

How to Develop Spark Applications on Windows 10 Using IntelliJ IDEA

Software Installation on WINDOWS 10

● Prerequisites:

1. The Spark commodity cluster has been built on Linux (CentOS is recommended)

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

Linux can be used for *developing, testing, and deploying* Spark applications

Both Mac OS X and WINDOWS can be only used for *developing and testing*

2. The PC should be **in the same LAN** with the Spark commodity cluster

● Software installed on WINDOWS 10:

1. Java version 1.8.0_131

use the CMD command “*java --version*” to validate whether Java can work

2. Scala version 2.11.11

use the CMD command “*scala --version*” to validate whether Scala can work

3. IntelliJ IDEA (the community version)

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

(NOTE that both the Java and Scala version on WINDOWS 10 should be the same as the corresponding version on the Spark commodity cluster.)

Configure System Environment Variables on WINDOWS 10

- Add the binding of the hostname and IP address of all the nodes on the Spark commodity cluster to the *hosts* file, which is usually located in:

C:\Windows\System32\drivers\etc\hosts

A sample for the *hosts* file:

```
# 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
...            ...
```

Configure System Environment Variables on WINDOWS 10

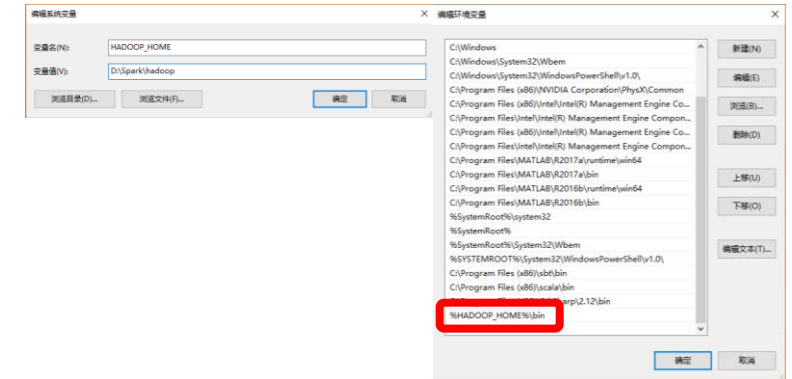
- Download a Hadoop component to a new folder (e.g., “D:\Spark\hadoop”). The website of the Hadoop component is presented below:

<https://github.com/srccodes/hadoop-common-2.2.0-bin>

Note that the new folder should not include other files.

- Configure the following system environment variables:

HADOOP_HOME
PATH

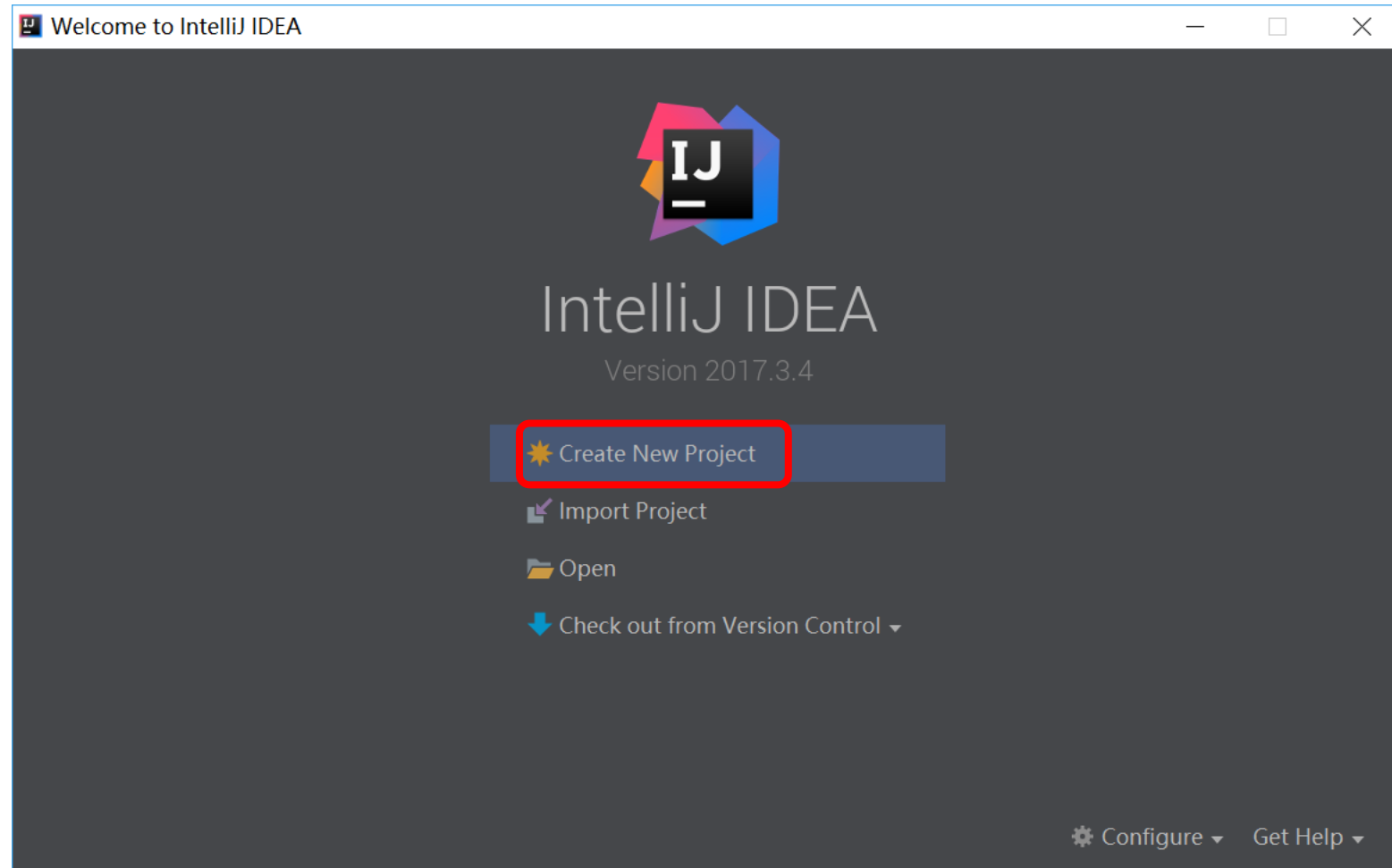


If you can't configure them correctly, the below error may be thrown:

java.io.IOException: Could not locate executable null\bin\winutils.exe in the Hadoop binaries.

