基于Spark集群在Windows 10下使用IntelliJ IDEA开发Spark应用程序

WINDOWS10软件安装

- 前提条件：

  1. 基于Linux的Spark集群已经搭建完成（考虑到稳定性，推荐使用CentOS）

     https://github.com/QiqiDuan257/parallel-pso-spark/blob/master/How-to-Install-Spark-on-CentOS7-Chinese.md

     Linux用于开发、测试，同时也是用于生产的标准（唯一）平台

     Mac OS X只用于开发、测试

     WINDOWS只用于开发、测试

  2. 采用WINDOWS10的个人电脑需要与Spark集群处在同一个局域网内

- 需要在WINDOWS10下安装的软件：

  1. java version 1.8.0_131

     （通过CMD命令“*java --version*”验证JAVA是否安装成功）

  2. Scala version 2.11.11

     （通过CMD命令“*scala --version*”验证Scala是否安装成功）

  3. IntelliJ IDEA（Community版本）

     https://www.jetbrains.com/idea/download/#section=windows

  特别注意：

  WINDOWS10下的Java、Scala版本需要与Spark集群下的版本保持一致

WINDOWS10系统环境变量配置

- 以管理员身份修改*hosts*配置文件，保存Spark集群中所有节点的IP地址与主机名之间的一一映射关系：

  *hosts*配置文件位置：*C:\Windows\System32\drivers\etc\hosts*

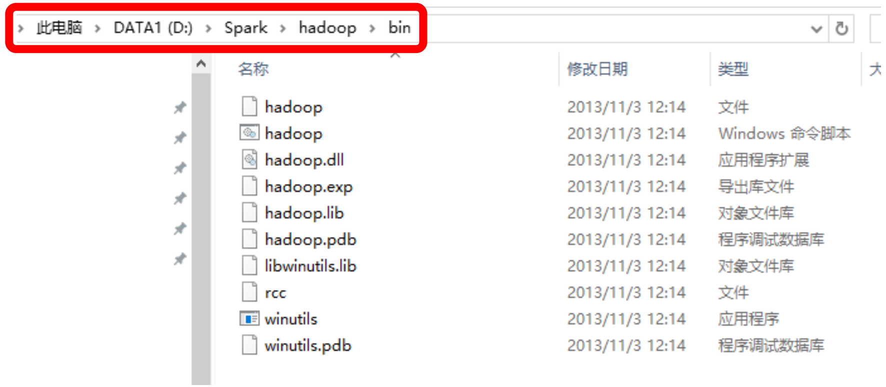
*hosts*配置文件示例：

```
# for the Spark commodity cluster
10.20.51.154    dc001.syhlab dc001
10.20.42.194    dc002.syhlab dc002
10.20.42.177    dc003.syhlab dc003
10.20.42.175    dc004.syhlab dc004
…               …
```

WINDOWS10系统环境变量配置

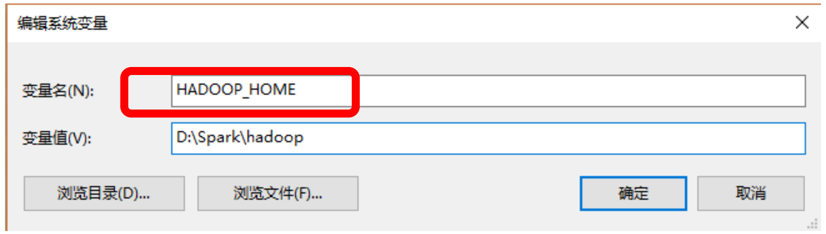
- 下载**Hadoop组件**到文件夹“**D:\Spark\hadoop**”中（此文件夹应只用于**Spark**开发）：
  https://github.com/srccodes/hadoop-common-2.2.0-bin

- 配置环境变量***HADOOP_HOME***与***PATH***。否则，会抛出以下异常：
  java.io.IOException: Could not locate executable null\bin\winutils.exe in the Hadoop binaries.



› 此电脑 › DATA1 (D:) › Spark › hadoop › bin

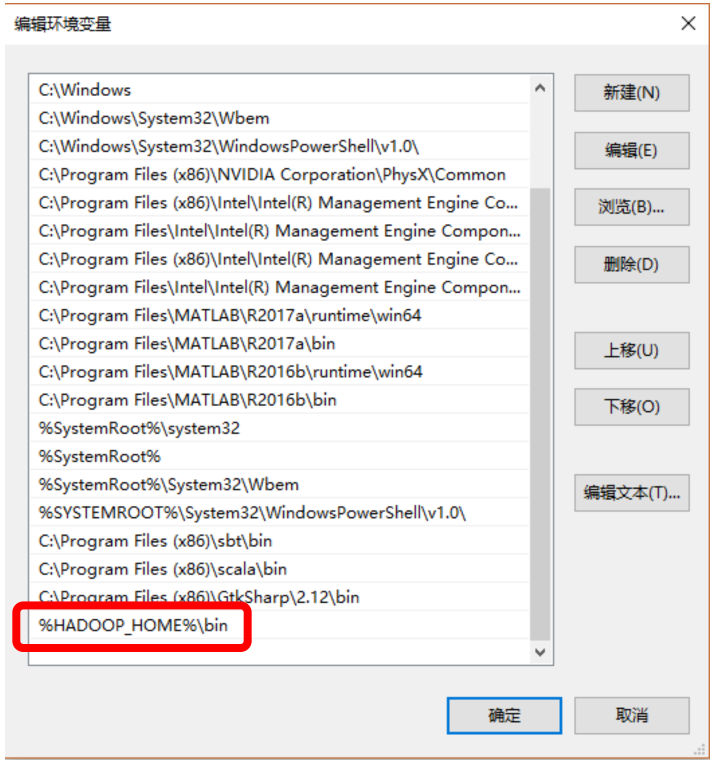
名称	修改日期	类型
hadoop	2013/11/3 12:14	文件
hadoop	2013/11/3 12:14	Windows 命令脚本
hadoop.dll	2013/11/3 12:14	应用程序扩展
hadoop.exp	2013/11/3 12:14	导出库文件
hadoop.lib	2013/11/3 12:14	对象文件库
hadoop.pdb	2013/11/3 12:14	程序调试数据库
libwinutils.lib	2013/11/3 12:14	对象文件库
rcc	2013/11/3 12:14	文件
winutils	2013/11/3 12:14	应用程序
winutils.pdb	2013/11/3 12:14	程序调试数据库



编辑系统变量

变量名(N): HADOOP_HOME

变量值(V): D:\Spark\hadoop

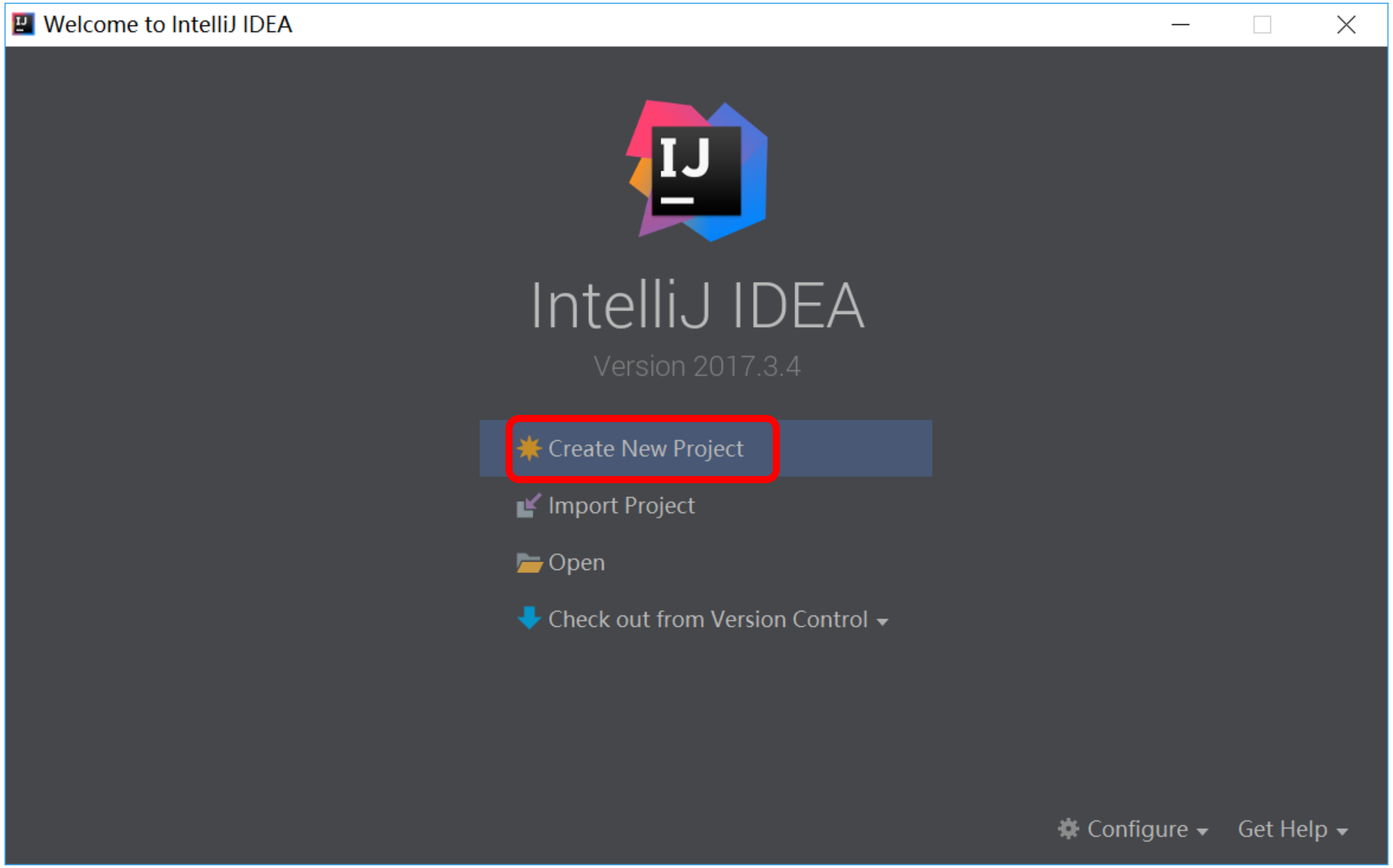


编辑环境变量

C:\Windows
C:\Windows\System32\Wbem
C:\Windows\System32\WindowsPowerShell\v1.0\
C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common
C:\Program Files (x86)\Intel\Intel(R) Management Engine Co...
C:\Program Files\Intel\Intel(R) Management Engine Compon...
C:\Program Files (x86)\Intel\Intel(R) Management Engine Co...
C:\Program Files\Intel\Intel(R) Management Engine Compon...
C:\Program Files\MATLAB\R2017a\runtime\win64
C:\Program Files\MATLAB\R2017a\bin
C:\Program Files\MATLAB\R2016b\runtime\win64
C:\Program Files\MATLAB\R2016b\bin
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
C:\Program Files (x86)\sbt\bin
C:\Program Files (x86)\scala\bin
C:\Program Files (x86)\GtkSharp\2.12\bin
%HADOOP_HOME%\bin

