MapReduce 算法设计

MapReduce 排序算法

利用框架会自动对 key 排序的性质

要自定义 partitioner 来划分 key，或者 Hadoop 内置 TotalOrderPartitioner

MapReduce 单词同现分析

单词同现矩阵是一个二维的 $N \times N$ 矩阵（N 是语料库中不同的单词总数）

m_{ij} 表示单词 i，j 在一定范围内同现的次数

算法思路：map 读入一篇文档，将同现的词组作为 key，同现次数作为 value 发射出去，由 reduce 接受并统计总体的次数

MapReduce 文档倒排索引

倒排索引：根据一个 term，给出所有含有这个 term 的文档列表

思路：map 将词语作为 key，文档作为 value 发射出去，由 reduce 统计并得出倒排索引

如果需要考虑单词的词频，位置等，就需要改进算法，将这些信息称为 payload，最终获得的结果是一个 term 对应一系列 posting 的列表，每个 posting 包含文档 id 和 payload

如果按照简单的方式处理，内存会成为瓶颈，reduce 节点的内存可能不能放下与一个 term 相关的所有文档。解决方法是将 term-docid 统一作为 key 发送，然后定制 partitioner 将 term 相同的键值对分到同一个 reducer，然后 reduce 每次读取 term-docid 对应的一系列 value 时只需要判断 term 与前一个 term 是否相同