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





第2章 Spark的设计与运行原理





2.3.3 RDD运行原理



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

2.3.3 RDD运行原理(RDD概念、操作和特性)



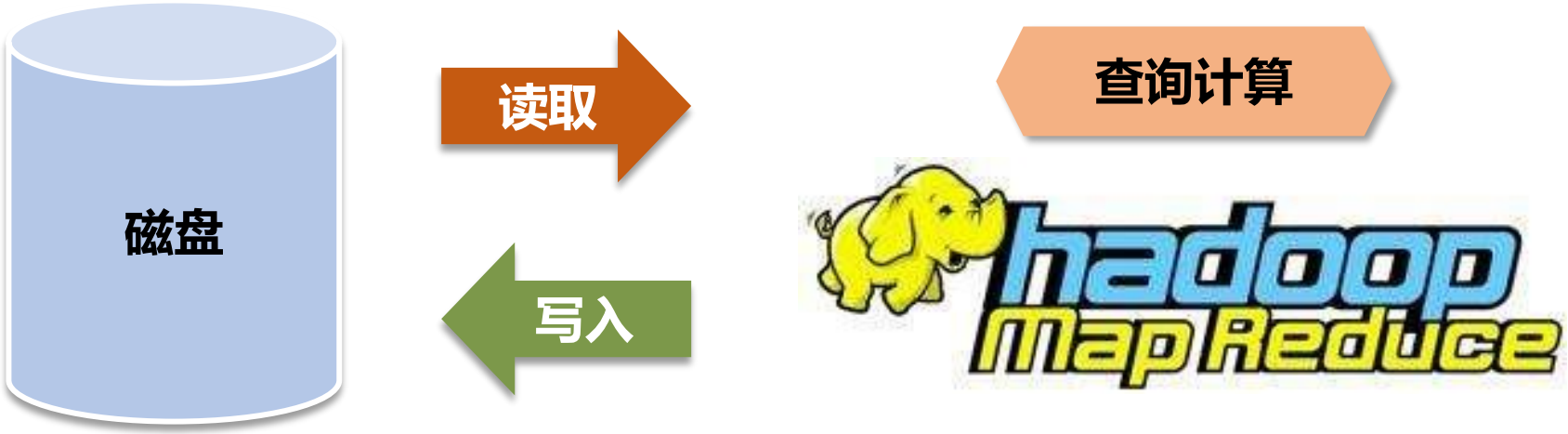


~~处理~~

迭代场景

- 逻辑斯蒂回归
- 模拟退火算法
- 遗传算法

2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)



迭代

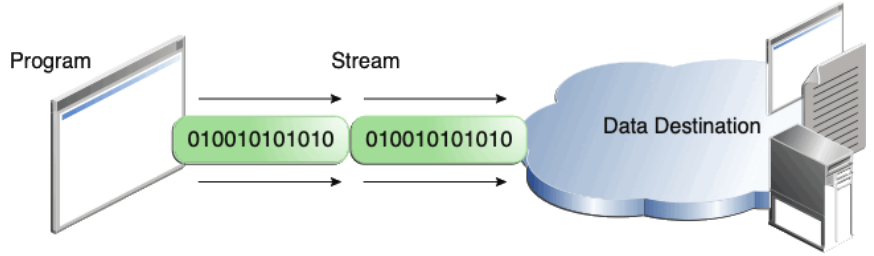
反复读写工作子集



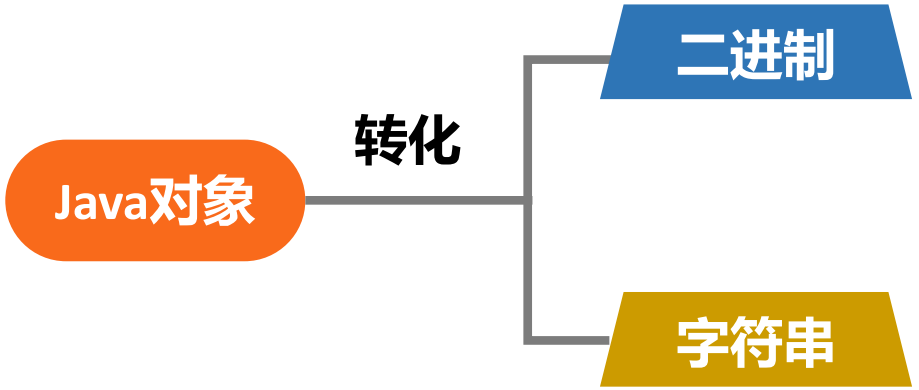
磁盘IO开销

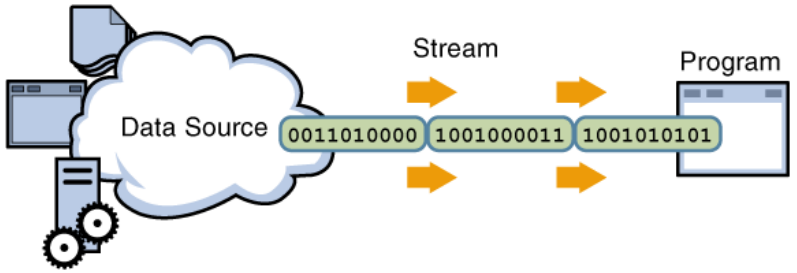
序列化和反序列化开销

2.3.3 RDD运行原理(RDD概念、操作和特性)

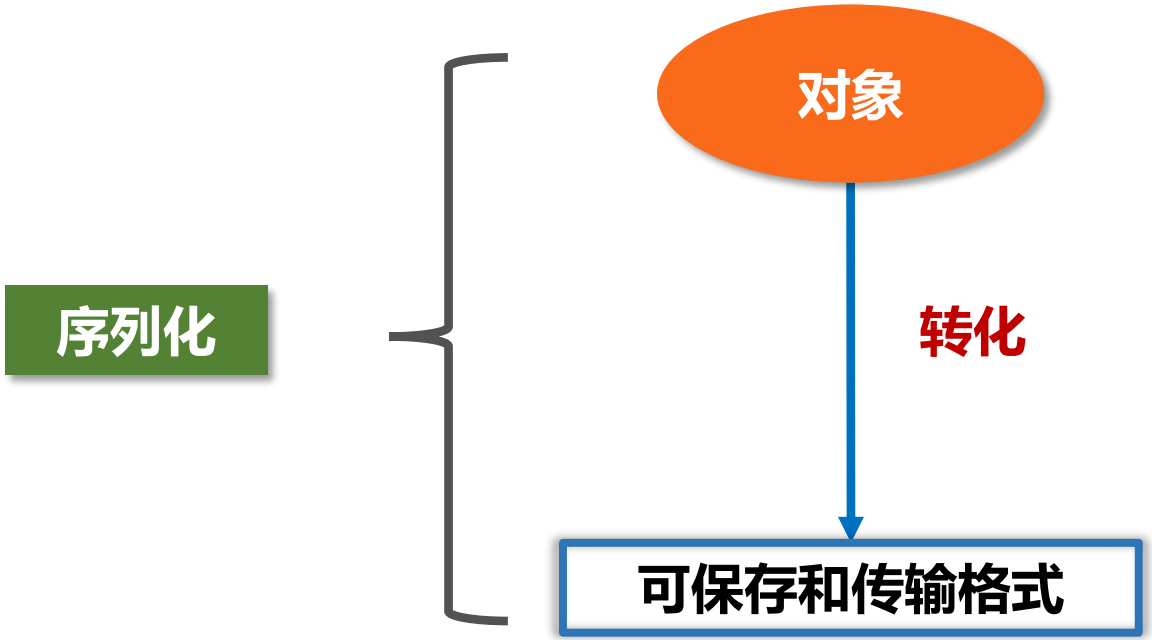


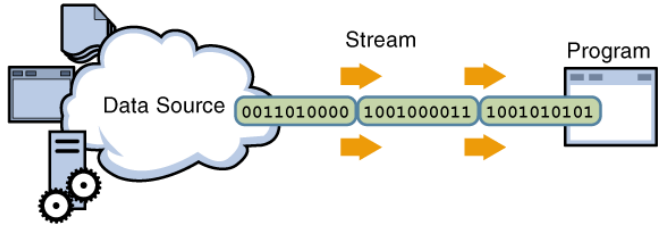






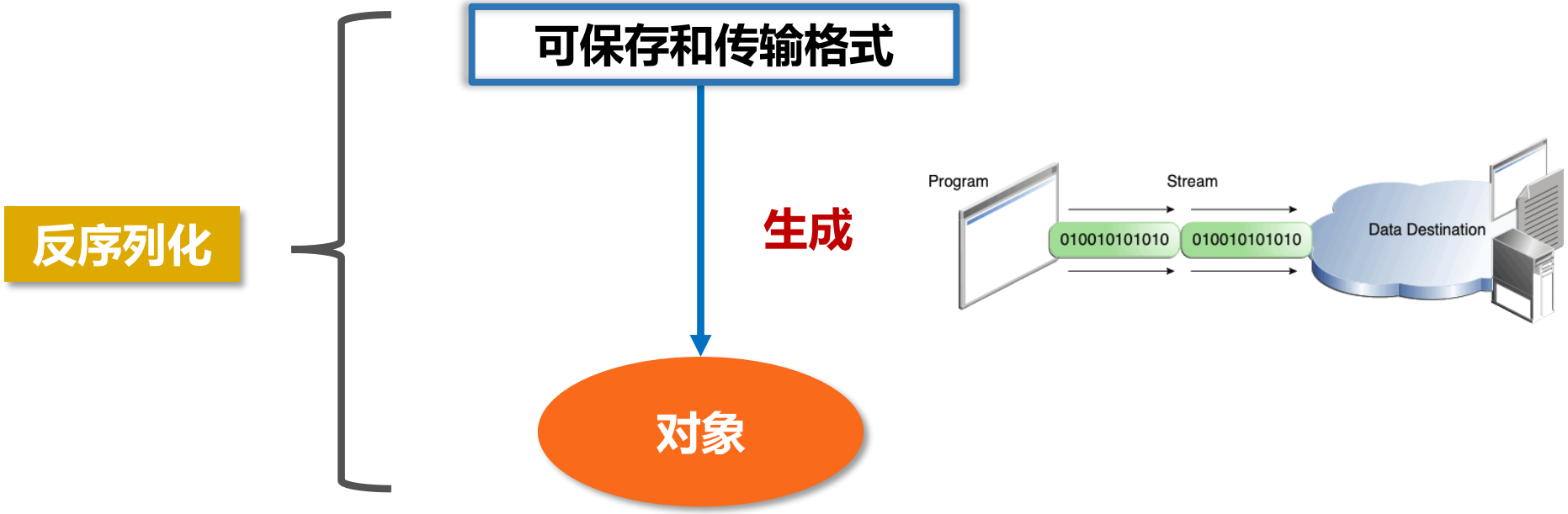
2.3.3 RDD运行原理(RDD概念、操作和特性)





