哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

大数据技术与实践

李春山
哈尔滨工业大学（威海）计算学部

内容提要

概述

MapReduce体系结构

MapReduce工作流程

实例分析：WordCount

MapReduce的具体应用

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

概述

概述

分布式并行编程

MapReduce模型简介

Map和Reduce函数

分布式并行编程

- “摩尔定律”， CPU性能大约每隔18个月翻一番
- 从2005年开始摩尔定律逐渐失效，需要处理的数据量快速增加，人们开始借助于分布式并行编程来提高程序性能
- 分布式程序运行在大规模计算机集群上，可以并行执行大规模数据处理任务，从而获得海量的计算能力
- 谷歌公司最先提出了分布式并行编程模型MapReduce，Hadoop MapReduce是它的开源实现，后者比前者使用门槛低很多

分布式并行编程

问题：在MapReduce出现之前，已经有像MPI这样非常成熟的并行计算框架了，那么为什么Google还需要MapReduce？MapReduce相较于传统的并行计算框架有什么优势？

	传统并行计算框架	MapReduce
集群架构/容错性	共享式(共享内存/共享存储)，容错性差	非共享式，容错性好
硬件/价格/扩展性	刀片服务器、高速网、SAN，价格贵，扩展性差	普通PC机，便宜，扩展性好
编程/学习难度	what-how，难	what，简单
适用场景	实时、细粒度计算、计算密集型	批处理、非实时、数据密集型

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

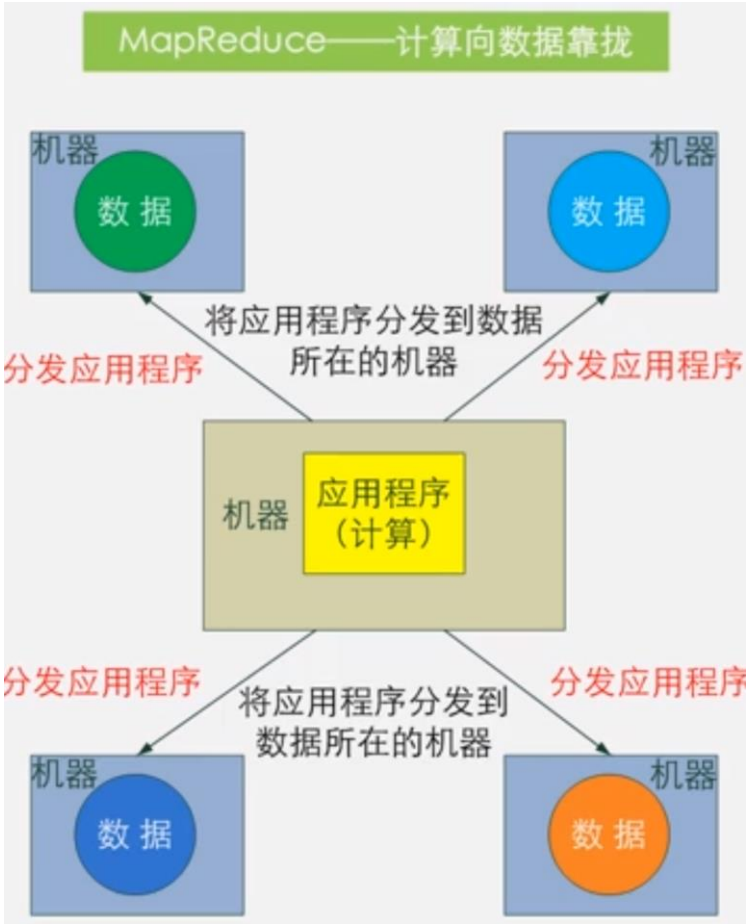
MapReduce模型简介

MapReduce模型简介

- MapReduce将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数：Map和Reduce
- 编程容易，不需要掌握分布式并行编程细节，也可以很容易把自己的程序运行在分布式系统上，完成海量数据的计算

MapReduce模型简介

- MapReduce设计的一个理念就是“**计算向数据靠拢**”，而不是“数据向计算靠拢”，因为，移动数据需要大量的网络传输开销



MapReduce模型简介

Map和Reduce函数

Map和Reduce

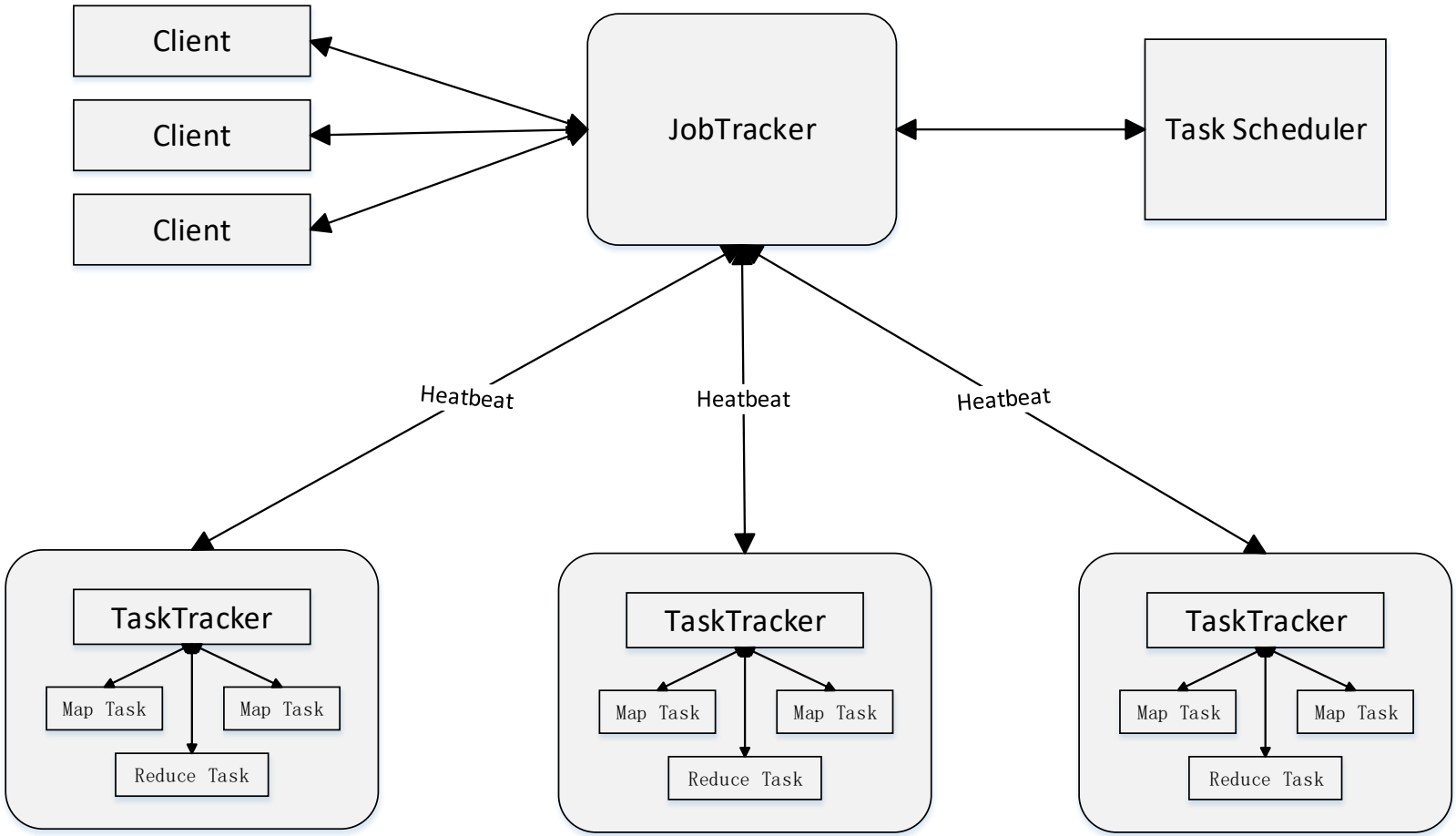
函数	输入	输出	说明
Map	<k_1,v_1> 如： <行号,"a b c">	List(<k_2,v_2>) 如： <"a",1> <"b",1> <"c",1>	1.将小数据集进一步解析成一批<key,value>对，输入Map函数中进行处理 2.每一个输入的<k_1,v_1>会输出一批<k_2,v_2>。<k_2,v_2>是计算的中间结果
Reduce	<k_2,List(v_2)> 如： <"a",<1,1,1>>	<k_3,v_3> <"a",3>	输入的中间结果<k_2,List(v_2)>中的List(v_2)表示是一批属于同一个k_2的value

