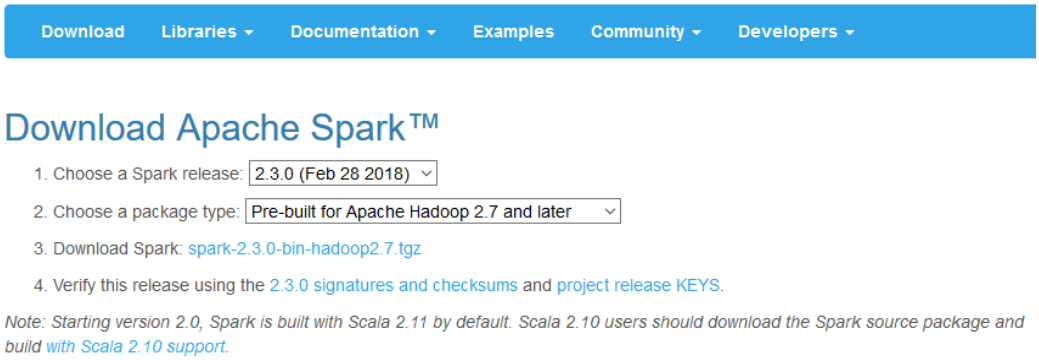
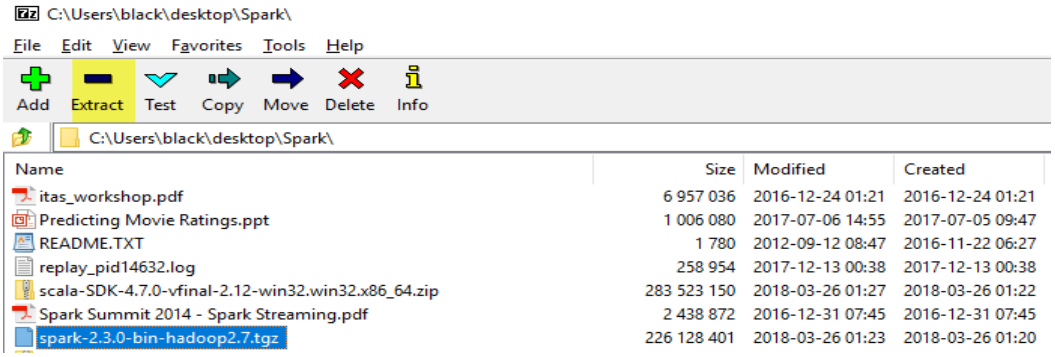
Steps to install Apache Spark in STANDALONE Mode

1. Go to Apache Spark Website.
2. Select version of Apache Spark. In this case, download version 2.3. Click on spark-2.3.0-bin-hadoop2.7.tgz

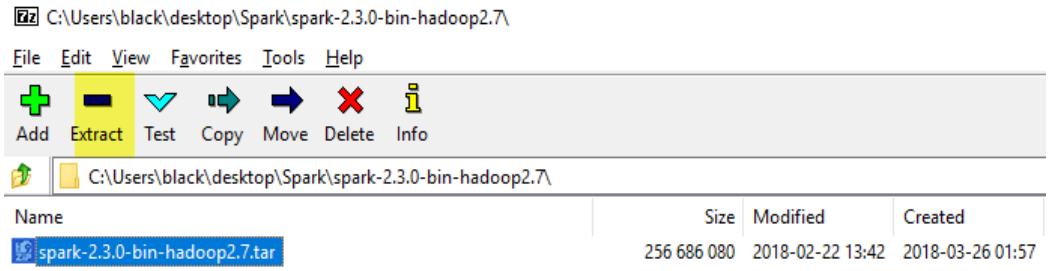


3. It will direct you to the Apache Software Foundation. Click on **http://www-us.apache.org/dist/spark/spark-2.3.0/spark-2.3.0-bin-hadoop2.7.tgz**. This will download a tgz file.

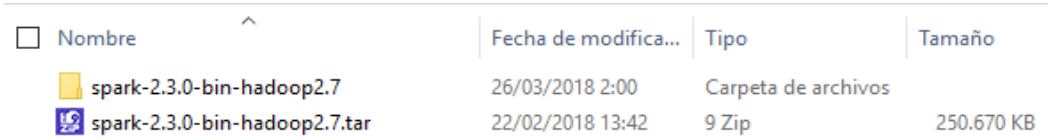
4. Move the tgz file to your desktop.
5. Use any tool to unzip the tgz file. In this course, it will be use 7 zip. Open 7 zip, select spark-2.3.0 and click on extract.



6. This will create a new folder with the name of the tgz. Inside the spark-2.3.0-bin-hadoop2.7 folder, it will be a tar file. Select it and click on Extract.



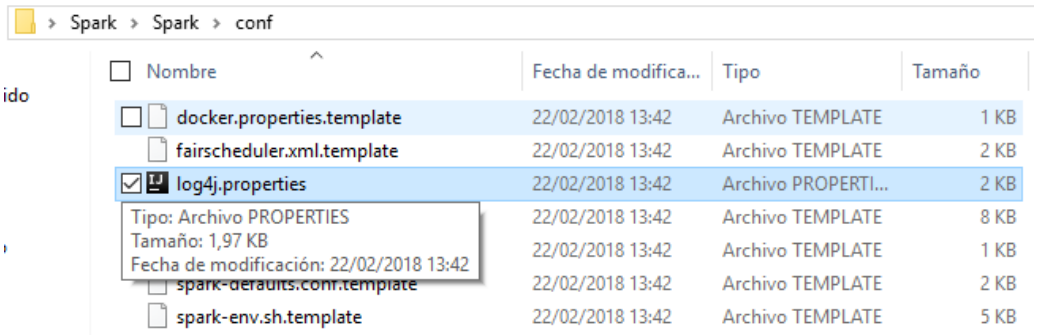
7. This will create a new folder with the name of spark-2.3.0-bin-hadoop2.7.



8. Click on the folder and there will be another folder with the same name. Click on it one more time and finally there will another folder with the same name (spark-2.3.0-bin-hadoop2.7).
   Example: **C:\Users\black\Desktop\Spark\spark-2.3.0-bin-hadoop2.7\spark-2.3.0-bin-hadoop2.7**

9. Rename the folder to "**Spark"** and move it to your desktop.

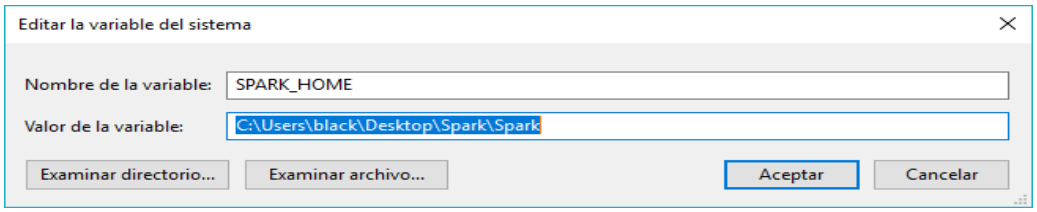
10. Inside the Spark folder, Go to the Conf folder and rename the log4j.properties.template to log4j.properties.