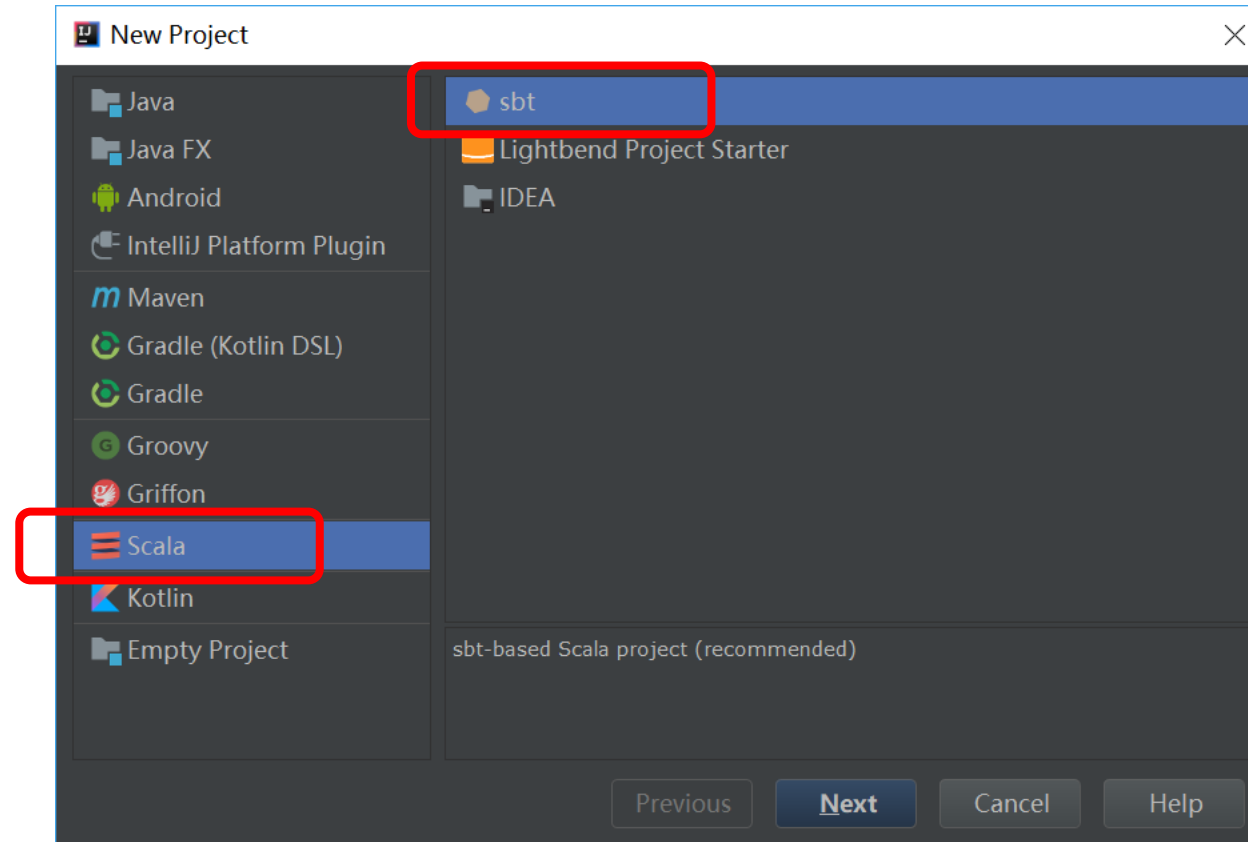
Use IntelliJ IDEA

- Create a Spark application: create new project



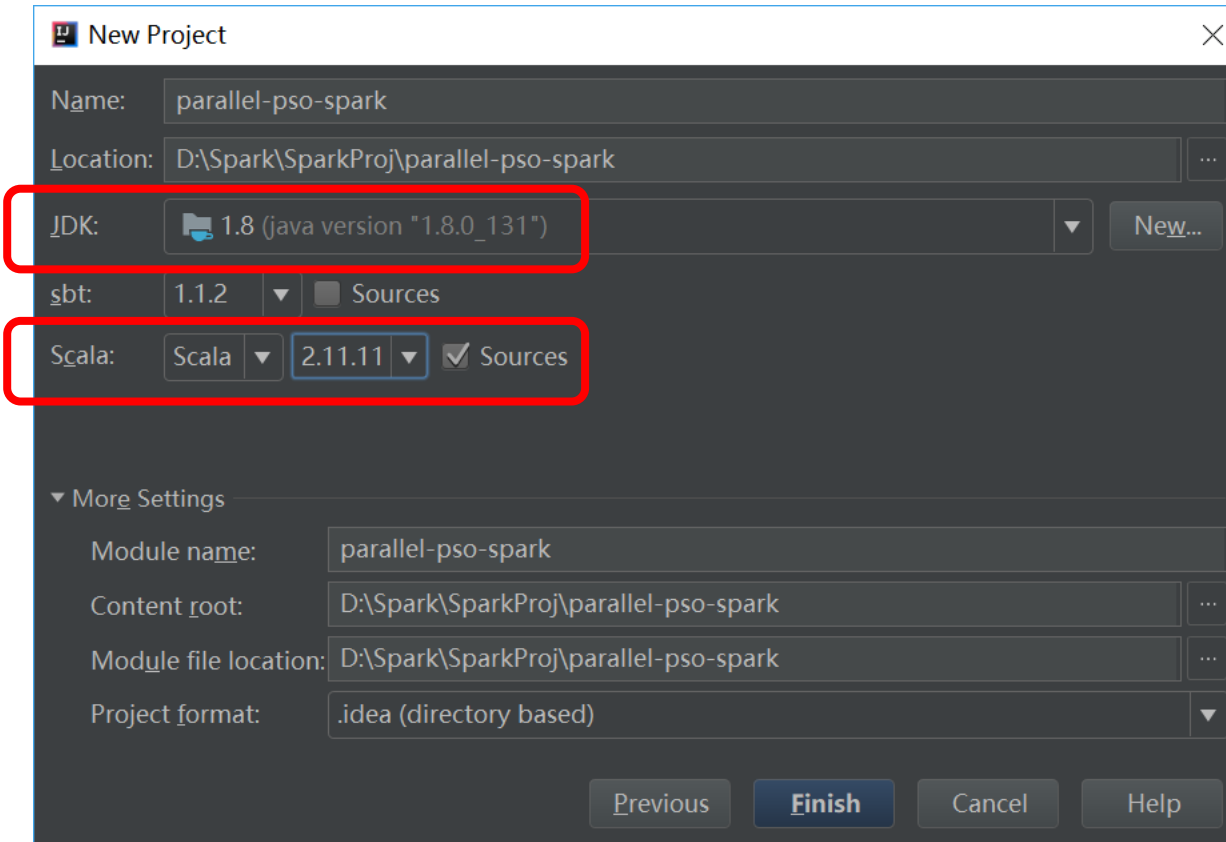
Use IntelliJ IDEA

- Create a Spark application: create a **sbt** project



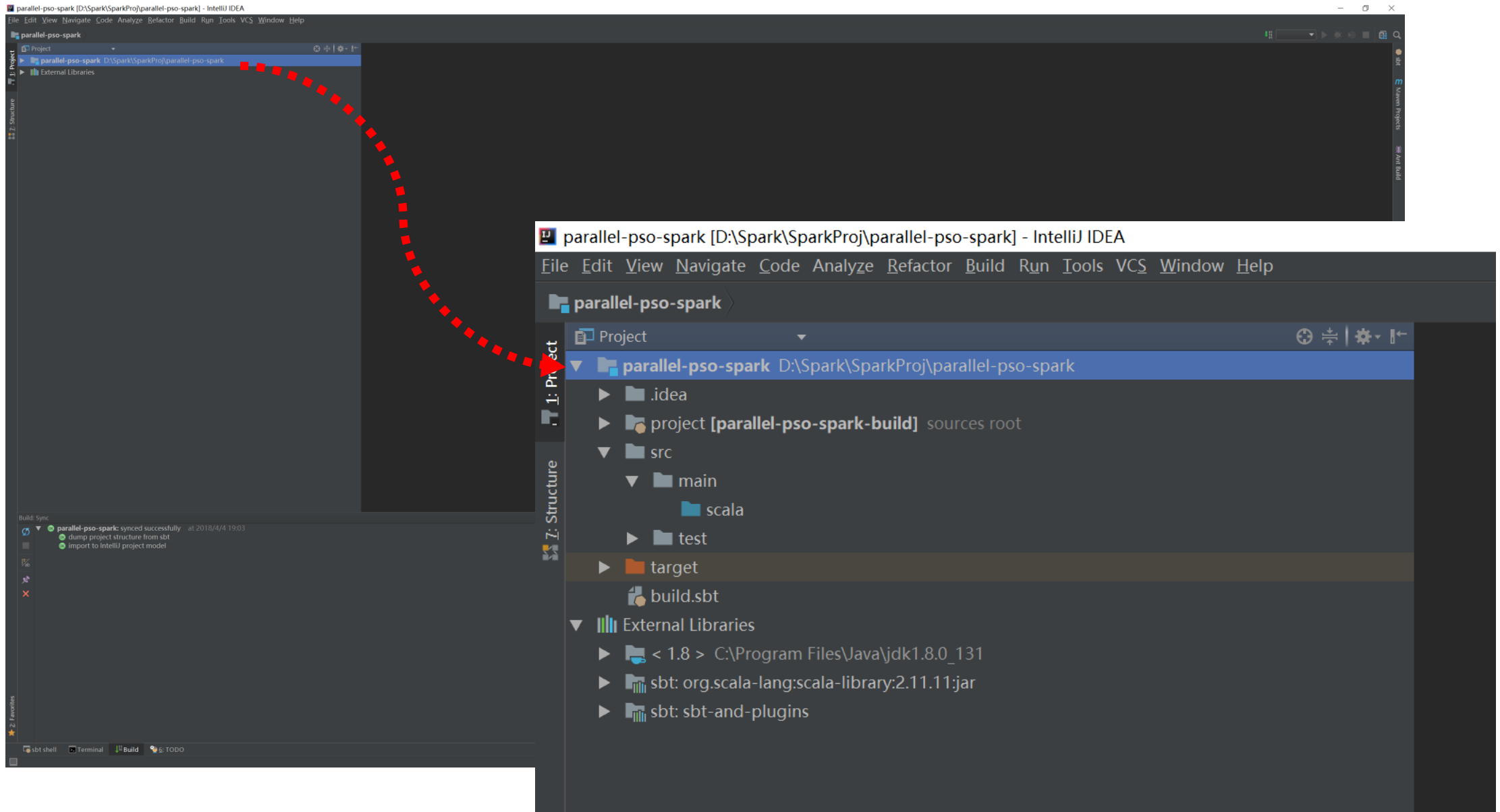
Use IntelliJ IDEA

- Create a Spark application: configure the Java, sbt, and Scala version



Use IntelliJ IDEA

- Create a Spark application: show the UI when starting the project