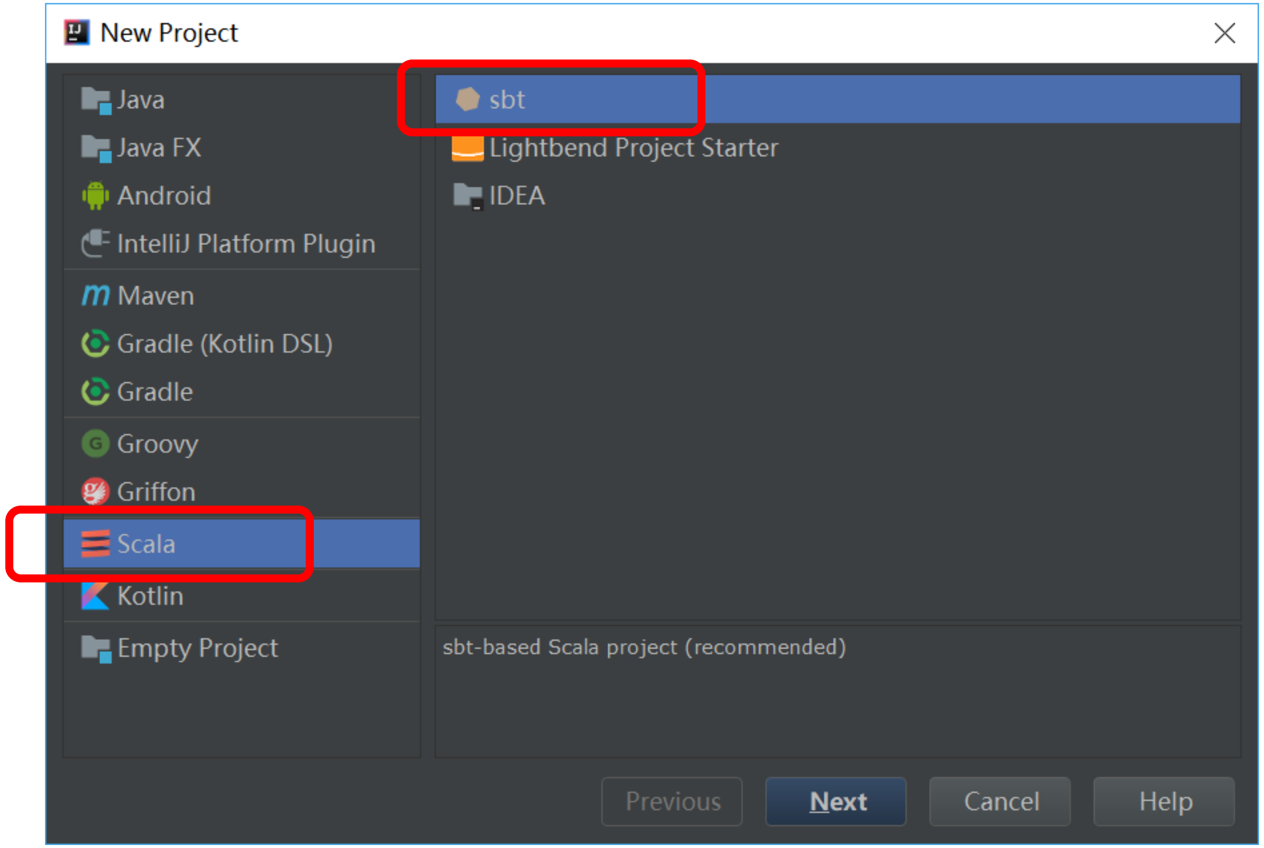
IntelliJ IDEA设置

- 创建基于Spark的应用程序：新建项目



IntelliJ IDEA设置

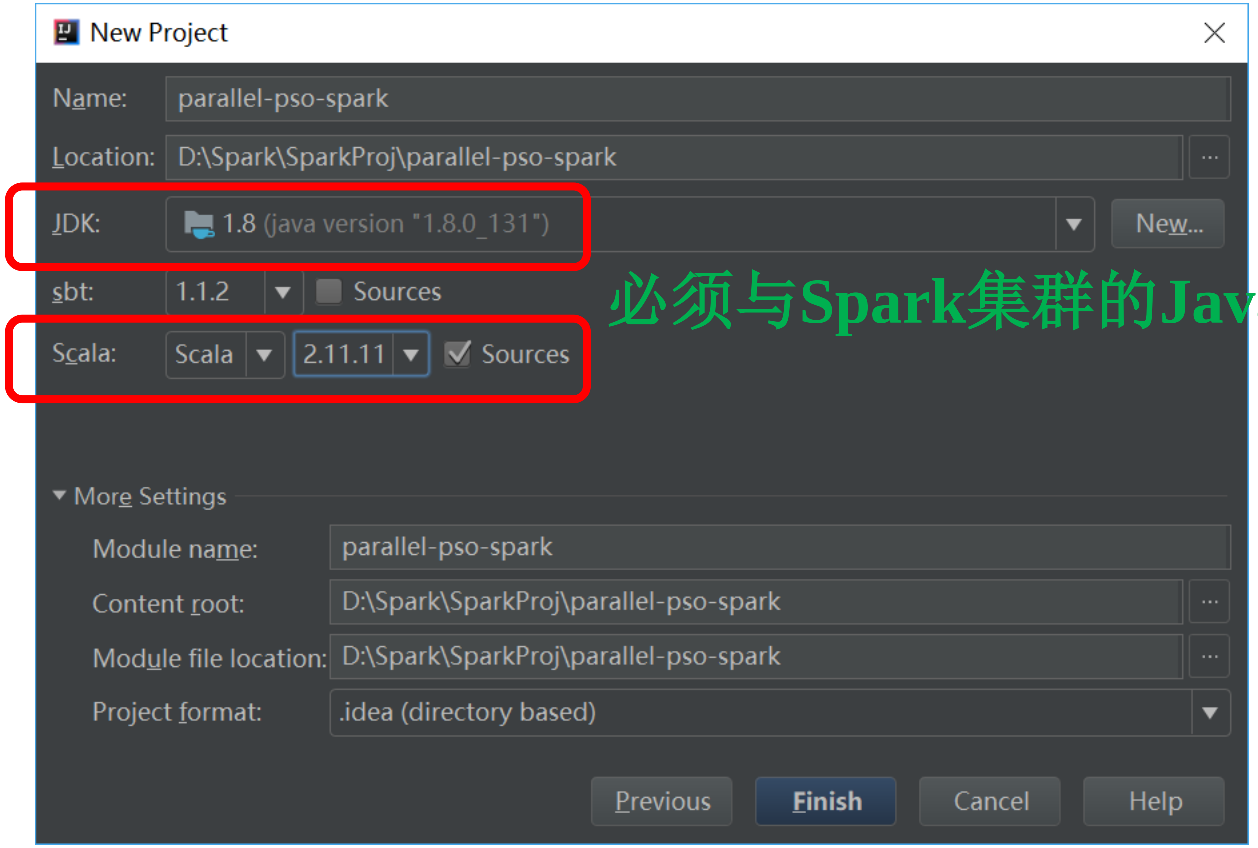
- 创建基于Spark的应用程序：创建基于Scala语言的sbt项目



sbt-based Scala project (recommended)

IntelliJ IDEA设置

- 创建基于Spark的应用程序：配置Java、sbt、Scala版本



New Project

Name: parallel-pso-spark

Location: D:\Spark\SparkProj\parallel-pso-spark

JDK: 1.8 (java version "1.8.0_131")

sbt: 1.1.2 □ Sources

Scala: Scala 2.11.11 ☑ Sources

More Settings

Module name: parallel-pso-spark

Content root: D:\Spark\SparkProj\parallel-pso-spark

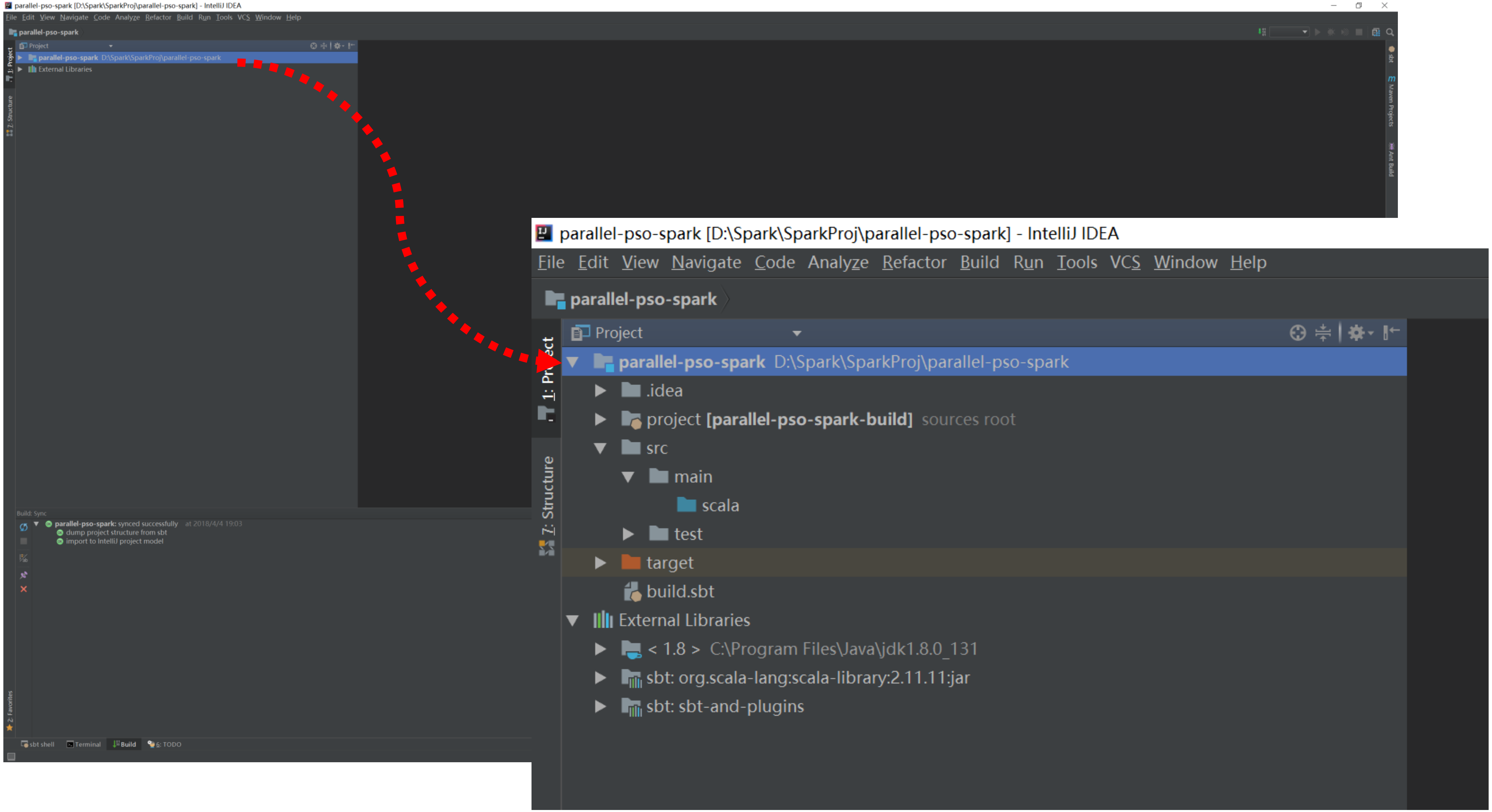
Module file location: D:\Spark\SparkProj\parallel-pso-spark

Project format: .idea (directory based)

