# GeoSpark

## 第一章 简介

### 一、总体介绍

GeoSpark是在Apache Spark第三方项目中的一个子项目，也是一个用于处理大规模空间数据的集群计算系统，目前最新版本更新到v1.1.0（2018年3月13日）。GeoSpark继承自Apache Spark，并拥有一系列具有创造性的空间弹性分布式数据集（SRDDs），借助机器这些数据集可以有效的加载，处理并分析大规模的空间数据。GeoSpark为Apache开发人员提供了API使得他们能够利用SRDDs方便的开发空间分析程序，这些程序为地理空间查询提供了有力的支持。

### 二、功能介绍

2.1 GeoSparkSQL

最新版本的GeoSpark主要是包含了新的SQL功能，新增了四叉树和R树的索引解析并修复了一些bug。这个版本最主要的就是包含了一个完整的GeoSparkSQL版本，GeoSparkSQL完全支持Apache Spark SQL。

2.2 SRDDs

支持特殊的SRDDs，包括PointRDD,RectangleRDD,PolygonRDD以及LineStringRDD。

2.3 Data Format

支持的数据格式有CSV、TSV、WKT、GeoJSON、NASA Earth Data NetCDF/HDF、ESRI ShapeFile。

2.4 空间分割

支持的空间分割技术有四叉树、KDB树、R树、沃罗诺伊图（Voronoi diagram）、均匀网络（Uniform grids）

2.5 空间索引

支持四叉树、R树、以及空间K近邻查询

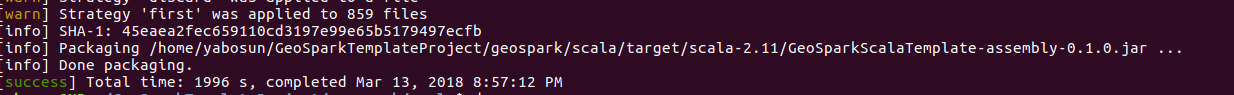
### 三、程序调试运行

**环境配置：**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 系统 | IDE | Spark | Scala | Maven | JDK | Sbt |
| Ubuntu16.04 | IntelliJ IDEA2017.3 | 2.2.1 | 2.11.12 | 3.5.3 | 1.8.0\_161 | 1.1.1 |