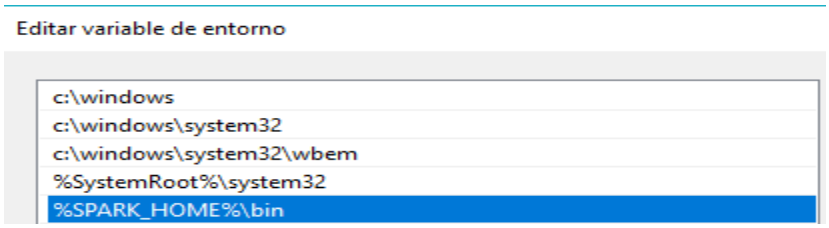
11. Open the properties and change log4j.rootCategory=INFO (line 19) to log4j.rootCategory=WARN

```
1   #
2   # Licensed to the Apache Software Foundation (ASF) under one or more
3   # contributor license agreements.  See the NOTICE file distributed with
4   # this work for additional information regarding copyright ownership.
5   # The ASF licenses this file to You under the Apache License, Version 2.0
6   # (the "License"); you may not use this file except in compliance with
7   # the License.  You may obtain a copy of the License at
8   #
9   #    http://www.apache.org/licenses/LICENSE-2.0
10  #
11  # Unless required by applicable law or agreed to in writing, software
12  # distributed under the License is distributed on an "AS IS" BASIS,
13  # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
14  # See the License for the specific language governing permissions and
15  # limitations under the License.
16  #
17
18  # Set everything to be logged to the console
19  log4j.rootCategory=WARN, console
20  log4j.appender.console=org.apache.log4j.ConsoleAppender
21  log4j.appender.console.target=System.err
22  log4j.appender.console.layout=org.apache.log4j.PatternLayout
23  log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss} %p %c{1}: %m%n
```

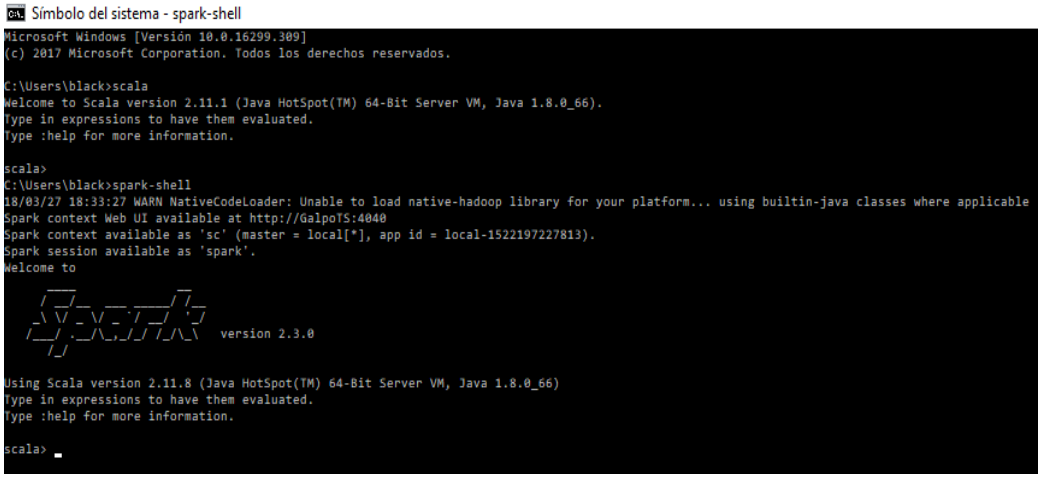
12. Go to Environment variables and create new System Variable. The new variable will have the name SPARK_HOME and the value of the path of the spark folder.



13. Add the %SPARK_HOME%\bin variable to PATH variable



14. Open a cmd window and enter **"spark-shell"**

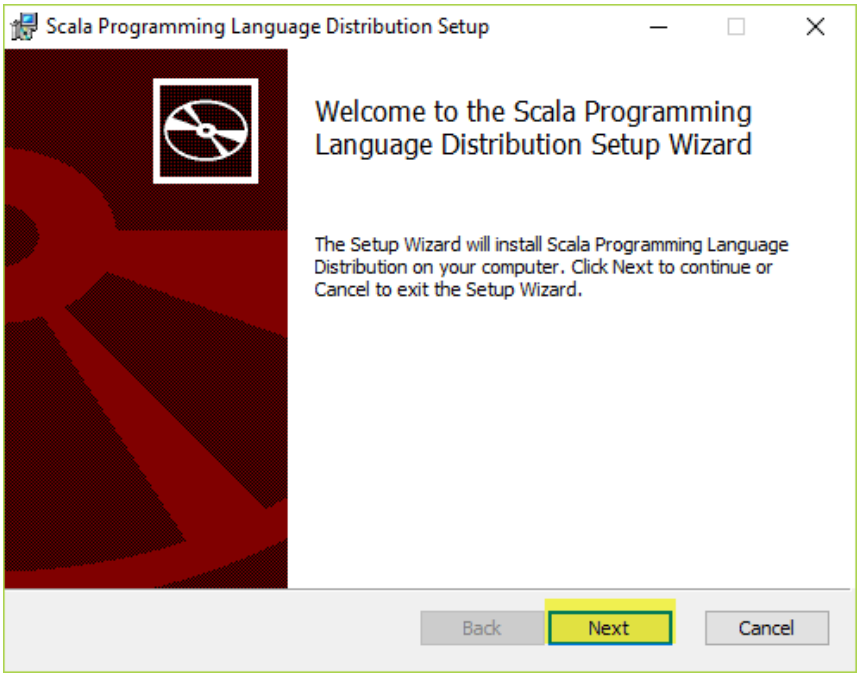


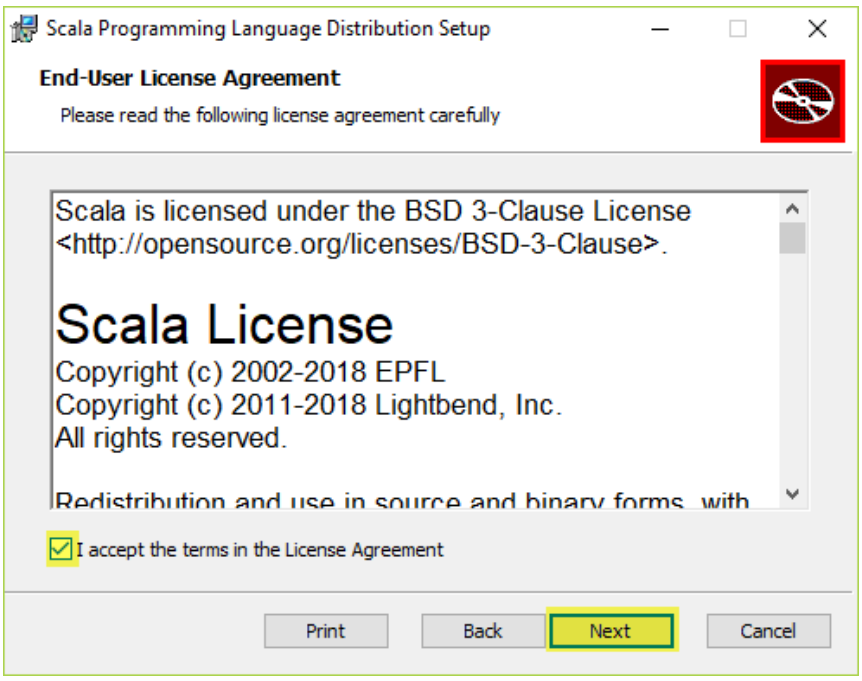
Steps to install Scala and Scala IDE

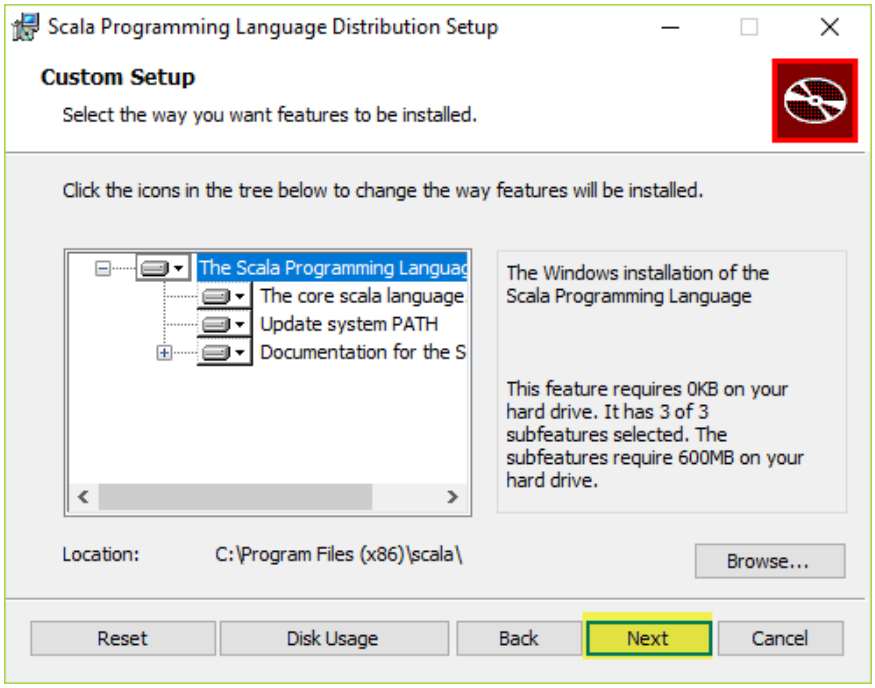
1. Go to scala-lang Website.

