



博客 (<http://www.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>) ...

写博客

发Chat

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

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

Spark中的序列化机制

原创2015年07月16日 13:54:24

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

5524

LambdaCC (<http://blog.csdn.net/u011491148>)

+ 关注

码云

未开通

(<https://github.com/LambdaCC>)

utm_source=...

原创23

粉丝7

喜欢1

Spark中的序列化机制

GitHub (<https://github.com/TaoXiao>)

标签：Spark Kryo

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Executor端的代码运行在不同的JVM中，甚至在不同的节点上，因此必然要有相应的序列化机制来支撑数据实例在不同的JVM或者节点之间的传输。

什么时候需要调用序列化?

先看一个自定义的类



- 他的最新文章
- 更多文章 (<http://blog.csdn.net/u011491148>)
- [Kerberos 安装](http://blog.csdn.net/u011491148/article/details/48545785) (<http://blog.csdn.net/u011491148/article/details/48545785>)
- [YUM配置及自定义](http://blog.csdn.net/u011491148/article/details/47019431) (<http://blog.csdn.net/u011491148/article/details/47019431>)
- [使用Kryo](http://blog.csdn.net/u011491148/article/details/46913115) (<http://blog.csdn.net/u011491148/article/details/46913115>)
- [HBase使用常见异常](http://blog.csdn.net/u011491148/article/details/46848673) (<http://blog.csdn.net/u011491148/article/details/46848673>)
- [HBase中由Reverse DNS引起的问题](http://blog.csdn.net/u011491148/article/details/46779227) (<http://blog.csdn.net/u011491148/article/details/46779227>)

- 文章分类
- [数据挖掘&机器学习](http://blog.csdn.net/) ([http://bl...](http://blog.csdn.net/)) 1篇
- [Phoenix](http://blog.csdn.net/) ([http://blog.csdn.n...](http://blog.csdn.net/)) 3篇
- [Python](http://blog.csdn.net/) ([http://blog.csdn.net...](http://blog.csdn.net/)) 1篇
- [Java](http://blog.csdn.net/) ([http://blog.csdn.net/u...](http://blog.csdn.net/)) 17篇
- [Hadoop](http://blog.csdn.net/) ([http://blog.csdn.n...](http://blog.csdn.net/)) 2篇
- 展开

```
1 package cn.gridx.spark.examples.serialization;
2
3 import org.apache.hadoop.hbase.util.Bytes;
4
5 public class UnserializableJavaClass {
6     public String ms;
7     public byte[] bytes;
8     public int n;
9
10    public UnserializableJavaClass() {
11        ms = "Uninitialized String";
12        bytes = null;
13        n = 0;
14    }
15
16    public UnserializableJavaClass(String s, byte[] bytes, int n) {
17        ms = s;
18        this.bytes = bytes.clone();
19        this.n = n;
20    }
21
22    public String getMs() {
23        return this.ms ;
24    }
25
26    public byte[] getBytes() { return this.bytes; }
27
28    public int getInt() { return this.n ; }
29
30    public String toString() { return "ms=" + ms + "\nbytes="
31        + Bytes.toStringBinary(bytes) + "\nn=" + n + "\n"; }
32
33    public void setInt(int n) { this.n = n; }
34 }
```

文章存档

2015年9月 (http://blog.csdn....)	1篇
2015年7月 (http://blog.csdn....)	6篇
2015年6月 (http://blog.csdn....)	5篇
2015年5月 (http://blog.csdn....)	7篇
2015年3月 (http://blog.csdn....)	3篇

展开

他的热门文章

利用Phoenix为HBase创建二级索引 (<http://blog.csdn.net/u011491148/article/details/45749807>)

12737

Scala中json4s的使用例子 (<http://blog.csdn.net/u011491148/article/details/44731265>)

8644

使用Phoenix的JDBC接口 (<http://blog.csdn.net/u011491148/article/details/45689109>)

6144

Spark中的序列化机制 (<http://blog.csdn.net/u011491148/article/details/46910803>)

5504

HDFS中的压缩与解压缩机制 (<http://blog.csdn.net/u011491148/article/details/9966369>)

