

2015-09-01 · BIG DATA

## **View Spark Source in Eclipse**

Reading source code is a great way to learn opensource projects. I used to read Java projects' source code on GrepCode for it is online and has very nice cross reference features. As for Scala projects such as Apache Spark, though its source code can be found on GitHub, it's quite necessary to setup an IDE to view the code more efficiently. Here's a howto of viewing Spark source code in Eclipse.

## Install Eclipse and Scala IDE Plugin

One can download Eclipse from here. I recommend the "Eclipse IDE for Java EE Developers", which contains a lot of daily-used features.

#### TAG CLOUD

algorithm analytics apache beam canal clojure crossfilter dc.js elasticsearch es6 etl flume frontend functional programming hbase hive java javascript kafka lodash mapreduce mysql ops pandas python scala scalatra source code spark spark streaming sql stream processing thrift vue vuex webjars websocket



Then go to Scala IDE's official site and install the plugin through update site or zip archive.

## **Generate Project File with Maven**

Spark is mainly built with Maven, so make sure you have Maven installed on your box, and download the latest Spark source code from here, unarchive it, and execute the following command:

\$ mvn -am -pl core dependency:resolve eclipse:eclipse

This command does a bunch of things. First, it indicates what modules should be built. Spark is a large project with multiple modules. Currently we're only interested in its core module, so <code>-pl</code> or <code>--projects</code> is used. <code>-</code> am or <code>--also-make</code> tells Maven to build core module's dependencies as well. We can see the module list in output:

#### **ARCHIVES**

April 2018

October 2017

September 2017

August 2017

July 2017

June 2017

March 2017

January 2017

January 2017

September 2015

May 2015

April 2015

May 2014

October 2013

April 2013

#### RECENT POSTS

Connect HBase with Python and Thriff

Form Handling in Vuex Strict Mode

Error Handling in RESTful API

Flume Source Code: Component

Lifecycle

Pandas and Tidy Data

[INFO] Spark Project Shuffle Streaming Service

[INFO] Spark Project Networking

```
[INFO] Spark Project Unsafe
      [INFO] Spark Project Core
  dependency: resolve tells Maven to download all dependencies. eclipse: eclipse will generate the
  .project and .classpath files for Eclipse. But the result is not perfect, both files need some fixes.
 Edit core/.classpath, change the following two lines:
     <classpathentry kind="src" path="src/main/scala" including="**/*.java"/>
     <classpathentry kind="src" path="src/test/scala" output="target/scala-2.10/test-cl</pre>
∢ .....
 to
     <classpathentry kind="src" path="src/main/scala" including="**/*.java|**/*.scala"</pre>
     <classpathentry kind="src" path="src/test/scala" output="target/scala-2.10/test-cl</pre>
 Edit core/.project, make it looks like this:
     <buildSpec>
       <buildCommand>
         <name>org.scala-ide.sdt.core.scalabuilder
       </buildCommand>
     </buildSpec>
     <natures>
       <nature>org.scala-ide.sdt.core.scalanature
```

```
8 <nature>org.eclipse.jdt.core.javanature</nature>
9 </natures>
```

Now you can import "Existing Projects into Workspace", including core, launcher, network, and unsafe.

#### **Miscellaneous**

### Access restriction: The type 'Unsafe' is not API

For module spark-unsafe, Eclipse will report an error "Access restriction: The type 'Unsafe' is not API (restriction on required library /path/to/jre/lib/rt.jar". To fix this, right click the "JRE System Library" entry in Package Explorer, change it to "Workspace default JRE".

#### **Download Sources and Javadocs**

Add the following entry into pom's project / build / plugins:

## build-helper-maven-plugin

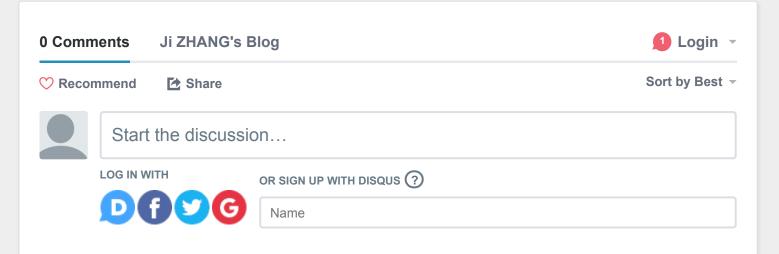
Since Spark is a mixture of Java and Scala code, and the maven-eclipse-plugin only knows about Java source files, so we need to use build-helper-maven-plugin to include the Scala sources, as is described here. Fortunately, Spark's pom.xml has already included this setting.

#### References

- http://docs.scala-lang.org/tutorials/scala-with-maven.html
- https://wiki.scala-lang.org/display/SIW/ScalaEclipseMaven
- https://cwiki.apache.org/confluence/display/SPARK/Useful+Developer+Tools

Python 2 to 3 Quick Guide

Spark Streaming Logging Configuration



Be the first to comment.

**ALSO ON JI ZHANG'S BLOG** 

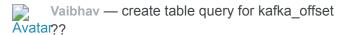
#### 使用 WebSocket 和 Python 编写日志查看器

1 comment • 9 months ago



# **How to Achieve Exactly-Once Semantics in Spark Streaming**

2 comments • 9 months ago



Clojure实战(4):编写Hadoop MapReduce脚

本

1 comment • 5 years ago



#### Nginx热升级 - Ji ZHANG's Blog

1 comment • 5 years ago

Ieon8693 — 好东西,赞一下,老板,把你链接 Avatam 我blog了。







**DISQUS** 



© 2018 Ji ZHANG Powered by Hexo

