Big Data

(Apache Spark)

Instructor: Trong-Hop Do

April 24th 2021

S3Lab
Smart Software System Laboratory


“Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming.”
– Chris Lynch, Vertica Systems

Apache Spark Installation on Windows

Install Java 8 or Later

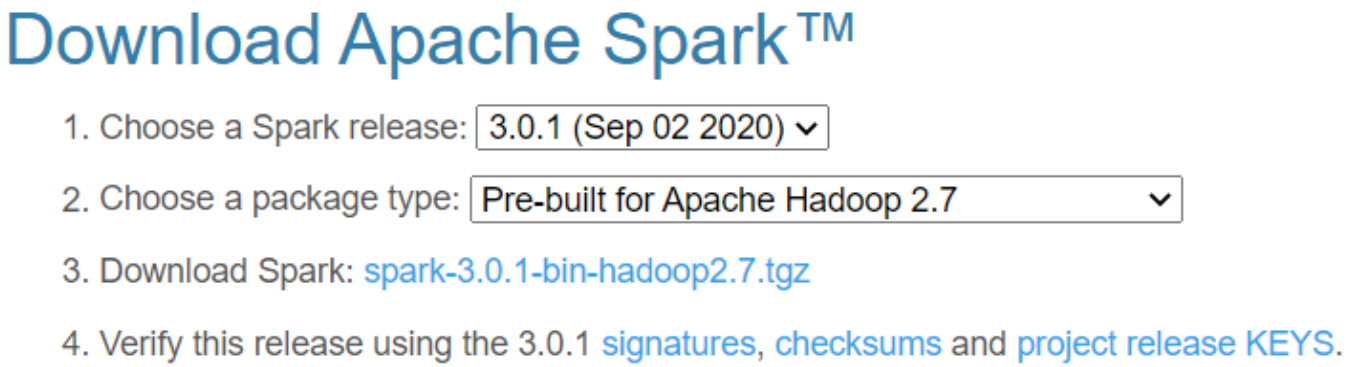
- To install Apache Spark on windows, you would need Java 8 or later version hence download the Java version from Oracle and install it on your system.

