

大数据技术-第四章：MapReduce分布式计算框架 MapReduce编程模型概述



CONTENTS

01. MapReduce概述 02 MapReduce基本设计思想
- 03 MapReduce优缺点 04 MapReduce编程模型简介



01

MapReduce概述



MapReduce是面向大数据并行处理的计算模型、框架和平台。具体包含以下3层含义。

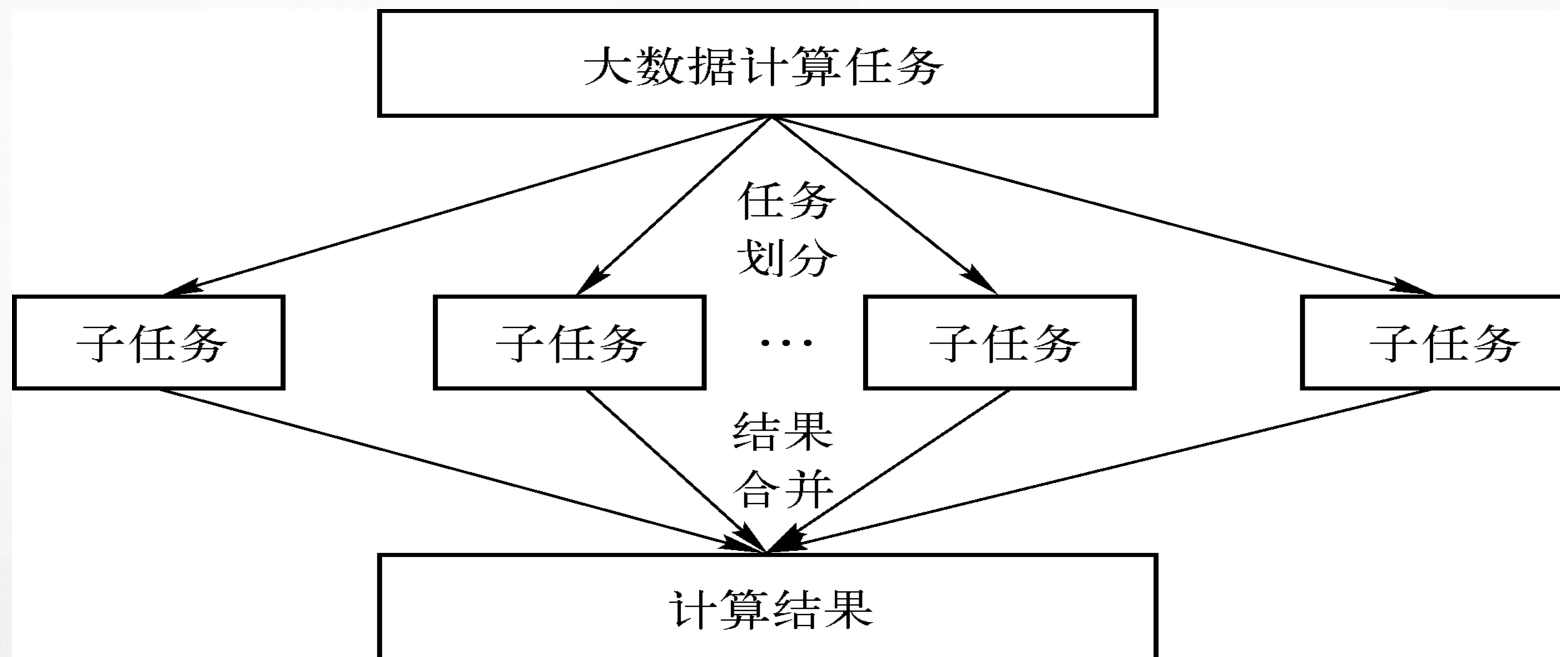
- MapReduce是一个并行程序的计算模型与方法
- MapReduce是一个并行程序运行的软件框架
- MapReduce是一个基于集群的高性能并行计算平台

02

MapReduce基本设计思想

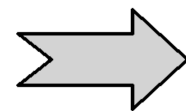


(1) 分而治之



(2) 抽象成模型

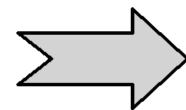
- 大量数据记录/元素进行重复处理
- 对每个数据记录/元素作感兴趣的
处理、获取感兴趣的中间结果信息