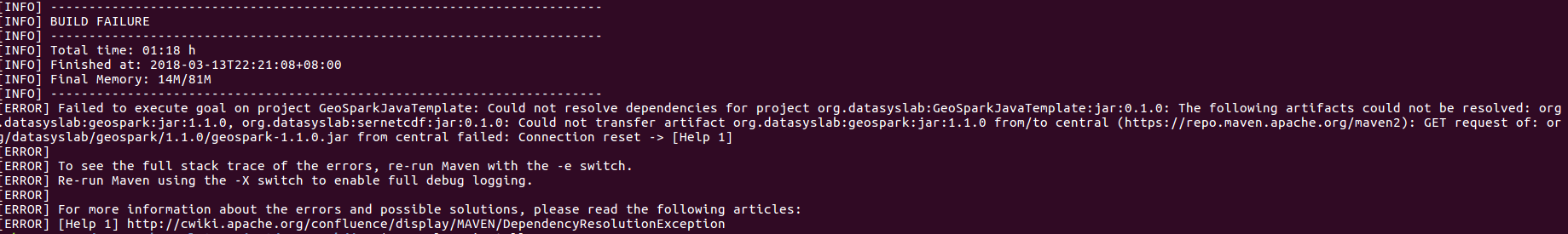
**编译GeoSpark Scala Template Project：**

成功编译

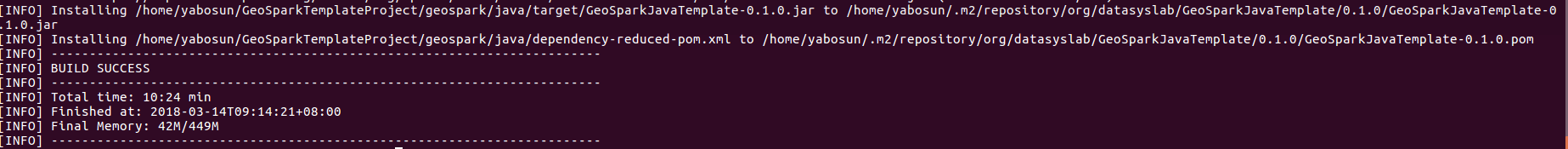


**编译GeoSpark Java Template Project：**

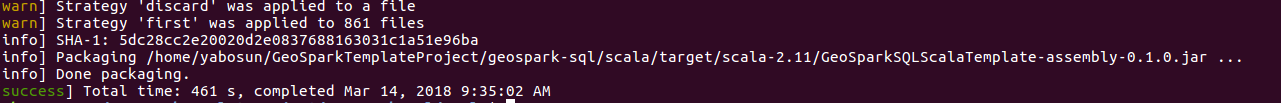
时间比较久，然后出现错误：



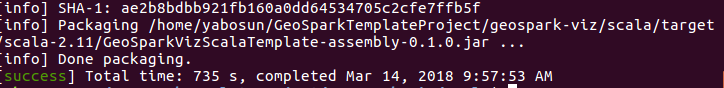
查看错误信息怀疑是网络问题，重新编译后显示成功



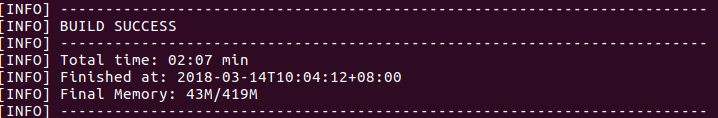
编译GeoSpark-SQL Scala Template Project成功



编译GeoSpark-Viz Scala Template Project成功



编译GeoSpark-Viz Java Template Project成功



[GeoSparkTemplateProject](https://github.com/jiayuasu/GeoSparkTemplateProject)

[GeoSparkGithub项目地址](https://github.com/DataSystemsLab/GeoSpark)

**调试过程中的坑：**

1. 安装Scala插件以及配置sdk

<http://dblab.xmu.edu.cn/blog/1327/>

如果遇到无法在idea中直接安装可以参考这篇文档：<https://blog.csdn.net/iamlihongwei/article/details/72783459?locationNum=10&fps=1>下载对应的安装包

1. 配置集群，查看devs文件
2. 安装依赖，直接去maven官网，搜索GeoSpark的依赖，拷贝到pom.xml中即可

需要简单了解的知识：

1. maven相关配置，简单了解一下
2. Scala语法
3. Linux有关命令
4. Spark入门与进阶

这个时候本地配置的都配置好了，而且运行简单的案例也成功了，但是运行GeoSpark还是出现问题IllegalAccessError

网上说是连接了Google的相关依赖无法连接，查询pom.xml文件没有对应的依赖，参考如下链接：

<https://blog.csdn.net/superzyl/article/details/53764731>

查阅官方api说明如下：

Thrown if an application attempts to access or modify a field, or to call a method that it does not have access to.

Normally, this error is caught by the compiler; this error can only occur at run time if the definition of a class has incompatibly changed.

后面查阅相关资料，怀疑是guava的版本冲突，但是具体对应的版本不清楚是哪个版本

Guava简介：

Guava工程包含了若干被Google的Java项目广泛依赖的核心库，集合、缓存、原生类型支持、并发库、通用注解、字符串处理、IO等

接下来删除相关的guava版本，只留了16.0.1，重新编译还是不行

最后把整个项目重新下载，这时候不知道为什么无法git clone，总是clone半截显示无法connect解决办法如下：

1. 删除原有项目，重启计算机
2. 查看git版本，没问题
3. 考虑到权限问题，把项目地址改了一个文件夹，然后进行sudo git clone，发现并不是这些的问题，而是早上网速抽风，坑

好了，现在clone好了，接下来使用idea导入项目进行编译

导入为maven项目过程中会显示项目是只读的无法修改

查询只能把项目中的所有文件递归给可读可写等权限

修改完成

导入Java项目，按照网上所说添加依赖

进行编译运行

仍然报错IllegalAccessError

在github上直接提问，大佬回复说是依赖冲突的问题，让我看

[http://datasystemslab.github.io/GeoSpark/download/project/#package-the-project](http://datasystemslab.github.io/GeoSpark/download/project/" \l "package-the-project)

试着先按照教程调试sparksql

引入sbt过程中一直在转，最后索性崩了

解决参考这篇：

<https://stackoverflow.com/questions/47697141/intellij-cannot-import-sbt-project> 未成功

看了一下dependency scope

[https://maven.apache.org/guides/introduction/introduction-to-dependency-mechanism.html#Dependency\_Scope](https://maven.apache.org/guides/introduction/introduction-to-dependency-mechanism.html" \l "Dependency_Scope)