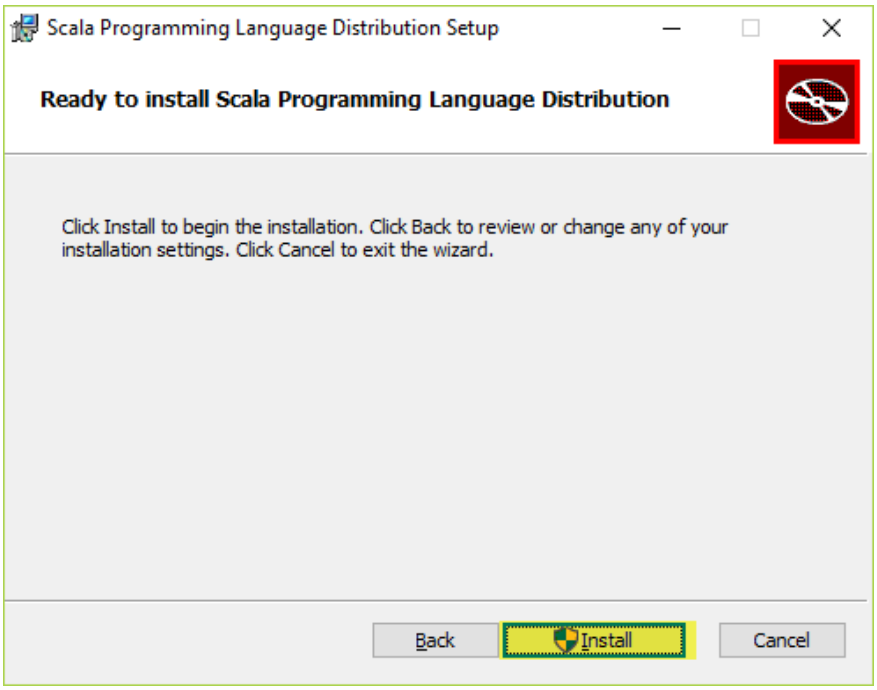
2. Go to Other ways to install Scala and click on Download the Scala binaries for windows. This will download the scala-2.12.5.msi file.

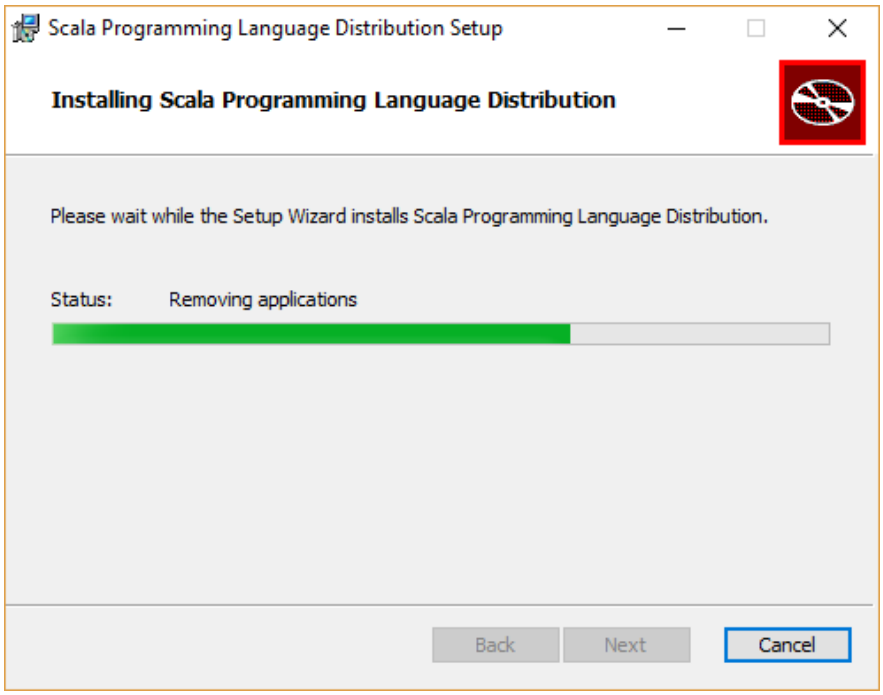
3. Install Scala Follow the steps to install it.

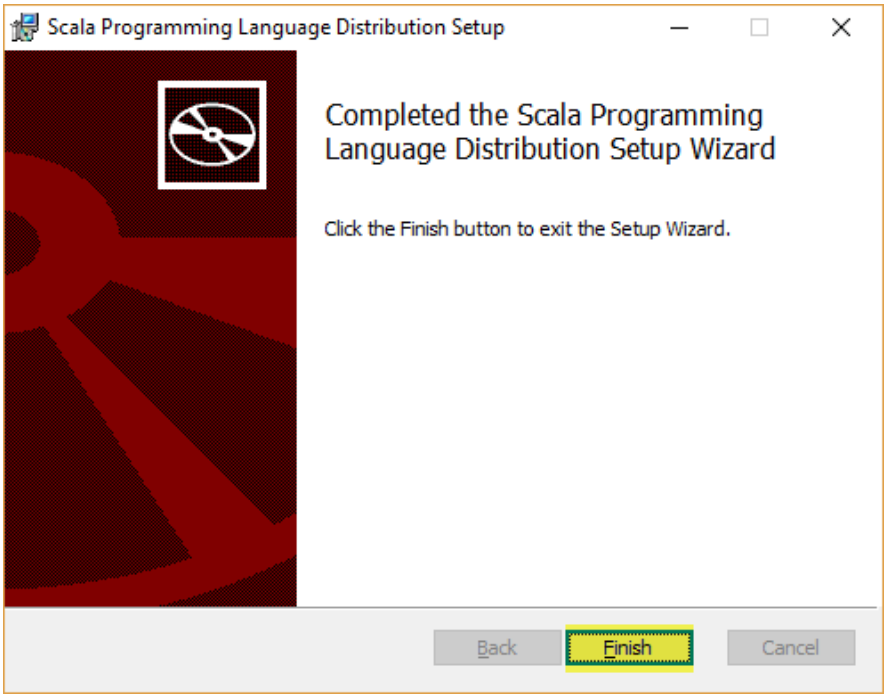




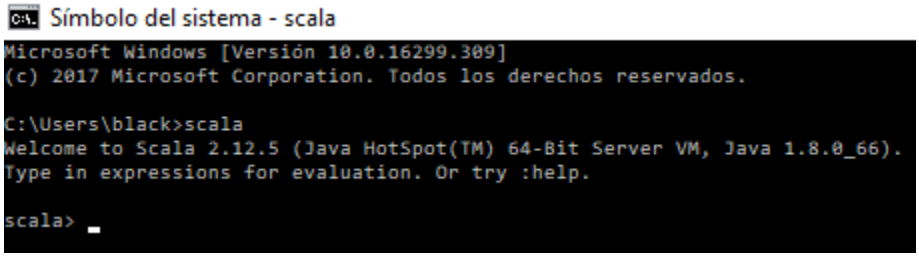

Scala Programming Language Distribution Setup
Custom Setup
Select the way you want features to be installed.
Click the icons in the tree below to change the way features will be installed.
The Scala Programming Languag
The core scala language
Update system PATH
Documentation for the S
The Windows installation of the Scala Programming Language
This feature requires 0KB on your hard drive. It has 3 of 3 subfeatures selected. The subfeatures require 600MB on your hard drive.
Location:
C:\Program Files (x86)\scala\
Browse...
Reset
Disk Usage
Back
Next
Cancel