2.3.3 RDD运行原理(RDD概念、操作和特性)



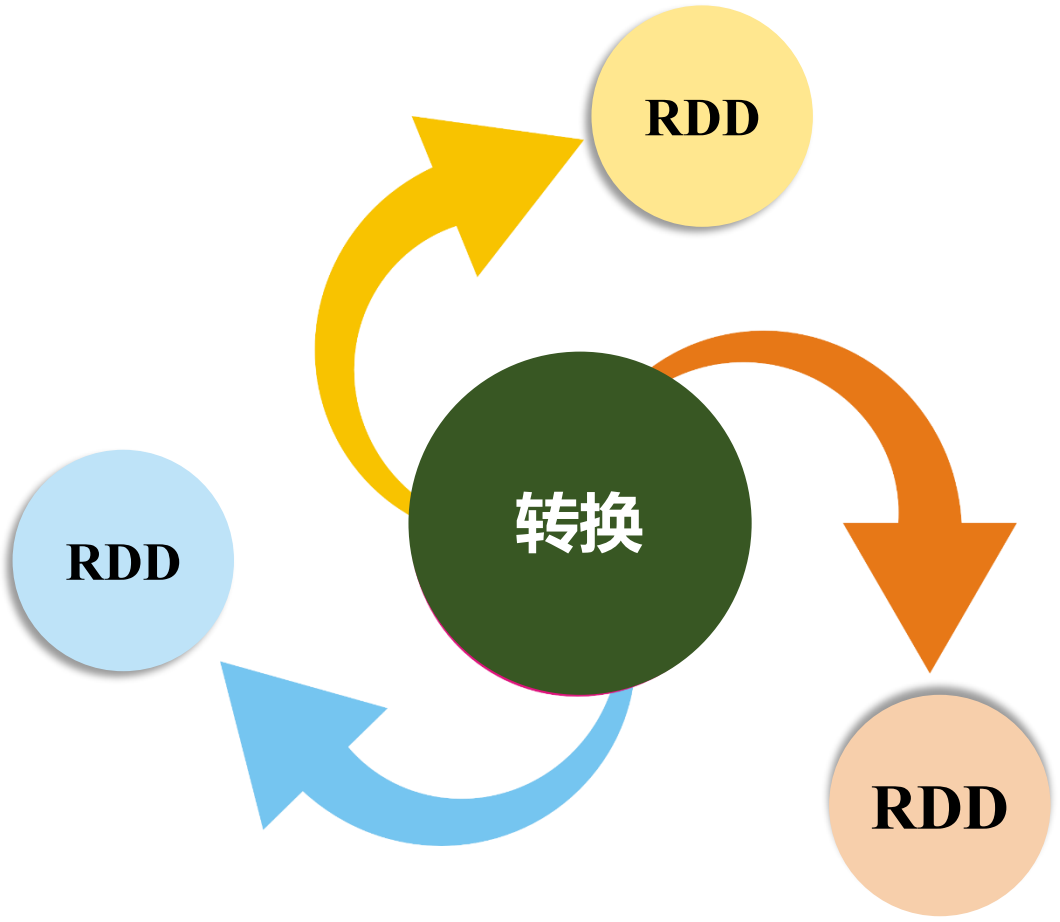


2.3.3 RDD运行原理(RDD概念、操作和特性)

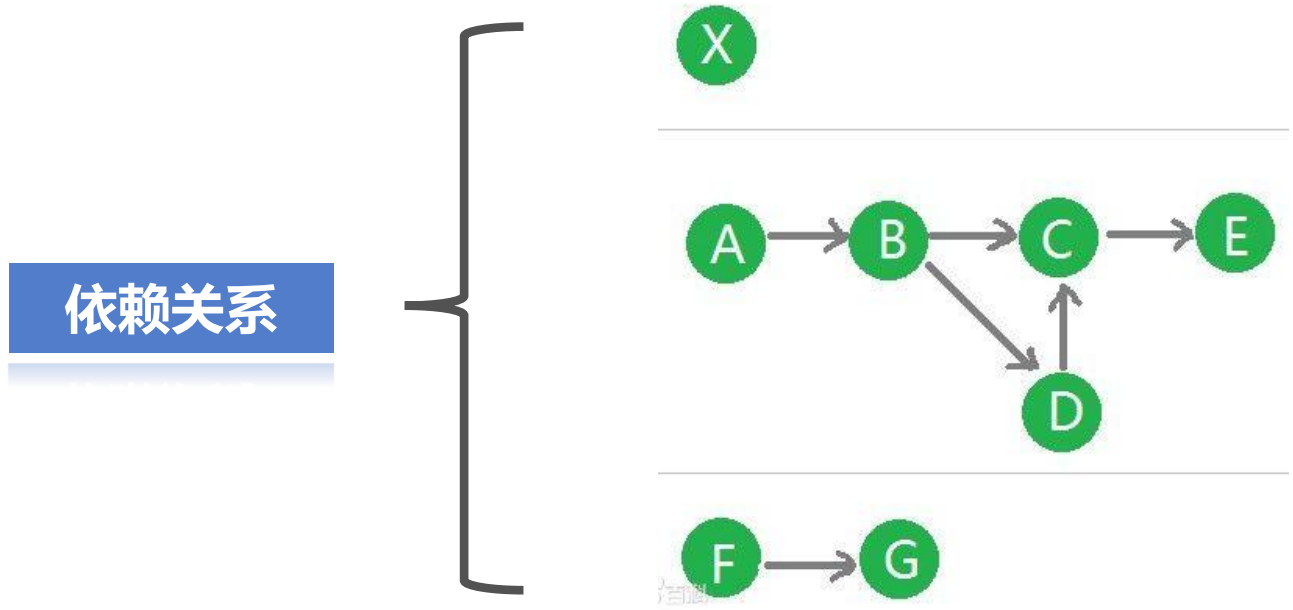


提供了抽象的数据结构

2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)

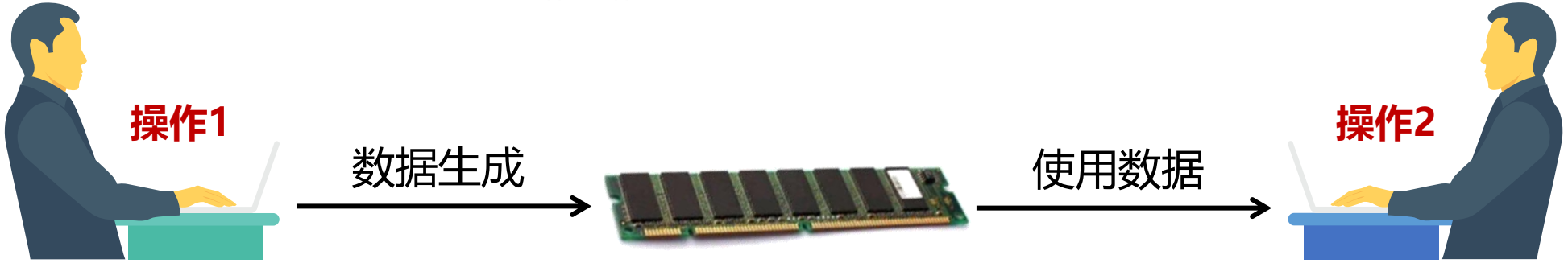


DAG图

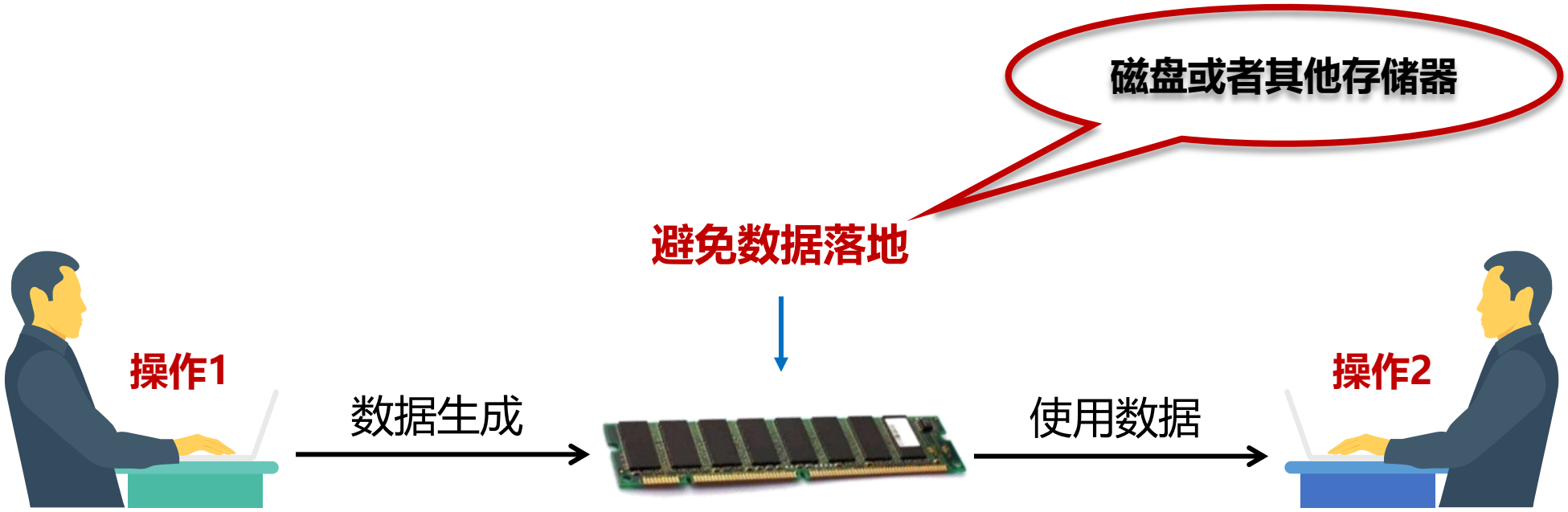
2.3.3 RDD运行原理(RDD概念、操作和特性)