- https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html

Windows x64	166.79 MB	jdk-8u271-windows-x64.exe

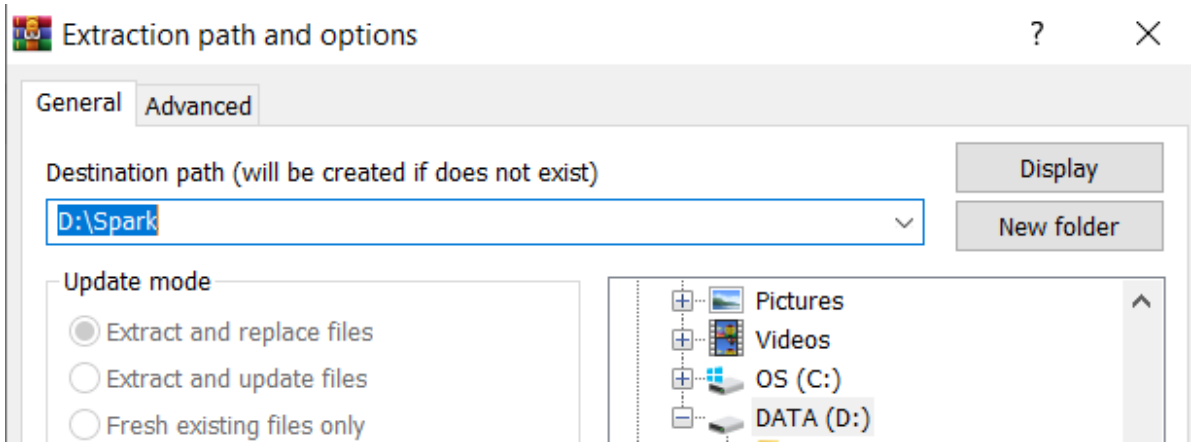
Apache Spark Installation on Windows

- Download Apache spark
- https://spark.apache.org/downloads.html



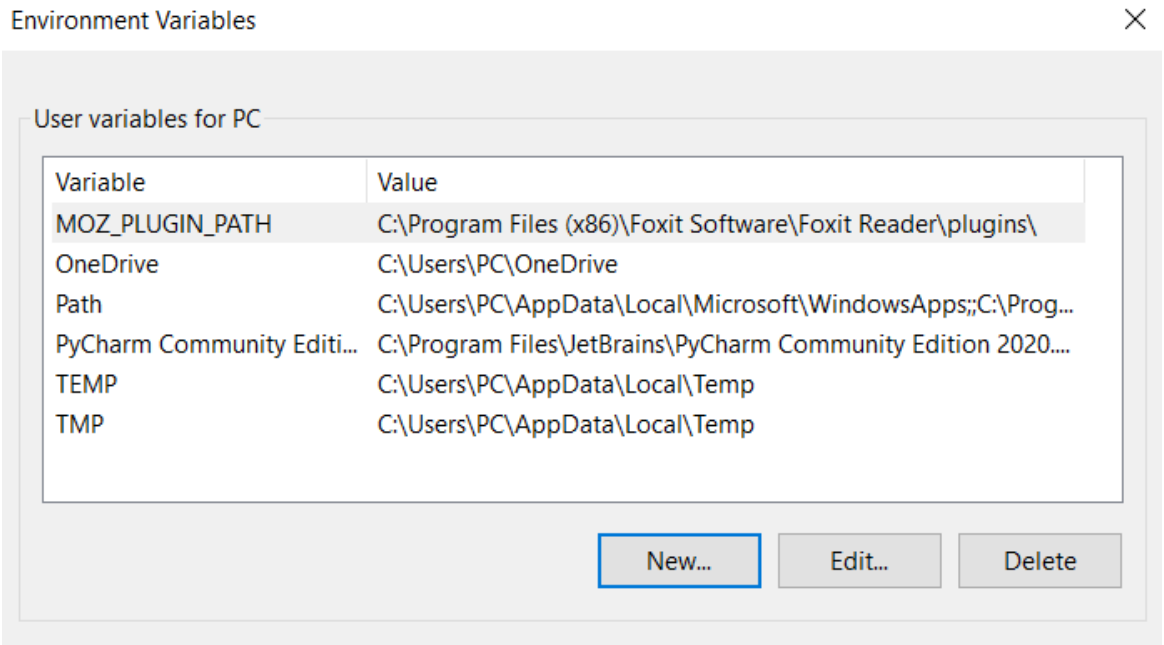
Apache Spark Installation on Windows

- Extract the zip file to any folder

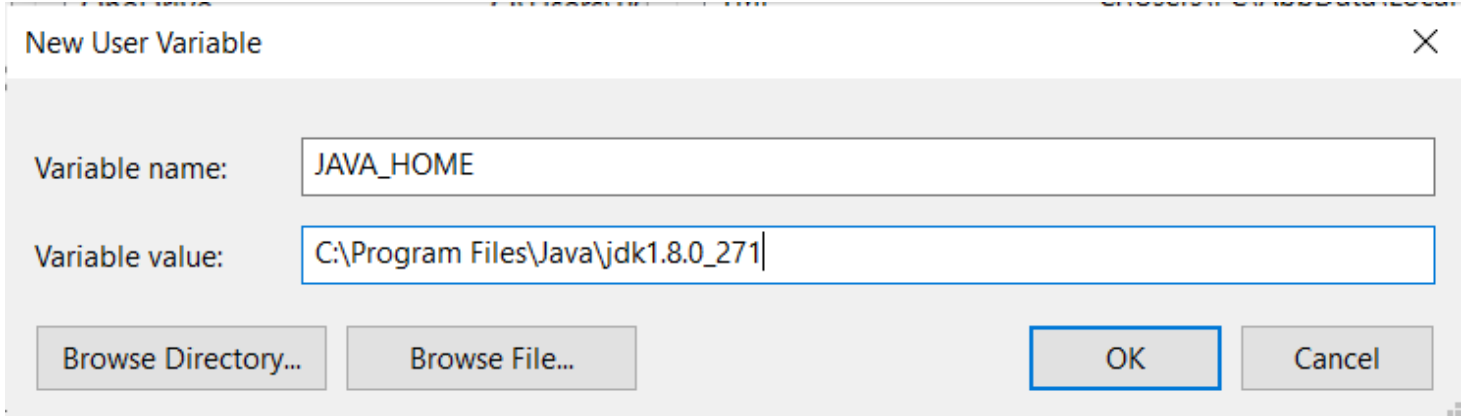


Environment Variables Setting

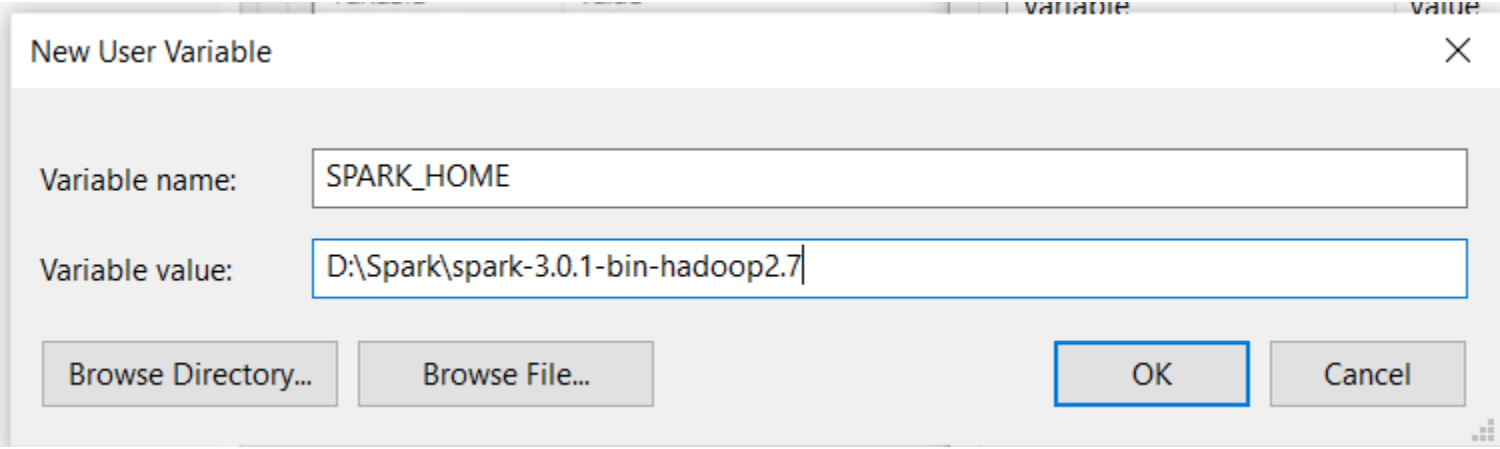
- Open System Environment Variables window and select Environment Variables.



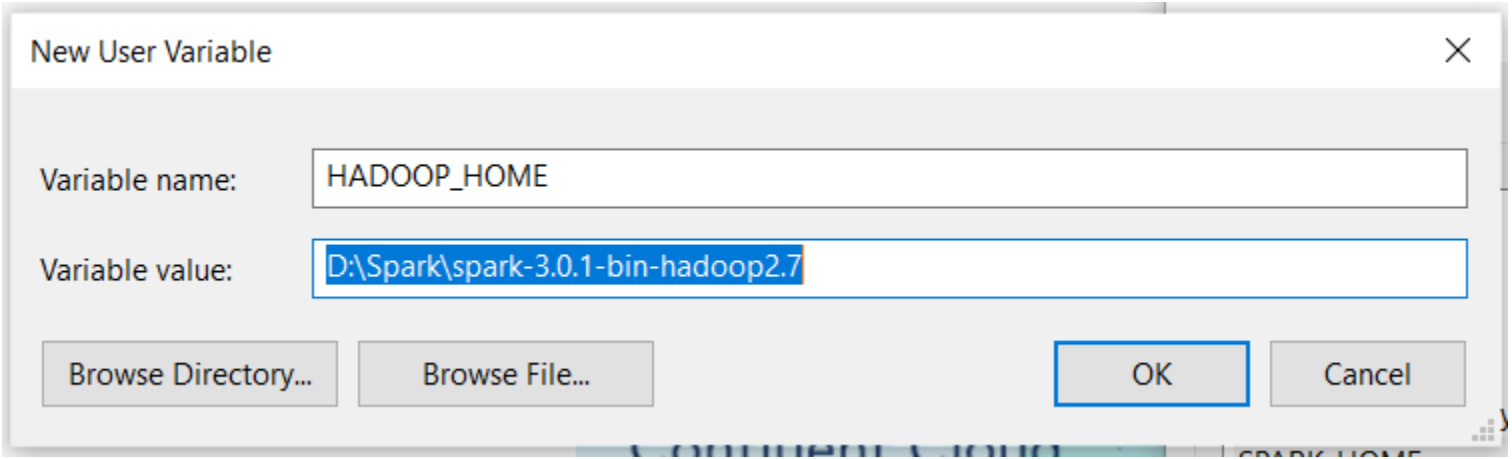
Environment Variables Setting



Apache Spark Installation on Windows

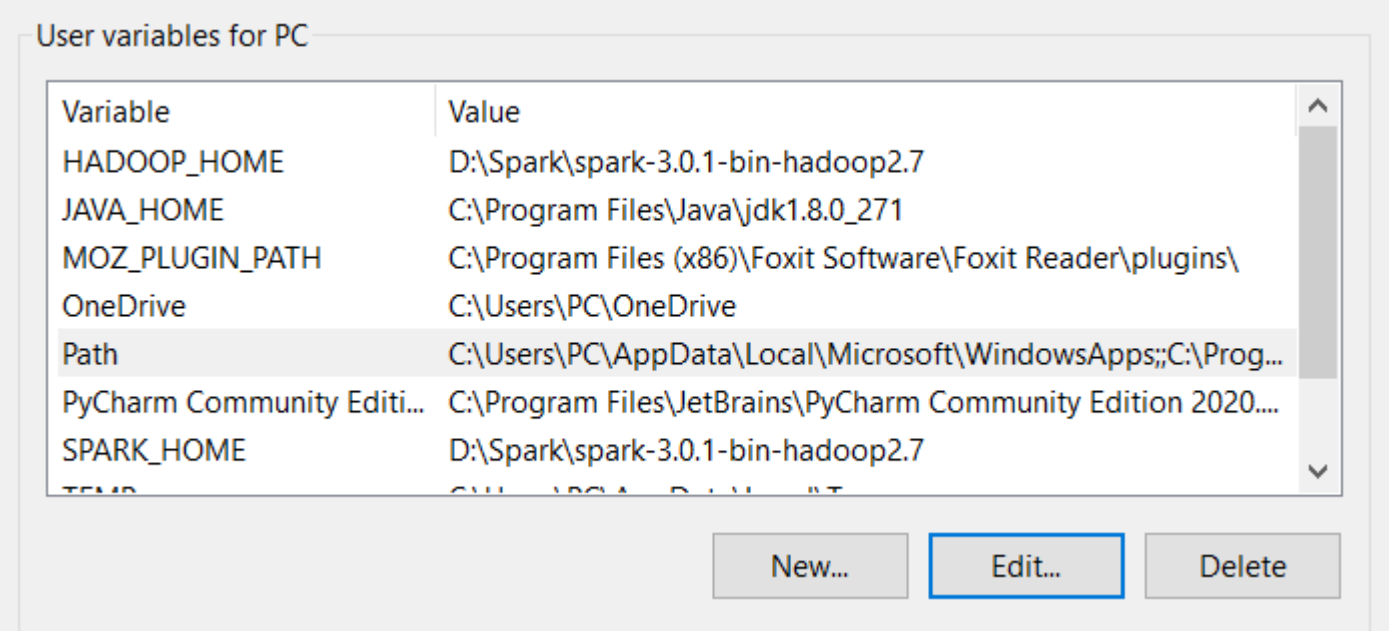


Apache Spark Installation on Windows



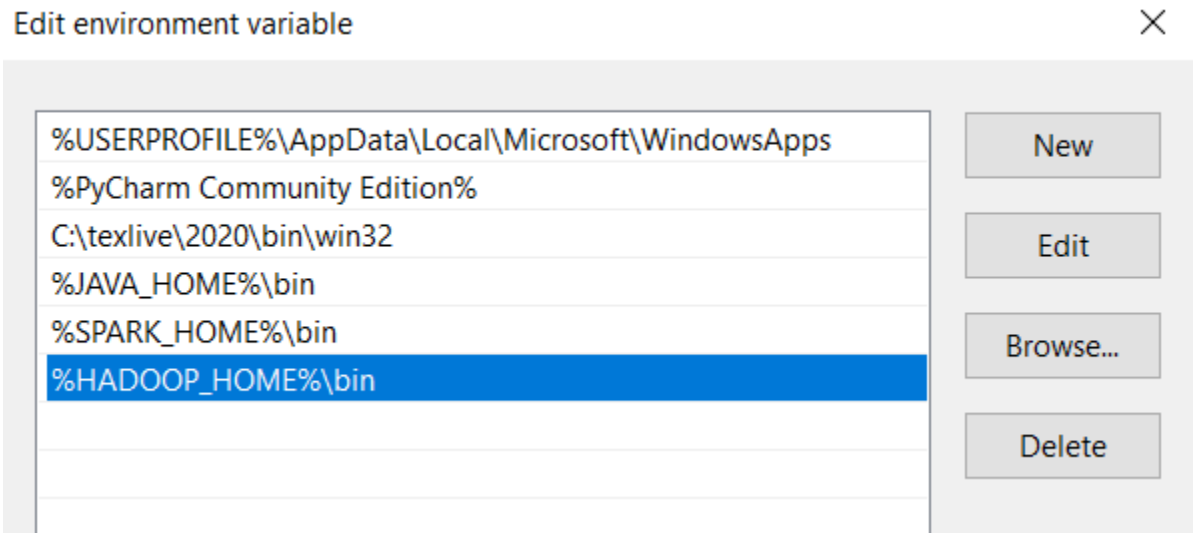
Apache Spark Installation on Windows

- Now Edit the PATH variable



Apache Spark Installation on Windows

- Add Spark, Java, and Hadoop bin location by selecting New option.



