



博客 (<http://blog.csdn.net/?ref=toolbar>)

学院 (<http://edu.csdn.net?ref=toolbar>)

下载 (<http://download.csdn.net?ref=toolbar>)

GitChat (<http://gitbook.cn/?ref=csdn>) ...

0

Q



登录 (<https://passport.csdn.net/account/login?ref=toolbar>)

注册 (<https://passport.csdn.net/account/mobile/register?ref=toolbar&action=mobileRegister>)

Spark 序列化问题全解

原创

2017年11月01日 10:56:42

标签 : spark (<http://so.csdn.net/so/search/s.do?q=spark&t=blog>) /

序列化 (<http://so.csdn.net/so/search/s.do?q=序列化&t=blog>)



437

在Spark应用开发中，很容易出现如下报错：

```
1 org.apache.spark.SparkException: Task not serializable
2   at org.apache.spark.util.ClosureCleaner$.ensureSerializable(ClosureCleaner.scala:304)
3   at org.apache.spark.util.ClosureCleaner$.org$apache$spark$util$ClosureCleaner$$clean(ClosureCleaner.scala:294)
4   at org.apache.spark.util.ClosureCleaner$.clean(ClosureCleaner.scala:122)
5   at org.apache.spark.SparkContext.clean(SparkContext.scala:2058)
6   ...
7   Caused by: java.io.NotSerializableException
```

该报错意思是用户代码的transformation操作中包含不可序列化的对象引用。

本文主要从以下三个方面解释Spark 应用中序列化问题。

- 加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！
- ✕
- 1、Java序列化含义？
- 2、Spark代码为什么需要序列化？
- 3、如何解决Spark序列化问题？

1、Java序列化含义？

Spark是基于JVM运行的，其序列化必然遵守Java的序列化规则。

序列化就是指将一个对象转化为二进制的byte流（注意，不是bit流），然后以文件的方式进行保存或通过网络传输，等待被反序列化读取出来。序列化常被用于数据存取和通信过程中。

对于java应用实现序列化一般方法：

- class实现序列化操作是让class 实现Serializable接口，但实现该接口不保证该class一定可以序列化，因为序列化必须保证该class引用的所有属性可以序列化。
- 这里需要明白，static和transient修饰的变量不会被序列化，这也是解决序列化问题的方法之一，让不能序列化的引用用static和transient来修饰。（static修饰的是类的状态，而不是对象状态，所以不存在序列化问题。transient修饰的变量，是不会被序列化到文件中，在被反序列化后，transient变量的值被设为初始值，如int是0，对象是null）



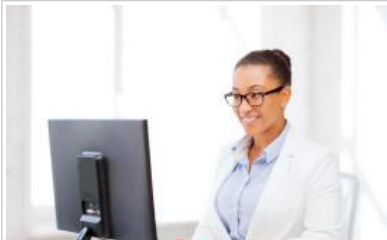
xwc35047 (<http://blog.csdn.net/xwc35047>)

+ 关注

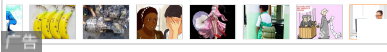
(<http://blog.csdn.net/xwc35047>)

码云

原创	粉丝	喜欢	未开通
31	30	19	(https://github.com/xwc35047)



在线律师



他的最新文章

更多文章 (<http://blog.csdn.net/xwc35047>)

Spark与Hive的Beeline运行机制 (<http://blog.csdn.net/xwc35047/article/details/7896469>)

2018小记 (<http://blog.csdn.net/xwc35047/article/details/78956729>)

spark添加JVM监控信息到grafana显示 (<http://blog.csdn.net/xwc35047/article/details/78915358>)

Spark ERROR: org.apache.spark.shuffle.FetchFailedException 问题追查 (<http://blog.csdn.net/xwc35047/article/details/78771148>)

Spark作业提交执行流程源码图 (<http://blog.csdn.net/xwc35047/article/details/78732738>)