管道化（流水线化）



2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)



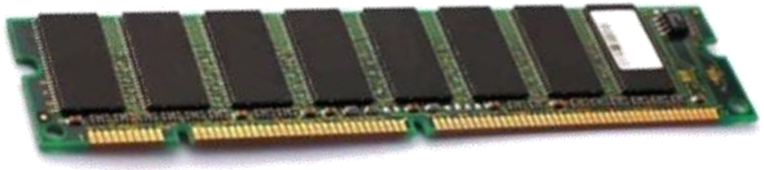
RDD

本质上是一个只读的分区记录集合

2.3.3 RDD运行原理(RDD概念、操作和特性)

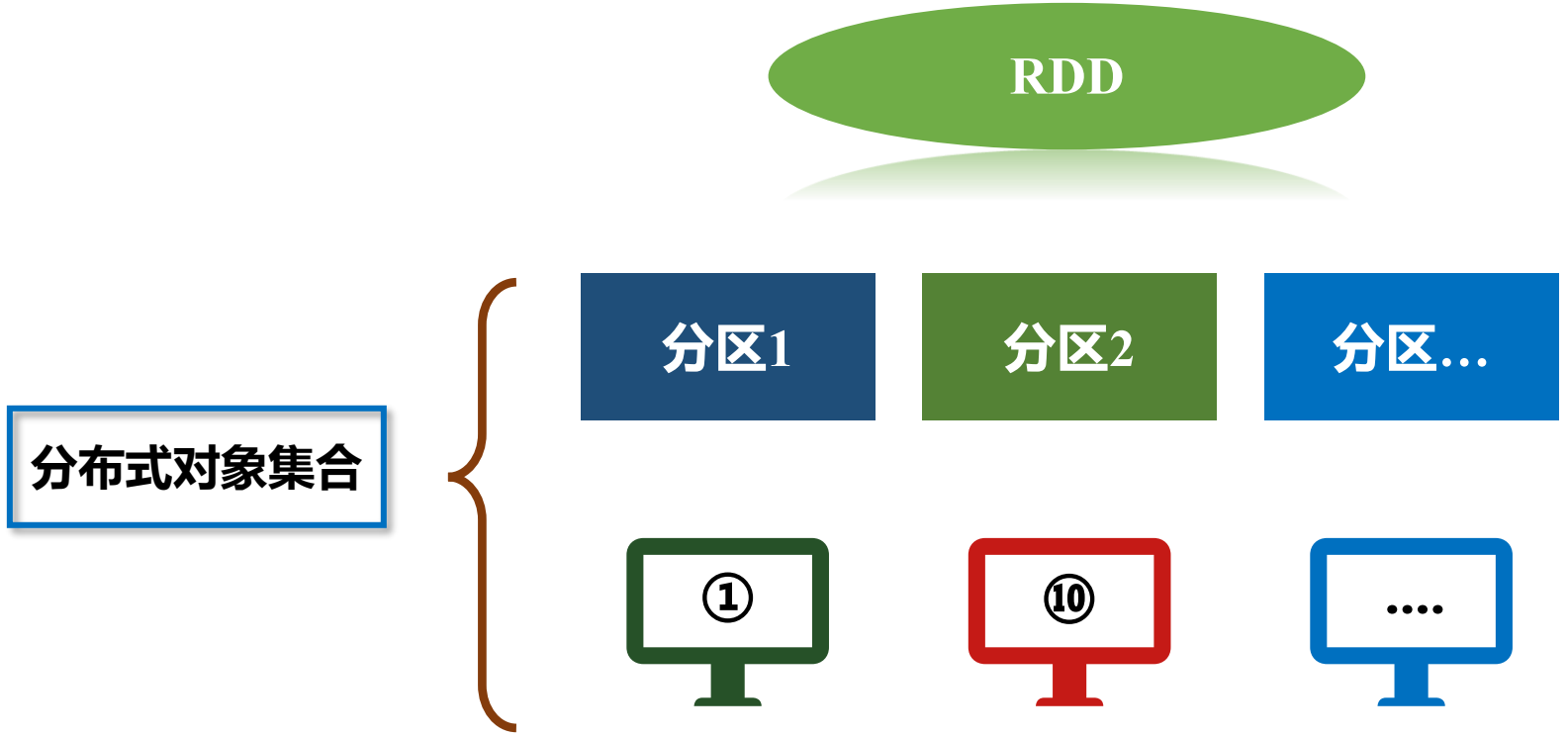
只读的分区集合



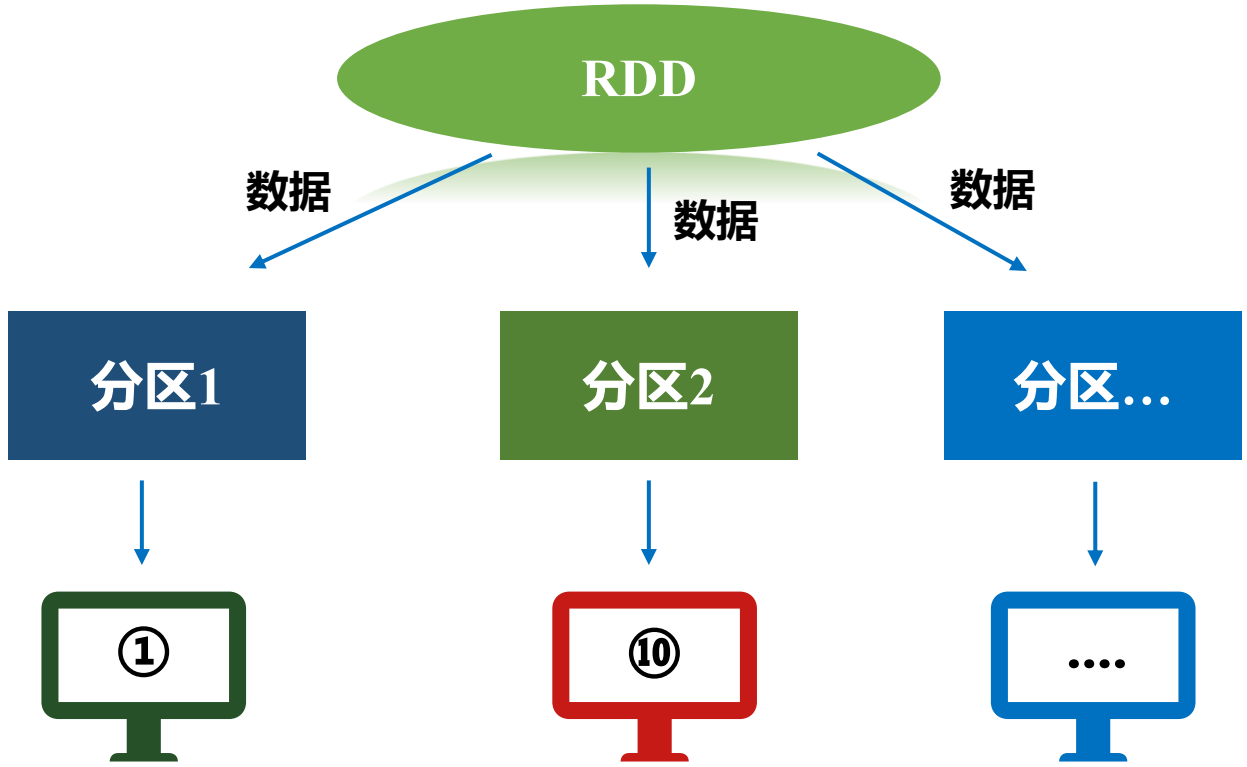


RDD

2.3.3 RDD运行原理(RDD概念、操作和特性)

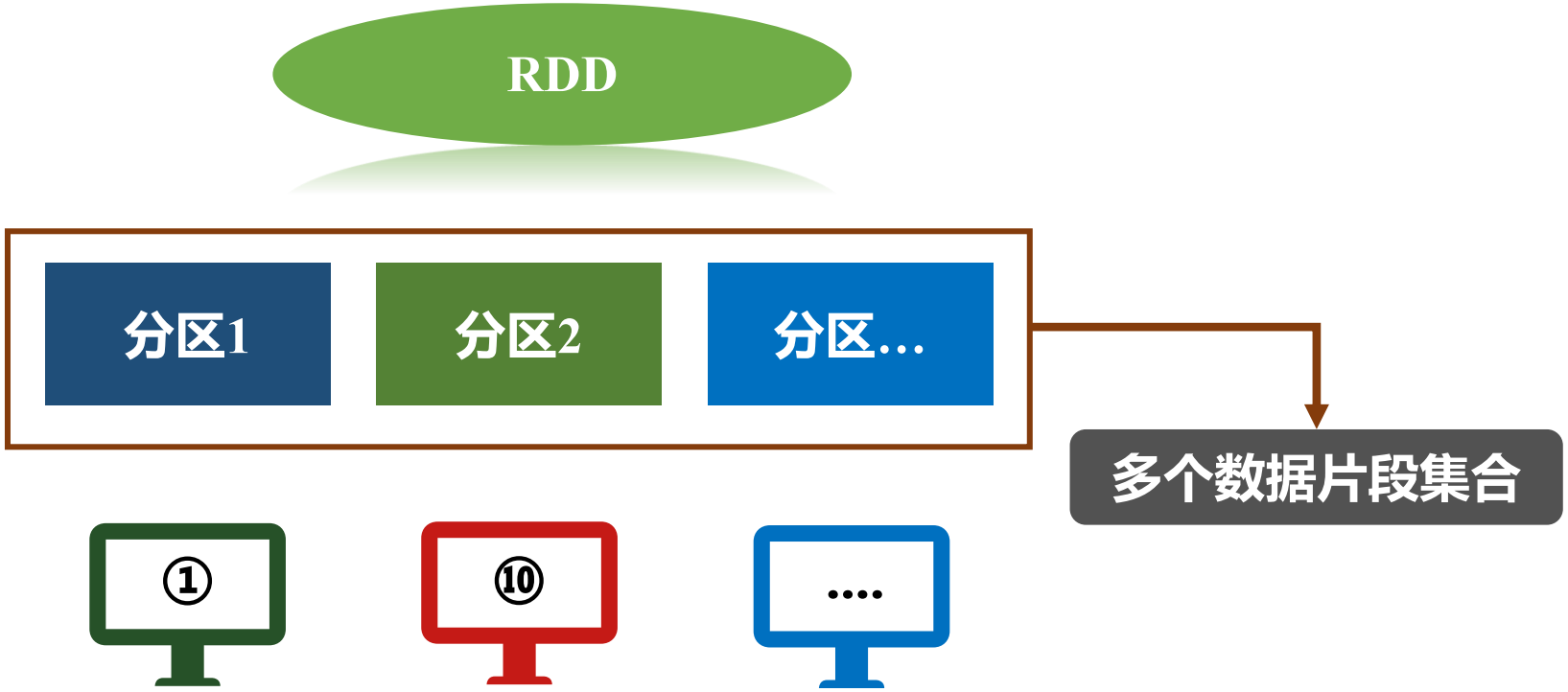


2.3.3 RDD运行原理(RDD概念、操作和特性)



分布式并行处理 高效并行计算

2.3.3 RDD运行原理(RDD概念、操作和特性)



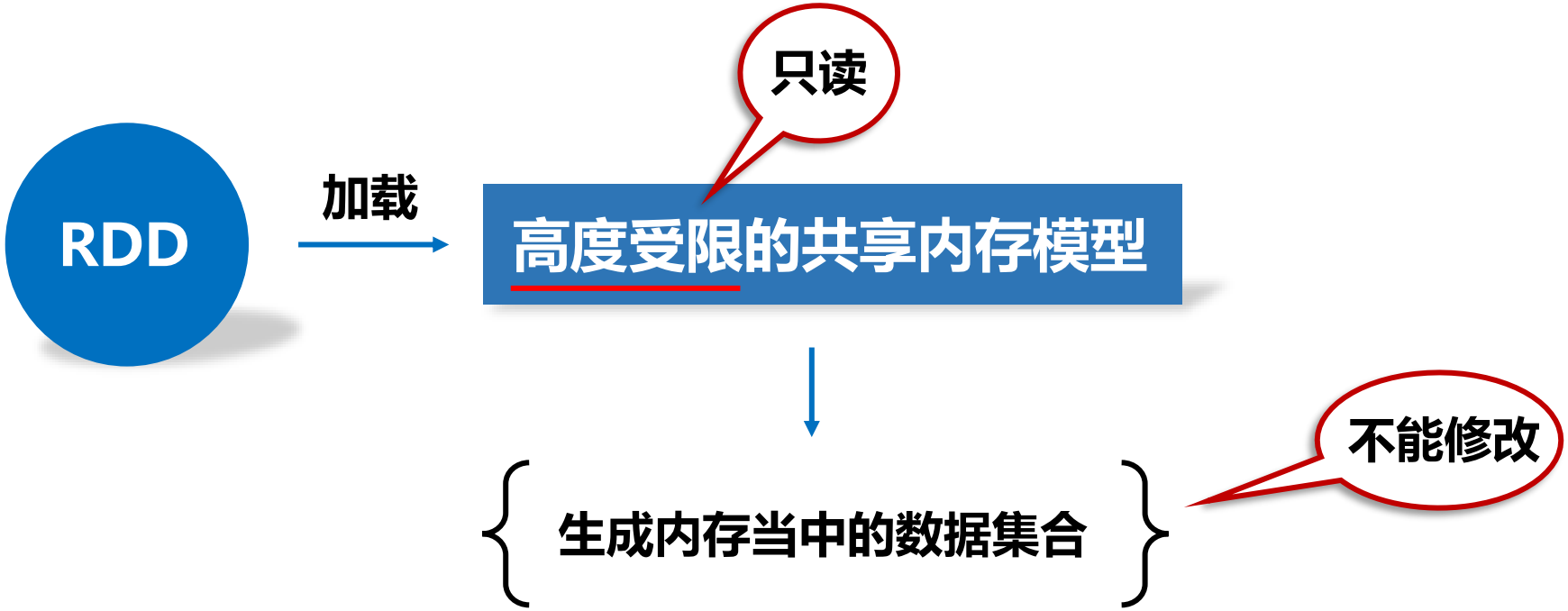
