CS522 – Big Data

SPARK PROJECT – PART 5

Maharishi University of Management

Department of Computer Science

**Students**: Thi Luong Dinh 985443

Dinh Tan Luong 985408

November 20th, 2016

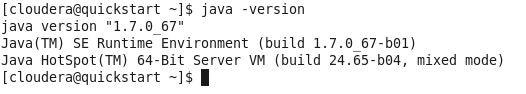
# First Scala Spark Project

Get the guide from below links:

<http://www.tutorialspoint.com//apache_spark/index.htm>

<http://scala-ide.org/docs/videos.html>

## Verifying Java Installation

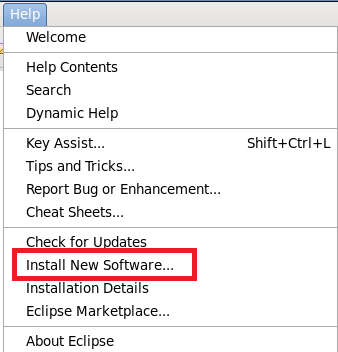


## Verifying Scala installation

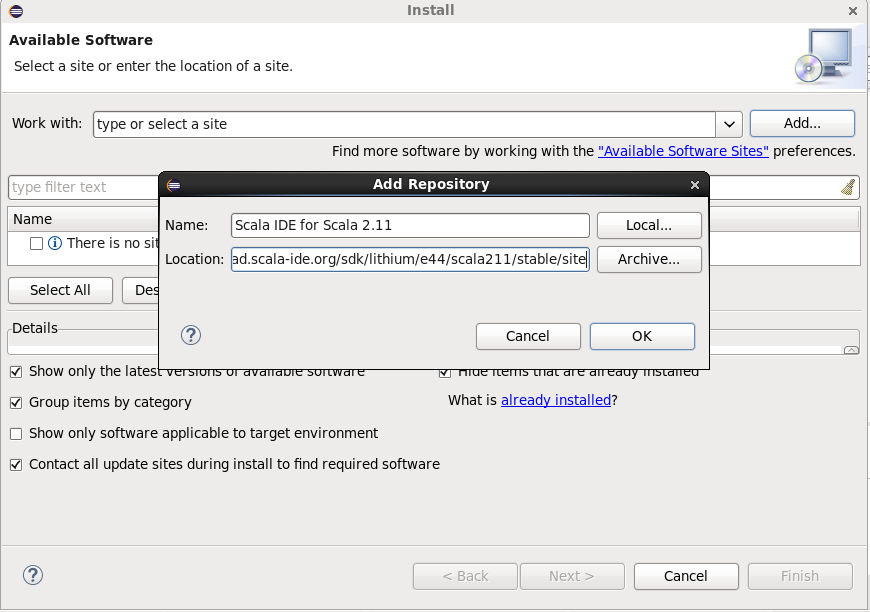


## Install Scala

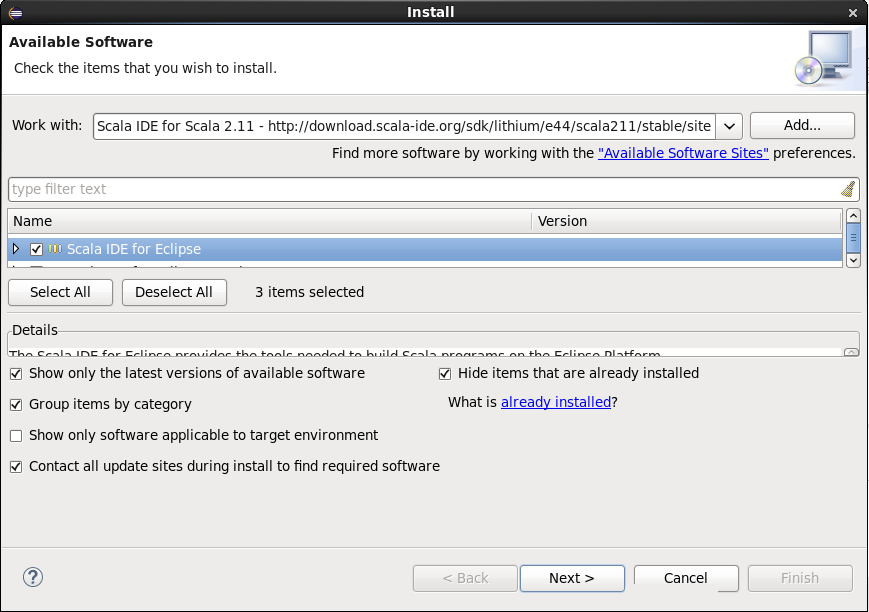
Install Scala from Eclipse by select **Help 🡪 Install New Software…**