文章分类

- | | |
|-----------------------------------------------------------------------------------------------------------------------------------------|-----|
| spark经验总结 (http://blog.csdn.net/xwc35047/article/details/7896469) | 19篇 |
| Spark入门教程 (http://blog.csdn.net/xwc35047/article/details/7896469) | 6篇 |
| zeppelin (http://blog.csdn.net/xwc35047/article/details/7896469) | 4篇 |
| 30分钟概览新技术 (http://blog.csdn.net/xwc35047/article/details/7896469) | 4篇 |
| antlr4介绍 (http://blog.csdn.net/xwc35047/article/details/7896469) | 1篇 |

- 此外还可以实现readObject()方法和writeObject()方法来自定义实现序列化。（具体用例见参考链接）

展开

2、Spark的transformation操作为什么需要序列化？

Spark是分布式执行引擎，其核心抽象是弹性分布式数据集RDD，其代表了分布在不同节点的数据。Spark的计算是在executor上分布式执行的，故用户开发的关于RDD的map，flatMap，reduceByKey等transformation操作（闭包）有如下执行过程：

1. 代码中对象在driver本地序列化
2. 对象序列化后传输到远程executor节点
3. 远程executor节点反序列化对象
4. 最终远程节点执行

故对象在执行中需要序列化通过网络传输，则必须经过序列化过程。

3、如何解决Spark序列化问题？

如果出现NotSerializableException报错，可以在spark-default.xml文件中加入如下参数来开启SerializationDebugger功能类，从而可以在日志中打印出序列化出问题的类和属性信息。

1	spark.executor.extraJavaOptions -Dsun.io.serialization.extendedDebugInfo=true
2	spark.driver.extraJavaOption -Dsun.io.serialization.extendedDebugInfo=true

对于scala语言开发，解决序列化问题主要如下几点：

- 在Object中声明对象（每个class对应有一个Object）
- 如果在闭包中使用SparkContext或者SqlContext，建议使用SparkContext.get() and SQLContext.getActiveOrCreate()
- 使用static或transient修饰不可序列化的属性从而避免序列化。
注：scala语言中，class的Object

对于java语言开发，对于不可序列化对象，如果本身不需要存储或传输，则可使用static或transient修饰；如果需要存储传输，则实现writeObject()/readObject()使用自定义序列化方法。

此外注意

对于Spark Streaming作业，注意哪些操作在driver，哪些操作在executor。因为在driver端（foreachRDD）实例化的对象，很可能不能在foreach中运行，因为对象不能从driver序列化传递到executor端（有些对象有TCP链接，一定不可以序列化）。所以这里一般在foreachPartitions或foreach算子中来实例化对象，这样对象在executor端实例化，没有从driver传输到executor的过程。

1	dstream.foreachRDD { rdd =>
2	val where1 = "on the driver"
3	rdd.foreach { record =>
4	val where2 = "on different executors"
5	}
6	}
7	}

参考资料：

Avoid NotSerializable Error in Spark Job (flystarhe.github.io/2016/09/05/feature-engineering/)
spark not serializable problem (https://stackoverflow.com/questions/22592811/task-not-serializable-java-io-notserializableexception-when-calling-function-ou)
Spark Streaming / Tips on Running Streaming Apps inside Databricks
(https://docs.cloud.databricks.com/docs/latest/databricks_guide/07%20Spark%20Streaming/17%

文章存档

2018年1月 (http://blog.csdn....	2篇
2017年12月 (http://blog.csd...	3篇
2017年11月 (http://blog.csd...	8篇
2017年8月 (http://blog.csdn....	1篇
2017年7月 (http://blog.csdn....	1篇

展开

他的热门文章

Spark入门教程（1）——spark是什么及发展趋势概述 (http://blog.csdn.net/xwc35047/article/details/51072145)
12814

Spark入门教程(2)---开发、编译配置 (http://blog.csdn.net/xwc35047/article/details/51119608)
5481

30分钟概览Spark Streaming 实时计算 (http://blog.csdn.net/xwc35047/article/details/55668963)
4258

spark入门教程（3）--Spark 核心API开发 (http://blog.csdn.net/xwc35047/article/details/51146622)
4212

