02

第2章 Spark的设计与运行原理

授课教师：[李阳源]
15138617705@163.com
主页：https://sites.google.com/view/dalguri



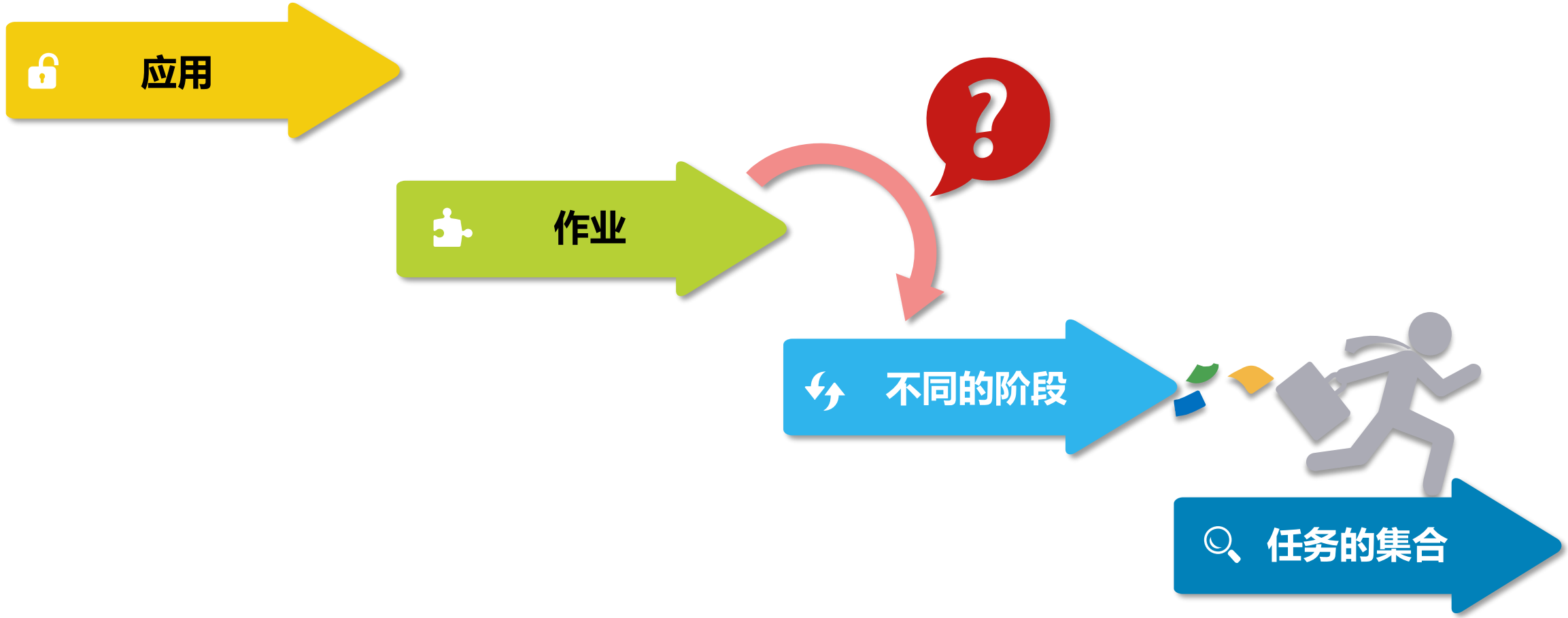


2.3.4 RDD运行原理

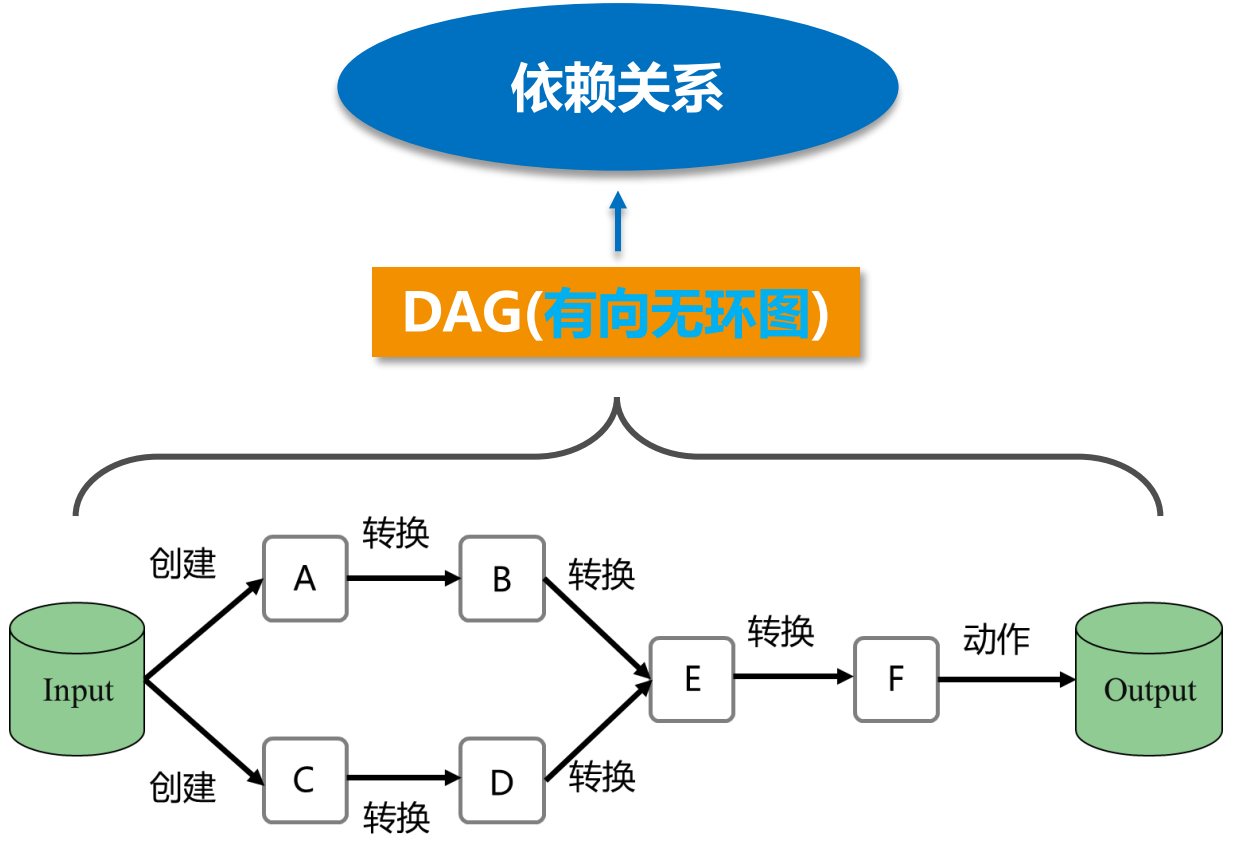
提纲

1. RDD设计背景
2. RDD概念
3. RDD特性
4. RDD之间的依赖关系
5. 阶段的划分
6. RDD的运行过程

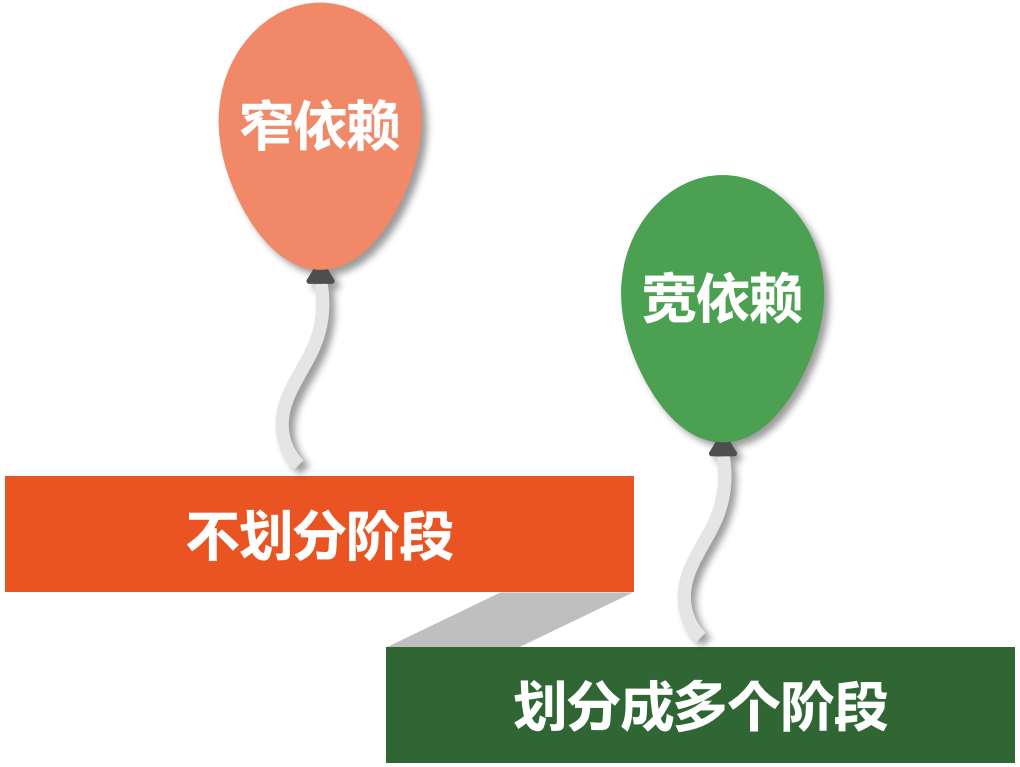
