SE256 : Scalable Systems for Data Science

Lab Session: 2

Maven setup:
Run the following commands to download and extract maven.

wget http://www.eu.apache.org/dist/maven/maven-3/3.3.9/binaries/apache-maven-3.3.9-bin.tar.gz

tar -xvf apache-maven-3.3.9-bin.tar.gz

Update proxy settings in the config file present in /apache-maven-3.3.9/conf/setting.xml
Make sure that you uncomment proxy section only if you are using college wifi. Below is snippet of the xml file.

```
 <!-- proxies
  | This is a list of proxies which can be used on this machine to connect to the network.
  | Unless otherwise specified (by system property or command-line switch), the first proxy
  | specification in this list marked as active will be used.
  |-->
 <proxies>
   <!-- proxy
    | Specification for one proxy, to be used in connecting to the network.
    |
   <proxy>
     <id>optional</id>
     <active>true</active>
     <protocol>http</protocol>
```

```
    <username>proxyuser</username> /* required only for wifi connection*/
    <password>proxypass</password> /* required only for wifi connection*/
    <host>proxy.iisc.ernet.in</host>
    <port>3128</port>
    <nonProxyHosts>local.net|some.host.com</nonProxyHosts>
  </proxy>

 </proxies>
```

Add the following lines to bashrc file

export M2_HOME=<Path to extracted folder>/apache-maven-3.3.9/
export PATH=$M2_HOME/bin:$PATH

Run the following command
source ~/.bashrc

Install **eclipse** from their website.

First set proxy in eclipse (only for HTTP and HTTPS) and then edit /etc/eclipse.ini as per following page

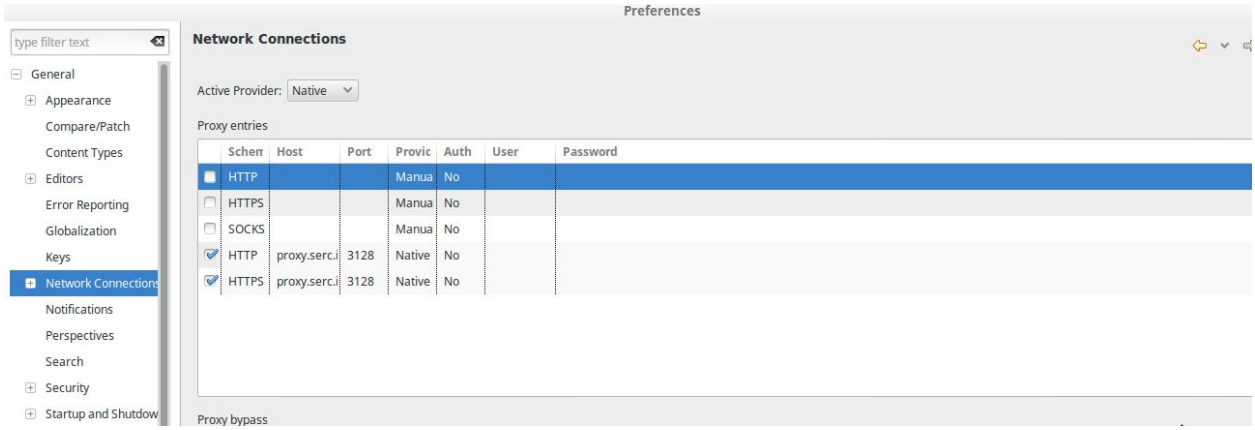
find / -name eclipse.ini /*command to search for eclipse.ini file*/

Append the following lines in the file eclipse.ini
-Dhttp.proxyPort=3128
-Dhttp.proxyHost=proxy.iisc.ernet.in
-Dhttp.nonProxyHosts=localhost|127.0.0.1

Additionally you can refer the following link for understanding how to configure proxy settings on Eclipse.

http://stackoverflow.com/questions/5857499/how-do-i-have-to-configure-the-proxy-settings-so-eclipse-can-download-new-plugin