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

MapReduce体系结构

MapReduce的体系结构

MapReduce体系结构主要由四个部分组成，分别是：Client、JobTracker、TaskTracker以及Task



MapReduce的体系结构

MapReduce主要有以下4个部分组成：

1）Client

•用户编写的MapReduce程序通过Client提交到JobTracker端

•用户可通过Client提供的一些接口查看作业运行状态

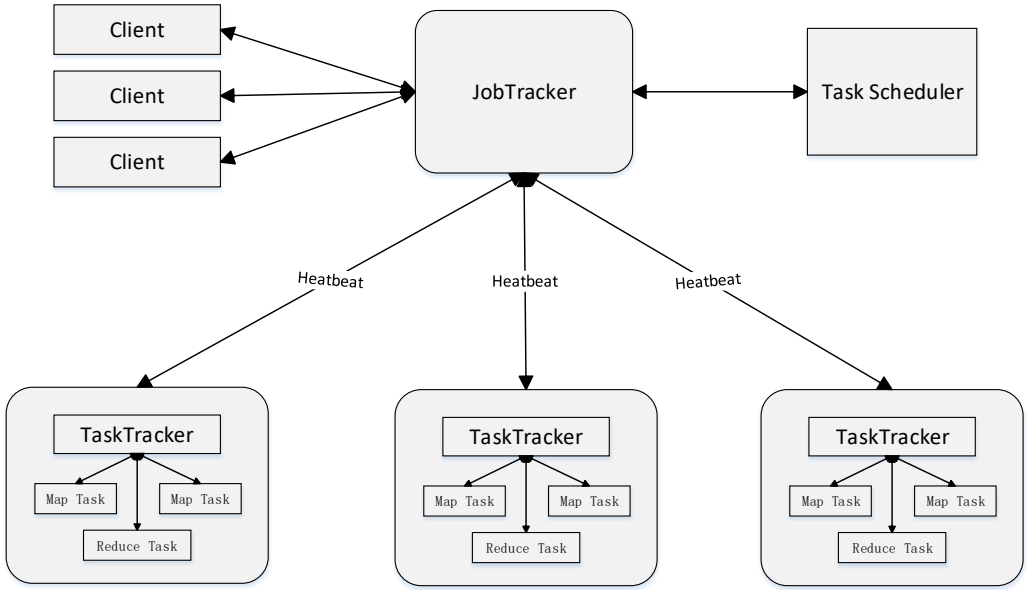
MapReduce的体系结构

2）JobTracker

•JobTracker负责资源监控和作业调度

•JobTracker 监控所有TaskTracker与Job的健康状况，一旦发现失败，就将相应的任务转移到其他节点

•JobTracker 会跟踪任务的执行进度、资源使用量等信息，并将这些信息告诉任务调度器（TaskScheduler），而调度器会在资源出现空闲时，选择合适的任务去使用这些资源



MapReduce的体系结构

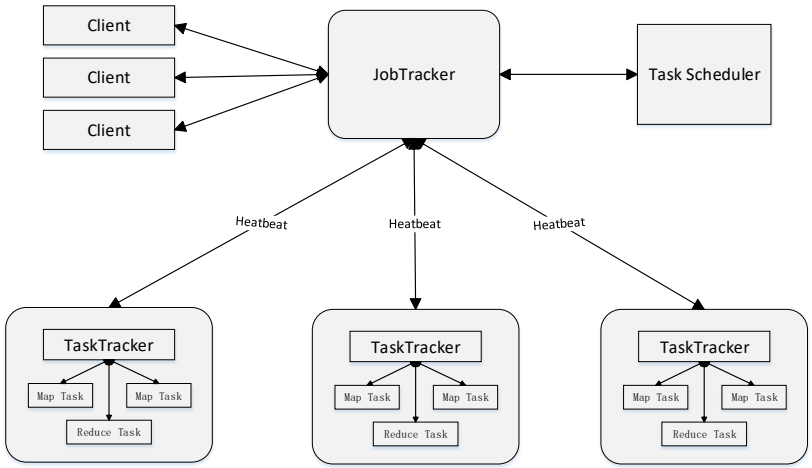
3）TaskTracker

•TaskTracker 会周期性地通过“心跳”将本节点上资源的使用情况和任务的运行进度汇报给JobTracker，同时接收JobTracker 发送过来的命令并执行相应的操作（如启动新任务、杀死任务等）

