# 课堂笔记

郑博 21:26:20

如果hashpartition后，不发生shuffle的个别机器又跑的很慢，这种情况咋办呢？

答：找到跑的慢的原因（查看负载）。

谢国贵：

老师，昨晚讲的RDD的懒计算和持久化不是很理解，假如我先从文件读取RDD数据，然后做一系列Transform，最后做了一个Action，此时scala才真正开始从文件读取RDD，然后Transform，最后Action吗？

如果此次Action后，我再来一个Action，并且我没有做持久化的话，那scala是不是又从读取文件开始直到我的这次Action？如果在Action前做了持久化，那就不用再去生成RDD数据，直接从内存或硬盘读取RDD数据，然后直接Action返回，是这样理解吗？@Cloudy

答：对的。

郑博 21:44:49

数据倾斜的问题咋解决

答：分布式计算引擎无法避免数据倾斜的问题。

1. 数据分布不均衡所致：map端倾斜
2. 聚合，key的值有的多有的少，造成Reduce端数据倾斜。

MR中：Group by 语法造成倾斜时，有个set...参数进行优化