必须与Spark集群的Java/Scala版本保持一致！

IntelliJ IDEA设置

- 创建基于Spark的应用程序：项目初始化界面



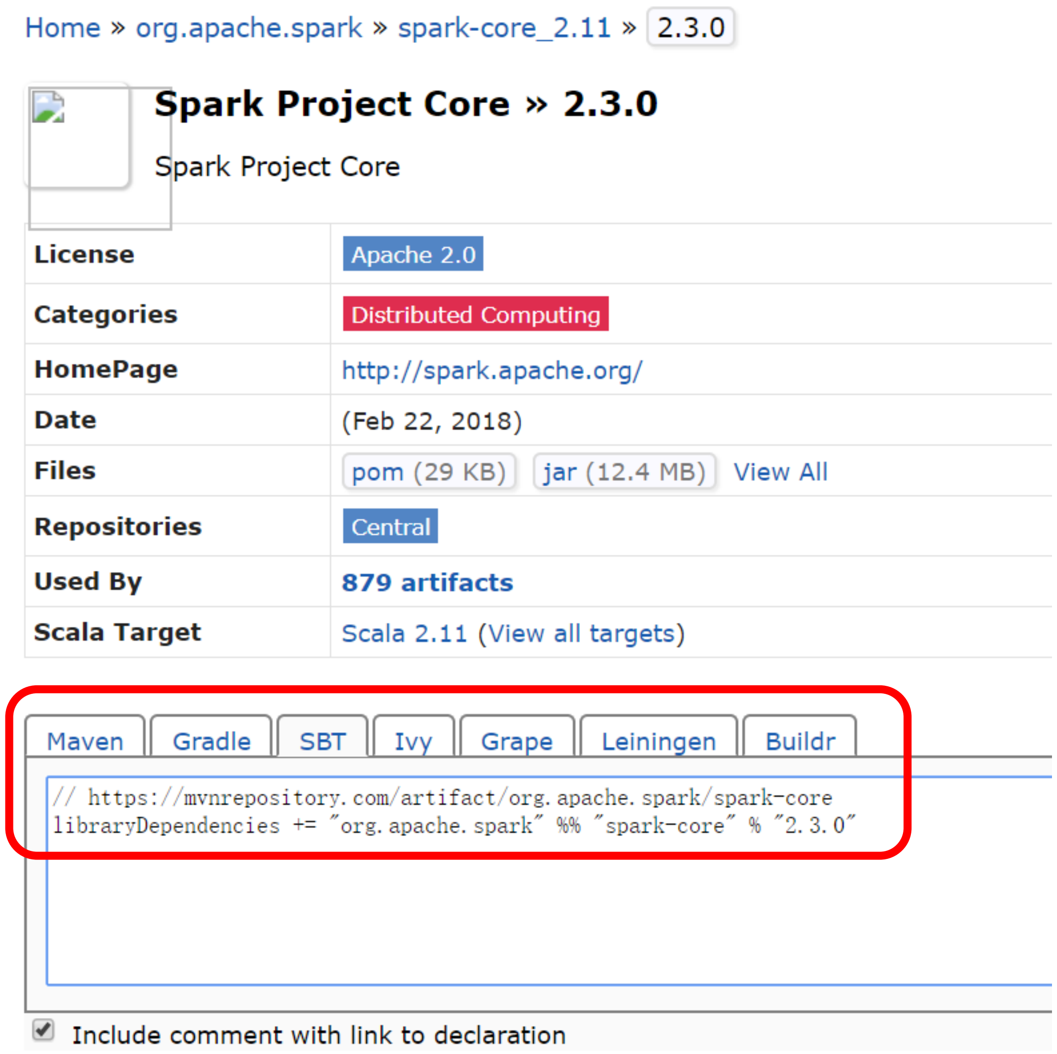
parallel-pso-spark [D:\Spark\SparkProj\parallel-pso-spark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
parallel-pso-spark
Project
1: Project
7: Structure
2: Favorites
sbt shell
Terminal
Build
6: TODO
Build: Sync

```
parallel-pso-spark: synced successfully   at 2018/4/4 19:03
    dump project structure from sbt
    import to IntelliJ project model
```

```
▼ parallel-pso-spark  D:\Spark\SparkProj\parallel-pso-spark
    ▶ .idea
    ▶ project [parallel-pso-spark-build]  sources root
    ▼ src
        ▼ main
              scala
        ▶ test
    ▶ target
      build.sbt
▼ External Libraries
    ▶ < 1.8 >  C:\Program Files\Java\jdk1.8.0_131
    ▶ sbt: org.scala-lang:scala-library:2.11.11:jar
    ▶ sbt: sbt-and-plugins
```

IntelliJ IDEA设置

- 从网站 https://mvnrepository.com/ 中获取Spark项目的sbt依赖包



Home » org.apache.spark » spark-core_2.11 » 2.3.0

Spark Project Core » 2.3.0

Spark Project Core

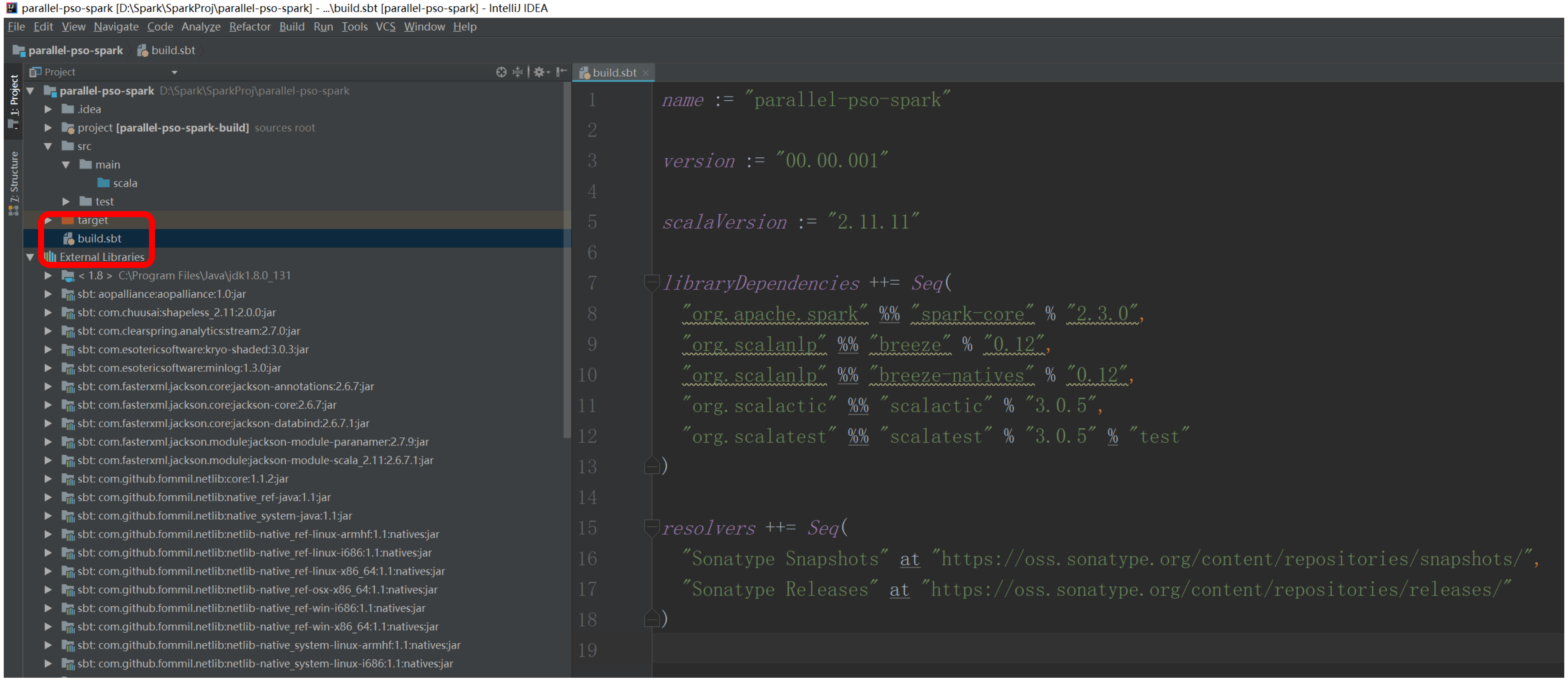
License	Apache 2.0
Categories	Distributed Computing
HomePage	http://spark.apache.org/
Date	(Feb 22, 2018)
Files	pom (29 KB) jar (12.4 MB) View All
Repositories	Central
Used By	879 artifacts
Scala Target	Scala 2.11 (View all targets)

```
// https://mvnrepository.com/artifact/org.apache.spark/spark-core
libraryDependencies += "org.apache.spark" %% "spark-core" % "2.3.0"
```

☑ Include comment with link to declaration

IntelliJ IDEA设置

- 创建基于Spark的应用程序：更新***build.sbt***文件，自动载入所有的依赖包
  可能等待较长时间（取决于网络速度）

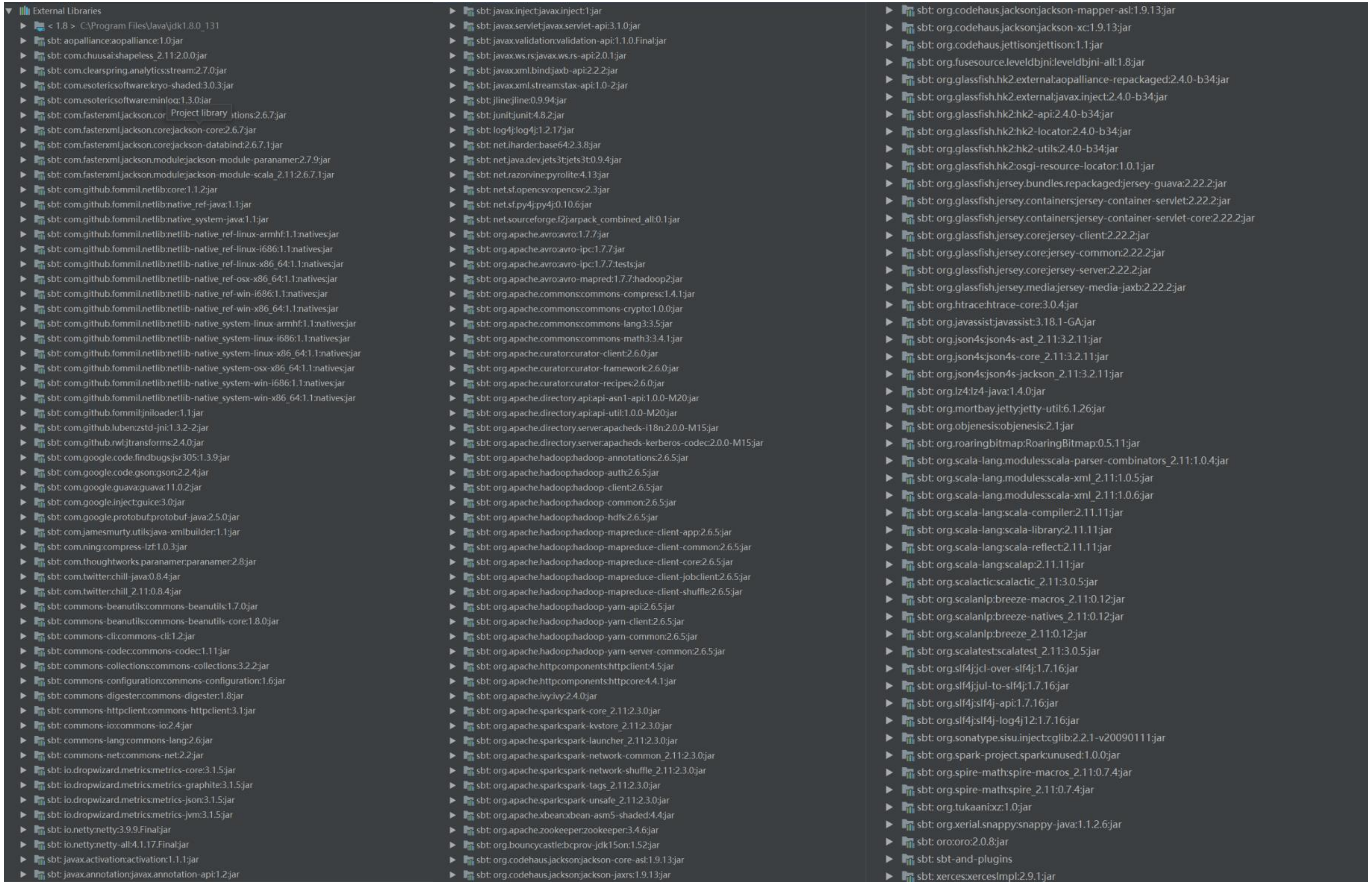


```
▼ parallel-pso-spark  D:\Spark\SparkProj\parallel-pso-spark
    ▶ .idea
    ▶ project [parallel-pso-spark-build]  sources root
    ▼ src
        ▼ main
              scala
        ▶ test
    ▶ target
      build.sbt
▼ External Libraries
    ▶ < 1.8 >  C:\Program Files\Java\jdk1.8.0_131
    ▶ sbt: aopalliance:aopalliance:1.0:jar
    ▶ sbt: com.chuusai:shapeless_2.11:2.0.0:jar
    ▶ sbt: com.clearspring.analytics:stream:2.7.0:jar
    ▶ sbt: com.esotericsoftware:kryo-shaded:3.0.3:jar
    ▶ sbt: com.esotericsoftware:minlog:1.3.0:jar
    ▶ sbt: com.fasterxml.jackson.core:jackson-annotations:2.6.7:jar
    ▶ sbt: com.fasterxml.jackson.core:jackson-core:2.6.7:jar
    ▶ sbt: com.fasterxml.jackson.core:jackson-databind:2.6.7.1:jar
    ▶ sbt: com.fasterxml.jackson.module:jackson-module-paranamer:2.7.9:jar
    ▶ sbt: com.fasterxml.jackson.module:jackson-module-scala_2.11:2.6.7.1:jar
    ▶ sbt: com.github.fommil.netlib:core:1.1.2:jar
    ▶ sbt: com.github.fommil.netlib:native_ref-java:1.1:jar
    ▶ sbt: com.github.fommil.netlib:native_system-java:1.1:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_ref-linux-armhf:1.1:natives:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_ref-linux-i686:1.1:natives:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_ref-linux-x86_64:1.1:natives:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_ref-osx-x86_64:1.1:natives:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_ref-win-i686:1.1:natives:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_ref-win-x86_64:1.1:natives:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_system-linux-armhf:1.1:natives:jar
    ▶ sbt: com.github.fommil.netlib:netlib-native_system-linux-i686:1.1:natives:jar
```

```scala
name := "parallel-pso-spark"

version := "00.00.001"

scalaVersion := "2.11.11"

libraryDependencies ++= Seq(
  "org.apache.spark" %% "spark-core" % "2.3.0",
  "org.scalanlp" %% "breeze" % "0.12",
  "org.scalanlp" %% "breeze-natives" % "0.12",
  "org.scalactic" %% "scalactic" % "3.0.5",
  "org.scalatest" %% "scalatest" % "3.0.5" % "test"
)

resolvers ++= Seq(
  "Sonatype Snapshots" at "https://oss.sonatype.org/content/repositories/snapshots/",
  "Sonatype Releases" at "https://oss.sonatype.org/content/repositories/releases/"
)