Scala Programming Language Distribution Setup
Ready to install Scala Programming Language Distribution
Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.
Back
Install
Cancel


Scala Programming Language Distribution Setup
Installing Scala Programming Language Distribution
Please wait while the Setup Wizard installs Scala Programming Language Distribution.
Status:
Removing applications
Back
Next
Cancel


Scala Programming Language Distribution Setup
Completed the Scala Programming Language Distribution Setup Wizard
Click the Finish button to exit the Setup Wizard.
Back
Finish
Cancel

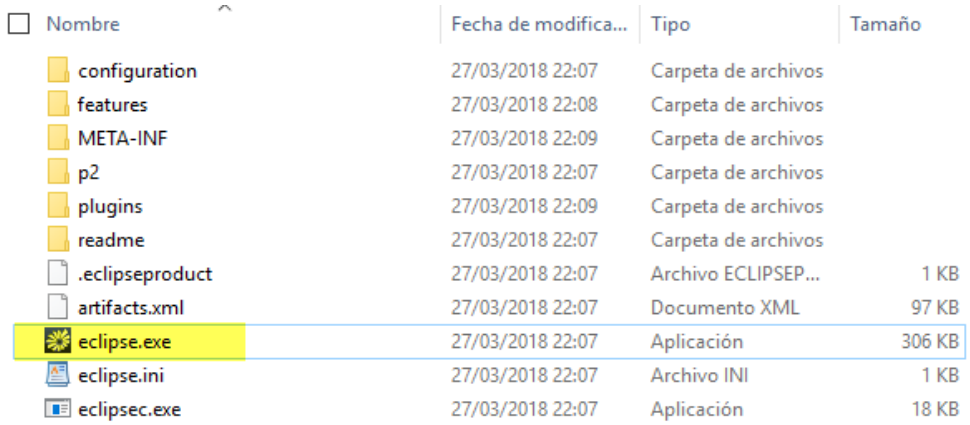
4. Open a cmd window and enter **"spark-shell"**



5. Go to Scala-ide website.

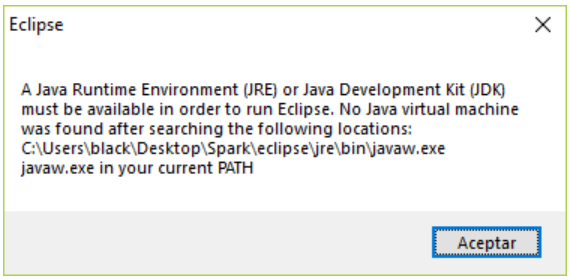
6. Click on Download IDE Windows - 64 bit

7. That will download a zip file. Unzip it and execute eclipse.exe

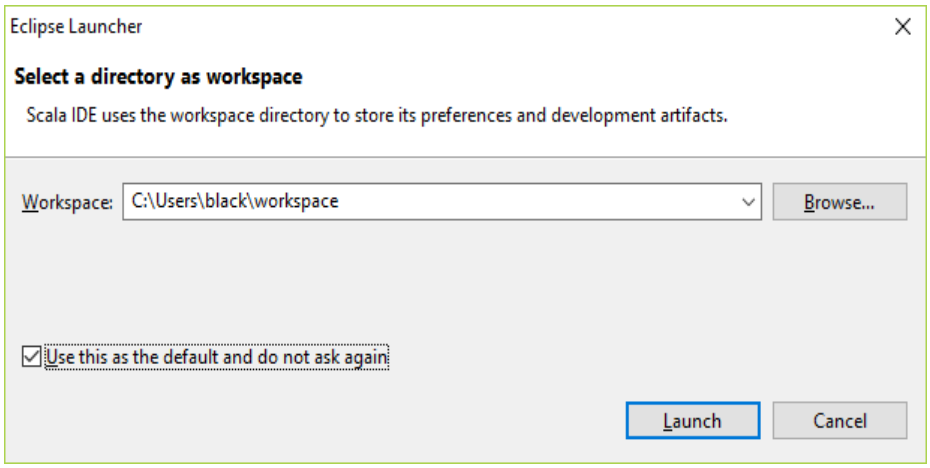


Nombre	Fecha de modifica...	Tipo	Tamaño
configuration	27/03/2018 22:07	Carpeta de archivos	
features	27/03/2018 22:08	Carpeta de archivos	
META-INF	27/03/2018 22:09	Carpeta de archivos	
p2	27/03/2018 22:07	Carpeta de archivos	
plugins	27/03/2018 22:09	Carpeta de archivos	
readme	27/03/2018 22:07	Carpeta de archivos	
.eclipseproduct	27/03/2018 22:07	Archivo ECLIPSEP...	1 KB
artifacts.xml	27/03/2018 22:07	Documento XML	97 KB
eclipse.exe	27/03/2018 22:07	Aplicación	306 KB
eclipse.ini	27/03/2018 22:07	Archivo INI	1 KB
eclipsec.exe	27/03/2018 22:07	Aplicación	18 KB

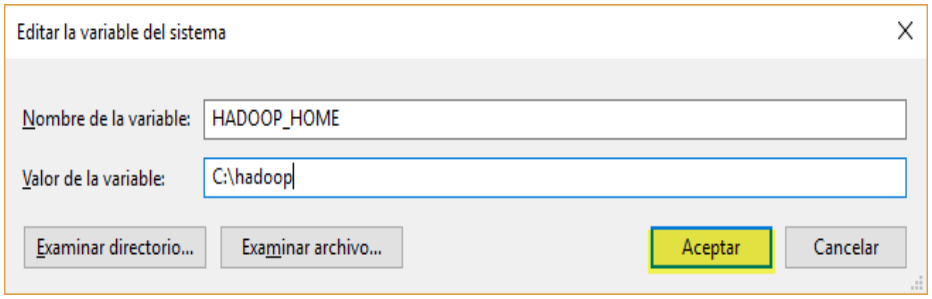
8. Add javaw.exe to your PATH system variable. Example: C:\Program Files\Java\jdk1.8.0_66\bin. This will help to avoid the next error:



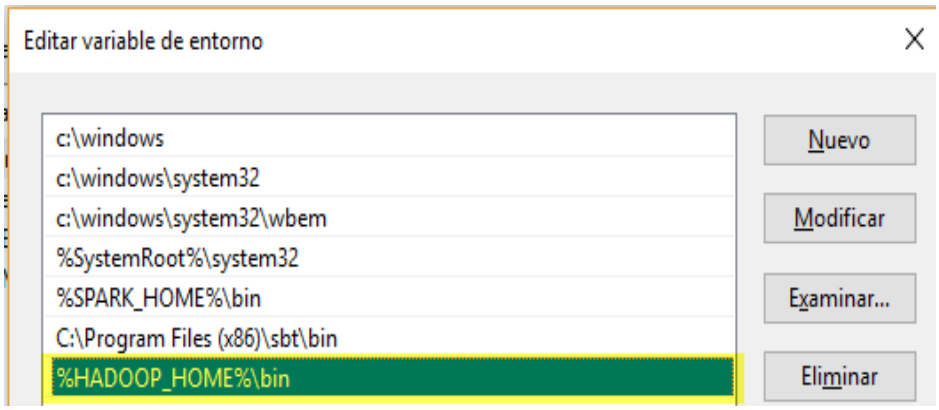
9. Select the workspace. Click Launch.



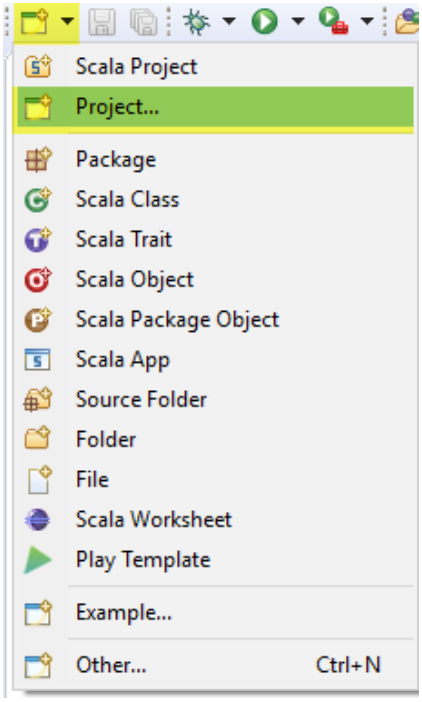
10. Create a new System Variable with name HADOOP_HOME.