就是说在本地IDE中运行不需要做任何改变，不需要打包

但是要是在集群模式下运行需要修改dependency scope并且打包

Scope几种模式：

1、test范围指的是测试范围有效，在编译和打包时都不会使用这个依赖

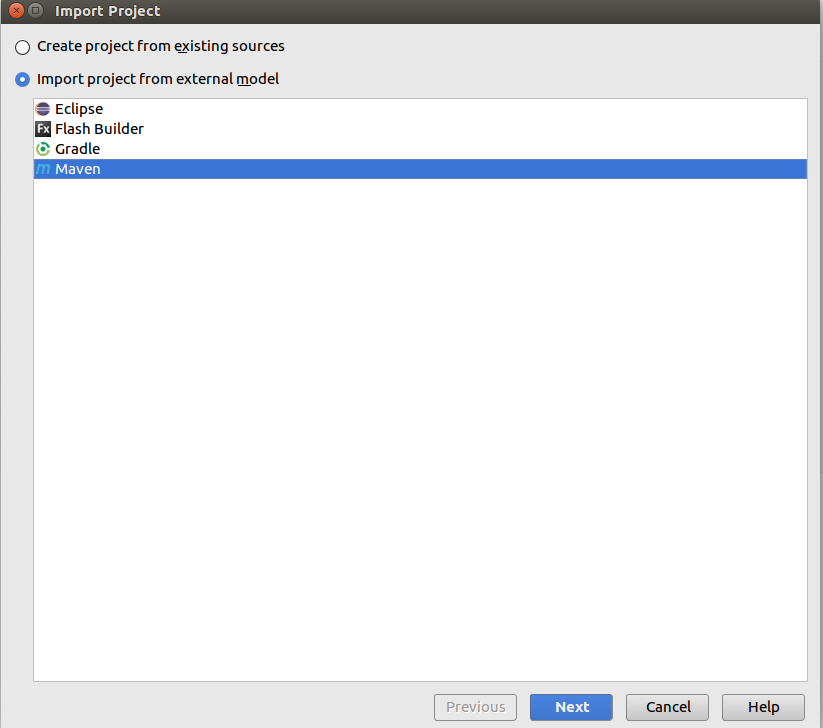
2、compile范围指的是编译范围有效，在编译和打包时都会将依赖存储进去

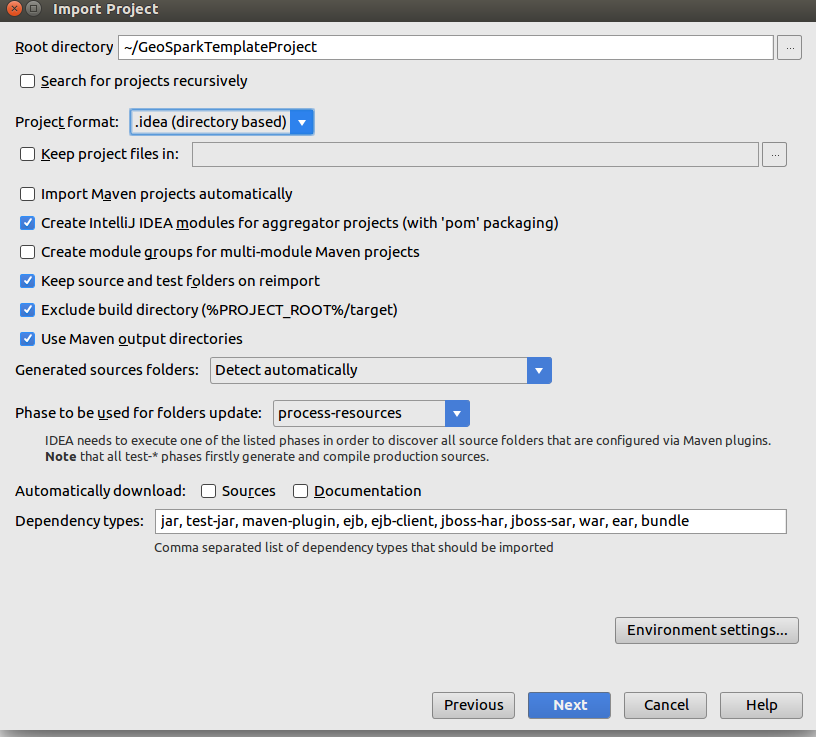
3、provided依赖：在编译和测试的过程有效，最后生成war包时不会加入，诸如：servlet-api，因为servlet-api，tomcat等web服务器已经存在了，如果再打包会冲突

