零基础学习Spark 1.x应用开发系列课程

如何使用IDEA开发程序

讲师-梦琪

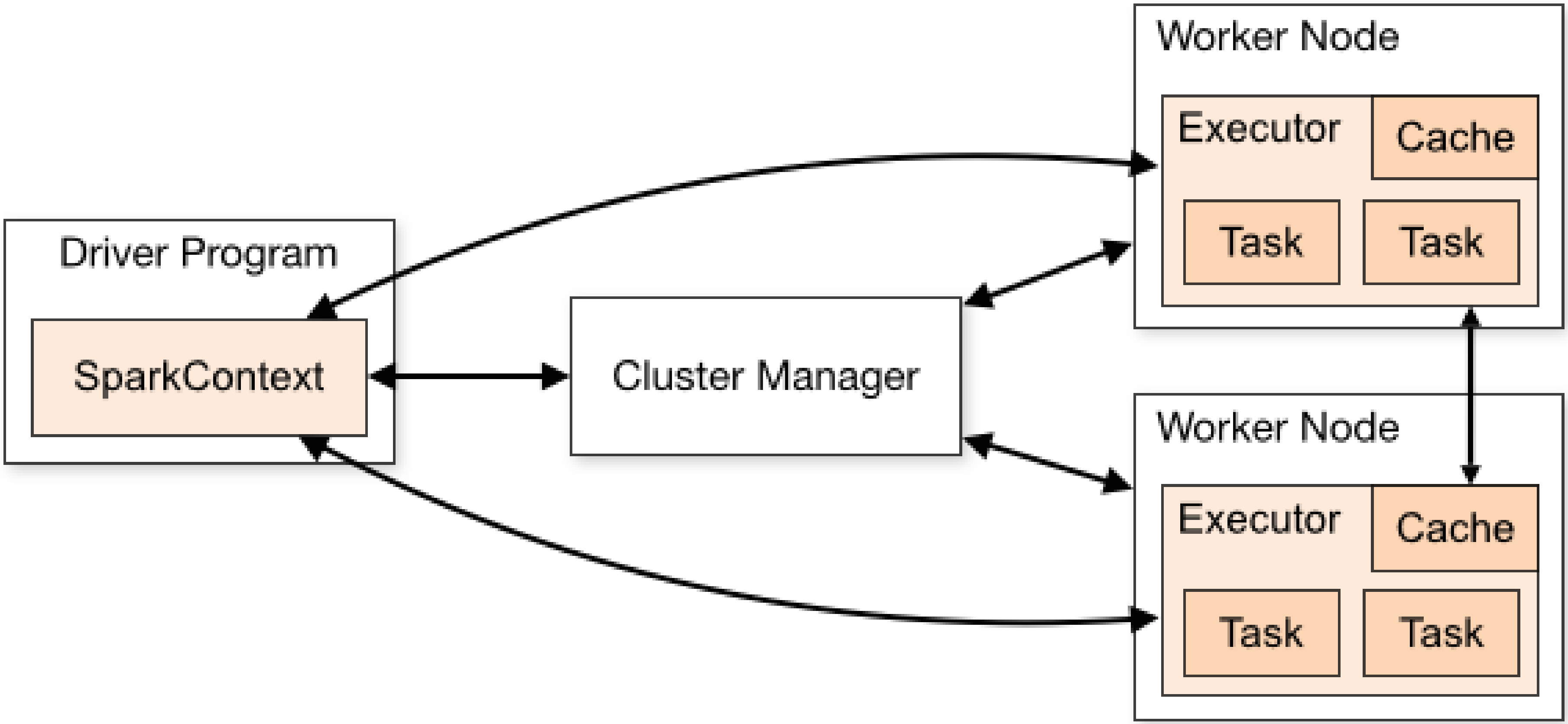
版权声明

【声明】本视频和讲义等均为云帆大数据网络课程的教学资料，所有资料只能在课程内使用，不允许在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问云帆大数据官方培训网站

http://www.cloudyhadoop.com

Spark Running Architecture



Spark Running Architecture

1、构建Spark Application运行环境；

在Driver Program中新建SparkContext（包含sparkcontext的程序称为Driver Program）；

Spark Application运行的表现方式为：在集群上运行着一组独立的executor进程，这些进程由sparkcontext来协调；

2、SparkContext向资源管理器申请运行Executor资源，并启动StandaloneExecutorBackend，executor向sparkcontent申请task；

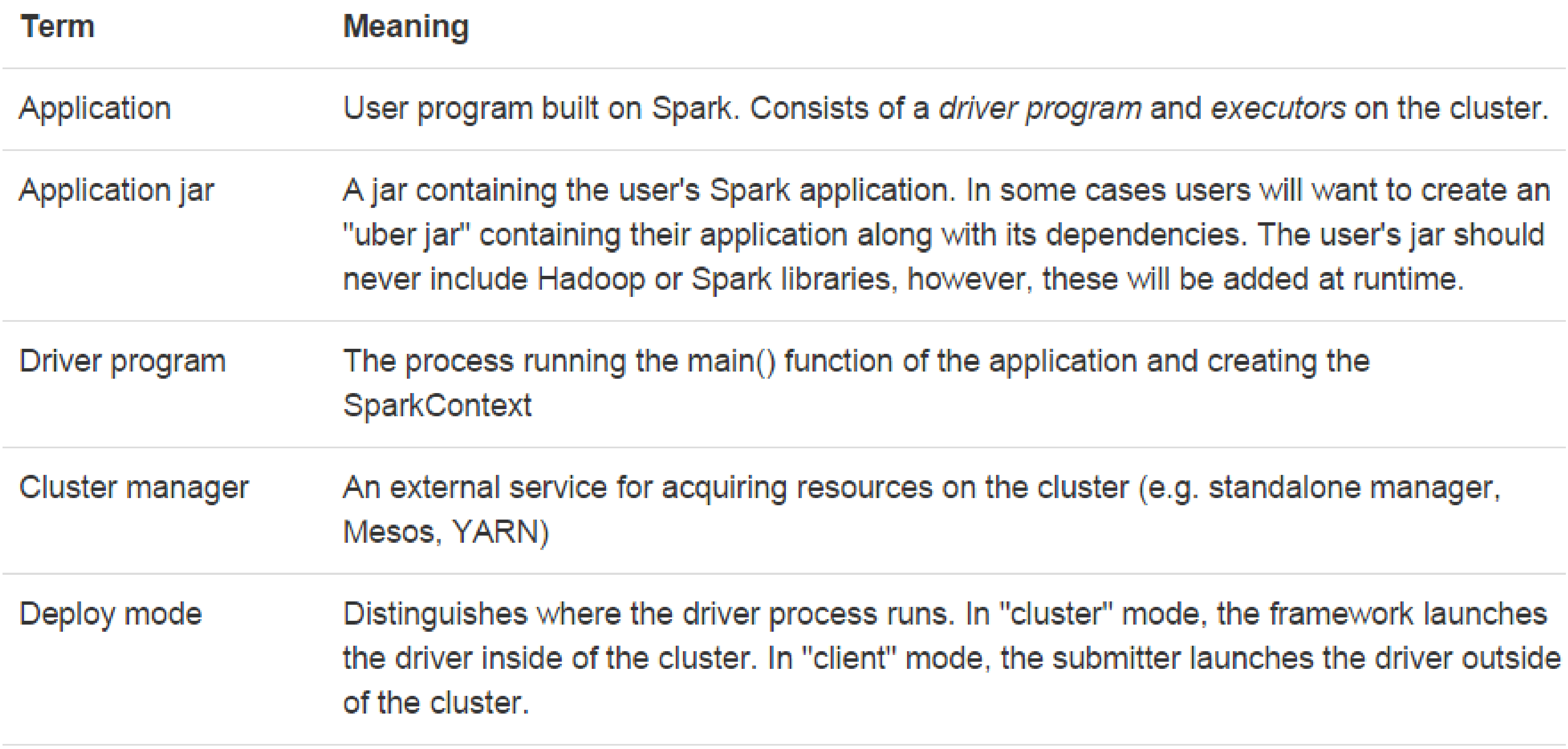
集群通过SparkContext连接到不同的cluster manager(standalone、yarn、mesos)，cluster manager为运行应用的Executor分配资源；一旦连接建立之后，Spark每个Application就会获得各个节点上的Executor（进程）；每个Application都有自己独立的executor进程；Executor才是真正运行在WorkNode上的工作进程，它们为应用来计算或者存储数据；

3、SparkContext获取到executor之后，Application的应用代码将会被发送到各个executor；

4、SparkContext构建RDD DAG图，将RDD DAG图分解成Stage DAG图，将Stage提交给TaskScheduler，最后由TaskScheduler将Task发送给Executor运行；

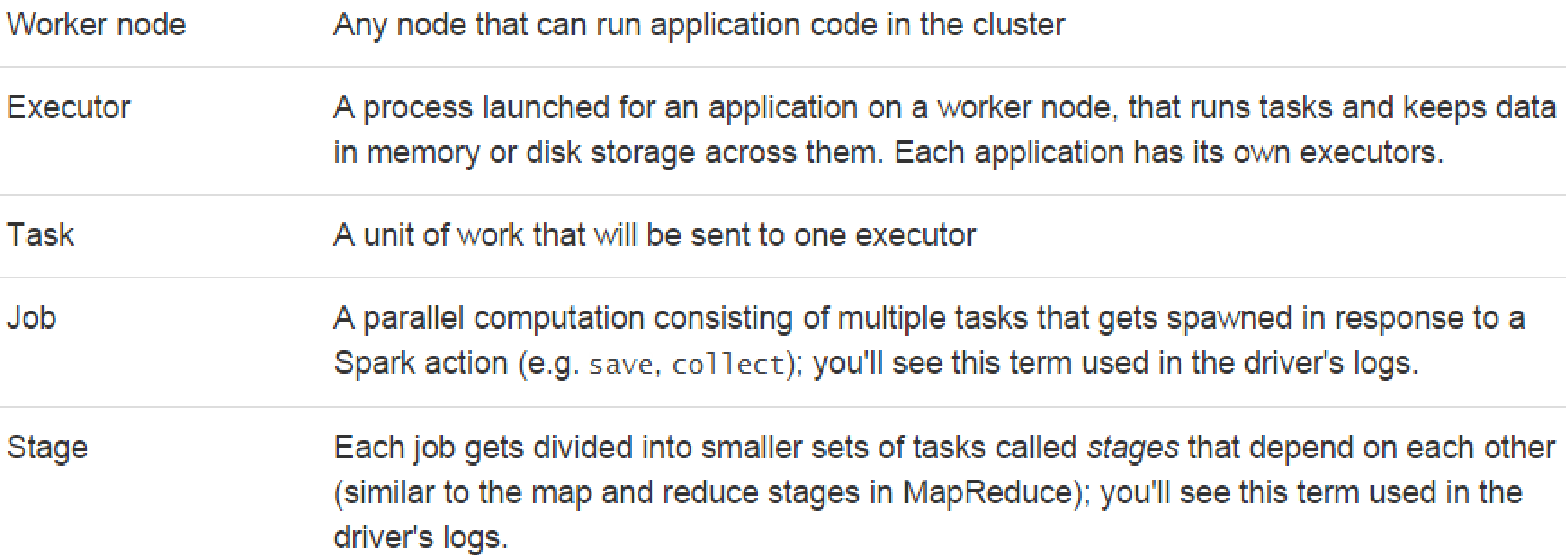
5、Task在Executor上运行，运行完毕后释放所有资源；

Cluster Concepts



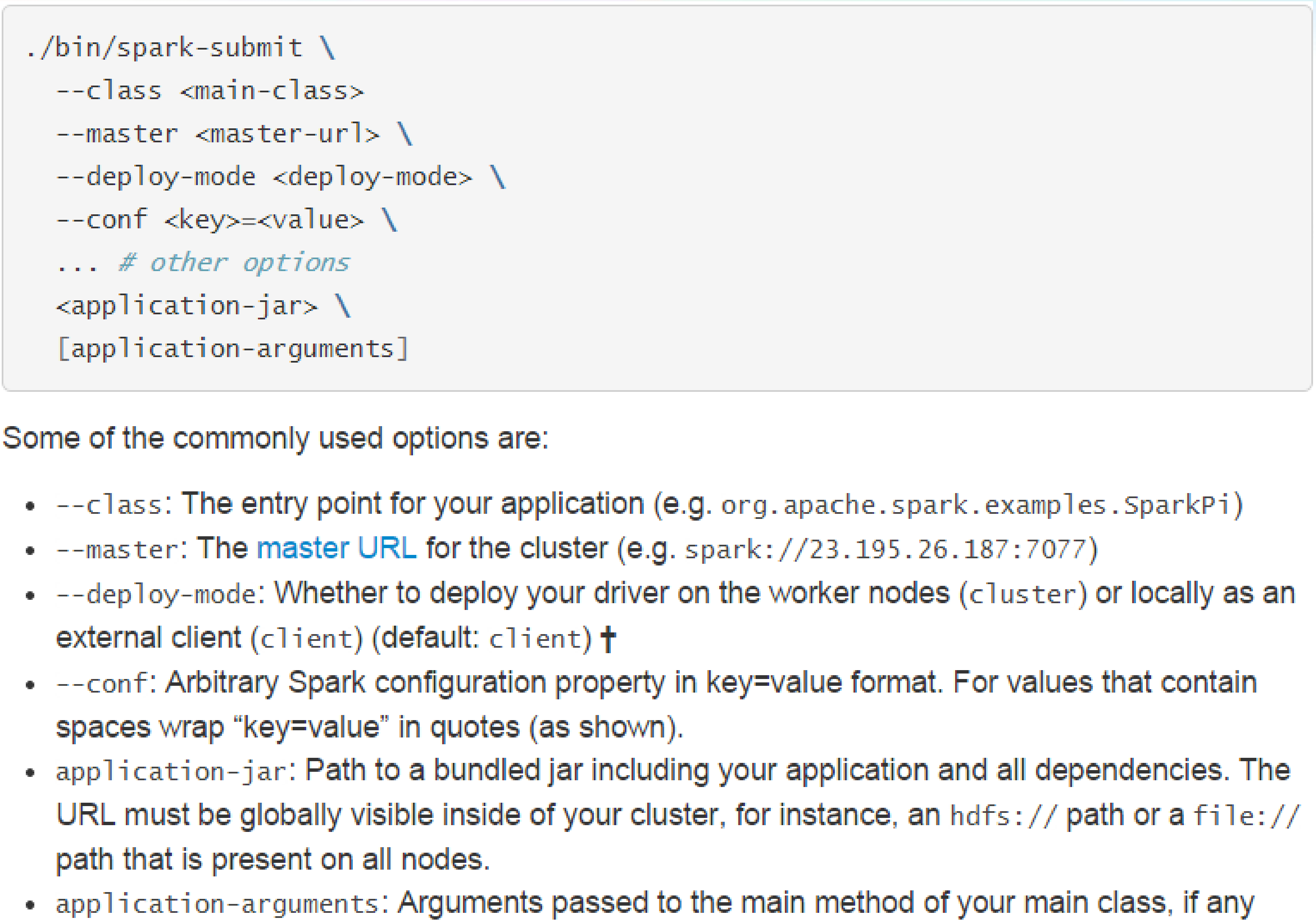
Term	Meaning
Application	User program built on Spark. Consists of a *driver program* and *executors* on the cluster.
Application jar	A jar containing the user's Spark application. In some cases users will want to create an "uber jar" containing their application along with its dependencies. The user's jar should never include Hadoop or Spark libraries, however, these will be added at runtime.
Driver program	The process running the main() function of the application and creating the SparkContext
Cluster manager	An external service for acquiring resources on the cluster (e.g. standalone manager, Mesos, YARN)
Deploy mode	Distinguishes where the driver process runs. In "cluster" mode, the framework launches the driver inside of the cluster. In "client" mode, the submitter launches the driver outside of the cluster.

Cluster Concepts



Worker node	Any node that can run application code in the cluster
Executor	A process launched for an application on a worker node, that runs tasks and keeps data in memory or disk storage across them. Each application has its own executors.
Task	A unit of work that will be sent to one executor
Job	A parallel computation consisting of multiple tasks that gets spawned in response to a Spark action (e.g. `save`, `collect`); you'll see this term used in the driver's logs.
Stage	Each job gets divided into smaller sets of tasks called *stages* that depend on each other (similar to the map and reduce stages in MapReduce); you'll see this term used in the driver's logs.

Launching Applications with spark-submit

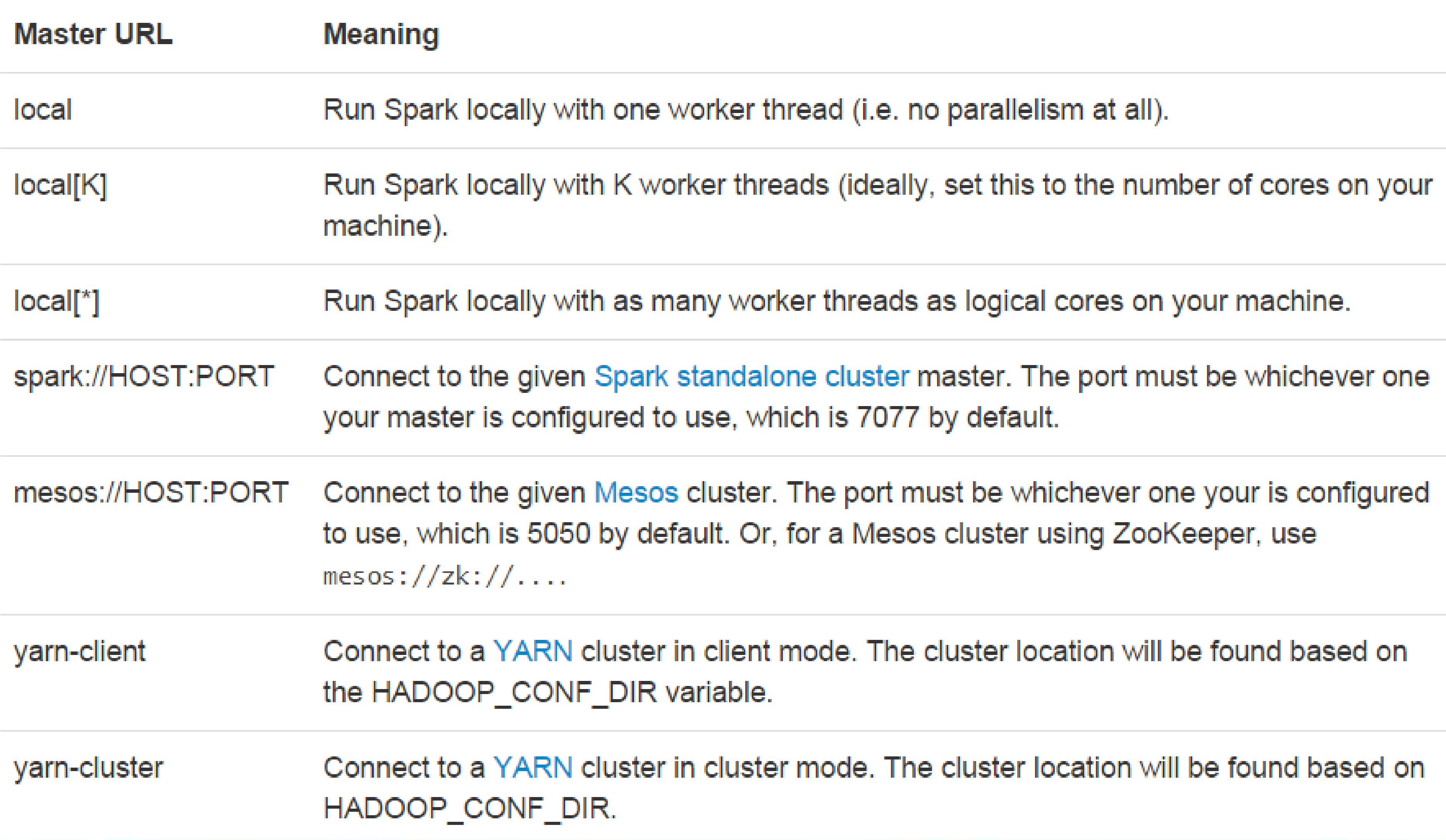


```
./bin/spark-submit \
  --class <main-class>
  --master <master-url> \
  --deploy-mode <deploy-mode> \
  --conf <key>=<value> \
  ... # other options
  <application-jar> \
  [application-arguments]
```

Some of the commonly used options are:

- `--class`: The entry point for your application (e.g. `org.apache.spark.examples.SparkPi`)
- `--master`: The master URL for the cluster (e.g. `spark://23.195.26.187:7077`)
- `--deploy-mode`: Whether to deploy your driver on the worker nodes (`cluster`) or locally as an external client (`client`) (default: `client`) †
- `--conf`: Arbitrary Spark configuration property in key=value format. For values that contain spaces wrap “key=value” in quotes (as shown).
- `application-jar`: Path to a bundled jar including your application and all dependencies. The URL must be globally visible inside of your cluster, for instance, an `hdfs://` path or a `file://` path that is present on all nodes.
- `application-arguments`: Arguments passed to the main method of your main class, if any

Master URLs



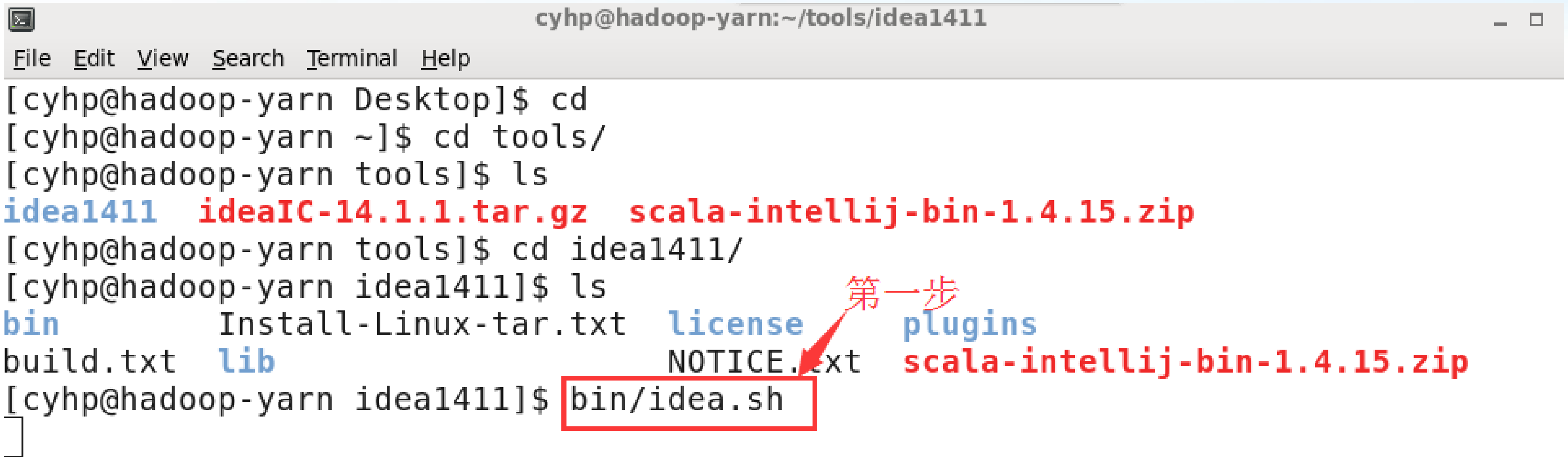
Master URL	Meaning
local	Run Spark locally with one worker thread (i.e. no parallelism at all).
local[K]	Run Spark locally with K worker threads (ideally, set this to the number of cores on your machine).
local[*]	Run Spark locally with as many worker threads as logical cores on your machine.
spark://HOST:PORT	Connect to the given Spark standalone cluster master. The port must be whichever one your master is configured to use, which is 7077 by default.
mesos://HOST:PORT	Connect to the given Mesos cluster. The port must be whichever one your is configured to use, which is 5050 by default. Or, for a Mesos cluster using ZooKeeper, use `mesos://zk://....`
yarn-client	Connect to a YARN cluster in client mode. The cluster location will be found based on the HADOOP_CONF_DIR variable.
yarn-cluster	Connect to a YARN cluster in cluster mode. The cluster location will be found based on HADOOP_CONF_DIR.

零基础学习Spark 1.x应用开发系列课程

安装IDEA及Scala插件

讲师-梦琪

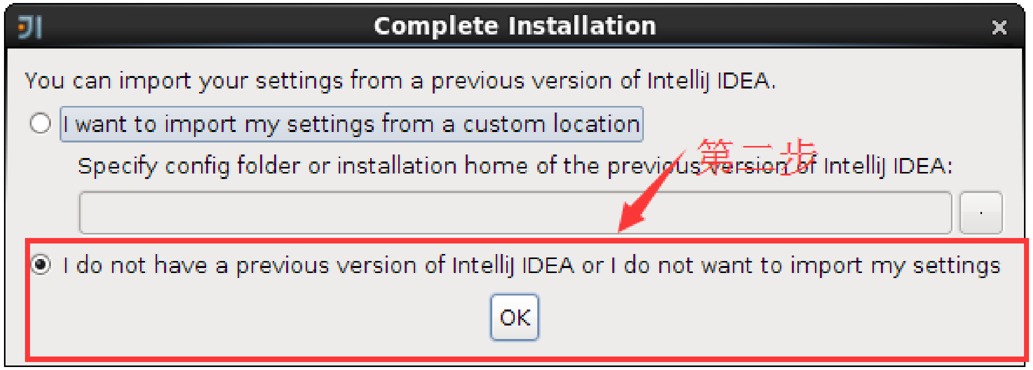
启动IDEA



cyhp@hadoop-yarn:~/tools/idea1411
File Edit View Search Terminal Help

```
[cyhp@hadoop-yarn Desktop]$ cd
[cyhp@hadoop-yarn ~]$ cd tools/
[cyhp@hadoop-yarn tools]$ ls
idea1411  ideaIC-14.1.1.tar.gz  scala-intellij-bin-1.4.15.zip
[cyhp@hadoop-yarn tools]$ cd idea1411/
[cyhp@hadoop-yarn idea1411]$ ls
bin          Install-Linux-tar.txt  license     plugins
build.txt  lib                      NOTICE.txt  scala-intellij-bin-1.4.15.zip
[cyhp@hadoop-yarn idea1411]$ bin/idea.sh
```

第一步



You can import your settings from a previous version of IntelliJ IDEA.