Test apache Spark shell

```
C:\Users\PC>spark-shell
20/12/07 16:50:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://192.168.56.1:4040
Spark context available as 'sc' (master = local[*], app id = local-1607334630777).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.0.1
      /_/

Using Scala version 2.12.10 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_271)
Type in expressions to have them evaluated.
Type :help for more information.

scala> 20/12/07 16:50:45 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of P
rocessTree metrics is stopped


scala>
```

Test apache Spark shell

- Type any command to test Spark shell

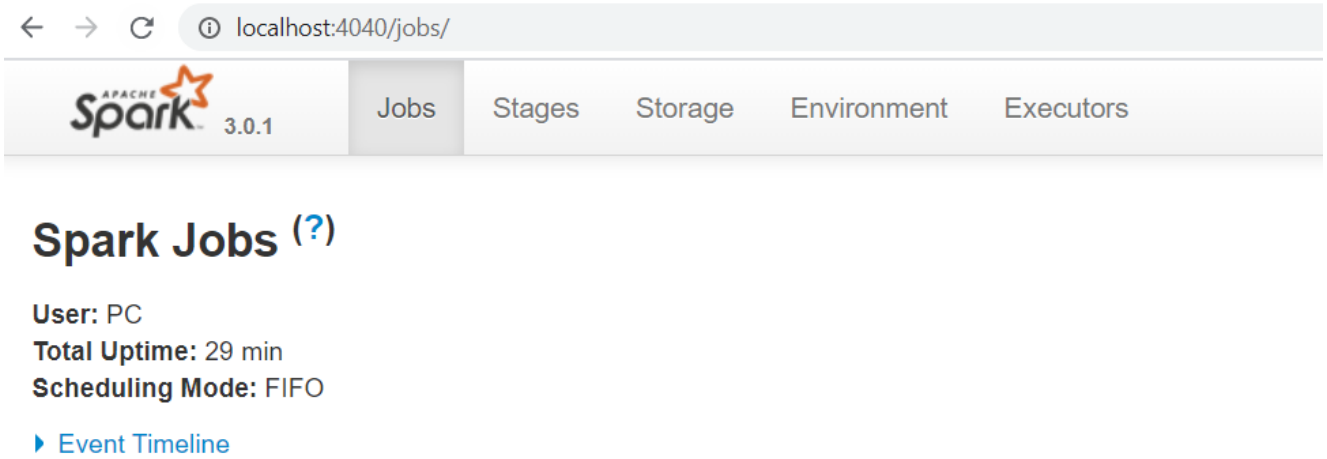
```
scala> spark.version
res0: String = 3.0.1

scala> 3+3
res1: Int = 6

scala>
```

Web UI on Windows

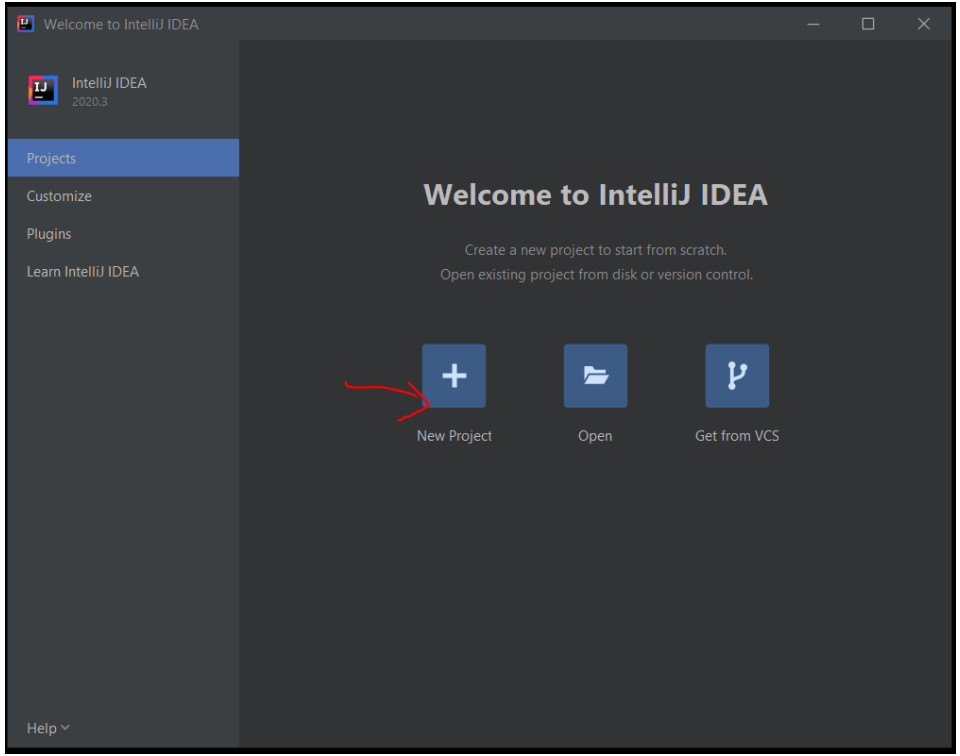
- Open a web browser and navigate to http://localhost:4040/



Spark Setup with Scala and Run in IntelliJ

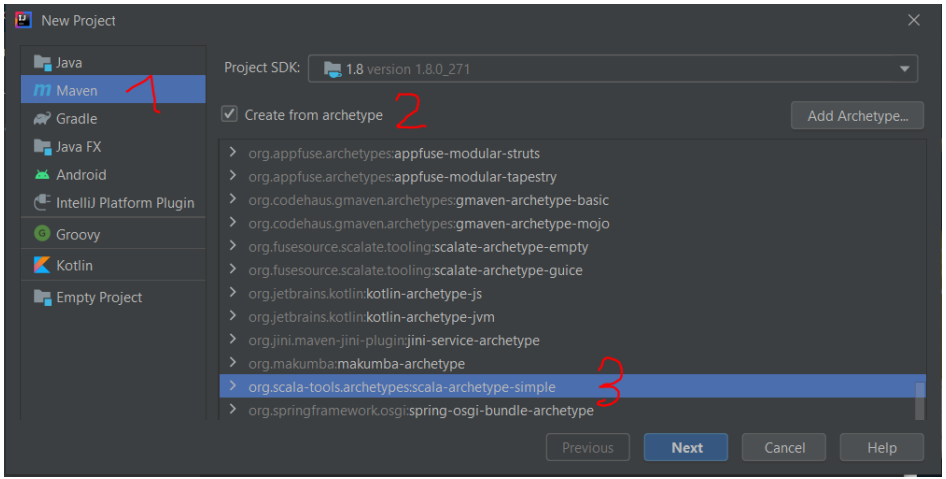
- download IntelliJ IDEA community edition
- https://www.jetbrains.com/idea/download/#section=windows