4648

下面给出几种Spark中对 UnserializableJavaClass 的用法，看看怎样使用会涉及到对该类的序列化。

Example 1 :

```
1 val conf = new SparkConf().setAppName("Test Serialization")
2 val sc = new SparkContext(conf)
3
4 val javaObj = new UnserializableJavaClass("I'm `UnserializableJavaClass`", Bytes.toBytes("Hello `UnserializableJavaClass`"), 1010)
5
6 val rdd = sc.parallelize(1 to 10, 4)
7
8 rdd.map(i => new UnserializableJavaClass("I'm `UnserializableJavaClass`", Bytes.toBytes("Hello `UnserializableJavaClass`"), i*100))
9
10    .map(x => { x.setInt(javaObj.n + 1); x })
11    .collect
12    .foreach(println)
13
14 sc.stop
```

运行结果：

HTTP/1.1 400 Bad Request

```
Exception in thread "main" org.apache.spark.SparkException: Task not serializable
    at org.apache.spark.util.ClosureCleaner$.ensureSerializable(ClosureCleaner.scala:166)
    at org.apache.spark.util.ClosureCleaner$.clean(ClosureCleaner.scala:158)
    at org.apache.spark.SparkContext.clean(SparkContext.scala:1440)
    at org.apache.spark.rdd.RDD.map(RDD.scala:271)
    at cn.gridx.spark.examples.serialization.TestSerialization$.main(TestSerialization.scala:22)
    at cn.gridx.spark.examples.serialization.TestSerialization.main(TestSerialization.scala)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:39)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:25)
    at java.lang.reflect.Method.invoke(Method.java:597)
    at org.apache.spark.deploy.SparkSubmit$.launch(SparkSubmit.scala:358)
    at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:75)
    at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.io.NotSerializableException: cn.gridx.spark.examples.serialization.UnserializableJavaClass
    at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1164)
    at java.io.ObjectOutputStream.defaultWriteFields(ObjectOutputStream.java:1518)
    at java.io.ObjectOutputStream.writeSerialData(ObjectOutputStream.java:1483)
    at java.io.ObjectOutputStream.writeOrdinaryObject(ObjectOutputStream.java:1400)
    at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1158)
    at java.io.ObjectOutputStream.writeObject(ObjectOutputStream.java:330)
    at org.apache.spark.serializer.JavaSerializationStream.writeObject(JavaSerializer.scala:42)
    at org.apache.spark.serializer.JavaSerializerInstance.serialize(JavaSerializer.scala:73)
    at org.apache.spark.util.ClosureCleaner$.ensureSerializable(ClosureCleaner.scala:164)
    ... 12 more
```



解释：在这行 `map(x => { x.setInt(javaObj.n + 1); x })` 中，闭包中引用了Driver端创建的实例 `javaObj`，因此需要将该实例序列化后通过网络传输至各个Executor。

如果把上例中稍加修改：

Example 2



内容举报



返回顶部

```
1  . . . . .
2
3  rdd.map(i => new UnserializableJavaClass("I'm `UnserializableJavaClass`", Bytes.toBytes("Hello `Unserializab
4  leJavaClass`"), i*100))
5  .map(x => { x.setInt(javaObj.n + 1); x }) // 这里改一下
6  .collect
   .foreach(println)
```

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

即将Example 1中的 `map(x => { x.setInt(javaObj.n + 1); x })` 换成 `map(x => { x.setInt(x.n + 1); x })`，那么，就可以正常地运行结束。

解释：在 `map(x => { x.setInt(x.n + 1); x })` 中，Executor没有引用到Driver的实例，因此 `javaObj` 不需要被从Driver传输到Executor，因而不需要将其序列化。

Example 3:

什么样的数据类型能够直接被Spark序列化

先看一个例子，在这个例子中，我们自定义了一个名为 `UnserializableClass` 的类，并将其用在了Spark中。

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/
- webmaster@csdn.net
- (mailto:webmaster@csdn.net)
- 400-660-0108