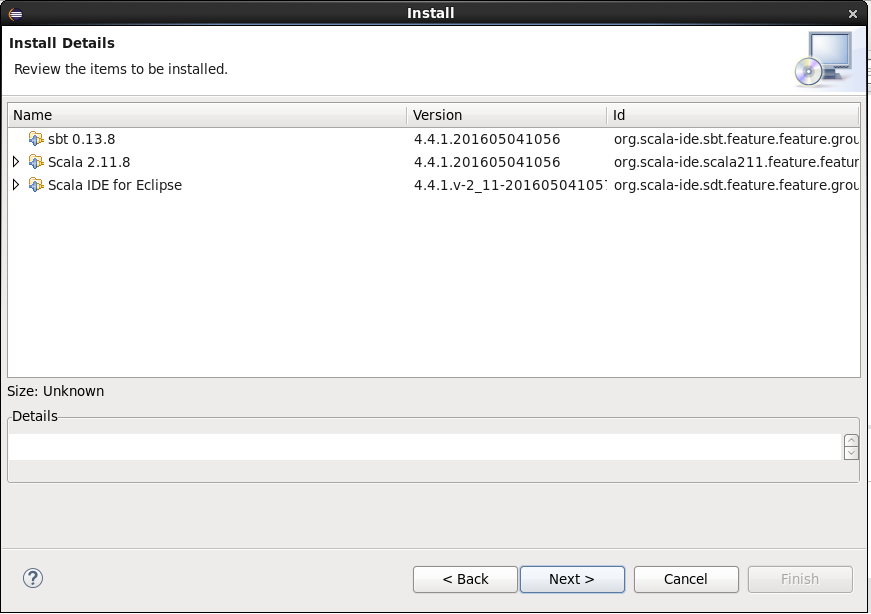


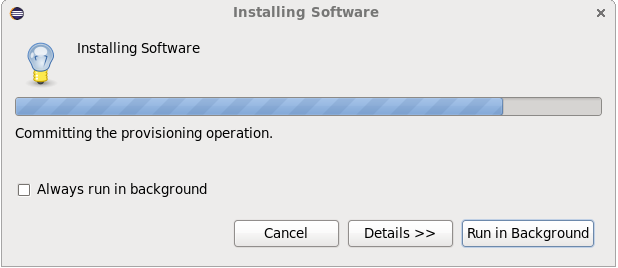
Give a name and put the location to **http://download.scala-ide.org/sdk/lithium/e44/scala211/stable/site**