•TaskTracker 使用“slot”等量划分本节点上的资源量（CPU、内存等）。一个Task 获取到一个slot 后才有机会运行，而Hadoop调度器的作用就是将各个TaskTracker上的空闲slot分配给Task使用。slot 分为Map slot 和Reduce slot 两种，分别供MapTask 和Reduce Task 使用

4）Task

Task 分为Map Task 和Reduce Task 两种，均由TaskTracker 启动

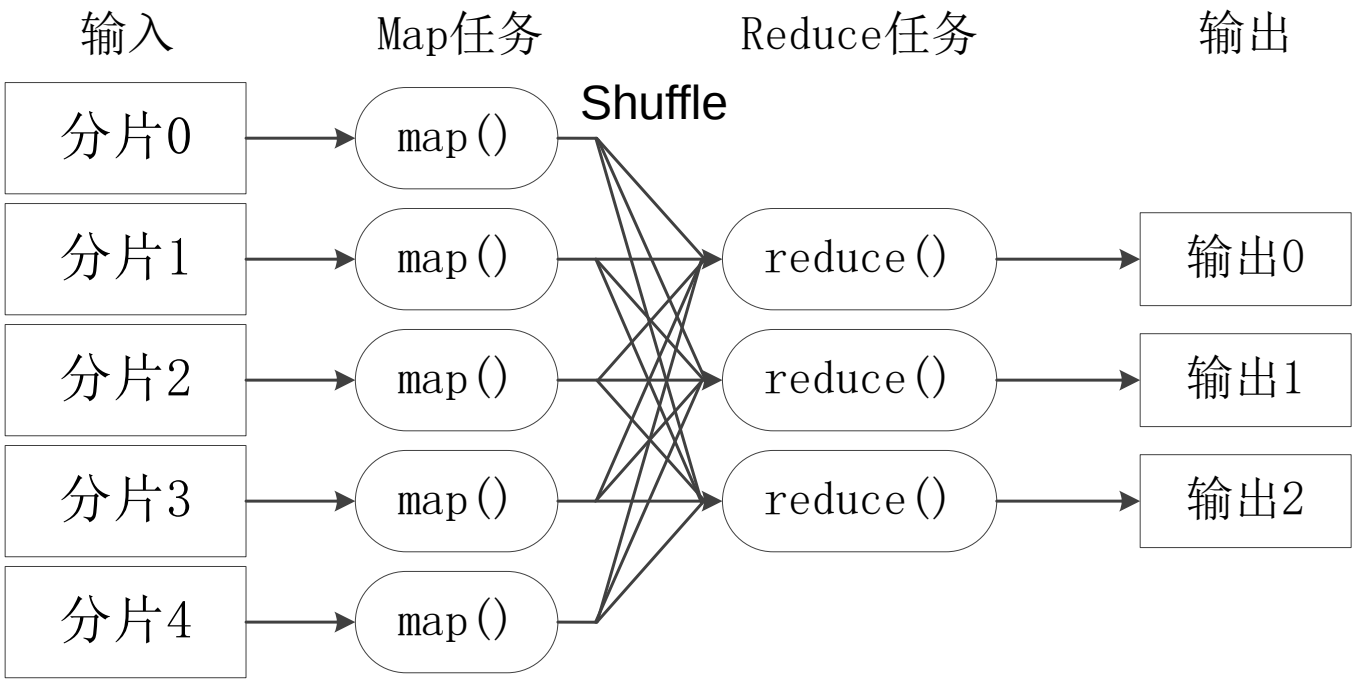


MapReduce工作流程

MapReduce工作流程

- 工作流程概述
- **MapReduce各个执行阶段**
- **Shuffle过程详解**

工作流程概述

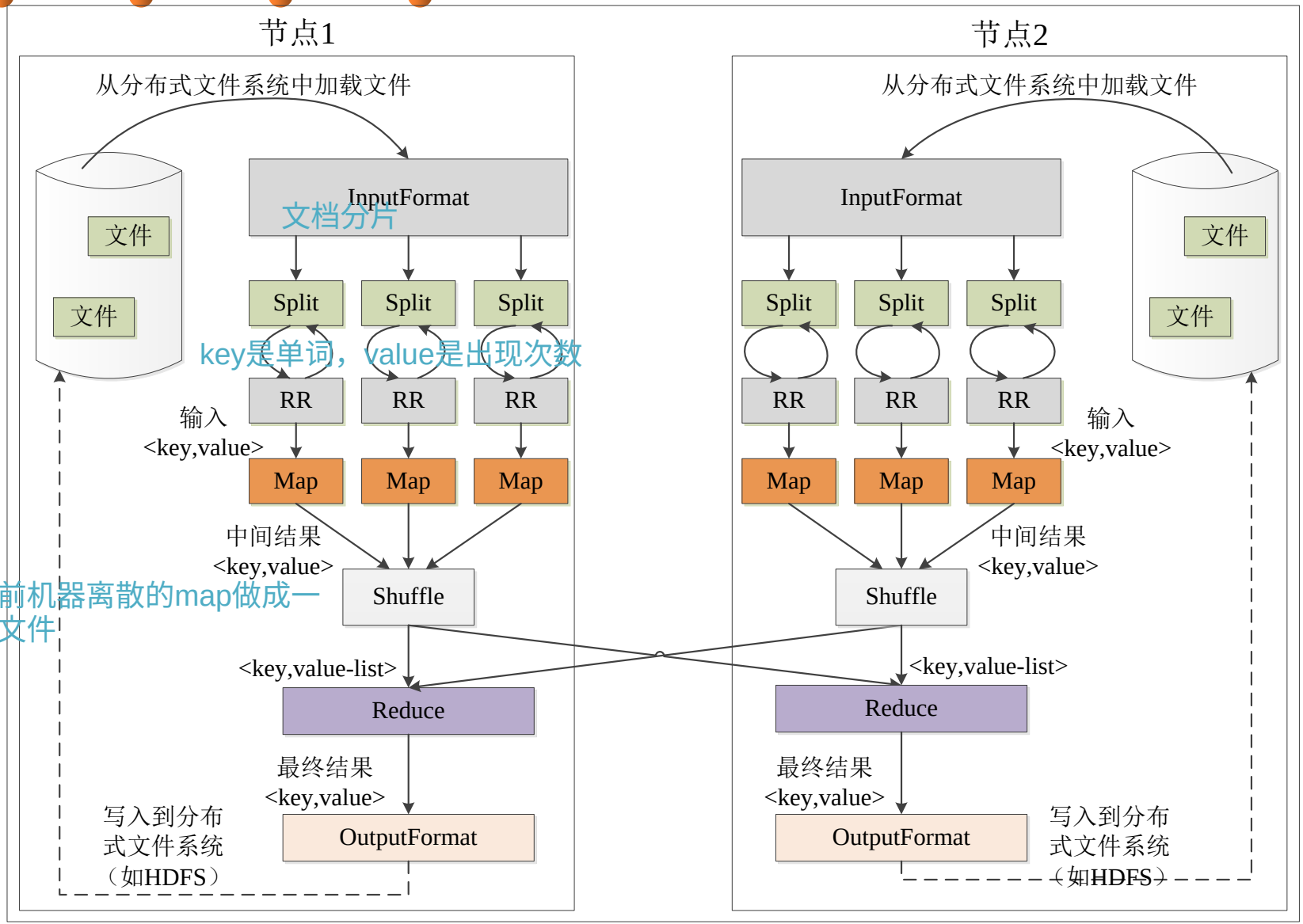