2.3.4 RDD运行原理（RDD之间的依赖关系）



2.3.4 RDD运行原理（RDD之间的依赖关系）



2.3.4 RDD运行原理（RDD之间的依赖关系）



2.3.4 RDD运行原理（RDD之间的依赖关系）



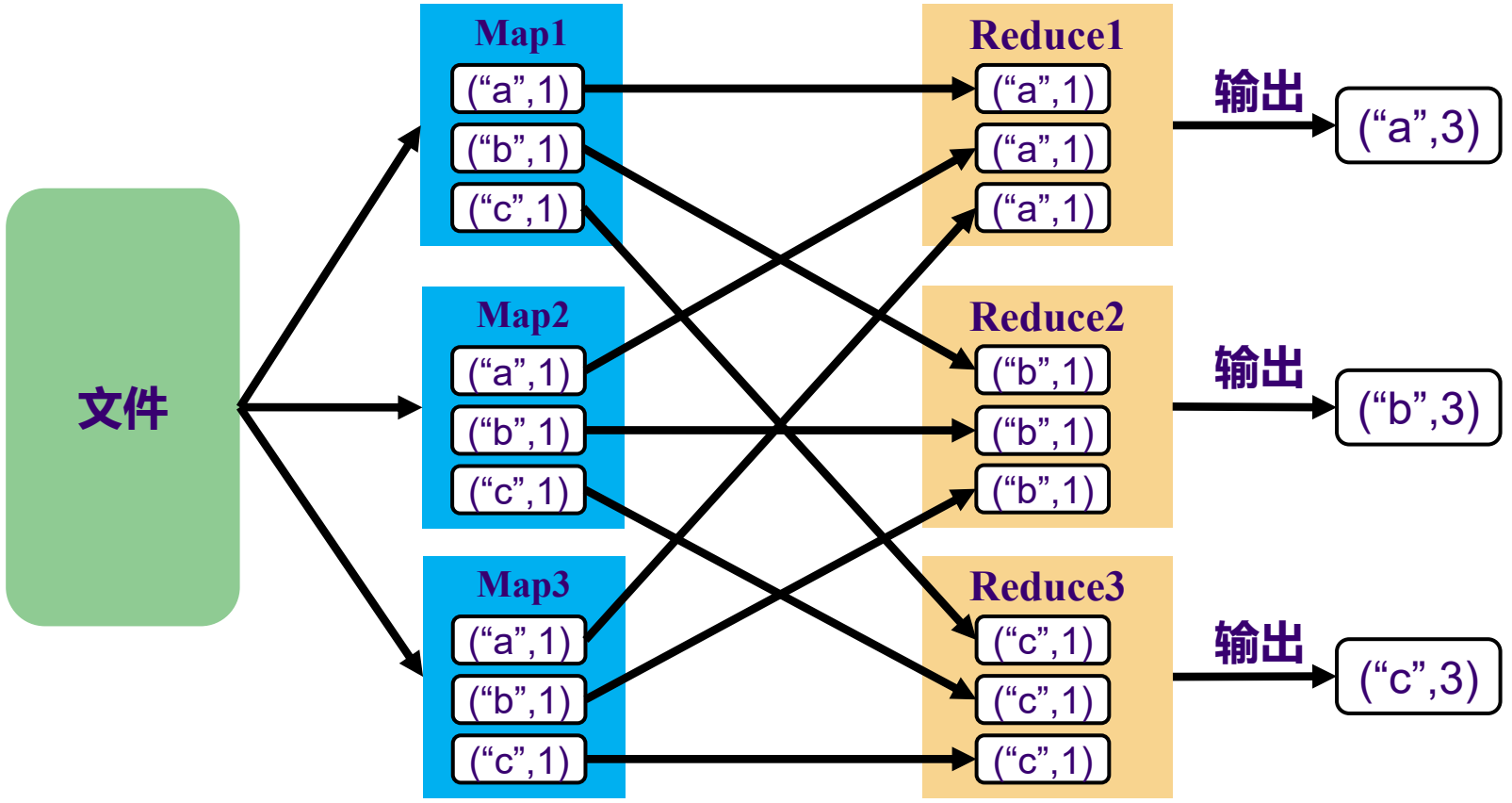
宽依赖



Shuffle 操作

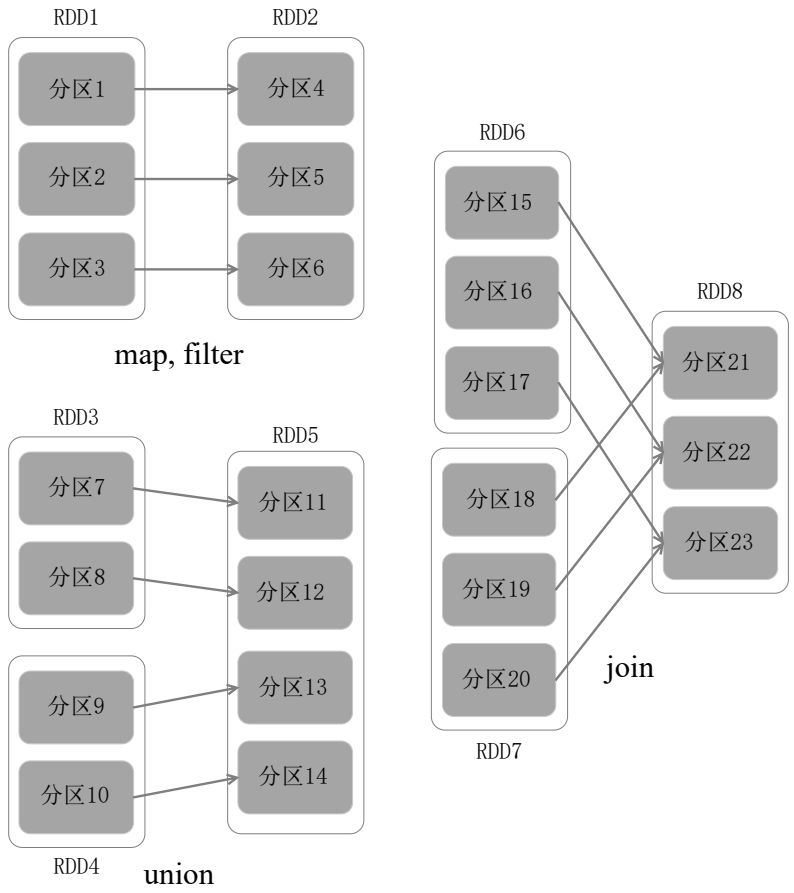
洗牌

2.3.4 RDD运行原理（RDD之间的依赖关系）



大量网络数据分发

2.3.4 RDD运行原理（RDD之间的依赖关系）