2.3.3 RDD运行原理(RDD概念、操作和特性)



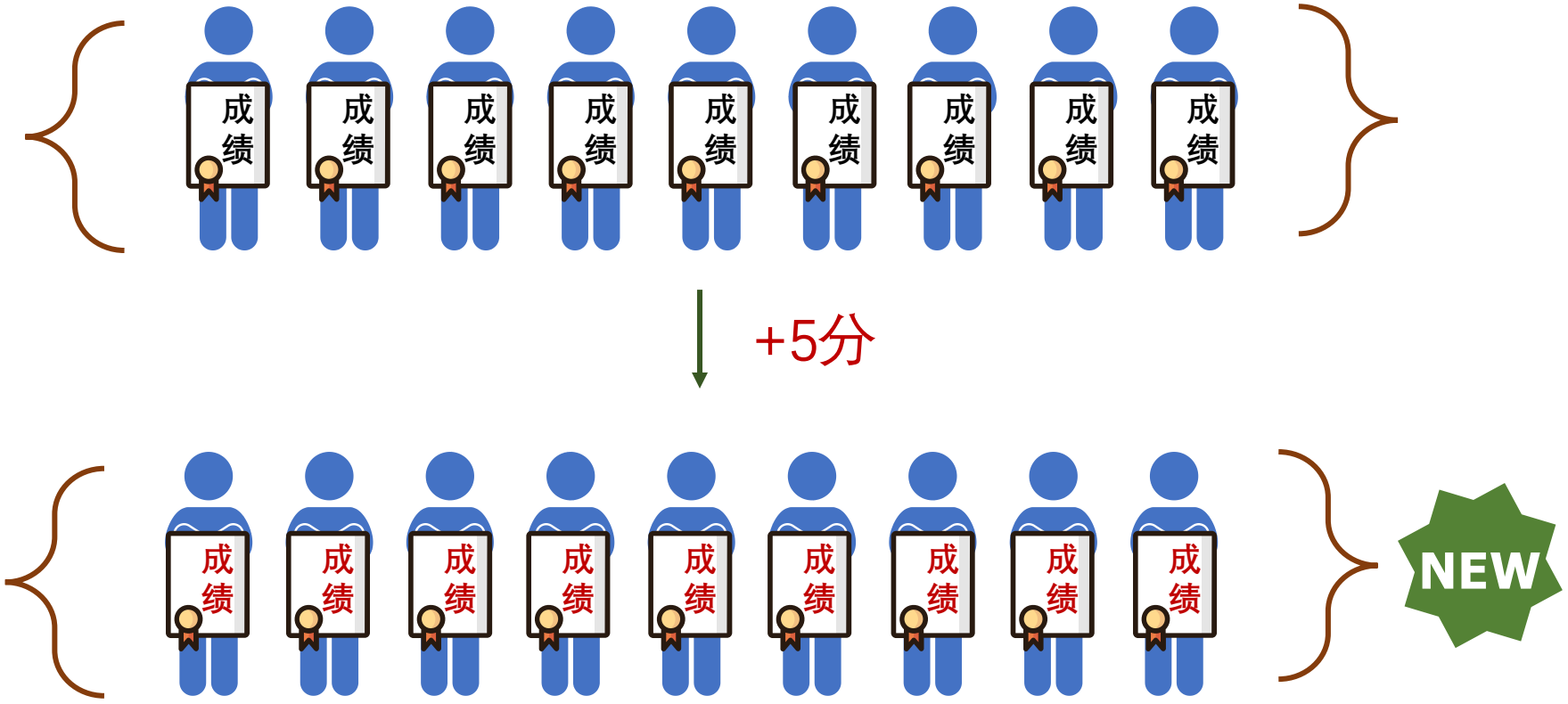
RDD

高度受限的共享内存模型

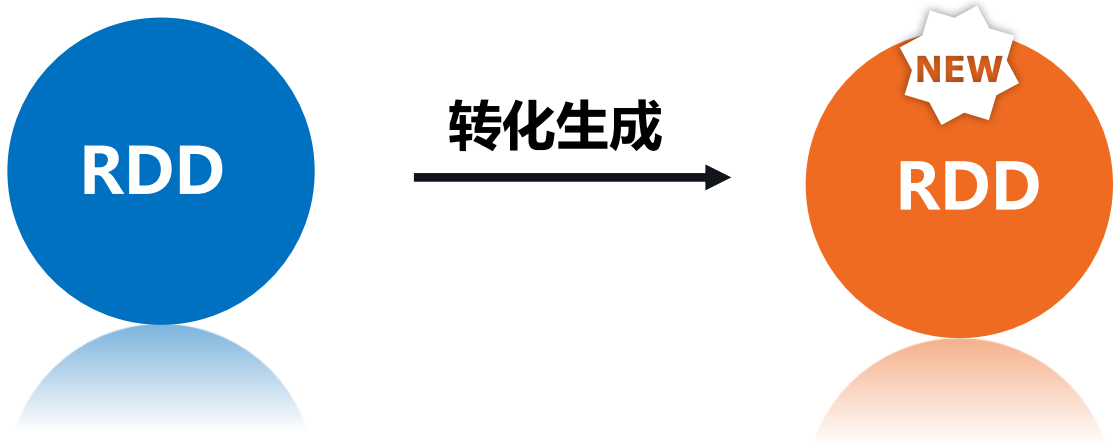
2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)



通过生成新的RDD来完成一个数据修改的目的

2.3.3 RDD运行原理(RDD概念、操作和特性)





2.3.3 RDD运行原理(RDD概念、操作和特性)



RDD典型的执行过程如下

RDD读入外部数据源进行创建

RDD经过一系列的转换（Transformation）操作

每一次都会产生不同的RDD供给下一个转换操作使用



最后一个RDD经过“动作”操作进行转换

并输出到外部数据源

2.3.3 RDD运行原理(RDD概念、操作和特性)



操作类型



(Action)