Create a Scala project In IntelliJ

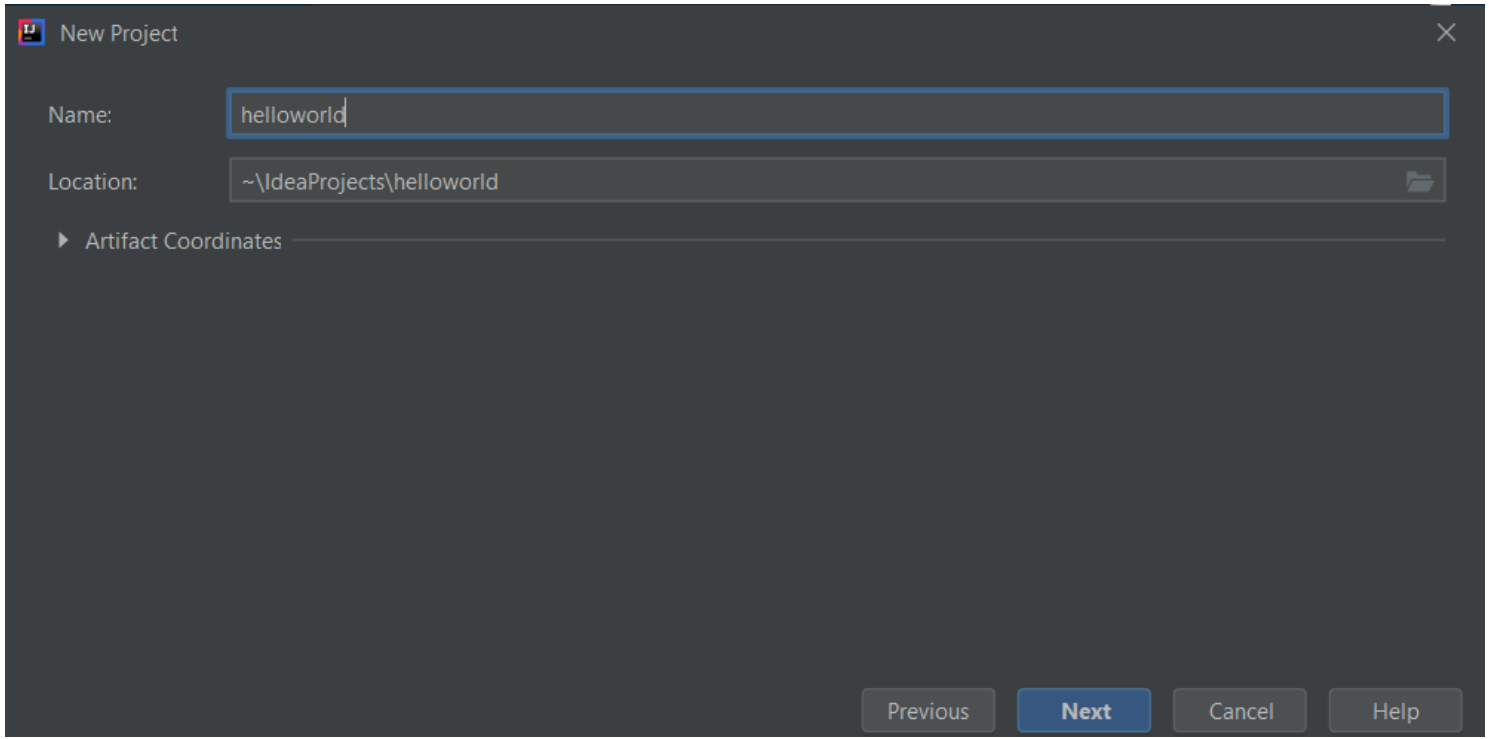


Create a Scala project In IntelliJ

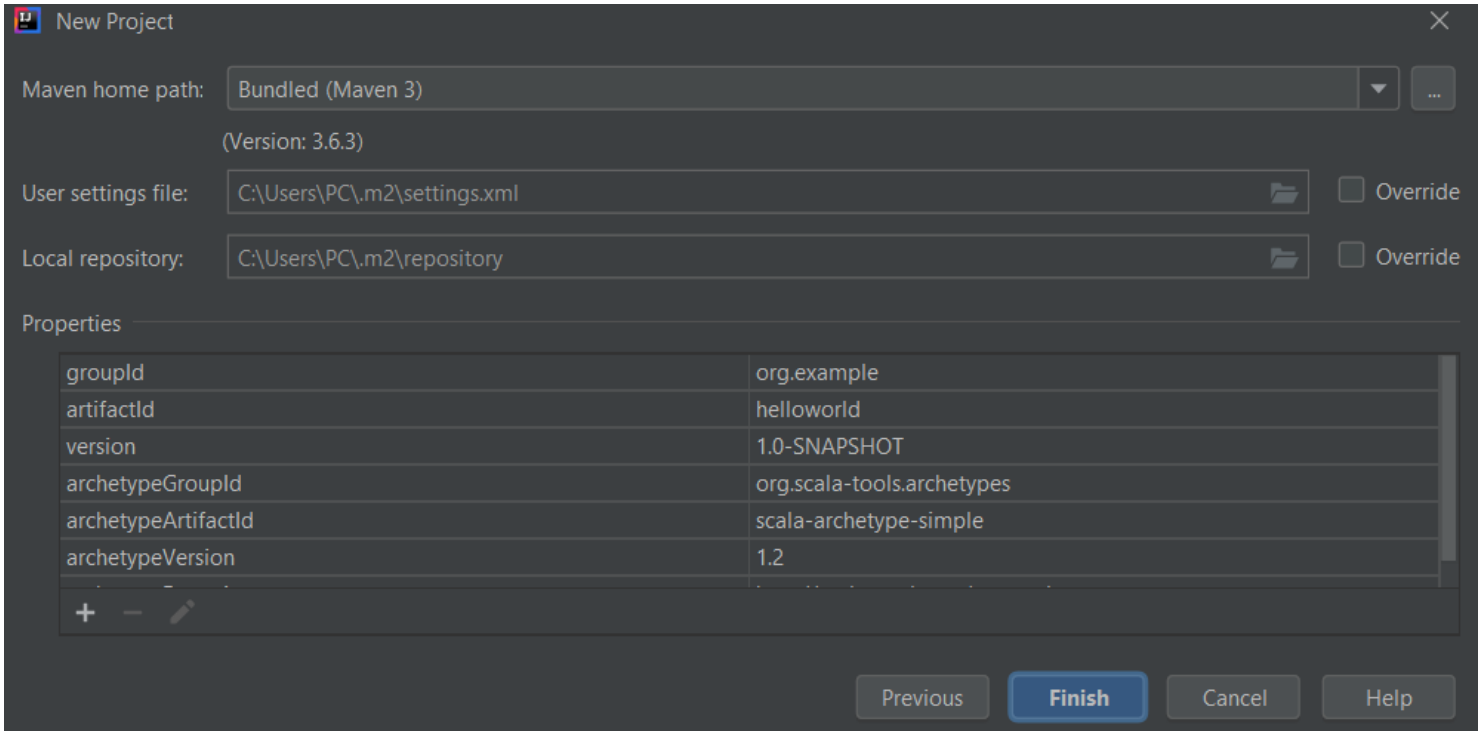
- Select Maven from the left panel
- Check option Create from archetype
- Select org.scala-tools.archetypes:scala-archetypes-simple.
    - The archetype is a kind of templates that creates the right directory structure and downloads the required default dependencies. Since we have selected Scala archetypes, it downloads all Scala dependencies and enables IntelliJ to write Scala code.



