

spark executor内存足够使用，却报错SparkOutOfMemory: Unable to acquire xxx bytes of memory，如何解决？是否与coalesce有关？

原创 慢点走 于 2020-12-23 21:56:13 发布 653 收藏

分类专栏: spark 项目 文章标签: executor 内存足够使用 内存溢出 spark coalesce

文章目录

[问题描述](#)

[解决办法](#)

问题描述

从 spark 查出来数据后可以再次保存到hive中。当数据量不是很大的时候，为了避免产生过多小文件，可以使用重分区来解决。

重分区有两个方法：coalesce 和 repartition。本文重点不是介绍这两个方法，所以大概说一下区别：coalesce 不需要 shuffle，而 repartition 需要 shuffle。而为了避免shuffle，则可以使用coalesce。

当代码如下时：

```
1 df.coalesce(1)
2 .write
3 .mode(SaveMode.Overwrite)
4 .saveAsTable("test.test")
```

即使executor的 内存 远远大于数据量时，有时仍然会报错：Unable to acquire xxx bytes of memory

我的情况是：executor内存是32G，核是8个，聚合后的数据量大概是50M左右，但是仍然出现了上述的错误。但可以肯定的是，不是内存不够了

解决办法

经过查证，coalesce在参数过小时（例如1），并行度不够，所以只会在很少的节点上运行。

而repartition(1)虽然会加入shuffle步骤，但是上游的分区会并行执行。

所以只要将代码中的coalesce(1)改为repartition(1)即可。

```
1 df.repartition(1)
2 .write
3 .mode(SaveMode.Overwrite)
4 .saveAsTable("test.test")
```