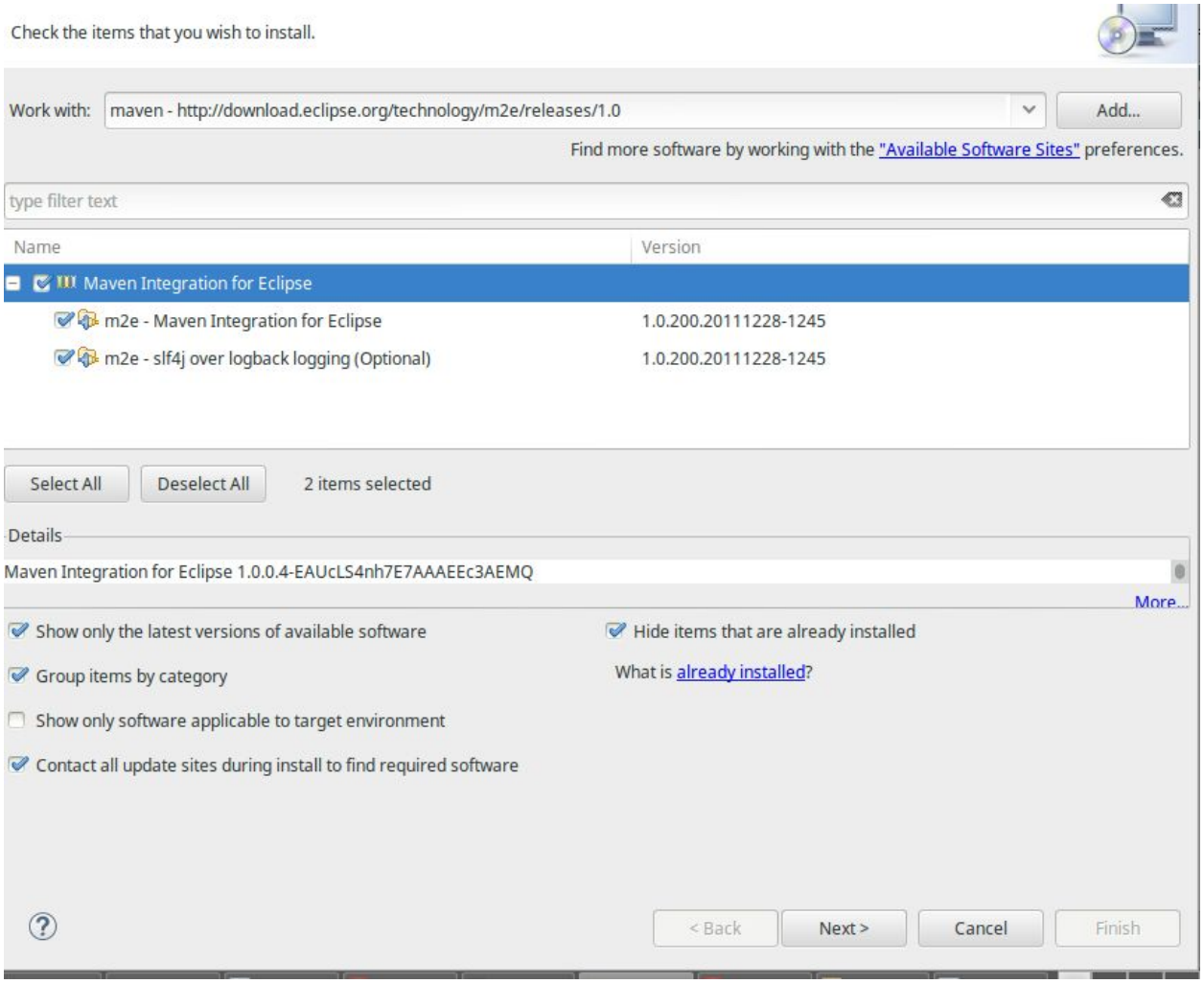
Under **windows->Preferences->General-> Network** add the proxy settings as follows:



then in eclipse install new software use link

In eclipse, go to **Help->Install New Software** , use the link given below to integrate maven in eclipse.