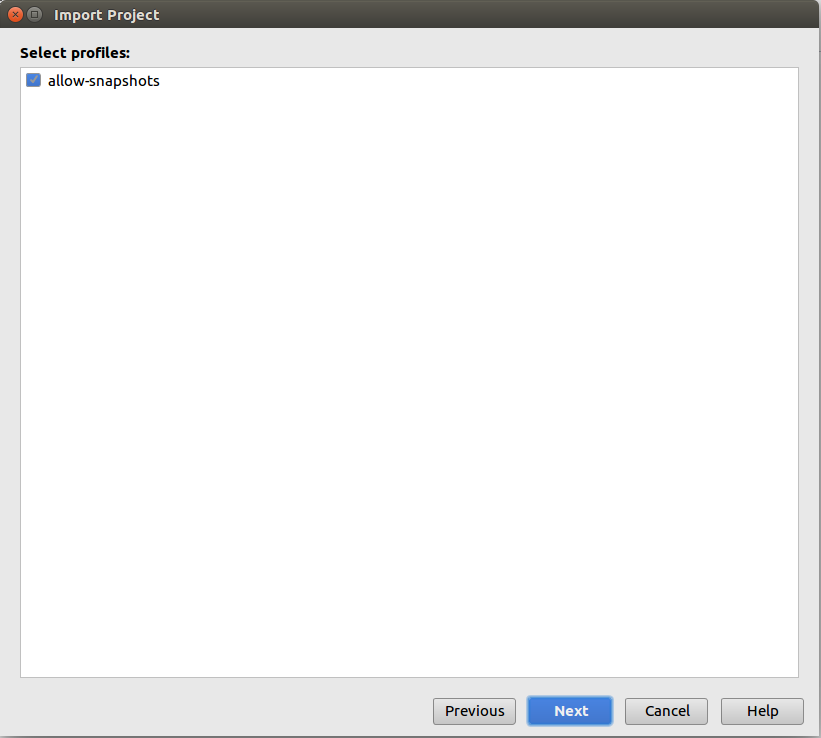
4、runtime在运行的时候依赖，在编译的时候不依赖

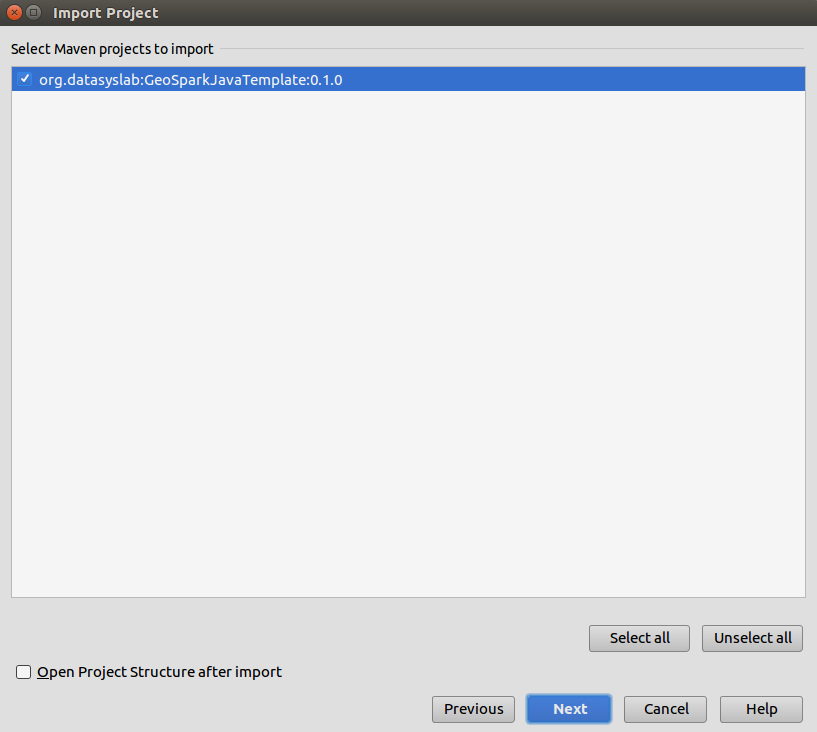
一般情况下默认是compile

重新卸载安装了idea，删除了所有个人配置开始







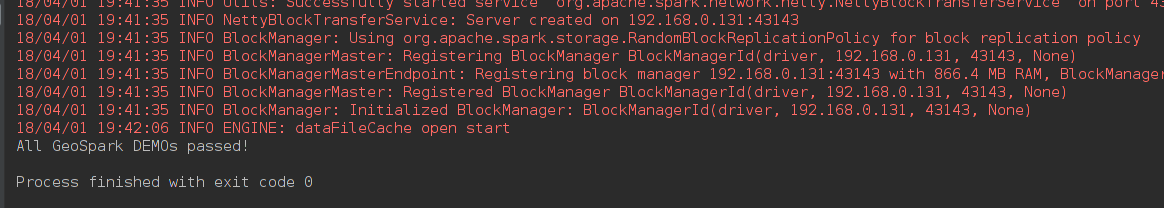


并无卵用。。。

后又请教大佬，回复如下：根本问题就是依赖冲突的问题，不同版本的IDEA对于依赖的处理不同，然后大佬又更新的项目的一些依赖，并给我说在自己机器上用IDEA2017.3没有问题

于是，卸载了我之前安装的IDEA版本，装了2017.3without jdk，重新clone代码，运行。。。

完美



**Sbt安装**

<https://www.imooc.com/wenda/detail/316526>

### 成功运行

# Spark知识

1. Unable to load native-hadoop library for your platform... using builtin-java classes where applicable 表示现在运行编译的是32位版本的Hadoop 而机器是64位，不过不影响运行