

# spark参数配置调优

原创

2018年04月19日 22:52:07

25705

SPARK配置参数的两个地方：

1. \$SPARK\_HOME/conf/spark-env.sh 脚本上配置。配置格式如下：

```
export SPARK_DAEMON_MEMORY=1024m
```

2. 编程的方式（程序中在创建SparkContext之前，使用System.setProperty（“xx”，“xxx”）语句设置相应系统属性值），即在spark-shell下配置

如：scala> System.setProperty("spark.akka.frameSize","10240m")

## 一、环境变量spark-env.sh配置项

SCALA\_HOME #指向你的scala安装路径

MESOS\_NATIVE\_LIBRARY #如果你要在Mesos上运行集群的话

SPARK\_WORKER\_MEMORY #作业可使用的内存容量，默认格式1000M或者 2G（默认：所有RAM去掉给操作系统用的1 GB）；每个作业独立的内存空间由SPARK\_MEM决定。

SPARK\_JAVA\_OPTS #添加JVM选项。你可以通过-D来获取任何系统属性 eg: SPARK\_JAVA\_OPTS+="-Dspark.kryoserializer.buffer.mb=1024"

SPARK\_MEM #设置每个节点所能使用的内存总量。他们应该和JVM ‘s -Xmx选项的格式保持一致（e.g.300m或1g）。注意：这个选项将很快被弃用支持系统属性spark.executor.memory，所以我们推荐将它使用在新代码中。

SPARK\_DAEMON\_MEMORY #分配给Spark master和worker守护进程的内存空间(默认512M)

SPARK\_DAEMON\_JAVA\_OPTS #Spark master和worker守护进程的JVM选项（默认：none）

## 二、System Properties

Property Name	Default	Meaning
spark.executor.memory	512m	Amount of memory to use per executor process, in the same format as JVM memory strings (e.g. '512m', '2g').
spark.akkaframeSize	10m	Maximum message size to allow in "control plane" communication (for serialized tasks and task results), in MB. Increase this if your tasks need to send back large results to the driver (e.g. using collect() on a large dataset).
spark.default.parallelism	8	Default number of tasks to use for distributed shuffle operations (groupByKey, reduceByKey, etc) when not set by user.



陈兴振

原创

粉丝

喜欢

55

64

32

等级：博客5

访问量：51

积分：3772

排名：1万



## 大数据可视化



## 博主最新文章

html页面跳转及参数传递

GET请求json示例

bootstrap table通过ajax获取后示在table

射线法判断地图上点是否在多边形

vs 2008下配置opencv

## 文章分类

mysql

mysql cluster

java

linux

lvs

流媒体

展开

## 文章存档

2017年8月

2016年3月

2014年12月



2014年7月

2014年5月

2014年3月











展开

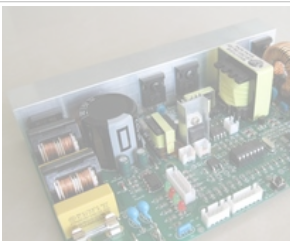

spark.akka.frameSize: 控制Spark中通信消息的最大容量（如 task 的输出结果），默认为10M。当处理大数据时，task 的输出可能会大于这个值，需要根据实际数据设置一个更高的值。如果是这个值不够大而产生的错误，可以从 worker 的日志 中进行排查。通常 worker 上的任务失败后，master 的运行日志上出现“Lost TID: ”的提示，可查看失败的 worker 的日志文件(\$SPARK\_HOME/worker/下面的log文件) 中记录的任务的 Serialized size of result 是否超过10M来确定。spark.default.parallelism: 控制Spark中的分布式shuffle过程默认使用的task数量，默认为8个。如果不做调整，数据量大时，就容易运行时间很长，甚至是出Exception，因为8个task要处理那么多的数据。注意这个值也不是说设置得越大越好。spark.local.dir : Spark 运行时的临时目录，map 的输出文件，保存在磁盘的 RDD 等都保存在这里。默认是 /tmp 这个目录，而一开始我们搭建的小集群上/tmp 这个目录的空间只有2G，大数据量跑起来就出 Exception（“No space left on device”）了。

- 参考：
-   
  

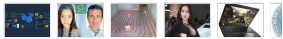
- <http://rdc.taobao.com/?p=533>
- <http://spark.incubator.apache.org/docs/0.7.3/configuration.html>
- <https://github.com/amplab/shark/blob/master/conf/shark-env.sh.template>
- <http://www.cnblogs.com/vincent-hv/p/3316502.html>
- [http://www.07net01.com/linux/Sparkdulibushumoshi\\_545676\\_1374481945.html](http://www.07net01.com/linux/Sparkdulibushumoshi_545676_1374481945.html)
- [https://groups.google.com/forum/#!searchin/spark-users/java.lang.OutOfMemoryError\\$3A\\$20GC\\$20overhead\\$20limit](https://groups.google.com/forum/#!searchin/spark-users/java.lang.OutOfMemoryError$3A$20GC$20overhead$20limit)

目前您尚未登录，请 [登录](#) 或 [注册](#) 后参与评论

- 浅谈mysql集群  
 67205
- Hive内置数据类型  
 64345
- spark从hdfs上读取文件运行woi  
 55451
- mysql插入表中的中文字符显示为  
问号的解决方法  
 51103
- spark参数配置调优  
 25697
- flash上制作一个按钮，控制动画  
停  
 19156
- Linux下利用rpm包安装mysql  
 16462
- mysql binlog使用  
 14918
- Hbase集群安装配置  
 10740
- mysql cluster的常见问题  
 8494







大功率稳压电源



联系我们



- 请扫描二维码联系
-  webmaster@csdn.net
-  400-660-0101
-  QQ客服 

关于 招聘 广告服务 

©1999-2018 CSDN版权所有  
京ICP证09002463号

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

spark 笔记(二) 参数设置和调优

 xyl520 2015年06月26日 10:17  2001

在迁移相关的spark程序到yarn的过程中间，对有些地方的配置进行了调整和优化，总结起来，常用的一些设置如下：1. spark.serializer 对象的序列化设置可以设置成spark的序列化...