```

IntelliJ IDEA设置

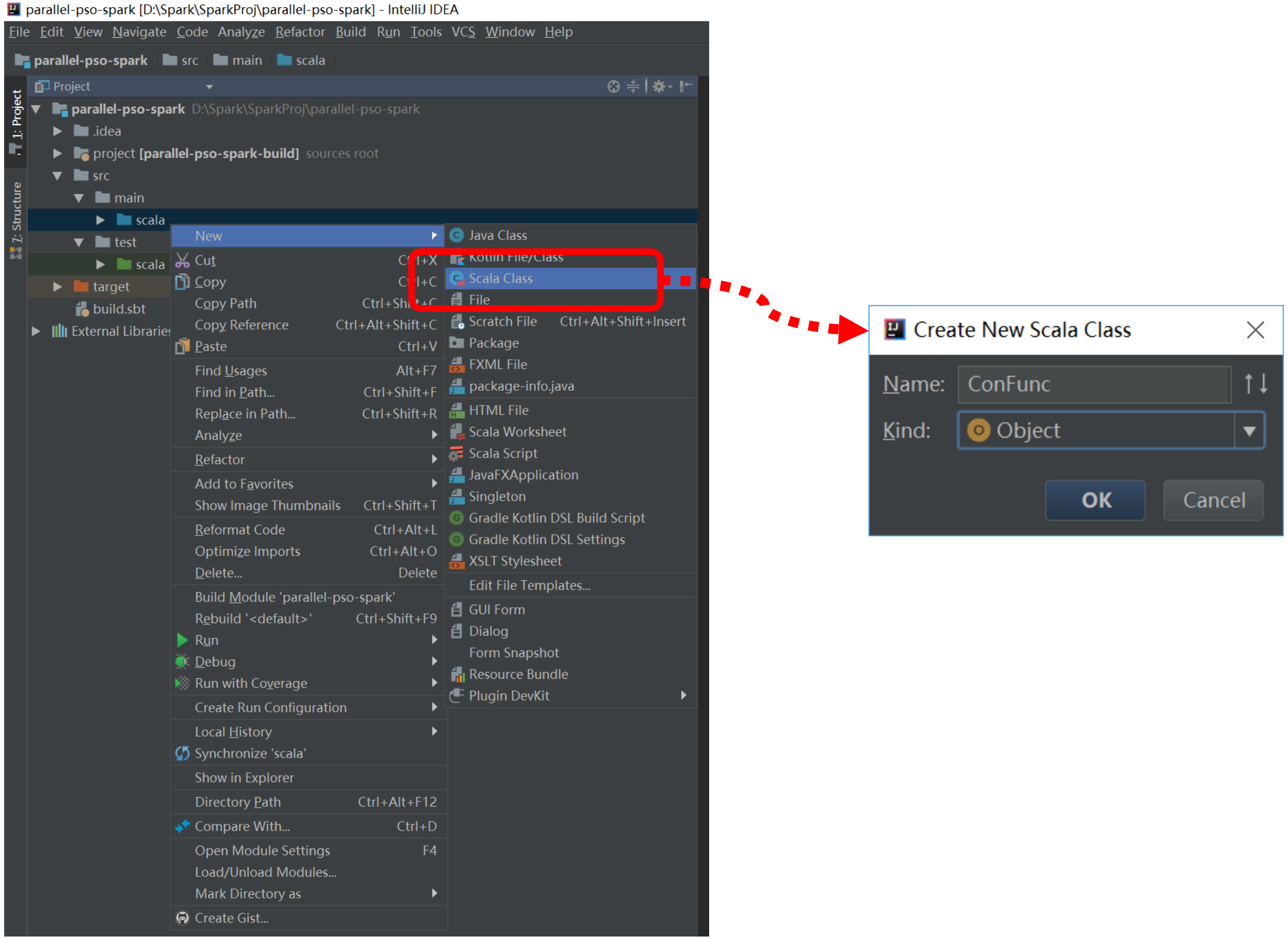
- 创建基于Spark的应用程序：载入后的依赖包



```
▼ External Libraries
  ▶ < 1.8 >  C:\Program Files\Java\jdk1.8.0_131
  ▶ sbt: aopalliance:aopalliance:1.0:jar
  ▶ sbt: com.chuusai:shapeless_2.11:2.0.0:jar
  ▶ sbt: com.clearspring.analytics:stream:2.7.0:jar
  ▶ sbt: com.esotericsoftware:kryo-shaded:3.0.3:jar
  ▶ sbt: com.esotericsoftware:minlog:1.3.0:jar
  ▶ sbt: com.fasterxml.jackson.cor Project library itions:2.6.7:jar
  ▶ sbt: com.fasterxml.jackson.core:jackson-core:2.6.7:jar
  ▶ sbt: com.fasterxml.jackson.core:jackson-databind:2.6.7.1:jar
  ▶ sbt: com.fasterxml.jackson.module:jackson-module-paranamer:2.7.9:jar
  ▶ sbt: com.fasterxml.jackson.module:jackson-module-scala_2.11:2.6.7.1:jar
  ▶ sbt: com.github.fommil.netlib:core:1.1.2:jar
  ▶ sbt: com.github.fommil.netlib:native_ref-java:1.1:jar
  ▶ sbt: com.github.fommil.netlib:native_system-java:1.1:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_ref-linux-armhf:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_ref-linux-i686:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_ref-linux-x86_64:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_ref-osx-x86_64:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_ref-win-i686:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_ref-win-x86_64:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_system-linux-armhf:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_system-linux-i686:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_system-linux-x86_64:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_system-osx-x86_64:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_system-win-i686:1.1:natives:jar
  ▶ sbt: com.github.fommil.netlib:netlib-native_system-win-x86_64:1.1:natives:jar
  ▶ sbt: com.github.fommil:jniloader:1.1:jar
  ▶ sbt: com.github.luben:zstd-jni:1.3.2-2:jar
  ▶ sbt: com.github.rwl:jtransforms:2.4.0:jar
  ▶ sbt: com.google.code.findbugs:jsr305:1.3.9:jar
  ▶ sbt: com.google.code.gson:gson:2.2.4:jar
  ▶ sbt: com.google.guava:guava:11.0.2:jar
  ▶ sbt: com.google.inject:guice:3.0:jar
  ▶ sbt: com.google.protobuf:protobuf-java:2.5.0:jar
  ▶ sbt: com.jamesmurty.utils:java-xmlbuilder:1.1:jar
  ▶ sbt: com.ning:compress-lzf:1.0.3:jar
  ▶ sbt: com.thoughtworks.paranamer:paranamer:2.8:jar
  ▶ sbt: com.twitter:chill-java:0.8.4:jar
  ▶ sbt: com.twitter:chill_2.11:0.8.4:jar
  ▶ sbt: commons-beanutils:commons-beanutils:1.7.0:jar
  ▶ sbt: commons-beanutils:commons-beanutils-core:1.8.0:jar
  ▶ sbt: commons-cli:commons-cli:1.2:jar
  ▶ sbt: commons-codec:commons-codec:1.11:jar
  ▶ sbt: commons-collections:commons-collections:3.2.2:jar
  ▶ sbt: commons-configuration:commons-configuration:1.6:jar
  ▶ sbt: commons-digester:commons-digester:1.8:jar
  ▶ sbt: commons-httpclient:commons-httpclient:3.1:jar
  ▶ sbt: commons-io:commons-io:2.4:jar
  ▶ sbt: commons-lang:commons-lang:2.6:jar
  ▶ sbt: commons-net:commons-net:2.2:jar
  ▶ sbt: io.dropwizard.metrics:metrics-core:3.1.5:jar
  ▶ sbt: io.dropwizard.metrics:metrics-graphite:3.1.5:jar
  ▶ sbt: io.dropwizard.metrics:metrics-json:3.1.5:jar
  ▶ sbt: io.dropwizard.metrics:metrics-jvm:3.1.5:jar
  ▶ sbt: io.netty:netty:3.9.9.Final:jar
  ▶ sbt: io.netty:netty-all:4.1.17.Final:jar
  ▶ sbt: javax.activation:activation:1.1.1:jar
  ▶ sbt: javax.annotation:javax.annotation-api:1.2:jar
  ▶ sbt: javax.inject:javax.inject:1:jar
  ▶ sbt: javax.servlet:javax.servlet-api:3.1.0:jar
  ▶ sbt: javax.validation:validation-api:1.1.0.Final:jar
  ▶ sbt: javax.ws.rs:javax.ws.rs-api:2.0.1:jar
  ▶ sbt: javax.xml.bind:jaxb-api:2.2.2:jar
  ▶ sbt: javax.xml.stream:stax-api:1.0-2:jar
  ▶ sbt: jline:jline:0.9.94:jar
  ▶ sbt: junit:junit:4.8.2:jar
  ▶ sbt: log4j:log4j:1.2.17:jar
  ▶ sbt: net.iharder:base64:2.3.8:jar
  ▶ sbt: net.java.dev.jets3t:jets3t:0.9.4:jar
  ▶ sbt: net.razorvine:pyrolite:4.13:jar
  ▶ sbt: net.sf.opencsv:opencsv:2.3:jar
  ▶ sbt: net.sf.py4j:py4j:0.10.6:jar
  ▶ sbt: net.sourceforge.f2j:arpack_combined_all:0.1:jar
  ▶ sbt: org.apache.avro:avro:1.7.7:jar
  ▶ sbt: org.apache.avro:avro-ipc:1.7.7:jar
  ▶ sbt: org.apache.avro:avro-ipc:1.7.7:tests:jar
  ▶ sbt: org.apache.avro:avro-mapred:1.7.7:hadoop2:jar
  ▶ sbt: org.apache.commons:commons-compress:1.4.1:jar
  ▶ sbt: org.apache.commons:commons-crypto:1.0.0:jar
  ▶ sbt: org.apache.commons:commons-lang3:3.5:jar
  ▶ sbt: org.apache.commons:commons-math3:3.4.1:jar
  ▶ sbt: org.apache.curator:curator-client:2.6.0:jar
  ▶ sbt: org.apache.curator:curator-framework:2.6.0:jar
  ▶ sbt: org.apache.curator:curator-recipes:2.6.0:jar
  ▶ sbt: org.apache.directory.api:api-asn1-api:1.0.0-M20:jar