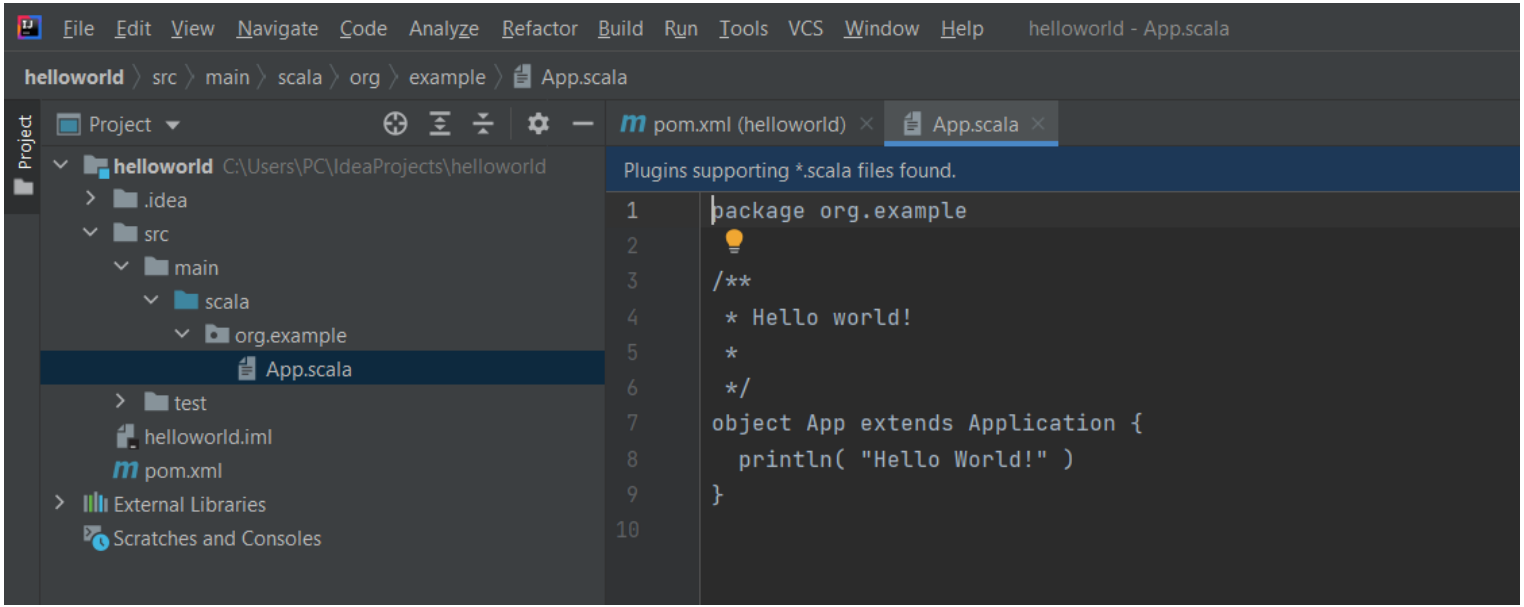
Create a Scala project In IntelliJ



Create a Scala project In IntelliJ

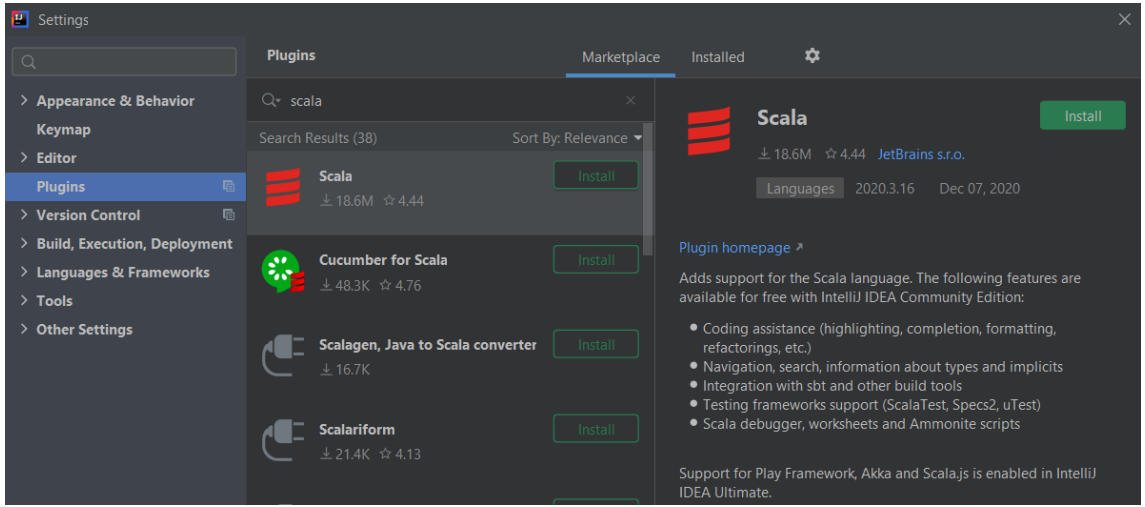


Create a Scala project In IntelliJ

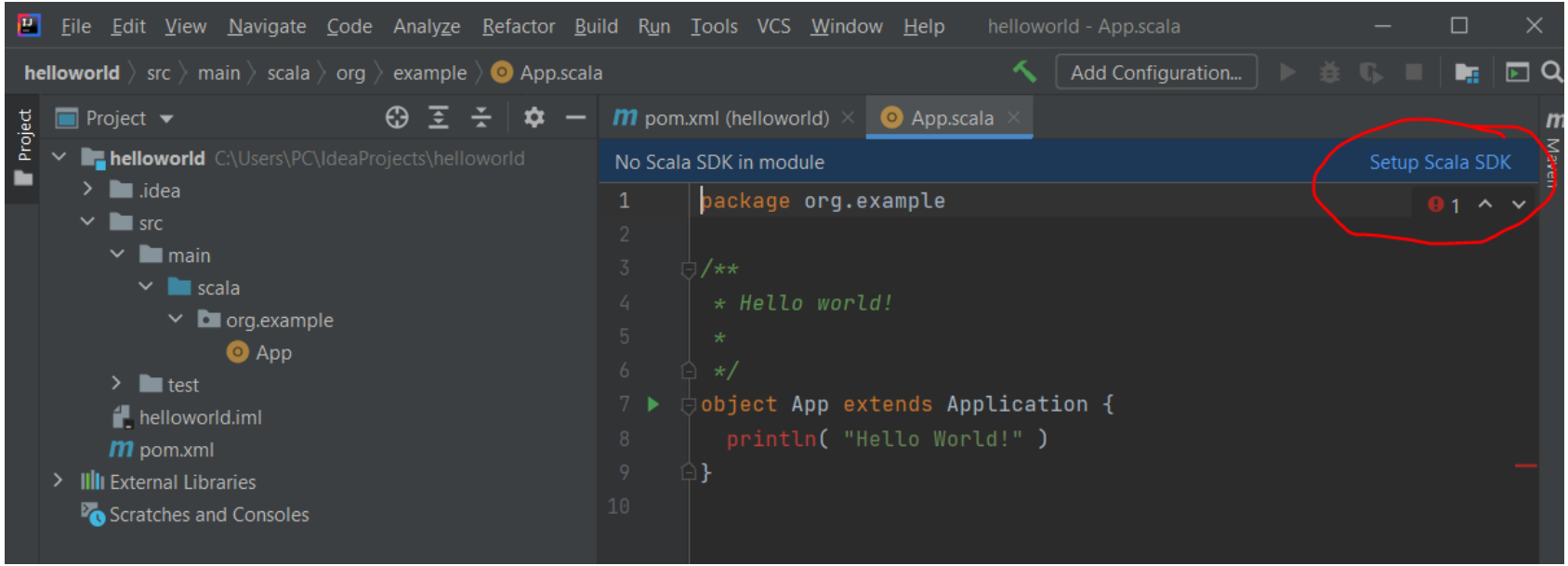


Install Scala Plugin

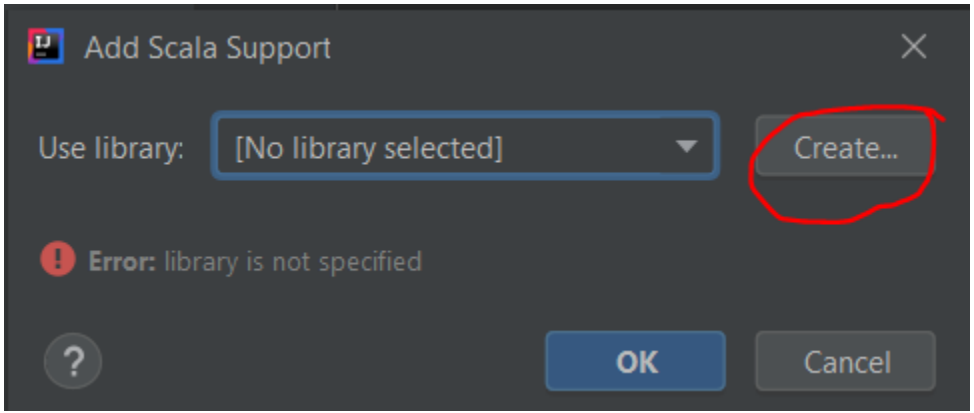
- Open File > Settings (or using shot keys Ctrl + Alt + s)
- Select the Plugins option from the left panel. This brings you Feature panel.
- Click on Install to install the Scala plugin.
- After plugin install, restart the IntelliJ IDE.