MapReduce工作流程

每一个子任务都可以独立的完成，
如果有依赖关系就不能这样分解

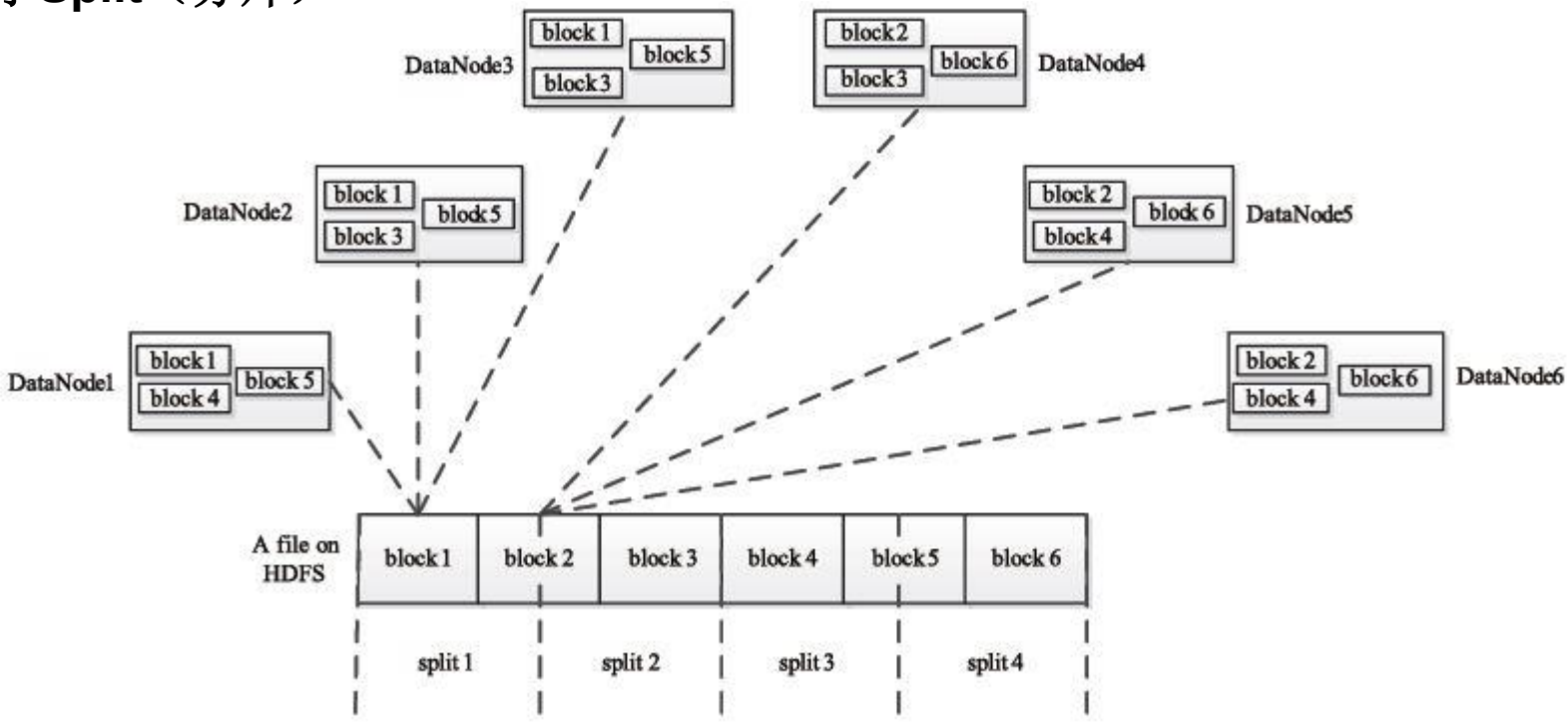
- 不同的Map任务之间不会进行通信
- 不同的Reduce任务之间也不会发生任何信息交换
- 用户不能显式地从一台机器向另一台机器发送消息
- 所有的数据交换都是通过MapReduce框架自身去实现的

MapReduce各个执行阶段



MapReduce各个执行阶段

关于**Split**（分片）

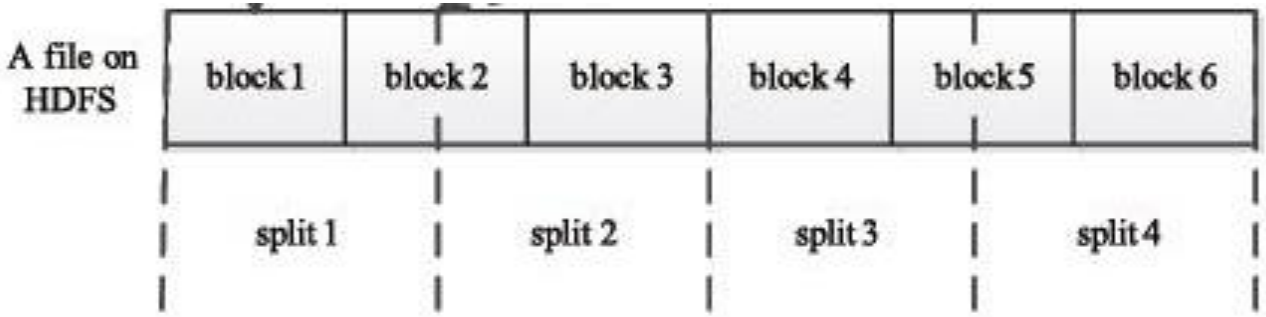


HDFS 以固定大小的block 为基本单位存储数据，而对于MapReduce 而言，其处理单位是split。split 是一个逻辑概念，它只包含一些元数据信息，比如数据起始位置、数据长度、数据所在节点等。它的划分方法完全由用户自己决定。