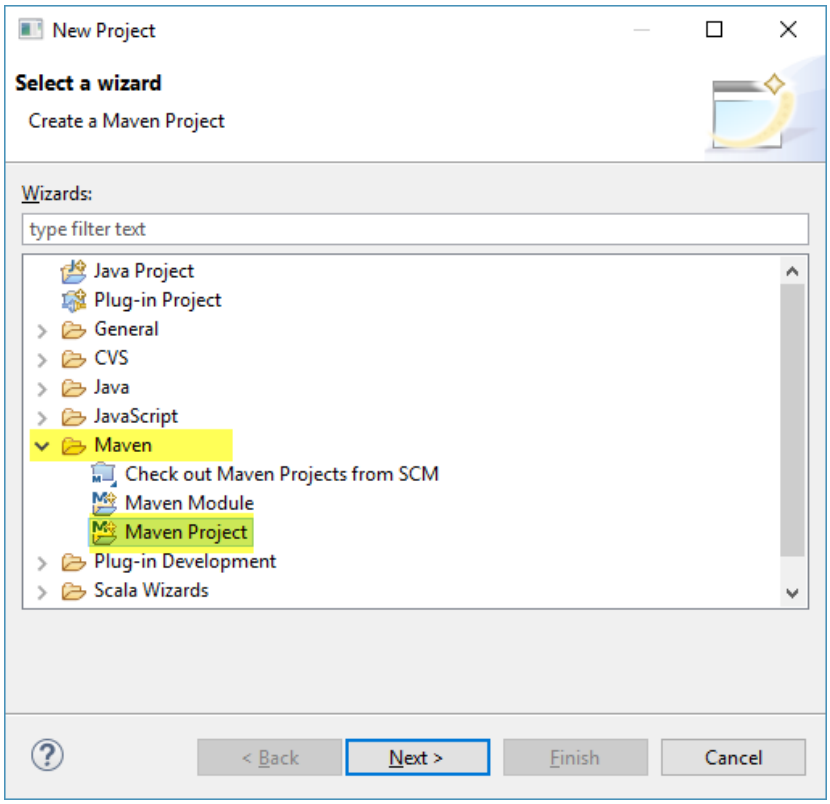
11. Add the HADOOP_HOME variable to the PATH System Variable. %HADOOP_HOME\bin%



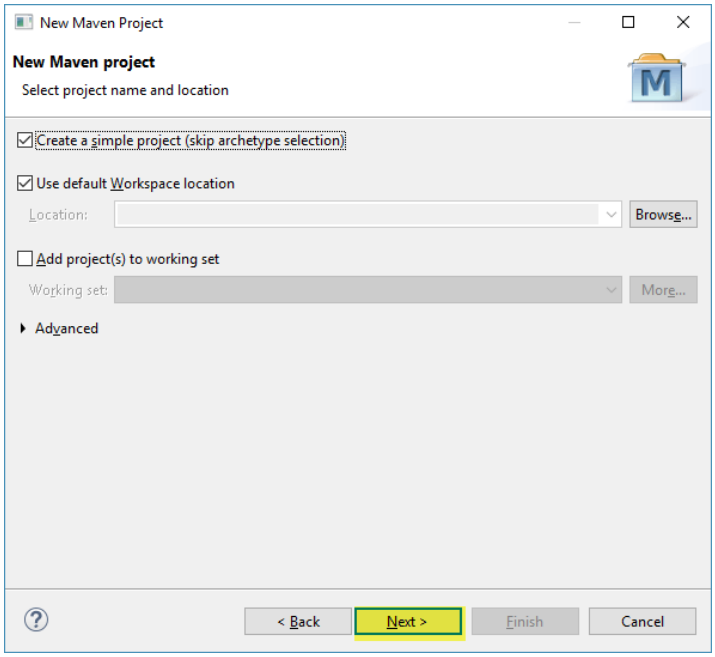
12. Create a new Maven project. Click in new and the select Project.



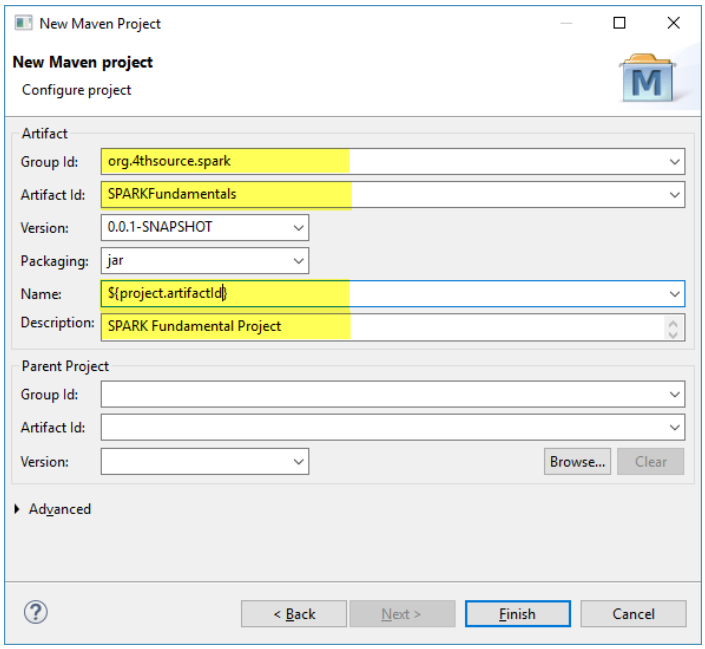
13. Go to Maven folder and select Maven Project.



14. Select “Create a simple project (skip archetype selection)” and click Next.



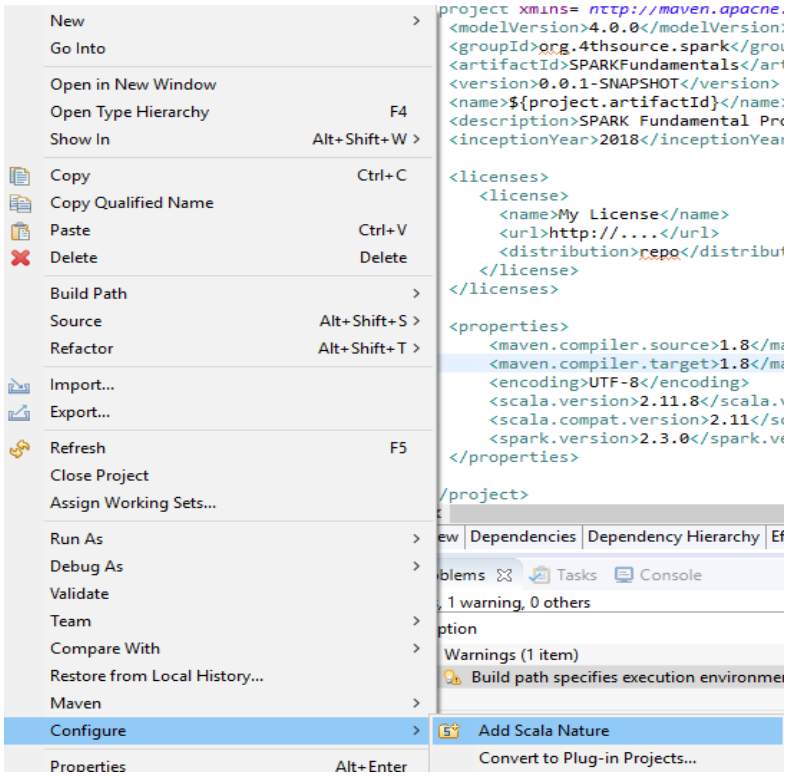
15. Enter the next field. Anything else leaves it blank. Click Finish.