Map output statuses were bytes which exceeds spark.akka.frameSize

spark.akka.frameSize 是worker和driver通信的每块数据大小，控制Spark中通信消息的最大容量（如 task 的输出结果），默认为10M。当处理大数据时，task 的输出...

 u012302488 2016年05月17日 15:27  1170

UI设计师凭什么拿下年薪40W？程序员的我不平衡！

牛逼的UI设计师是这么炼成的？



Spark技术内幕：Shuffle Map Task运算结果的处理

Shuffle Map Task运算结果的处理这个结果的处理，分为两部分，一个是在Executor端是如何直接处理Task的结果的；还有就是Driver端，如果在接到Task运行结束的消息时，如何对S...

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录

 注册

## 关于SPARK\_WORKER\_MEMORY和SPARK\_MEM

在spark中最容易混淆的是各种内存关系。本篇讲述一下SPARK\_WORKER\_MEMORY和SPARK\_MEM。SPARK\_WORKER\_MEMORY是计算节点worker所能支配的内存，各个节点可...

 book\_mmicky 2014年05月13日 15:23 4357

## spark性能调优

 lihaitao000 2016年06月21日 18:33 3171

spark性能调优有一些措施，下面说说我用的一些调优手段。1.RDD分片数和executor个数的协调 要想充分的使数据并行执行，并且能充分的利用每一个executor，则在rdd的个数与exec...

## 开源商城系统

盘点8款好用的开源商城系统

百度广告



## Spark 常用配置项与优化配置项

 u012307002 2015年03月12日 22:01 5709

Spark 常用配置项与优化配置项 1、配置加载顺序：SparkConf方式 > 命令行参数方式 > 文件配置方式。应用程序SparkConf 优先级高 2.sp...

## spark history server内存不足服务自动挂掉

 levy\_cui 2016年05月13日 17:33 2123

版本：Spark 1.5.2 built for Hadoop 2.4.0 今天spark的history server自己挂掉了，查看日志：16/05/13 14:12:30 WARN DFS C...

## Spark运行模式（一） - - - - Spark独立模式

除了可以在Mesos或者YARN集群管理器上运行Spark外，Spark还提供了独立部署模式。你可以通过手动启动一个master和workers,或者使用提供的脚本来手动地启动单独的集群模式。你也可以...

 happyAnger6 2015年07月26日 22:12 32436

## Spark集群基于Zookeeper的HA搭建部署笔记

 panguoyuan 2015年01月26日 15:46 2768

1.环境介绍（1）操作系统RHEL6.2-64（2）两个节点：spark1(192.168.232.147),spark2(192.168.232.152)（3）两个节点上都装好了Hado...

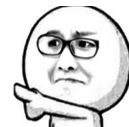
## spark2.0 history server 内存溢出解决

 houzhizhen 2016年12月20日 10:52 902

从下面命令构建类可以看到，如果你想增加history server的内存，只能设置SPARK\_DAEMON\_MEMORY。在SPARK\_HISTORY\_OPTS里设置 "-Xmx"不起作用。 ...

## 码农怎能不懂英语？！试试这个数学公式

老司机教你一个数学公式秒懂天下英语



## 【问题跟踪】KryoException: java.io.IOException: No space left on device

今天在对LDA进行不同参数训练模型，记录其avglogLikelihood和logPerplexity，以便判断模型训练是否收敛时，产生了一个令人极度崩溃的事儿：程序在辛辛苦苦跑了7.3h后...挂了...

 yhao2014 2016年04月15日 10:35 11336

## 关于 “No space left on device的原因”

 liudayu\_hikvision 2011年01月17日 19:59 36763

看到这个错误，第一个反应是磁盘空间满了；但df一看，每个分区的空间都还富余的很。从munin 的监控图表上看 Filesystem usage 也很平稳，但下面的 Inode usage 就有问题了，...

在实现“马踏棋盘”问题时，因为程序出错不停循环写文件耗尽硬盘空间。#df/dev/sdc1	20799540	19751436	0
100% /home/s...			

Spark : java.io.IOException: No space left on device

This is because Spark create some temp shuffle files under /tmp directory of you local system.You ca...

 dupihua

 2

 04月12日 15:01

 191

增加spark worker的内存和datanode的内存方法

 wonder4 2016年09月08日 22:43  1180

spark:\$SPARK\_HOME/bin/spark-env.sh修改SPARK\_WORKER\_MEMORY=4g\$HADOOP\_HOME/etc/hadoop/hadoop-env.sh修改HADOOP\_HEAPSIZE=4g

如何学习大数据

大数据学习路线

百度广告