MapReduce各个执行阶段

Map任务的数量

•Hadoop为每个split创建一个Map任务，split 的多少决定了Map任务的数目。大多数情况下，理想的分片大小是一个HDFS块



Reduce任务的数量

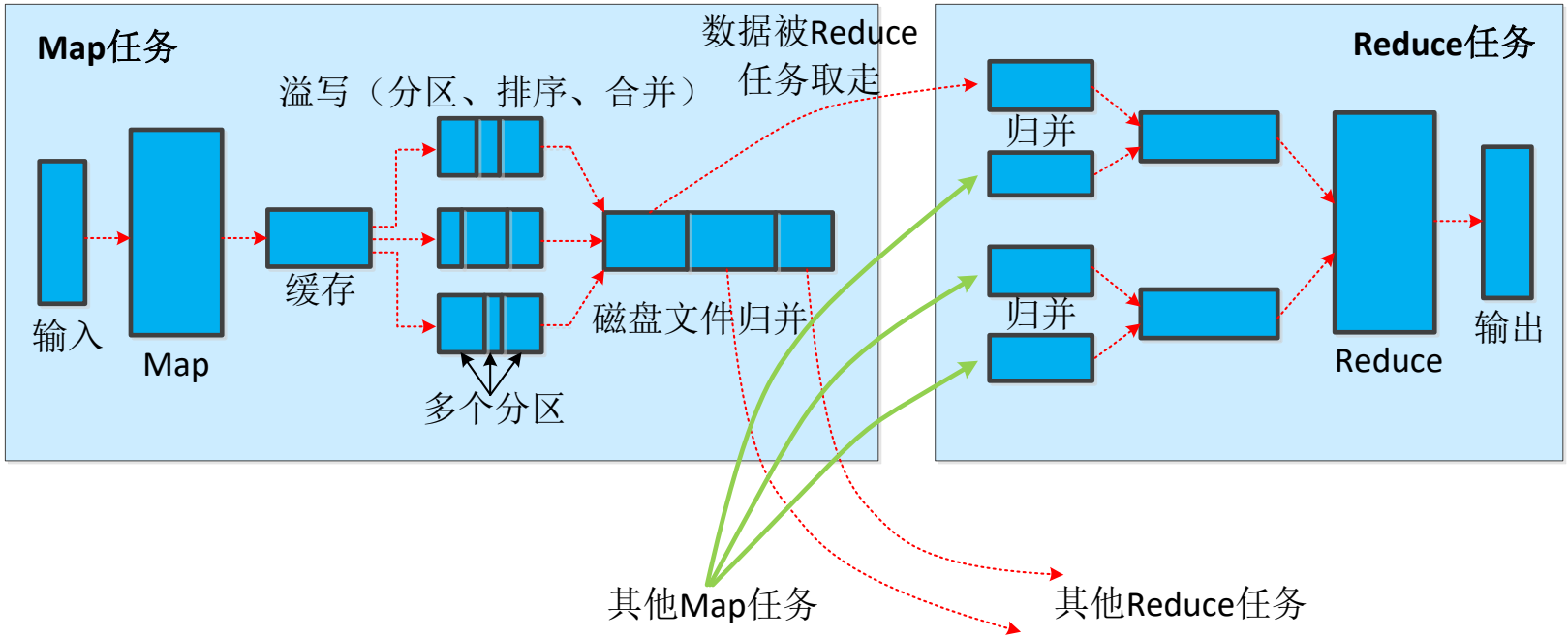
•最优的Reduce任务个数取决于集群中可用的reduce任务槽(slot)的数目

•通常设置比reduce任务槽数目稍微小一些的Reduce任务个数（这样可以预留一些系统资源处理可能发生的错误）

Shuffle过程原理

Shuffle过程详解

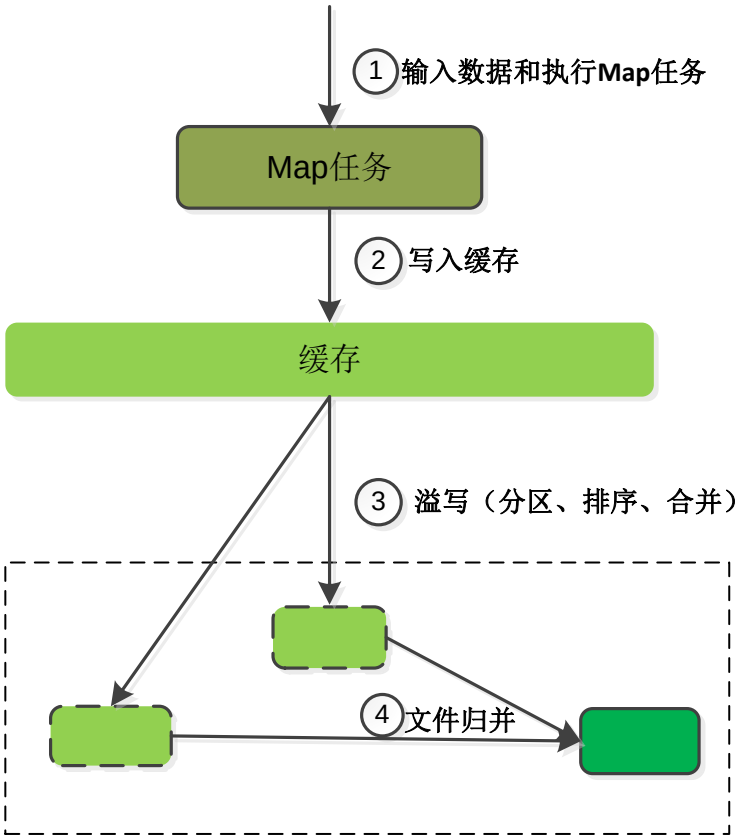
1. Shuffle过程简介



Shuffle过程

Shuffle过程详解

2. Map端的Shuffle过程



- 每个Map任务分配一个缓存
- MapReduce默认100MB缓存

- 设置溢写比例0.8
- 分区默认采用哈希函数
- 排序是默认的操作
- 排序后可以合并（Combine）
- 合并不能改变最终结果

- 在Map任务全部结束之前进行归并
- 归并得到一个大的文件，放在本地磁盘
- 文件归并时，如果溢写文件数量大于预定值（默认是3）则可以再次启动Combiner，少于3不需要
- JobTracker会一直监测Map任务的执行，并通知Reduce任务来领取数据