- 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
- (https://passpc

```
1 package cn.gridx.spark.examples.serialization;
2
3 public class UnserializableClass {
4     public String ms;
5
6     public UnserializableClass() {
7         ms = "Uninitialized String";
8     }
9
10    public UnserializableClass(String s) {
11        ms = s;
12    }
13
14    public String addPrefix(String s) {
15        return ms + ":" + s;
16    }
17
18    public String getMs() { return ms; }
19 }
```

下面在一个Spark程序中使用该类的实例

```
1 // Driver
2 def main(args: Array[String]) {
3     val conf = new SparkConf()
4     .setAppName("Test Spark Serialization")
5     val sc = new SparkContext(conf)
6
7     val rdd = sc.parallelize(1 to 10, 4)
8     val obj = new UnserializableClass("Hello")
9
10    rdd.map(i => obj.addPrefix(i.toString))
11        .collect
12        .foreach(println)
13
14    sc.stop
15 }
```

运行后，报异常：

```
Exception in thread "main" org.apache.spark.SparkException: Task not serializable
    at org.apache.spark.util.ClosureCleaner$.ensureSerializable(ClosureCleaner.scala:166)
    at org.apache.spark.util.ClosureCleaner$.clean(ClosureCleaner.scala:158)
    at org.apache.spark.SparkContext.clean(SparkContext.scala:1440)
    at org.apache.spark.rdd.RDD.map(RDD.scala:271)
    at cn.gridx.spark.examples.serialization.TestKryo$.main(TestKryo.scala:24)
    at cn.gridx.spark.examples.serialization.TestKryo.main(TestKryo.scala)
    at sun.reflect.NativeMethodAccessorImpl.invoke(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:39)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:25)
    at java.lang.reflect.Method.invoke(Method.java:597)
    at org.apache.spark.deploy.SparkSubmit$.launch(SparkSubmit.scala:358)
    at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:75)
    at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: java.io.NotSerializableException: cn.gridx.spark.examples.serialization.UnserializableClass
    at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1164)
    at java.io.ObjectOutputStream.defaultWriteFields(ObjectOutputStream.java:1518)
    at java.io.ObjectOutputStream.writeSerialData(ObjectOutputStream.java:1483)
    at java.io.ObjectOutputStream.writeOrdinaryObject(ObjectOutputStream.java:1400)
    at java.io.ObjectOutputStream.writeObject0(ObjectOutputStream.java:1158)
    at java.io.ObjectOutputStream.writeObject(ObjectOutputStream.java:330)
    at org.apache.spark.serializer.JavaSerializationStream.writeObject(JavaSerializer.scala:42)
    at org.apache.spark.serializer.JavaSerializerInstance.serialize(JavaSerializer.scala:73)
    at org.apache.spark.util.ClosureCleaner$.ensureSerializable(ClosureCleaner.scala:164)
    ... 12 more
```

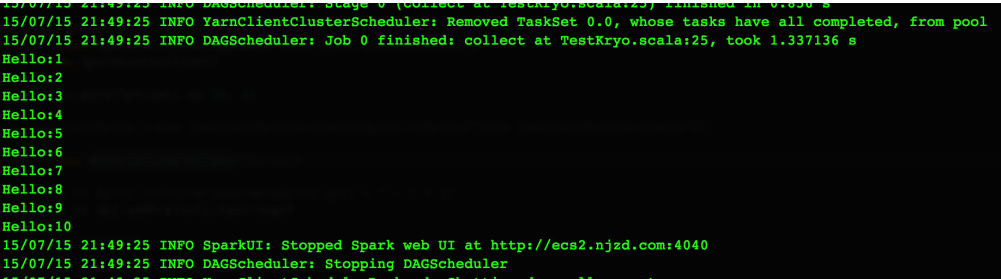
可见，对于自己定义的普通类，Spark是无法直接对其序列化的。

根据 Tuning Spark (<http://spark.apache.org/docs/1.2.1/tuning.html>)，Spark可以使用Java的序列化框架。『**只要一个class实现了 java.io.Serializable 接口，那么Spark就能使用Java的 ObjectOutputStream 来序列化该类**』。

验证：令 UnserializableClass 的声明实现接口 java.io.Serializable，但是类 UnserializableClass 的定义不做任何修改，即：

```
1 package cn.gridx.spark.examples.serialization;
2
3 public class UnserializableClass implements java.io.Serializable{
4     . . . . .
5     . . . . .
6 }
```

