**学习日志**

姓名： 陈昭宇 学号：1812977 日期：6.25

**学习内容安排：**

上午：

MapReduce工作原理

下午：

MapReduce的类型与格式

**学习反馈：**

已掌握知识：

作业的工作原理：

1. 请求
2. 检查输出说明
3. 对作业输入划分
4. 将资源输入划分
5. 准备执行

初始化

任务的分配

失败（尝试也可被杀死这个和失败时杀死时不同的，不计入运势尝试次数；）

1. Tasktracker失败如果失败发送心跳停止或者很少，会从tasktracker池中移除
2. Jobtracker失败 Hadoop没有处理这种失败的机制

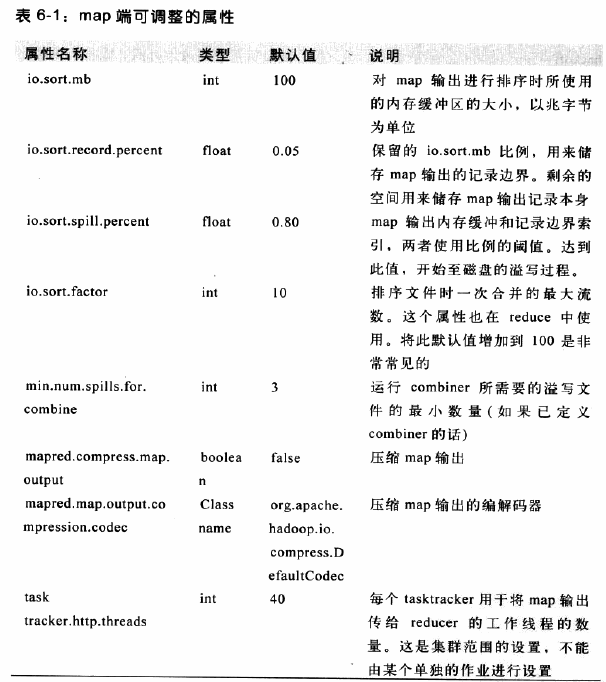
作业的调度

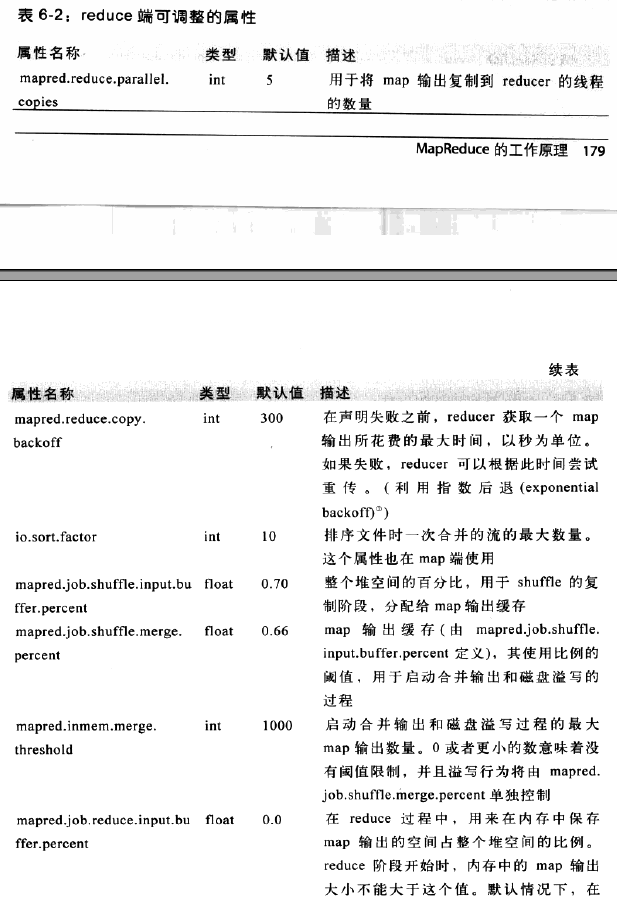
可选择调度器

1. Fair Scheduler 让每个用户公平的共享集群，即公平分配；支持抢占

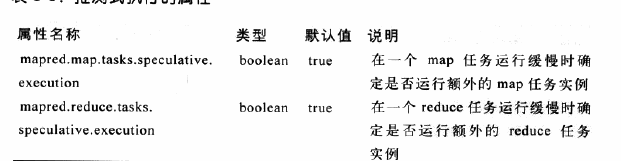
Shuffle和排序 map输出传到reducer作为后者的输入即称为shuffle。