Use IntelliJ IDEA

- Download all the sbt library dependencies from the website <https://mvnrepository.com/> for the Spark application

[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)

[Maven](#)

[Gradle](#)

[SBT](#)

[Ivy](#)

[Grape](#)

[Leiningen](#)

[Buildr](#)

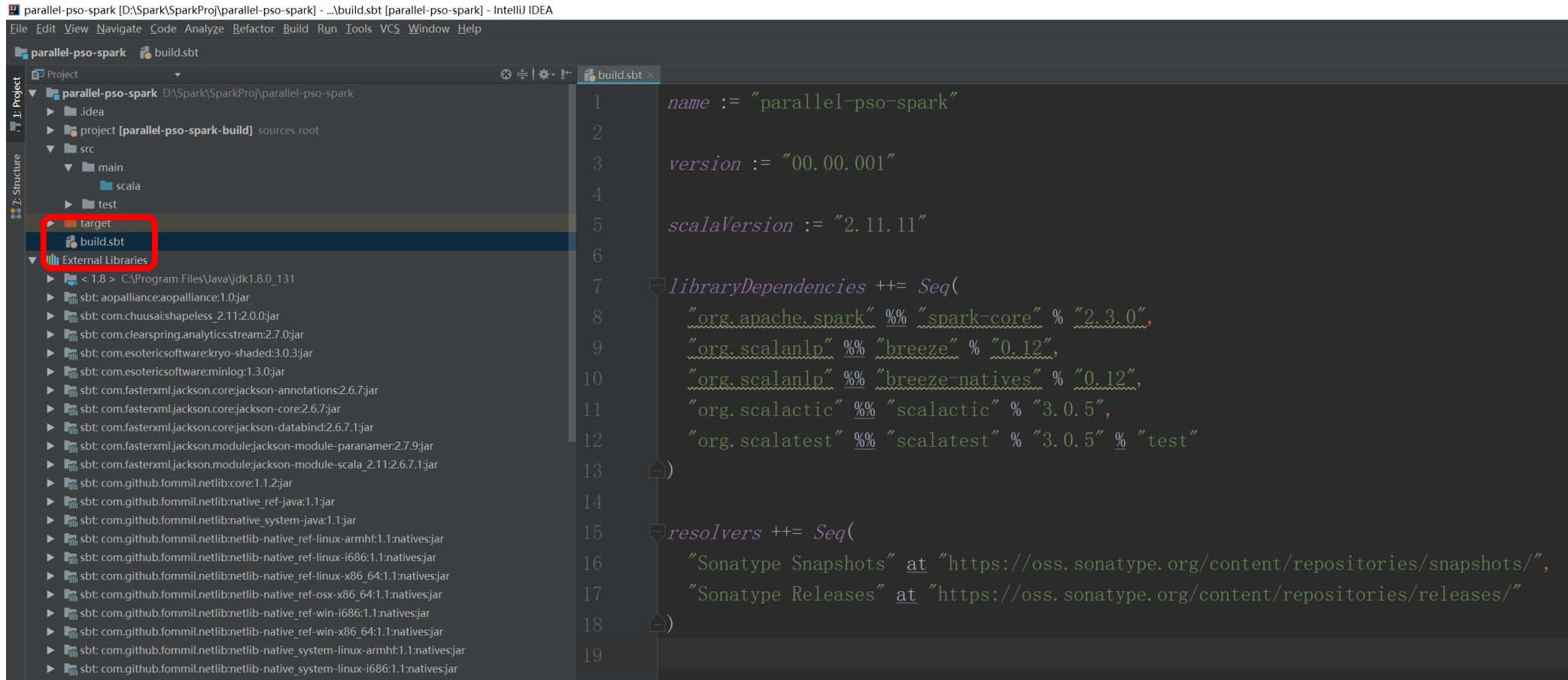
```
// https://mvnrepository.com/artifact/org.apache.spark/spark-core
libraryDependencies += "org.apache.spark" %% "spark-core" % "2.3.0"
```