自己执行

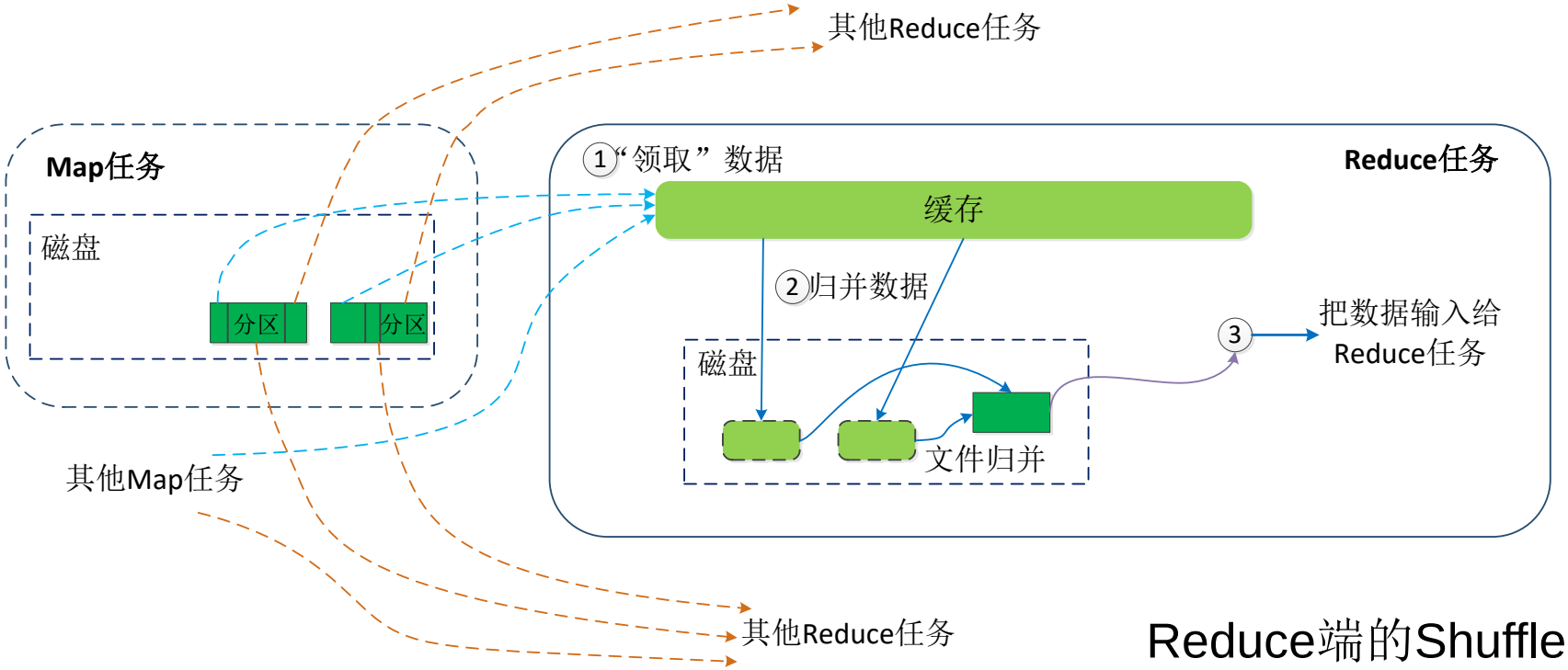
合并（Combine）和归并（Merge）的区别：

两个键值对<“a”,1>和<“a”,1>，如果合并，会得到<“a”,2>，如果归并，会得到<“a”,<1,1>>

Shuffle过程详解

3. Reduce端的Shuffle过程

- Reduce任务通过RPC向JobTracker询问Map任务是否已经完成，若完成，则领取数据
- Reduce领取数据先放入缓存，来自不同Map机器，先归并，再合并，写入磁盘
- 多个溢写文件归并成一个或多个大文件，文件中的键值对是排序的
- 当数据很少时，不需要溢写到磁盘，直接在缓存中归并，然后输出给Reduce

