并行计算

MAPREDUCE

孙 晖

15130120141

题目 1：

老师在课堂上讲了 MapReduce 和 Hadoop 的相关技术，并布置了以下作业：

有一组文件（比如 100 个.txt 文件），每个文件中的内容为中文、英文、西班牙文、法文四种文字的字/单词。

要求：

（1）使用 MapReduce 并行计算框架完成所有文件中每种文字的单词个数的统计。

（2）中文每一个字视作一个单词。

（3）写出使用 MapReduce 完成本次作业任务的方案。

（4）不要求写代码。

（5）作业要在 7.9 之前上交电子版。

答：

首先将问题简化后画图。再做步骤的详细描述。首先可以明确的是有四个 key 值，分别为：

key-中文

key-英文

key-西班牙文

key-法文

100 个.txt 文件把每一个文件看作一个 split，当然也有其它的分法，这里这样分后并作简化，以其中三个.txt 文件为例来画图（实际可能就是整体输入 100 个文件然后按照每一个文件一个 split 来划分，这里为了画图方便而做简化。）。

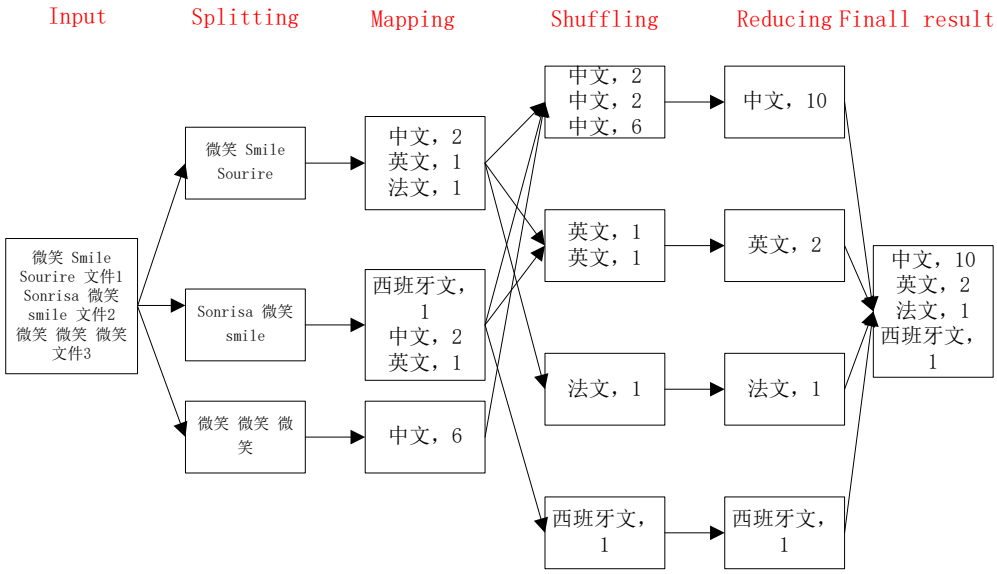
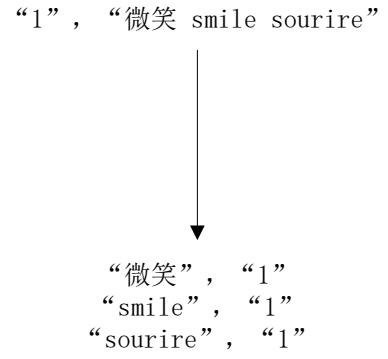


图 1 用 MapReduce 的过程

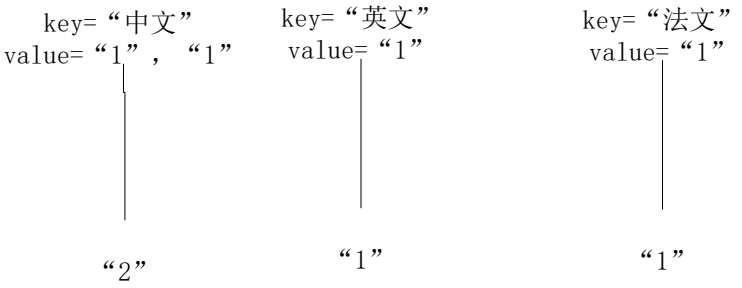
按照老师要求，对步骤进行如下列举：

方案：首先，分别统计每个文件中四种语言的单词出现的次数，然后累加不同文件中同一种语言出现的次数。说明：input_key 为 split 编号，key 为 shuffle 后的 key，也就是要输出的 key 的值。

[1] 输入：100 个.txt 文件。
[2] 用户实现 map 函数，输入为：
Input_key={0,1,2…,99}
value=文本的内容
[3] map 输出 key/value 值
例如：对文件 1 而言：
“1”“微笑 smile sourire”



[4] MapReduce 运行系统把所用相同的 key 的记录收集到一起（shuffle/sort）
[5] 用户实现 reduce 函数对一个 key 对应的 values 计算求和



[6] Reduce 输出<key,sum>：
“中文”，“2”
“英文”，“1”
“法文”，“1”
整体的步骤是这样的，对 100 个.txt 文件进行整体输入，且在输入过程进行

map 操作<input_key，value>分别对应着<1,文件 1 内容>，<2，文件 2 内容>…，然后 100 个文件输入进去后就分成 100 个 split，对于每一个 split 进行 map 操作后产生这样一个 list，<key,value>，分别对应<中文，数量>，<英文，数量>，<法文，数量>，<西班牙文，数量>。然后进行 shuffle 操作，把相同的 key 的中间结果汇集到相同结点上，然后进行 reduce 操作，reduce 操作就是把同一个 key 的 value 值加起来，然后输出结果。