spark性能调优--jvm调优

2017年05月22日 20:10:00 kequanjade 阅读数：393

版权声明：本文为博主原创文章，未经博主允许不得转载。 https://blog.csdn.net/keyuquan/article/details/72629605

一.问题切入

调用spark 程序的时候，在获取数据库连接的时候总是报 内存溢出 错误

（在ideal上运行的时候设置jvm参数 -Xms512m -Xmx1024m -XX:PermSize=512m -XX:MaxPermSize=1024M，不会报

二.jvm参数 和 saprk 参数 和内存四区 解读

1.内存四区

1、栈区（stack）：由编译器自动分配释放 ，存放函数的参数值，局部变量的值等。其操作方式类似于数据结构中的栈。

2、堆区（heap：一般由程序员分配释放， 若程序员不释放，程序结束时可能由OS回 收 。注意它与数据结构中的堆是两回事，分配方式倒是类似于

3、数据区：主要包括静态全局区和常量区，如果要站在汇编角度细分的话还可以分为很多小的区。

全局区（静态区）（static）：全局变量和静态变量的存储是放在一块的，初始化的全局变量和静态变量在一块区域，未初始化的全局变量和未初始
量在相邻的 另一块区域。 程序结束后有系统释放

常量区：常量字符串就是放在这里的。 程序结束后由系统释放

4、代码区：存放函数体的二进制代码。

参考： http://blog.csdn.net/wu5215080/article/details/38899259

2.jvm 参数

-Xms512m -Xmx1024m-XX:PermSize=512m -XX:MaxPermSize=1024M

-Xms	JVM初始分配的堆内存	默认是设备物理内存的 1/64
-Xmx	JVM最大允许分配的堆内存，按需分配	默认是设备物理内存的 1/4
-XX:PermSize	JVM初始分配的非堆内存	默认是设备物理内存的 1/64
-XX:MaxPermSize	JVM最大允许分配的非堆内存	默认是设备物理内存的 1/4

参考： http://www.cnblogs.com/mingforyou/archive/2012/03/03/2378143.html

3.spark参数

-- driver-memory ： driver运行的内存大小，默认1G driver：sparkcontext ，sqlContext等运行的地方， sparkcontext ，sqlContext 一般运行在栈内存

-- executor-memory : executor的内存大小，默认1G executor： rdd 等运行的地方 ,rdd 一般运行在栈内存

conf spark.storage.memoryFraction=0.3 spark用于缓存rdd的内存百分比(剩下的内存用来保证任务运行时各种其它内存空间的需要)，默认0.6（和运
没有关系）

得出：

栈内存 正比于 driver-memory : 内存被 sparkcontext，sqlContext 等固定占用，和数据库连接没有多大关系

栈内存 正比于 executor-memory ； executor-memory 分两种： rdd 和其他（包含获取获取 数据库连接的内存）

开发者调查 AI开发者大会日程曝光 全场课程特价最后 1 天 hadoop hdfs Apache Hado-- 登录 注册

三.问题分析和解决

方向：增大executor-memory 和减小 conf spark.storage.memoryFraction 的值 ,根据具体环境而定

命令方式：

```
nohup spark-submit \
--masteryarn  \
--executor-memory 1024M \
--confspark.storage.memoryFraction=0.3 \
--classcom.xiaopeng.bi.gamepublish.GamePublishKpi \
/home/hduser/projs/xiaopeng_bi.jar60 >> /home/hduser/projs/logs/gamepublishkpi.log &
```

代码方式：

```
 val sparkConf = newSparkConf().setAppName(this.getClass.getName.replace("$",""))
     .set("spark.default.parallelism", "60") //  1. 调节并行度
     .set("spark.serializer","org.apache.spark.serializer.KryoSerializer") // 3.序列化方式
     .set("spark.shuffle.consolidateFiles", "true")// 4. shuffle 过程中  合并小文件
     .set("spark.storage.memoryFraction", "0.4");// 5.cache占用的内存占比
     .set("spark.sql.shuffle.partitions", "60")// 6.shuffle 时 partion的个数
```

想对作者说点什么？ 我来说一句

ctelinla： 厉害了,我的哥!! (1年前 #1楼)

spark调优 JVM调优 84
我们的堆内存分为：新生代，和年老代， 年轻代又分为：Eden区，幸存一区，幸存二区， 每一次访对象的时候，… 来自： mn_kw的博客

spark之jvm调优 186
转自:https://blog.csdn.net/lxhandlbb/article/details/52987928一、性能调优分类：1.常规性能调优： 分配资源，并… 来自： weixin_41804049的博客

Spark性能调优之——JVM调优之原理概述 以及降低cache操作的内存占比 2669
性能调优分成好几块：1.常规性能调优： 分配资源，并行度。。等。2.JVM调优：JVM相关的参数。【没有大家想… 来自： coderlaw's study

spark JVM调优之原理概述以及降低cache操作的内存占比 1546
每一次放对象的时候，都是放入eden区域，和其中一个survivor区域；另外一个survivor区域是空闲的。 当eden区… 来自： 涛涛的专栏