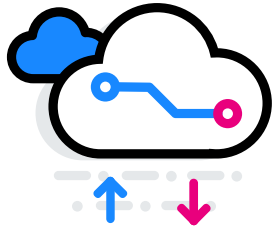
(Transformation)



2.3.3 RDD运行原理(RDD概念、操作和特性)



（Transformation）



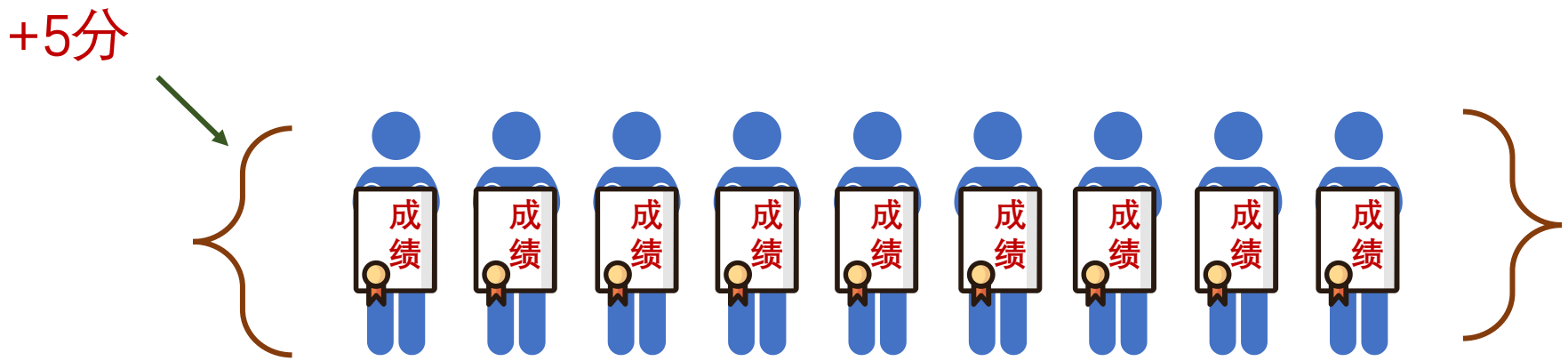
只记录转换的轨迹 ✓



发生计算 ✗

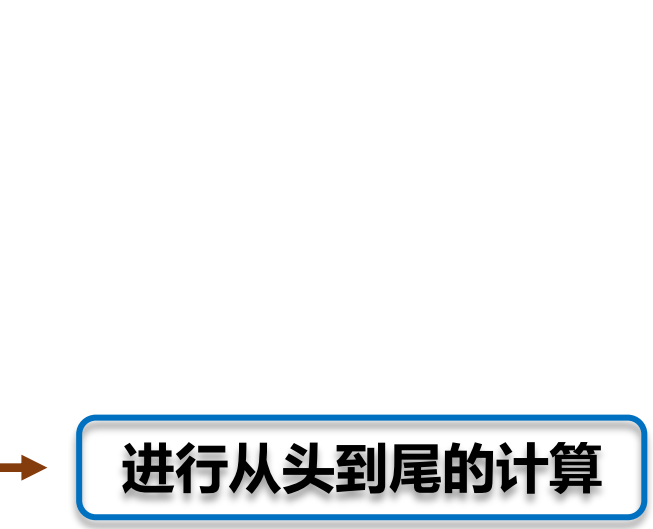
2.3.3 RDD运行原理(RDD概念、操作和特性)



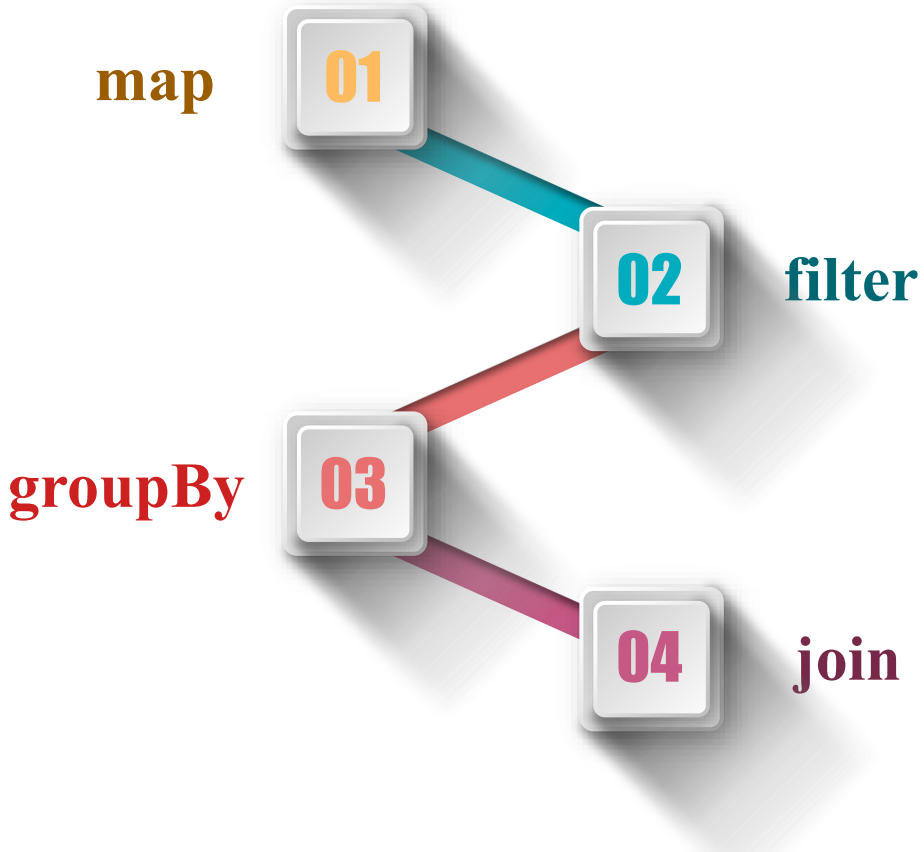


转换操作 → **记录加5分,并不会真正地发生计算**

2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)



RDD提供的转换操作

2.3.3 RDD运行原理(RDD概念、操作和特性)

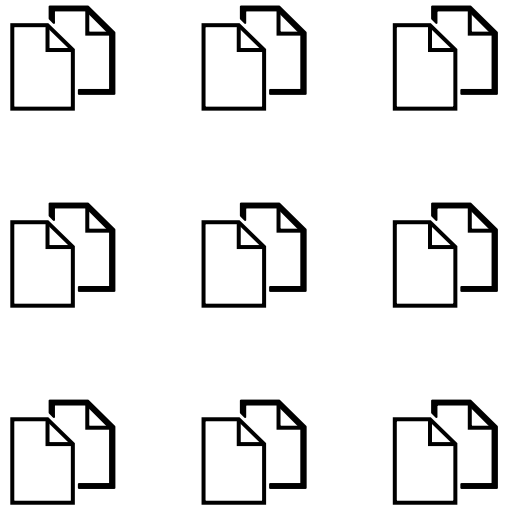


粗粒度

- 一次只能针对RDD全集进行转换



2.3.3 RDD运行原理(RDD概念、操作和特性)



数据库

+5分

95001学生

2.3.3 RDD运行原理(RDD概念、操作和特性)



粗粒度



支持粗粒度修改　一次只能针对RDD全集进行转换

2.3.3 RDD运行原理(RDD概念、操作和特性)



粗粒度







不支持细粒度修改

不适合数据库对单条进行修改

2.3.3 RDD运行原理(RDD概念、操作和特性)



高度受限的共享内存模型会不会影响表达能力?

2.3.3 RDD运行原理(RDD概念、操作和特性)





能力非常强大

高度受限的共享内存模型

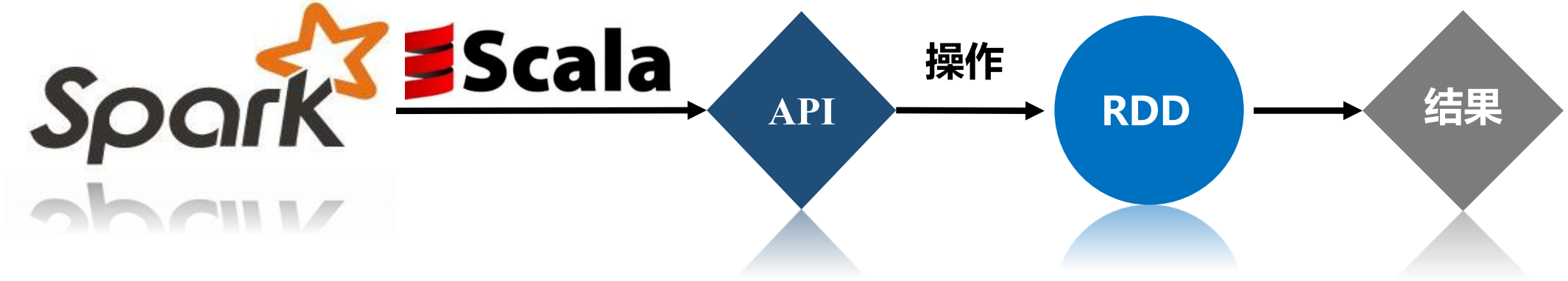
不会影响表达能力

2.3.3 RDD运行原理(RDD概念、操作和特性)



	操作1	操作2	操作3
hadoop MapReduce	√		×
PreGel Your passion. Our ingredients.		√	
Spark	√	√	√