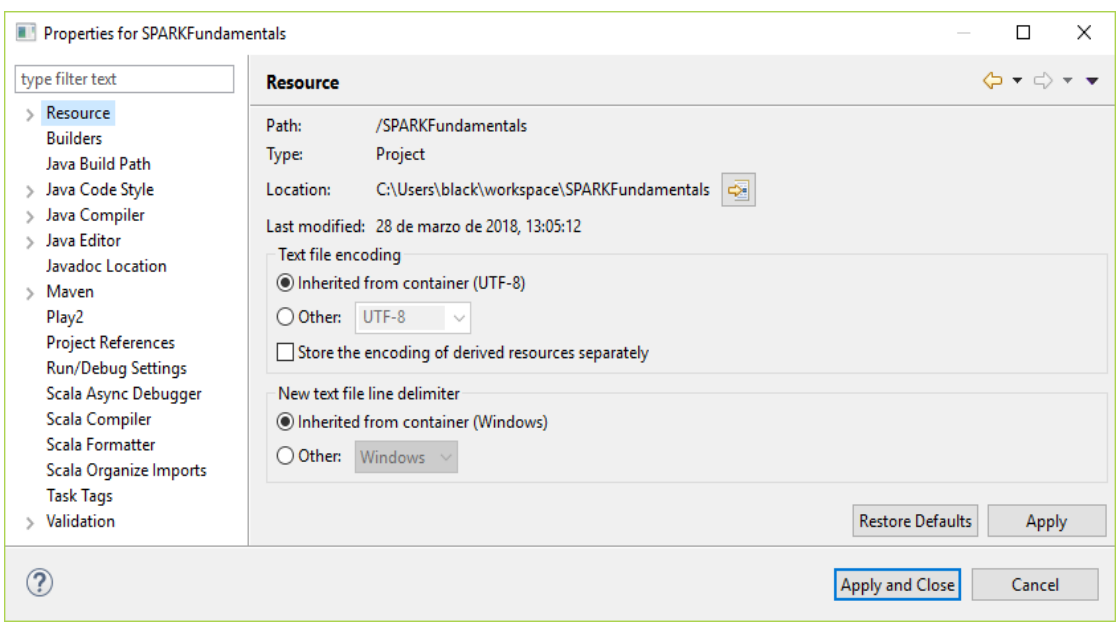
16. Overwrite the pom.xml with this pom.xml. Before doing that, make sure to put the GroupId and ArtifactID that you wrote in step 15.

pom.xml

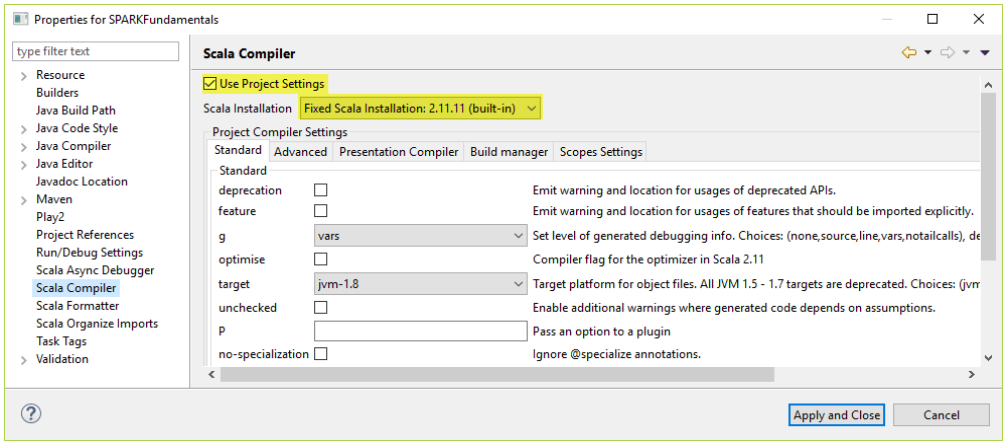
17. Next, we need to add a Scala Nature to our project. Right click under your project, go to Configure and click on Add Scala Nature.



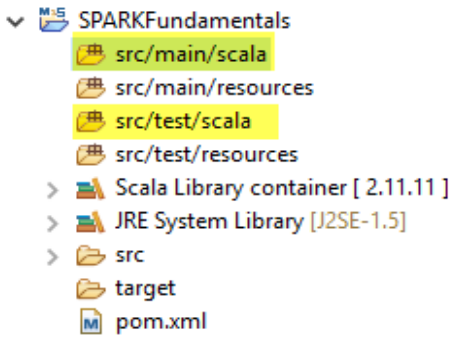
18. To select the Scala Compiler. Right click on the project and go to properties.



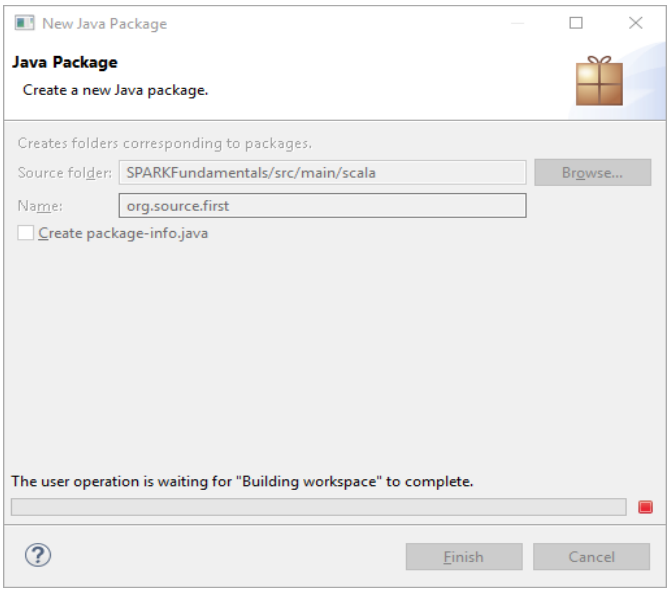
19. Select Scala Compiler; select “Use Project Settings”. Finally select the correct version of Scala. In this case is 2.11.11. Apply the changes and then click OK.



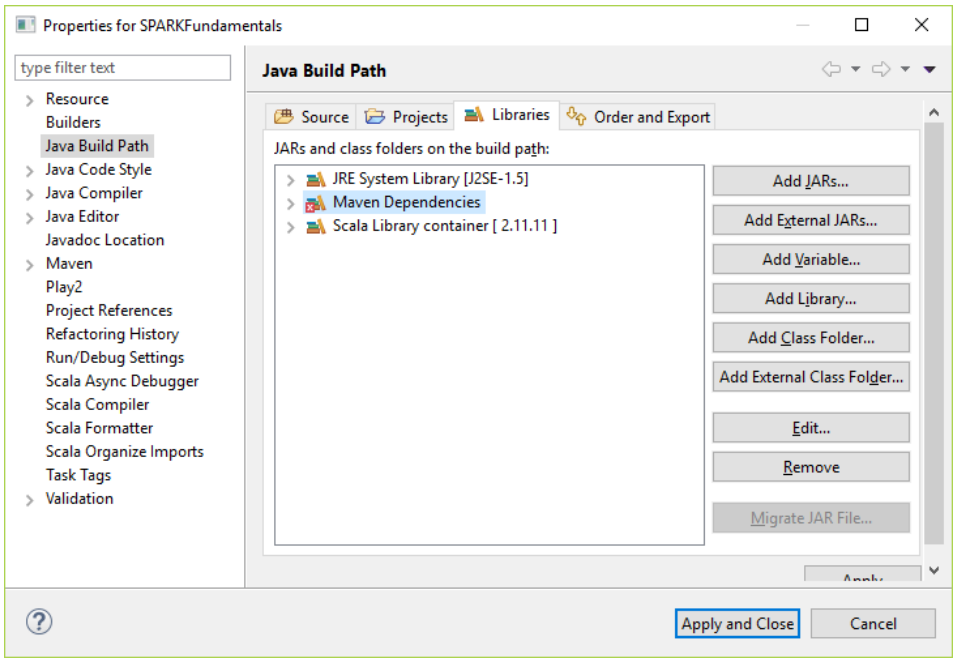
20. Change the src/main/java and src/test/java to src/main/scala and src/test/scala. Right click on each, select Refactor option, change the name and click OK.