2.3.4 RDD运行原理（RDD之间的依赖关系）



窄依赖

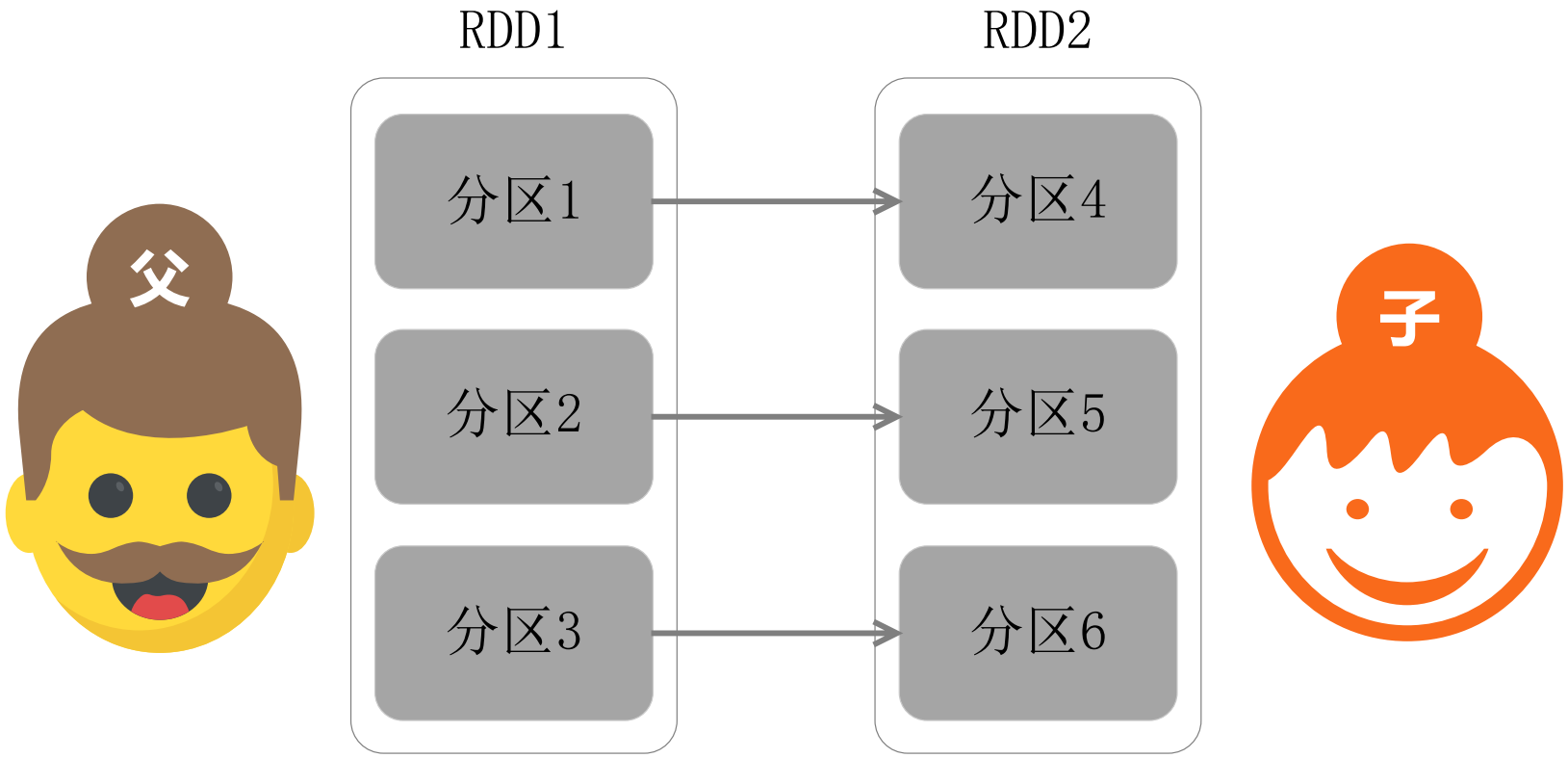


对应



父RDD的分区　　子RDD的分区

2.3.4 RDD运行原理（RDD之间的依赖关系）

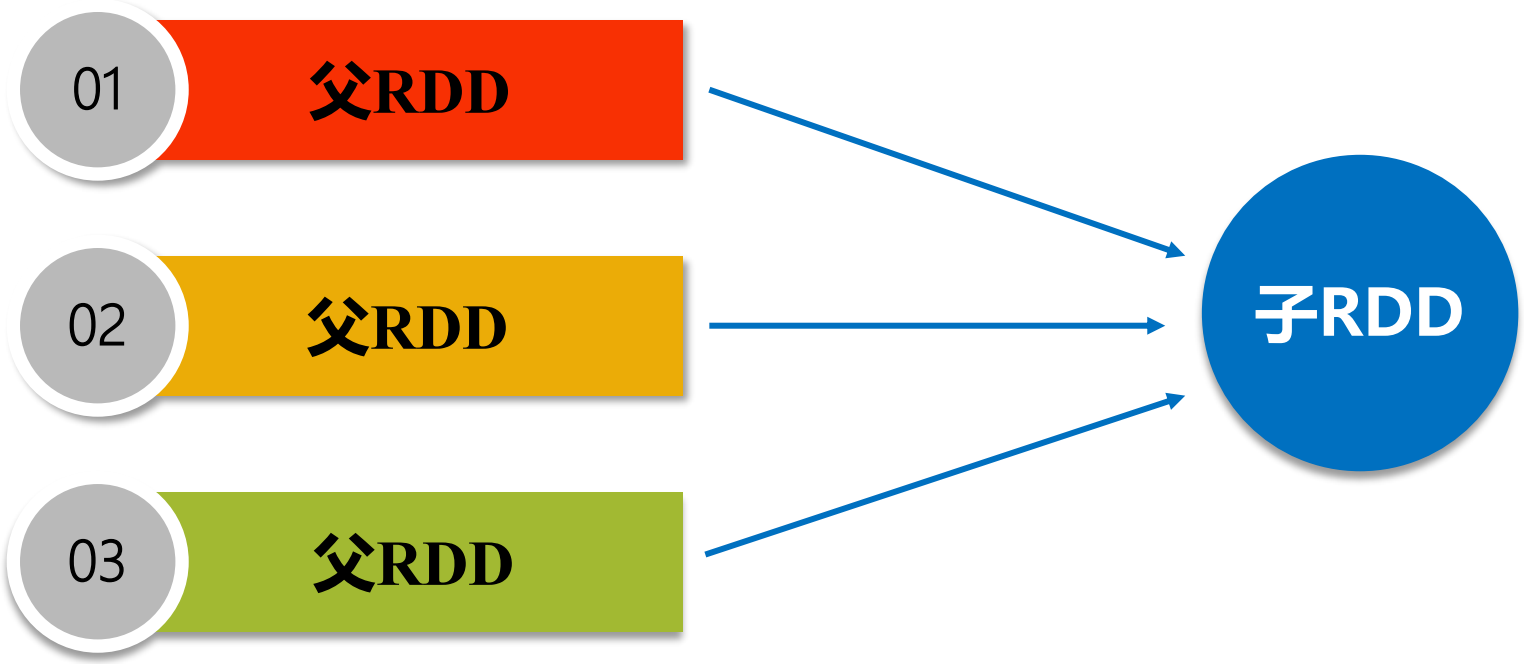


(a)窄依赖

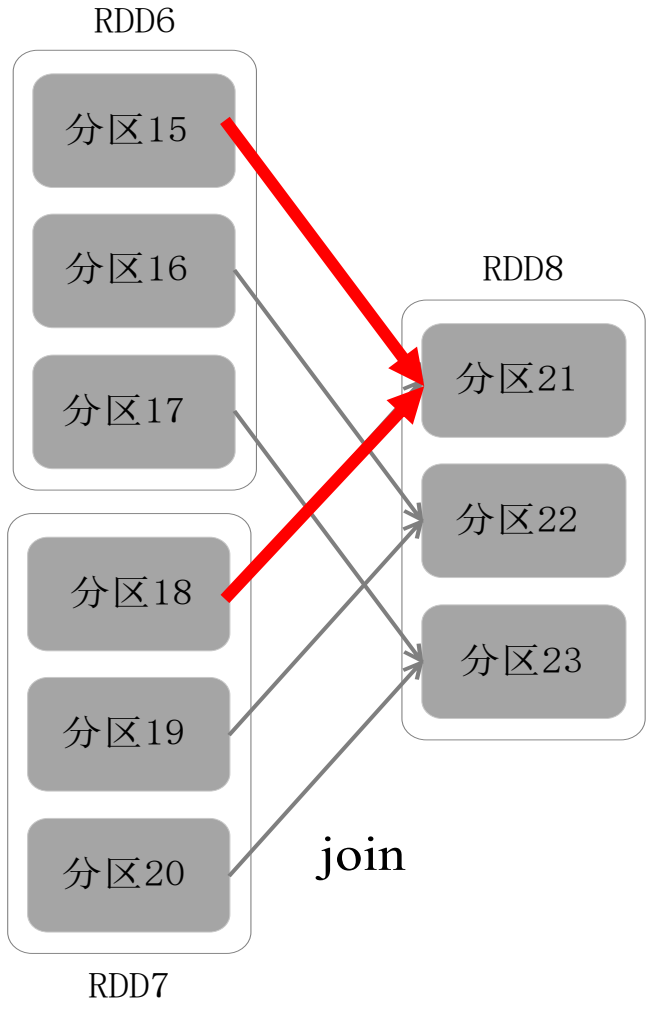
2.3.4 RDD运行原理（RDD之间的依赖关系）



窄依赖



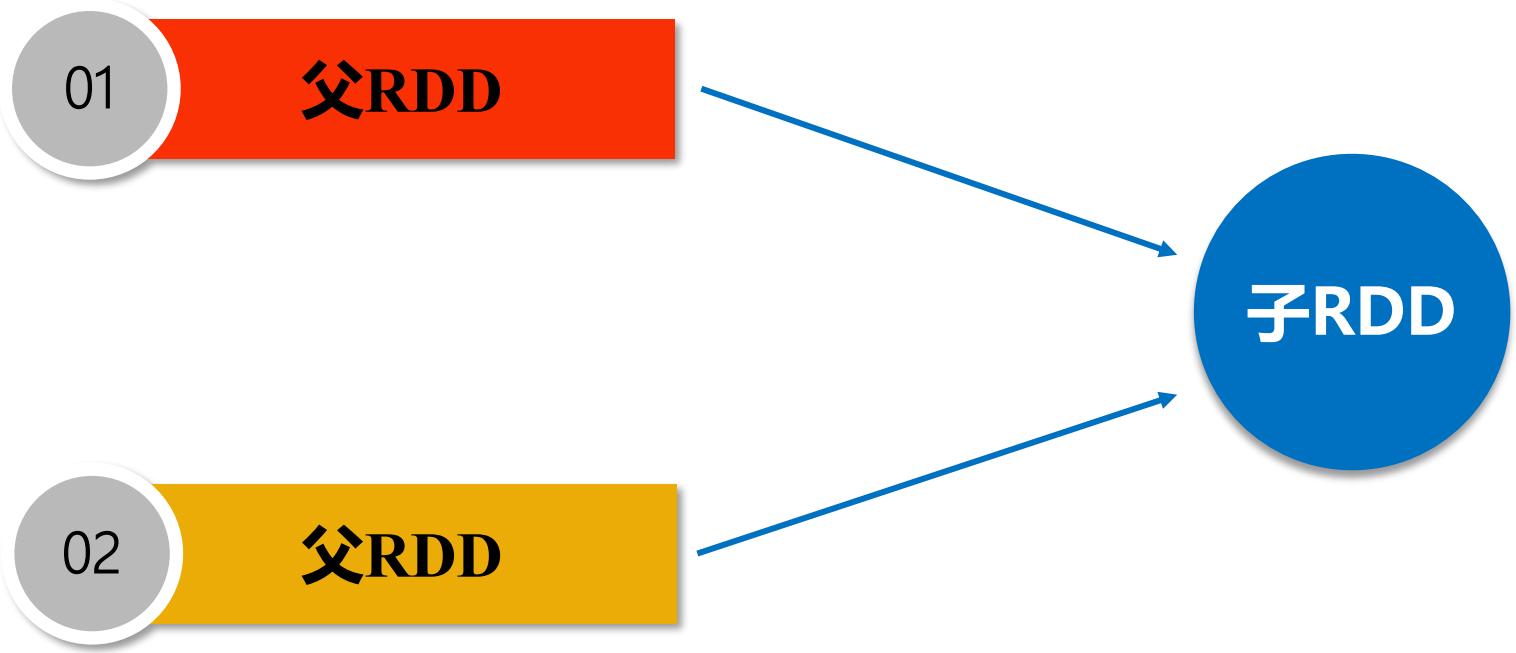
2.3.4 RDD运行原理（RDD之间的依赖关系）