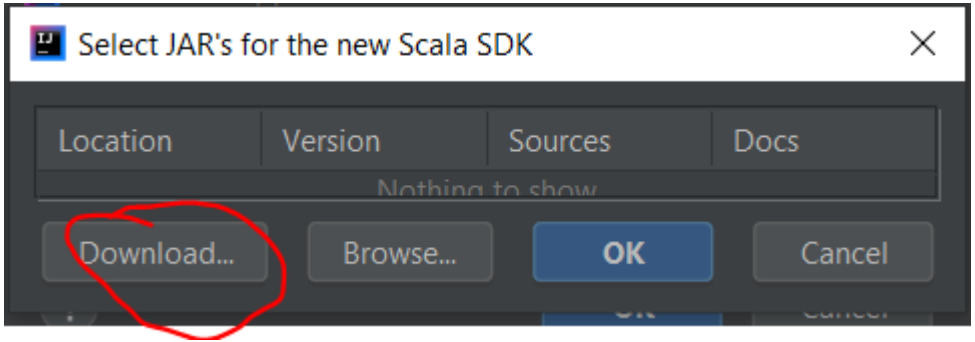
Setup Scala SDK



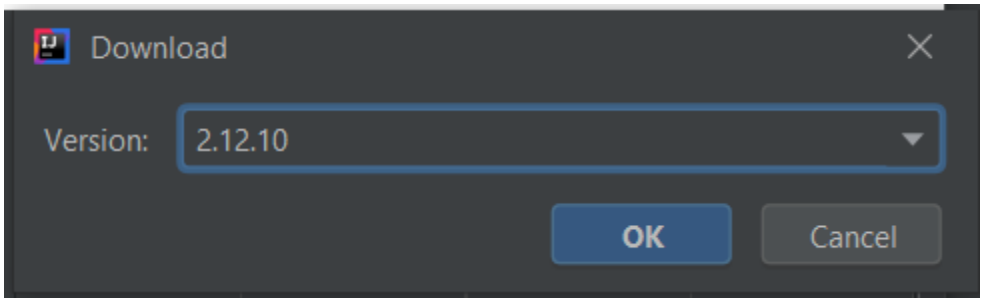
Setup Scala SDK



Setup Scala SDK

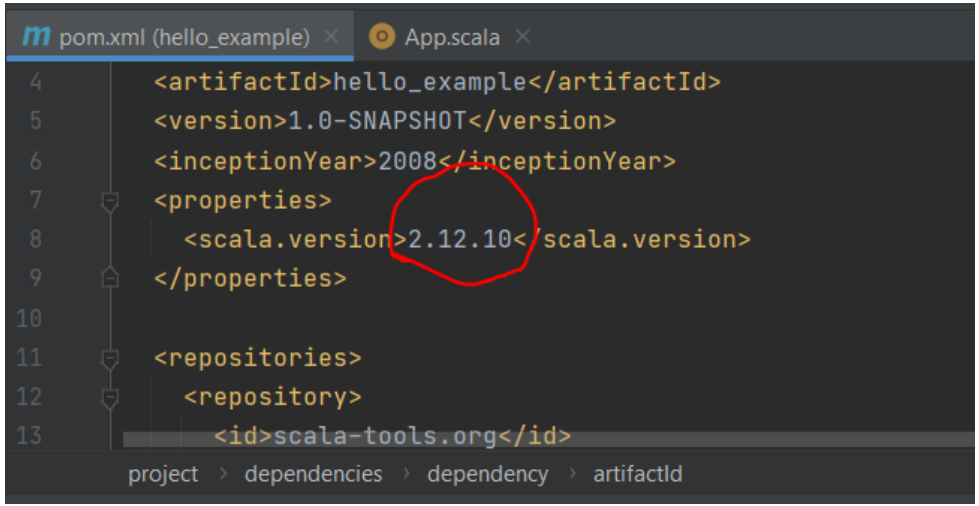


Setup Scala SDK

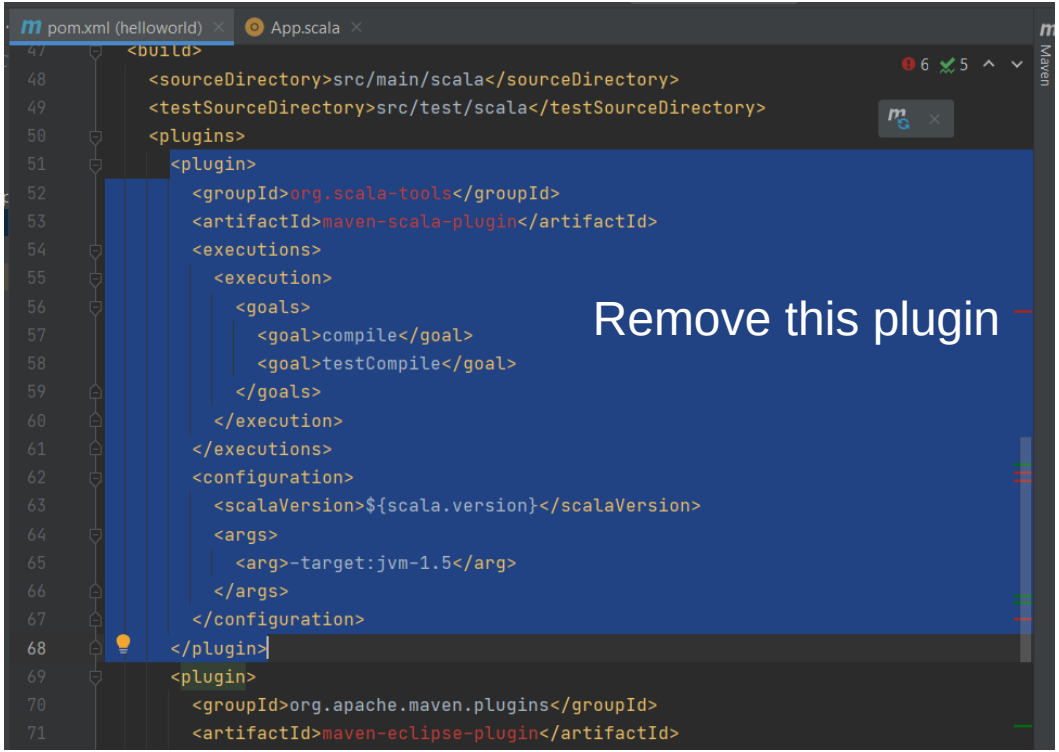


Make changes to pom.xml file

- First, change the Scala version to the latest version

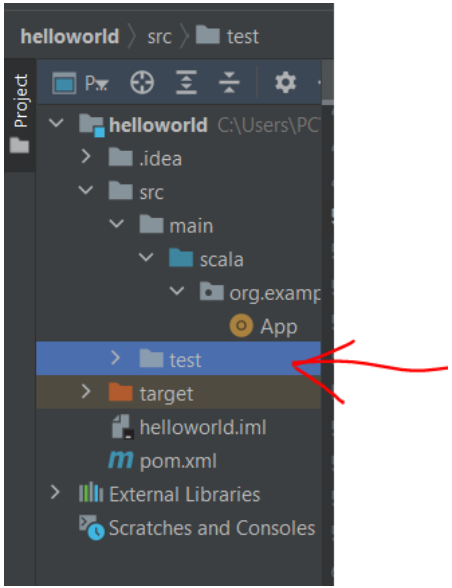


Make changes to pom.xml file



Delete Unnecessary Files

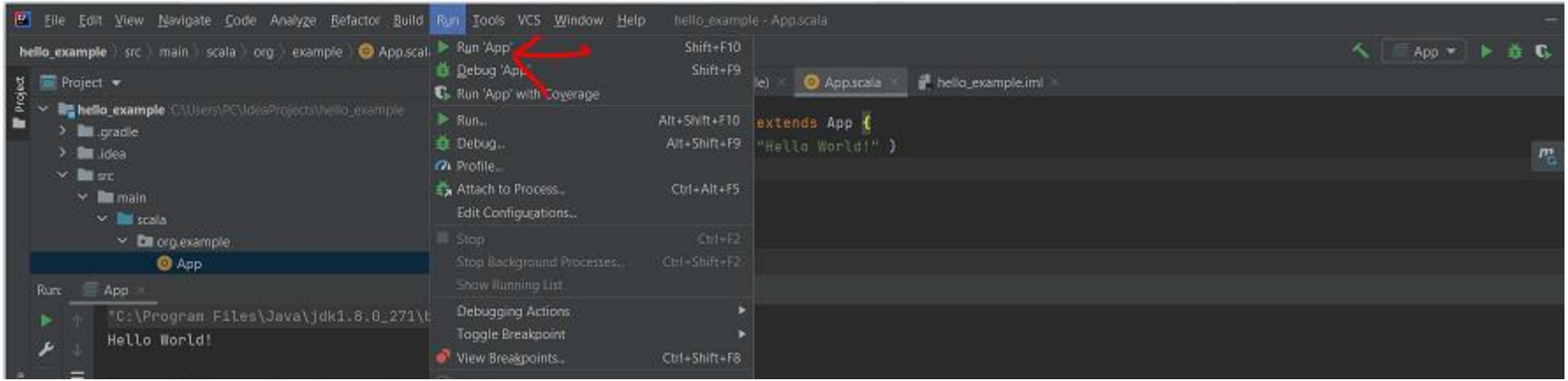
- Delete src/test



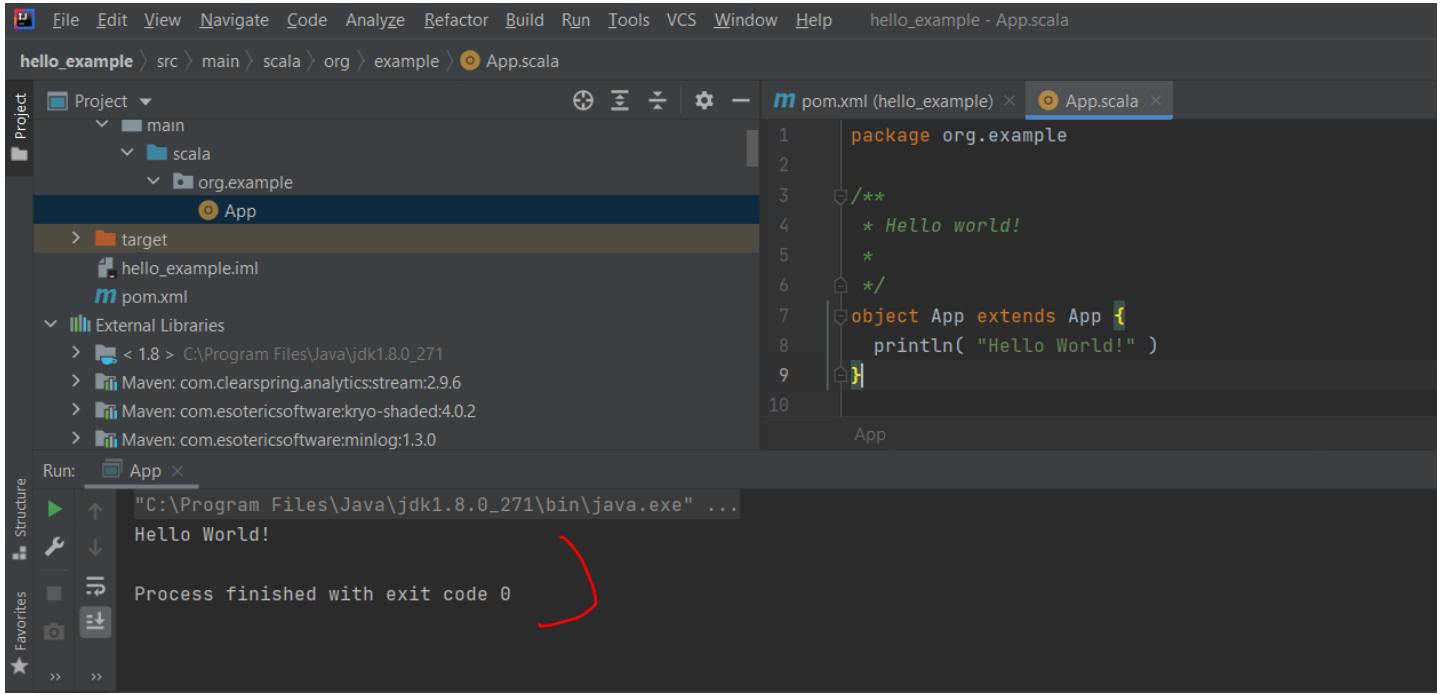
Add Spark Dependencies to Maven pom.xml File

```
