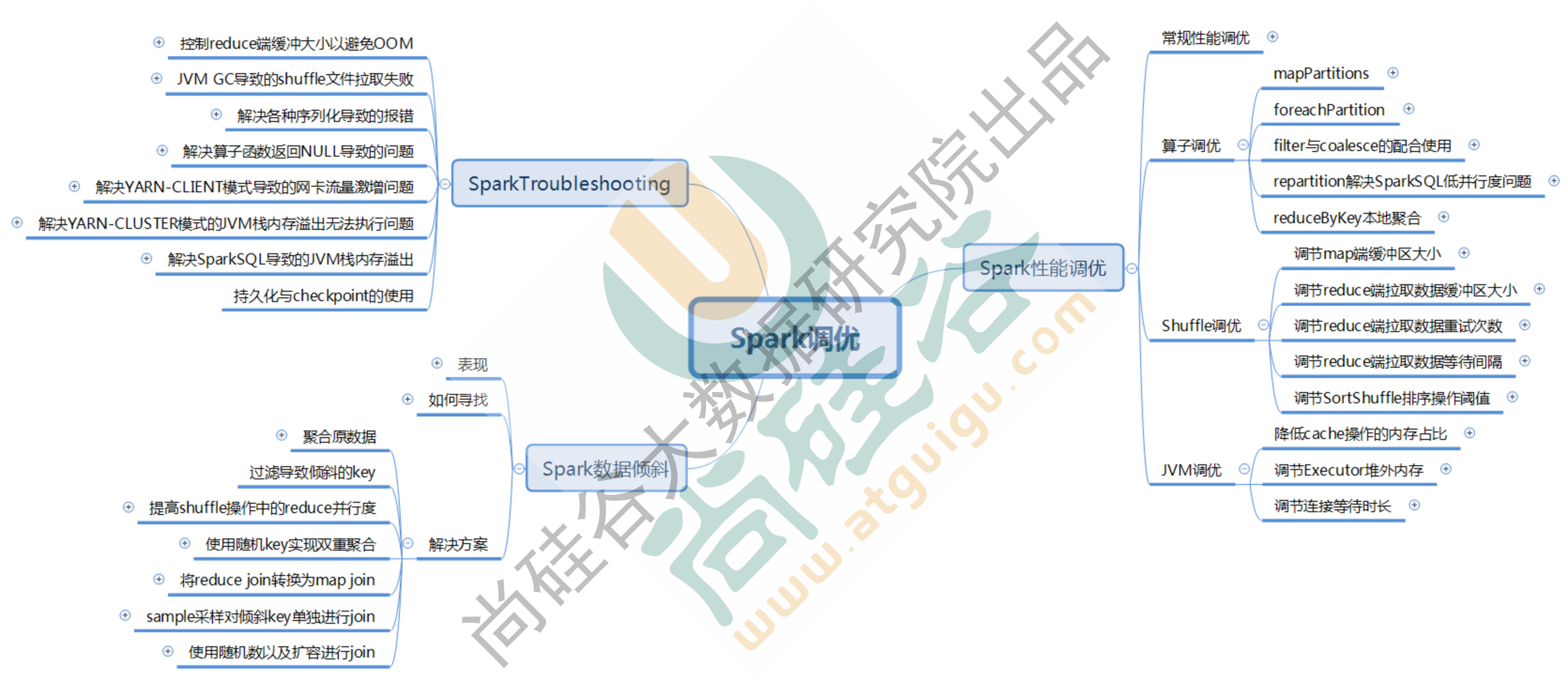
Spark调优
SparkTroubleshooting
控制reduce端缓冲大小以避免OOM
JVM GC导致的shuffle文件拉取失败
解决各种序列化导致的报错
解决算子函数返回NULL导致的问题
解决YARN-CLIENT模式导致的网卡流量激增问题
解决YARN-CLUSTER模式的JVM栈内存溢出无法执行问题
解决SparkSQL导致的JVM栈内存溢出
持久化与checkpoint的使用
Spark数据倾斜
表现
如何寻找
解决方案
聚合原数据
过滤导致倾斜的key
提高shuffle操作中的reduce并行度
使用随机key实现双重聚合
将reduce join转换为map join
sample采样对倾斜key单独进行join
使用随机数以及扩容进行join
Spark性能调优
常规性能调优
算子调优
mapPartitions
foreachPartition
filter与coalesce的配合使用
repartition解决SparkSQL低并行度问题
reduceByKey本地聚合
Shuffle调优
调节map端缓冲区大小
调节reduce端拉取数据缓冲区大小
调节reduce端拉取数据重试次数
调节reduce端拉取数据等待间隔
调节SortShuffle排序操作阈值
JVM调优
降低cache操作的内存占比
调节Executor堆外内存
调节连接等待时长
尚硅谷大数据研究院出品
www.atguigu.com