Map

- 排序和整理中间结果以利后续处理

- 收集整理中间结果
- 产生最终结果输出



Reduce

(3) 上升到框架

MapReduce以统一构架为程序员隐藏系统底层细节。该统一框架可负责自动完成以下系统底层主要相关的处理：

- 计算任务的自动划分和调度。
- 数据的自动化分布存储和划分。
- 处理数据与计算任务的同步。
- 结果数据的收集整理（Sorting（排序），Combining（合并），Partitioning（分区）等）。
- 系统通信、负载平衡、计算性能优化处理。
- 处理系统结点出错检测和失效恢复。

03

MapReduce优缺点



MapReduce优缺点

1. MapReduce的优点

- (1) MapReduce 易于编程
- (2) 良好的扩展性
- (3) 高容错性
- (4) 适合 PB 级以上海量数据的离线处理

2. MapReduce的缺点

- (1) 不适合实时计算
- (2) 不适合流式计算
- (3) 不适合DAG（有向无环图）计算

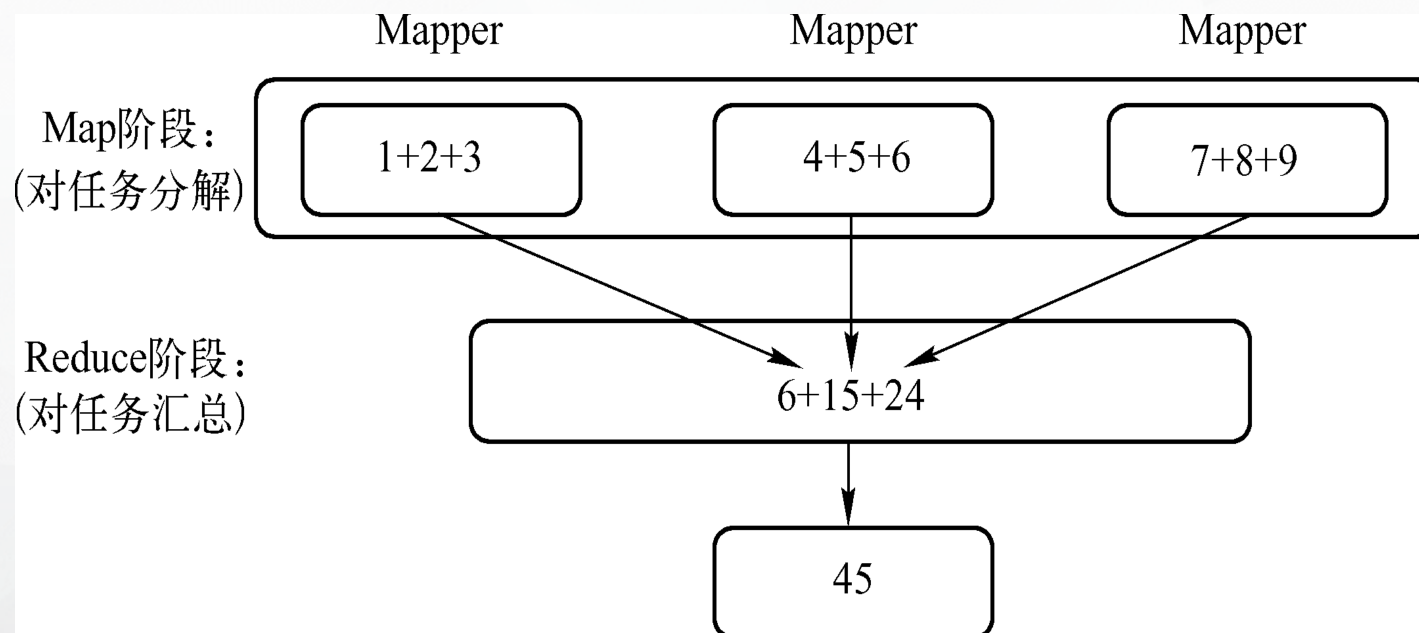
04

MapReduce编程模型简介

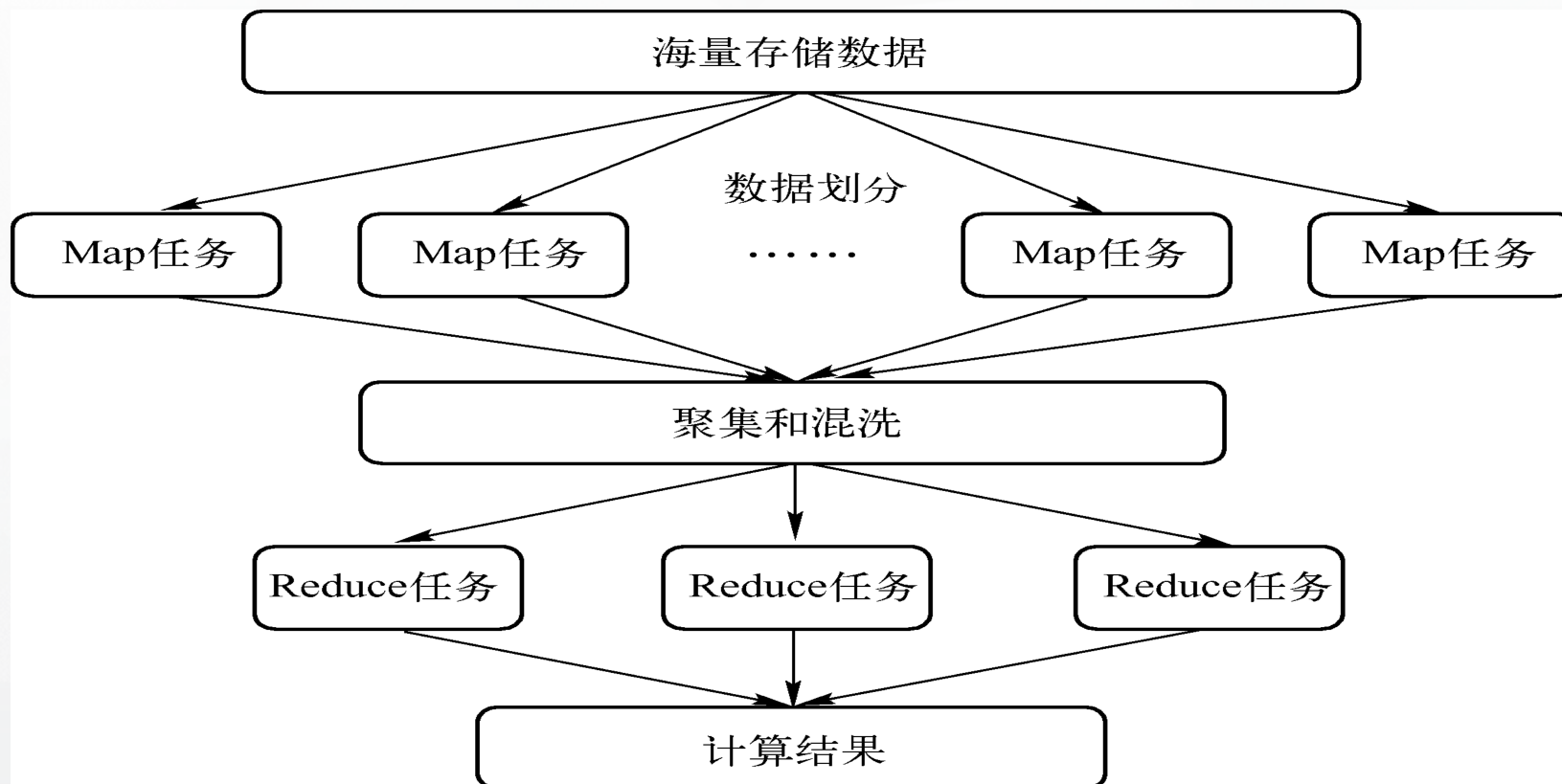


MapReduce编程模型简介

从MapReduce 自身的命名特点可以看出，MapReduce 由两个阶段组成：Map 和Reduce。用户只需实现Mapper和Reducer这两个抽象类，编写map() 和reduce() 两个函数，即可完成简单的分布式程序的开发。这就是最简单的MapReduce编程模型。



MapReduce编程模型简介



Turing AI 万维图灵 | 大数据系列课程

大数据

BIG
DATA

智 / 能 / 科 / 技 放 / 眼 / 未 / 来