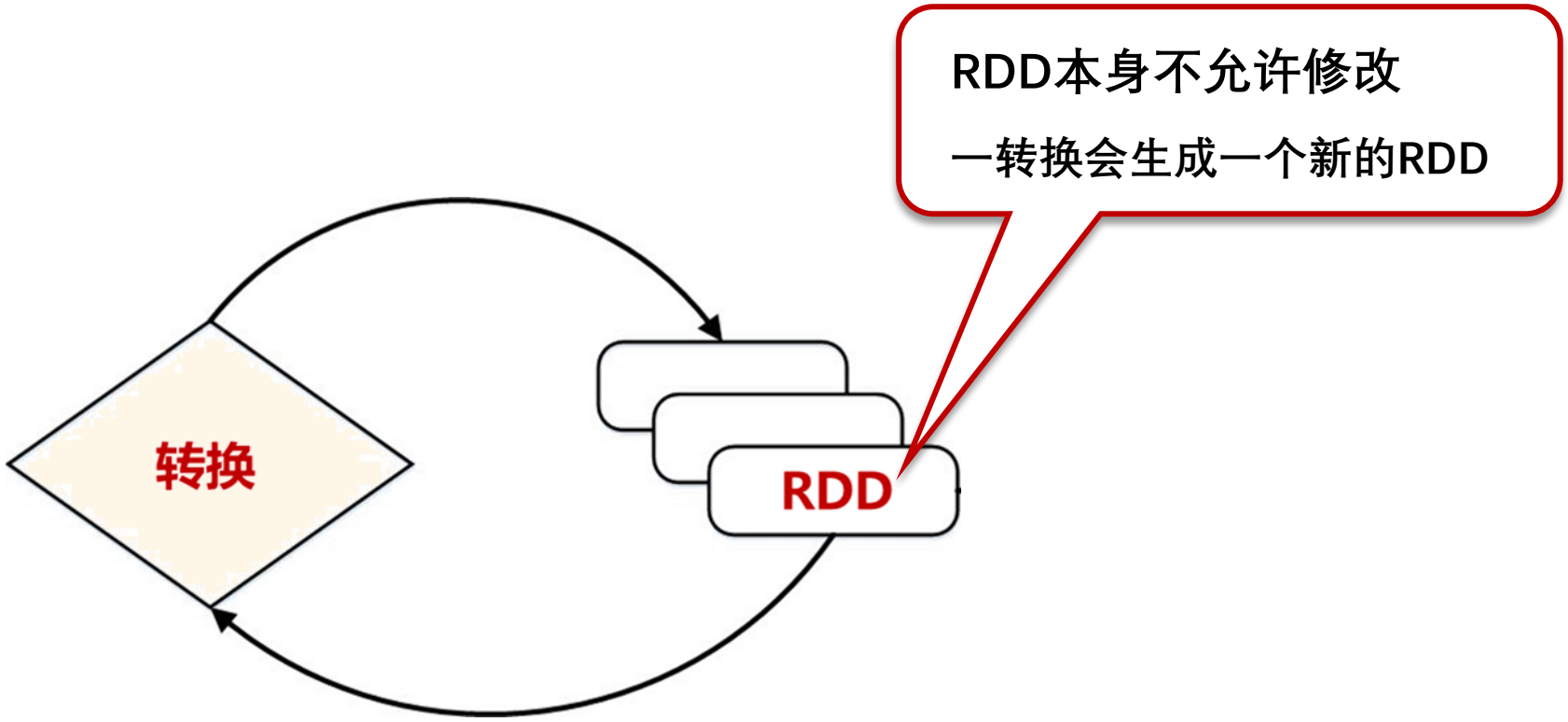
2.3.3 RDD运行原理(RDD概念、操作和特性)



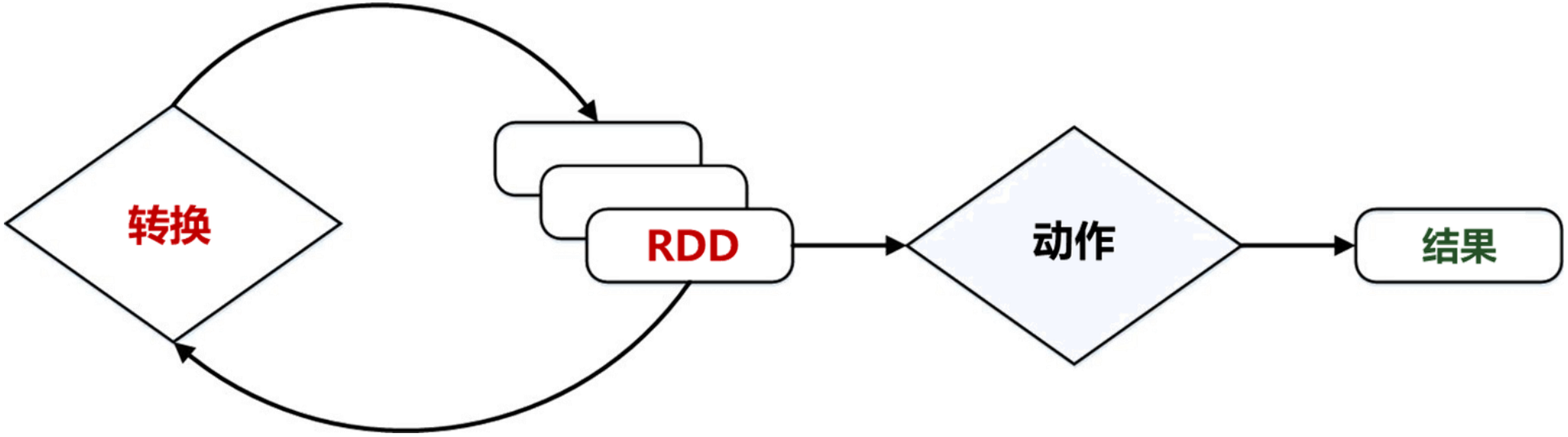
2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)

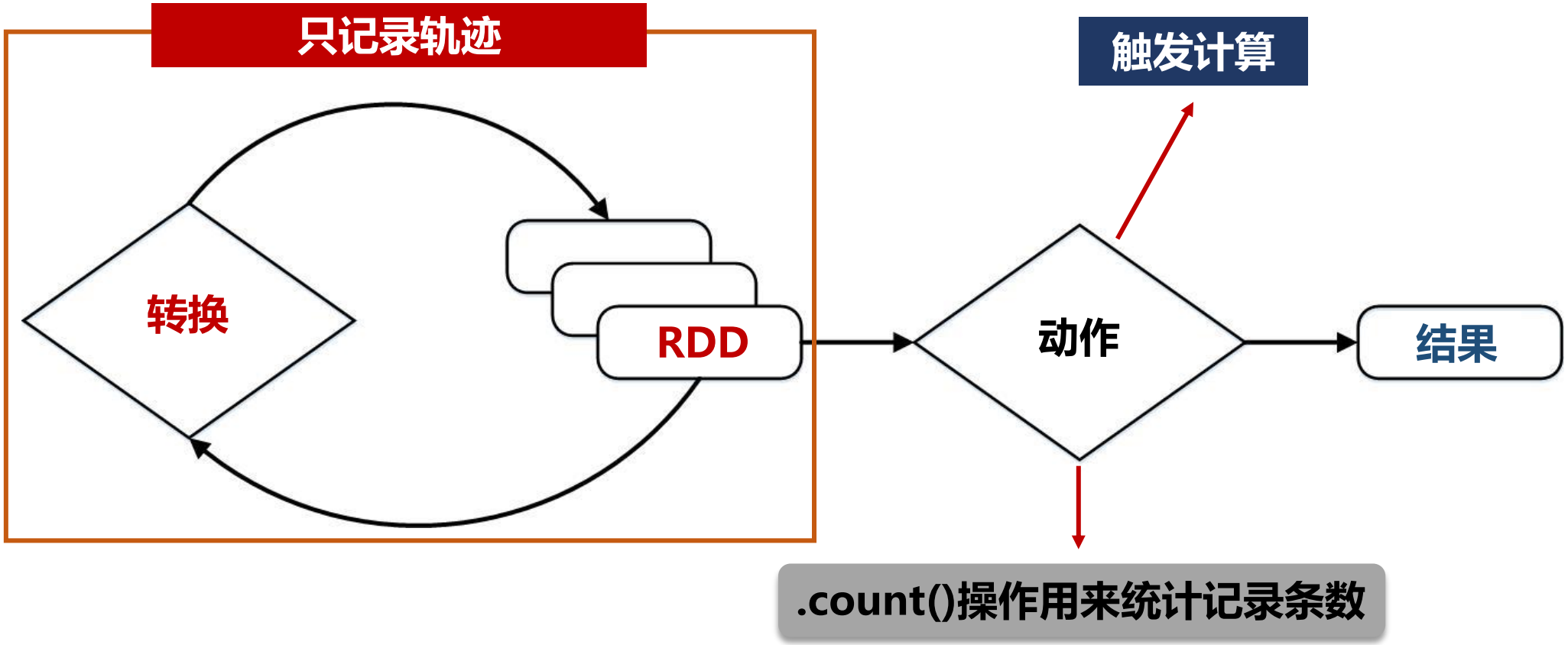


2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)

惰性调用机制



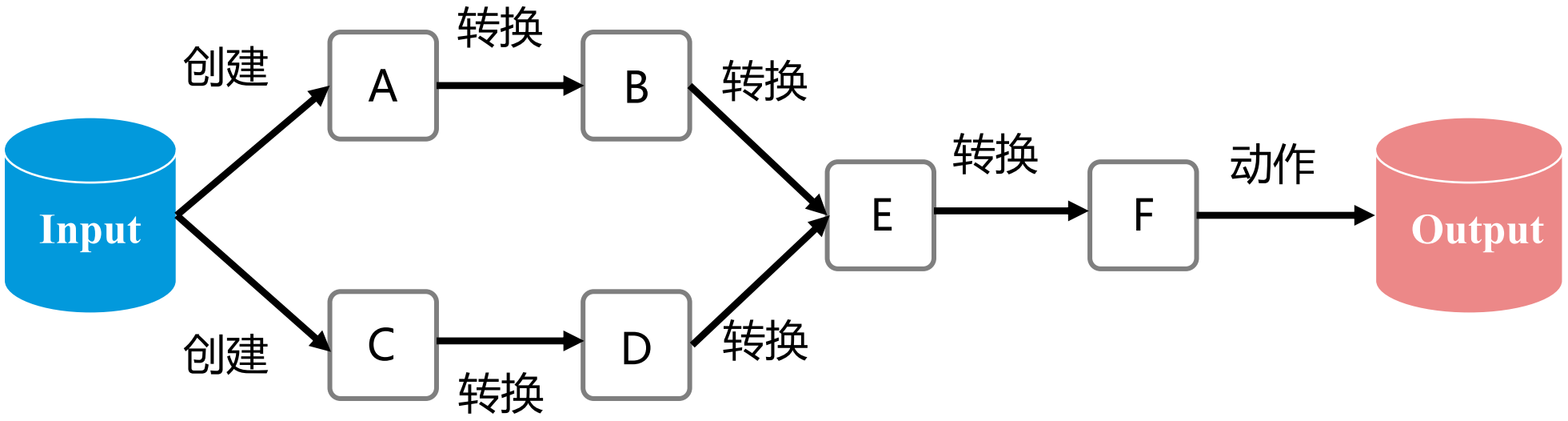
2.3.3 RDD运行原理(RDD概念、操作和特性)