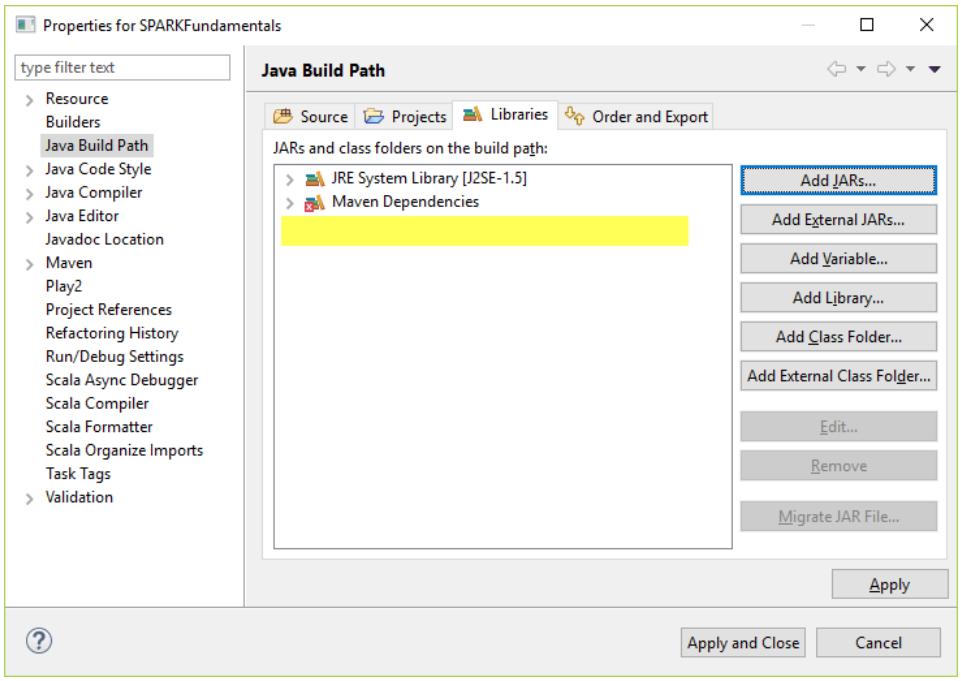
21. Now create a new package under scr/main/scala. Right click under the folder, select New. Finally click on Package. Enter the name of the package and click Finish.



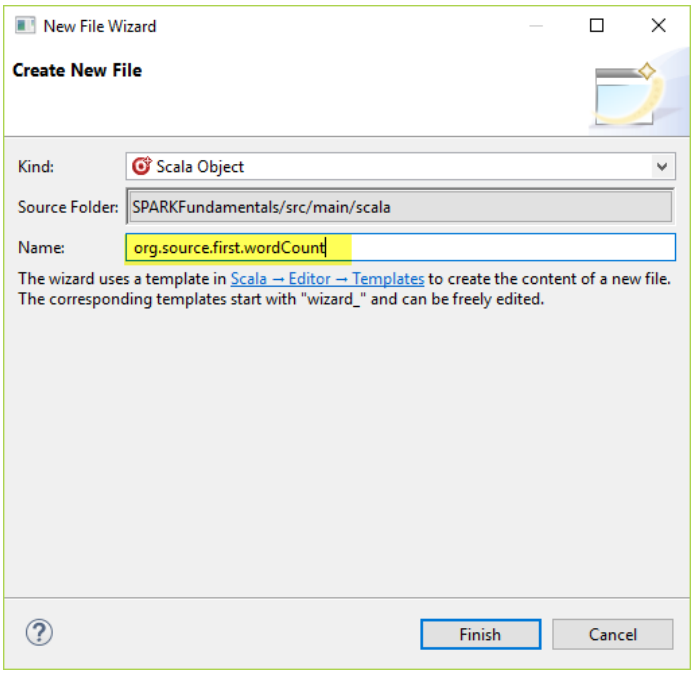
22. Next, go to Configure Build Path. Right click on your project, then Build Path. Finally, Configure Build Path.



23. Go to Libraries and Remove Scala Library container [2.11.11]. Apply the changes and click OK.



24. Create a new Scala Object under your new package. Right click under it and select New. Finally click on Scala Object. Enter the name of the Scala Object and click Finish.



25. Copy the file below to your project.

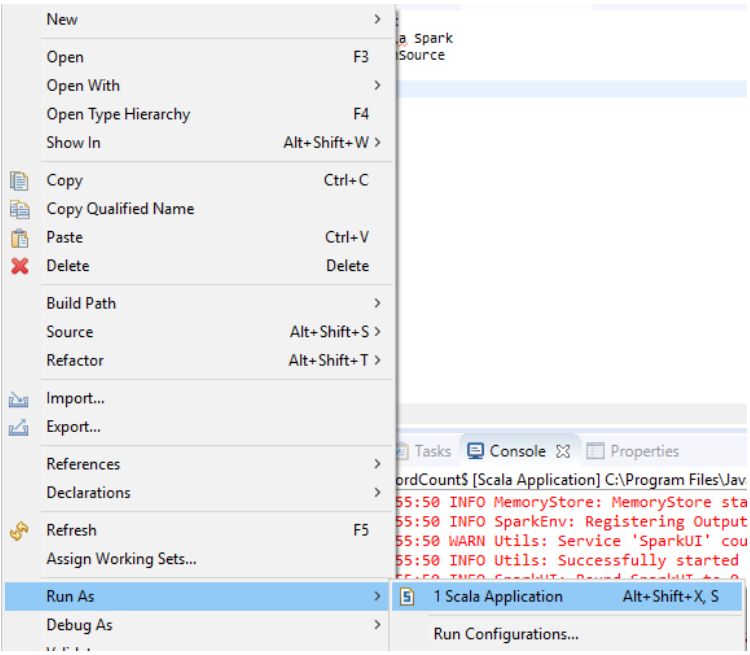
first_scala

26. Write your first Spark-Scala code

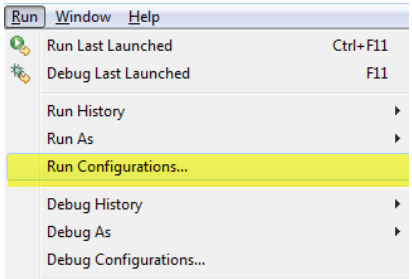
```
package org.source.first

import org.apache.spark._
import org.apache.spark.SparkContext._
import org.apache.spark.sql.SparkSession

object wordCount {
  val spark = SparkSession.builder()
                          .appName("Wordcount")
                          .config("spark.master","local")
                          .getOrCreate()

   spark.sparkContext.setLogLevel("ERROR")

   def main(args: Array[String]) = {
//    val conf = new SparkConf()
//      .setAppName("wordCount")
//      .setMaster("local")
//
//    val sc = new SparkContext(conf)
```

```
    val textrdd = spark.sparkContext.textFile("first_scala.txt")

    val newRDD = textrdd.flatMap(line => line.split(" ")).map(word =>
(word,1)).reduceByKey(_ + _).saveAsTextFile("food.count.txt")
    //.take(3)
    //map(word => (word,1)).reduceByKey(_ +
_).saveAsTextFile("food.count.txt")

  }
}
```

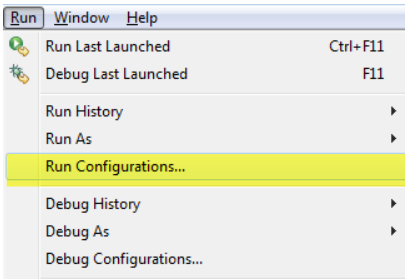
27. Right click on wordCount.scala, go to Run As and click on 1 Scala Application.