Spark的Driver端代码不做任何修改，此时重新运行Driver程序，可以正常运行结束。



除了Java序列化之外，还有什么其他选择？

上面说了，『**只要一个class实现了 java.io.Serializable 接口，那么Spark就能使用Java的 ObjectOutputStream 来序列化该类**』。


那么问题来了

- 1. 很多第三方的class没有实现 java.io.Serializable* 接口，我们也无法去改变这些第三方class的定义；
- 2. Java的序列化框架效率很低，很慢，性能很差。

实际上，Spark还支持另一种序列化框架 —— Kryo (<https://github.com/EsotericSoftware/kryo>)。Kryo是一个高效的序列化框架（可以比Java的序列化快10倍以上）。但是Kryo并不能支持所有的 Serilizable class，因此需要在使用Kryo前对目标类进行注册（register）。




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>)

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>)


看了年度账单，才知道我特么这么会省钱！！

在学习这条路上，我省钱的技巧足够出一本书.....



(http://www.baidu.com/cb.php?c=IgF_pyfqHmknjb3nHD0IZ0qnfK9ujYzP16sP1f30Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1Y3uHwbuHcznvNWPWDsnAfk0AwY5HDdnHRdPHbknWf0IgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfEmvqVmhq15LNYUNq1ULNzmvRqmhkEu1Ds0ZFb5Hns0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cwrHn:


Spark 序列化问题全解

 xwc35047 2017年11月01日 10:56 437

在Spark应用开发中，很容易出现如下报错：org.apache.spark.SparkException: Task not serializable at org.apache.spark.k.u...

(<http://blog.csdn.net/xwc35047/article/details/78411749>)

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)

[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>)

迈出成为抢手机器学习工程师第1步



【免费试听】迈出成为抢手机器学习工程师第1步

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 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/qz_14950717/article/details/51427207)



Task not serializable exception while running apache spark job

spark出现task不能序列化错误的解决方法 org.apache.spark.SparkException: Task not serializable 出现 “task not ser...

 zy_zhengyang 2015年09月29日 09:59  1305

(http://blog.csdn.net/zy_zhengyang/article/details/48803105)



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  u014487509 2015年11月23日 14:52  3390
1/23 14:43:47 ERROR Executor: Exception in task 0.0 in stage
4...

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



Spark Q&A : Task/Object not serializable 任务不能序列化

databricks的github io上针对Spark任务经常遇到的一些问题做了一些总结, 这里对关于任务和对象序列化这一章进行翻译.
原链接 Job aborted due to stage f...

 Edin_BlackPoint 2017年06月05日 16:32  1652

(http://blog.csdn.net/Edin_BlackPoint/article/details/72868621)



spark出现task不能序列化错误的解决方法

 javastart 2016年03月10日 13:18  2863

应用场景：使用JavaHiveContext执行SQL之后，希望能得到其字段名及相应的值，但却出现"Caused by: java.io.NotSerializableException: org.a...

(<http://blog.csdn.net/javastart/article/details/50845767>)



Spark闭包与序列化

 xiaolang85 2016年07月13日 13:47  718

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！Spark的...



(<http://blog.csdn.net/xiaolang85/article/details/51897345>)

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.KryoSeria  lxhandlbb 2016年10月31日 22:24  2081
lizer") 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 性能相关参数配置详解 - 压缩与序列化篇

 colorant 2014年08月19日 14:47  17832

随着Spark的逐渐成熟完善，越来越多的可配置参数被添加到Spark中来, 本文试图通过阐述这其中部分参数的工作原理和配置思路, 和大家一起探讨一下如何根据实际场合对Spark进行配置优化。- 压缩...

(<http://blog.csdn.net/colorant/article/details/38681581>)

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)

Spark中的序列化机制



u011491148

2015年07月16日 13:54

5524

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>)

javabean里序列化机制和构造函数的作用20170621

full constructor 和 minimal constructor default constructor是缺省构造



Ape55

2017年06月21日 11:38

240

造函数，用于平时的new XXX(); minimal constr...

(<http://blog.csdn.net/Ape55/article/details/73530502>)