hadoop mr优化

摘要: hadoop mapreducer 优化

1，比较器尽量不要让mr产生序列化和反序列化的转换，参考WritableComparable类

2，reducer严重的数据倾斜，可以考虑自定义分区器

     但是之前可以尝试使用combiner来压缩数据看是否能解决问题

3，Map阶段不使用正则表达式

4，split使用StringUtils，据测试性能远高于（String,Scanner,StringTokenizer），writableutils等工具类可以尝试去用一种

5，hadoop产生的数据要压缩，减少数据传输和I/O开销

6，使用结果或中间数据采用紧凑二进制的文件格式avro,sequencefiles，而不是使用txt，也能减少I/O和网络开销

7，能不用reducer就不用，map直接完成，也是为shuffle/sort阶段的开销

8，hive解释器太烂，尽量自己写mr,当然hiveql直观和读性比java程序来得高