2.3.4 RDD运行原理（RDD之间的依赖关系）

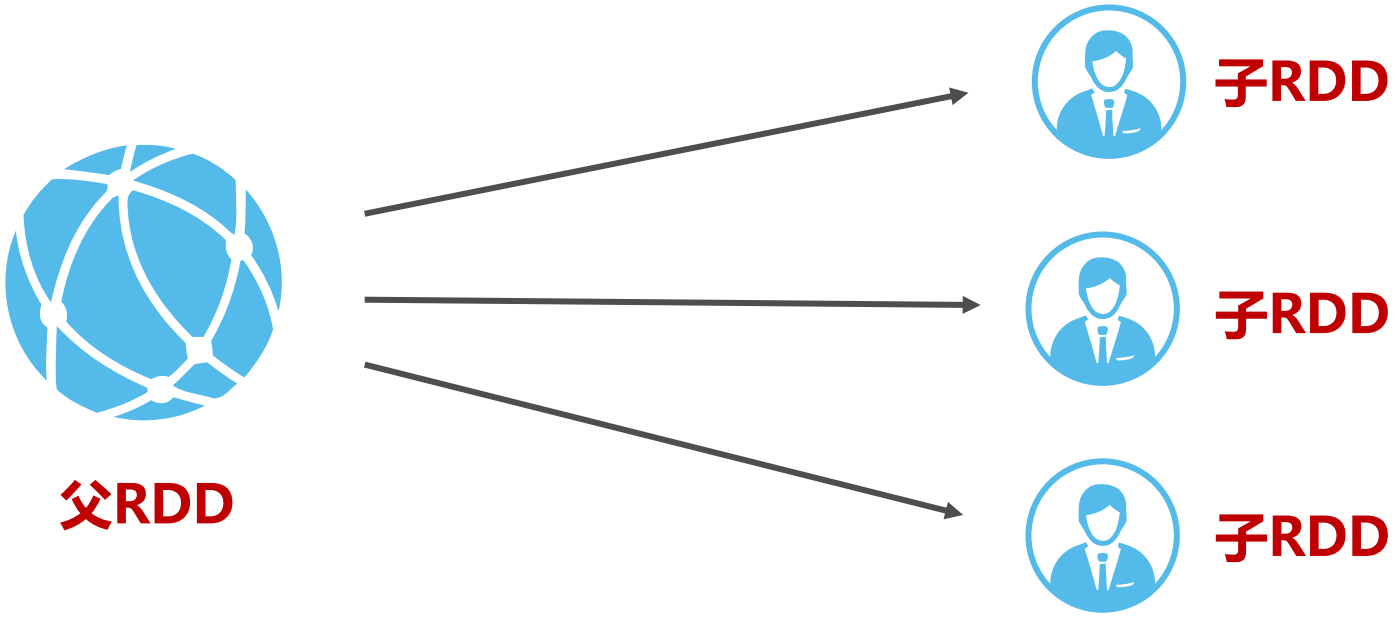


窄依赖

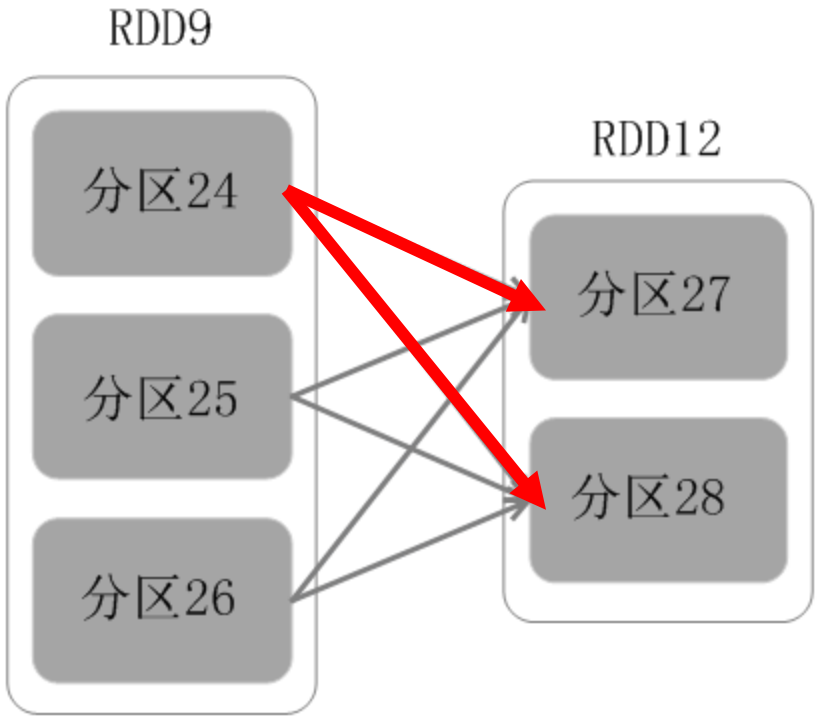


2.3.4 RDD运行原理（RDD之间的依赖关系）

宽依赖



2.3.4 RDD运行原理（RDD之间的依赖关系）

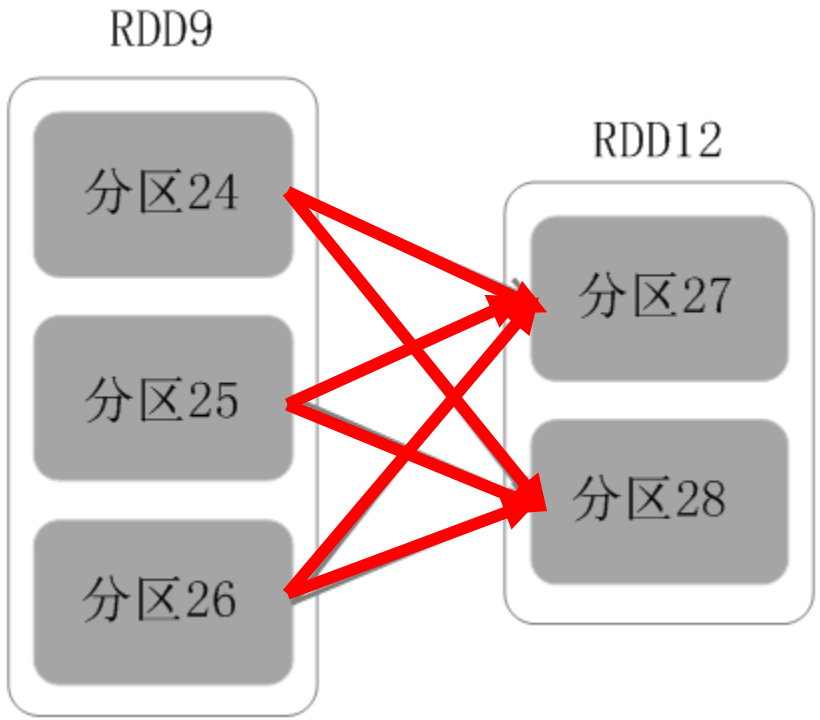


Shuffle 操作

2.3.4 RDD运行原理（RDD之间的依赖关系）



宽依赖

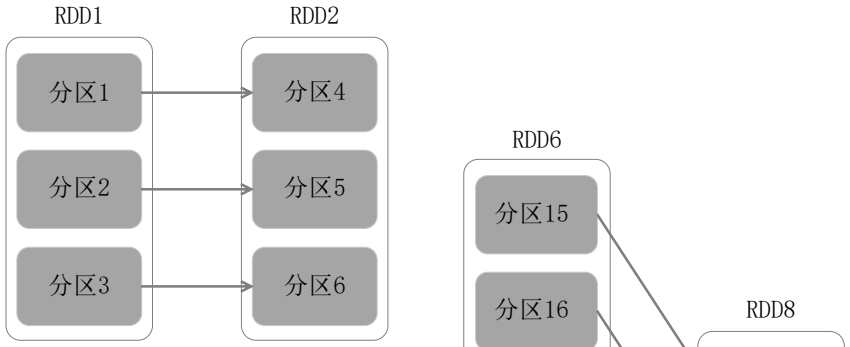


