**Project 2**

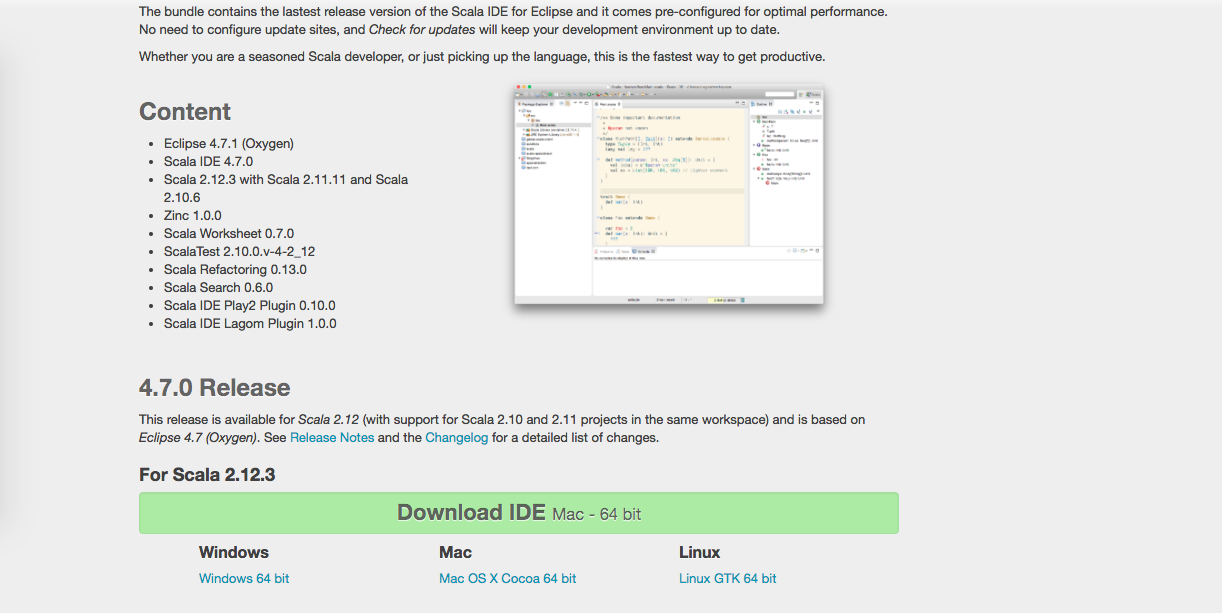
Vinh Tuong Hoang – 986023

1. Download and Install Scala IDE

Somehow, none of the version of Scala IDE works on my Cloudera virtual machine, So I need to use the Scala IDE on my Mac, and then copy the file into my Cloudera machine to execute bash file. Bellow are the steps that I’ve done:

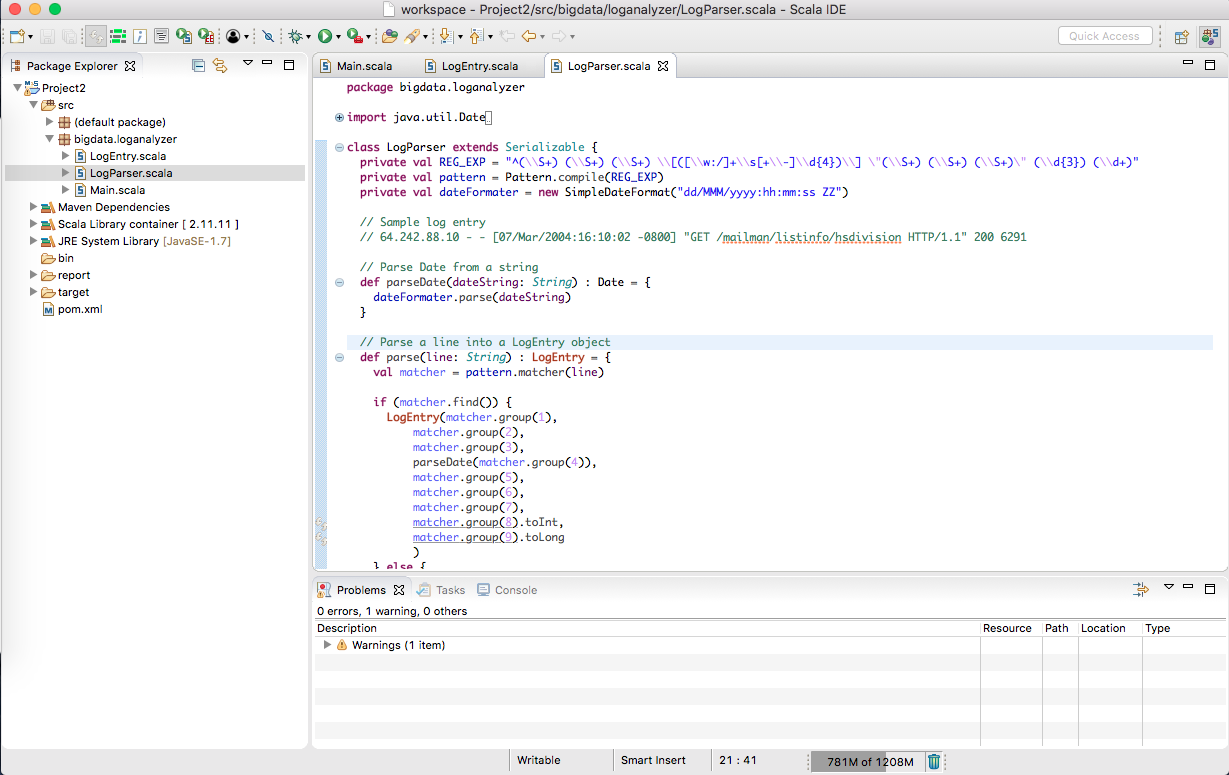
Steps:

* Go to <http://scala-ide.org/download/sdk.html> and download version for Mac OS X.
* Once download finished, unzip and double click ‘Eclipse.app’ to launch the Scala IDE.



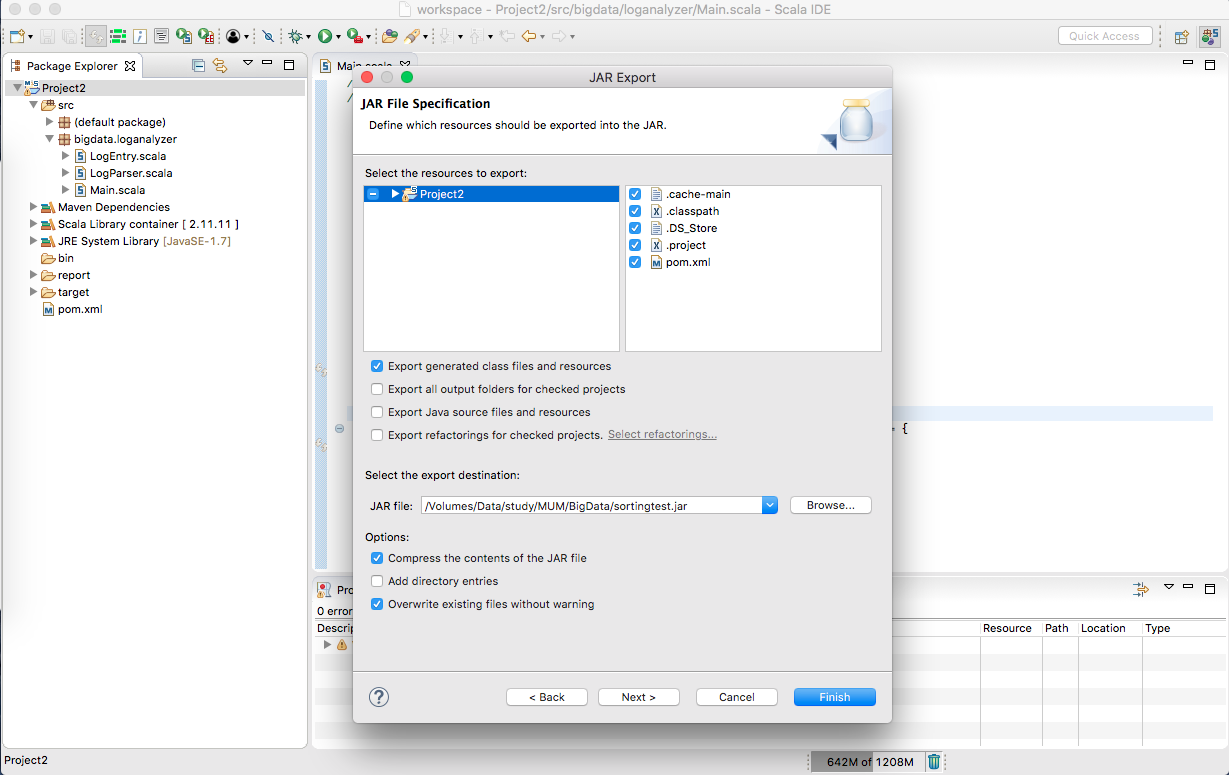
1. On Scala IDE, create new project File -> New -> Scala Project

* Create classes and packages to implement the problem.
* There 3 classes:
  + LogEntry.scala: entity which corressonding to a line of log.
  + LogParser.scala: Parses a line of log into a LogEntry object.
  + Main.scala: Execute the program with its functions



1. Export project into jar file

* Right click on the project, choose Export… -> Jar file -> Next -> Finish (as the image) to make a jar file, named project2.jar
* Copy the jar file into Cloudera virtual machine, via Google docs sharing or shared folder of Virtual Box.



1. Execute the jar file with spark-submit.

* Create a bash file, named project.sh with the content as the image. The bash file basically does the followings:
  + Copy the input file (access\_log) into HDFS
  + Call spark-submit with the class name and the jar file name
* The output files (average.output.txt) of the program is located at output/ folder on local file system.

