Apache Spark with Scala – Course Materials

Install the Course Materials

The scripts and data for this course may be downloaded at

**<https://github.com/PacktPublishing/Apache-Spark-with-Scala---Hands-On-with-Big-Data-/tree/master/SparkScalaCourse/SparkScalaCourse>**

Download and un-zip this file and move the SparkScalaCourse folder (which contains another SparkScalaCourse folder) to a path you’ll remember.

Next, download the MovieLens 100K dataset from:

<https://github.com/PacktPublishing/Apache-Spark-with-Scala---Hands-On-with-Big-Data-/tree/master/ml-100k>

Unzip it and **move the resulting ml-100k folder into your SparkScalaCourse/data folder.**

Install IntelliJ and Apache Spark

Make sure you have **JDK 8 or 11** installed. **Apache Spark is not compatible with Java 16**. Enter

java -version

from a command or terminal prompt to see what version, if any, you have installed already. If you need to get it, [download the JDK from Oracle](https://www.oracle.com/java/technologies/javase-jdk14-downloads.html) (you’ll need to create an account with them first.)

Next, install [IntelliJ IDEA Community Edition](https://www.jetbrains.com/idea/download), after selecting your platform (Windows, Mac, or Linux). During the installation process, choose to **install the Scala plugin** for IntelliJ, using **Scala 2.12**. If you are not prompted to do so while installing, choose the “Configure” dropdown menu from the IntelliJ welcome screen, then select “plugins,” and add the Scala plugin from there.

Also from the “Configure” dropdown menu, select “Structure for new projects” and confirm a valid JDK is selected.

**WINDOWS ONLY:** Create a C:\hadoop\bin directory and copy the winutils.exe file found inside your SparkScalaCourse folder into it. Create a new environment variable (enter “environment variables” in the Windows search bar, click on “Add Environment Variables,” and add a new system variable) named HADOOP\_HOME with a value of C:\Hadoop. Next select the PATH environment variable, and APPEND a new entry, separated by a semi-colon, of %HADOOP\_HOME%\bin Now, restart IntelliJ to make sure the new environment variables are picked up.

Import the Course Project

From the IntelliJ welcome screen, select “**Open or Import**“

Select your **SparkScalaCourse/SparkScalaCourse** folder.

Try it Out

Expand the project’s tree view to show the SparkScalaCourse/src/main/scala/com/sundogsoftware/spark folder.

Right click on “HelloWorld” and select “Run HelloWorld”

You should see a message like:

Hello world! The u.data file has 100000 lines.

But, you might see a “class not found” error. If so, just quit IntelliJ, restart it, and try again. It’s just a bug in IntelliJ.

Once you see the “Hello World” message, everything is set up successfully! If not, go back and look for a step you may have missed. Sometimes IntelliJ just gets confused – you might need to refresh the SBT configuration as shown in the setup video, or even re-add the dstream-twitter library.