☒ Include comment with link to declaration

Use IntelliJ IDEA

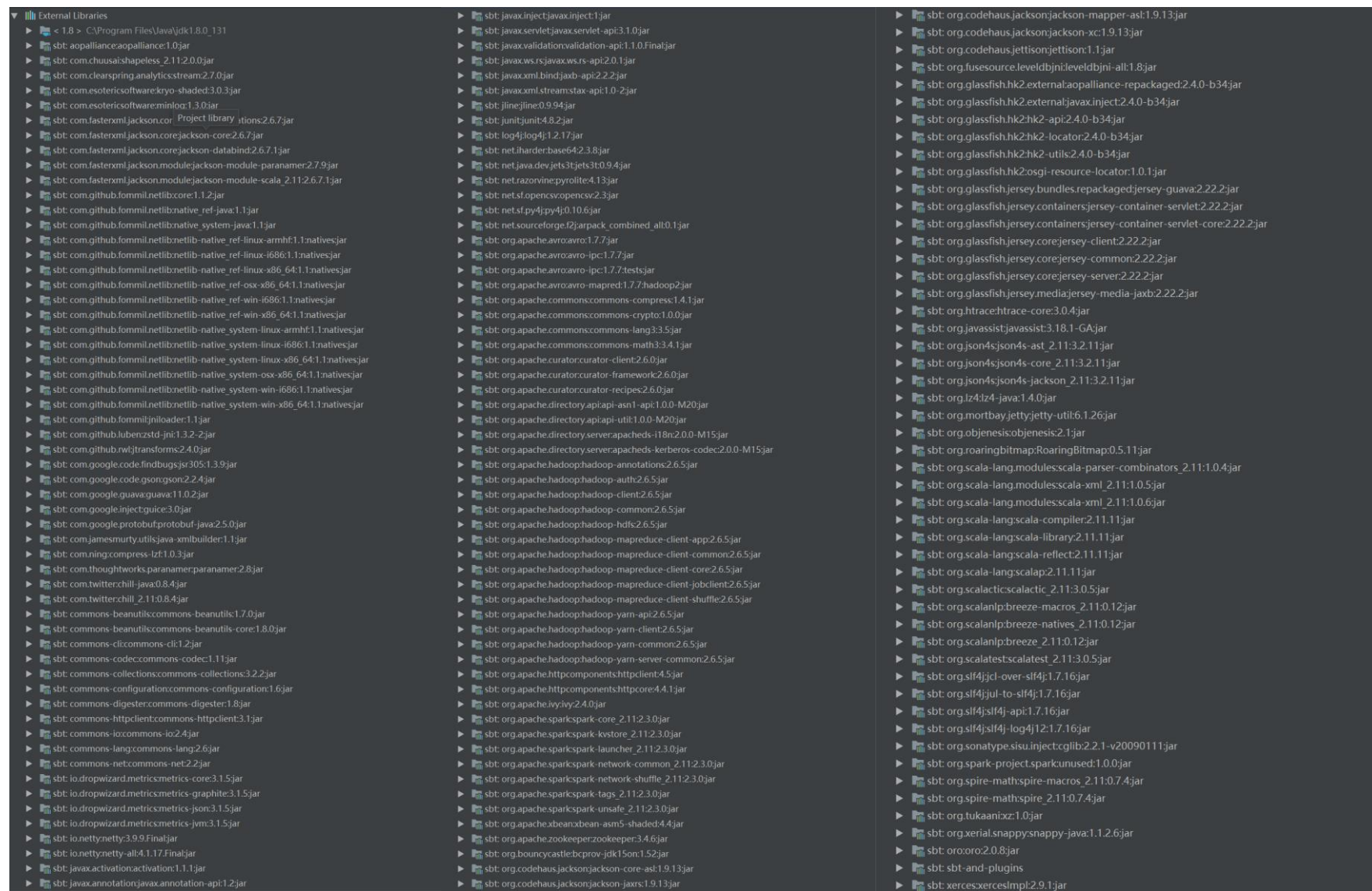
- Create a Spark application: update the *build.sbt* file, which can automatically load all the library dependencies