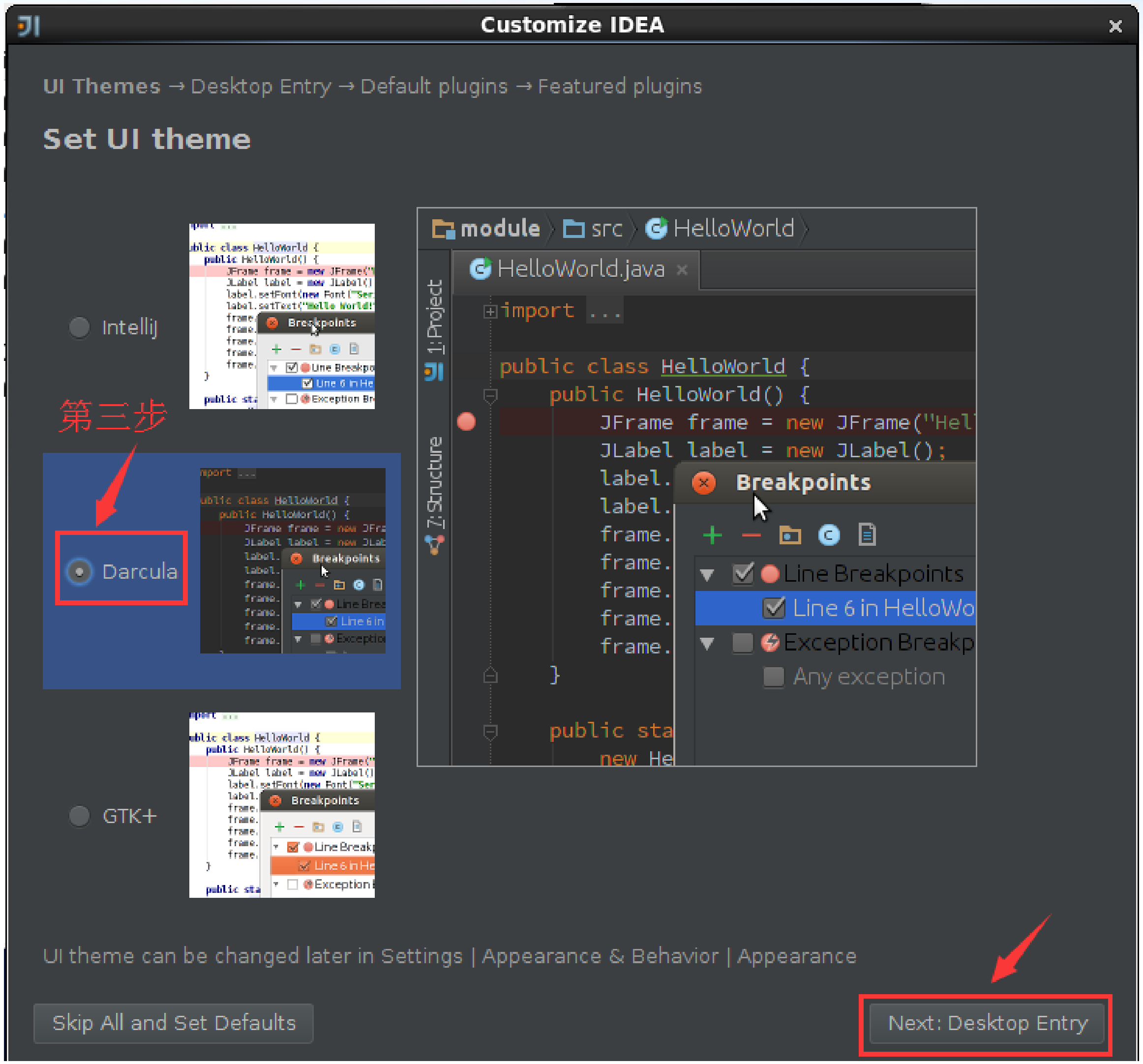
○ I want to import my settings from a custom location

Specify config folder or installation home of the previous version of IntelliJ IDEA:

● I do not have a previous version of IntelliJ IDEA or I do not want to import my settings

第二步

选择IDEA UI Theme



Customize IDEA

UI Themes → Desktop Entry → Default plugins → Featured plugins

Set UI theme

○ IntelliJ

第三步

● Darcula

○ GTK+

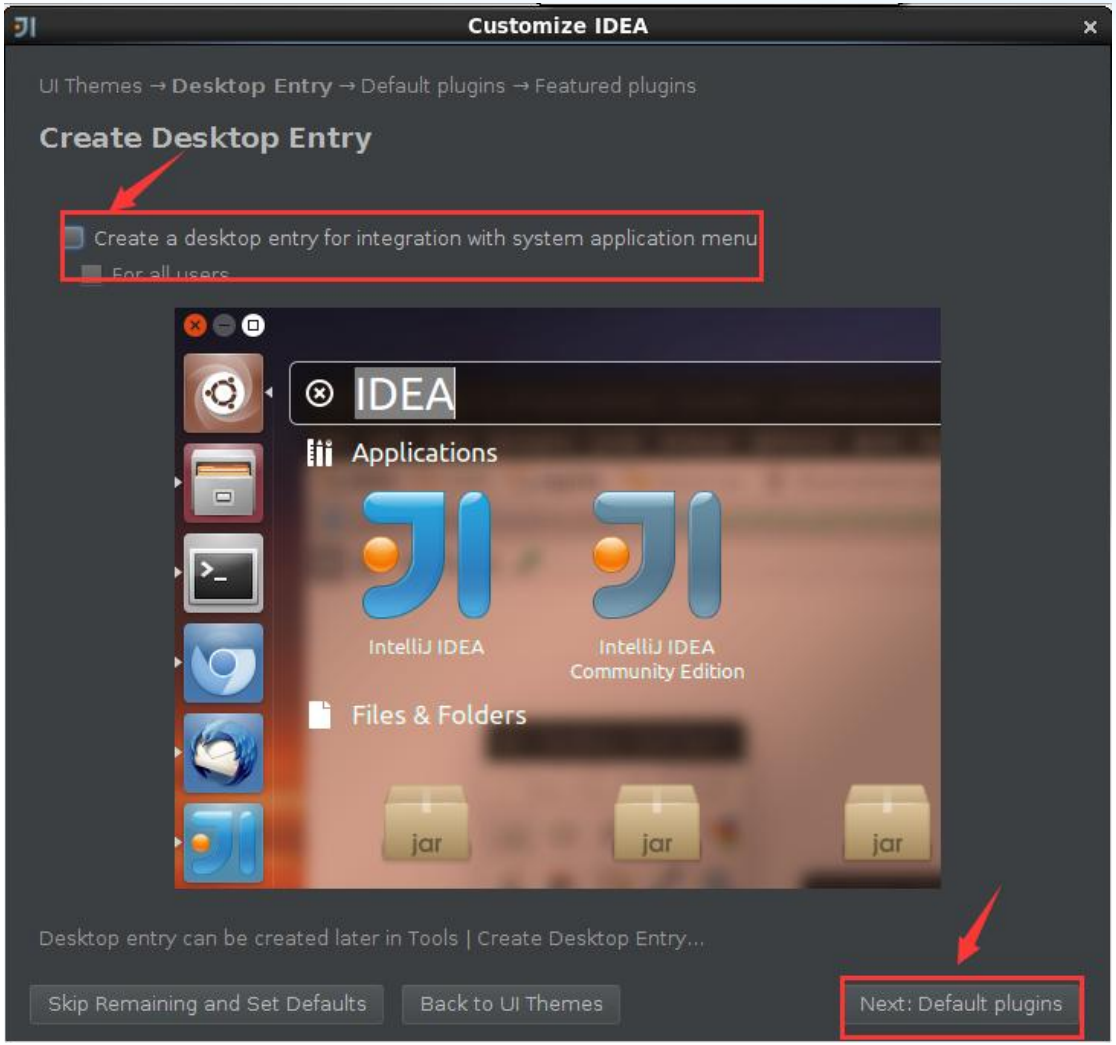
```
module 〉 src 〉 HelloWorld 〉
HelloWorld.java ×
import ...

public class HelloWorld {
    public HelloWorld() {
        JFrame frame = new JFrame("Hel
        JLabel label = new JLabel();
        label.
        label.
        frame.
        frame.
        frame.
        frame.
        frame.
    }

    public sta
        new He
```

UI theme can be changed later in Settings | Appearance & Behavior | Appearance

安装IDEA



Customize IDEA

UI Themes → **Desktop Entry** → Default plugins → Featured plugins

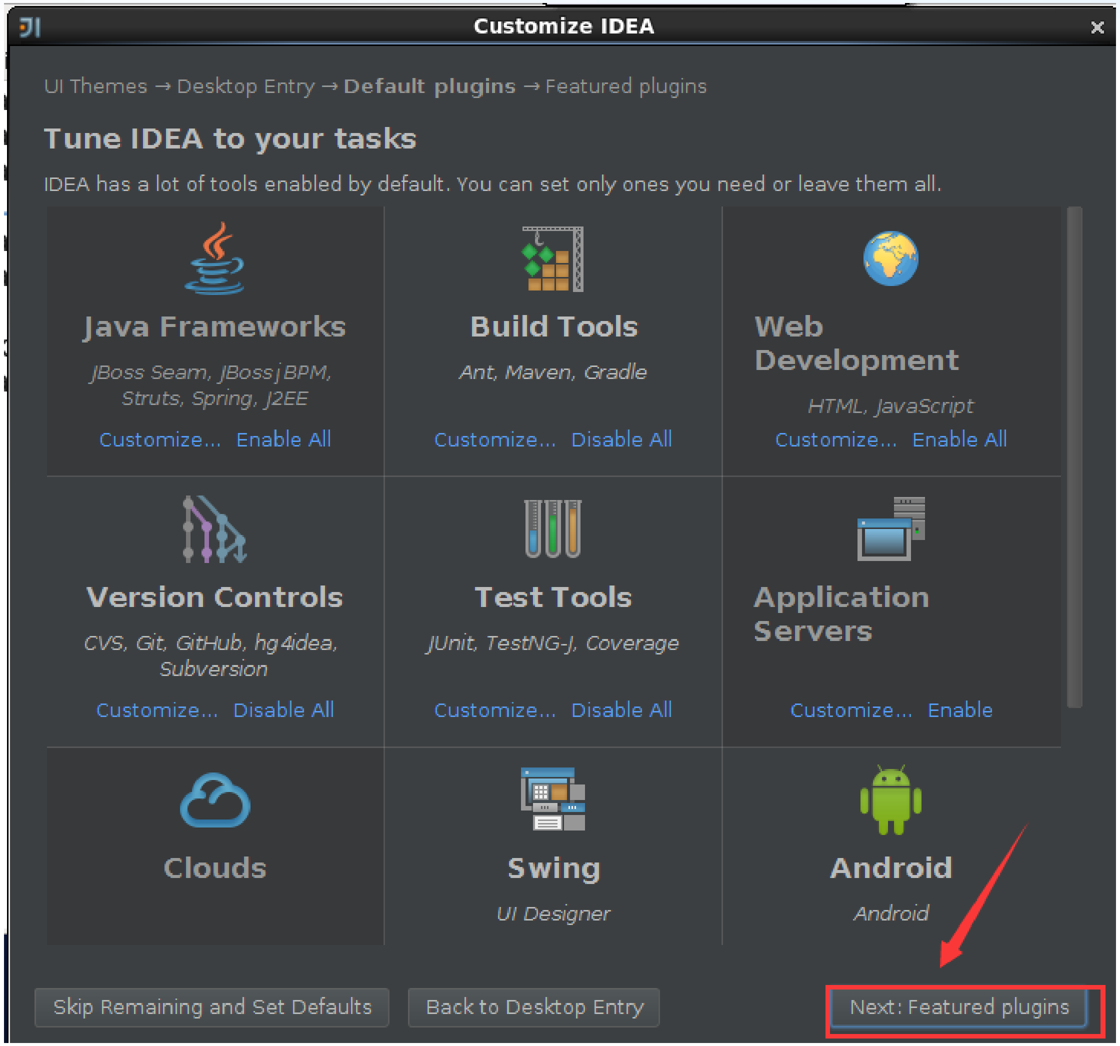
Create Desktop Entry

□ Create a desktop entry for integration with system application menu

□ For all users

Desktop entry can be created later in Tools | Create Desktop Entry...

安装IDEA



Customize IDEA

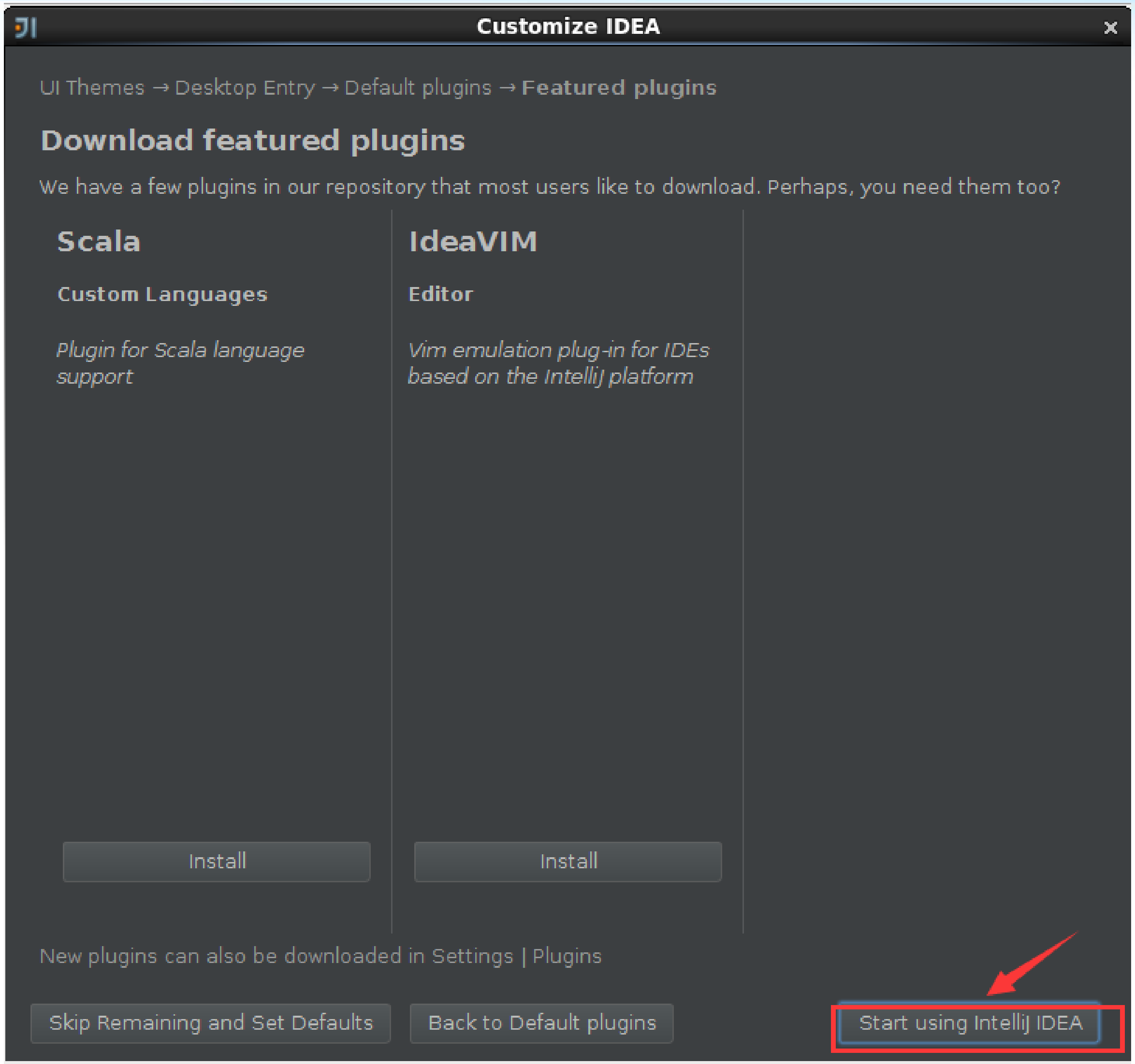
UI Themes → Desktop Entry → **Default plugins** → Featured plugins

Tune IDEA to your tasks

IDEA has a lot of tools enabled by default. You can set only ones you need or leave them all.

Java Frameworks *JBoss Seam, JBoss jBPM, Struts, Spring, J2EE* Customize... Enable All	**Build Tools** *Ant, Maven, Gradle* Customize... Disable All	**Web Development** *HTML, JavaScript* Customize... Enable All
Version Controls *CVS, Git, GitHub, hg4idea, Subversion* Customize... Disable All	**Test Tools** *JUnit, TestNG-J, Coverage* Customize... Disable All	**Application Servers** Customize... Enable
Clouds	**Swing** *UI Designer*	**Android** *Android*

安装IDEA



Customize IDEA

UI Themes → Desktop Entry → Default plugins → **Featured plugins**

Download featured plugins

We have a few plugins in our repository that most users like to download. Perhaps, you need them too?

Scala

Custom Languages

Plugin for Scala language support

Install

IdeaVIM

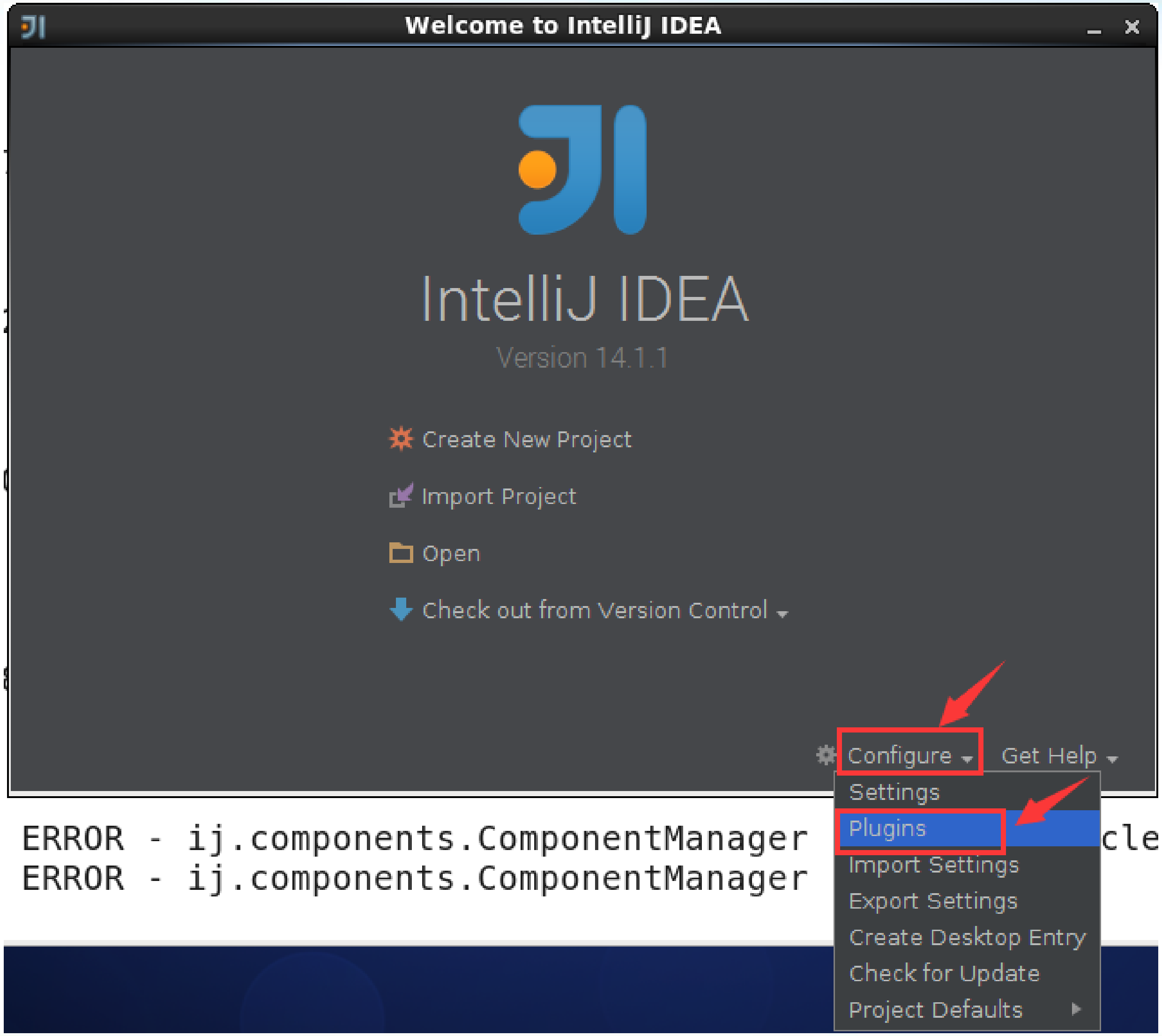
Editor

Vim emulation plug-in for IDEs based on the IntelliJ platform

Install

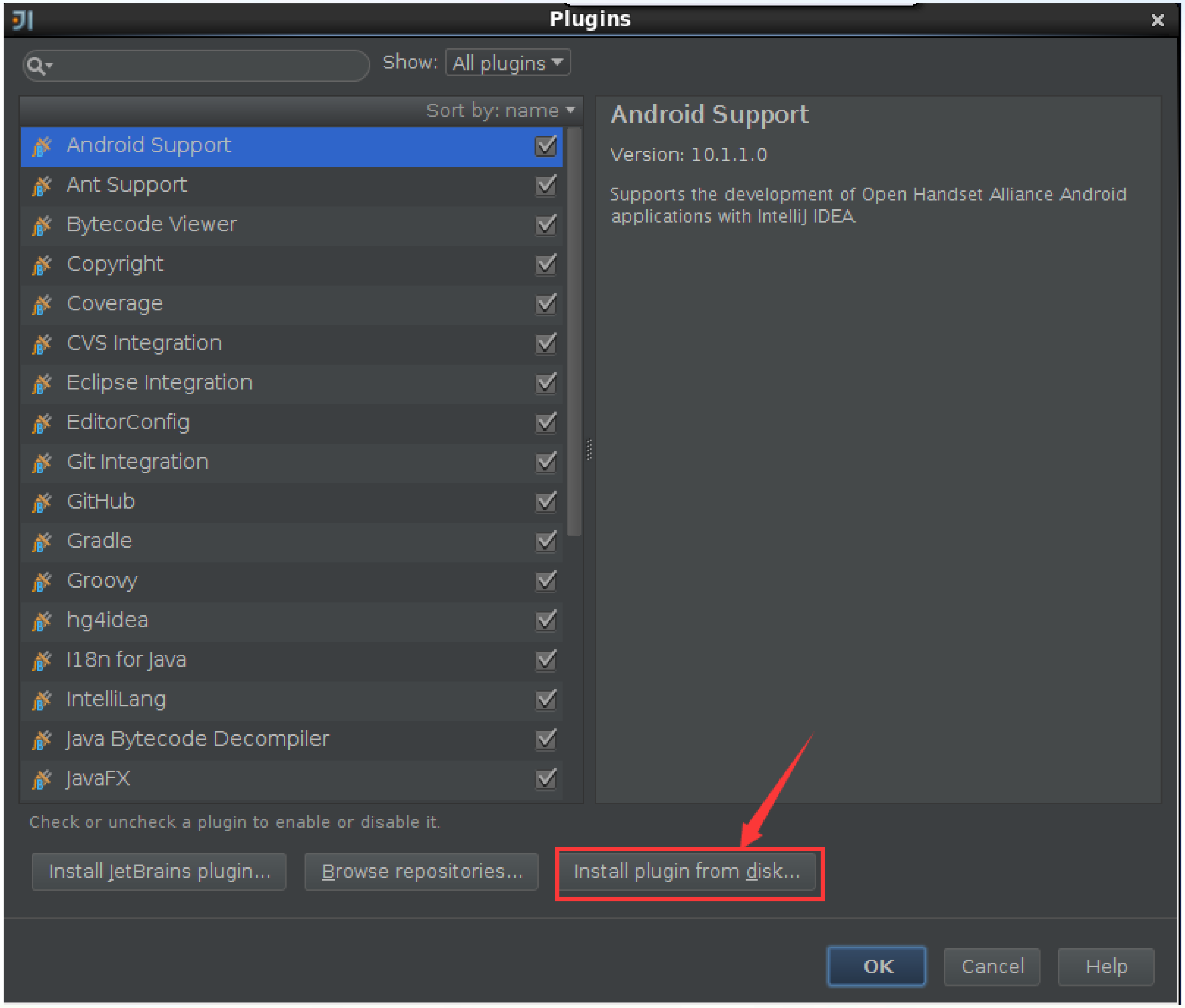
New plugins can also be downloaded in Settings | Plugins

启动IDEA



```
ERROR - ij.components.ComponentManager                    cle
ERROR - ij.components.ComponentManager
```

安装Scala 插件



- Android Support ☑
- Ant Support ☑
- Bytecode Viewer ☑
- Copyright ☑
- Coverage ☑
- CVS Integration ☑
- Eclipse Integration ☑
- EditorConfig ☑
- Git Integration ☑
- GitHub ☑
- Gradle ☑
- Groovy ☑
- hg4idea ☑
- I18n for Java ☑
- IntelliLang ☑
- Java Bytecode Decompiler ☑
- JavaFX ☑

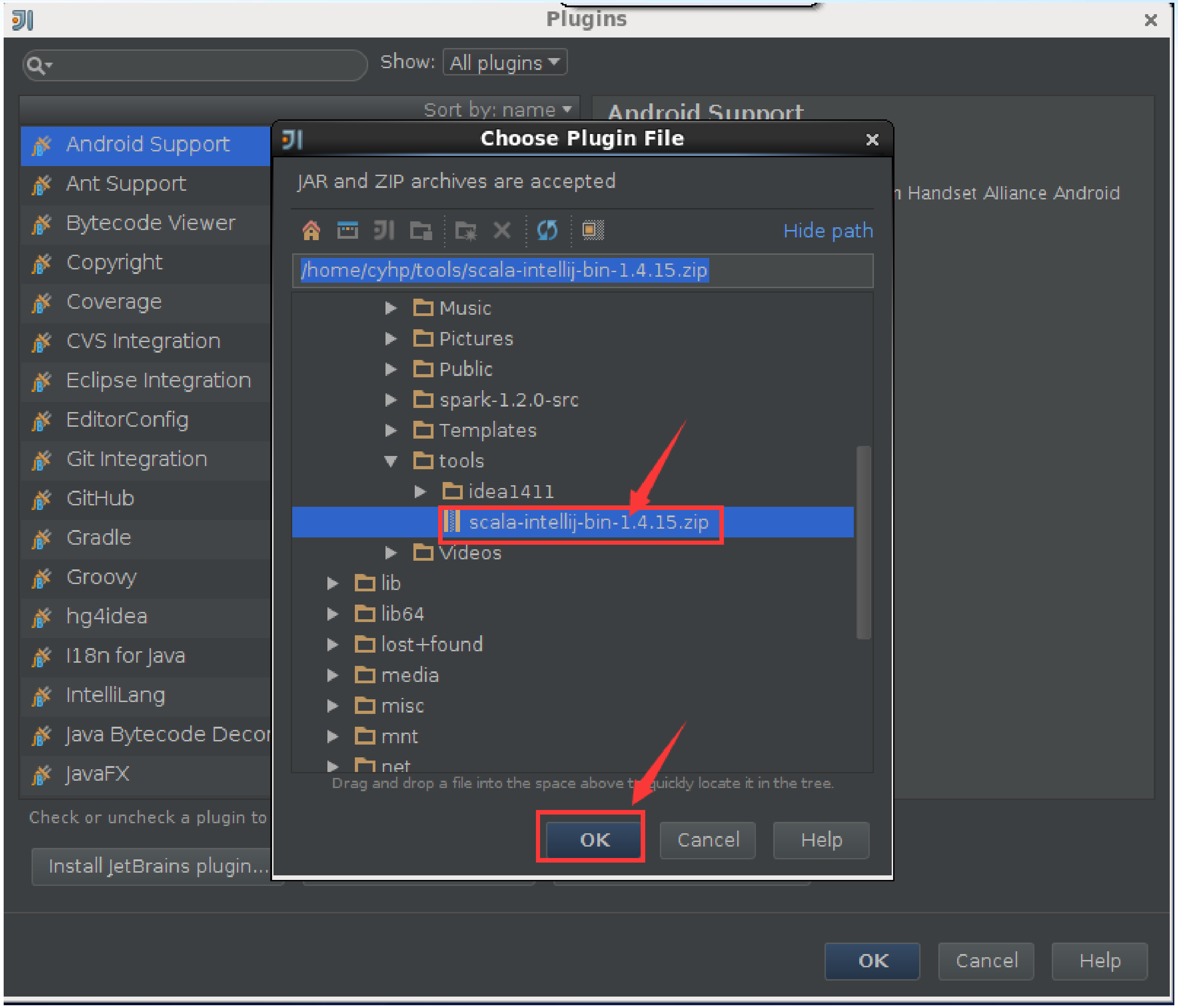
Android Support

Version: 10.1.1.0

Supports the development of Open Handset Alliance Android applications with IntelliJ IDEA.

Check or uncheck a plugin to enable or disable it.

