
MapReduce总结
InputFormat类
hadoop默认使用的是TextInputForamat类，即一次读取一行文本，让后将偏移量作为key，一行line文本作为value返回
CombineTextInputFormat类可以把多个小文件合并为一个切片，提高处理的效率
逻辑处理接口Mapper类
用户自定义的Mapper根据需求设计map(),setup(),cleanup()方法
Partitioner类分区
如果用户不设置分区，hadoop使用默认的分区，分区号=key.hashCode()&Integer.MAXVALUE % numReduce
用户可以自定义分区，注意有几个分区应该有几个reducetask，每个reduceTask负责处理一个特定的分区
Comparable排序
用自定义Bean对象作为map的key输出时，要指明key的排序规则，必须通过实现WritableComparable接口，重写其中大的compateTo方法。
部分排序 shuffle阶段
全排序 通常只有一个reduceTask，对拉取来的一个分区进行全排序作为reduce的输入
二次排序 归并排序？
Combiner聚合操作
合理的利用combiner可以减少reduce的处理负担，还可以降低磁盘的存储量
Reducer类
用户根据自定义的reducer实现reduce(),setup(),clearup()方法
OutputFormat类
hadoop默认使用TextOutputFormat类，每一个kv作为一个line输出
用户可以使用自定义的OutputFormat类