note that you may need wait a long time, which depends on the network speed



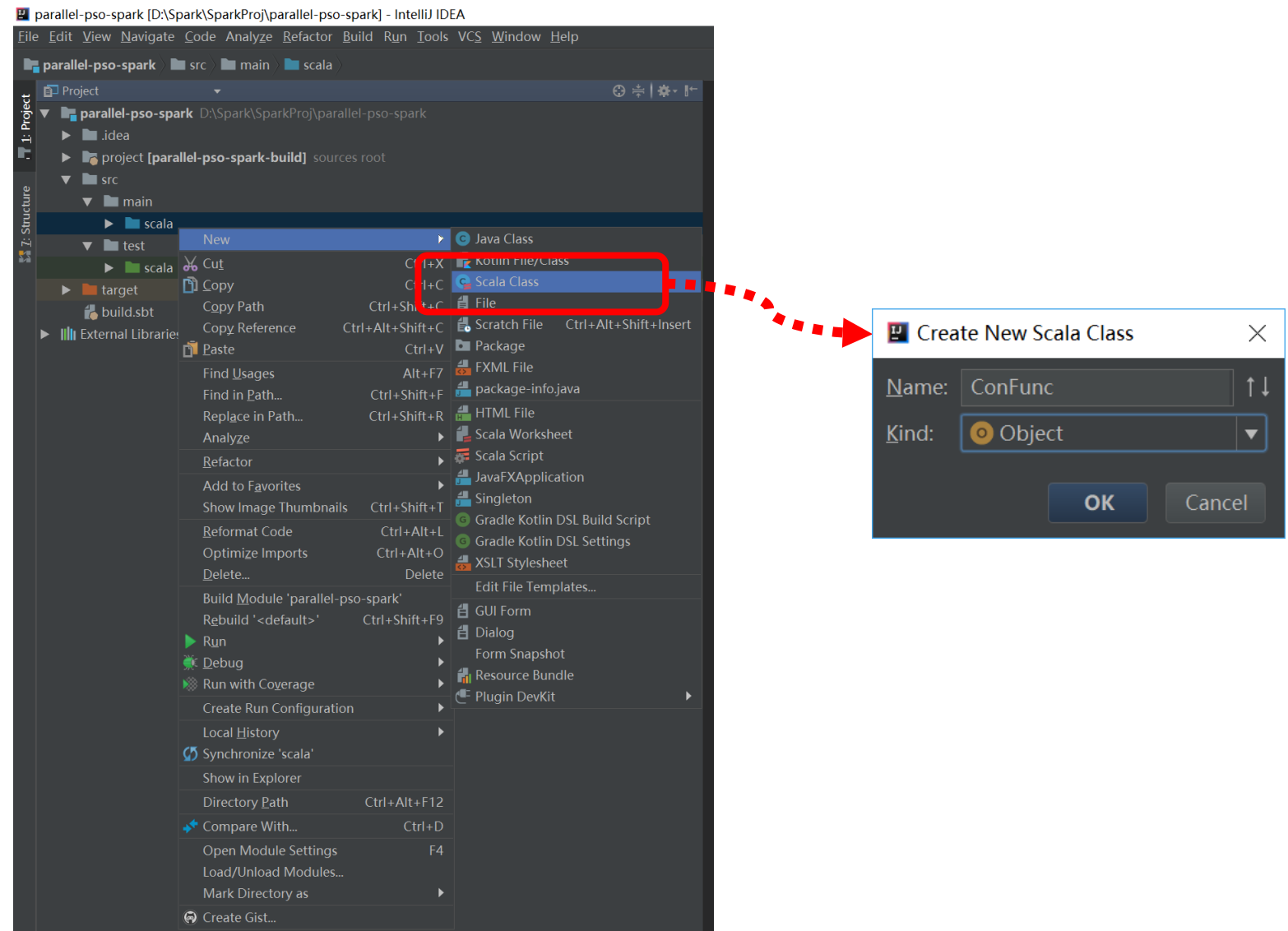
Use IntelliJ IDEA

- **Create a Spark application: check all the library dependencies automatically loaded**



Use IntelliJ IDEA

- Create a Spark application: create *Object* (rather than *class*)



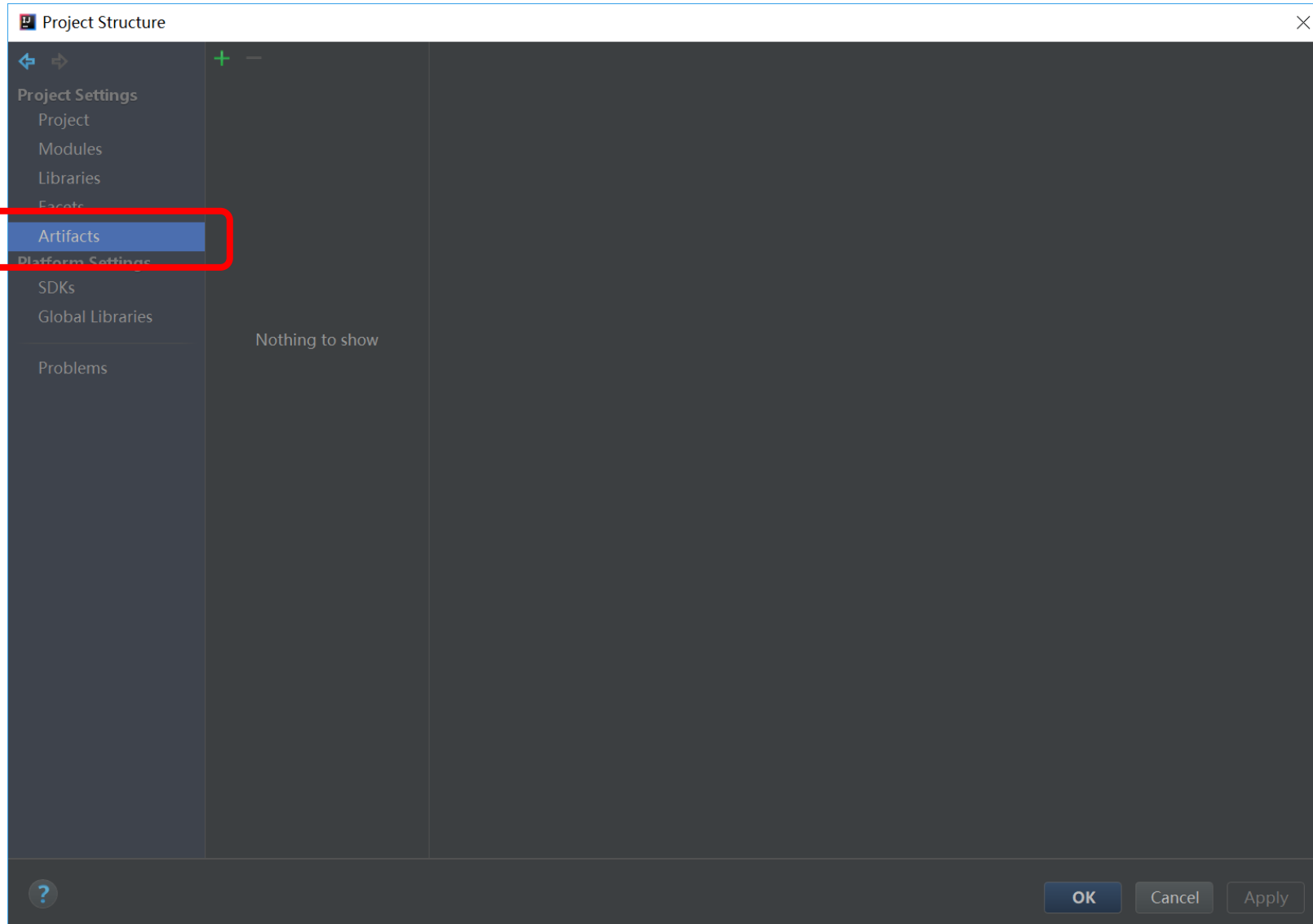
Use IntelliJ IDEA

- **Create a Spark application:** create a simple *demo* program to show how to connect the Spark commodity cluster

```
1 package peas
2
3 import org.apache.spark. {SparkConf, SparkContext}
4
5 /**
6  * Demo to Connect the Spark Commodity Cluster.
7  */
8 object DemoConnectSparkCluster extends App {
9     // configure and start the connection to the Spark commodity cluster
10    val sparkConf = new SparkConf()
11        .setAppName("DemoConnectSparkCluster")
12        .setMaster("spark://dc001.syhlab:7077")
13        .setJars(List("D:\\Spark\\SparkProj\\parallel-pso-spark\\out\\artifacts\\parallel_pso_spark_jar\\parallel-pso-spark.jar"))
14    val sc = new SparkContext(sparkConf)
15
16    // calculate the value of PI
17    val numSamples = 1000000000
18    val count = sc.parallelize(1 to numSamples, 100)
19        .filter({_ =>
20        val x = math.random
21        val y = math.random
22        x * x + y * y < 1
23    }).count()
24    println(f"PI ~= ${4.0 * count / numSamples}%9.7f")
25
26    // close the connection to the Spark commodity cluster
27    sc.stop()
28 }
```

