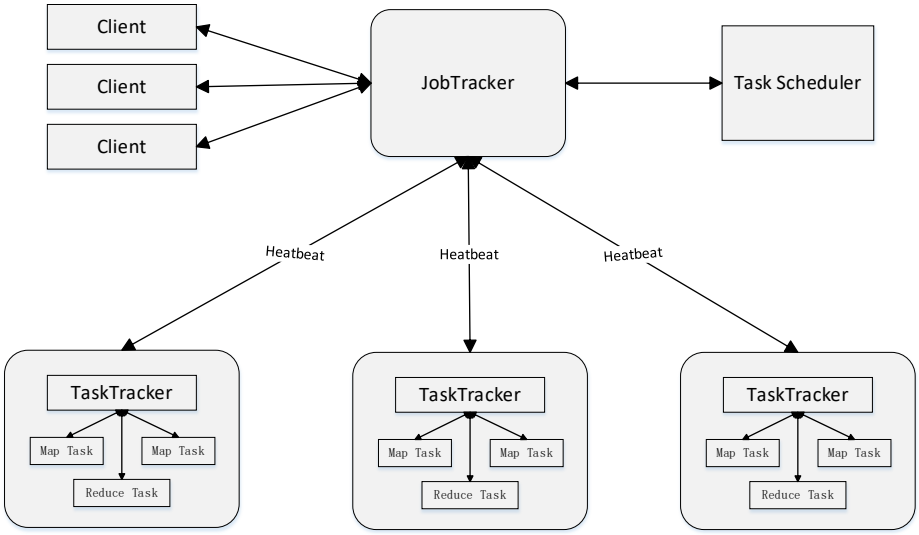
第七章

1. 分布式并行编程
   a) MPI: MPI 框架只是提供了通信机制，即任务之间同步和通信的手段。计算任务怎么分解，数据怎么划分，计算怎么实现，任务怎么合并等等问题都由程序开发者自己决定。
   b) 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 + Reduce。采用“分而治之”策略，一个存储在分布式文件系统中的大规模数据集，会被切分成许多独立的分片（split），这些分片可以被多个 Map 任务并行处理

	MPI	**MapReduce**
优点	1) 效率高 (job 都在内存中，不需要存储中间结果)； 2）速度快（如果资源充足，会启动所有 instance）	1) MapReduce job 可以起很多 instance，各个 instance 在计算的过程中互不干扰； 2）MapReduce job 没有 instance 间通信开销； 3）某个 instance 计算 failed，调度系统会自动重试，再次计算，并不影响其他结果，也不需要所有 instance 重新计算
缺点	1）如果资源不够，则 job 一直等待； 2）某个 instance 的失败会导致重新计算； 3) job instance 不能起太多，否则进程间通信开销大	1）MapReduce job 的计算的中间结果是以文件形式存储，效率较低
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

