针对一个分区中的数据进行操作，输出一个或多个 record

图 4.14 展示了 flatMap() 和 mapPartitionsWithIndex() 操作的 stage 和 task 划分图，这两个操作都生成了一个 stage，stage 中不同颜色的箭头表示不同的 task。每个 task 负责处理一个分区，进行流水线计算，且计算逻辑清晰。这两个操作唯一不同的是 flatMap() 每读入一条 record 就处理和输出一条，而 mapPartitionsWithIndex() 等到全部 record 都处理完后再输出 record。图 4.14 右图中的 mapPartitionsWithIndex() 是计算每个分区中奇数的和及偶数的和。

RangeDependency 类型的操作

典型操作	语义特点
在一般情况下的 union() 操作是指参与 union() 的 RDD 的 partitioner 不相同，详见第 3 章介绍的 union() 操作	将多个 RDD 的分区直接合并在一起