Spark调优 JVM调优 170
Spark调优 JVM调优 占个位置 以后补上 来自： 九师兄-梁川川

Spark调优的策略 1327
1. RDD的持久化 cahce() persist() checkpoint() 2. 避免创建重复的RDD 3.尽可能复用同一个RDD 类似于多个RDD... 来自： fanyao4144的博客

spark学习-60-源代码：ContextCleaner清理器 711
Spark运行的时候，会产生一堆临时文件，临时数据，比如持久化的RDD数据在磁盘上，没有持久化的在内存中，… 来自： 九师兄-梁川川

Spark性能优化：JVM参数调优 1万

Spark性能优化：JVM参数调优 年轻代：主要是用来存放新生的对象。 老年代：主要存放应用程序中生命周期长的… 来自： Ganymede的Hadoop世界

spark-调优-JVM 197

开启日志调试 在让G1 GC跑起来之后，我们下一步就是需要根据GC log，来进一步进行性能调优。首先，我们要让… 来自： program哲学

相关热词 SPARK spark和 sparK spark to spark as

Spark开发性能调优 2320

Spark开发性能调优 标签（空格分隔）： Spark –Write By Vin 1. 分配资源调优 Spark性能调优的王道就是分配资源,… 来自： fly_li的博客

博主推荐

段智华 关注 827篇文章

Mr_Smile2014 关注 103篇文章

jaryle 关注 306篇文章

sparksql调优之第一弹 1904

1，jvm调优这个是扯不断，理还乱。建议能加内存就加内存，没事调啥JVM，你都不了解JVM和你的任务数据。sp… 来自： Spark高级玩法

Spark性能调优1-测试记录 2341

Spark作为Zeppelin的SQL底层执行引擎，通过Thriftserver处理jdbc连接，为提高硬件资源利用率、IO带宽和内存利… 来自： 简单就好

Spark面对OOM问题的解决方法及优化总结 3.2万

分布式计算系统最常见的问题就是OOM问题，本文主要讲述Spark中OOM问题的原因和解决办法，并结合笔者实践… 来自： 拱头的专栏

spark性能调优（一）JVM调优 486

性能调优 JVM调优原理概述 不够炫但是很有用 够炫听起来高端的 1、常规性能调优：分配资源、并行度。。。等 2… 来自： 谭谦的博客

JVM学习笔记（四）------内存调优 4.2万

首先需要注意的是在对JVM内存调优的时候不能只看操作系统级别Java进程所占用的内存，这个数值不能准确的反… 来自： 走向架构师之路

spark性能调优之调节数据本地化等待时长 3061

本地化级别 PROCESS_LOCAL：进程本地化，代码和数据在同一个进程中，也就是在同一个executor中；计算数… 来自： 涛涛的专栏

Spark JVM调优之调节executor堆外内存与连接等待时长 133

executor堆外内存的调优 有时候，如果你的spark作业处理的数据量特别特别大，大约在几亿的数据量，然后spark… 来自： mn_kw的博客

“戏”说spark---spark 内存管理详解 178

Spark 作为一个基于内存的分布式计算引擎，其内存管理模块在整个系统中扮演着非常重要的角色。理解 Spark 内… 来自： weixin_35602748的博客

Spark性能调优-总结分享 3915

1、Spark调优背景 目前Zeppelin已经上线一段时间，Spark作为底层SQL执行引擎，需要进行整体性能调优，来提… 来自： shieh的专栏

JVM系列:解决JVM最大内存设置问题 7633

你知道JVM内存最大能调多大吗，这里和大家分享一下JVM最大内存方面的内容，Java虚拟机具有一个堆，堆是运… 来自： liugw_768的博客

kequanjade 关注

原创	粉丝	喜欢	评论
38	2	0	1

等级： 博客 3 访问： 2万+

积分： 577 排名： 10万+

最新文章

clickhouse 性能测试

clickhouse 部署

es 分词器

es query string

es 分页搜索 和 deep paging 问题

个人分类

elasticsearch学习笔记	12篇
spark	15篇
hadoop	4篇
协作框架	2篇
android&机顶盒	2篇

展开

归档

2018年9月	2篇
2018年8月	12篇
2018年2月	2篇
2017年12月	1篇
2017年10月	1篇

展开

热门文章

spark序列化溢出
阅读量：2742

hadoop任务卡死
阅读量：2577

Spark任务卡死
阅读量：2467

Rdd的 foreach 和 foreachPartition
阅读量：2378

hbase问题总结
阅读量：1127

最新评论

spark性能调优--jvm调优
ctelinla：厉害了,我的哥!!

联系我们

扫码联系客服 下载CSDN APP

QQ客服 kefu@csdn.net
客服论坛 400-660-0108
工作时间 8:00-22:00

关于我们 招聘 广告服务 网站地图
百度提供站内搜索 京ICP证09002463号
©2018 CSDN版权所有

网络110报警服务 经营性网站备案信息
北京互联网违法和不良信息举报中心
中国互联网举报中心