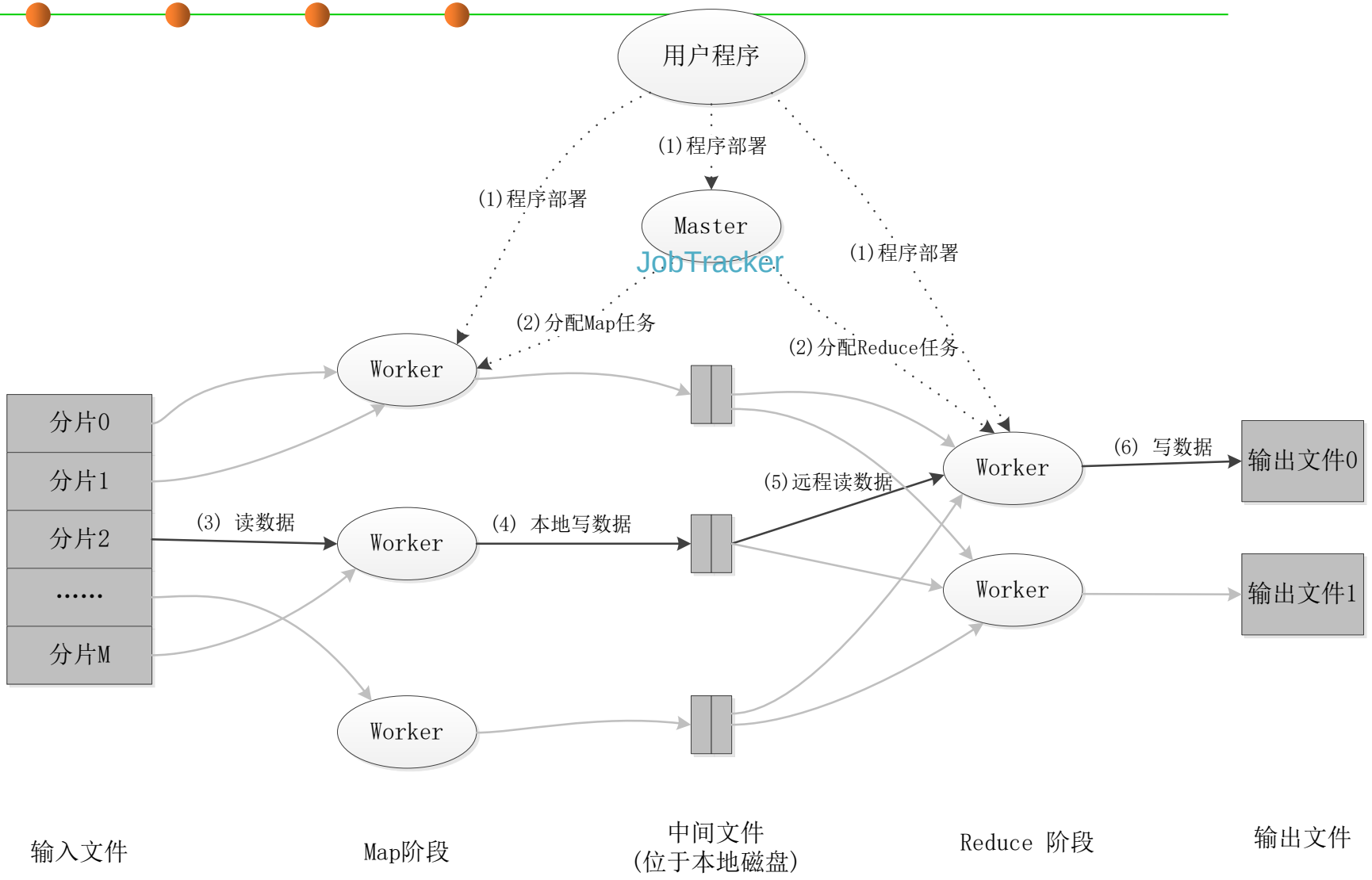


Reduce端的Shuffle过程

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

MapReduce应用程序执行过程

MapReduce应用程序执行过程



实例分析：WordCount

实例分析：WordCount

- **WordCount程序任务**
- **WordCount设计思路**
- **一个WordCount执行过程的实例**

WordCount程序任务

WordCount程序任务

程序	WordCount
输入	一个包含大量单词的文本文件
输出	文件中每个单词及其出现次数（频数），并按照单词字母顺序排序，每个单词和其频数占一行，单词和频数之间有间隔

一个WordCount的输入和输出实例

输入	输出
Hello World Hello Hadoop Hello MapReduce	Hadoop 1 Hello 3 MapReduce 1 World 1

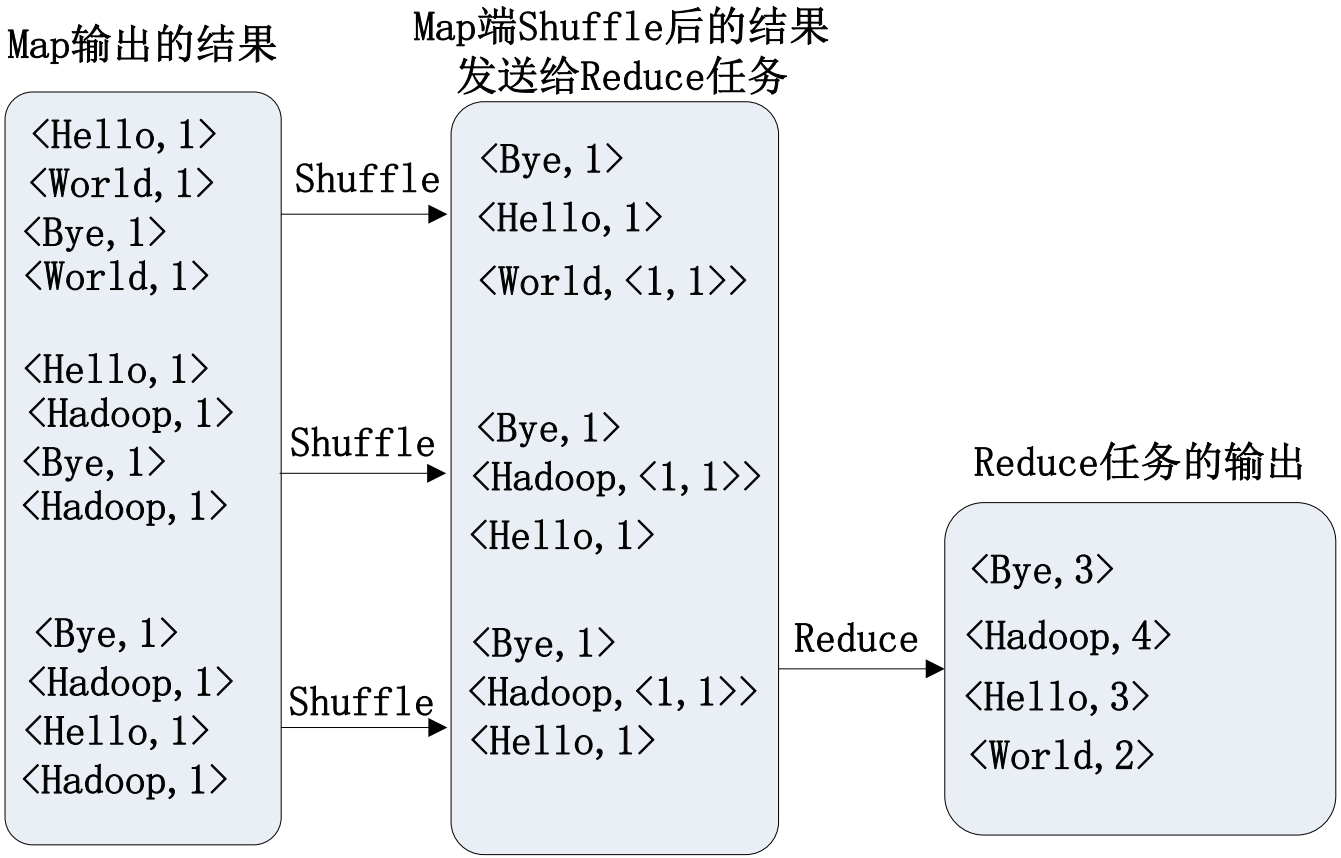
WordCount设计思路

- 首先，需要检查WordCount程序任务是否可以采用MapReduce来实现
- 其次，确定MapReduce程序的设计思路
- 最后，确定MapReduce程序的执行过程