2. MapReduce 体系架构



Client：提交作业；查看运行状态

JobTracker：资源监控和作业调度

TaskTracker：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

Tsk: Map Task 和 Reduce Task

MapReduce 执行过程

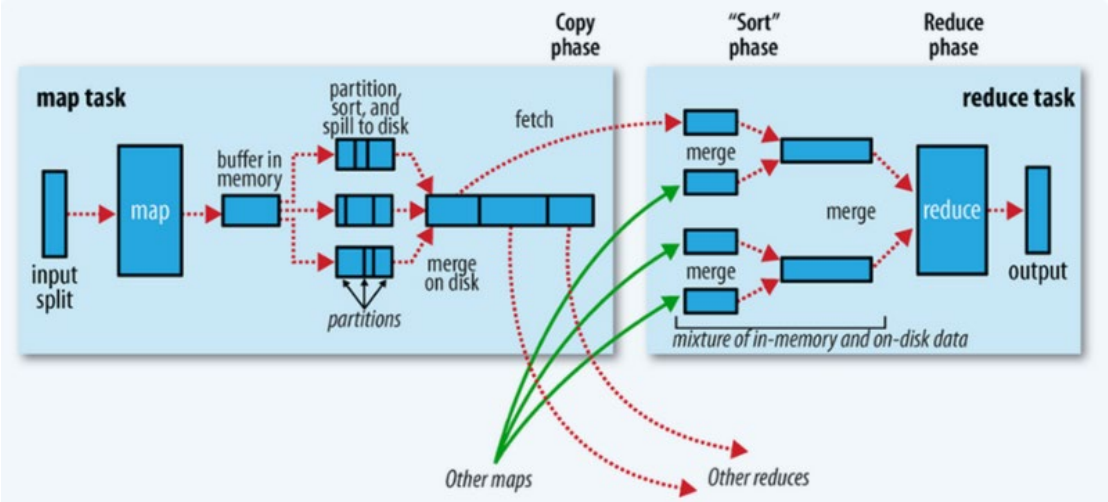
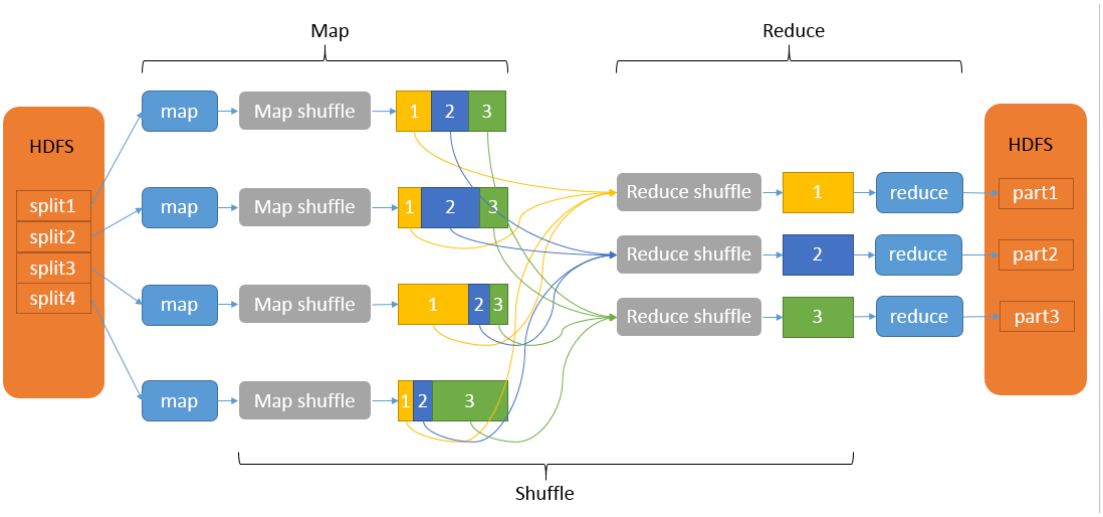
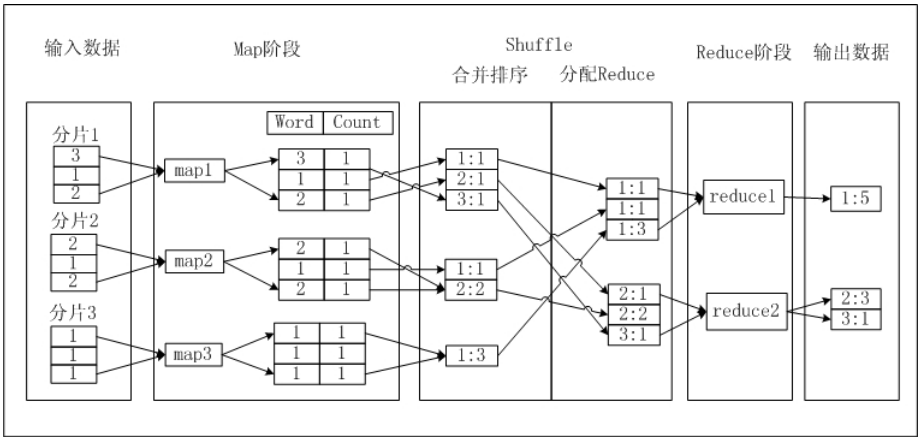


Figure 6-4 Shuffle and sort in MapReduce





1）输入数据：对文本进行分片，将每片内的数据作为单个 Map Task 的输入。

2) Map 阶段: Map 处理输入，每获取一个数字，将数字的 Count 设置为 1，并将此<Word, Count>对输出，此时以 Word 作为输出数据的 Key。

3）Shuffle > 合并排序：在 Shuffle 阶段前期，首先对每个 Map Task 的输出，按照 Key 值（即 Word 值）进行排序。排序后进行 Combiner 操作，即将 Key 值（Word 值）相同的 Count 累加，构成一个新的<Word, Count>对。此过程被称为合并排序。

