天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

Spark初级课程

讲师：平常心

数据倾斜

数据倾斜是并行处理的数据集中，某一部分(partition)的数据显著多于其他部分，从而使得该部分的处理速度成为整个数据集处理的瓶颈。

原因：

在Stage中包含的N个Task，Task可以并行处理，如果其中的Task耗时几秒中结束，而耗时最多的Task需要分钟时间，则这个Stage所耗费的时间主要由最慢的那个Task决定。因为同一个Stage的Task执行相同的计算，不同Task之间耗时的差异主要由该Task所处理的数据量决定。Stage的数据来源主要是①.外部数据源，如HDFS,Kafka;②.读取上一个Stage的shuffle数据。

天善智能
TIANSHAN SOFT

解决方法

1.增加并行度分散数据：默认的HashPartitioner使大量不同的Key对应数据分配到同一个Task，使该Task所处理的数据远远大于其他Task，如果增加Shuffle时的并行度，可以降低原Task的数据量，缓解数据倾斜现象。

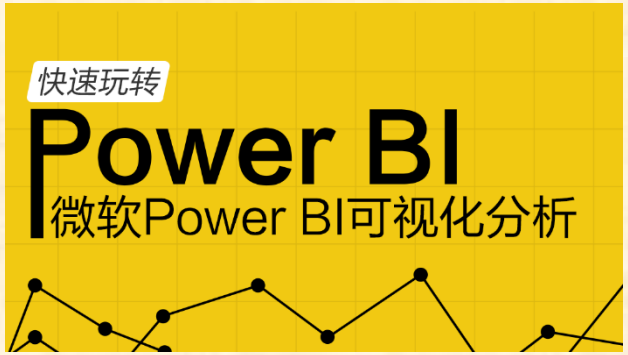
2.自定义Partitioner代替HashPartitioner，将所有不同的Key分配到不同的Task中。

3. Spark SQL中通过广播机制转化为Map Join，避免Shuffle带来的数据倾斜。

4.为倾斜的Key增加随机前/后缀，尤其是数据量特别大的Key,让原来相同的Key数据变为Key不同的数据。

天善智能
TIANSHAN SOFT

更多商业智能BI和大数据精品视频尽在 www.hellobi.com



BI、商业智能
数据挖掘 大数据
数据分析
R Python
机器学习
Tableau
QLIKVIEW
Hive Hadoop
BIWORK
BAO胖子 seng
贝克汉姆
曹浩

天善智能
TIANSHAN SOFT