大数据开发技术

东北林业大学

卢洋

第三章

MapReduce框架原理

WritableComparable

排序

排序的分类

(1) 部分排序

MapReduce根据输入记录的键对数据集进行排序，保证输出的每个文件内部有序。

(2) 全排序

最终输出结果只有一个文件，且文件内部有序。实现方式是只设置一个ReduceTask。该方法在处理大型文件时效率极低，因为一台机器处理所有文件，完全丧失了MapReduce所提供的并行架构。

(3) 辅助排序(GroupingComparator分组)

在Reduce端对key进行分组。应用于：在接收key为bean对象时，想让一个或多个字段相同(全部字段比较不相同)的key进入到同一个reduce方法时，可采用分组排序。

(4) 二次排序

在自定义排序过程中，如果compareTo中判断条件为两个即为二次排序。

自定义排序

- 原理分析

bean对象作为key传输，需要实现WritableComparable接口，重写compareTo方法，就可以实现排序