在网络当中大规模地来回传输数据

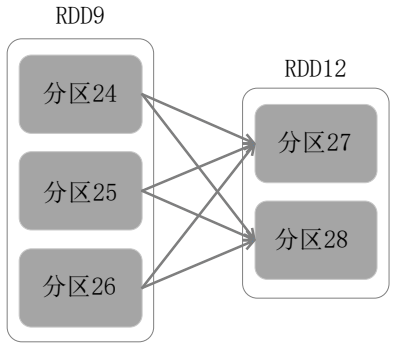
不同节点之间互相传数据

Shuffle 操作 **洗牌操作**

2.3.4 RDD运行原理（RDD之间的依赖关系）



(a)窄依赖



(b)宽依赖

2.3.4 RDD运行原理理解测验（选择题）



문제 1: 以下哪个操作属于窄依赖?

A. groupByKey()

B. reduceByKey()

C. map()

D. join()

2.3.4 RDD运行原理理解测验（选择题）



문제 2: 宽依赖会导致什么操作?

A. 本地计算

B. 内存缓存

C. Shuffle操作

D. 数据压缩

2.3.4 RDD运行原理理解测验（选择题）



문제 3: 在RDD依赖关系中，DAG代表什么?

A. 数据聚合图

B. 有向无环图

C. 分布式算法图

D. 动态分配图

2.3.4 RDD运行原理理解测验（选择题）



문제 4: 以下关于Shuffle操作的描述，哪个是正确的?

A. Shuffle操作只在本地节点内进行

B. Shuffle操作不影响性能

C. Shuffle操作需要大量网络数据传输

D. Shuffle操作只在窄依赖中发生

2.3.4 RDD运行原理理解测验（选择题）