安装Scala 插件



Plugins

Show: All plugins

Sort by: name

Android Support

Choose Plugin File

JAR and ZIP archives are accepted

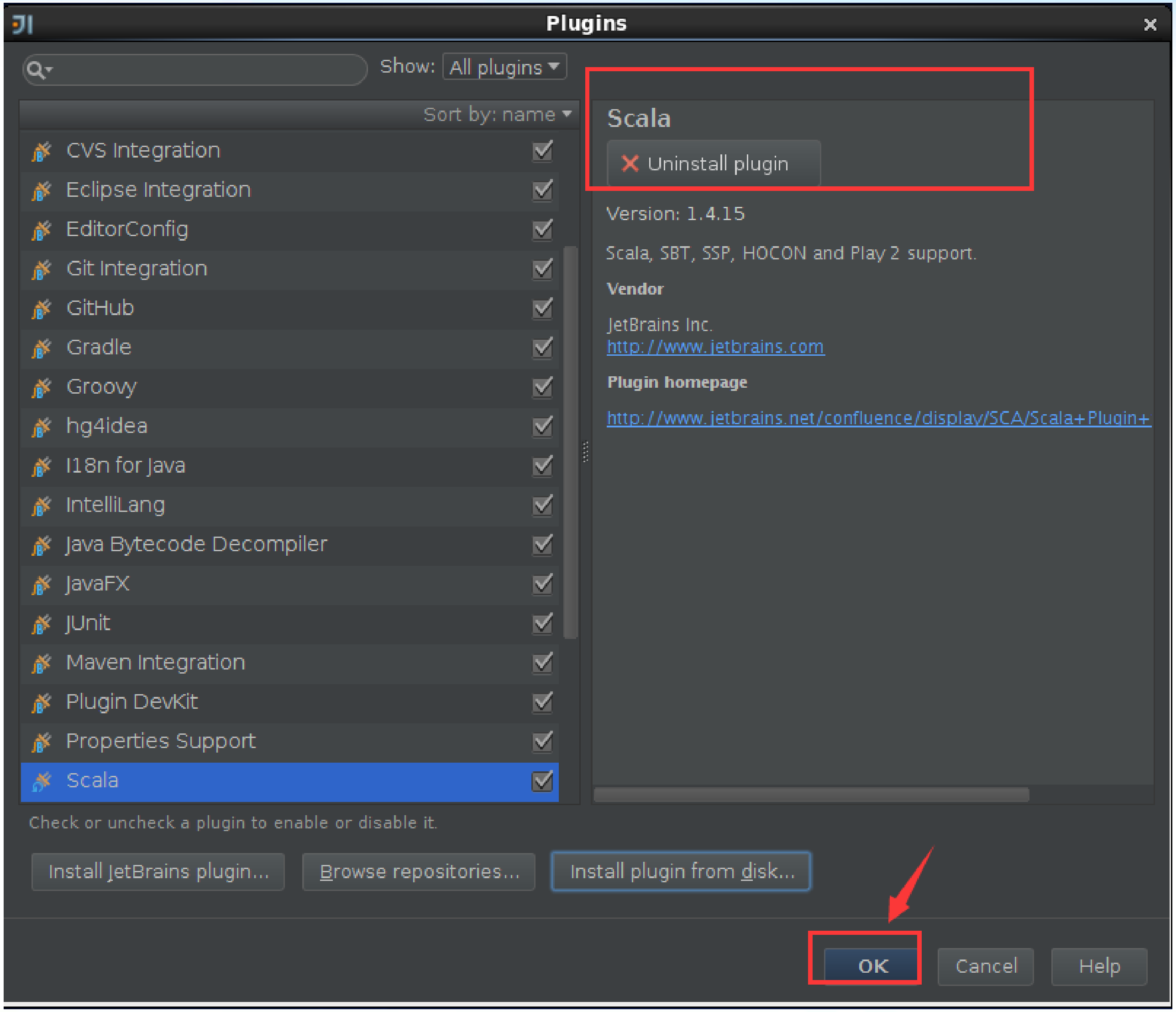
Hide path

/home/cyhp/tools/scala-intellij-bin-1.4.15.zip

- Music
- Pictures
- Public
- spark-1.2.0-src
- Templates
- tools
  - idea1411
  - scala-intellij-bin-1.4.15.zip
- Videos
- lib
- lib64
- lost+found
- media
- misc
- mnt
- net

Drag and drop a file into the space above to quickly locate it in the tree.

安装Scala插件



Plugins

Show: All plugins

Sort by: name

- CVS Integration ☑
- Eclipse Integration ☑
- EditorConfig ☑
- Git Integration ☑
- GitHub ☑
- Gradle ☑
- Groovy ☑
- hg4idea ☑
- I18n for Java ☑
- IntelliLang ☑
- Java Bytecode Decompiler ☑
- JavaFX ☑
- JUnit ☑
- Maven Integration ☑
- Plugin DevKit ☑
- Properties Support ☑
- Scala ☑

Scala

Uninstall plugin

Version: 1.4.15

Scala, SBT, SSP, HOCON and Play 2 support.

Vendor

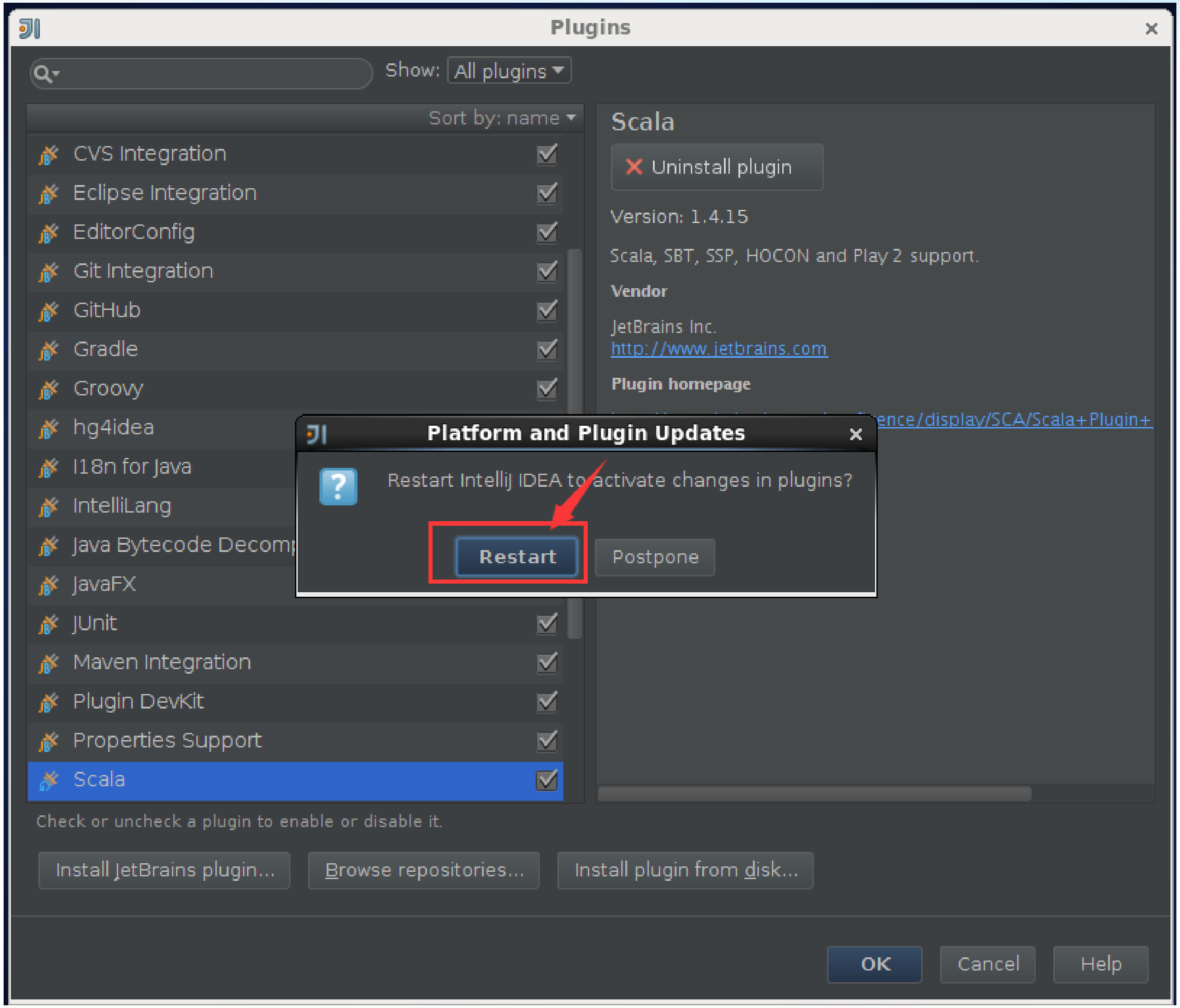
JetBrains Inc.
http://www.jetbrains.com

Plugin homepage

http://www.jetbrains.net/confluence/display/SCA/Scala+Plugin+

Check or uncheck a plugin to enable or disable it.

重启IDEA



Plugins

Show: All plugins

Sort by: name

- CVS Integration ☑
- Eclipse Integration ☑
- EditorConfig ☑
- Git Integration ☑
- GitHub ☑
- Gradle ☑
- Groovy ☑
- hg4idea
- I18n for Java
- IntelliLang
- Java Bytecode Decomp
- JavaFX
- JUnit ☑
- Maven Integration ☑
- Plugin DevKit ☑
- Properties Support ☑
- Scala ☑

Scala

✕ Uninstall plugin

Version: 1.4.15

Scala, SBT, SSP, HOCON and Play 2 support.

Vendor

JetBrains Inc.
http://www.jetbrains.com

Plugin homepage

ence/display/SCA/Scala+Plugin+

Platform and Plugin Updates

Restart IntelliJ IDEA to activate changes in plugins?

Check or uncheck a plugin to enable or disable it.

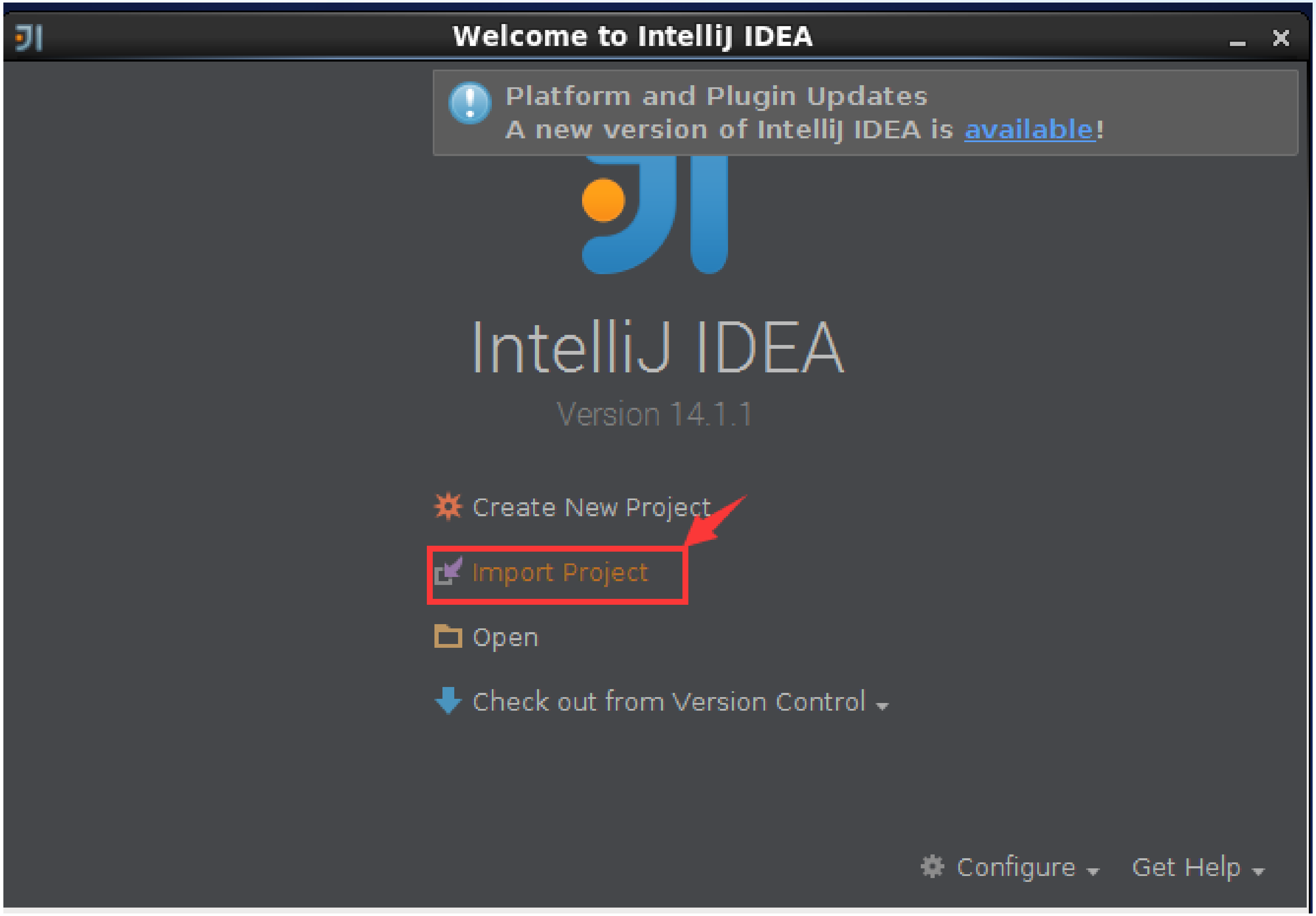
重启IDEA

零基础学习Spark 1.x应用开发系列课程

导入Spark源码

讲师-梦琪

导入Spark源码

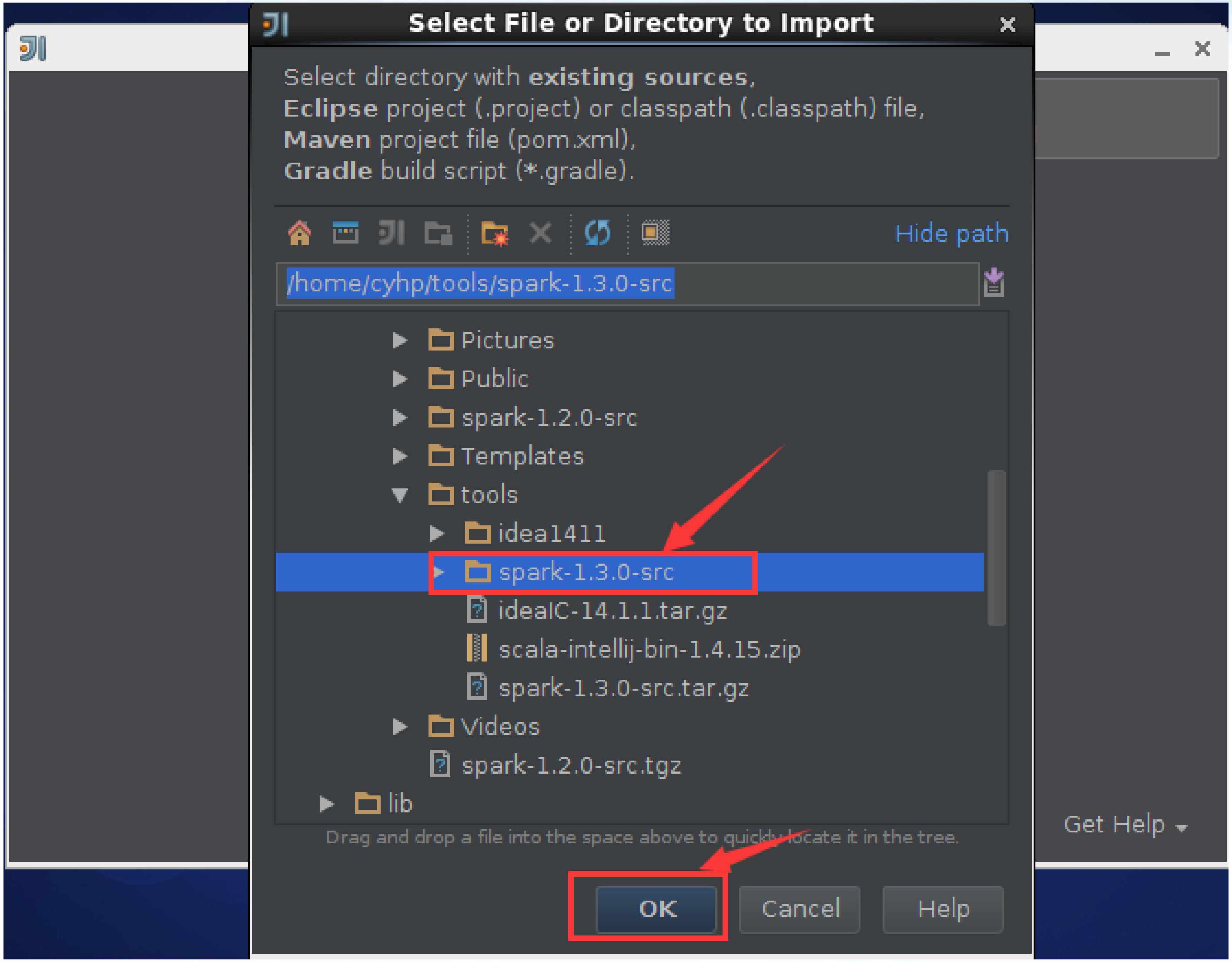


Welcome to IntelliJ IDEA

Platform and Plugin Updates
A new version of IntelliJ IDEA is available!

IntelliJ IDEA
Version 14.1.1

导入Spark源码



Select File or Directory to Import

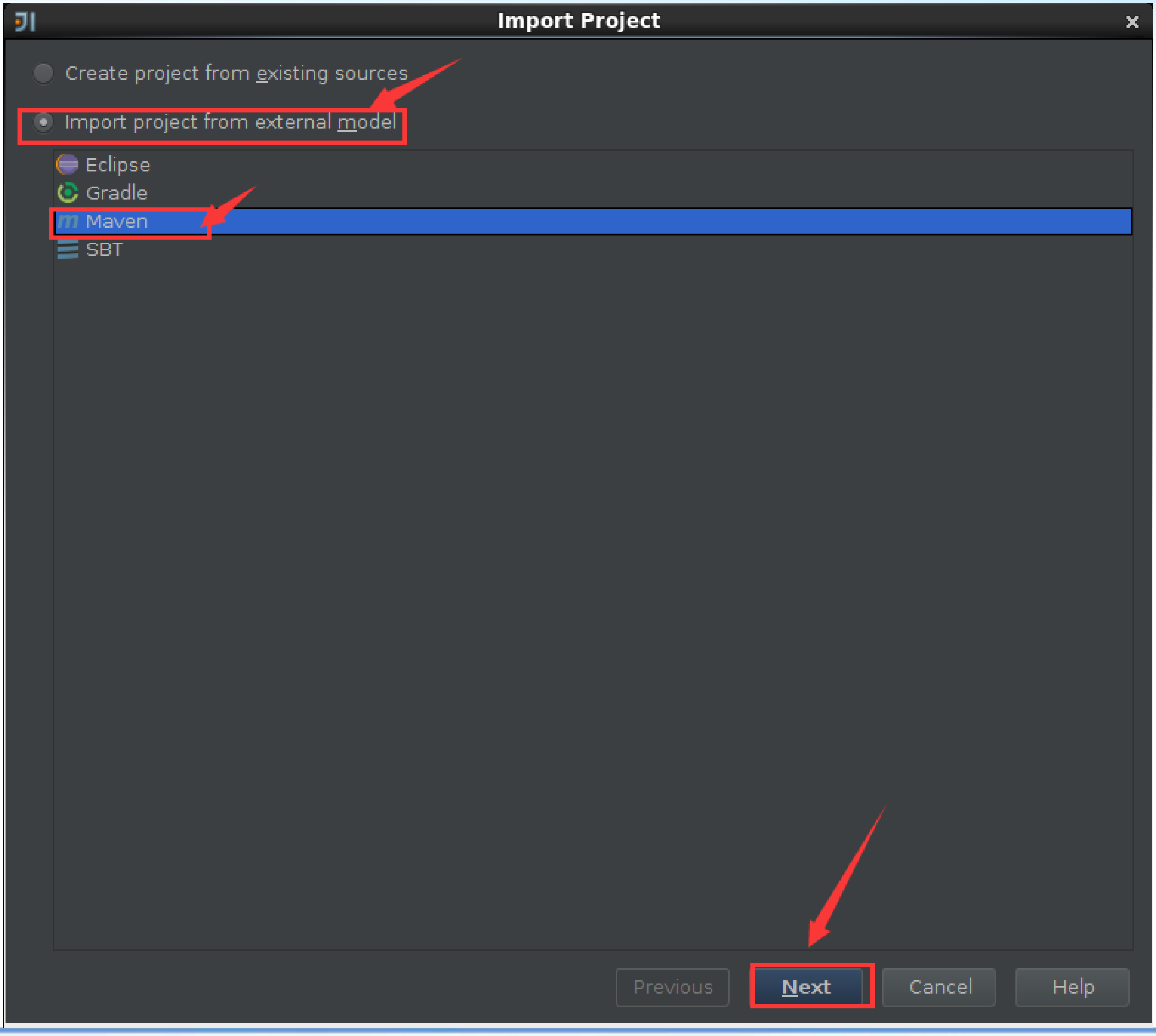
Select directory with **existing sources**,
Eclipse project (.project) or classpath (.classpath) file,
Maven project file (pom.xml),
Gradle build script (*.gradle).

```
/home/cyhp/tools/spark-1.3.0-src
```

- Pictures
- Public
- spark-1.2.0-src
- Templates
- tools
  - idea1411
  - spark-1.3.0-src
  - idealC-14.1.1.tar.gz
  - scala-intellij-bin-1.4.15.zip
  - spark-1.3.0-src.tar.gz
- Videos
- spark-1.2.0-src.tgz
- lib

Drag and drop a file into the space above to quickly locate it in the tree.

导入Spark源码



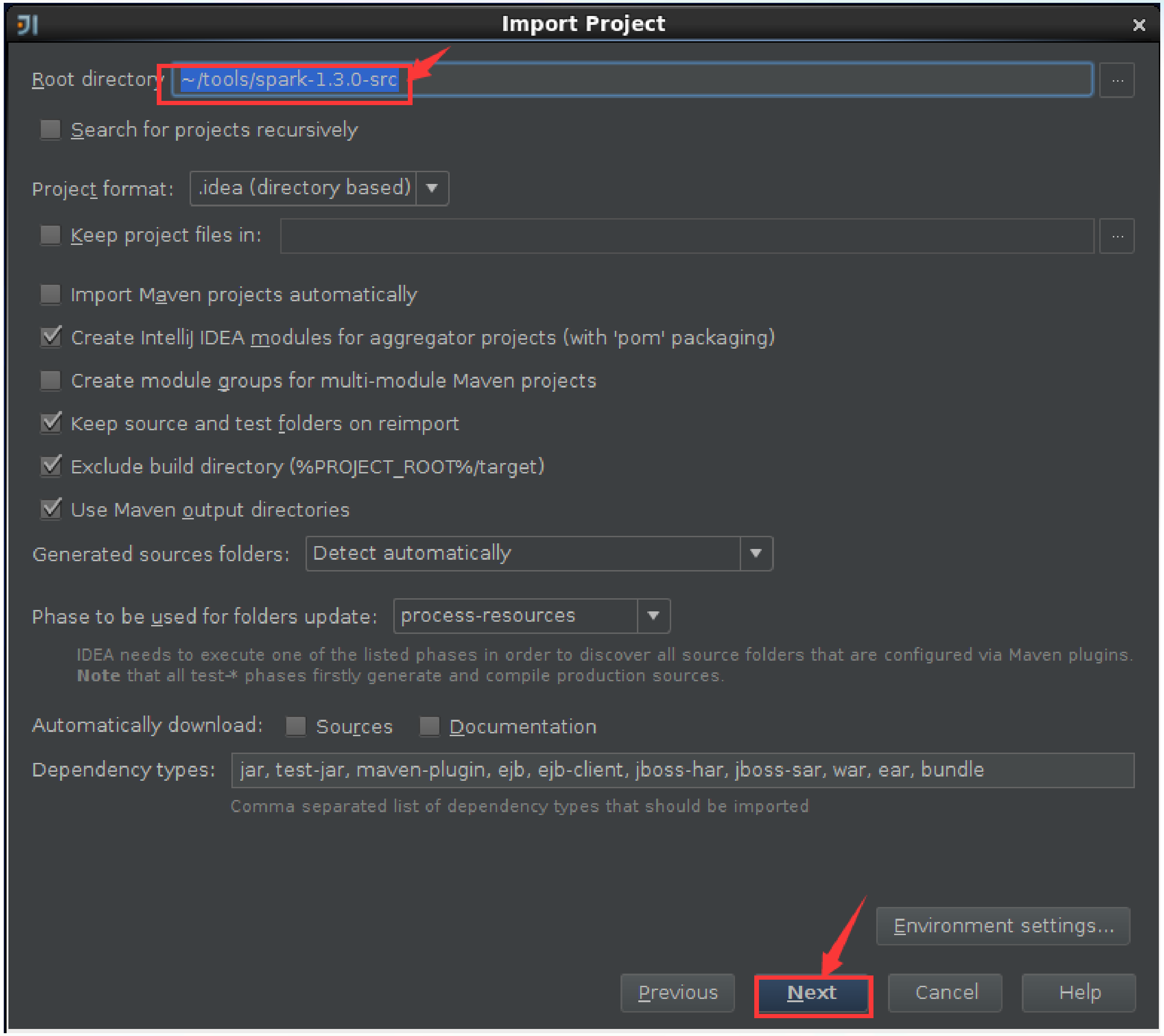
Import Project

○ Create project from existing sources

● Import project from external model

- Eclipse
- Gradle
- Maven
- SBT

导入Spark源码



Import Project

Root directory: ~/tools/spark-1.3.0-src

☐ Search for projects recursively

Project format: .idea (directory based)

☐ Keep project files in:

☐ Import Maven projects automatically

☑ Create IntelliJ IDEA modules for aggregator projects (with 'pom' packaging)

☐ Create module groups for multi-module Maven projects

☑ Keep source and test folders on reimport

☑ Exclude build directory (%PROJECT_ROOT%/target)

☑ Use Maven output directories

Generated sources folders: Detect automatically

Phase to be used for folders update: process-resources

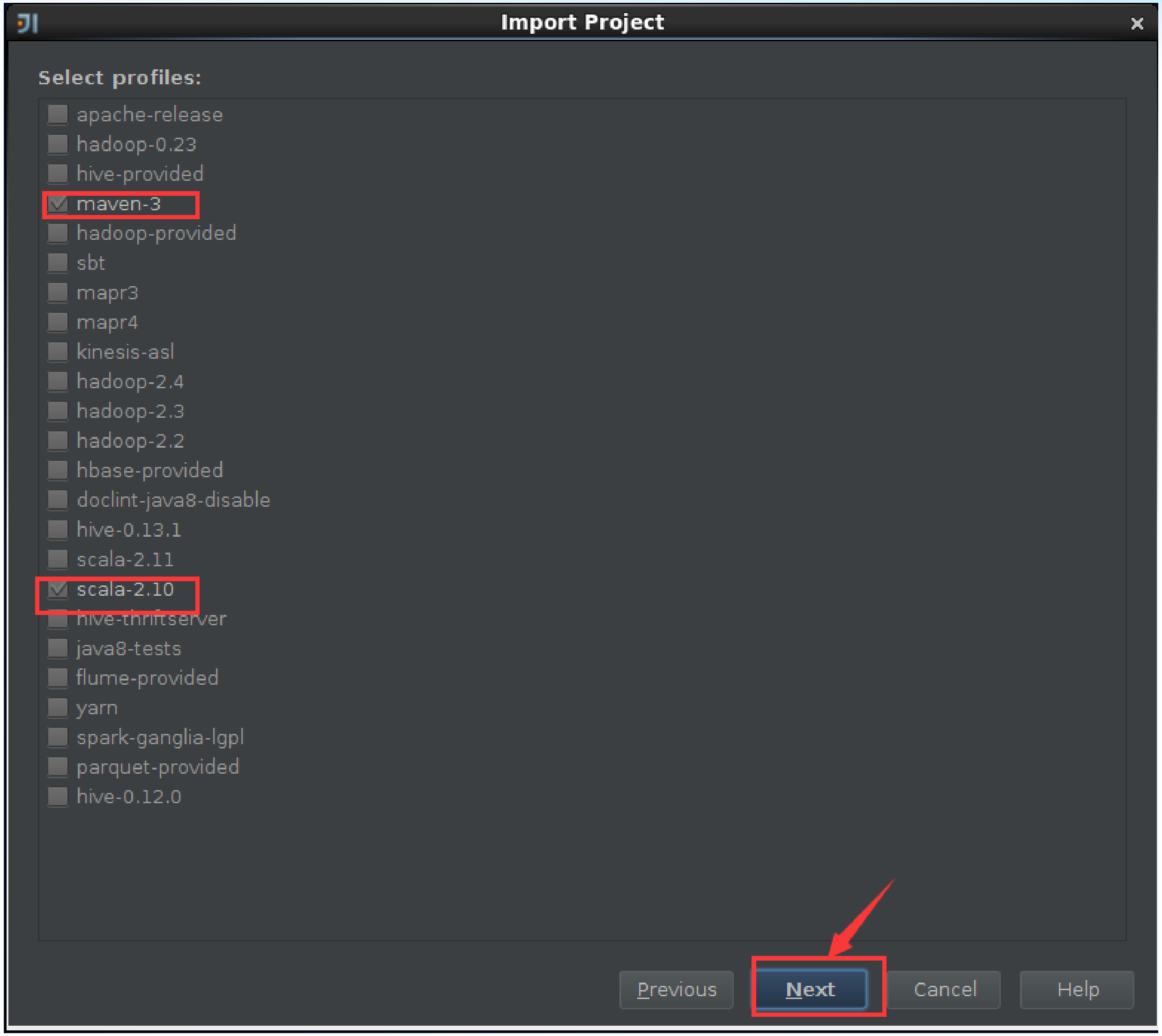
IDEA needs to execute one of the listed phases in order to discover all source folders that are configured via Maven plugins.
Note that all test-* phases firstly generate and compile production sources.