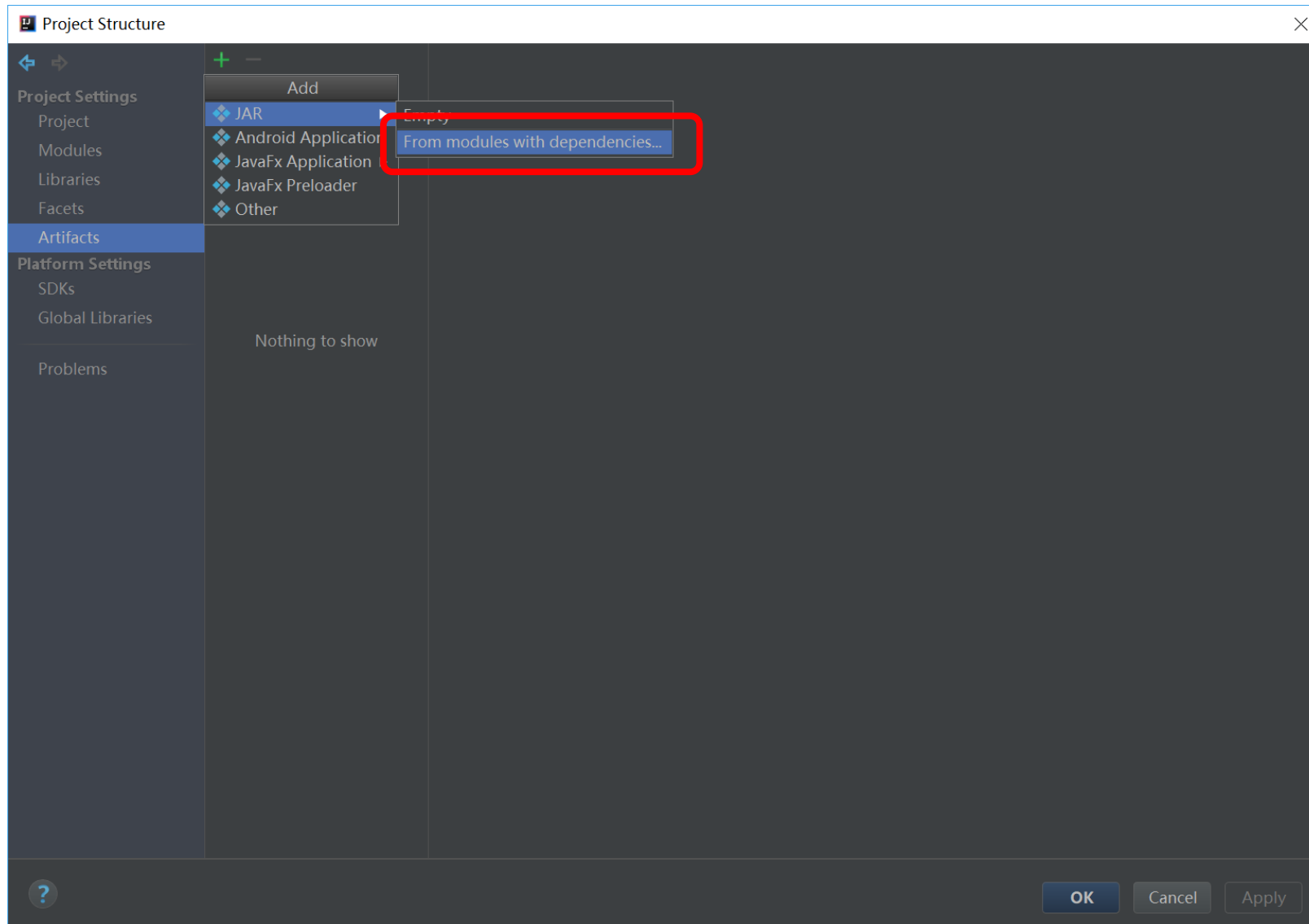
Use IntelliJ IDEA

- Create a Spark application: configure the project



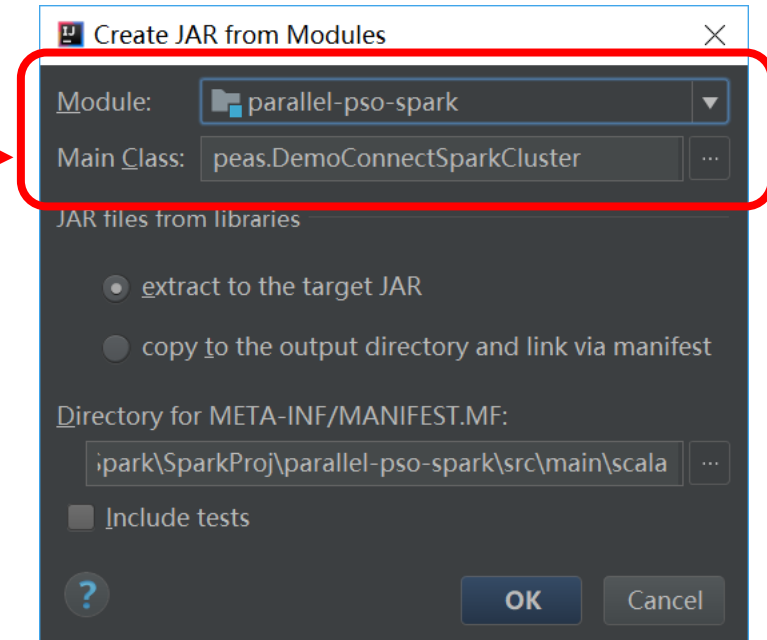
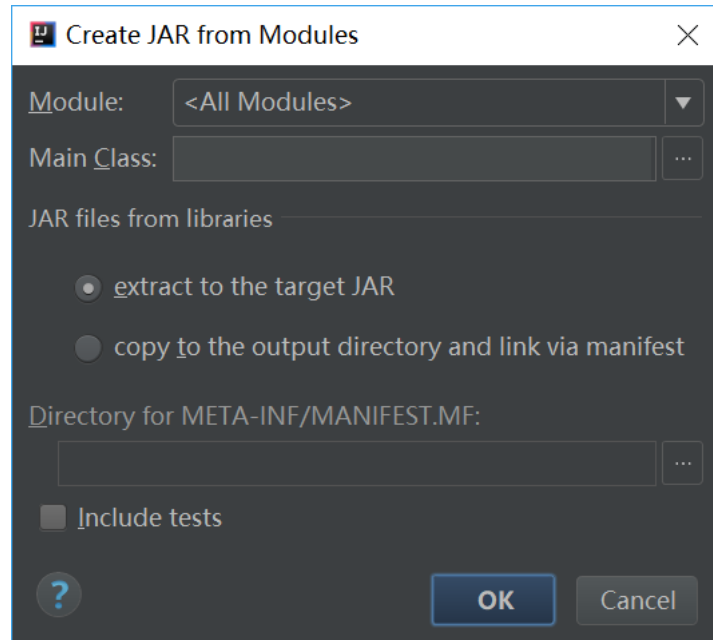
Use IntelliJ IDEA