每个map任务都有一个环形内存缓冲区超出会溢写，新建一个溢写文件，完成后会合并



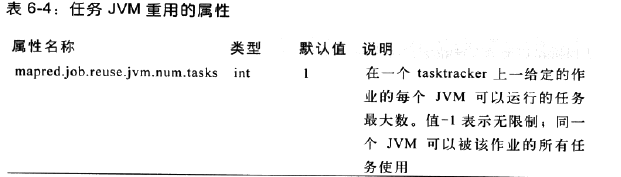


推测式执行（优化拖后腿的任务）



会减少集群效率

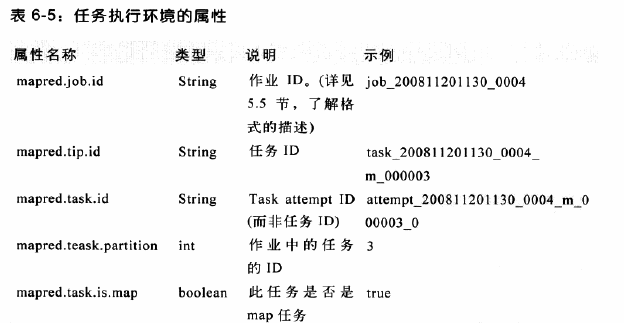
任务JVM重用



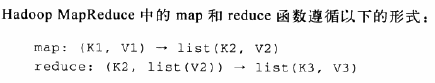
重用后不会在一个JVM中运行，而是单独的JVM

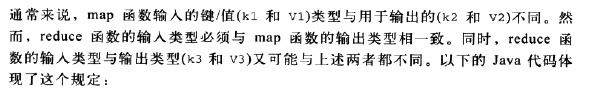
跳过坏记录

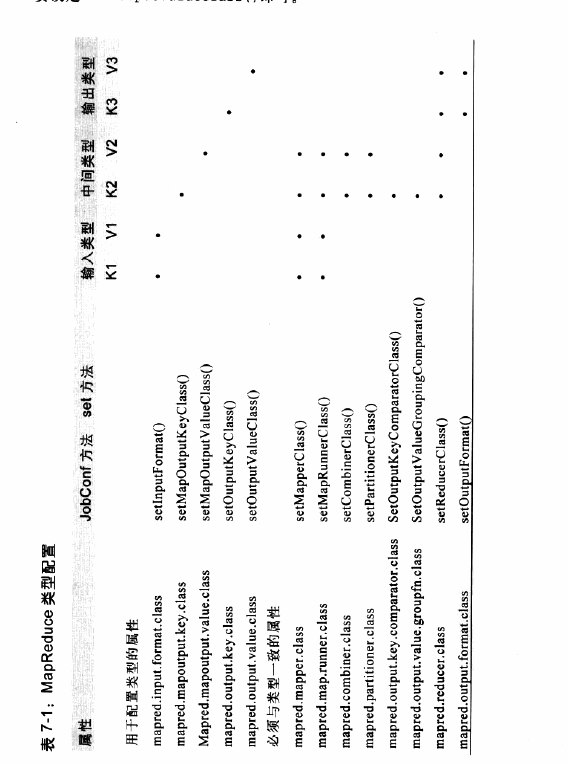
任务执行环境

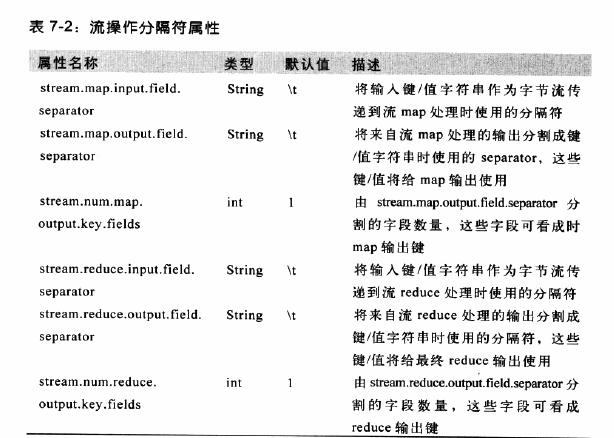


MapReduce 的类型与格式



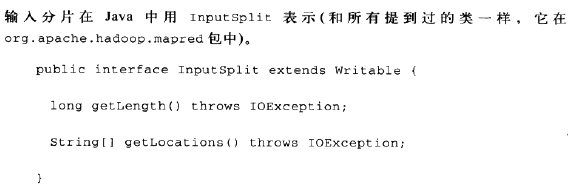




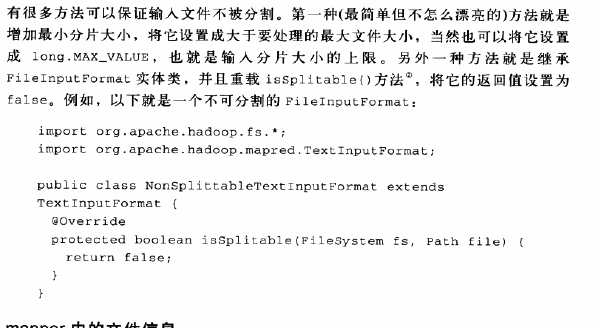


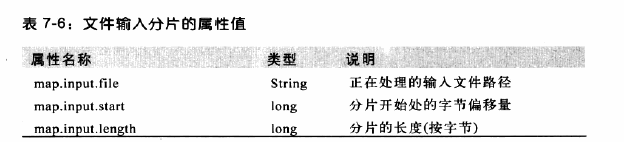
输入格式

输入分片和记录



避免分割





文本输入

一次收入多行文本

二进制输入

数据库格式的输入

输出格式

未掌握知识： 没有看例子中的具体代码

学习心得： 对大量数据的处理和流处理。是hadoop的特点。今天主要学的是原理（即它对数据处理的具体操作和类型与格式，刚好的帮助我了解了hadoop的运行逻辑。