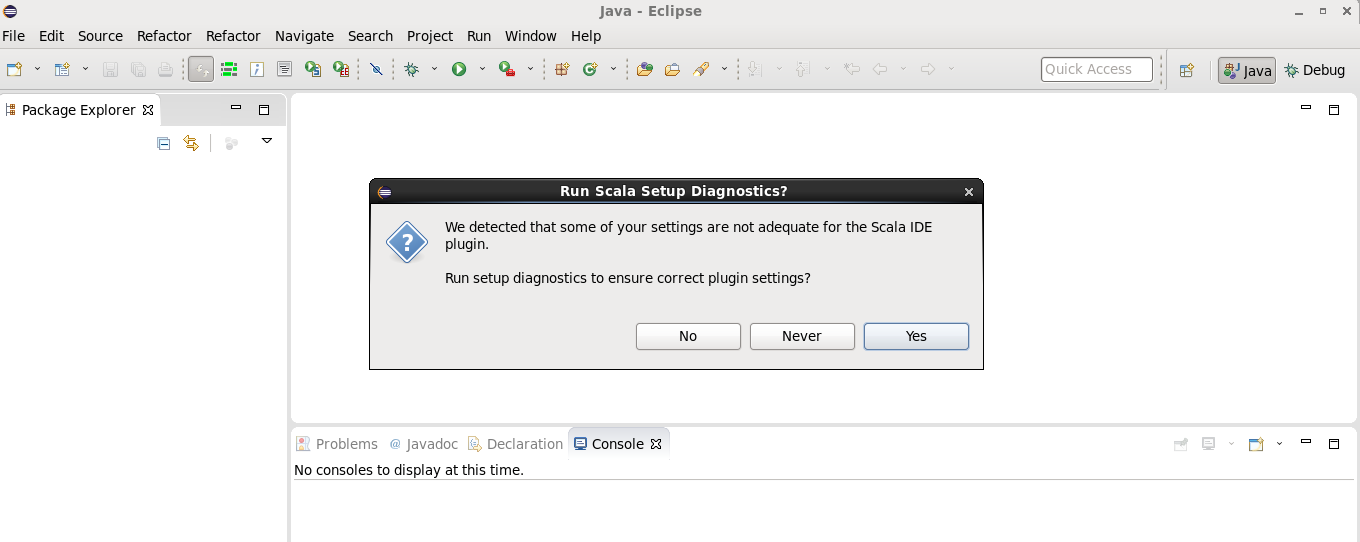


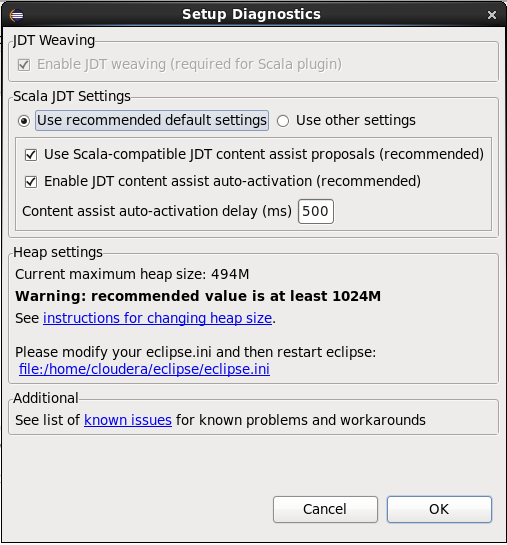
Select Scala IDE for Eclipse:





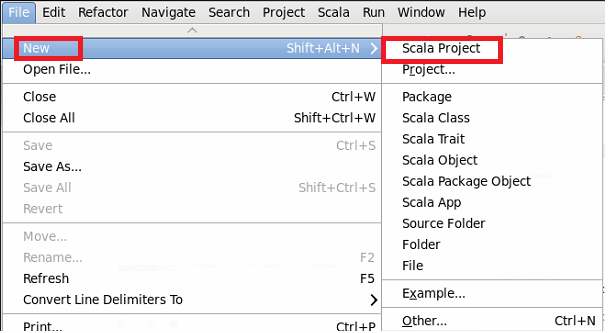




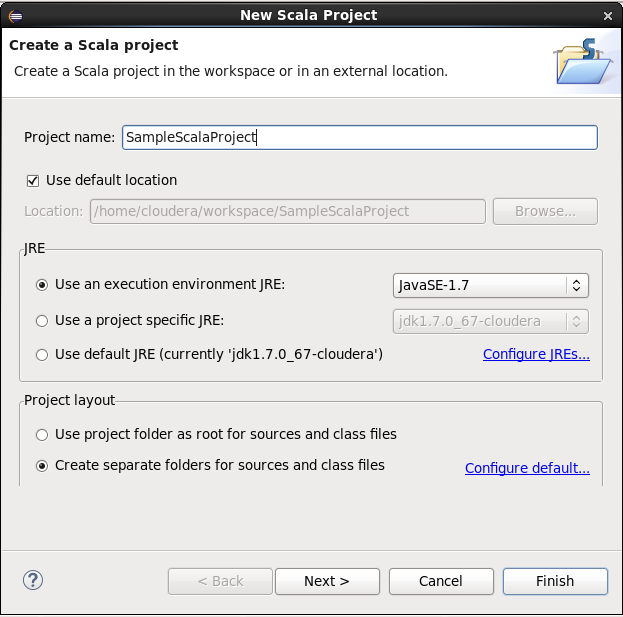


## Create a sample Scala project

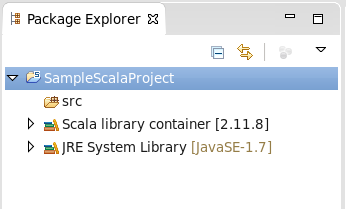
Select **File 🡪 New 🡪 Scala Project**



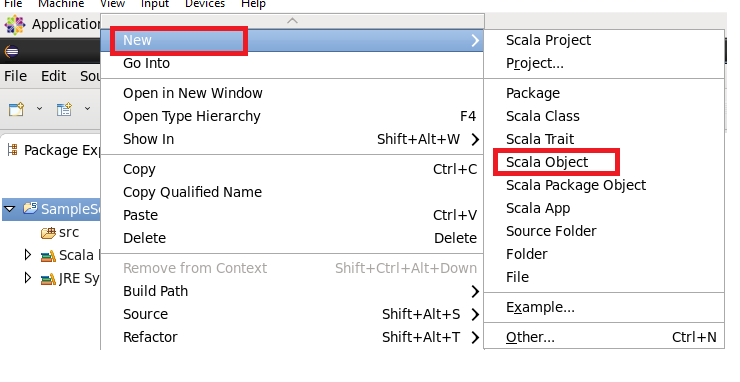
Enter the project name and click **Finish**



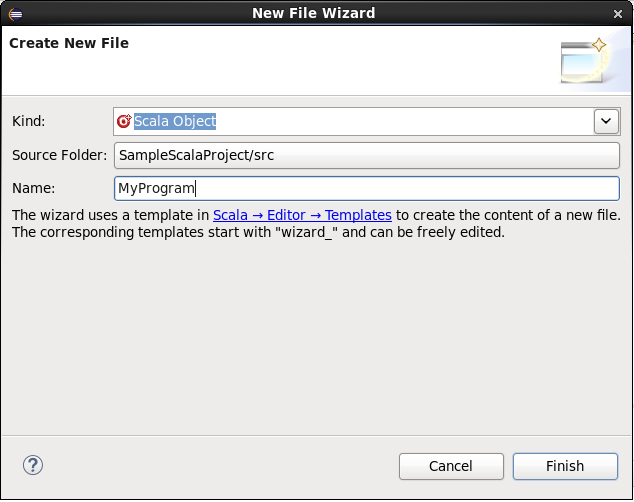
The new project was created:



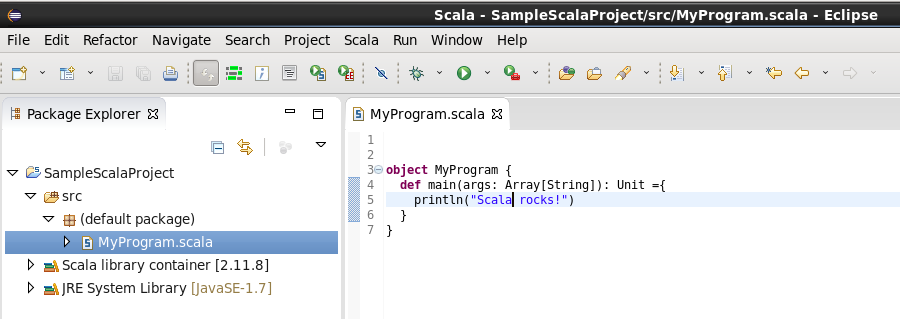
Add new Scala Object:



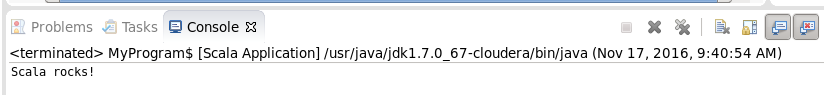
Enter name for the object:



Enter the code to print out the text “Scala rocks!”



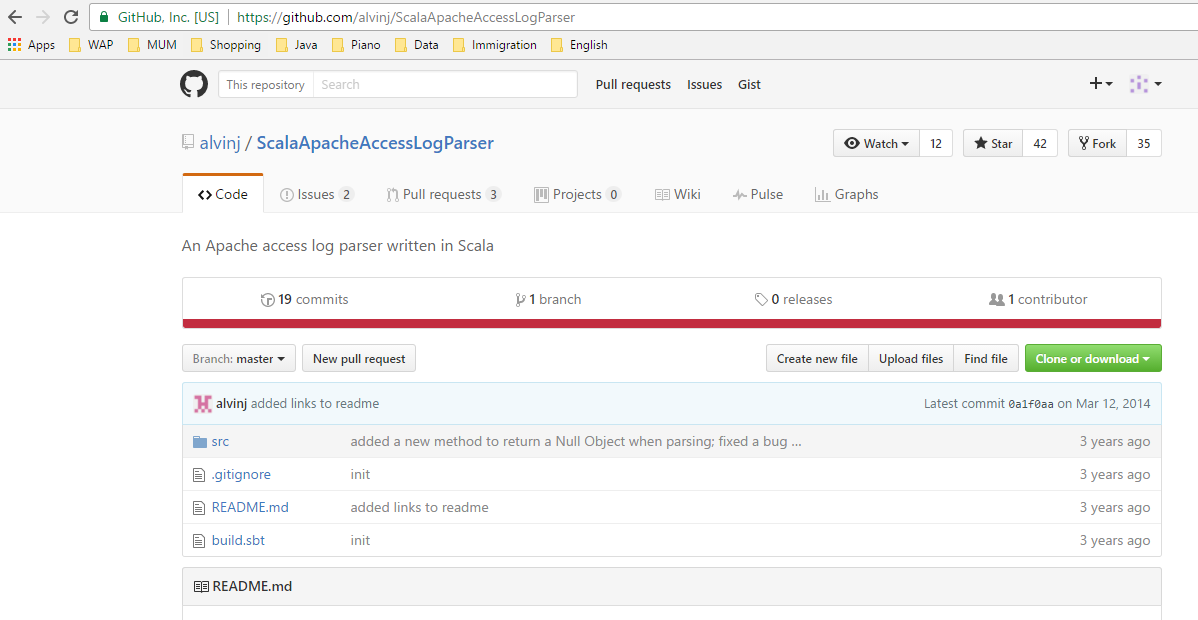
Run the project by clicking button Run:



# Spark Project

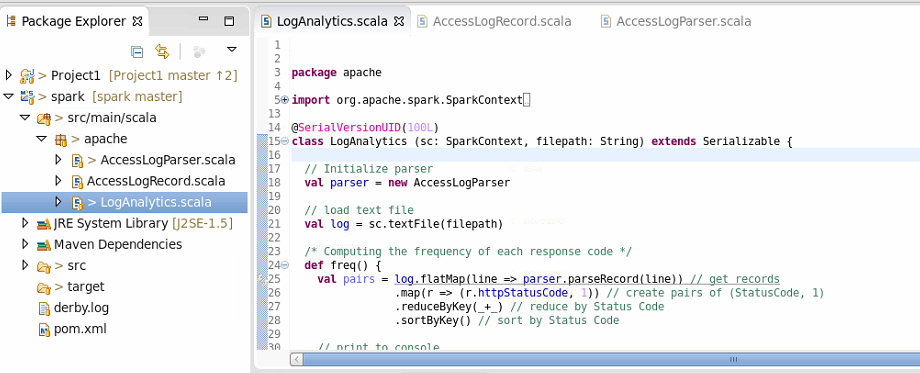
## Get the sample code:

Get the sample code from: https://github.com/alvinj/ScalaApacheAccessLogParser



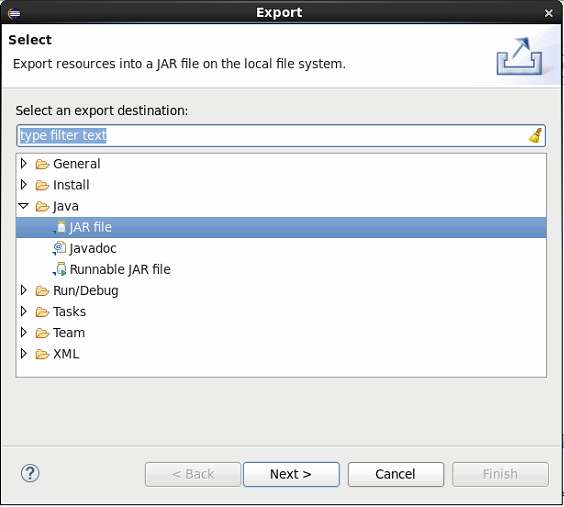
## Develop code

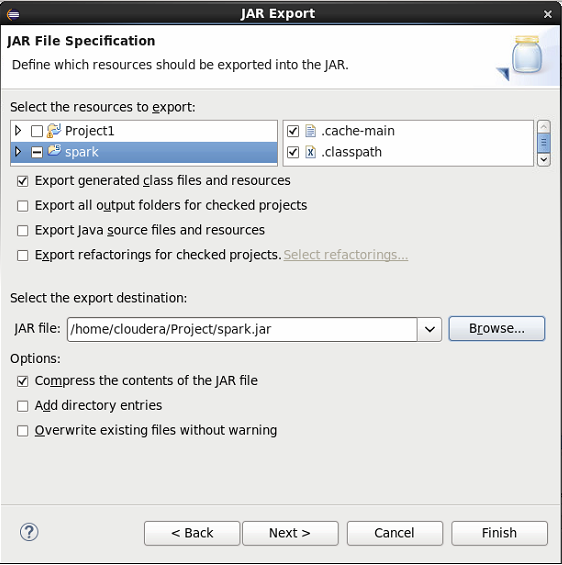
Write **LogAnalytics.scala** to do some calculation on the log:



## Run the project

Export the project to jar file:





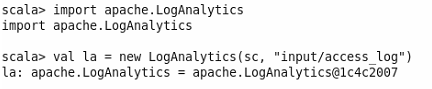
Use spark-shell command to deploy the jar file:



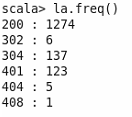
**Run some functions of in the LogAnaylytics.scala:**

Get the access\_log file from <http://www.monitorware.com/en/logsamples/apache.php> and put it the input folder of Hadoop.

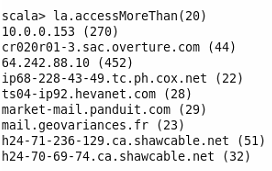
Import the Scala object LogAnalystics and use it:



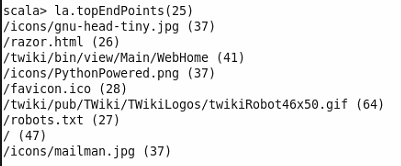
**Compute the frequency of each response code**



**All IP Addresses that have accessed this server more than 20 times**



**The top 25 endpoints requested by count**



**Calculate the average, max, min size of the content size**

