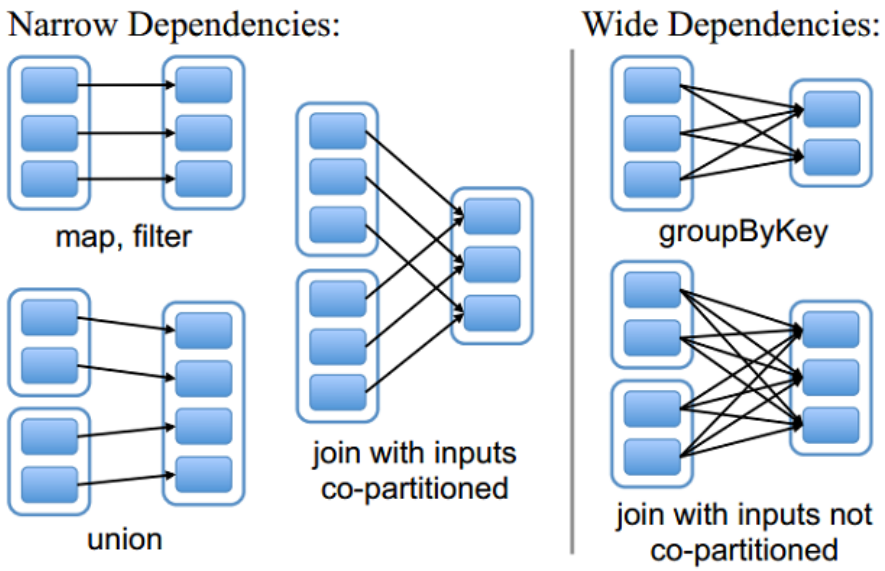
1. RDD 的依赖关系

RDD 和它依赖的父 RDD 之间的关系有两种，即窄依赖（narrow dependency）和宽依赖（wide dependency）。



1.1. 窄依赖

窄依赖：父 RDD 的 Partition 最多被子 RDD 的一个 Partition 依赖

总结：窄依赖形象比喻为**独生子女（一对一、多对一）**

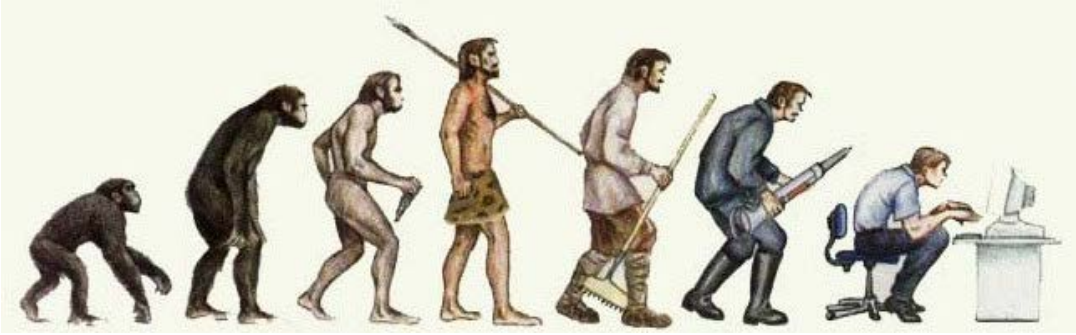
1.2. 宽依赖

宽依赖：子 RDD 的多个 Partition 会依赖父 RDD 的同一个 Partition

总结：宽依赖形象比喻为**超生（一对多）**

1.3. Lineage

Spark 将创建 RDD 的一系列 Lineage（即血统）记录下来，以便恢复丢失的分区数据。RDD 的 Lineage 会记录 RDD 的元数据信息和转换行为，当该 RDD 的部分分区数据丢失时，可以根据这些信息重新计算并恢复丢失的分区数据。



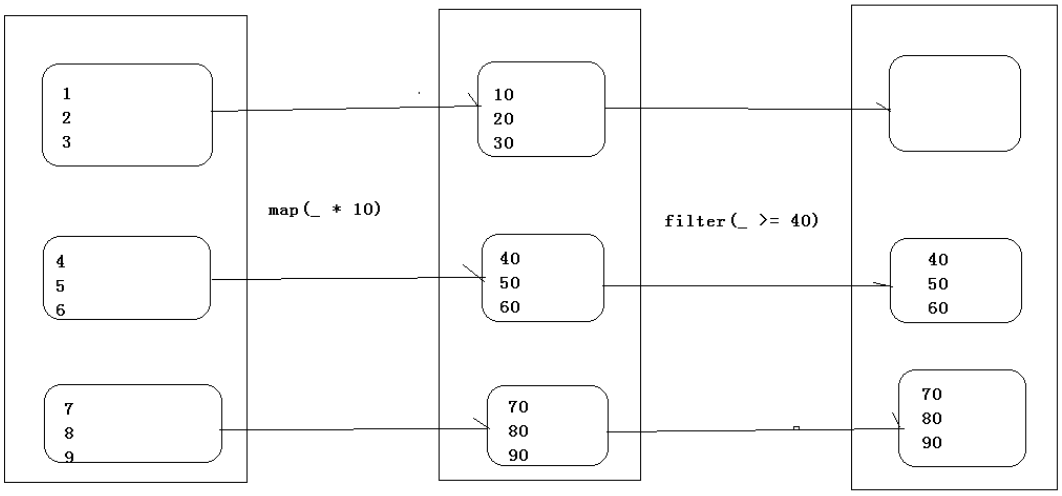
补充：

1、 窄依赖不发生 shuffle，宽依赖会发生 shuffle。

2、 join 大多数情况下是宽依赖；若已分组，且没有改变分区数量，则为窄依赖。

3、 发生 shuffle 时会划分 stage。

窄依赖：



特殊的窄依赖：