30分钟概览Spark分布式计算引擎 (http://blog.csdn.net/xwc35047/article/details/60330528)
3908



联系我们

网站客服 微博客服
(http://wpa.qq.com/msgrd?v=3&uin=2431299880&site=qq&r
(http://e.weibo.com/csdnsupport/f
webmaster@csdn.net
(mailto:webmaster@csdn.net)
400-660-0108

20Tips%20on%20Running%20Streaming%20Apps%20inside%20Databricks.html)
Java 序列化的高级认识 (https://www.ibm.com/developerworks/cn/java/j-lo-serial/index.html)
什么是writeObject 和readObject ? 可定制的序列化过程 (http://bluepopopo.iteye.com/blog/486548)

版权声明：本文为博主原创文章，未经博主允许不得转载。 -- Bruce Xu

京ICP证09002463号
(http://www.miibeian.gov.cn/)
关于
(http://www.csdn.net/company/about.h
招聘
(http://www.csdn.net/company/recruit.f
广告服务
(http://www.csdn.net/company/marketi
阿里云
Copyright © 1999-2018
CSDN.NET, All Rights Reserved



Spark中的序列化机制 u011491148 2015年07月16日 13:54 5523
Spark中的序列化机制 在写Spark的应用时，尝尝会碰到序列化的问题。例如，在Driver端的程序中创建了一个对象，而在各个Executor中会用到这个对象 —— 由于Driver端代码与Exec...
(http://blog.csdn.net/u011491148/article/details/46910803)

Spark闭包与序列化 bluishglc 2016年03月21日 11:27 11780
本文原文出处: http://blog.csdn.net/bluishglc/article/details/50945032 严禁任何形式的转载，否则将委托CSDN官方维护权益！在Spark的官方文...
(http://blog.csdn.net/bluishglc/article/details/50945032)

Spark 中的序列化 u013494310 2016年05月18日 10:34 4447
1.序列化常用于网络传输和数据持久化以便于存储和传输，Spark通过两种方式来创建序列化器 val serializer = instantiateClassFromConf[Serializer]...
(http://blog.csdn.net/u013494310/article/details/51441883)

spark2.0 Java序列化和Kyro序列化测试 houzhizhen 2016年11月23日 17:36 716
先列出测试结果，后面附有测试代码，可以发现，改成Kyro序列化之后，可以节约大量空间。 JavaSerial KyroSerial Int 812 ...
(http://blog.csdn.net/houzhizhen/article/details/53308239)

spark性能调优之使用Kryo序列化 hutao_hadoop 2016年09月28日 21:58 1414
在SparkConf中设置一个属性，spark.serializer，org.apache.spark.serializer.KryoSerializer类；注册你使用到的，需要通过Kryo序列化的， ...
(http://blog.csdn.net/hutao_hadoop/article/details/52694374)

多邮箱一次绑定√



网易邮箱大师，提高工作效率的法宝

[Spark优化]在Spark中使用Kryo序列化 lovebyz 2016年05月10日 21:26 6101
conf.set("spark.serializer" , "org.apache.spark.serializer.KryoSerializer") conf.registerKryoClasses...
(http://blog.csdn.net/lovebyz/article/details/51366782)

Spark性能优化第四季-序列化 u011007180 2016年07月17日 12:34 1840
一：Spark性能调优之序列化 1、之所以进行序列化，最重要的原因是内存空间有限（减少GC的压力，最大化避免Full GC的产生，因为一旦产生Full GC，则整个Task处于停止状态！）、减少磁盘...

(<http://blog.csdn.net/u011007180/article/details/51931771>)

SparkTask未序列化(Tasknotserializable)问题分析

问题描述及原因分析 在编写Spark程序中，由于在map等算子内部使用了外部定义的变量和函数，从而引发Task未序列化问题。然而，Spark算子在计算过程中使用外部变量在许多情形下确实在所难免，...