  ▶ sbt: org.apache.directory.api:api-util:1.0.0-M20:jar
  ▶ sbt: org.apache.directory.server:apacheds-i18n:2.0.0-M15:jar
  ▶ sbt: org.apache.directory.server:apacheds-kerberos-codec:2.0.0-M15:jar
  ▶ sbt: org.apache.hadoop:hadoop-annotations:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-auth:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-client:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-common:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-hdfs:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-mapreduce-client-app:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-mapreduce-client-common:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-mapreduce-client-core:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-mapreduce-client-jobclient:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-mapreduce-client-shuffle:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-yarn-api:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-yarn-client:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-yarn-common:2.6.5:jar
  ▶ sbt: org.apache.hadoop:hadoop-yarn-server-common:2.6.5:jar
  ▶ sbt: org.apache.httpcomponents:httpclient:4.5:jar
  ▶ sbt: org.apache.httpcomponents:httpcore:4.4.1:jar
  ▶ sbt: org.apache.ivy:ivy:2.4.0:jar
  ▶ sbt: org.apache.spark:spark-core_2.11:2.3.0:jar
  ▶ sbt: org.apache.spark:spark-kvstore_2.11:2.3.0:jar
  ▶ sbt: org.apache.spark:spark-launcher_2.11:2.3.0:jar
  ▶ sbt: org.apache.spark:spark-network-common_2.11:2.3.0:jar
  ▶ sbt: org.apache.spark:spark-network-shuffle_2.11:2.3.0:jar
  ▶ sbt: org.apache.spark:spark-tags_2.11:2.3.0:jar
  ▶ sbt: org.apache.spark:spark-unsafe_2.11:2.3.0:jar
  ▶ sbt: org.apache.xbean:xbean-asm5-shaded:4.4:jar
  ▶ sbt: org.apache.zookeeper:zookeeper:3.4.6:jar
  ▶ sbt: org.bouncycastle:bcprov-jdk15on:1.52:jar
  ▶ sbt: org.codehaus.jackson:jackson-core-asl:1.9.13:jar
  ▶ sbt: org.codehaus.jackson:jackson-jaxrs:1.9.13:jar
  ▶ sbt: org.codehaus.jackson:jackson-mapper-asl:1.9.13:jar
  ▶ sbt: org.codehaus.jackson:jackson-xc:1.9.13:jar
  ▶ sbt: org.codehaus.jettison:jettison:1.1:jar
  ▶ sbt: org.fusesource.leveldbjni:leveldbjni-all:1.8:jar
  ▶ sbt: org.glassfish.hk2.external:aopalliance-repackaged:2.4.0-b34:jar
  ▶ sbt: org.glassfish.hk2.external:javax.inject:2.4.0-b34:jar
  ▶ sbt: org.glassfish.hk2:hk2-api:2.4.0-b34:jar
  ▶ sbt: org.glassfish.hk2:hk2-locator:2.4.0-b34:jar
  ▶ sbt: org.glassfish.hk2:hk2-utils:2.4.0-b34:jar
  ▶ sbt: org.glassfish.hk2:osgi-resource-locator:1.0.1:jar
  ▶ sbt: org.glassfish.jersey.bundles.repackaged:jersey-guava:2.22.2:jar
  ▶ sbt: org.glassfish.jersey.containers:jersey-container-servlet:2.22.2:jar
  ▶ sbt: org.glassfish.jersey.containers:jersey-container-servlet-core:2.22.2:jar
  ▶ sbt: org.glassfish.jersey.core:jersey-client:2.22.2:jar
  ▶ sbt: org.glassfish.jersey.core:jersey-common:2.22.2:jar
  ▶ sbt: org.glassfish.jersey.core:jersey-server:2.22.2:jar
  ▶ sbt: org.glassfish.jersey.media:jersey-media-jaxb:2.22.2:jar
  ▶ sbt: org.htrace:htrace-core:3.0.4:jar
  ▶ sbt: org.javassist:javassist:3.18.1-GA:jar
  ▶ sbt: org.json4s:json4s-ast_2.11:3.2.11:jar
  ▶ sbt: org.json4s:json4s-core_2.11:3.2.11:jar
  ▶ sbt: org.json4s:json4s-jackson_2.11:3.2.11:jar
  ▶ sbt: org.lz4:lz4-java:1.4.0:jar
  ▶ sbt: org.mortbay.jetty:jetty-util:6.1.26:jar
  ▶ sbt: org.objenesis:objenesis:2.1:jar
  ▶ sbt: org.roaringbitmap:RoaringBitmap:0.5.11:jar
  ▶ sbt: org.scala-lang.modules:scala-parser-combinators_2.11:1.0.4:jar
  ▶ sbt: org.scala-lang.modules:scala-xml_2.11:1.0.5:jar
  ▶ sbt: org.scala-lang.modules:scala-xml_2.11:1.0.6:jar
  ▶ sbt: org.scala-lang:scala-compiler:2.11.11:jar
  ▶ sbt: org.scala-lang:scala-library:2.11.11:jar
  ▶ sbt: org.scala-lang:scala-reflect:2.11.11:jar
  ▶ sbt: org.scala-lang:scalap:2.11.11:jar
  ▶ sbt: org.scalactic:scalactic_2.11:3.0.5:jar
  ▶ sbt: org.scalanlp:breeze-macros_2.11:0.12:jar
  ▶ sbt: org.scalanlp:breeze-natives_2.11:0.12:jar
  ▶ sbt: org.scalanlp:breeze_2.11:0.12:jar
  ▶ sbt: org.scalatest:scalatest_2.11:3.0.5:jar
  ▶ sbt: org.slf4j:jcl-over-slf4j:1.7.16:jar
  ▶ sbt: org.slf4j:jul-to-slf4j:1.7.16:jar
  ▶ sbt: org.slf4j:slf4j-api:1.7.16:jar
  ▶ sbt: org.slf4j:slf4j-log4j12:1.7.16:jar
  ▶ sbt: org.sonatype.sisu.inject:cglib:2.2.1-v20090111:jar
  ▶ sbt: org.spark-project.spark:unused:1.0.0:jar
  ▶ sbt: org.spire-math:spire-macros_2.11:0.7.4:jar
  ▶ sbt: org.spire-math:spire_2.11:0.7.4:jar
  ▶ sbt: org.tukaani:xz:1.0:jar
  ▶ sbt: org.xerial.snappy:snappy-java:1.1.2.6:jar
  ▶ sbt: oro:oro:2.0.8:jar
  ▶ sbt: sbt-and-plugins
  ▶ sbt: xerces:xercesImpl:2.9.1:jar
```