Automatically download: ☐ Sources ☐ Documentation

Dependency types: jar, test-jar, maven-plugin, ejb, ejb-client, jboss-har, jboss-sar, war, ear, bundle

Comma separated list of dependency types that should be imported

导入Spark源码

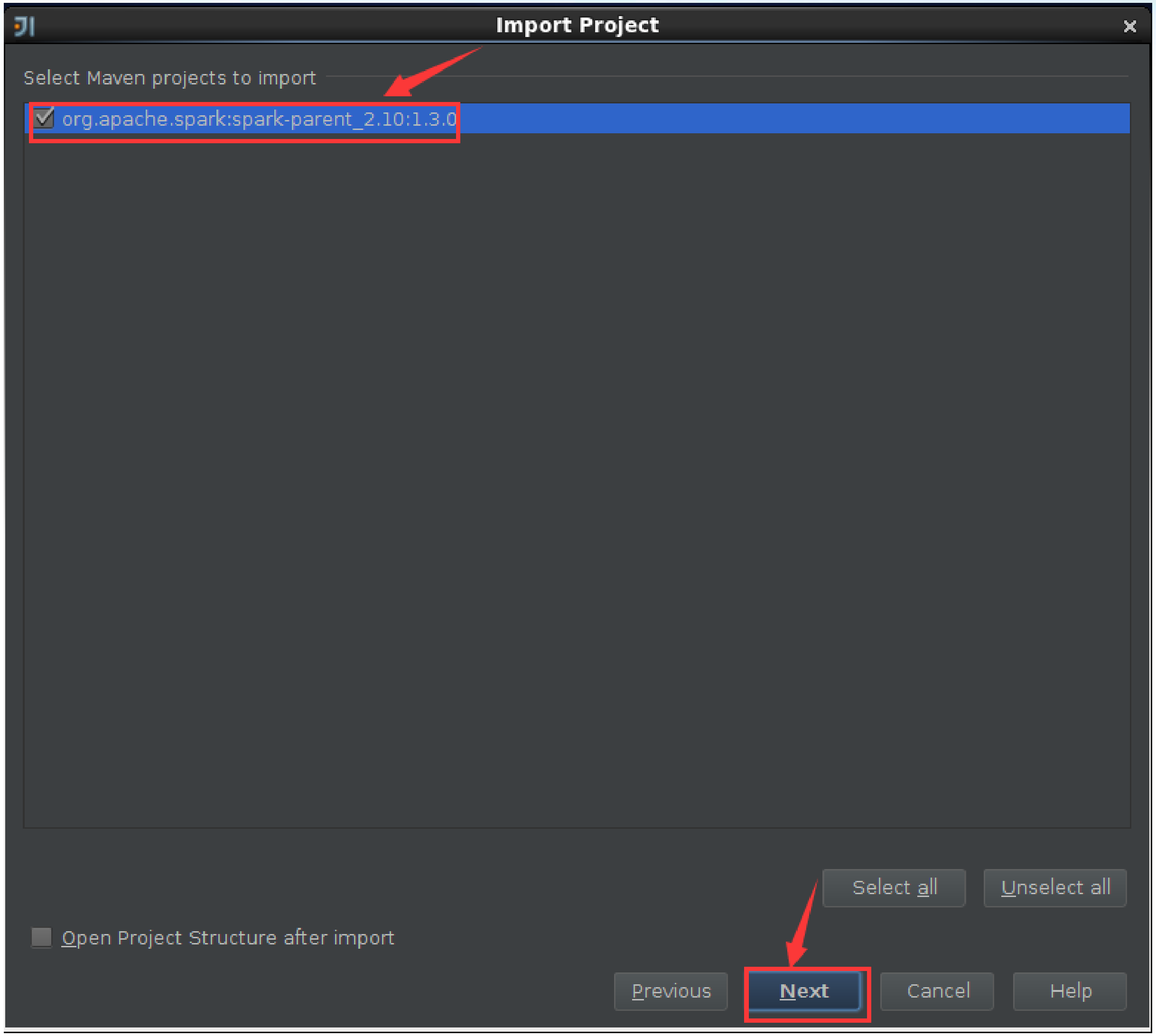


Import Project

Select profiles:

- □ apache-release
- □ hadoop-0.23
- □ hive-provided
- ☑ maven-3
- □ hadoop-provided
- □ sbt
- □ mapr3
- □ mapr4
- □ kinesis-asl
- □ hadoop-2.4
- □ hadoop-2.3
- □ hadoop-2.2
- □ hbase-provided
- □ doclint-java8-disable
- □ hive-0.13.1
- □ scala-2.11
- ☑ scala-2.10
- □ hive-thriftserver
- □ java8-tests
- □ flume-provided
- □ yarn
- □ spark-ganglia-lgpl
- □ parquet-provided
- □ hive-0.12.0

导入Spark源码



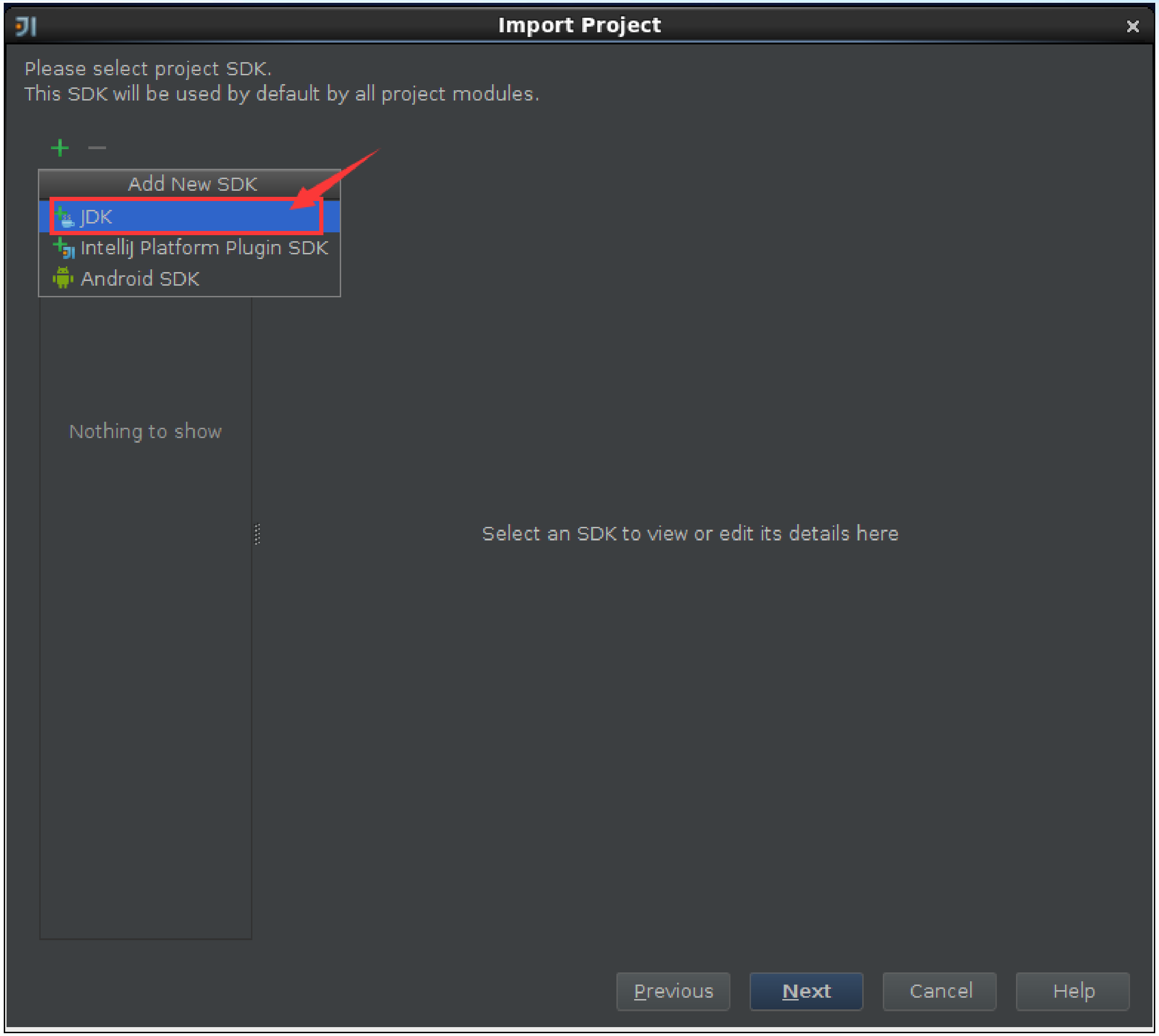
Import Project

Select Maven projects to import

☑ org.apache.spark:spark-parent_2.10:1.3.0

□ Open Project Structure after import

Add JDK SDK

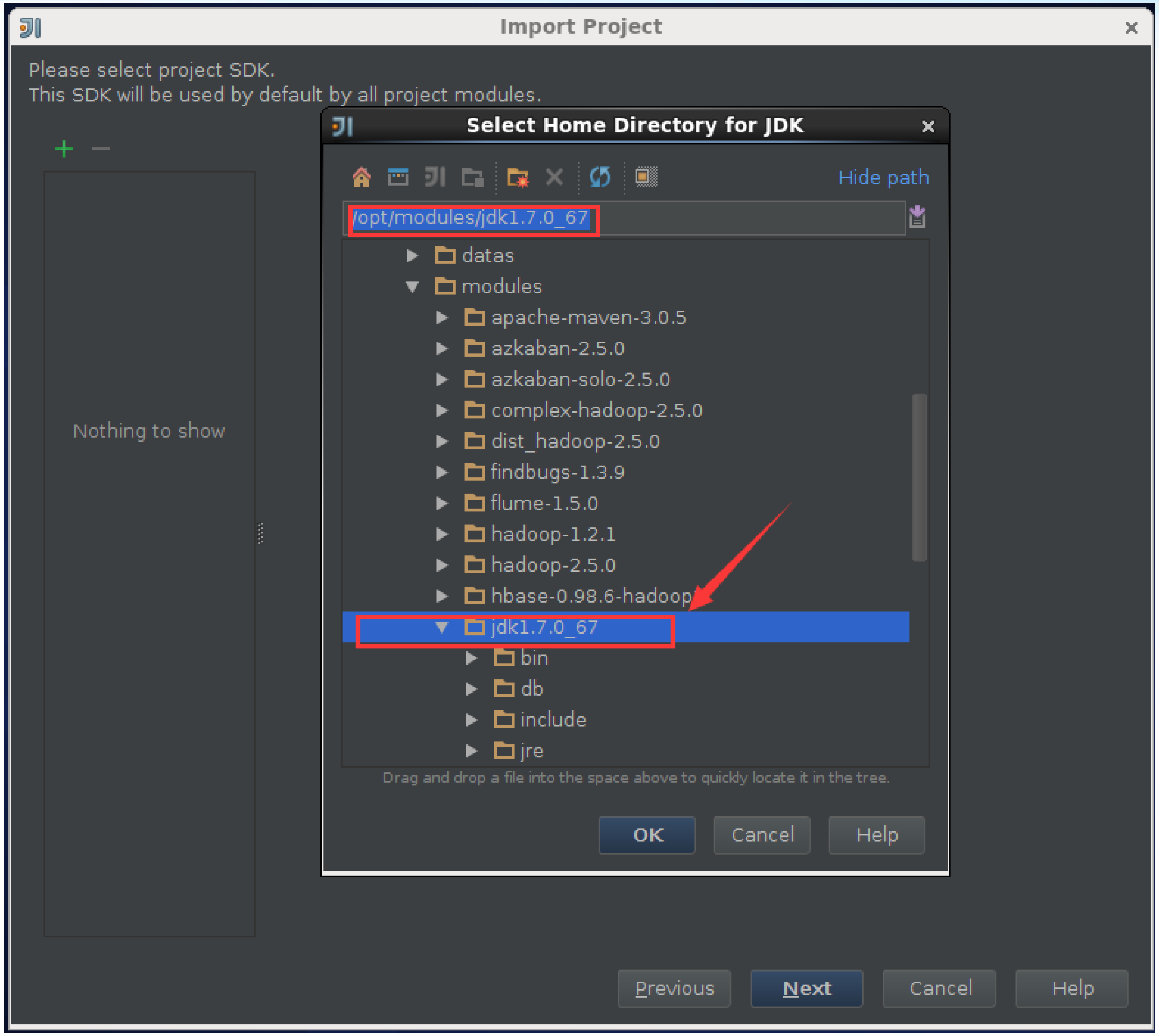


Please select project SDK.
This SDK will be used by default by all project modules.

Nothing to show

Select an SDK to view or edit its details here

Add JDK SDK



Import Project

Please select project SDK.
This SDK will be used by default by all project modules.

Nothing to show

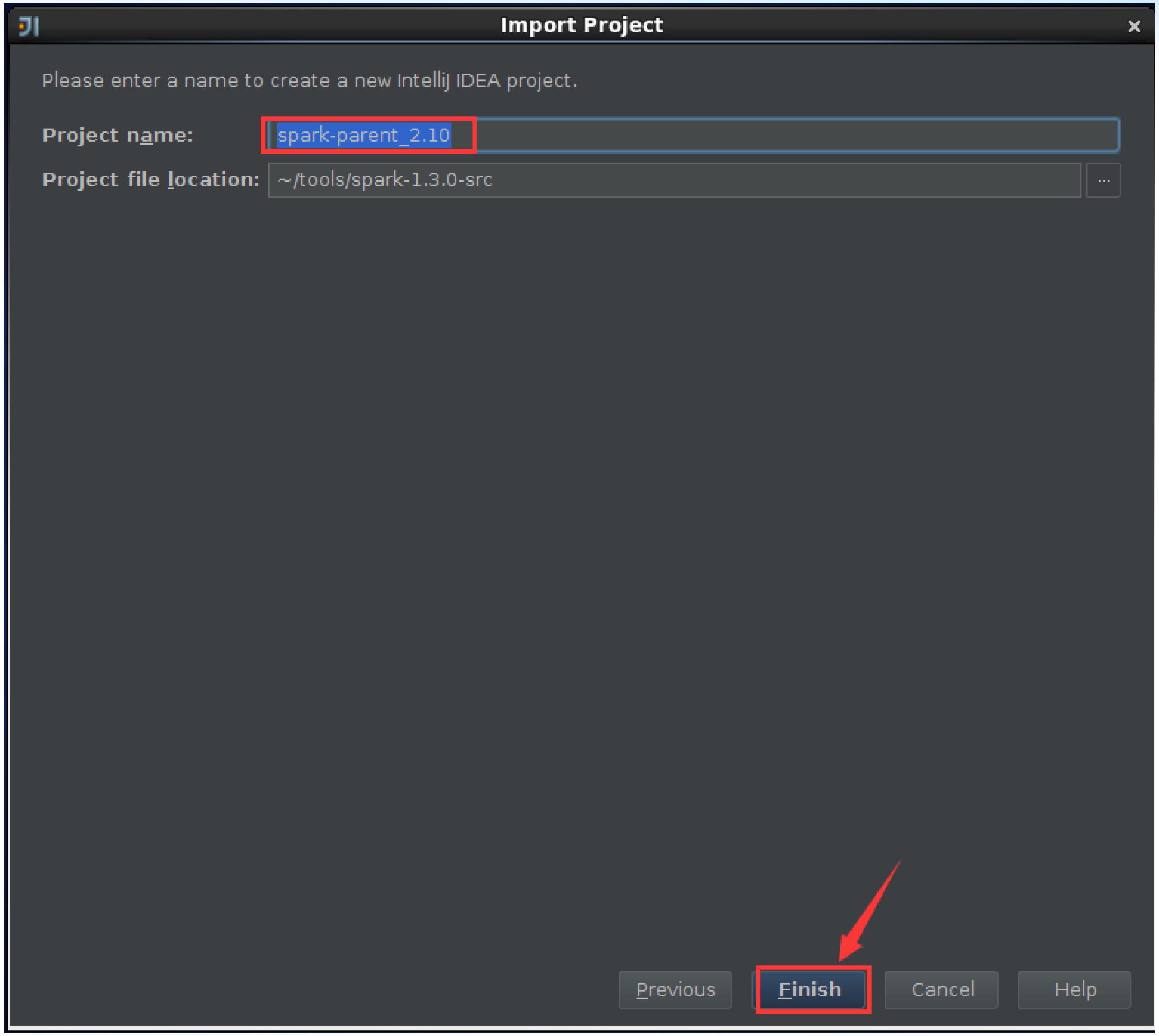
Select Home Directory for JDK

/opt/modules/jdk1.7.0_67

```
▶ datas
▼ modules
    ▶ apache-maven-3.0.5
    ▶ azkaban-2.5.0
    ▶ azkaban-solo-2.5.0
    ▶ complex-hadoop-2.5.0
    ▶ dist_hadoop-2.5.0
    ▶ findbugs-1.3.9
    ▶ flume-1.5.0
    ▶ hadoop-1.2.1
    ▶ hadoop-2.5.0
    ▶ hbase-0.98.6-hadoop
    ▼ jdk1.7.0_67
        ▶ bin
        ▶ db
        ▶ include
        ▶ jre
```

Drag and drop a file into the space above to quickly locate it in the tree.

导入Spark源码



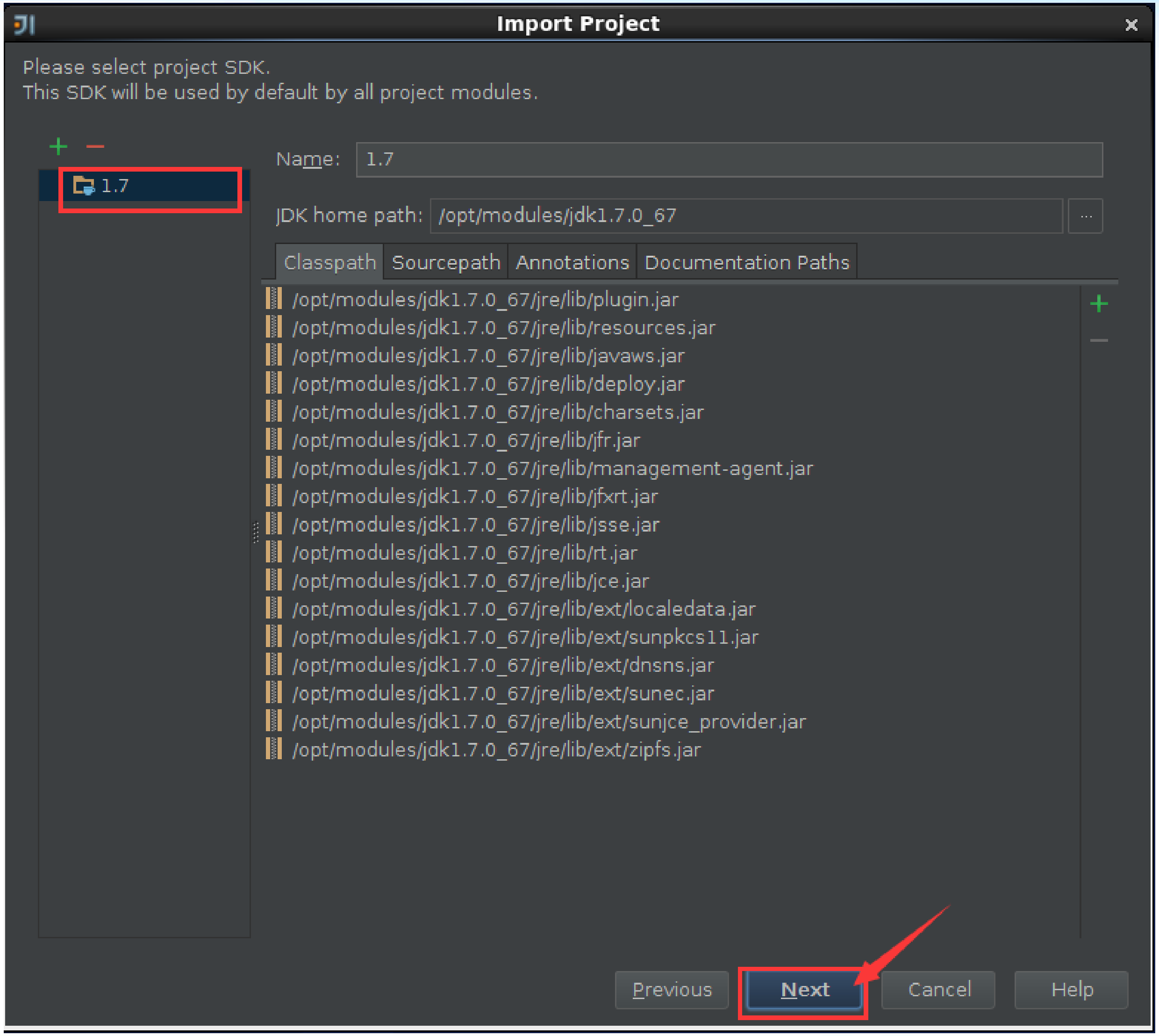
Import Project

Please enter a name to create a new IntelliJ IDEA project.

Project name: spark-parent_2.10

Project file location: ~/tools/spark-1.3.0-src

导入Spark源码



Import Project

Please select project SDK.
This SDK will be used by default by all project modules.

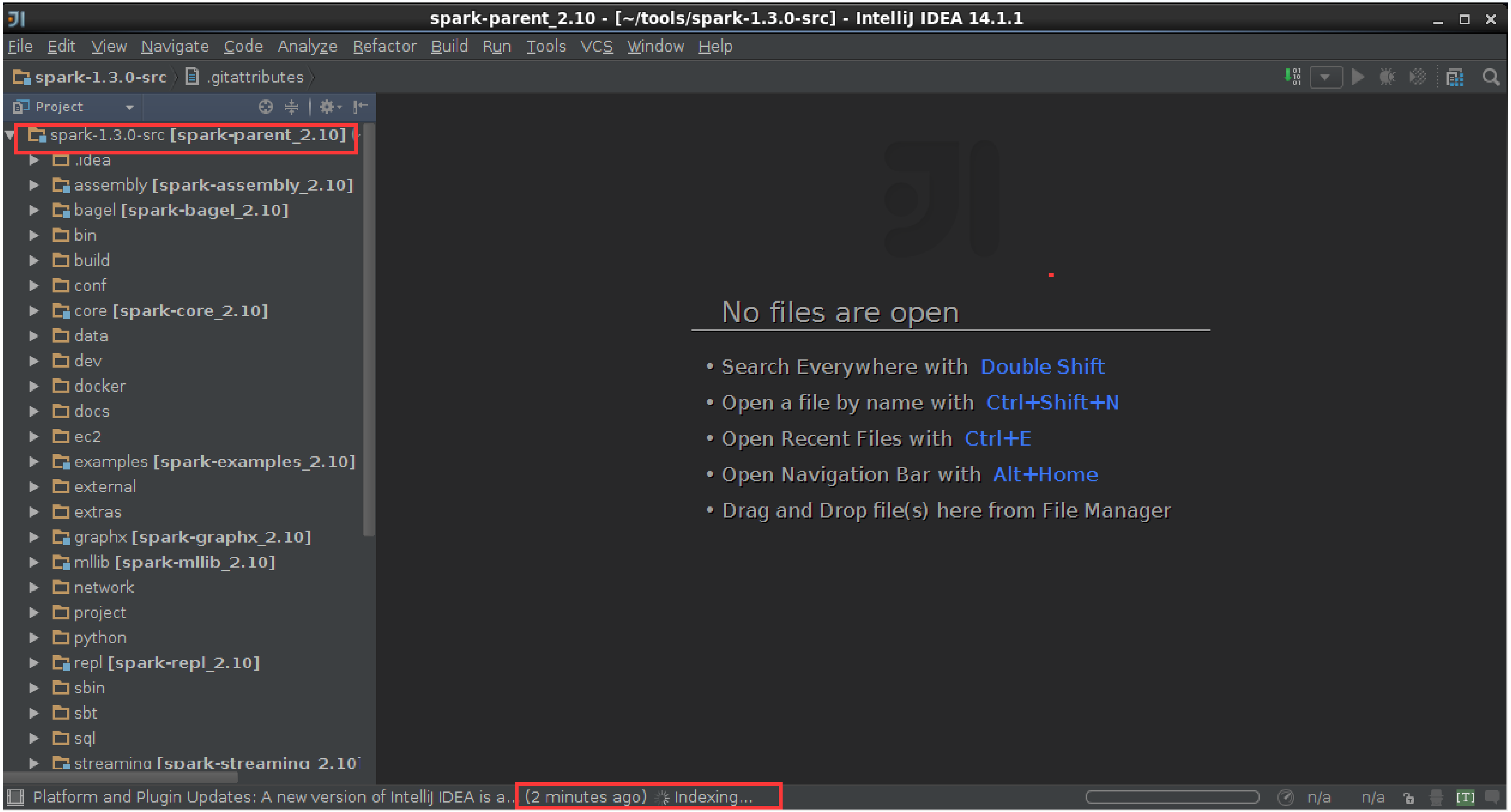
Name: 1.7

JDK home path: /opt/modules/jdk1.7.0_67

Classpath | Sourcepath | Annotations | Documentation Paths

```
/opt/modules/jdk1.7.0_67/jre/lib/plugin.jar
/opt/modules/jdk1.7.0_67/jre/lib/resources.jar
/opt/modules/jdk1.7.0_67/jre/lib/javaws.jar
/opt/modules/jdk1.7.0_67/jre/lib/deploy.jar
/opt/modules/jdk1.7.0_67/jre/lib/charsets.jar
/opt/modules/jdk1.7.0_67/jre/lib/jfr.jar
/opt/modules/jdk1.7.0_67/jre/lib/management-agent.jar
/opt/modules/jdk1.7.0_67/jre/lib/jfxrt.jar
/opt/modules/jdk1.7.0_67/jre/lib/jsse.jar
/opt/modules/jdk1.7.0_67/jre/lib/rt.jar
/opt/modules/jdk1.7.0_67/jre/lib/jce.jar
/opt/modules/jdk1.7.0_67/jre/lib/ext/localedata.jar
/opt/modules/jdk1.7.0_67/jre/lib/ext/sunpkcs11.jar
/opt/modules/jdk1.7.0_67/jre/lib/ext/dnsns.jar
/opt/modules/jdk1.7.0_67/jre/lib/ext/sunec.jar
/opt/modules/jdk1.7.0_67/jre/lib/ext/sunjce_provider.jar
/opt/modules/jdk1.7.0_67/jre/lib/ext/zipfs.jar
```

导入Spark源码



spark-parent_2.10 - [~/tools/spark-1.3.0-src] - IntelliJ IDEA 14.1.1

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
spark-1.3.0-src .gitattributes
Project

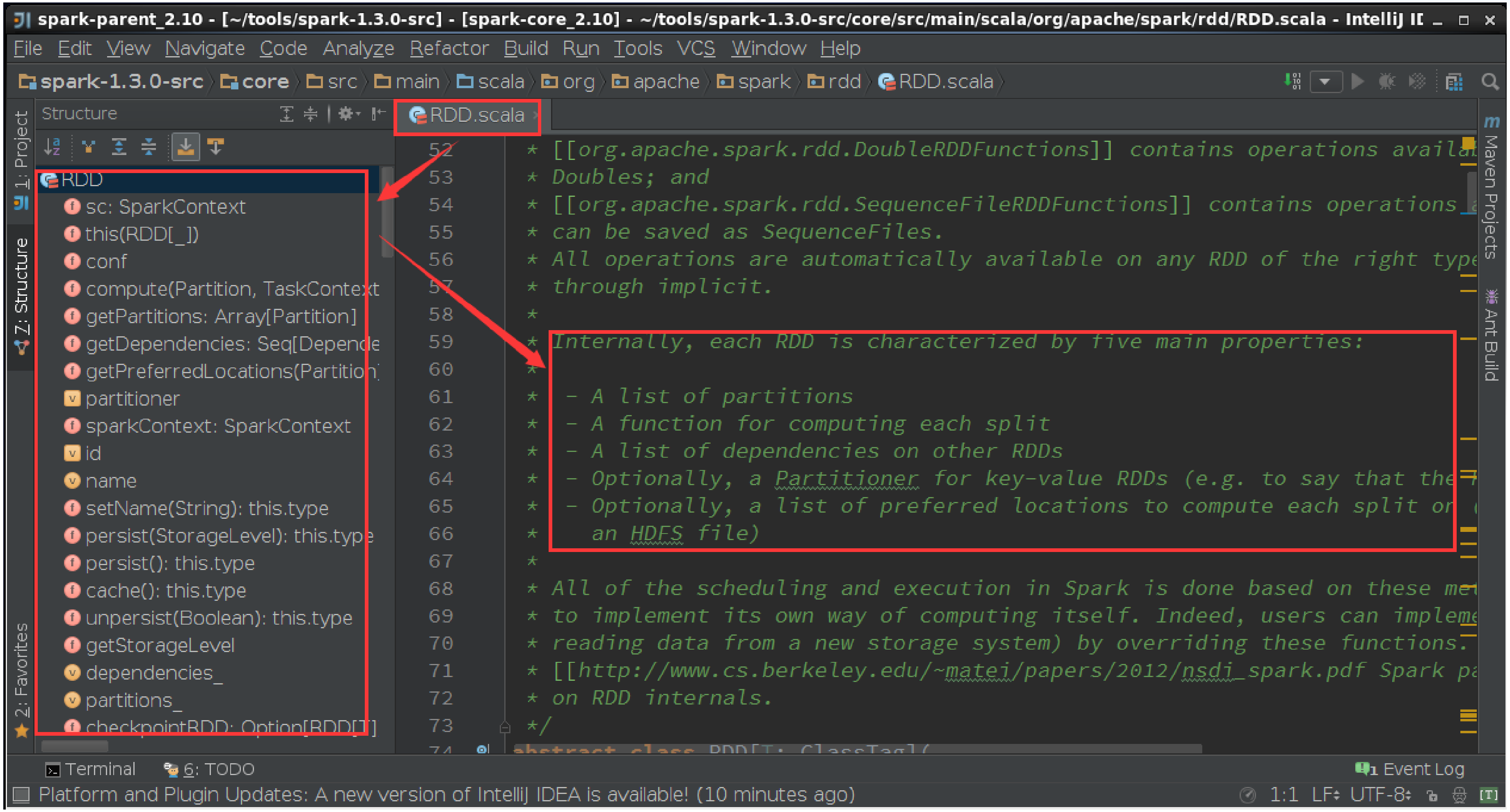
```
spark-1.3.0-src [spark-parent_2.10]
  .idea
  assembly [spark-assembly_2.10]
  bagel [spark-bagel_2.10]
  bin
  build
  conf
  core [spark-core_2.10]
  data
  dev
  docker
  docs
  ec2
  examples [spark-examples_2.10]
  external
  extras
  graphx [spark-graphx_2.10]
  mllib [spark-mllib_2.10]
  network
  project
  python
  repl [spark-repl_2.10]
  sbin
  sbt
  sql
  streaming [spark-streaming_2.10]
```

No files are open

- Search Everywhere with Double Shift
- Open a file by name with Ctrl+Shift+N
- Open Recent Files with Ctrl+E
- Open Navigation Bar with Alt+Home
- Drag and Drop file(s) here from File Manager

Platform and Plugin Updates: A new version of IntelliJ IDEA is a... (2 minutes ago) Indexing...