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-core_2.12</artifactId>
    <version>3.0.1</version>
    <scope>compile</scope>
  </dependency>
  <dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-sql_2.12</artifactId>
    <version>3.0.1</version>
    <scope>compile</scope>
  </dependency>
```

Run the applicatioin

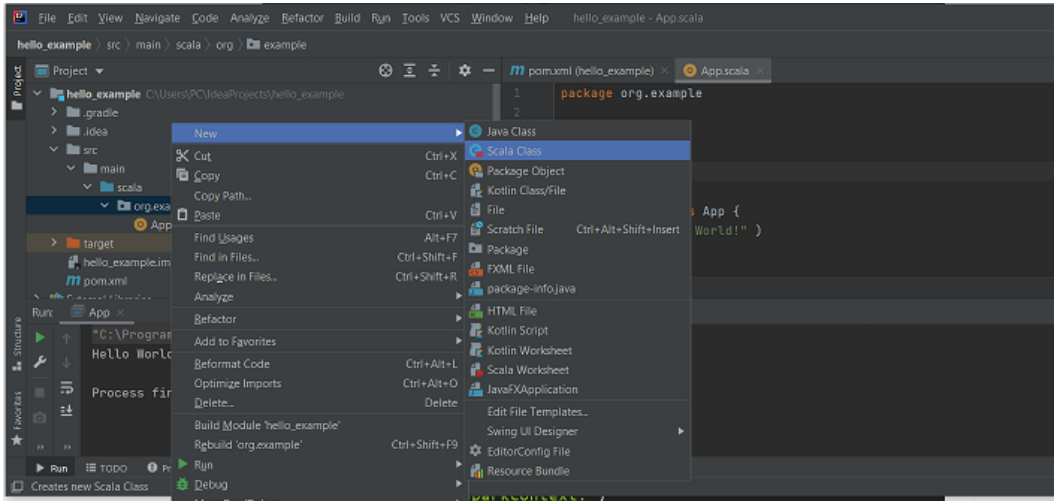


Run the applicatioin



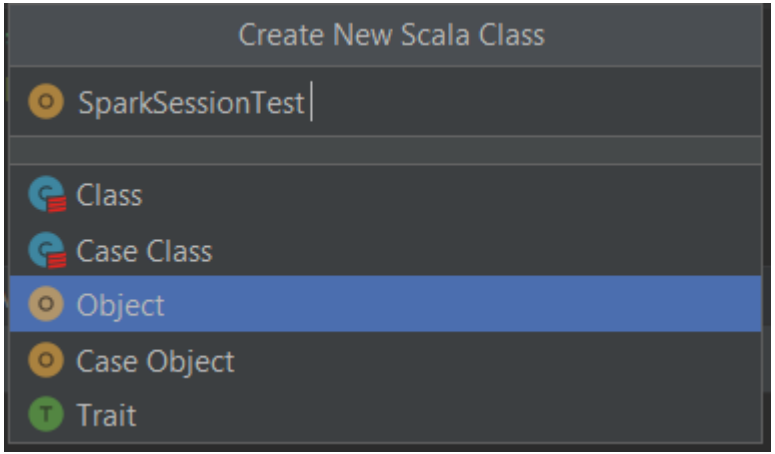
Create new Spark application

- Right click org.example -> New -> Scala Class



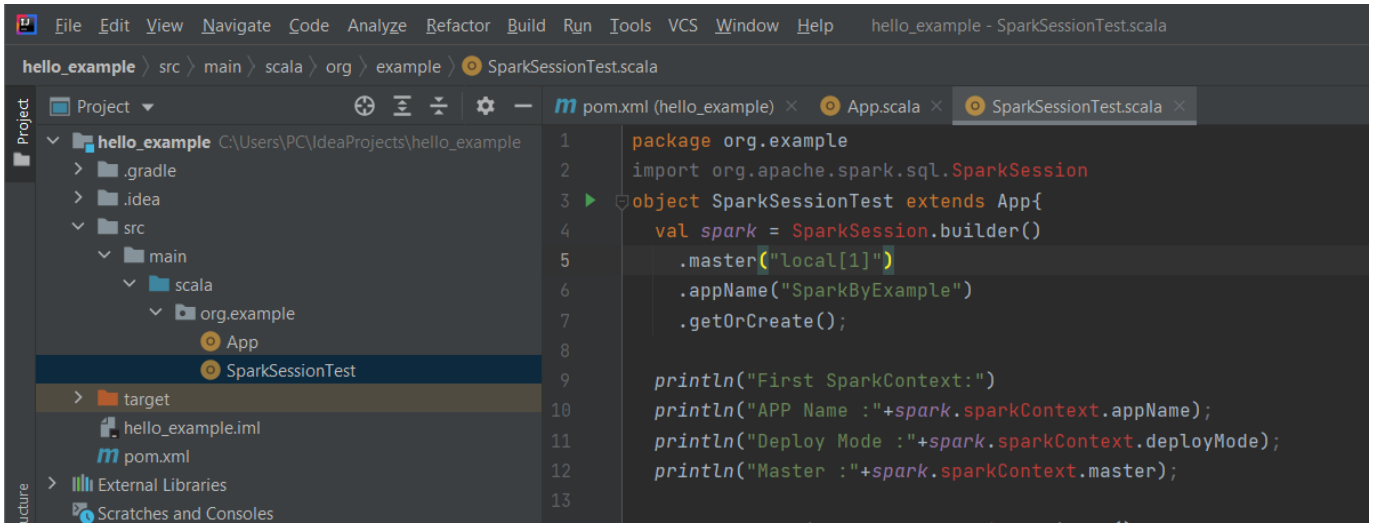
Create new Spark application

- Enter “SparkSessionTest”
- Choose “Object”