(<http://blog.csdn.net/javastart/article/details/51206715>)

Spark Executor Driver资源调度小结



u014388509 2014年08月23日 01:08 16329

Spark中Executor的生成策略

(<http://blog.csdn.net/u014388509/article/details/38763985>)

第35课Spark Master、Worker、Driver、Executor工作流程详解

第35课Spark Master、Worker、Driver、Executor工作流程详解



zhumr 2016年09月12日 23:10 8793

(<http://blog.csdn.net/zhumr/article/details/52518506>)

spark算子中用到scalal类，由于未序列化报错

由于spark算子用到的class没有实现序列化，报错如下所示 15/1 1/23 14:43:47 ERROR Executor: Exception in task 0.0 in stage 4...

(<http://blog.csdn.net/u014487509/article/details/49994965>)

Spark性能调优之——在实际项目中使用Kryo序列化

set (“spark.serializer” , “ org.apache.spark.serializer.KryoSerializer”) Java的序列化机制，ObjectOutputStream/Ob...

(<http://blog.csdn.net/lxhandlbb/article/details/52987863>)

spark-java-task未序列化



taizitj 2016年12月19日 11:18 769

原文链接-spark编程task未序列化 问题描述及原因分析 在编写Spark程序中，由于在map等算子内部使用了外部定义的变量和函数，从而引发Task未序列化问题。然而，Spark算子在...

(<http://blog.csdn.net/taizitj/article/details/53736763>)

spark Task序列化问题



qq_14950717 2016年05月16日 18:45 1137

1、问题描述及原因分析 在编写Spark程序中，由于在map，foreachPartition等算子内部使用了外部定义的变量和函数，从而引发Task未序列化问题。然而，Spark算子在计算过程中使用...

(http://blog.csdn.net/qq_14950717/article/details/51427207)

为什么XML需要序列化和反序列化



mobei1983 2016年06月03日 20:22 1285

为什么需要序列化 注意：“为避免编译错误，为可序列化的类添加了无参数构造函数。” MSDN的定义：序列化是将对象状态转换为可保持或可传输的形式过程。序列化的补集是反序列化，后者将流转换为...

(<http://blog.csdn.net/mobei1983/article/details/51581499>)

什么是序列化，为什么要序列化。



u011215133 2016年04月18日 11:10 8795

转自：网络--（忘记从哪看到了）整理：Bob 在学习分布式计算的时候，老师上课提到序列化这个概念。当时有些懵逼，不知道什么是序列化，下来查了一下，原来在Java里面，序列化就是和S...

(<http://blog.csdn.net/u011215133/article/details/51177843>)

spark序列化问题解决



zhanghytc 2017年05月10日 17:45 210

saprk 未序列化 Exception in thread "main" org.apache.spark.SparkException: Task not serializable

(<http://blog.csdn.net/zhanghytc/article/details/71554445>)



Spark Task未序列化(Task not serializable)问题分析及解决

在编写Spark程序中，在map等算子内部由于使用了外部定义的变量和函数，从而导致出现Task未序列化问题，而由于Spark算子内部往往需要根据外部指定的配置进行计算，因此使用外部变量有时在所难免。为...

 sogerno1 2015年05月23日 16:07  11697

(<http://blog.csdn.net/sogerno1/article/details/45935159>)

spark Task序列化问题



 qq_14950717 2016年05月16日 18:45  1137

1、问题描述及原因分析 在编写Spark程序中，由于在map，foreachPartition等算子内部使用了外部定义的变量和函数，从而引发Task未序列化问题。然而，Spark算子在计算过程中使用...

(http://blog.csdn.net/qq_14950717/article/details/51427207)

ria + prism 难解问题之 "返回类型必须是实体或复杂类型、复杂类型的集合或预定义..."

最近在做ria+siverlight 的项目，遇到了很多无解的问题。实在头疼不已，思量再三还是改一下懒惰的习性，写篇文章记录一下。如果有哪位世外高人能路过此地，指点一二，就更感激不尽了。。闲话少说，上...

 sorus 2011年11月02日 11:56  927

(<http://blog.csdn.net/sorus/article/details/6927834>)