导入Spark源码



spark-parent_2.10 - [~/tools/spark-1.3.0-src] - [spark-core_2.10] - ~/tools/spark-1.3.0-src/core/src/main/scala/org/apache/spark/rdd/RDD.scala - IntelliJ IE
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
spark-1.3.0-src core src main scala org apache spark rdd RDD.scala
Structure
RDD.scala
1: Project
Z: Structure
2: Favorites
Maven Projects
Ant Build
Terminal
6: TODO
Event Log
1:1 LF UTF-8

```
RDD
  sc: SparkContext
  this(RDD[_])
  conf
  compute(Partition, TaskContext
  getPartitions: Array[Partition]
  getDependencies: Seq[Depende
  getPreferredLocations(Partition
  partitioner
  sparkContext: SparkContext
  id
  name
  setName(String): this.type
  persist(StorageLevel): this.type
  persist(): this.type
  cache(): this.type
  unpersist(Boolean): this.type
  getStorageLevel
  dependencies_
  partitions_
  checkpointRDD: Option[RDD[T]]
```

```
 * [[org.apache.spark.rdd.DoubleRDDFunctions]] contains operations availa
 * Doubles; and
 * [[org.apache.spark.rdd.SequenceFileRDDFunctions]] contains operations
 * can be saved as SequenceFiles.
 * All operations are automatically available on any RDD of the right type
 * through implicit.
 *
 * Internally, each RDD is characterized by five main properties:
 *
 *  - A list of partitions
 *  - A function for computing each split
 *  - A list of dependencies on other RDDs
 *  - Optionally, a Partitioner for key-value RDDs (e.g. to say that the
 *  - Optionally, a list of preferred locations to compute each split on
 *    an HDFS file)
 *
 * All of the scheduling and execution in Spark is done based on these me
 * to implement its own way of computing itself. Indeed, users can impleme
 * reading data from a new storage system) by overriding these functions.
 * [[http://www.cs.berkeley.edu/~matei/papers/2012/nsdi_spark.pdf Spark pa
 * on RDD internals.
 */
abstract class RDD[T: ClassTag](
```

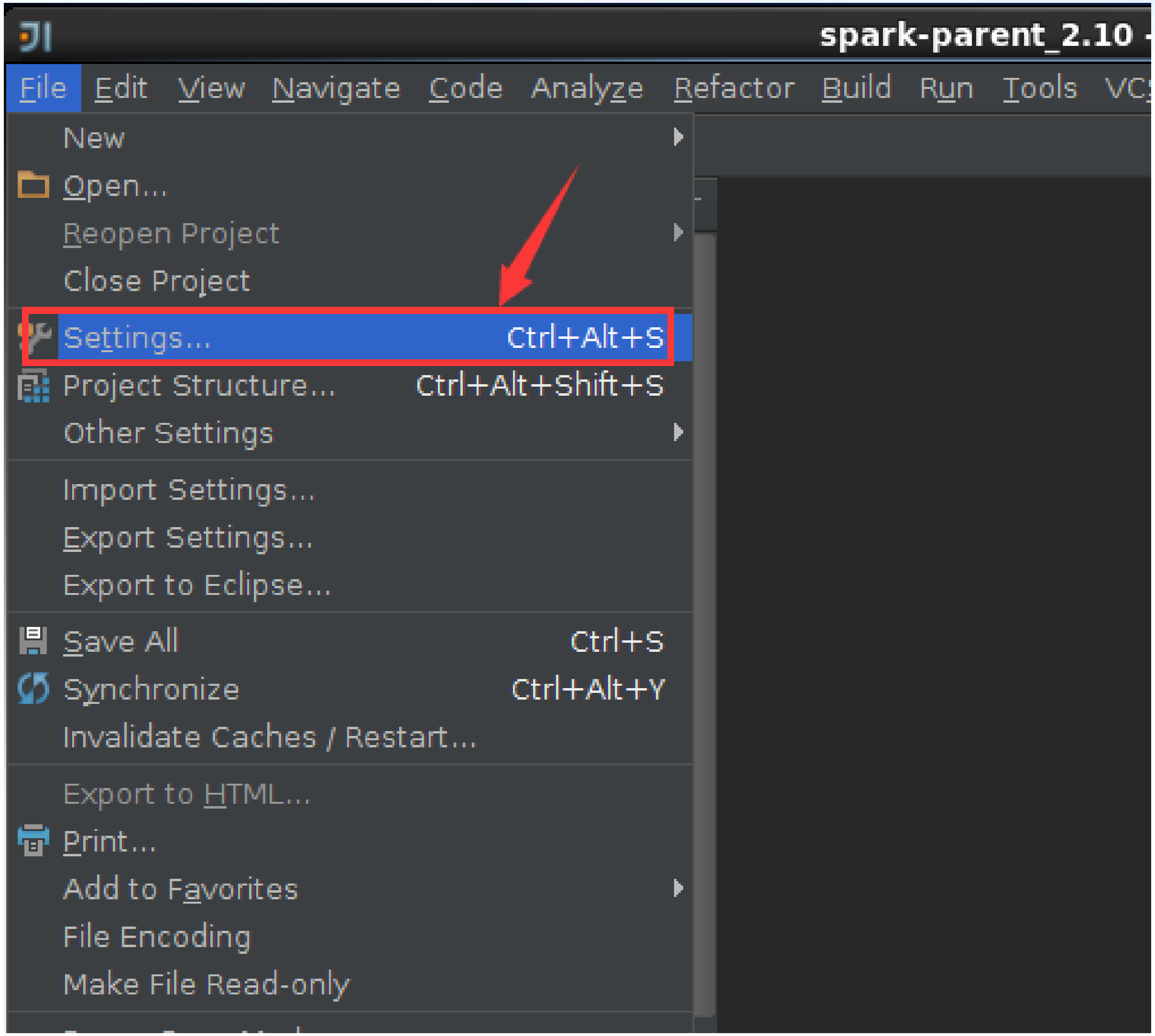
Platform and Plugin Updates: A new version of IntelliJ IDEA is available! (10 minutes ago)

零基础学习Spark 1.x应用开发系列课程

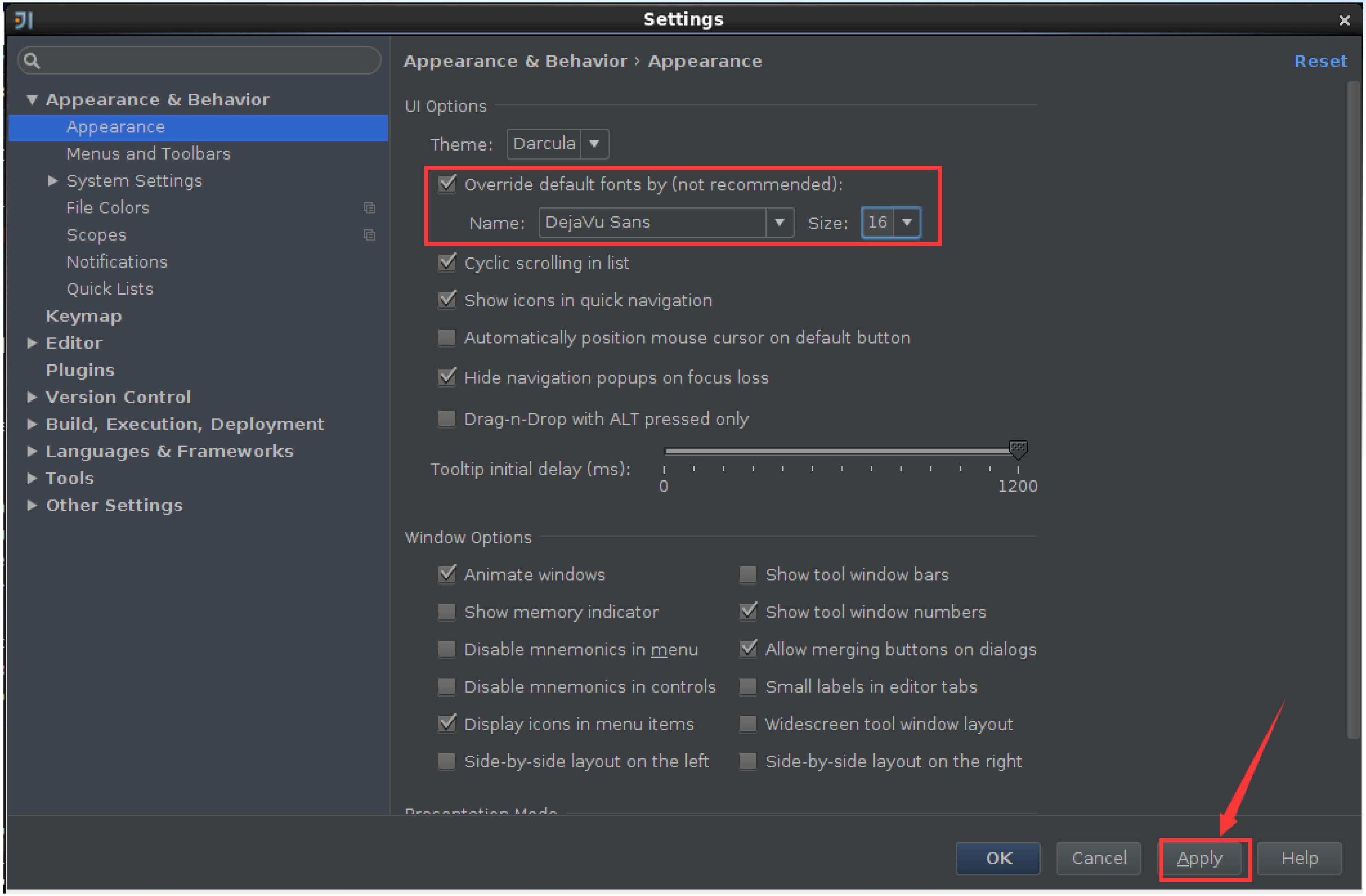
IDEA基本配置

讲师-梦琪

IDEA基本配置


spark-parent_2.10
File
Edit
View
Navigate
Code
Analyze
Refactor
Build
Run
Tools
VC
New
Open...
Reopen Project
Close Project
Settings... Ctrl+Alt+S
Project Structure... Ctrl+Alt+Shift+S
Other Settings
Import Settings...
Export Settings...
Export to Eclipse...
Save All Ctrl+S
Synchronize Ctrl+Alt+Y
Invalidate Caches / Restart...
Export to HTML...
Print...
Add to Favorites
File Encoding
Make File Read-only

IDEA 基本配置



Appearance & Behavior › Appearance

UI Options

Theme: Darcula

☑ Override default fonts by (not recommended):

Name: DejaVu Sans Size: 16

☑ Cyclic scrolling in list

☑ Show icons in quick navigation

☐ Automatically position mouse cursor on default button

☑ Hide navigation popups on focus loss

☐ Drag-n-Drop with ALT pressed only

Tooltip initial delay (ms): 0 – 1200

Window Options

☑ Animate windows

☐ Show tool window bars

☐ Show memory indicator

☑ Show tool window numbers

☐ Disable mnemonics in menu

☑ Allow merging buttons on dialogs

☐ Disable mnemonics in controls

☐ Small labels in editor tabs

☑ Display icons in menu items

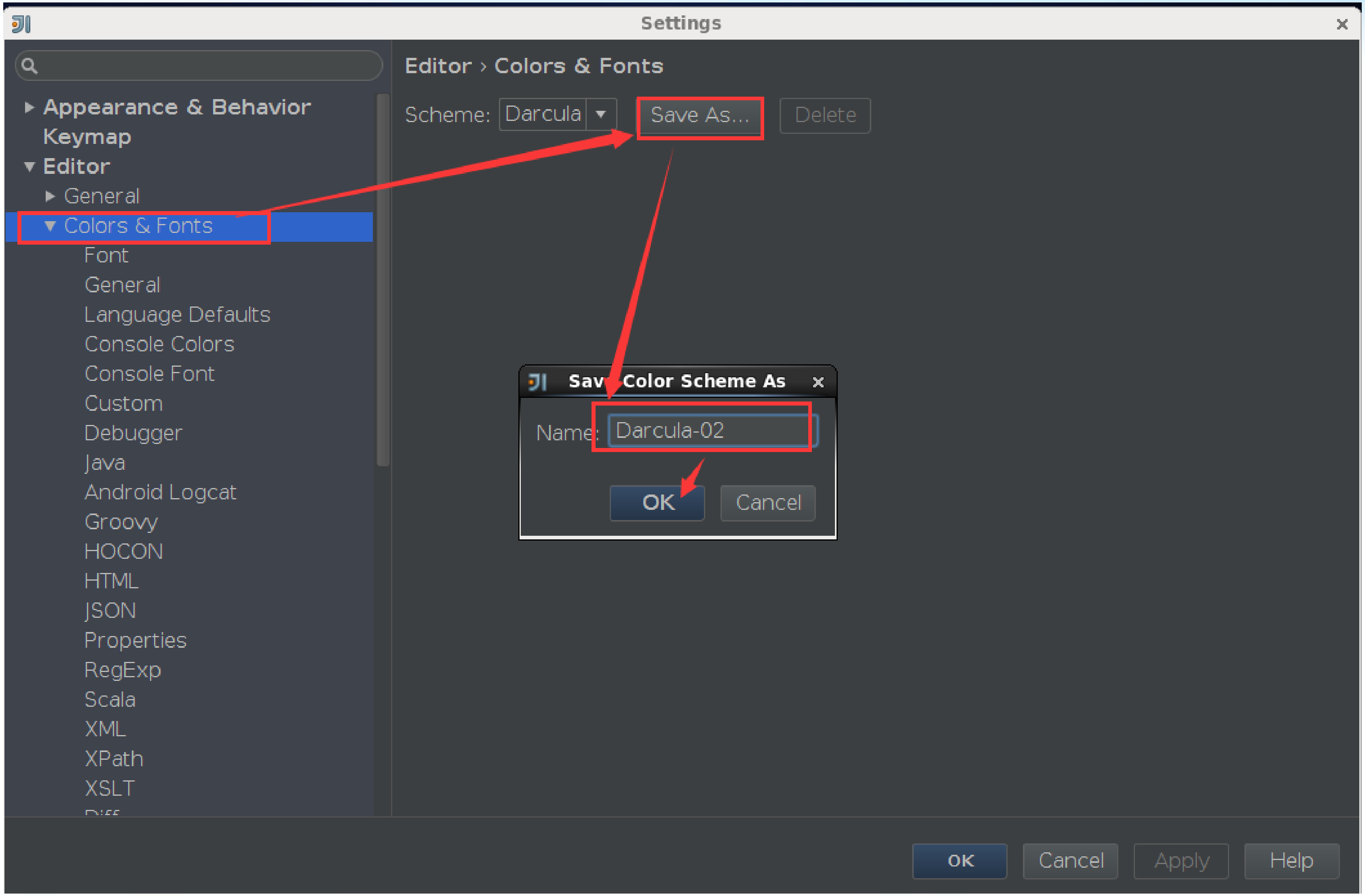
☐ Widescreen tool window layout

☐ Side-by-side layout on the left

☐ Side-by-side layout on the right

Presentation Mode

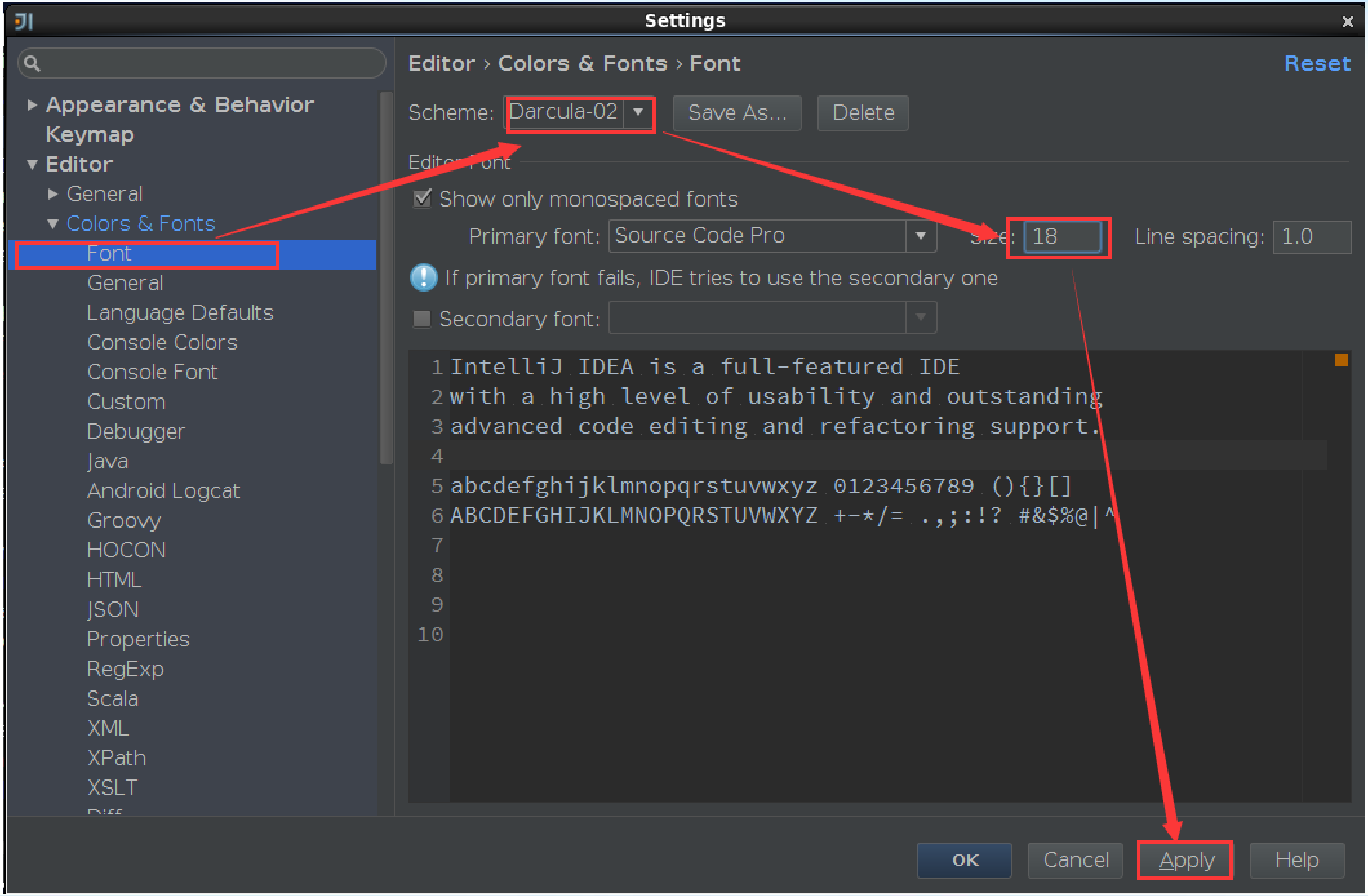
IDEA基本配置



Scheme: Darcula

Name: Darcula-02

IDEA 基本配置



Editor › Colors & Fonts › Font

Scheme: Darcula-02

Editor Font

☑ Show only monospaced fonts

Primary font: Source Code Pro Size: 18 Line spacing: 1.0

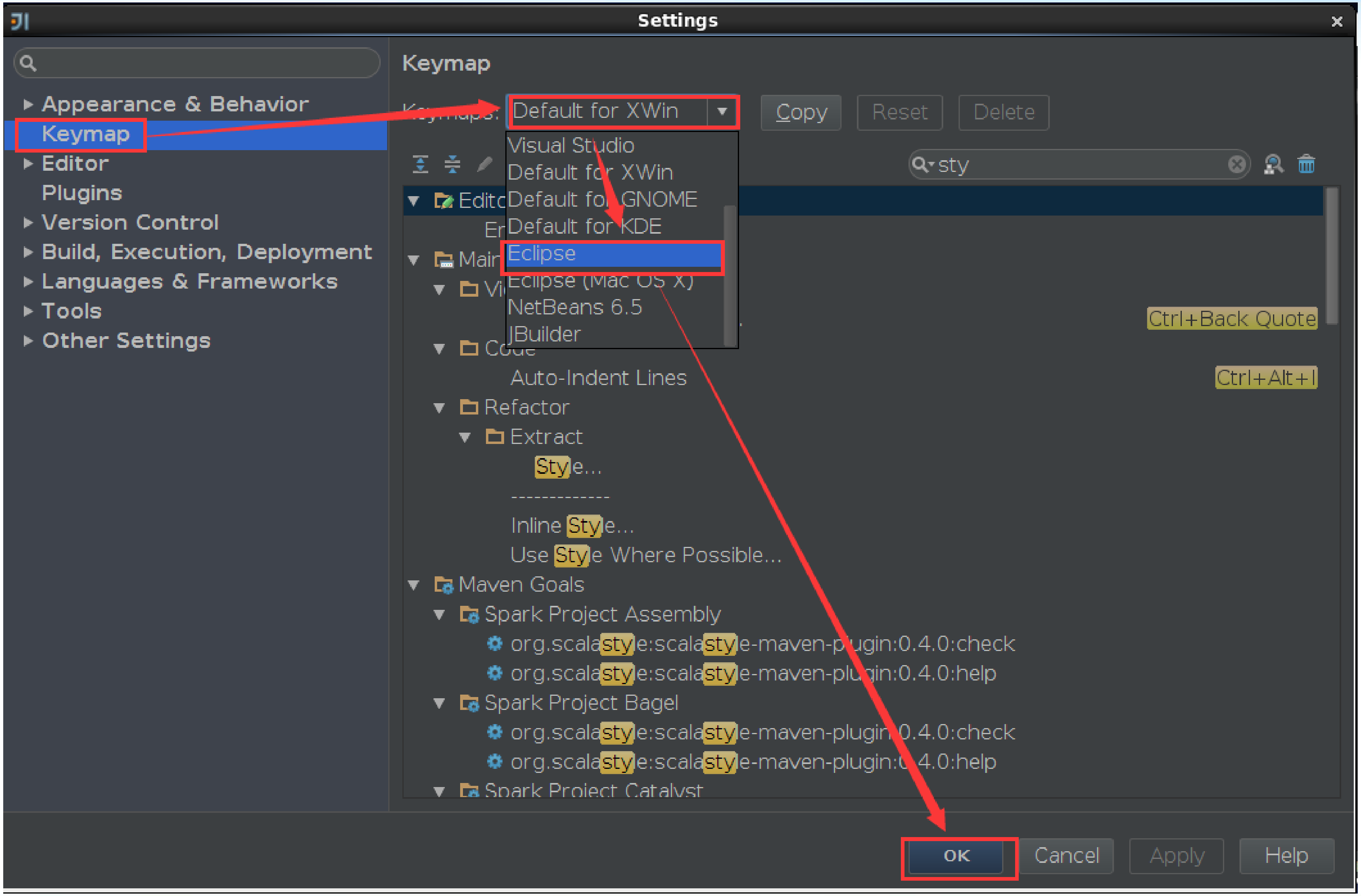
If primary font fails, IDE tries to use the secondary one

□ Secondary font:

```
IntelliJ IDEA is a full-featured IDE
with a high level of usability and outstanding
advanced code editing and refactoring support.

abcdefghijklmnopqrstuvwxyz 0123456789 (){}[]
ABCDEFGHIJKLMNOPQRSTUVWXYZ +-*/= .,;:!? #&$%@|^
```

IDEA 基本配置



Settings

- Appearance & Behavior
- Keymap
- Editor
- Plugins
- Version Control
- Build, Execution, Deployment
- Languages & Frameworks
- Tools
- Other Settings

Keymap

Keymaps: Default for XWin ▾ Copy Reset Delete

Visual Studio
Default for XWin
Default for GNOME
Default for KDE
Eclipse
Eclipse (Mac OS X)
NetBeans 6.5
JBuilder