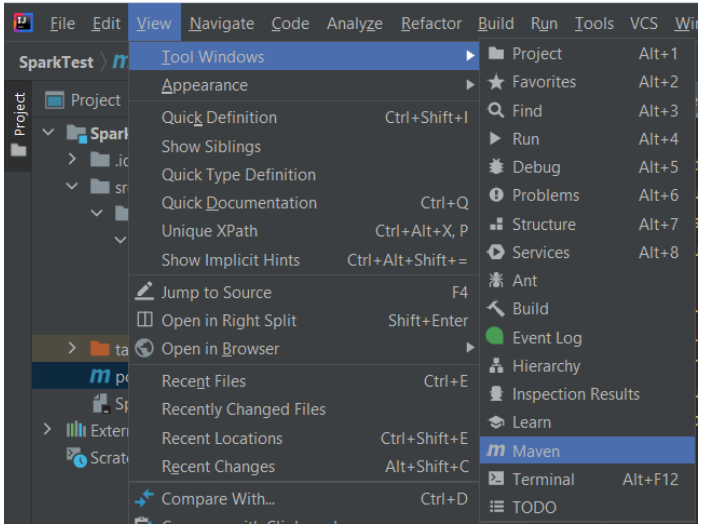


Create new Spark application

- Copy and past the code

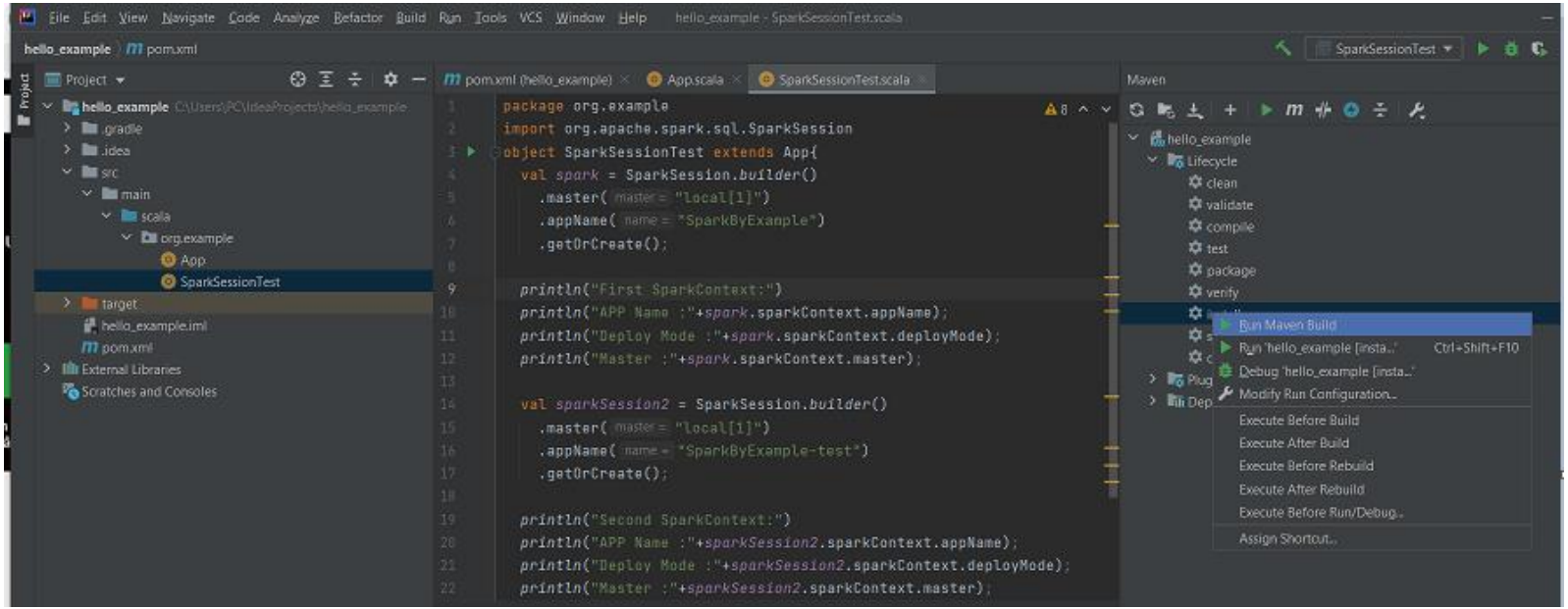


Run Maven Build



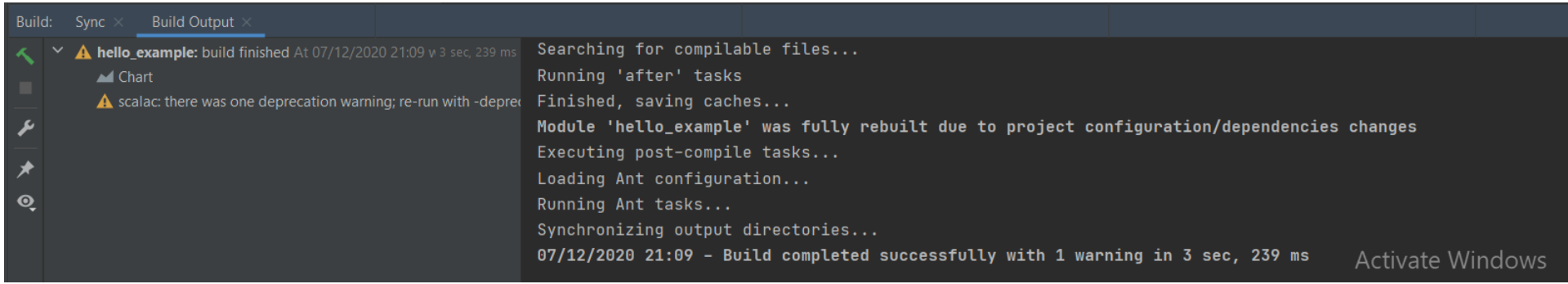
Run Maven Build

- Maven -> Lifecycle > install, right-click, and select Run Maven Build.



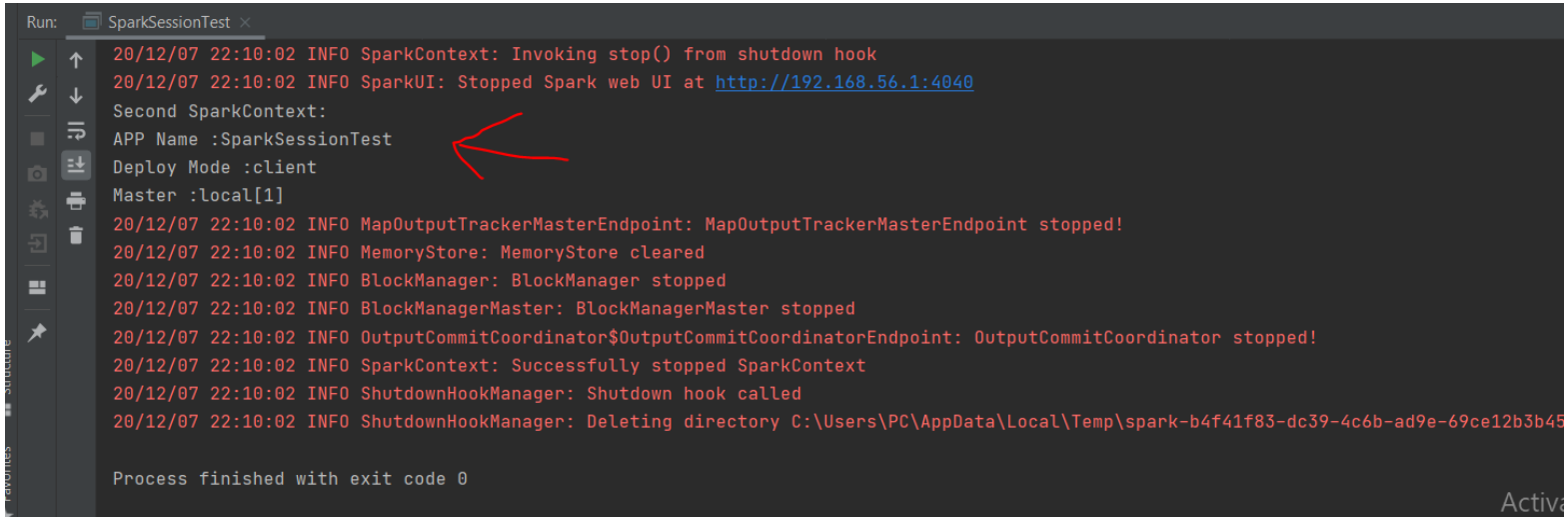
Create new Spark application

- Or just Load Maven Changes



Create new Spark application

- Run the application



Q & A

Cảm ơn đã theo dõi

Chúng tôi hy vọng cùng nhau đi đến thành công.