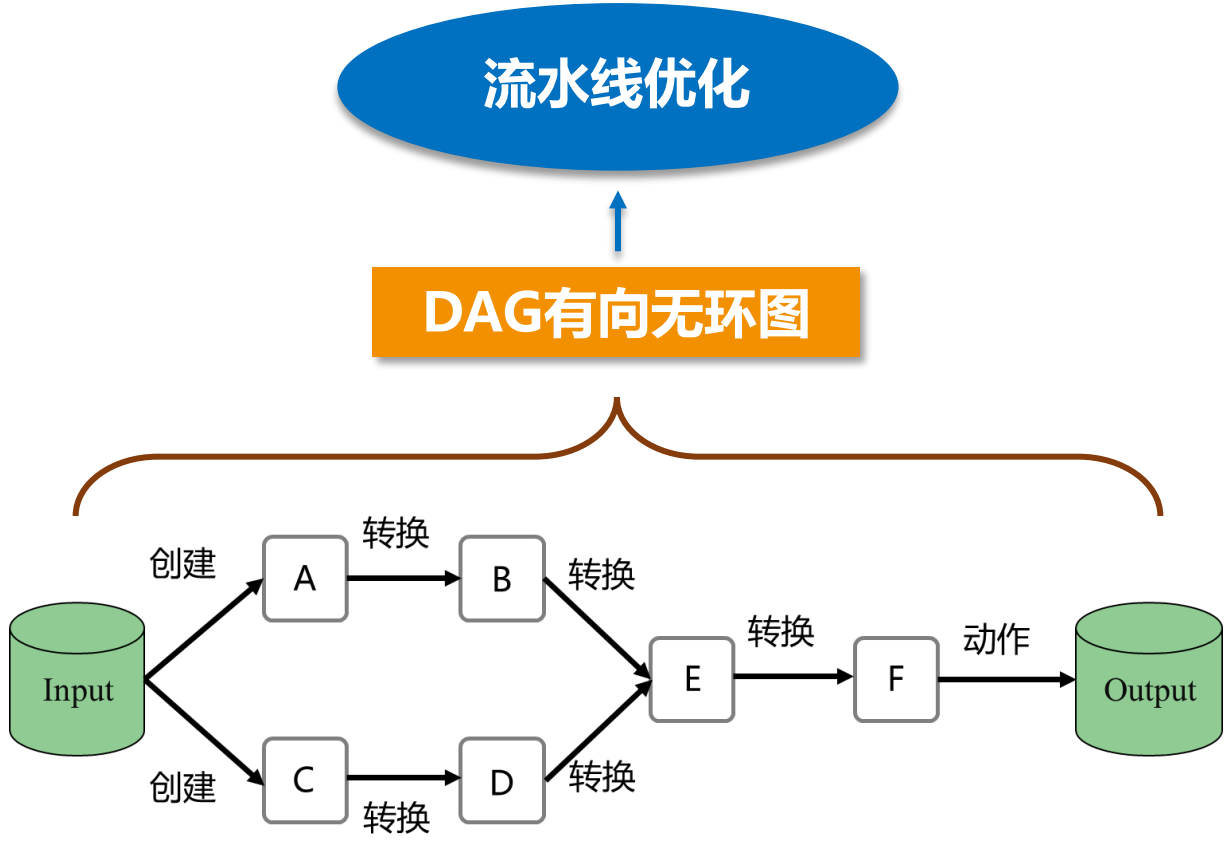
Lineage 血缘关系

恢复数据

寻亲



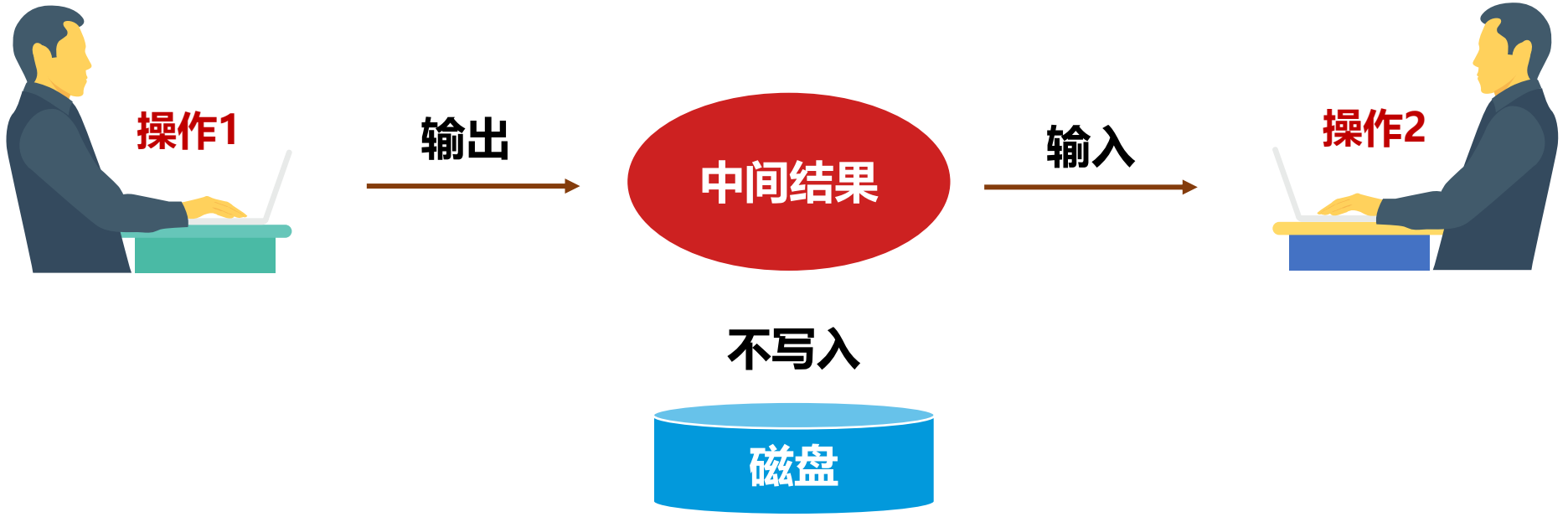
2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)