문제 5: union操作属于以下哪种依赖类型?

A. 宽依赖

B. 窄依赖

C. 混合依赖

D. 无依赖

2.3.4 RDD运行原理理解测验（选择题）



문제 6: 在窄依赖中，父RDD的一个分区对应子RDD的几个分区?

A. 0个

B. 1个

C. 多个

D. 不确定

2.3.4 RDD运行原理理解测验（选择题）



문제 7: 以下哪个操作会产生宽依赖?

A. filter()

B. map()

C. groupByKey()

D. mapPartitions()

2.3.4 RDD运行原理理解测验（选择题）



문제 8:在Spark中，阶段（Stage）的划分是基于什么?

A. 数据大小

B. 节点数量

C. 宽依赖的位置

D. 内存使用量

2.3.4 RDD运行原理理解测验（选择题）



문제 9: 以下关于filter操作的描述，哪个是正确的?

A. filter操作会产生Shuffle

B. filter操作属于宽依赖

C. filter操作属于窄依赖

D. filter操作不改变分区数量

2.3.4 RDD运行原理理解测验（选择题）



문제 10: 在宽依赖中，为什么需要进行Shuffle操作?

A. 为了提高计算速度

B. 为了节省内存

C. 为了将相关数据重新分布到同一分区

D. 为了减少网络传输



Thank you