- Create a Spark application: configure the project



Use IntelliJ IDEA

- Create a Spark application: configure the project

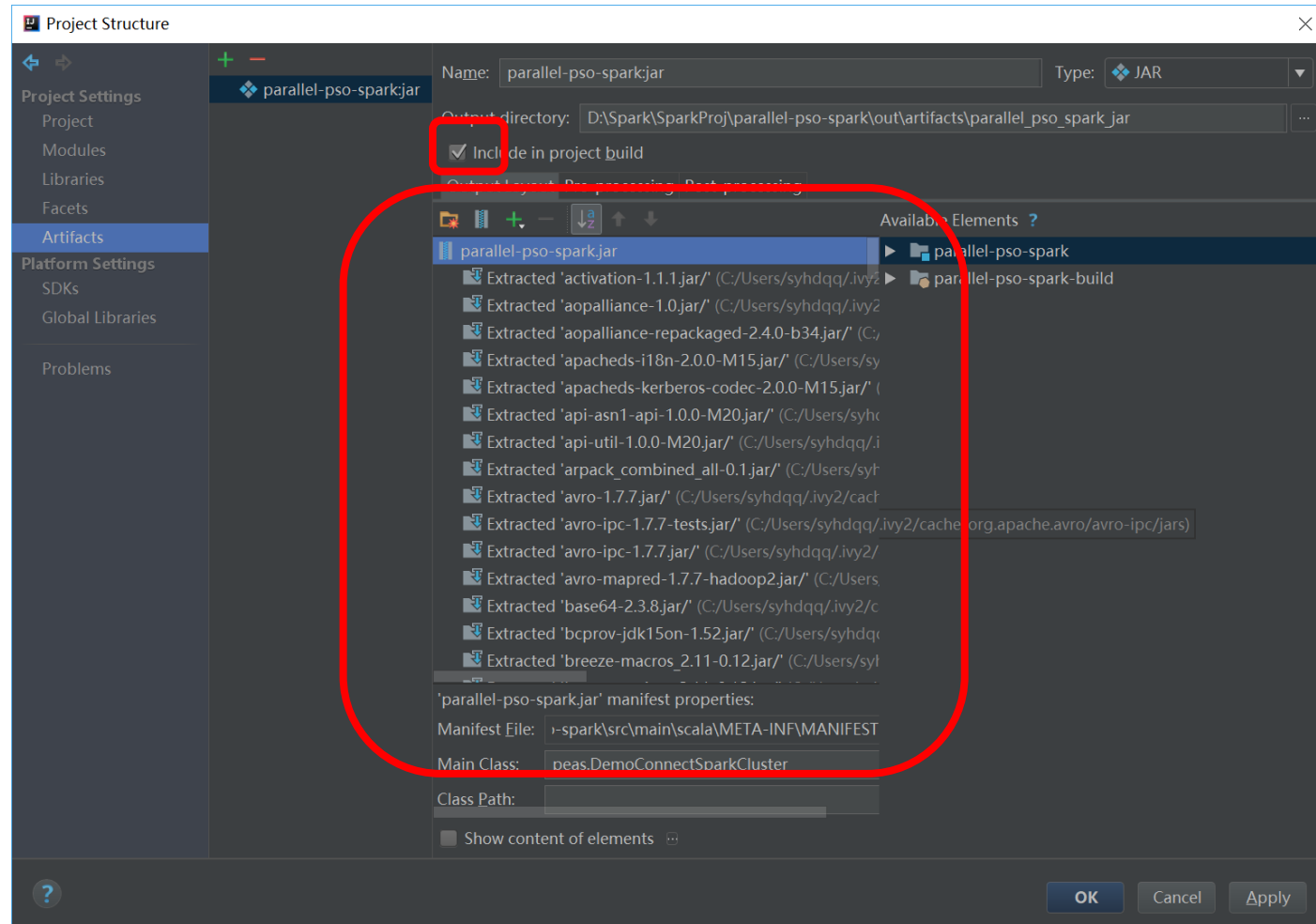


Use IntelliJ IDEA

- Create a Spark application: configure the project

If you can't correctly configure the settings, the below error will be thrown:

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

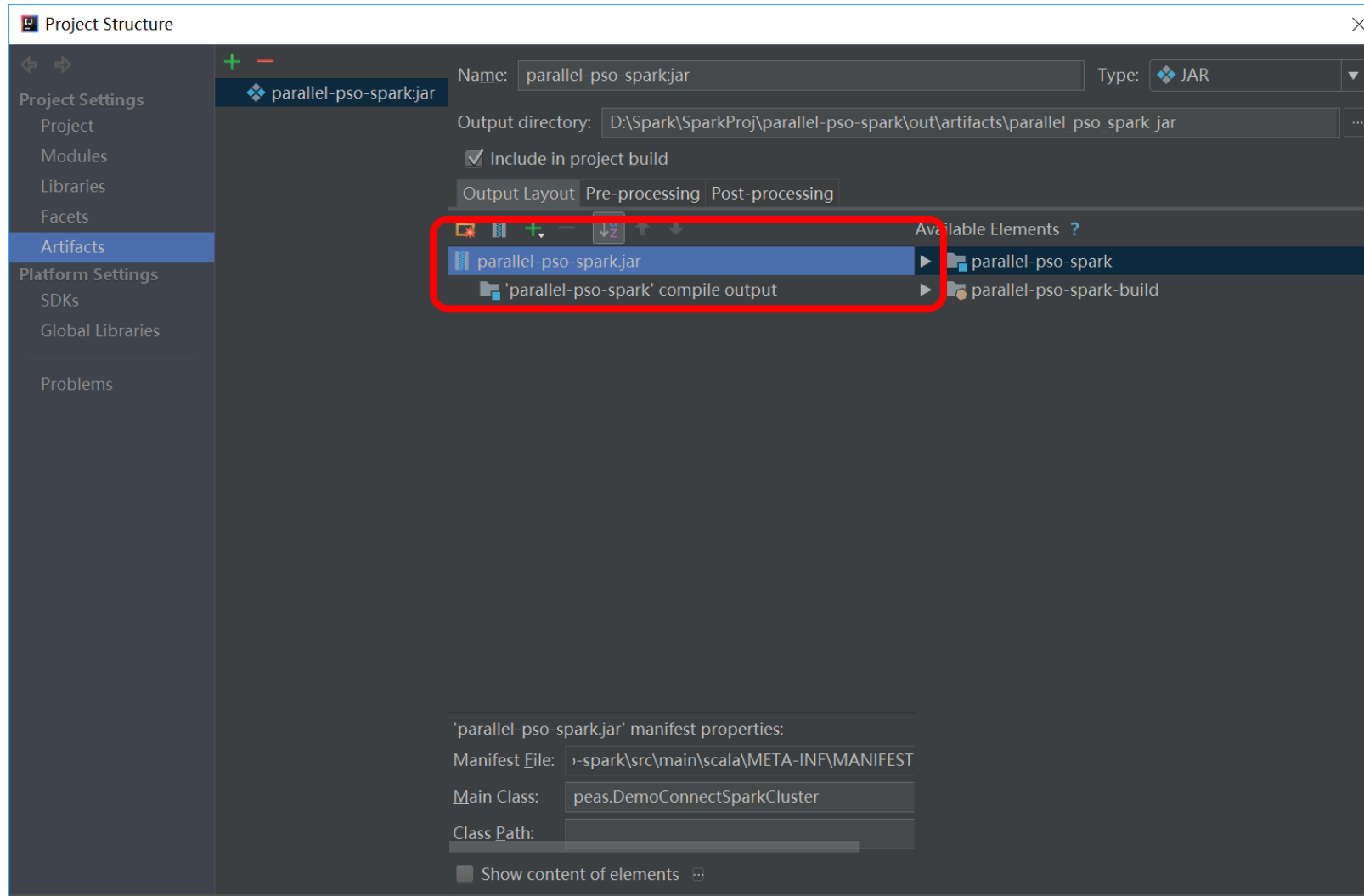


Use IntelliJ IDEA

- Create a Spark application: configure the project

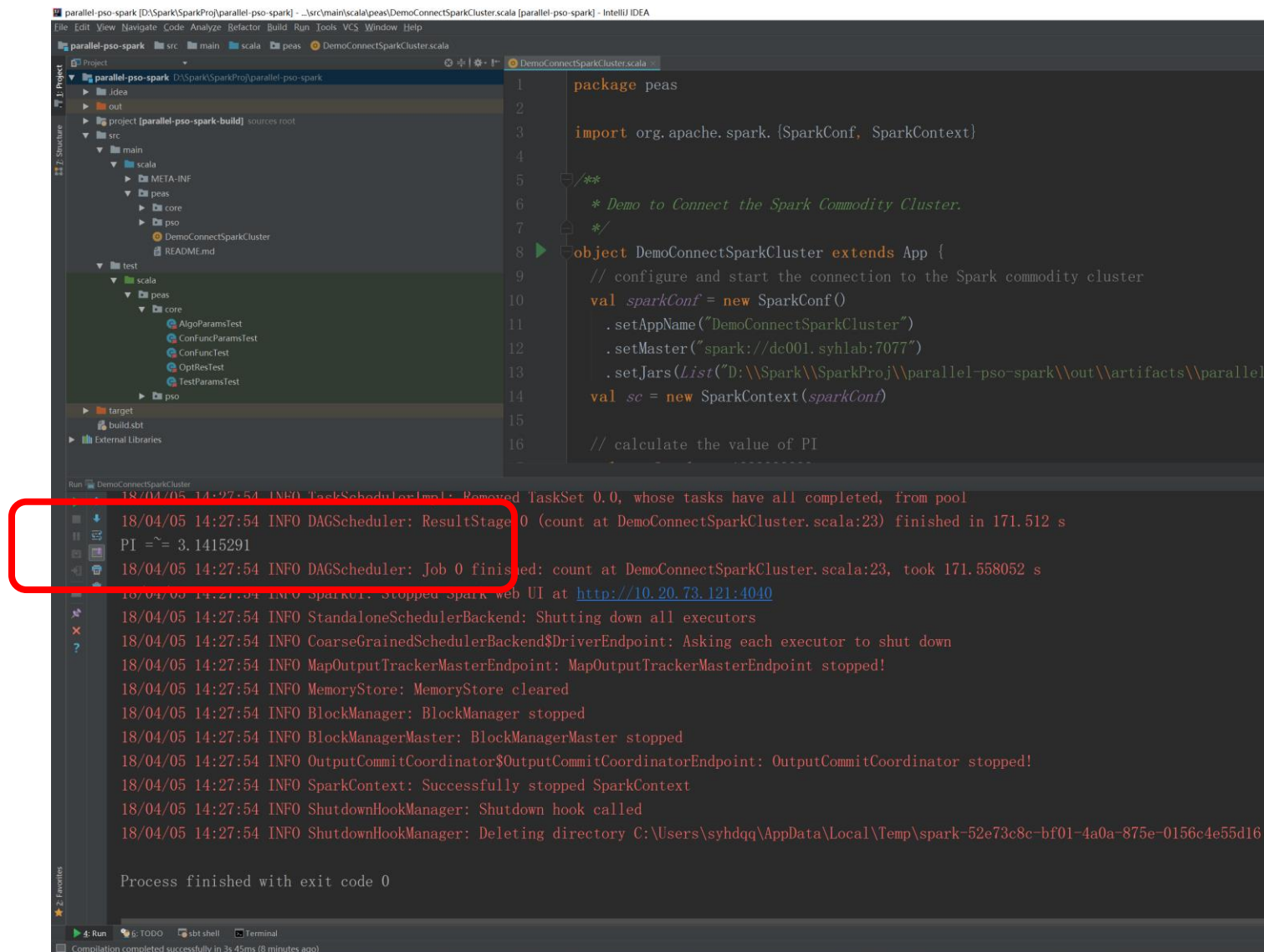
If you can't correctly configure the settings, the below error will be thrown:

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



Use IntelliJ IDEA

- Create a Spark application: show the output of the *demo* program



The screenshot displays the IntelliJ IDEA interface. On the left, the Project Structure view shows a project named 'parallel-pso-spark' with a source root 'src/main/scala' containing a package 'peas'. The 'DemoConnectSparkCluster' object is highlighted. The main editor shows the Scala code for 'DemoConnectSparkCluster.scala', which defines an object extending 'App' that configures Spark and calculates the value of PI. The bottom panel shows the Run output for 'DemoConnectSparkCluster', with a red box highlighting the first three lines of the log:

```
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
```

The status bar at the bottom indicates 'Compilation completed successfully in 3s 45ms (8 minutes ago)'.

Use IntelliJ IDEA

- Create a Spark application: monitor the Spark commodity cluster via *WEB UI*

Spark Master at spark://dc001:7077

URL: spark://dc001:7077
REST URL: spark://dc001:8080 (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