IntelliJ IDEA设置

- 创建基于Spark的应用程序：新建object

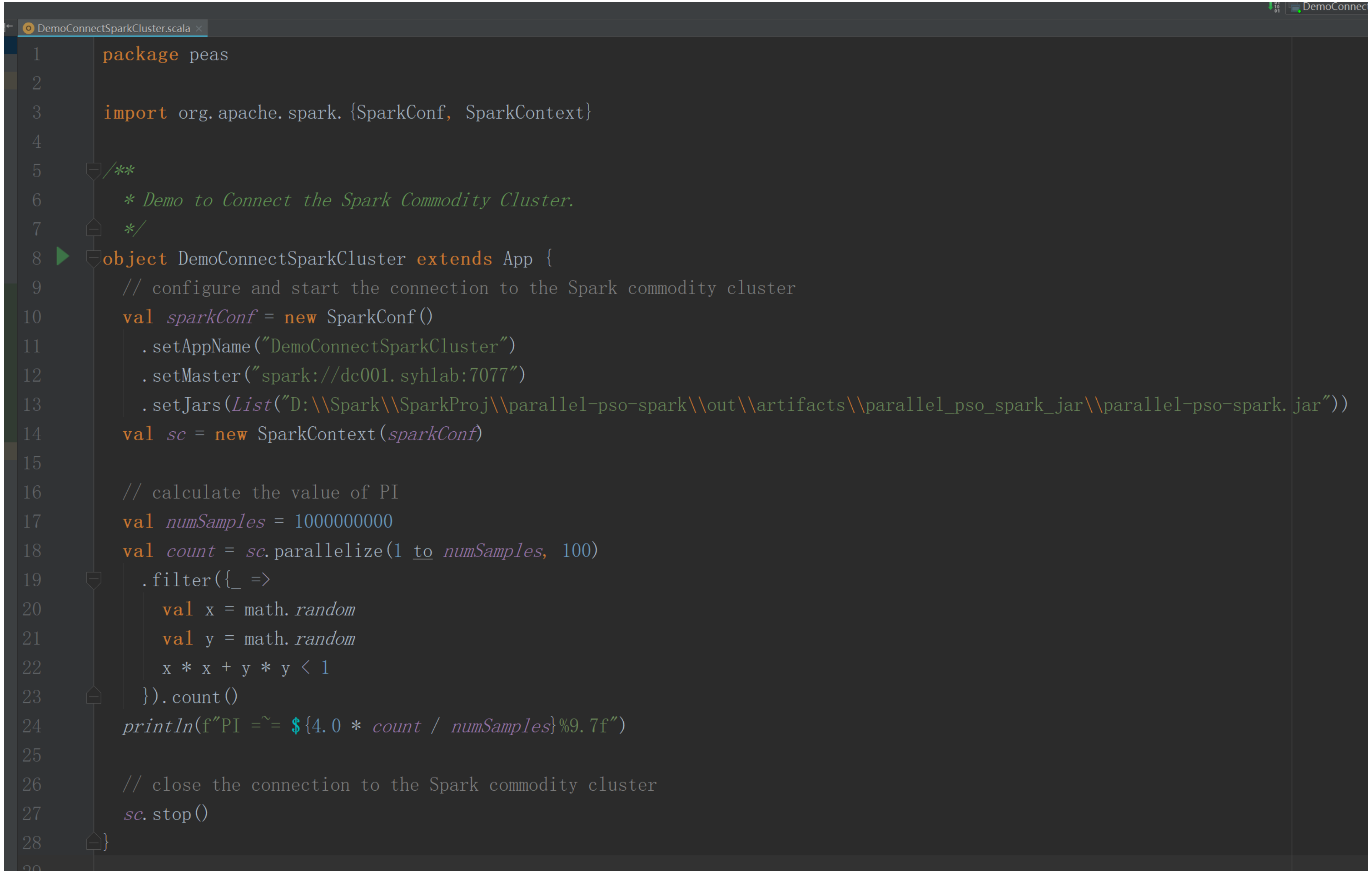


Name: ConFunc

Kind: Object

IntelliJ IDEA设置

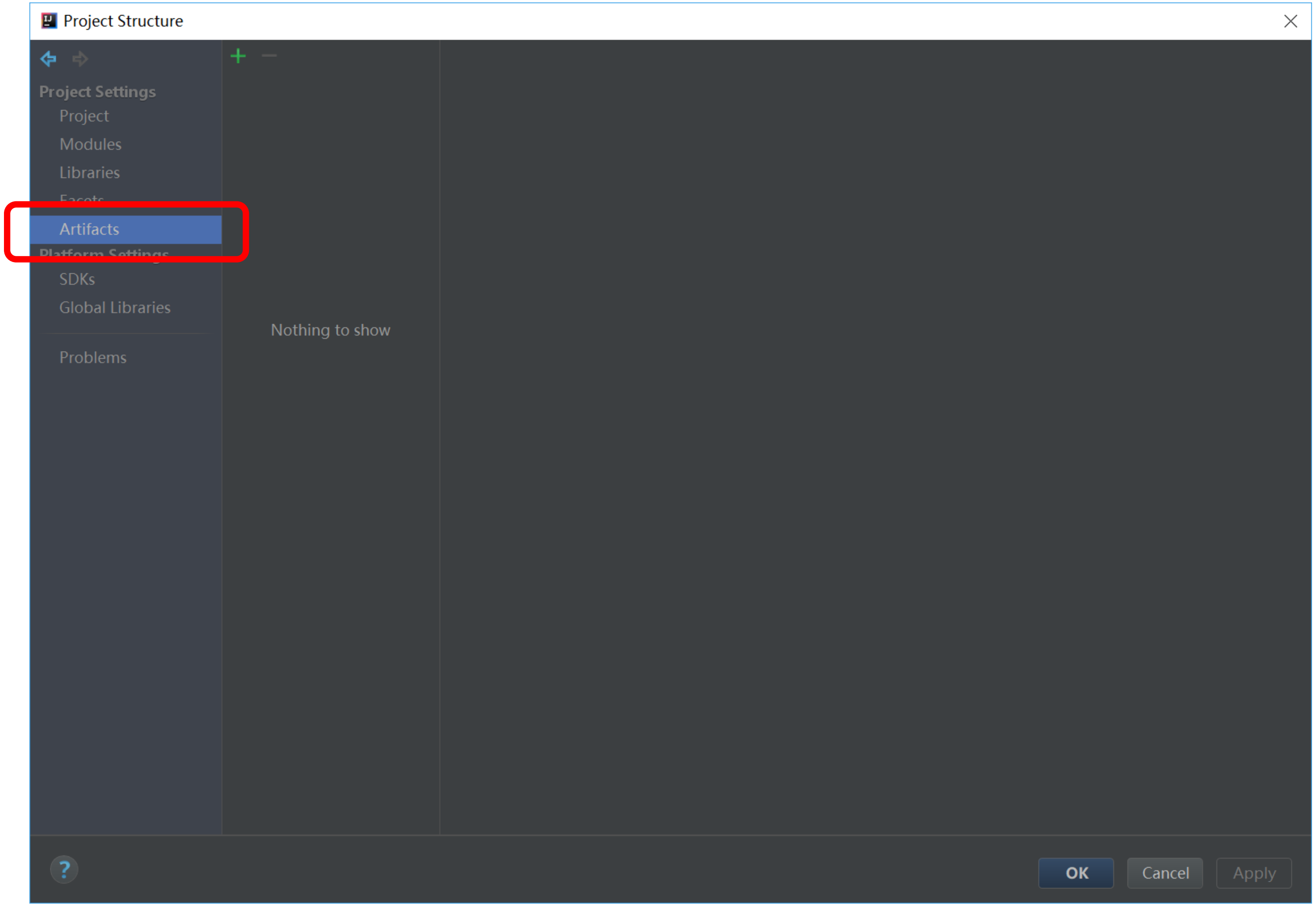
- 创建基于Spark的应用程序：创建一个简单的示例，验证是否能够连接到Spark集群



```scala
package peas

import org.apache.spark.{SparkConf, SparkContext}

/**
  * Demo to Connect the Spark Commodity Cluster.
  */
object DemoConnectSparkCluster extends App {
  // configure and start the connection to the Spark commodity cluster
  val sparkConf = new SparkConf()
    .setAppName("DemoConnectSparkCluster")
    .setMaster("spark://dc001.syhlab:7077")
    .setJars(List("D:\\Spark\\SparkProj\\parallel-pso-spark\\out\\artifacts\\parallel_pso_spark_jar\\parallel-pso-spark.jar"))
  val sc = new SparkContext(sparkConf)

  // calculate the value of PI
  val numSamples = 1000000000
  val count = sc.parallelize(1 to numSamples, 100)
    .filter({_ =>
      val x = math.random
      val y = math.random
      x * x + y * y < 1
    }).count()
  println(f"PI =~= ${4.0 * count / numSamples}%9.7f")

  // close the connection to the Spark commodity cluster
  sc.stop()
}
```

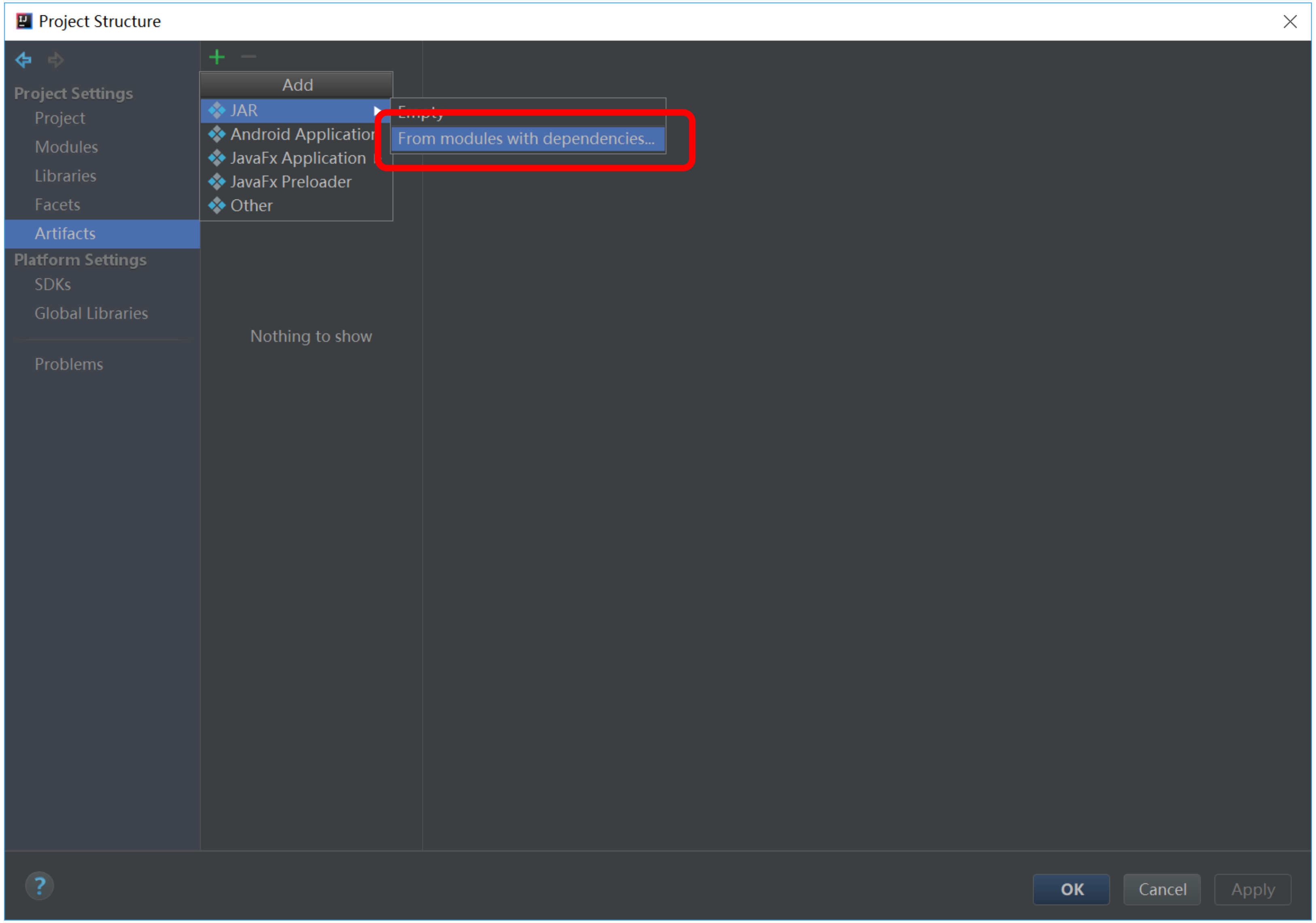
IntelliJ IDEA设置

- 创建基于Spark的应用程序：配置项目输出



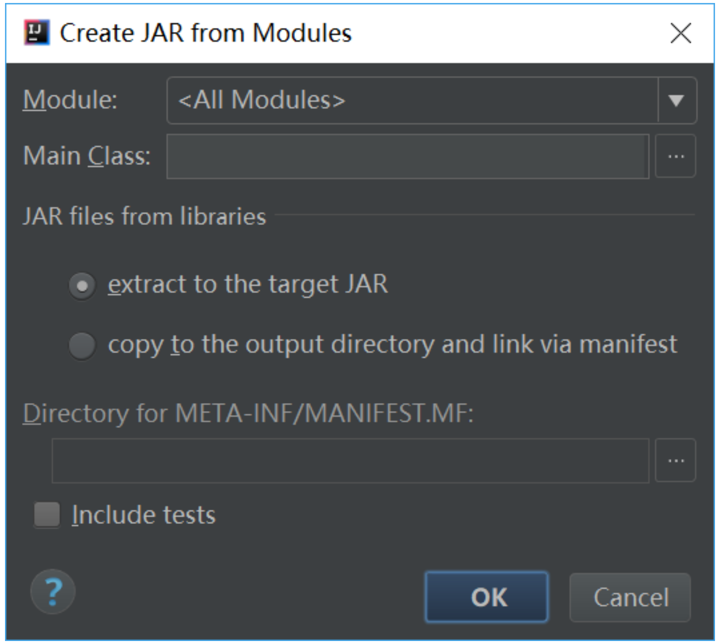
IntelliJ IDEA设置

- 创建基于Spark的应用程序：配置项目输出



IntelliJ IDEA设置

- 创建基于Spark的应用程序：配置项目输出



Create JAR from Modules

Module: <All Modules>

Main Class:

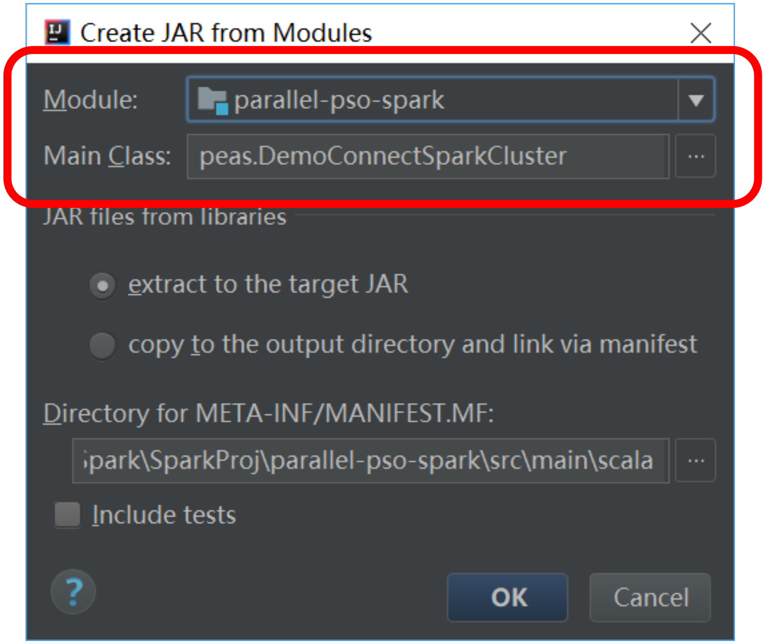
JAR files from libraries

● extract to the target JAR

○ copy to the output directory and link via manifest

Directory for META-INF/MANIFEST.MF:

□ Include tests



Create JAR from Modules

Module: parallel-pso-spark

Main Class: peas.DemoConnectSparkCluster

JAR files from libraries

● extract to the target JAR

○ copy to the output directory and link via manifest

Directory for META-INF/MANIFEST.MF:

;park\SparkProj\parallel-pso-spark\src\main\scala

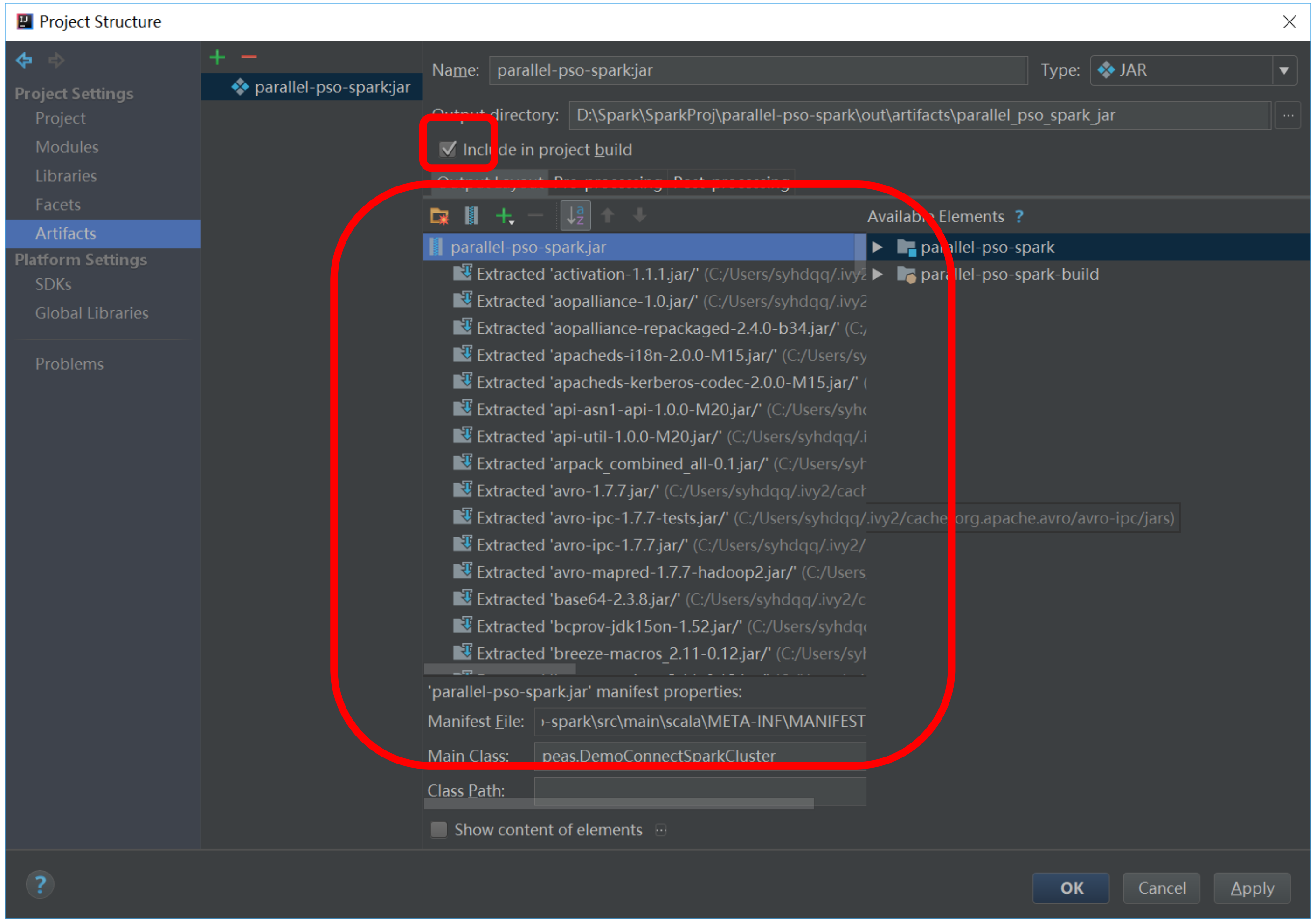
□ Include tests

IntelliJ IDEA设置

- 创建基于Spark的应用程序：配置项目输出

  如果以下项未能配置正确，将出现以下错误提示：

  *ERROR SparkContext: Failed to add *.jar to Spark environment java.lang.ClassNotFoundException*



Name: parallel-pso-spark:jar

Type: JAR

Output directory: D:\Spark\SparkProj\parallel-pso-spark\out\artifacts\parallel_pso_spark_jar

☑ Include in project build

parallel-pso-spark.jar

Extracted 'activation-1.1.1.jar/' (C:/Users/syhdqq/.ivy2

Extracted 'aopalliance-1.0.jar/' (C:/Users/syhdqq/.ivy2

Extracted 'aopalliance-repackaged-2.4.0-b34.jar/' (C:/

Extracted 'apacheds-i18n-2.0.0-M15.jar/' (C:/Users/sy

Extracted 'apacheds-kerberos-codec-2.0.0-M15.jar/' (

Extracted 'api-asn1-api-1.0.0-M20.jar/' (C:/Users/syh

Extracted 'api-util-1.0.0-M20.jar/' (C:/Users/syhdqq/.i

Extracted 'arpack_combined_all-0.1.jar/' (C:/Users/syh

Extracted 'avro-1.7.7.jar/' (C:/Users/syhdqq/.ivy2/cach

Extracted 'avro-ipc-1.7.7-tests.jar/' (C:/Users/syhdqq/.ivy2/cache/org.apache.avro/avro-ipc/jars)

Extracted 'avro-ipc-1.7.7.jar/' (C:/Users/syhdqq/.ivy2/

Extracted 'avro-mapred-1.7.7-hadoop2.jar/' (C:/Users

Extracted 'base64-2.3.8.jar/' (C:/Users/syhdqq/.ivy2/c

Extracted 'bcprov-jdk15on-1.52.jar/' (C:/Users/syhdq

Extracted 'breeze-macros_2.11-0.12.jar/' (C:/Users/syh

'parallel-pso-spark.jar' manifest properties:

Manifest File: -spark\src\main\scala\META-INF\MANIFEST

Main Class: peas.DemoConnectSparkCluster

Class Path:

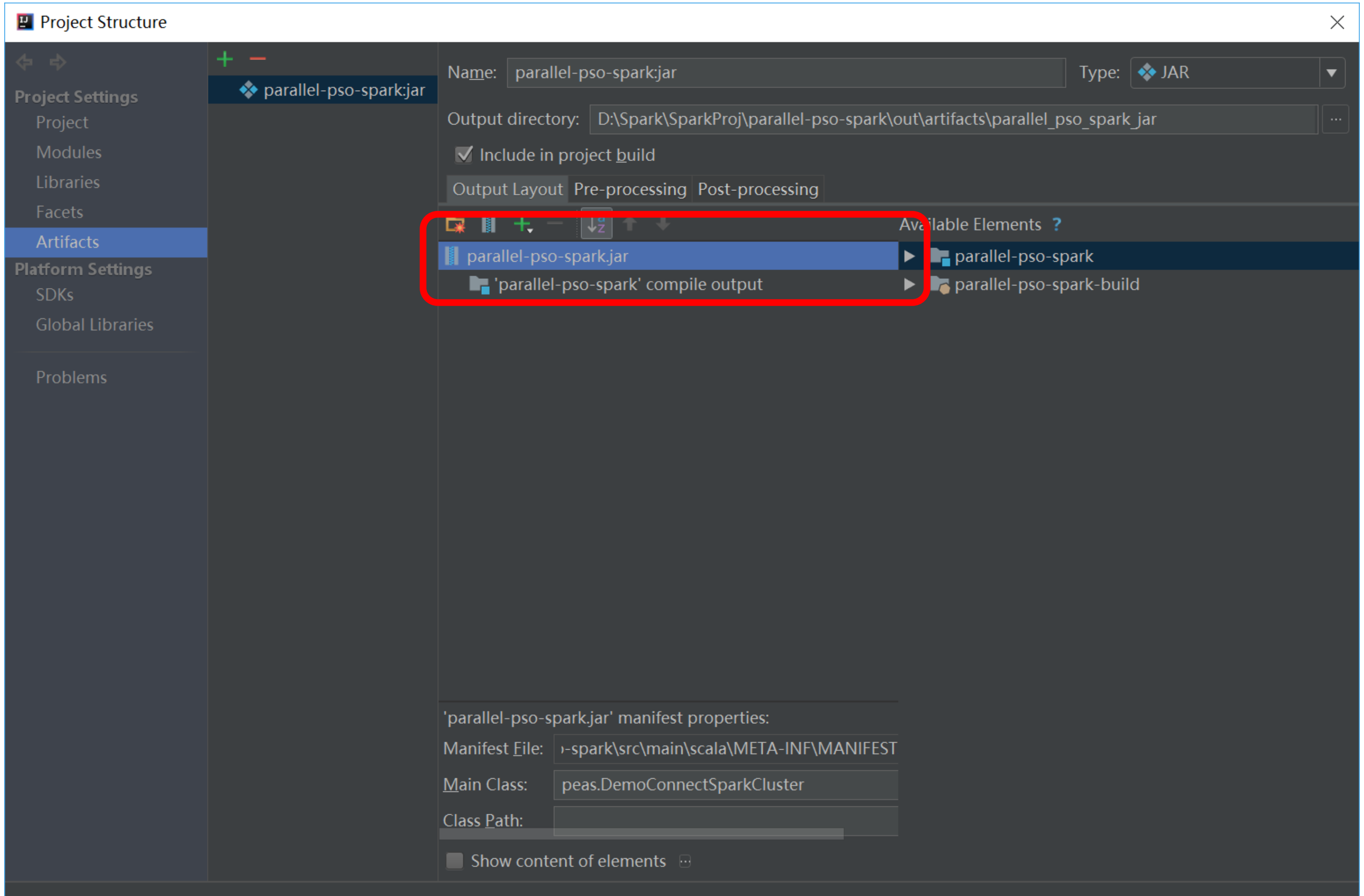
□ Show content of elements

IntelliJ IDEA设置

- 创建基于Spark的应用程序：配置项目输出

  如果以下项未能配置正确，将出现以下错误提示：

  *ERROR SparkContext: Failed to add *.jar to Spark environment java.lang.ClassNotFoundException*


Project Structure

parallel-pso-spark:jar

Name: parallel-pso-spark:jar

Type: JAR

Output directory: D:\Spark\SparkProj\parallel-pso-spark\out\artifacts\parallel_pso_spark_jar

☑ Include in project build

parallel-pso-spark.jar
'parallel-pso-spark' compile output

parallel-pso-spark
parallel-pso-spark-build

'parallel-pso-spark.jar' manifest properties:

Manifest File: ›-spark\src\main\scala\META-INF\MANIFEST

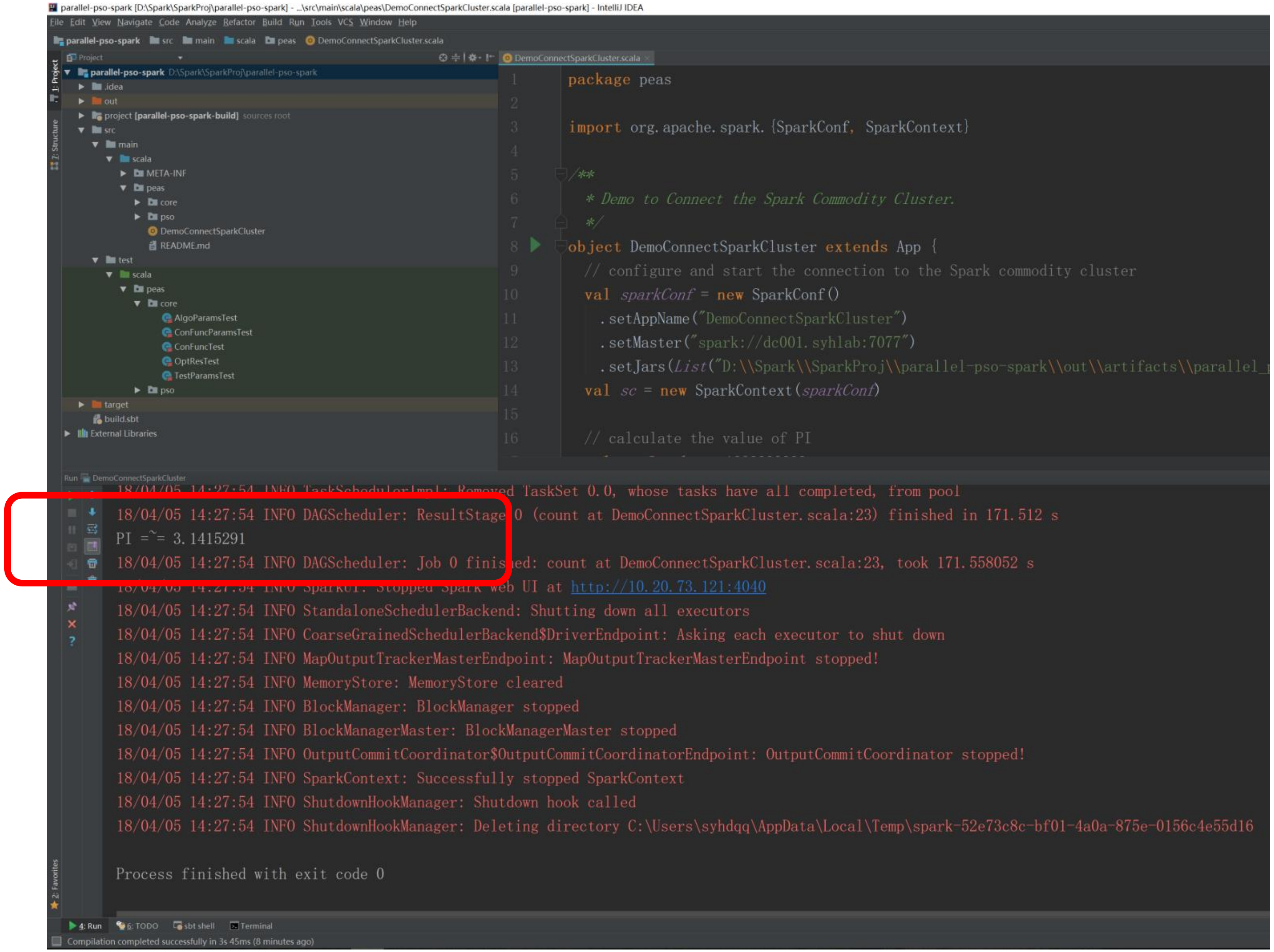
Main Class: peas.DemoConnectSparkCluster

Class Path:

☐ Show content of elements

IntelliJ IDEA设置

- 创建基于Spark的应用程序：程序输出（部分）


parallel-pso-spark [D:\Spark\SparkProj\parallel-pso-spark] - ...\src\main\scala\peas\DemoConnectSparkCluster.scala [parallel-pso-spark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
parallel-pso-spark src main scala peas DemoConnectSparkCluster.scala
Project
DemoConnectSparkCluster.scala
parallel-pso-spark D:\Spark\SparkProj\parallel-pso-spark
.idea
out
project [parallel-pso-spark-build] sources root
src
main
scala
META-INF
peas
core
pso
DemoConnectSparkCluster
README.md
test
scala
peas
core
AlgoParamsTest
ConFuncParamsTest
ConFuncTest
OptResTest
TestParamsTest
pso
target
build.sbt
External Libraries
1: Project
2: Structure
2: Favorites
Run DemoConnectSparkCluster
4: Run
6: TODO
sbt shell
Terminal

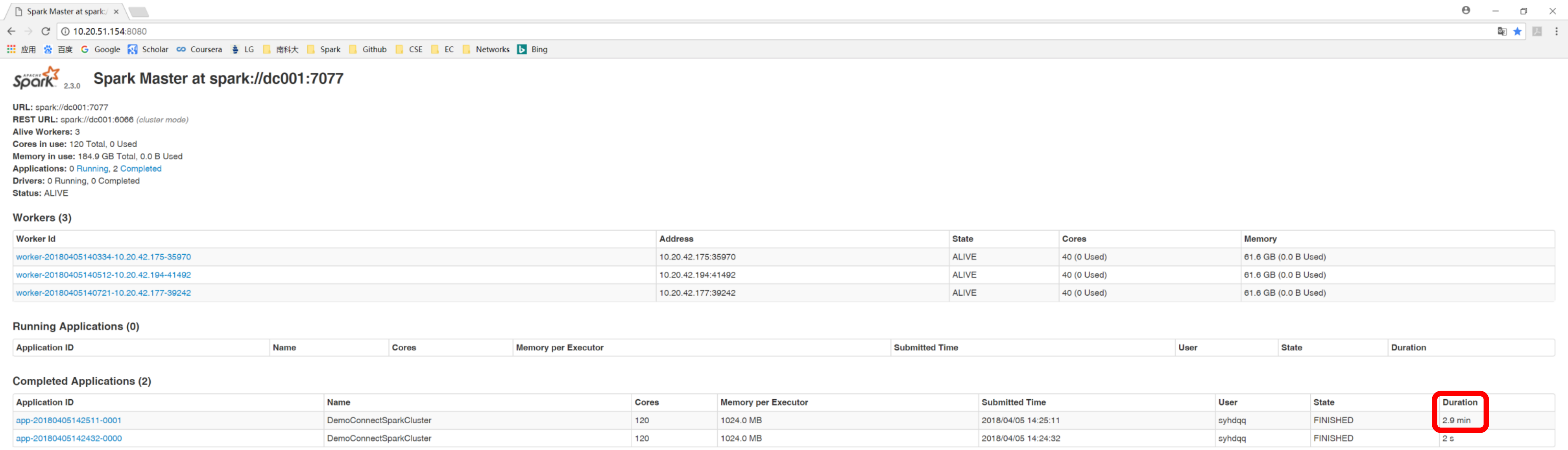
```scala
package peas

import org.apache.spark.{SparkConf, SparkContext}

/**
  * Demo to Connect the Spark Commodity Cluster.
  */
object DemoConnectSparkCluster extends App {
  // configure and start the connection to the Spark commodity cluster
  val sparkConf = new SparkConf()
    .setAppName("DemoConnectSparkCluster")
    .setMaster("spark://dc001.syhlab:7077")
    .setJars(List("D:\\Spark\\SparkProj\\parallel-pso-spark\\out\\artifacts\\parallel_
  val sc = new SparkContext(sparkConf)

  // calculate the value of PI
```

```
18/04/05 14:27:54 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/04/05 14:27:54 INFO DAGScheduler: ResultStage 0 (count at DemoConnectSparkCluster.scala:23) finished in 171.512 s
PI =~= 3.1415291
18/04/05 14:27:54 INFO DAGScheduler: Job 0 finished: count at DemoConnectSparkCluster.scala:23, took 171.558052 s
18/04/05 14:27:54 INFO SparkUI: Stopped Spark web UI at http://10.20.73.121:4040
18/04/05 14:27:54 INFO StandaloneSchedulerBackend: Shutting down all executors
18/04/05 14:27:54 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
18/04/05 14:27:54 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/04/05 14:27:54 INFO MemoryStore: MemoryStore cleared
18/04/05 14:27:54 INFO BlockManager: BlockManager stopped
18/04/05 14:27:54 INFO BlockManagerMaster: BlockManagerMaster stopped
18/04/05 14:27:54 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/04/05 14:27:54 INFO SparkContext: Successfully stopped SparkContext
18/04/05 14:27:54 INFO ShutdownHookManager: Shutdown hook called
18/04/05 14:27:54 INFO ShutdownHookManager: Deleting directory C:\Users\syhdqq\AppData\Local\Temp\spark-52e73c8c-bf01-4a0a-875e-0156c4e55d16

Process finished with exit code 0
```

Compilation completed successfully in 3s 45ms (8 minutes ago)

IntelliJ IDEA设置

- 创建基于Spark的应用程序：程序监控（通过WEB UI）



Spark Master at spark://dc001:7077

URL: spark://dc001:7077
REST URL: spark://dc001:6066 *(cluster mode)*
Alive Workers: 3
Cores in use: 120 Total, 0 Used
Memory in use: 184.9 GB Total, 0.0 B Used
Applications: 0 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory
worker-20180405140334-10.20.42.175-35970	10.20.42.175:35970	ALIVE	40 (0 Used)	61.6 GB (0.0 B Used)
worker-20180405140512-10.20.42.194-41492	10.20.42.194:41492	ALIVE	40 (0 Used)	61.6 GB (0.0 B Used)
worker-20180405140721-10.20.42.177-39242	10.20.42.177:39242	ALIVE	40 (0 Used)	61.6 GB (0.0 B Used)

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20180405142511-0001	DemoConnectSparkCluster	120	1024.0 MB	2018/04/05 14:25:11	syhdqq	FINISHED	2.9 min
app-20180405142432-0000	DemoConnectSparkCluster	120	1024.0 MB	2018/04/05 14:24:32	syhdqq	FINISHED	2 s

常见问题汇总

- **TaskSchedulerImpl: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources.**

解决方案之一（造成此问题的原因可能有多种，这里只给出我个人遇到的情况）：

确保**Windows**开发端与**Spark**集群处于同一级的局域网内：

特别是在**Windows**开发端使用路由器（例如**TP-LINK**）进行网络连接时，当**Spark**集群处于上一级**IP**地址（例如*10.20.2.5*）而**Windows**开发端处于下一级**IP**地址时（例如**192.168.7.9**），需要将路由器转化为交换机模式。

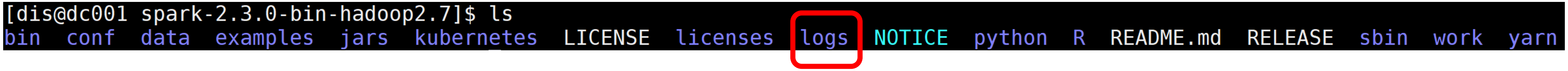
- **WARN TransportChannelHandler: Exception in connection from 10.20.2.5:54321 java.io.IOException: Connection reset by peer**

解决方案之一（造成此问题的原因可能有多种，这里只给出我个人遇到的情况）：

关闭**Windows**防火墙　　*可选*

常见问题汇总

*查看、分析log日志*是对Spark应用程序进行排错的重要手段。



```
[dis@dc001 spark-2.3.0-bin-hadoop2.7]$ ls
bin  conf  data  examples  jars  kubernetes  LICENSE  licenses  logs  NOTICE  python  R  README.md  RELEASE  sbin  work  yarn
```