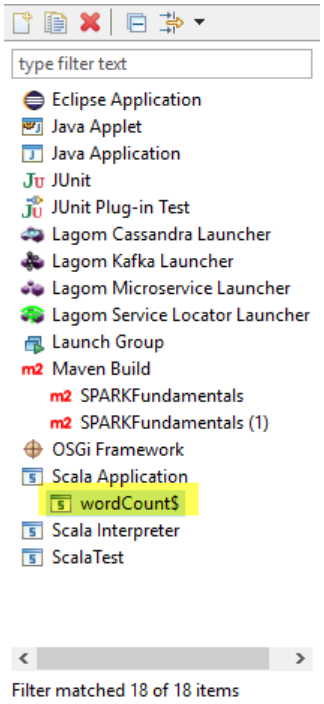
28. asdasas



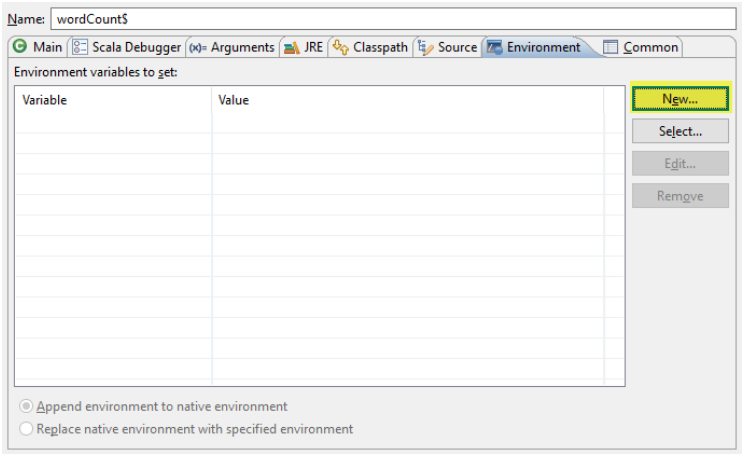
29. If there is a related error to winutil. Follow the next steps.
    a. Go to Run and click on Run Configurations.



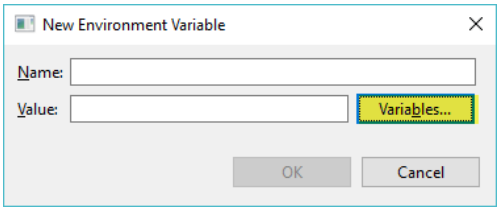
    b. Go to Scala Application-> wordCount$



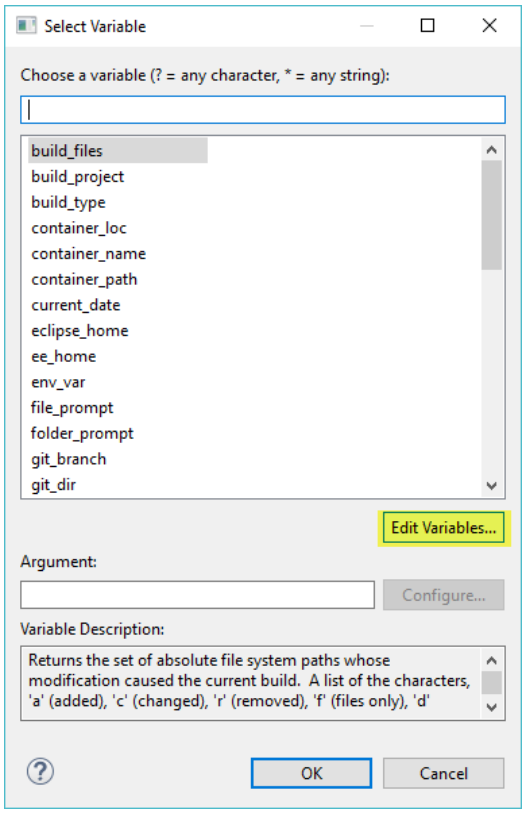
    c. Go to Environment and create a new variable. Click on New.



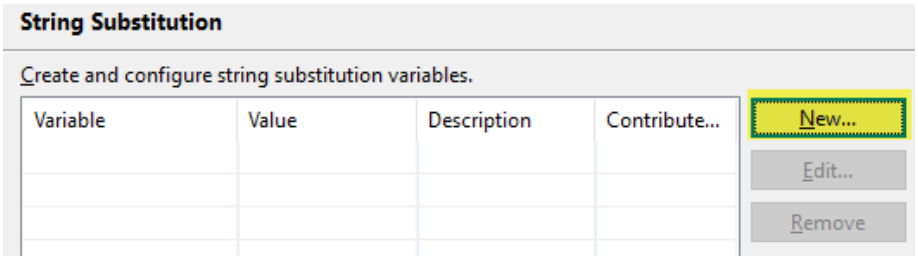
d. Click on the Variables button.



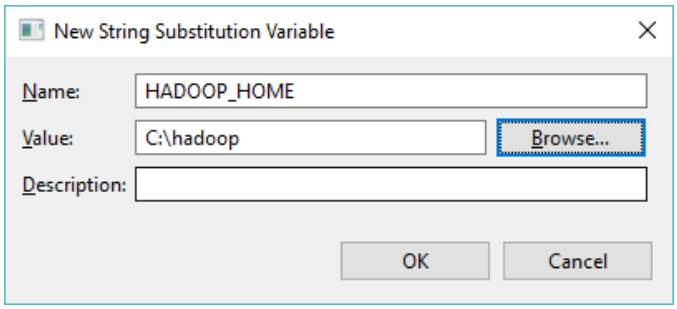
e. If this is the first time here, you need to click on Edit Variables. If not, the select the HADOOP_HOME variable and go to the final step.



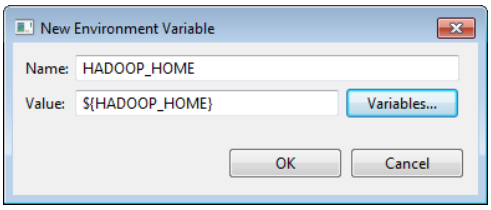
f. We are going to create a new String Substitution. Click on New button.



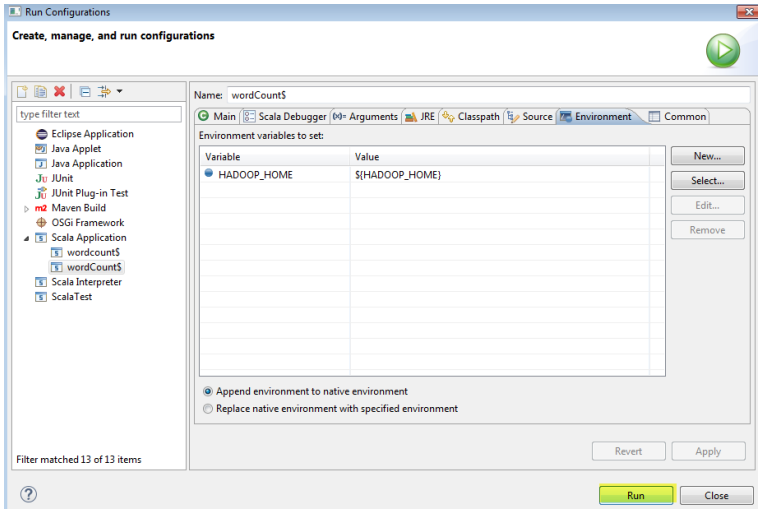
g. Enter the name of the new String Substitution Variable, browse for the bin folder. The bin folder has the winutil.exe. After that, click OK.



h. Click the button OK that is in the String Substitution. Then Select the new HADOOP_HOME variable and Click OK.



i. Now, the assigned HADOOP_HOME variable is in our project. Click Apply and then click Run. This will generate a folder with the name you provided in your Scala code.



j. Go to the folder, if everything was correct, you will see two files. One with the name _SUCCESS and the second with the name part-00000. Open the second file.

```
(Spark,3)
(Apache,1)
(Eclipse,1)
(4thSource,2)
(Scala,2)
```