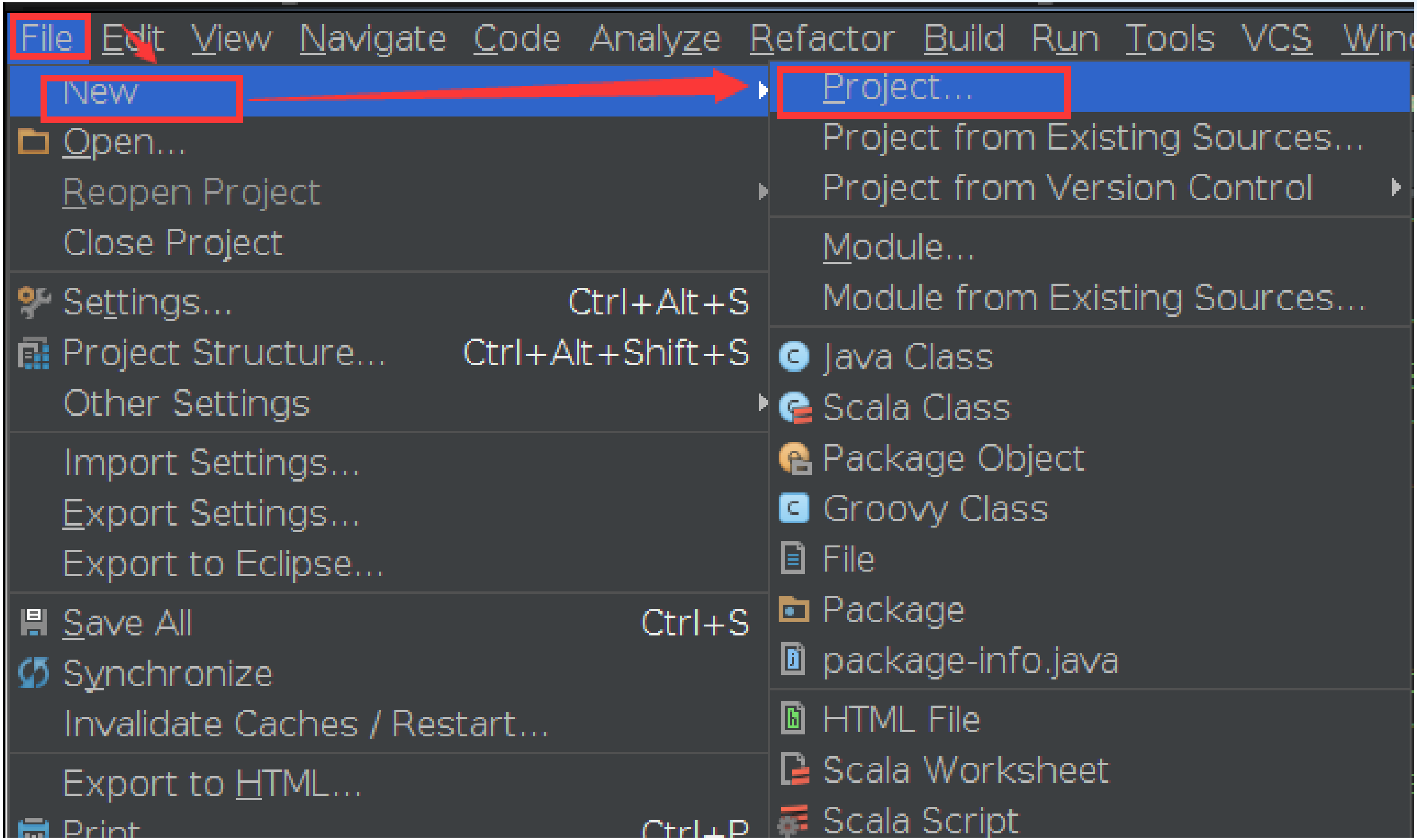
```
▼ Edito
    Er
▼ Main
  ▼ Vi                                              Ctrl+Back Quote
  ▼ Code
      Auto-Indent Lines                             Ctrl+Alt+I
  ▼ Refactor
    ▼ Extract
        Style...
      -------------
      Inline Style...
      Use Style Where Possible...
▼ Maven Goals
  ▼ Spark Project Assembly
      org.scalastyle:scalastyle-maven-plugin:0.4.0:check
      org.scalastyle:scalastyle-maven-plugin:0.4.0:help
  ▼ Spark Project Bagel
      org.scalastyle:scalastyle-maven-plugin:0.4.0:check
      org.scalastyle:scalastyle-maven-plugin:0.4.0:help
  ▼ Spark Project Catalyst
```

零基础学习Spark 1.x应用开发系列课程

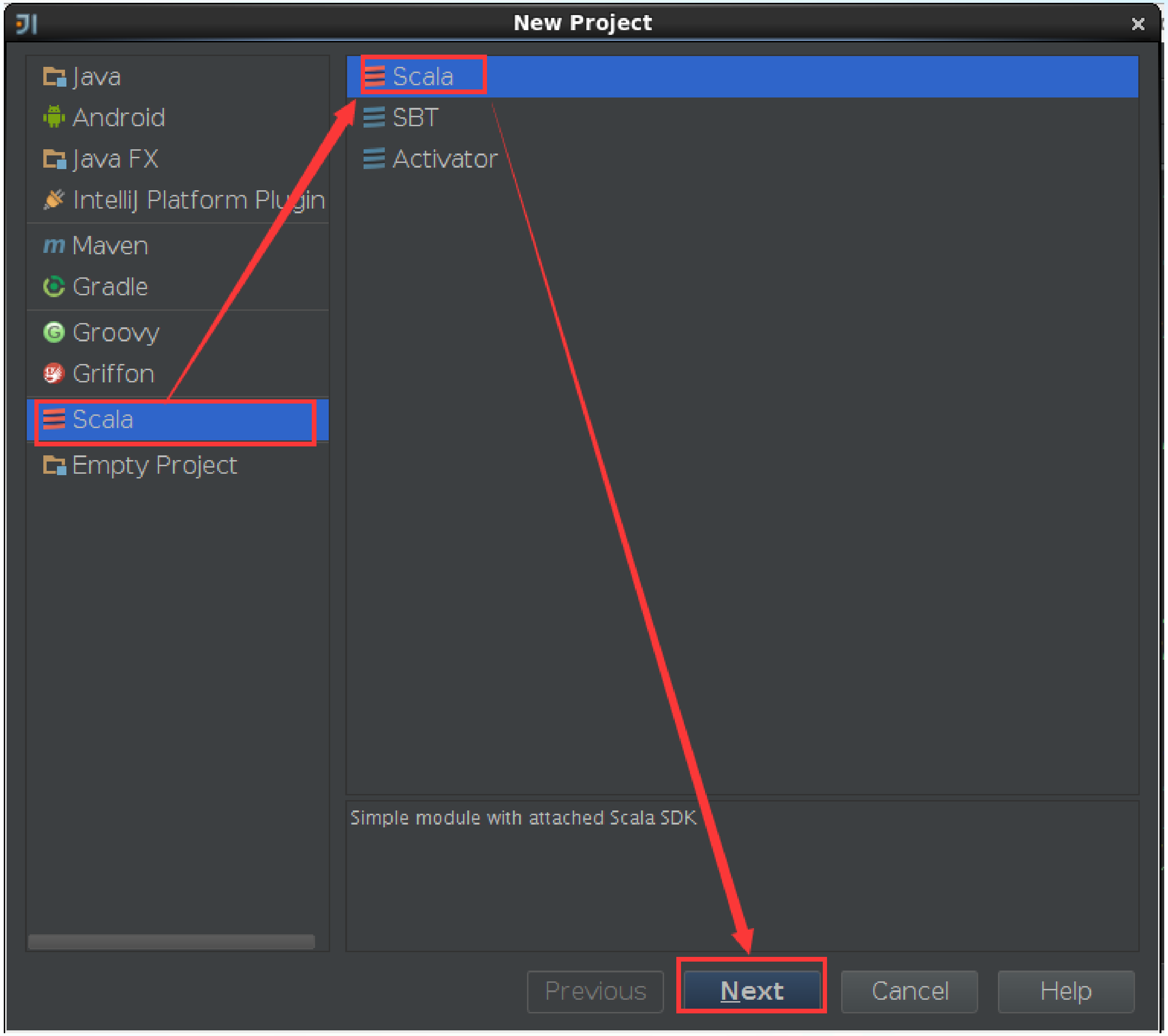
创建Scala Project

讲师-梦琪

Create Scala Project

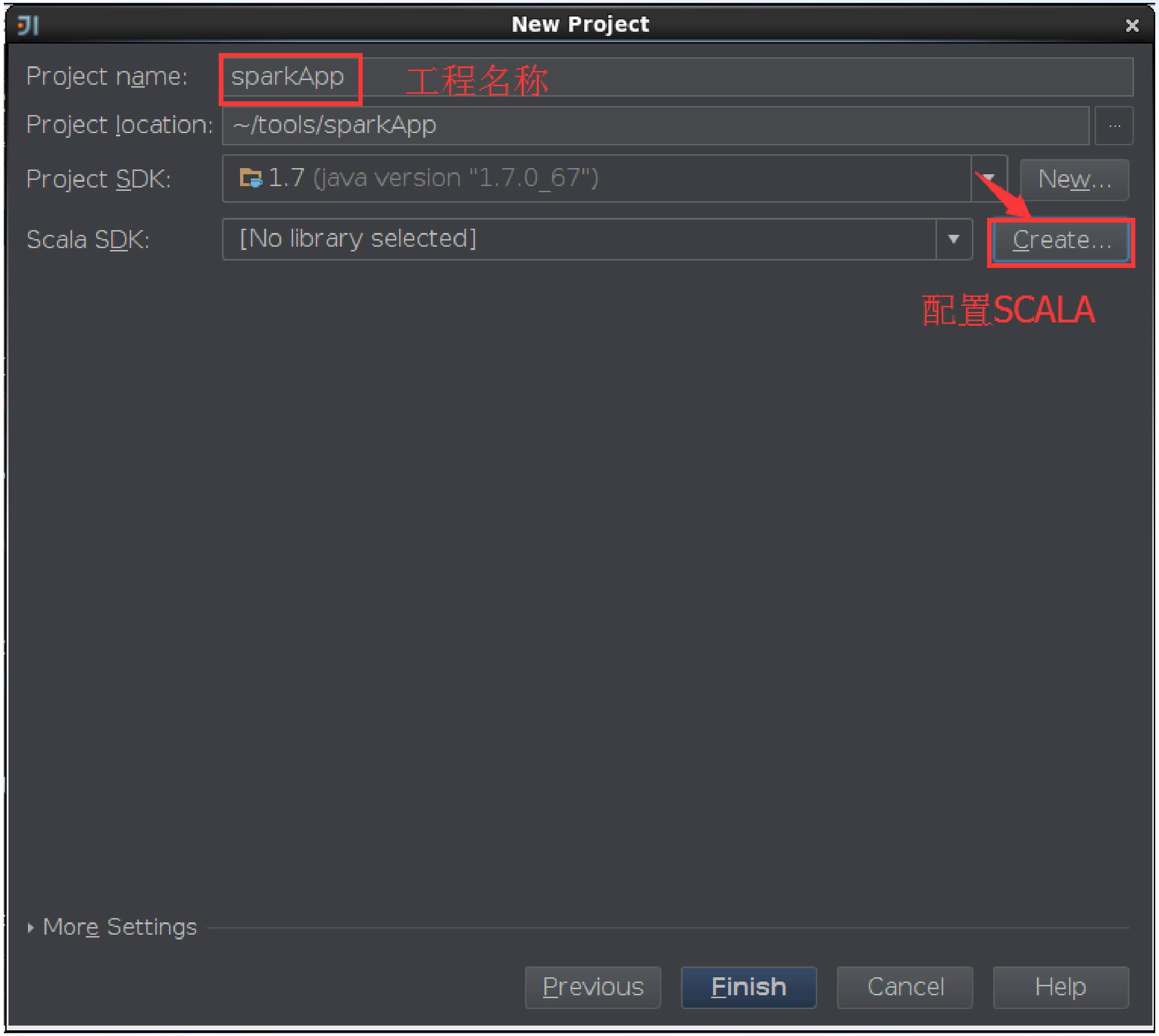


Create Scala Project



Simple module with attached Scala SDK

Create Scala Project



New Project

Project name: sparkApp

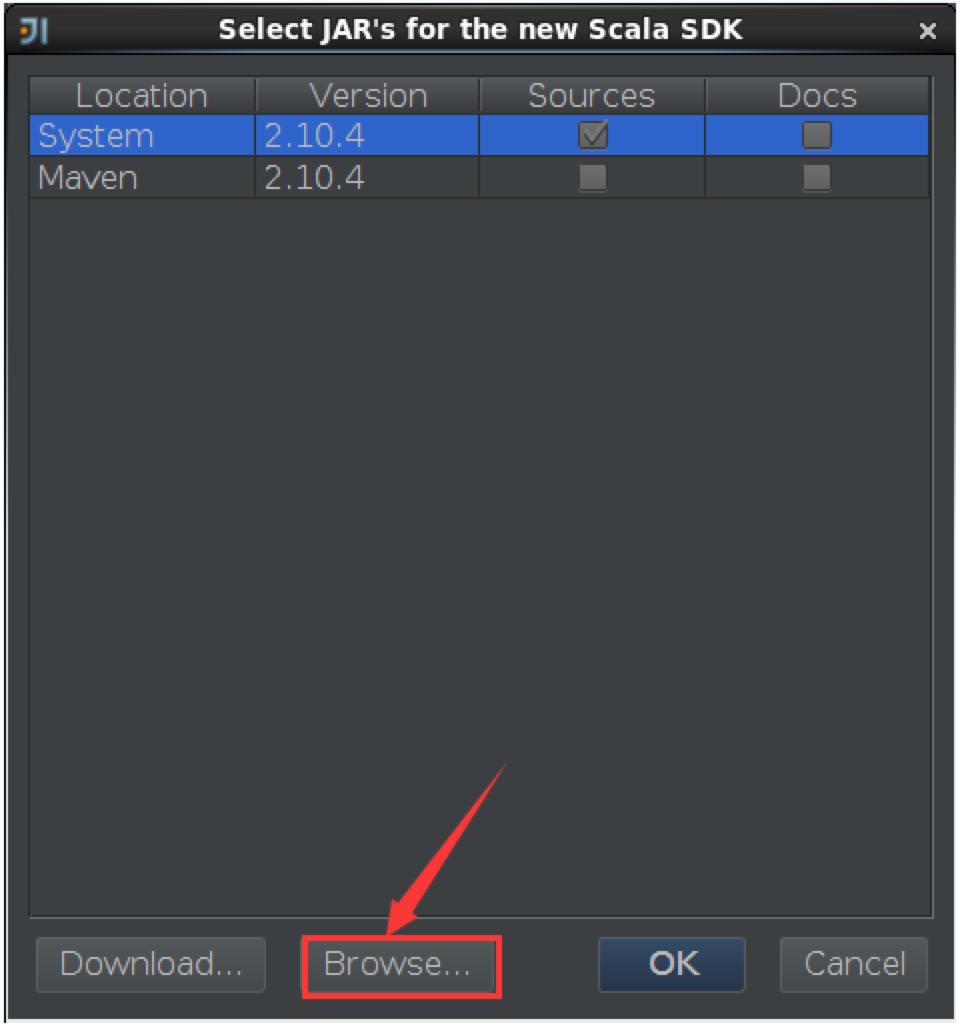
工程名称

Project location: ~/tools/sparkApp

Project SDK: 1.7 (java version "1.7.0_67")

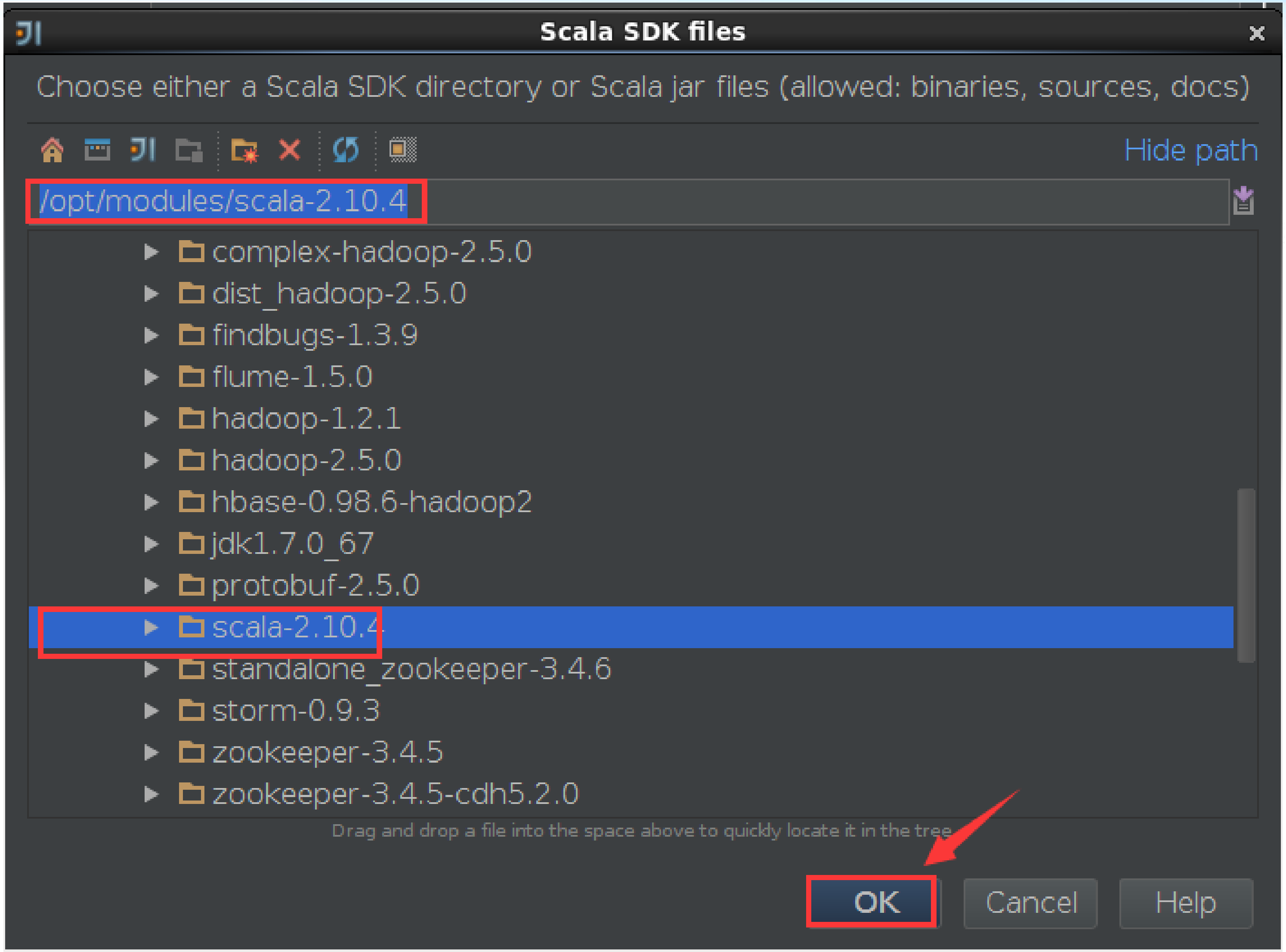
Scala SDK: [No library selected]

Set Scala SDK



Location	Version	Sources	Docs
System	2.10.4	☑	□
Maven	2.10.4	□	□

Set Scala SDK



Scala SDK files

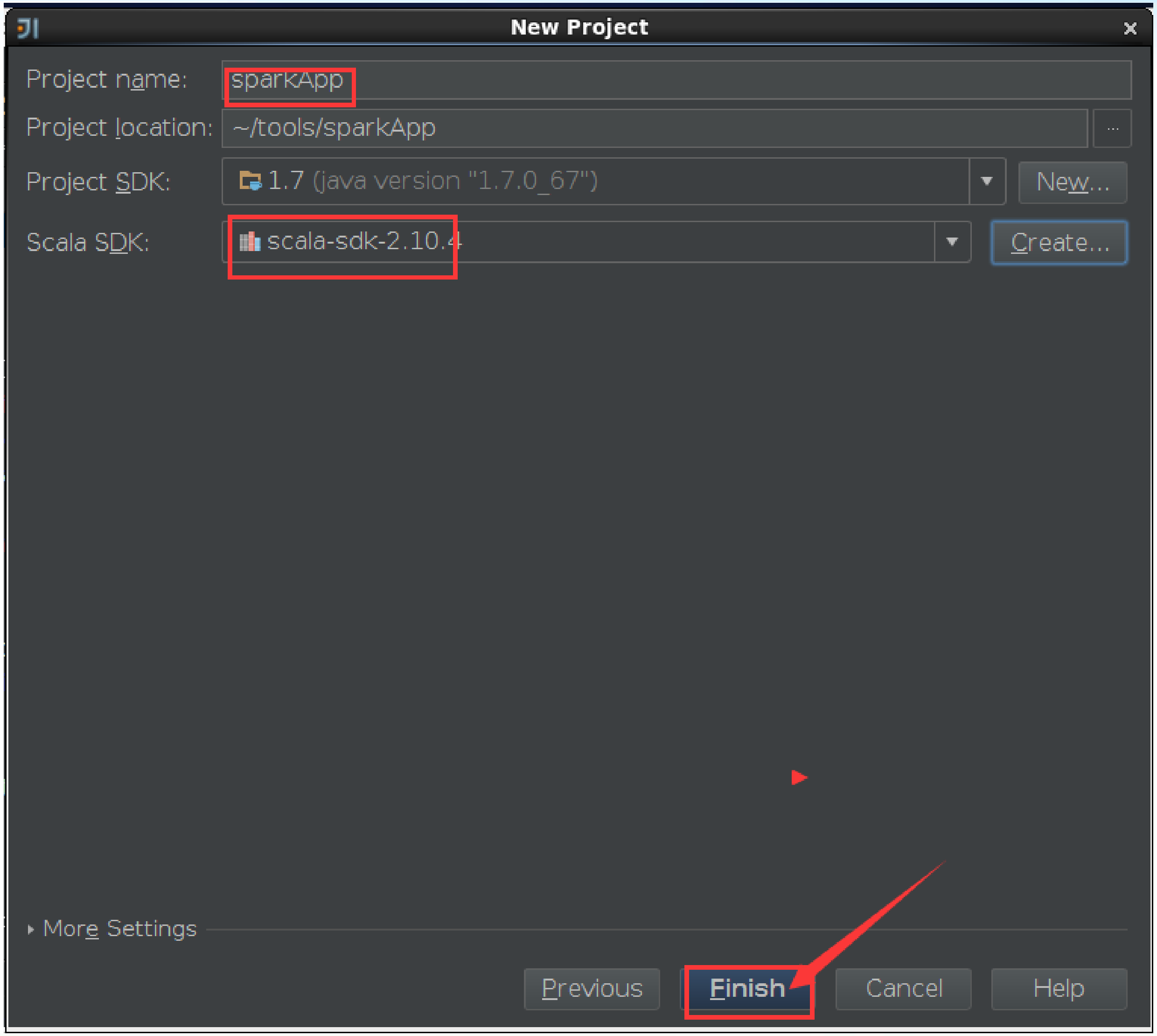
Choose either a Scala SDK directory or Scala jar files (allowed: binaries, sources, docs)

/opt/modules/scala-2.10.4

- complex-hadoop-2.5.0
- dist_hadoop-2.5.0
- findbugs-1.3.9
- flume-1.5.0
- hadoop-1.2.1
- hadoop-2.5.0
- hbase-0.98.6-hadoop2
- jdk1.7.0_67
- protobuf-2.5.0
- scala-2.10.4
- standalone_zookeeper-3.4.6
- storm-0.9.3
- zookeeper-3.4.5
- zookeeper-3.4.5-cdh5.2.0

Drag and drop a file into the space above to quickly locate it in the tree

Create Scala Project



New Project

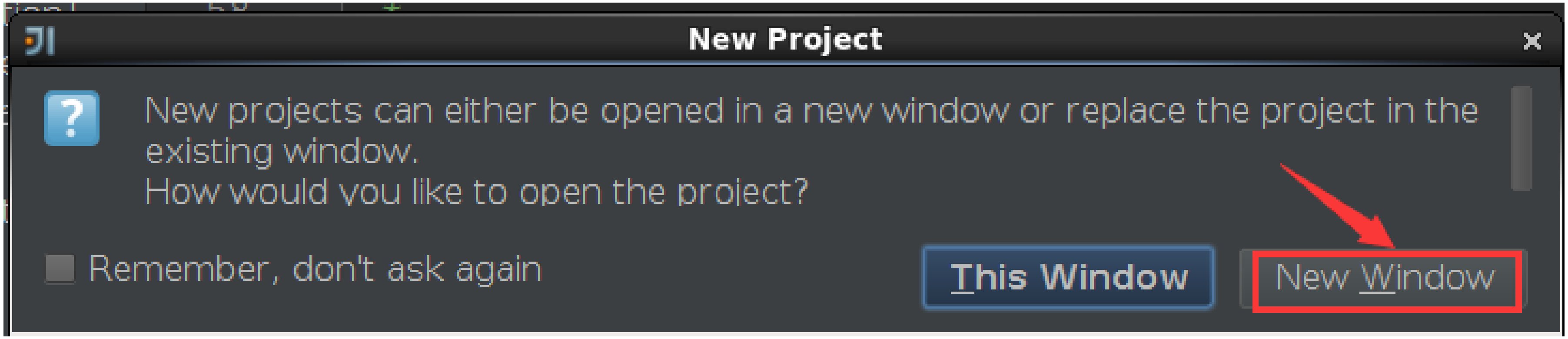
Project name: sparkApp

Project location: ~/tools/sparkApp

Project SDK: 1.7 (java version "1.7.0_67")