http://download.eclipse.org/technology/m2e/releases/1.0

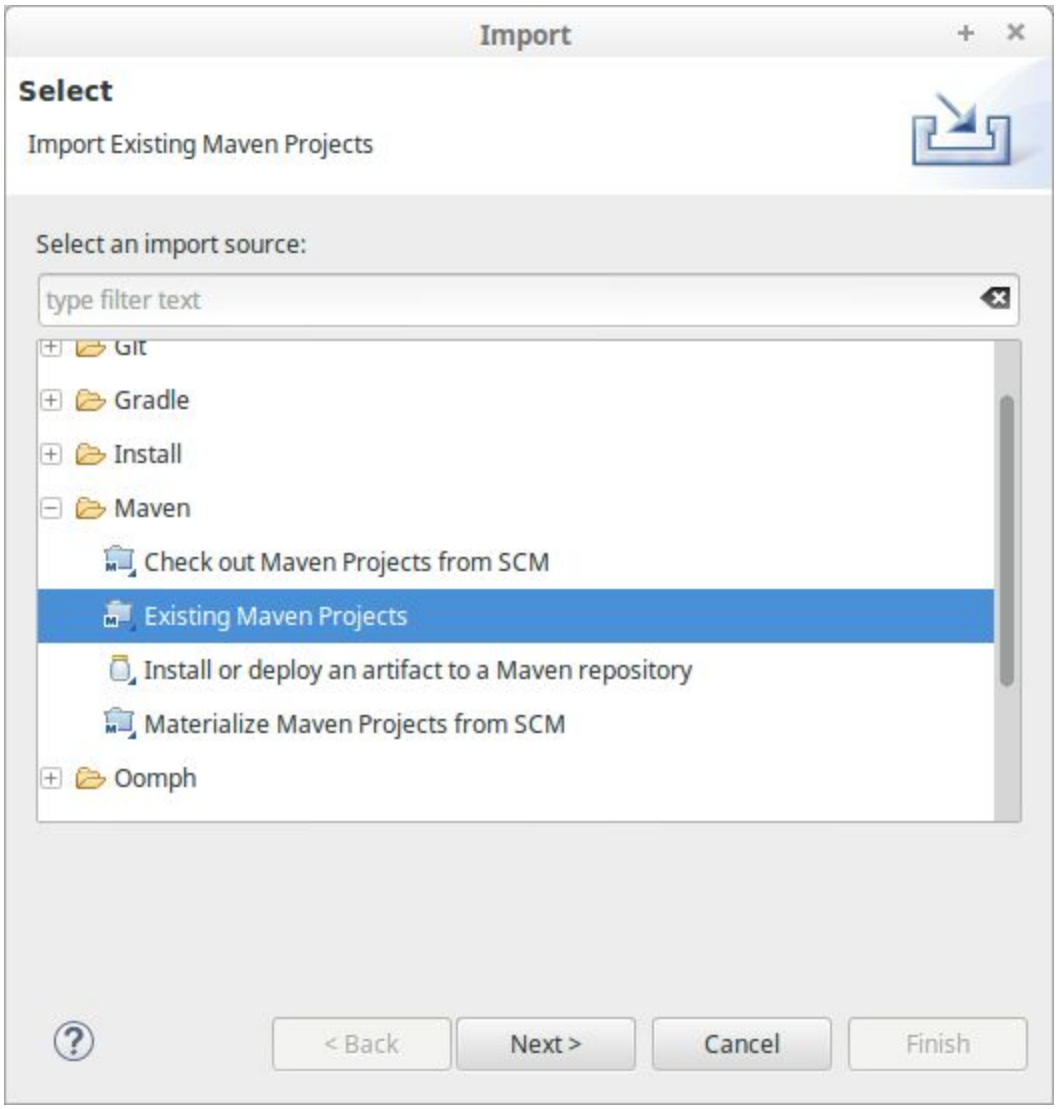


After performing the above step, go to **Window->Preferences**, expand maven from left menu, select Installations and add the path to your maven folder which you had extracted in the initial steps.