流水线优化



2.3.3 RDD运行原理(RDD概念、操作和特性)



Map函数

Reduce函数

应用复杂则需要写入非常复杂的代码

2.3.3 RDD运行原理(RDD概念、操作和特性)







2.3.3 RDD运行原理(RDD概念、操作和特性)



高效的容错性



- **天然的容错性**

2.3.3 RDD运行原理(RDD概念、操作和特性)

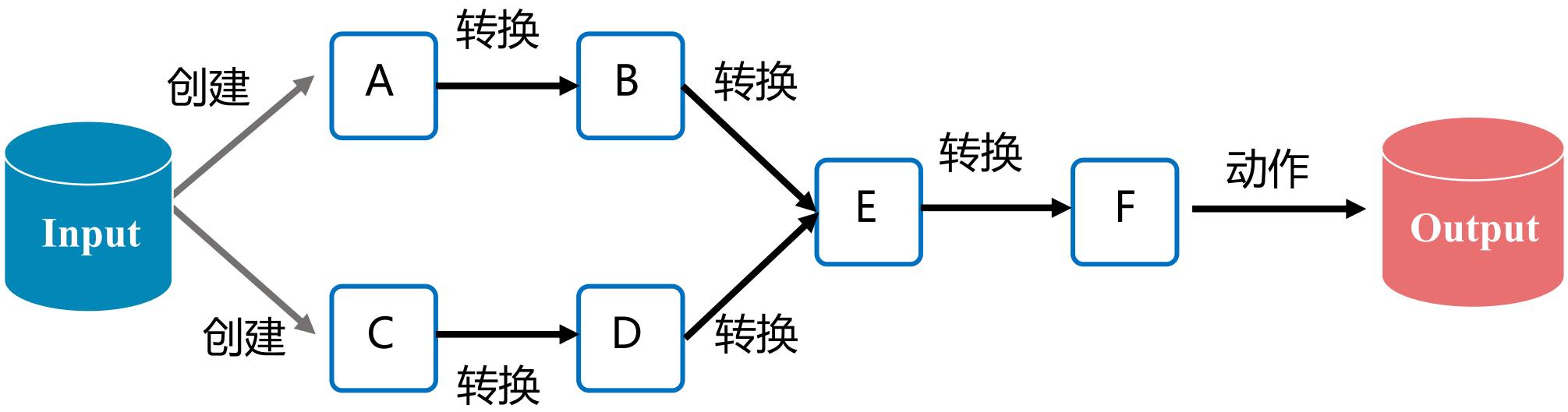


高效的容错性

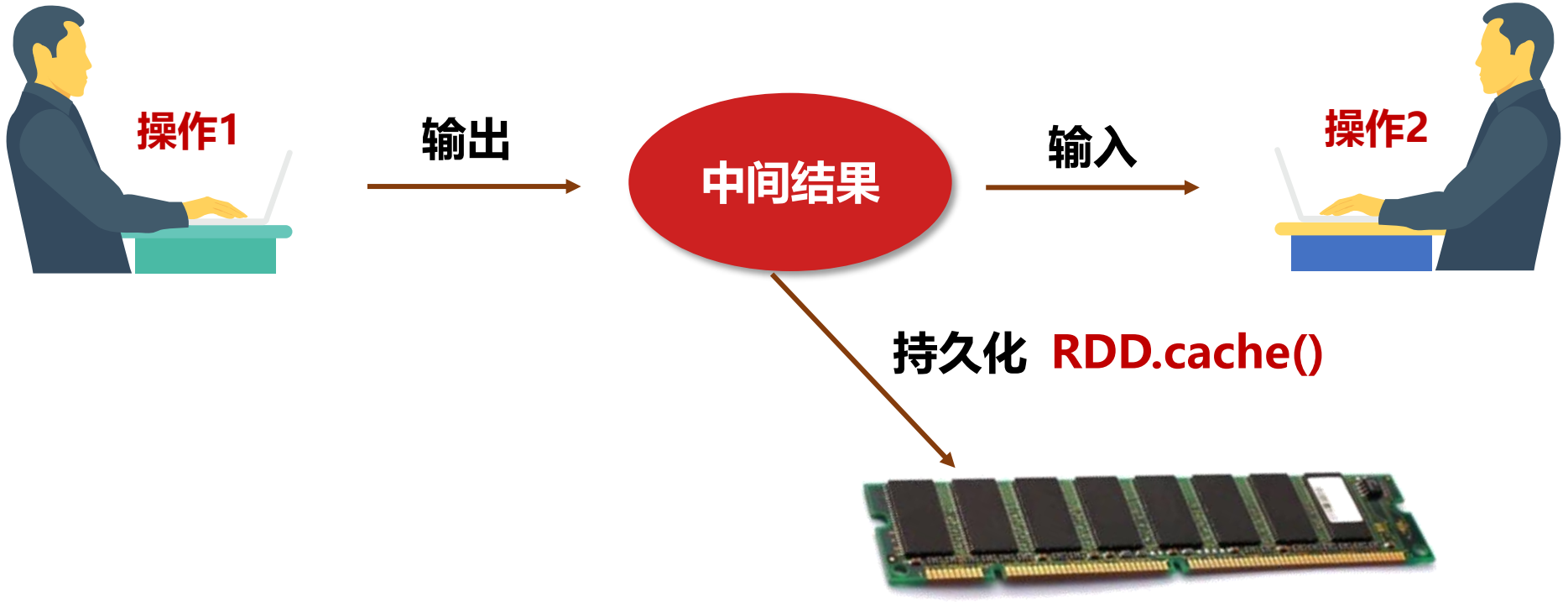
现有容错机制

数据复制	记录日志
数据备份	事务机制 日志 操作

2.3.3 RDD运行原理(RDD概念、操作和特性)

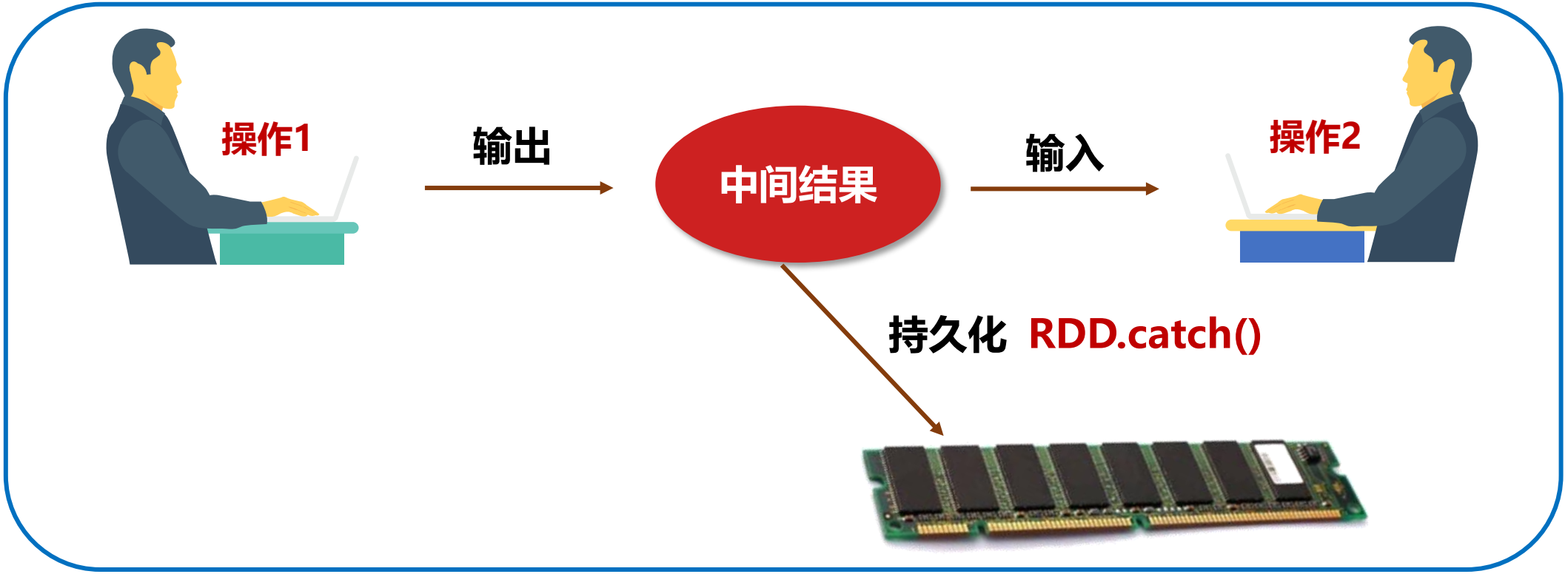


2.3.3 RDD运行原理(RDD概念、操作和特性)



2.3.3 RDD运行原理(RDD概念、操作和特性)

避免了不必要的读写磁盘开销

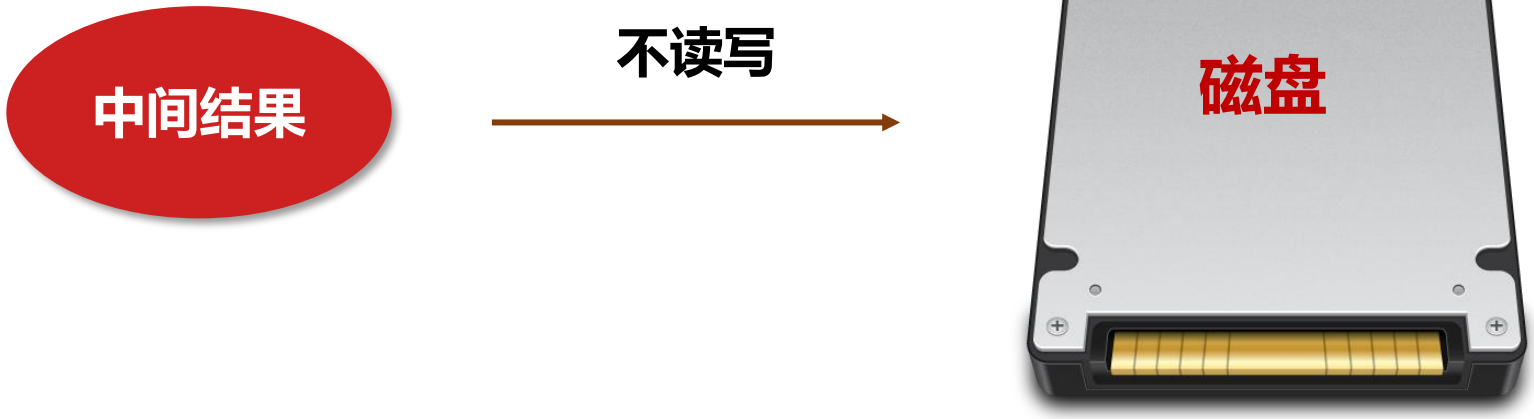


数据在内存的多个RDD之间进行传递的操作

2.3.3 RDD运行原理(RDD概念、操作和特性)



避免不必要的序列化和反序列化



2.3.3 RDD运行原理理解测验（选择题）



题目 1 : MapReduce的主要局限性中最恰当的是？

1.无法进行分布式处理

2.迭代计算时每次都会发生磁盘I/O

3.不支持并行处理

4.无法处理大容量数据

5.不支持函数式编程

2.3.3 RDD运行原理理解测验（选择题）



题目 2 ： RDD最核心的特性是什么？

1.可变的(Mutable)

2.基于磁盘存储

3.不可变性(Immutable)

4.只能在单个节点上运行

5.只能进行实时处

2.3.3 RDD运行原理理解测验（选择题）



题目 3： 在RDD中需要修改数据时采用什么方式处理？

1.直接修改现有RDD

2.生成新的RDD

3.在数据库中更新

4.初始化内存

5.临时存储到磁盘

2.3.3 RDD运行原理理解测验（选择题）



题目 4 ： RDD的两种主要操作类型是什么？

1.Map和Reduce

2.Read和Write

3.Input和Output

4.Transformation和Action

5.Create和Delete

2.3.3 RDD运行原理理解测验（选择题）



题目 5 ： RDD的Transformation操作的正确特征是？

1.立即执行

2.直接返回结果

3.只记录操作轨迹（惰性求值）

4.使用大量内存

5.将结果存储到磁盘

2.3.3 RDD运行原理理解测验（选择题）



题目 6 ： RDD支持的操作粒度是？

1.细粒度 (Fine-grained)

2.粗粒度 (Coarse-grained)

3.中等粒度 (Medium-grained)

4.可变粒度 (Variable-grained)

5.混合粒度 (Mixed-grained)

2.3.3 RDD运行原理理解测验（选择题）



题目 7 ： RDD的Lineage（血缘关系）的主要目的是？

1.性能提升

2.节省内存

3.故障恢复（容错性）

4.节省磁盘空间

5.提高网络速度

2.3.3 RDD运行原理理解测验（选择题）



题目 8 ： RDD流水线优化的最恰当优点是？

1.减少CPU使用量

2.避免中间结果的磁盘存储

3.增加网络带宽

4.减少内存使用量

5.降低代码复杂度

2.3.3 RDD运行原理理解测验（选择题）



题目 9 ： RDD.cache()方法的主要功能是？

1.将RDD存储到磁盘

2.压缩RDD

3.将RDD持久化到内存

4.删除RDD

5.复制RDD

2.3.3 RDD运行原理理解测验（选择题）



题目 10 :基于RDD的系统相比MapReduce不具备的主要优势是？

1.迭代算法中的性能提升

2.通过内存处理提高速度

3.单条记录修改的高效性

4.基于DAG的优化

5.提供丰富的高级API



Thank you