Scala SDK: scala-sdk-2.10.4

Create Scala Project

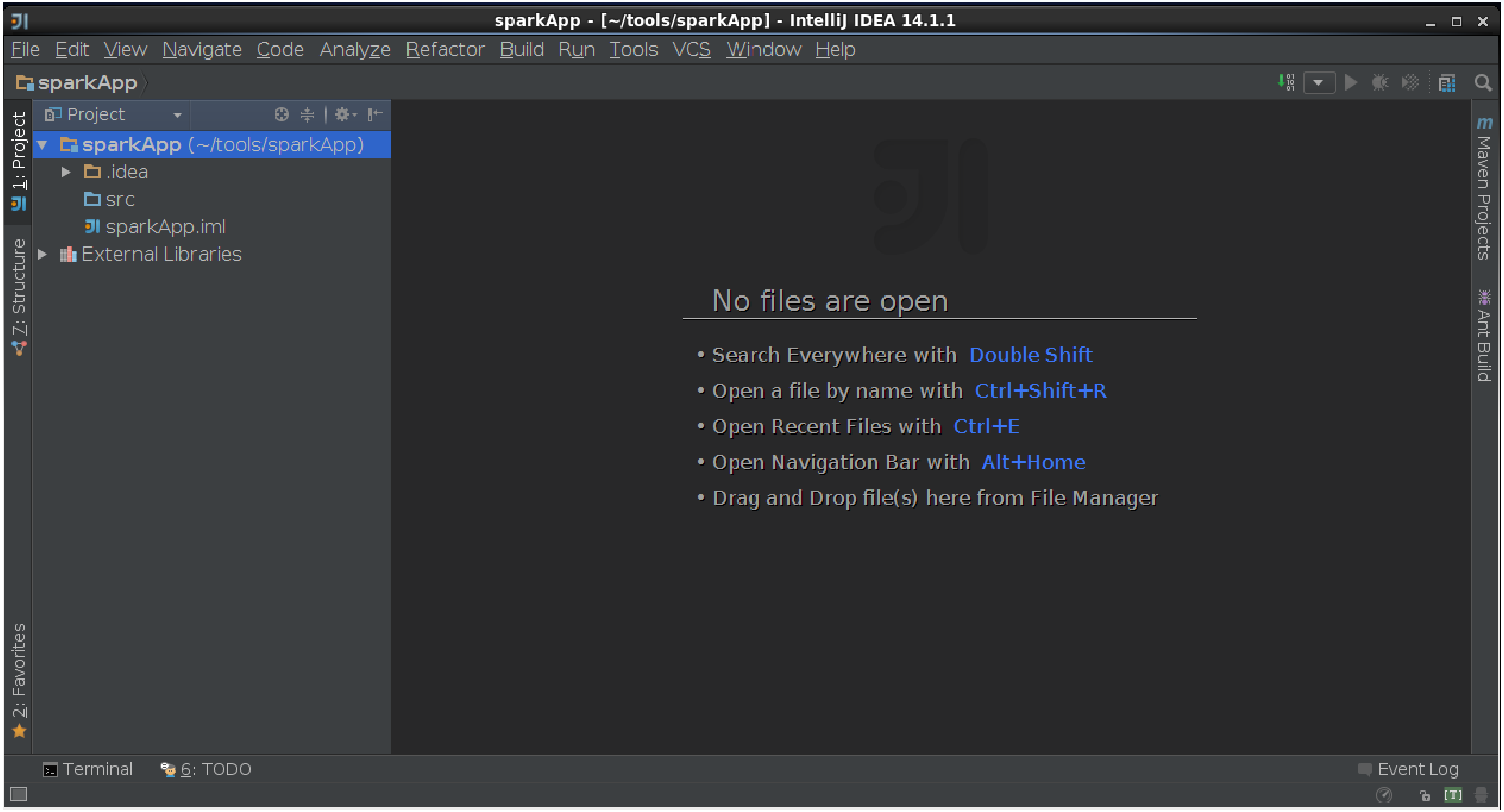


New Project

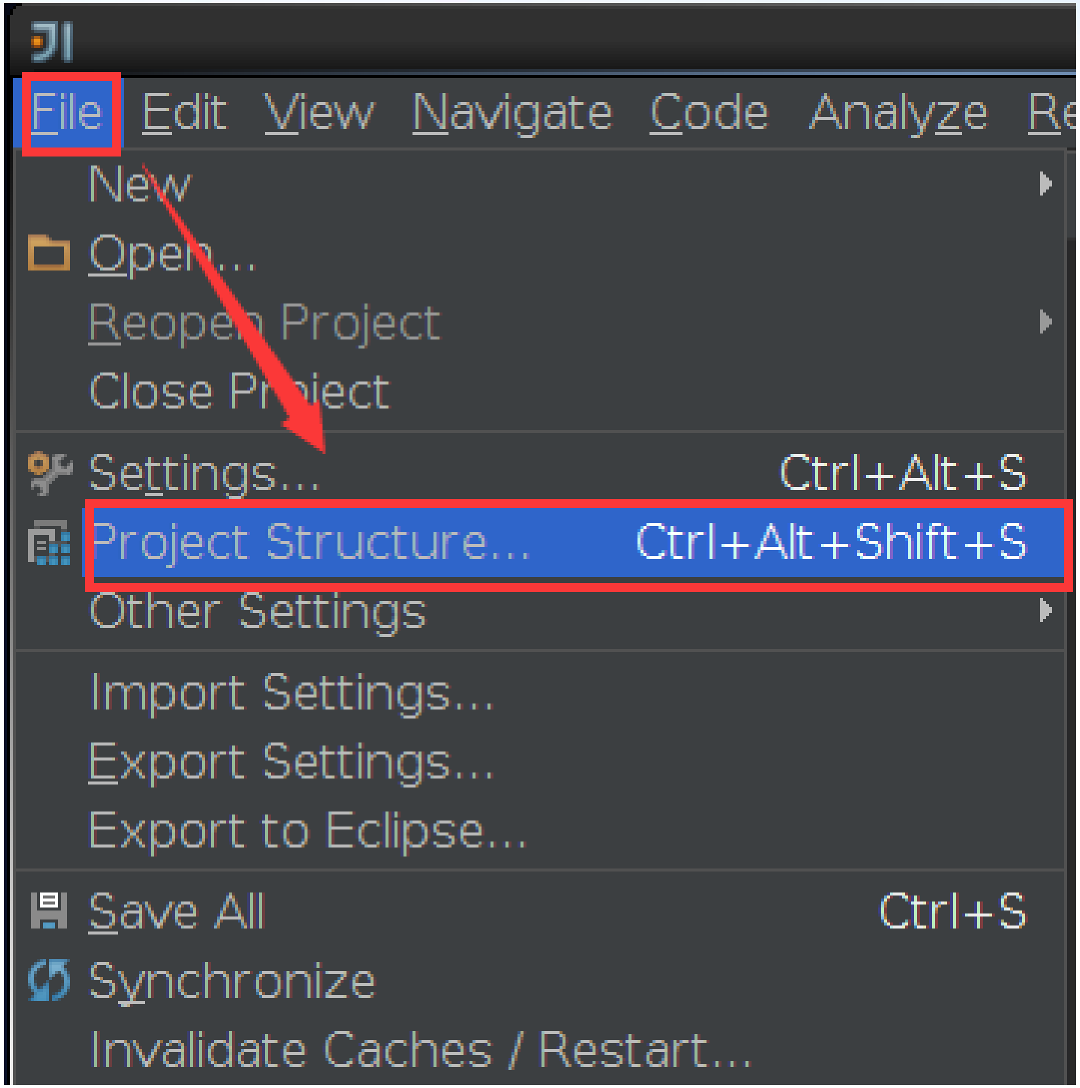
New projects can either be opened in a new window or replace the project in the existing window.
How would you like to open the project?

□ Remember, don't ask again

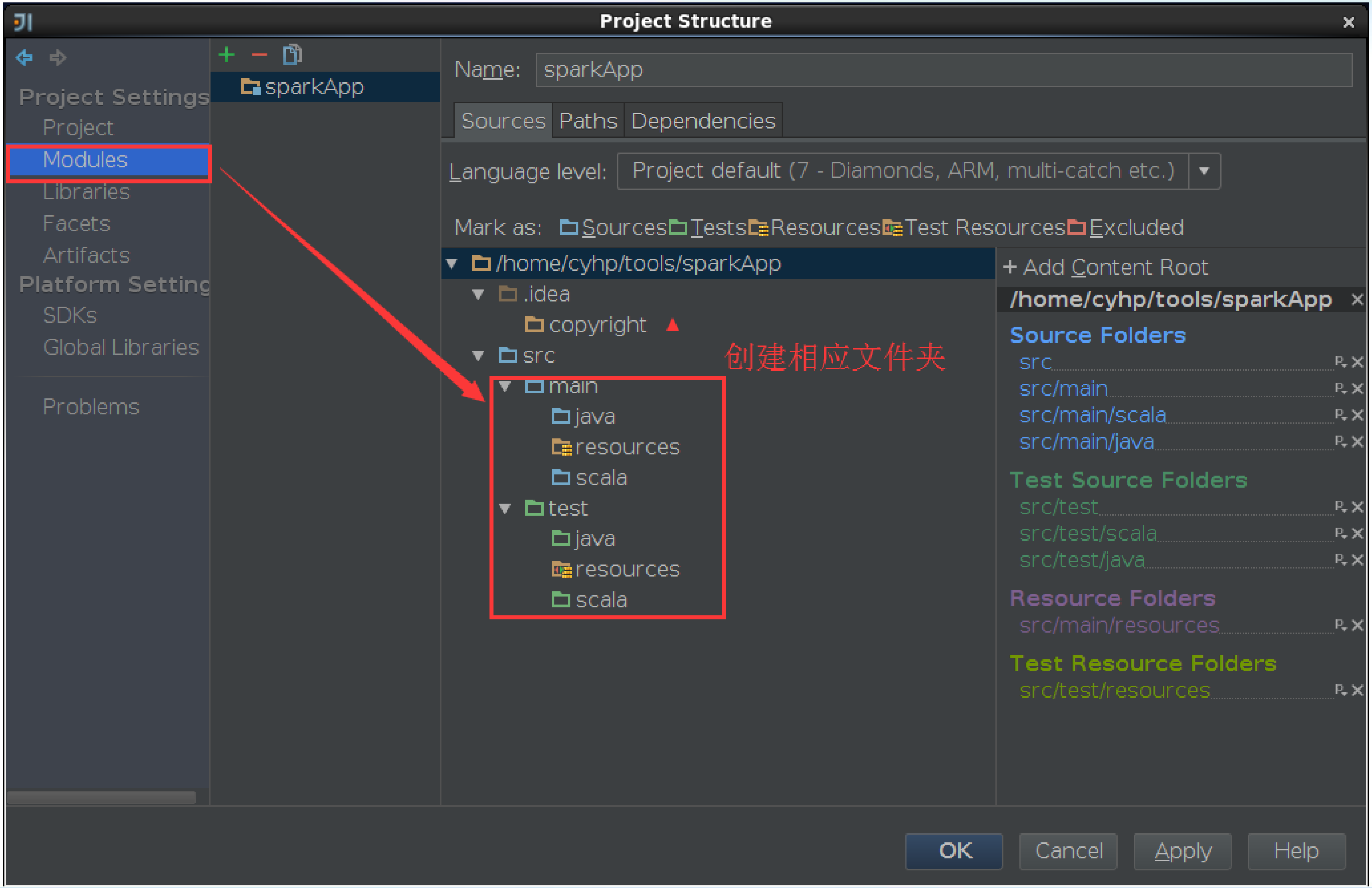
Create Scala Project


sparkApp - [~/tools/sparkApp] - IntelliJ IDEA 14.1.1
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
sparkApp
Project
sparkApp (~/tools/sparkApp)
.idea
src
sparkApp.iml
External Libraries
1: Project
7: Structure
2: Favorites
Maven Projects
Ant Build
No files are open
• Search Everywhere with Double Shift
• Open a file by name with Ctrl+Shift+R
• Open Recent Files with Ctrl+E
• Open Navigation Bar with Alt+Home
• Drag and Drop file(s) here from File Manager
Terminal
6: TODO
Event Log

Set Scala Project



Set Scala Project



Name: sparkApp

Language level: Project default (7 - Diamonds, ARM, multi-catch etc.)

Mark as: Sources Tests Resources Test Resources Excluded

```
▼ /home/cyhp/tools/sparkApp
  ▼ .idea
      copyright
  ▼ src
    ▼ main
        java
        resources
        scala
    ▼ test
        java
        resources
        scala
```

创建相应文件夹

/home/cyhp/tools/sparkApp

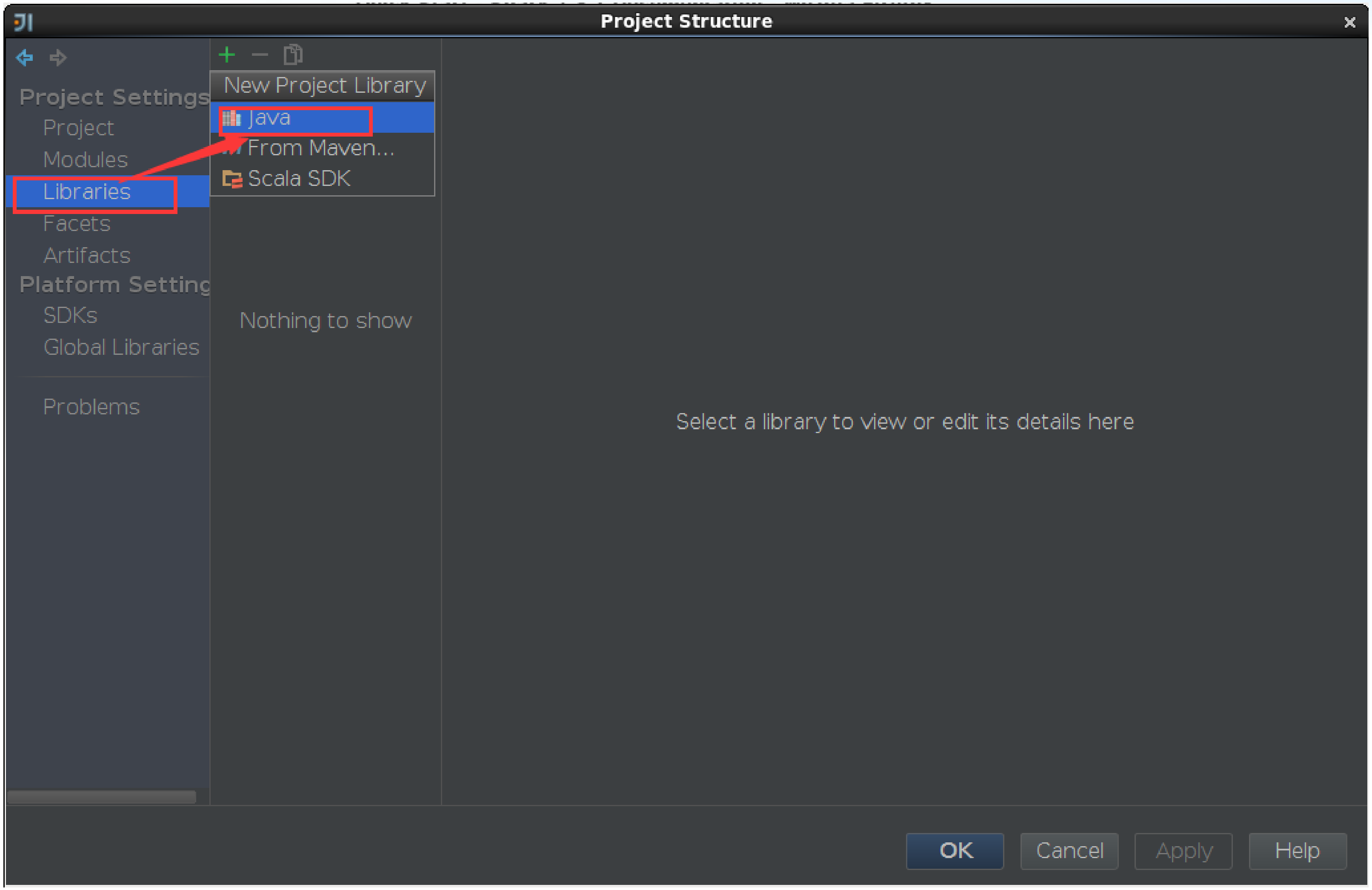
Source Folders
src
src/main
src/main/scala
src/main/java

Test Source Folders
src/test
src/test/scala
src/test/java

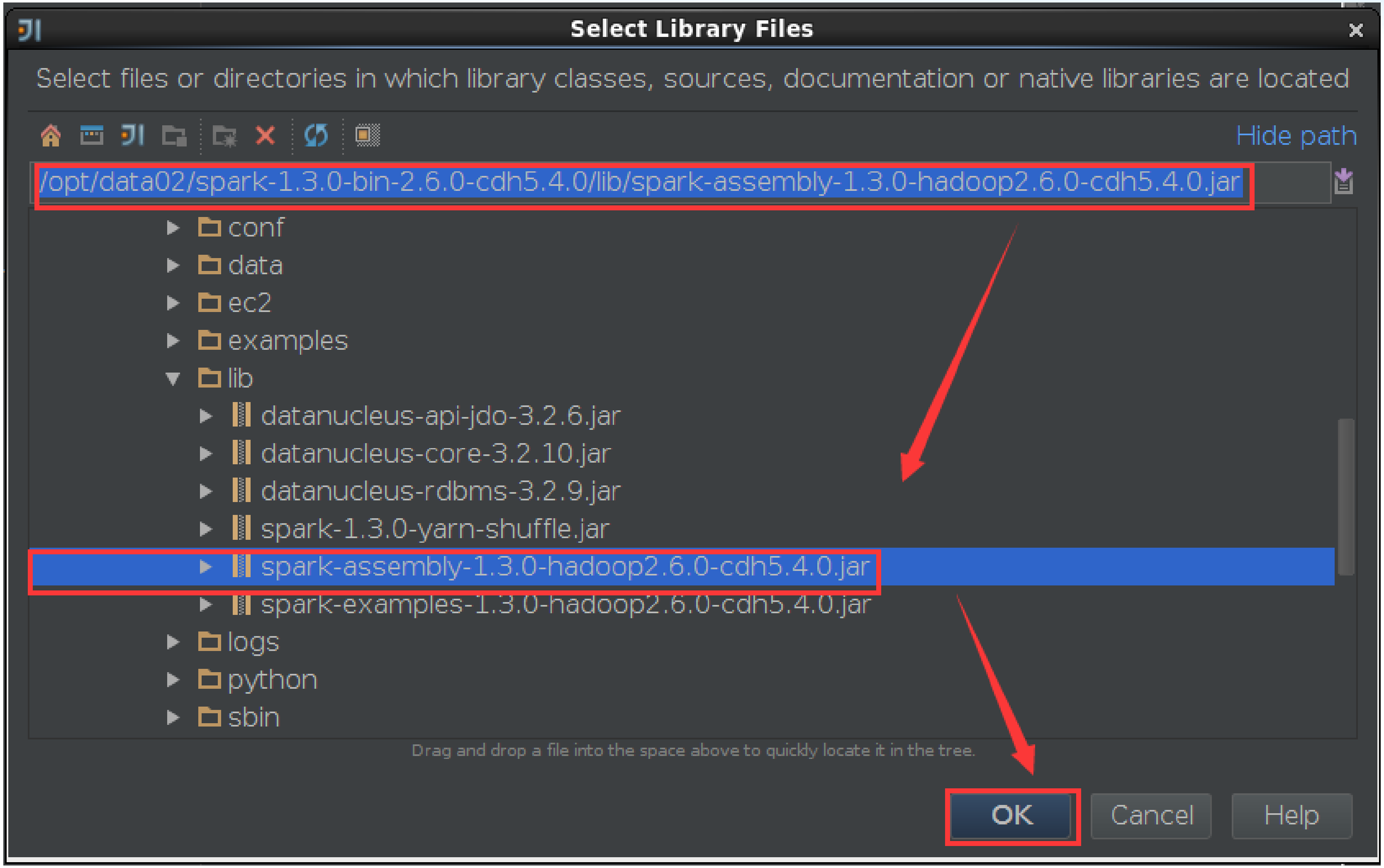
Resource Folders
src/main/resources

Test Resource Folders
src/test/resources

Add Spark JARS



Add Spark JARS



Select Library Files

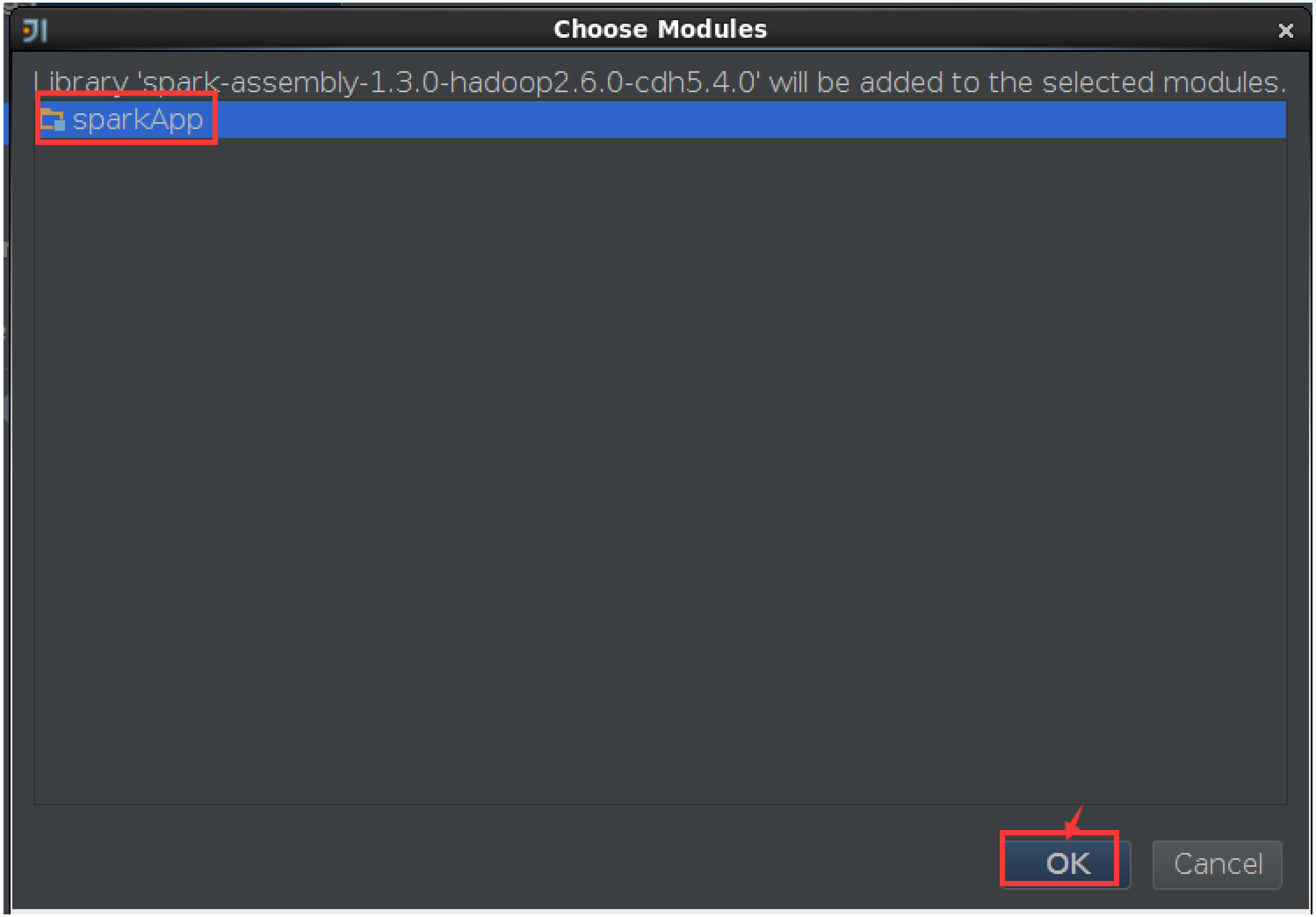
Select files or directories in which library classes, sources, documentation or native libraries are located

```
/opt/data02/spark-1.3.0-bin-2.6.0-cdh5.4.0/lib/spark-assembly-1.3.0-hadoop2.6.0-cdh5.4.0.jar
```

```
▶ conf
▶ data
▶ ec2
▶ examples
▼ lib
    ▶ datanucleus-api-jdo-3.2.6.jar
    ▶ datanucleus-core-3.2.10.jar
    ▶ datanucleus-rdbms-3.2.9.jar
    ▶ spark-1.3.0-yarn-shuffle.jar
    ▶ spark-assembly-1.3.0-hadoop2.6.0-cdh5.4.0.jar
    ▶ spark-examples-1.3.0-hadoop2.6.0-cdh5.4.0.jar
▶ logs
▶ python
▶ sbin
```

Drag and drop a file into the space above to quickly locate it in the tree.

Add Spark JARS



Choose Modules

Library 'spark-assembly-1.3.0-hadoop2.6.0-cdh5.4.0' will be added to the selected modules.