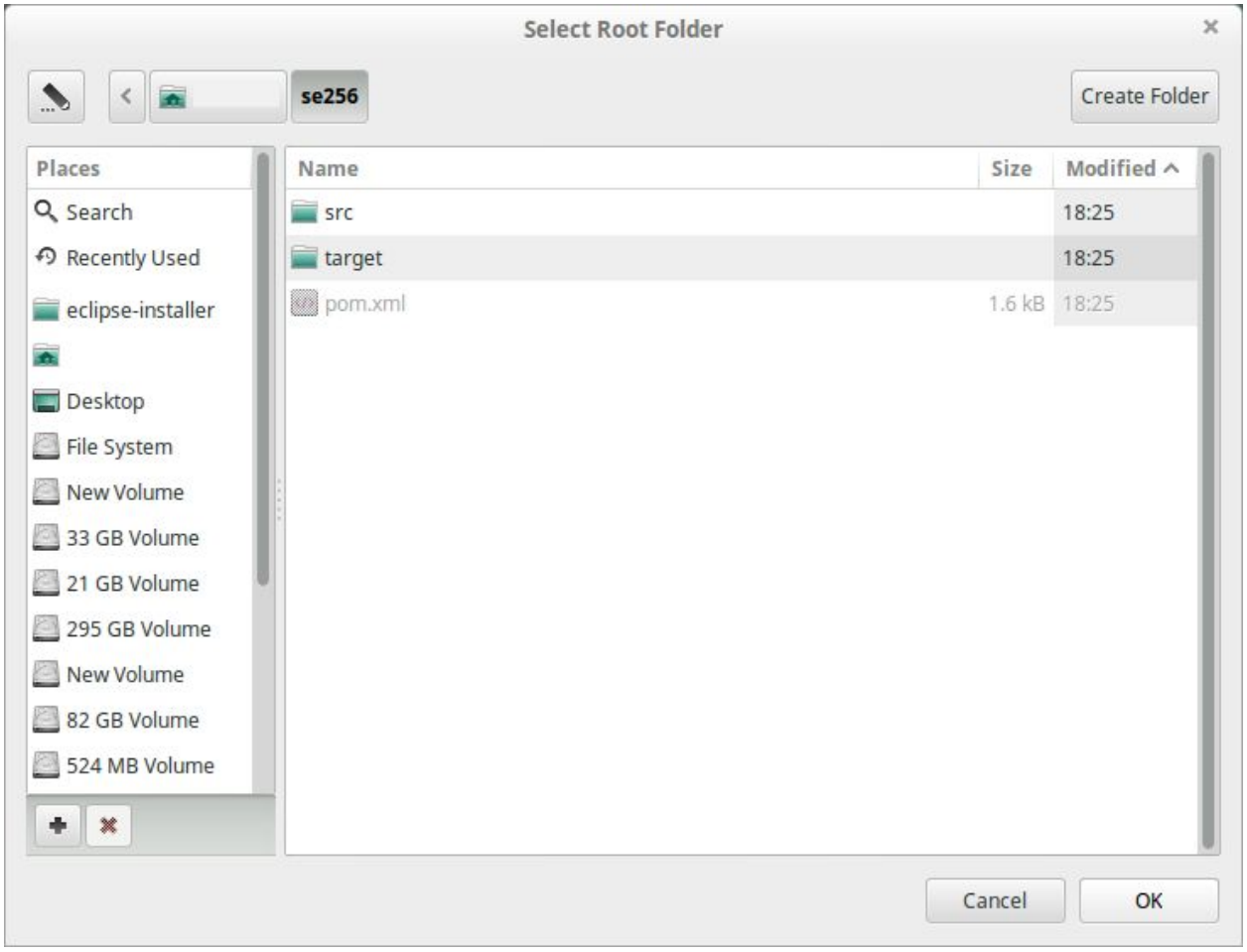
Within the maven Menu, select User Settings, now in the form opened, change user settings to point to the settings.xml file that we changed. Now apply the settings, it may take some time.

Refer to the following screenshot for this.