4) Shuffle > 分配 Reduce：在 Shuffle 阶段后期，数据被发送到 Reduce 端。Reduce Worker 收到数据后依赖 Key 值再次对数据排序。

5）Reduce 阶段：每个 Reduce Task 对数据进行处理时，采用与 Combiner 相同的逻辑，将 Key 值（Word 值）相同的 Count 累加，得到输出结果。

6）输出结果数据。

第九章 数据仓库 Hive

1. 数据库与数据仓库的区别？例子？

数据库：传统的关系型数据库的主要应用，主要是**基本的、日常的事务处理**，例如银行交易。

数据仓库：数据仓库系统的主要应用主要是 **OLAP**（On-Line Analytical Processing），**支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果**。

例子：电商行业发展

2. Hive 和传统数据库的对比

Hive与传统数据库的对比

对比项	Hive	传统数据库
数据插入	支持批量导入，不可单条导入	支持单挑和批量导入
数据更新	不支持	支持
索引	有限索引功能，不像RDBMS有键的概念，可在某些列上建索引，加速一些查询操作。创建的索引数据，会被保存在另外的表中	支持
分区	支持，Hive表示分区形式进行组织的，根据“分区列”的值对表进行粗略划分，加快数据的查询速度	支持，提供分区功能来改善大型表以及具有各种访问模式的表的可伸缩性、可管理性，以及提高数据库效率
执行延迟	高，构建在HDFS和MR之上，比传统数据库延迟要高	低，传统SQL语句的延迟一般少于1秒，而HQL语句延迟可达分钟级。
扩展性	好，基于Hadoop集群，有很好的横向扩展性	有限，RDBMS非分布式，横向扩展（分布式添加节点）难实现，纵向扩展（扩展内存，CPU等）也很有限

3. 用 MapReduce 实现连接操作的基本原理（ch09_10, p22, p23）
4. 用 MapReduce 实现分组操作的基本原理（ch09_10, p24, p25）
5. Impala 的组成部分（ch09_10, p42）

第十章 Spark

1. 2013 年 Spark 加入 Apache 孵化器项目后发展迅猛，如今已成为 Apache 软件基金会最重要的**三大分布式计算系统开源项目**之一（Hadoop、Spark、Storm）
2. Spark 的特点（ch09_10, p72）
3. Spark 主要的四大功能/组件是什么？（ch09_10, p80， p109）
4. RDD 的概念及特点（ch09_10, p90, p93）

第十一章 流计算 Storm & Spark Streaming

1. 流计算的实际例子有哪些？（ch11_ch12, p6, p7）
2. 流数据的特征（ch11_ch12, p9）
3. 批量计算与实时计算（ch11_ch12, p10）
4. 流计算的概念 （ch11_ch12, p13）
5. 流计算处理的三个阶段（ch11_ch12, p19）
6. Spark Streaming 与 Storm 的对比（ch11_ch12, p55）

第十二章 Flink

1. Flink 基本介绍（ch11_ch12, p60）
2. Flink 的优势（ch11_ch12, p70）
3. Flink 编程模型（ch11_ch12, p83）

第十三章 图计算

1. 针对图数据的经典数据分析任务有哪些？p11
2. 后 Hadoop 时代的新三驾马车指的什么？p25
3. 图计算工具 Pregel 求解最大值实例 p32
4. Pregel 实现容错的机制 p46
5. Pregel 计算模型中每个顶点包含的成员变量有哪些？p83

6. 针对图计算中的图节点的相似性，有哪些方案来计算节点相似性，说出至少一种方案。P105

第十四章 数据可视化

1. 在可视化的发展历程中，有两个图对可视化的发展有重要作用，这两个图分别叫什么？p6，P14
2. 数据可视化的重要作用有哪些？p22-p25

第十五章

1. 工业大数据跟商业大数据在哪些方面存在不同？p4
2. 工业大数据有哪些特点，请分别从数据角度和应用角度进行解释 p5, p6
3. 工业大数据的典型分析场景有哪三类？p8
4. 经典的工业大数据分析方法 CRISP-DM（Cross-Industry Standard Process for Data Mining）的流程有哪些？简要解释每一步的含义。P19
5. 工业大数据分析的三类工程方法分别是哪些？p20-p25
6. 工业设备分类 p29