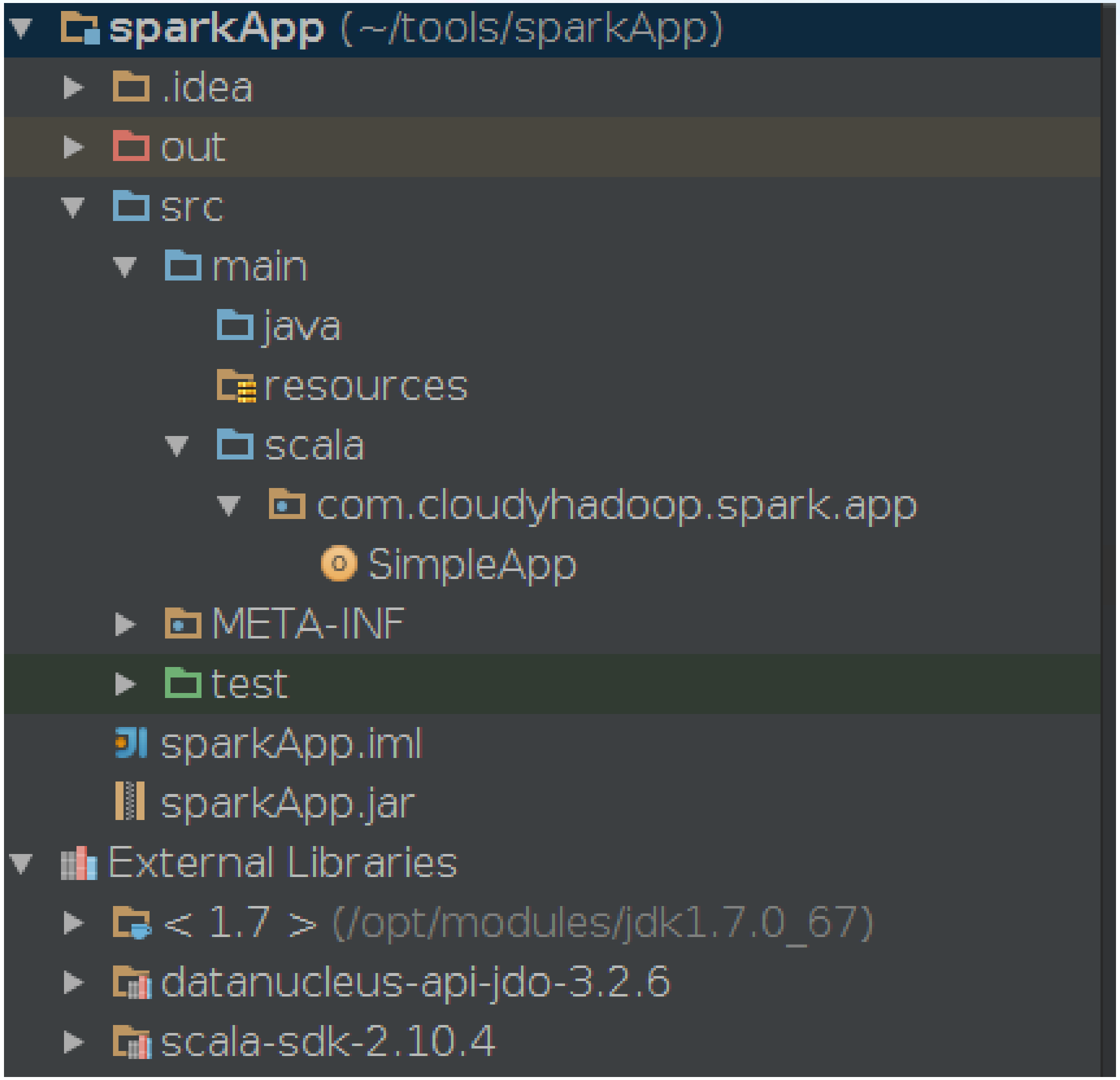
零基础学习Spark 1.x应用开发系列课程

IDEA打包Spark Application

讲师-梦琪

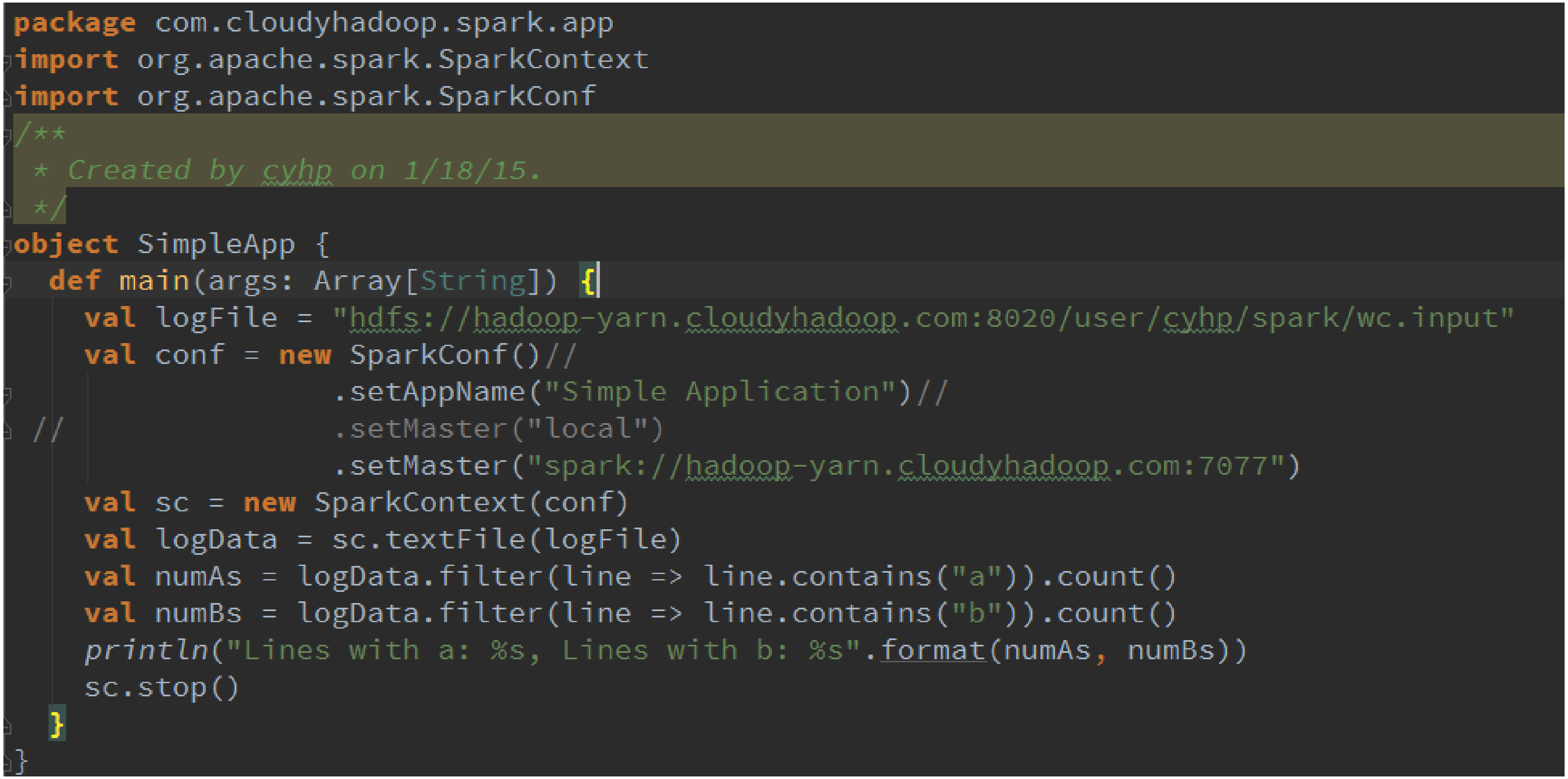
版权所有云帆大数据 www.cloudyhadoop.com

First Spark Application



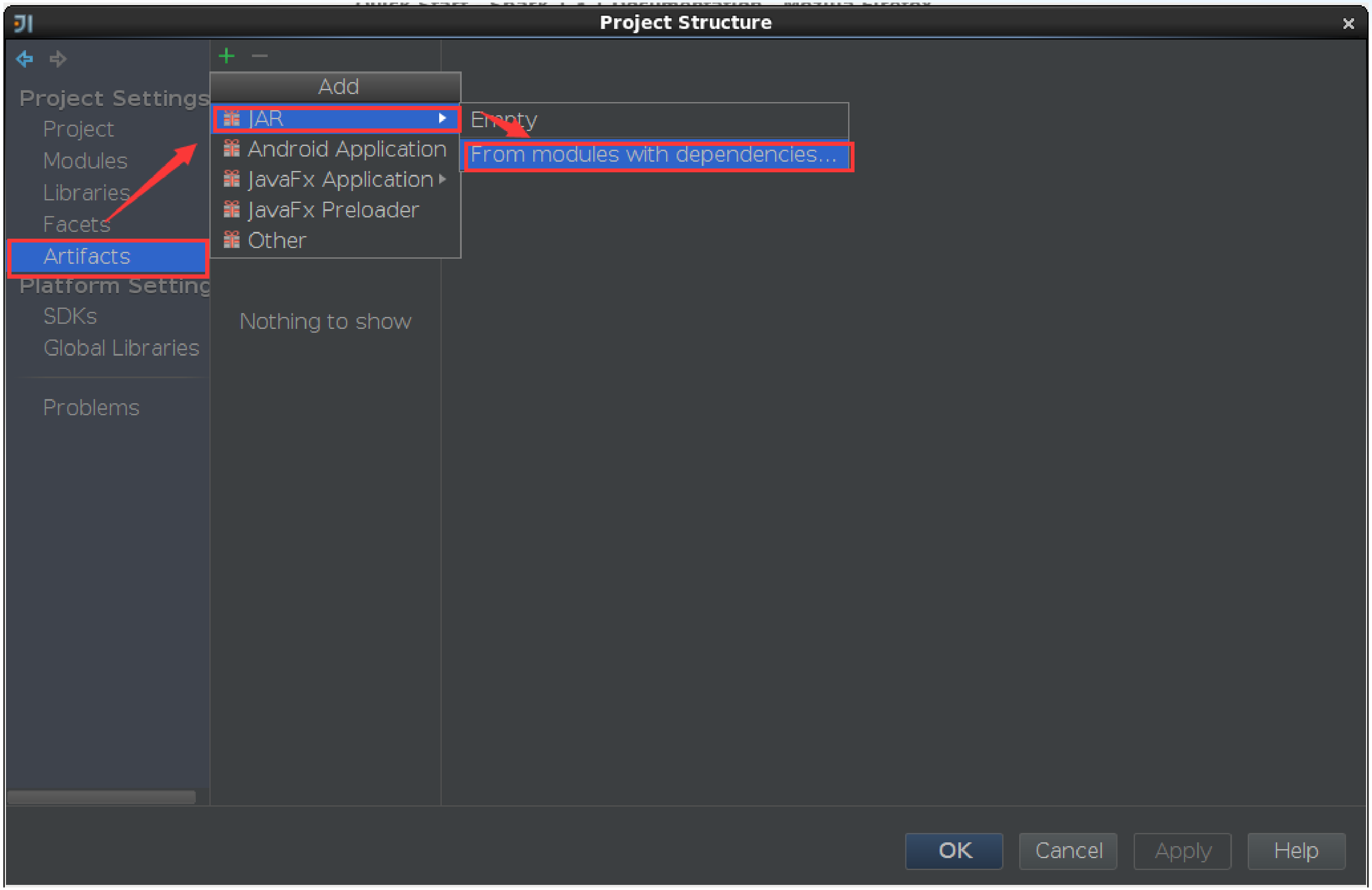
```
▼ sparkApp (~/tools/sparkApp)
    ▶ .idea
    ▶ out
    ▼ src
        ▼ main
            java
            resources
            ▼ scala
                ▼ com.cloudyhadoop.spark.app
                    SimpleApp
        ▶ META-INF
        ▶ test
    sparkApp.iml
    sparkApp.jar
▼ External Libraries
    ▶ < 1.7 > (/opt/modules/jdk1.7.0_67)
    ▶ datanucleus-api-jdo-3.2.6
    ▶ scala-sdk-2.10.4
```

First Spark Application

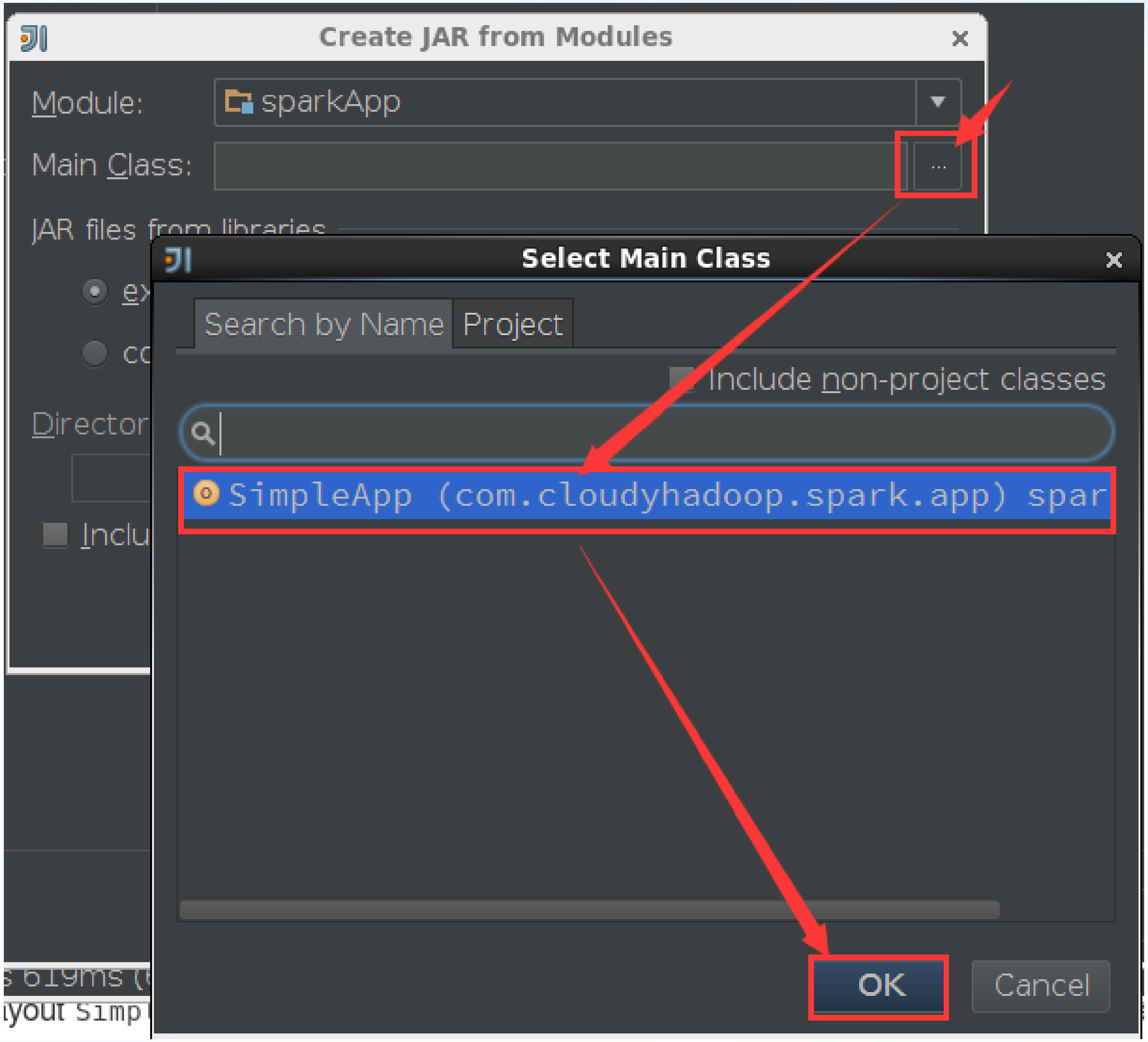


```
package com.cloudyhadoop.spark.app
import org.apache.spark.SparkContext
import org.apache.spark.SparkConf
/**
 * Created by cyhp on 1/18/15.
 */
object SimpleApp {
  def main(args: Array[String]) {
    val logFile = "hdfs://hadoop-yarn.cloudyhadoop.com:8020/user/cyhp/spark/wc.input"
    val conf = new SparkConf()//
                  .setAppName("Simple Application")//
//                .setMaster("local")
                  .setMaster("spark://hadoop-yarn.cloudyhadoop.com:7077")
    val sc = new SparkContext(conf)
    val logData = sc.textFile(logFile)
    val numAs = logData.filter(line => line.contains("a")).count()
    val numBs = logData.filter(line => line.contains("b")).count()
    println("Lines with a: %s, Lines with b: %s".format(numAs, numBs))
    sc.stop()
  }
}
```

Select Artifact



Create JAR form Modules



Create JAR from Modules

Module: sparkApp

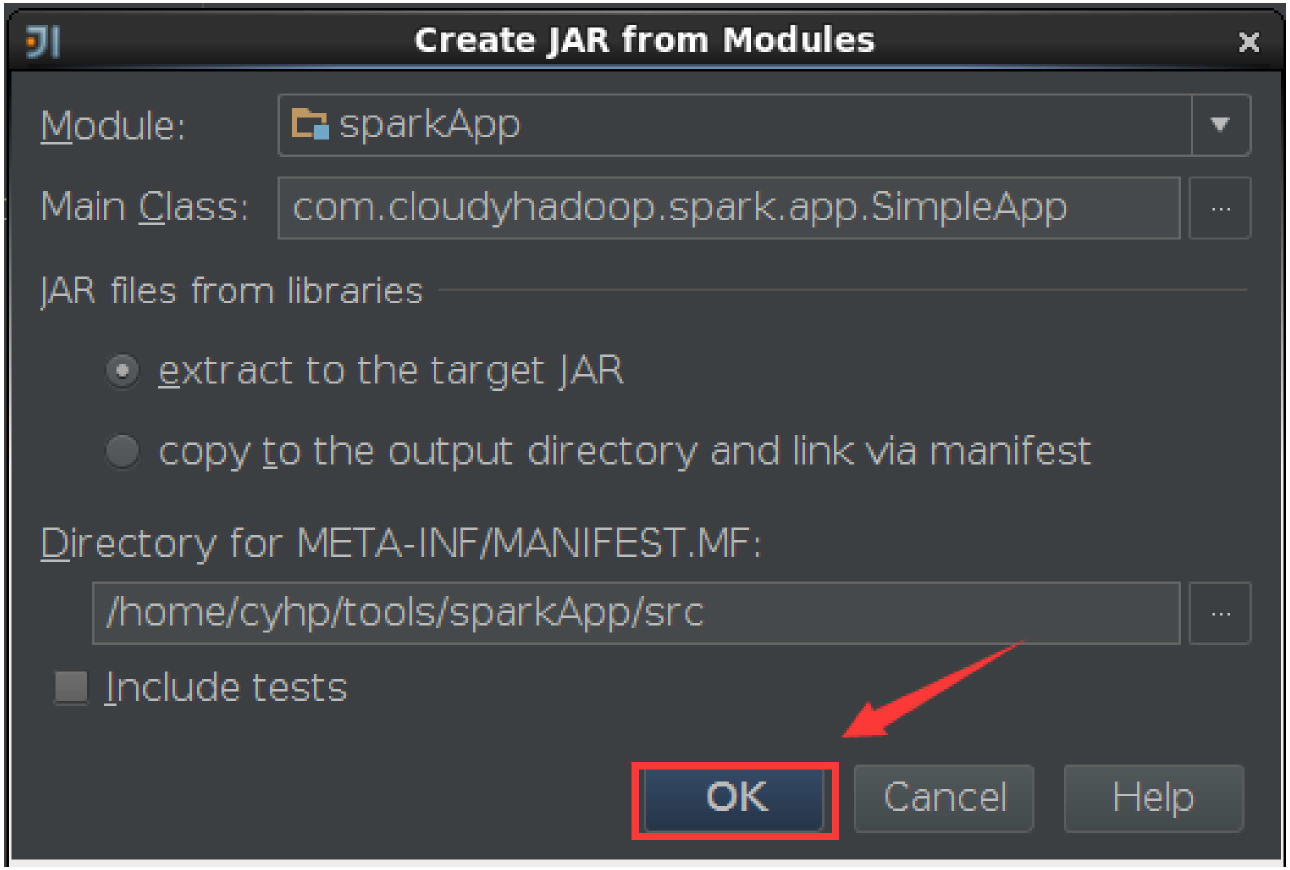
Main Class:

JAR files from libraries

Select Main Class

SimpleApp (com.cloudyhadoop.spark.app) spar

Create JAR form Modules



Create JAR from Modules

Module: sparkApp

Main Class: com.cloudyhadoop.spark.app.SimpleApp

JAR files from libraries

● extract to the target JAR

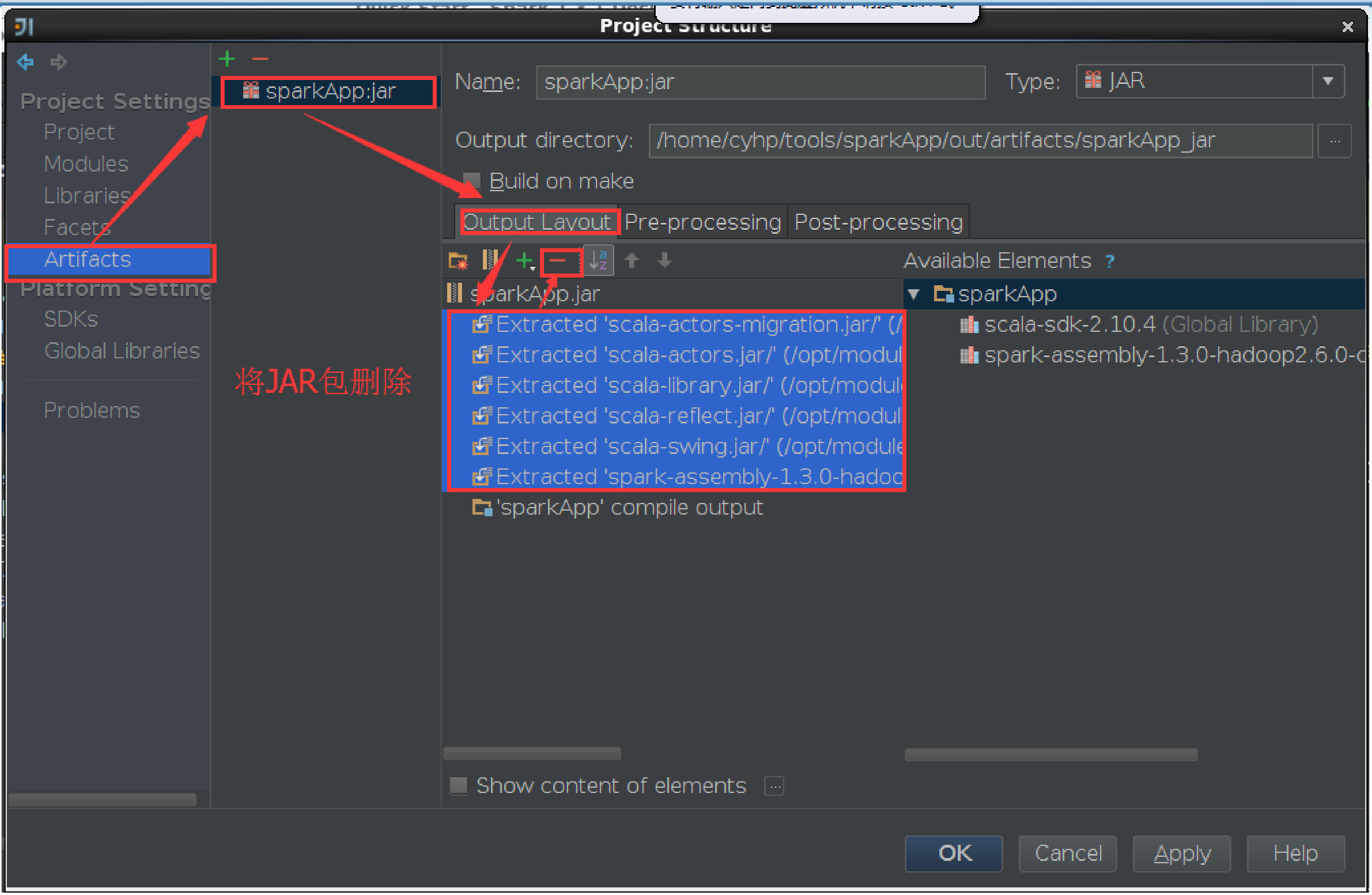
○ copy to the output directory and link via manifest

Directory for META-INF/MANIFEST.MF:

/home/cyhp/tools/sparkApp/src

□ Include tests

Delete Spark JARS



Project Structure

Name: sparkApp:jar

Type: JAR

Output directory: /home/cyhp/tools/sparkApp/out/artifacts/sparkApp_jar

□ Build on make

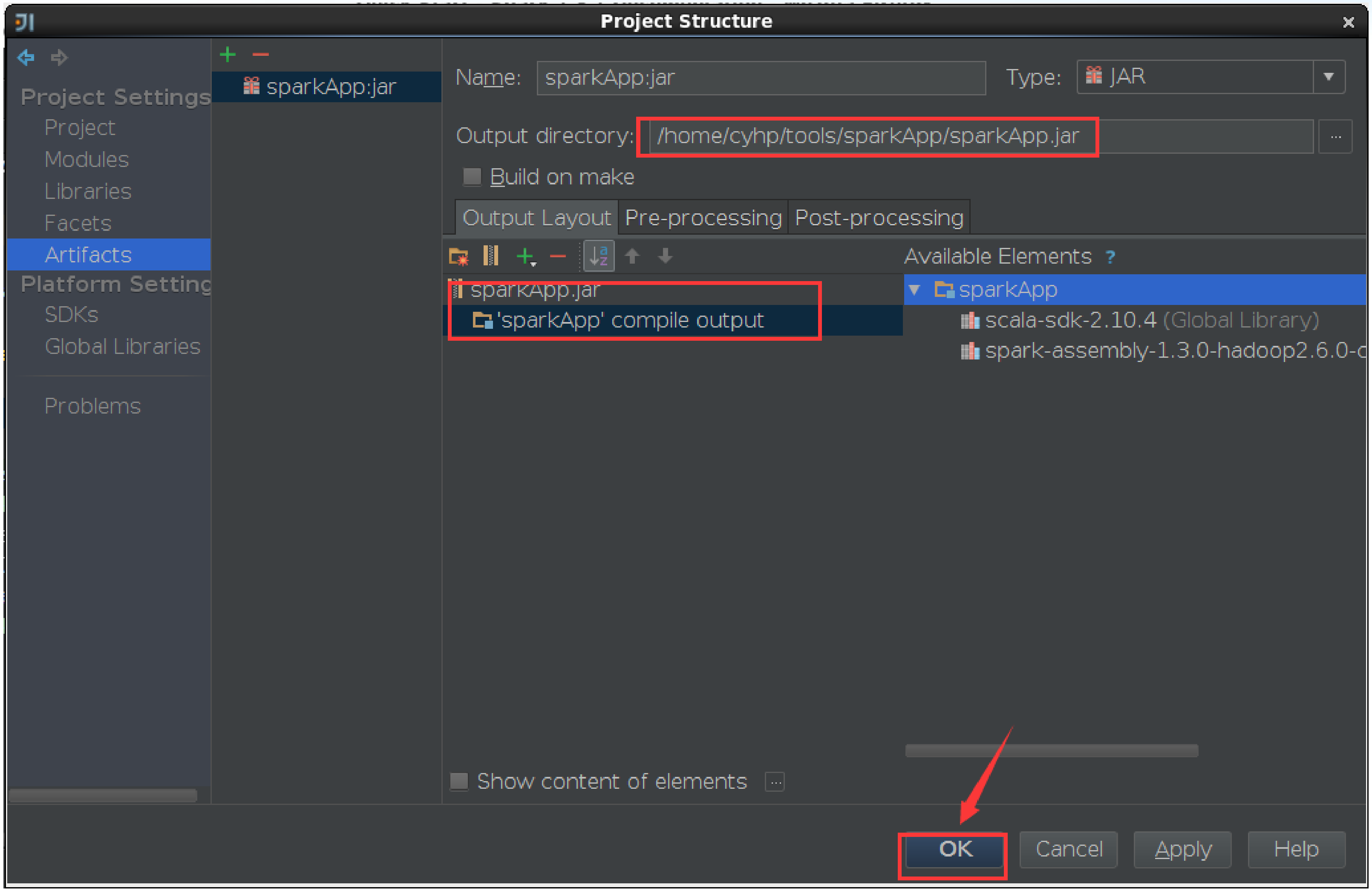
```
sparkApp.jar
  Extracted 'scala-actors-migration.jar/' (/
  Extracted 'scala-actors.jar/' (/opt/modul
  Extracted 'scala-library.jar/' (/opt/modul
  Extracted 'scala-reflect.jar/' (/opt/modul
  Extracted 'scala-swing.jar/' (/opt/module
  Extracted 'spark-assembly-1.3.0-hadoo
  'sparkApp' compile output
```

```
▼ sparkApp
    scala-sdk-2.10.4 (Global Library)
    spark-assembly-1.3.0-hadoop2.6.0-c
```

□ Show content of elements

将JAR包删除

Set Output



Project Structure

Project Settings
- Project
- Modules
- Libraries
- Facets
- Artifacts

Platform Settings
- SDKs
- Global Libraries

Problems

sparkApp:jar

Name: sparkApp:jar

Type: JAR

Output directory: /home/cyhp/tools/sparkApp/sparkApp.jar

□ Build on make

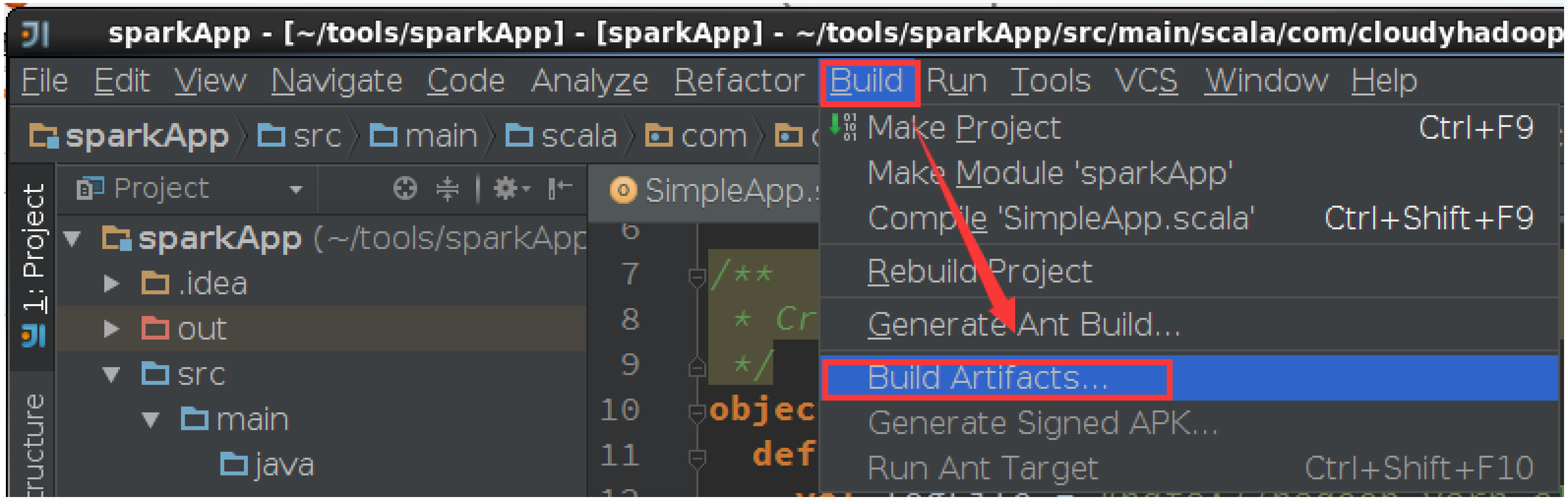
- sparkApp.jar
  - 'sparkApp' compile output

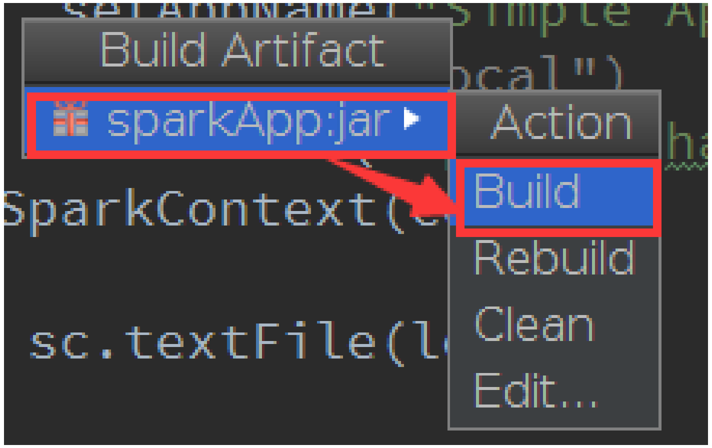
Available Elements ?

- sparkApp
  - scala-sdk-2.10.4 (Global Library)
  - spark-assembly-1.3.0-hadoop2.6.0-c

□ Show content of elements

Build Artifacts


sparkApp - [~/tools/sparkApp] - [sparkApp] - ~/tools/sparkApp/src/main/scala/com/cloudyhadoop
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
sparkApp src main scala com
Project
sparkApp (~/tools/sparkApp
.idea
out
src
main
java
SimpleApp.s
Make Project Ctrl+F9
Make Module 'sparkApp'
Compile 'SimpleApp.scala' Ctrl+Shift+F9
Rebuild Project
Generate Ant Build...
Build Artifacts...
Generate Signed APK...
Run Ant Target Ctrl+Shift+F10



云帆大数据
CloudyHadoop

云帆大数据网络课程

- 云帆大数据是国内首家坚持实时在线授课、提供高端开发课程网络培训机构。采用新兴的互联网教育模式，坚持实时在线授课模式，既继承传统教育的学习交流特点，又发挥互联网的无处不在的时空特性，将天南地北有志向的人才组织在一起学习交流，使原先孤立的个体学习，组合成有组织的学习探讨，并且把原先的学费用降低到十分之一左右，使更多的人能学习到最新的高端课程技术。云帆大数据同时是一个平台，如果你是一个学员，可以尽情的学习和交流；如果你是一个有梦想有才华的人，可以联系我们，给你提供一片驰骋的原野。
- 云帆大数据已推出国内首家《企业级Hadoop 2.x 应用开发课程》、《企业级Hadoop 2.x 项目实战课程》和《企业级Spark 1.x 应用开发课程》，更多其他详情，请登录我们的培训网站http://www.cloudyhadoop.com。

实时在线授课，专业课程辅导