一个WordCount执行过程的实例

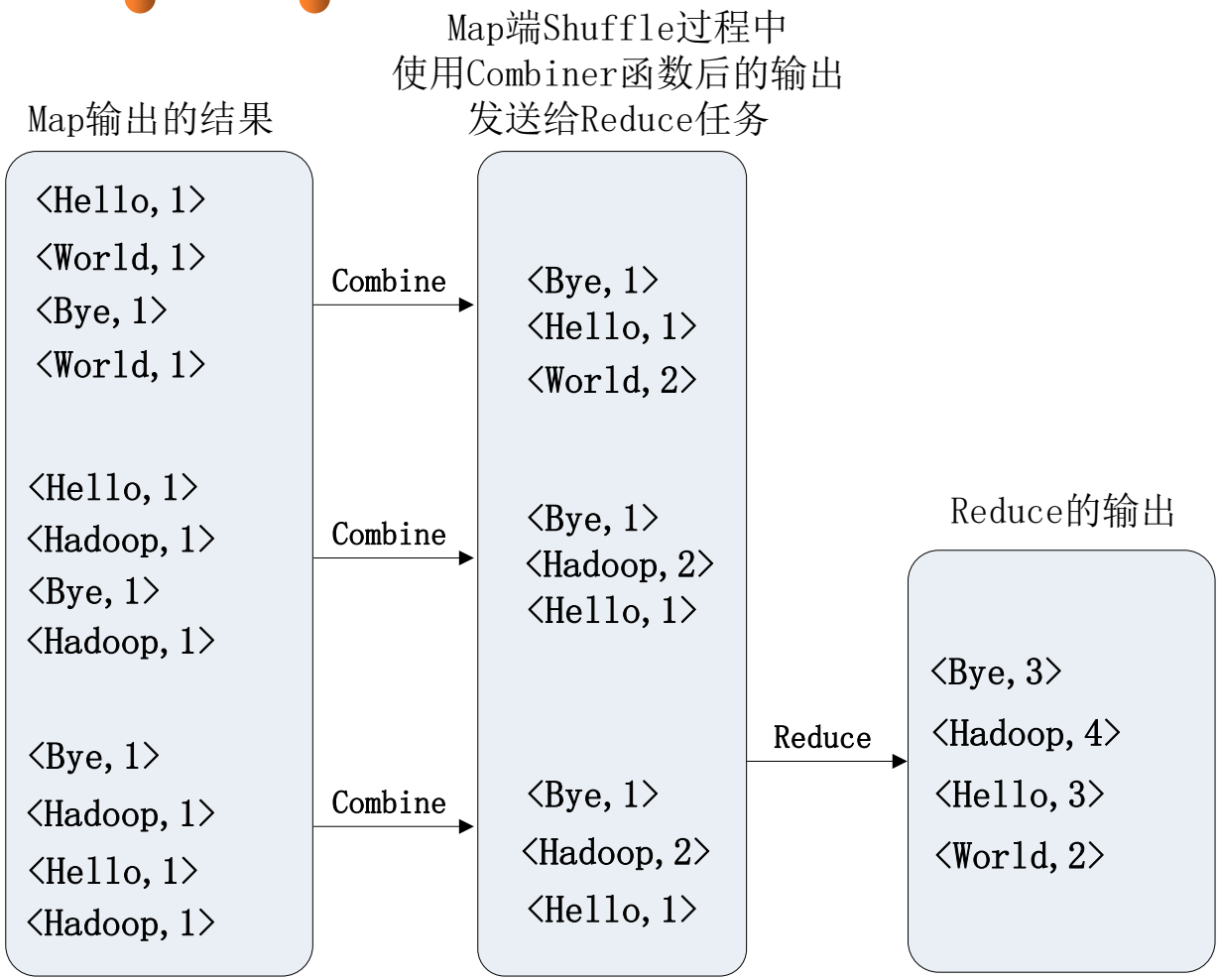
Map过程示意图

一个WordCount执行过程的实例



用户没有定义Combiner时的Reduce过程示意图

一个WordCount执行过程的实例



用户有定义Combiner时的Reduce过程示意图

MapReduce的具体应用

MapReduce的具体应用

MapReduce可以很好地应用于各种计算问题

- 关系代数运算（选择、投影、并、交、差、连接）
- 分组与聚合运算
- 矩阵-向量乘法
- 矩阵乘法

MapReduce的具体应用

用MapReduce实现关系的自然连接

雇员

Name	EmpId	DeptName
Harry	3415	财务
Sally	2241	销售
George	3401	财务
Harriet	2202	销售

部门

DeptName	Manager
财务	George
销售	Harriet
生产	Charles

雇员 ⋈ 部门

Name	EmpId	DeptName	Manager
Harry	3415	财务	George
Sally	2241	销售	Harriet
George	3401	财务	George
Harriet	2202	销售	Harriet

- 假设有关系R(A，B)和S(B,C)，对二者进行自然连接操作
- 使用Map过程，把来自R的每个元组<a,b>转换成一个键值对<b,<R,a>>，其中的键就是属性B的值。把关系R包含到值中，这样做使得我们可以在Reduce阶段，只把那些来自R的元组和来自S的元组进行匹配。类似地，使用Map过程，把来自S的每个元组<b,c>，转换成一个键值对<b,<S,c>>
- 所有具有相同B值的元组被发送到同一个Reduce进程中，Reduce进程的任务是，把来自关系R和S的、具有相同属性B值的元组进行合并
- Reduce进程的输出则是连接后的元组<a,b,c>，输出被写到一个单独的输出文件中

MapReduce的具体应用

用MapReduce实现关系的自然连接

Order

Orderid	Account	Date
1	a	d1
2	a	d2
3	b	d3

Map

Key	Value
1	“Order”,(a,d1)
2	“Order”,(a,d2)
3	“Order”,(b,d3)

Item

Orderid	Itemid	Num
1	10	1
1	20	3
2	10	5
2	50	100
3	20	1

Map

Key	Value
1	“Item”,(10,1)
1	“Item”,(20,3)
2	“Item”,(10,5)
2	“Item”,(50,100)
3	“Item”,(20,1)

Reduce

(1,a,d1,10,1)

(1,a,d1,20,3)

(2,a,d2,10,5)

(2,a,d2,50,100)

(3,b,d3,20,1)

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

致谢

部分图表、文字来自互联网，在此表示感谢！如有版权要求请联系：lics@hit.edu.cn，谢谢