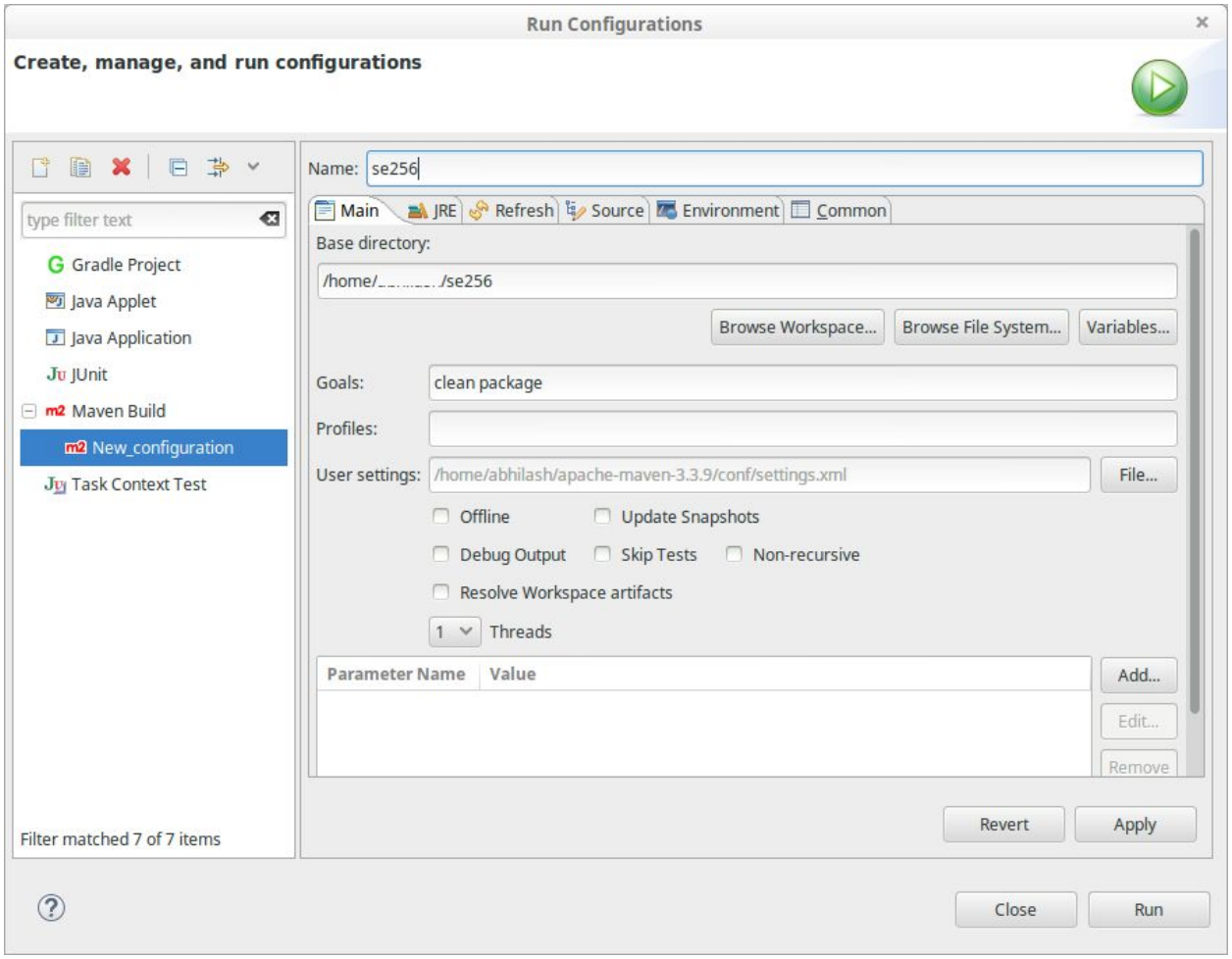
After this, we can now import a maven project, Go to **File->Import**,



Now point to path of project that you want to import, it is to be noted that the path should contain a pom.xml file which is used by maven. Refer screenshot below for example.



Now before building, go to **Run->Run configurations,** select maven build from left menu and create a configuration by entering parameters as given below.



Click Run which builds the project by creating a jar file in the target folder.

Now we can use this jar file to run hadoop application, make sure all the processes are running.

$jps

11899 DataNode
27396 Jps
11800 NameNode
12033 ResourceManager
12284 NodeManager

Now run the following commands:

Now we are going to run the wordCount example , we studied in last lab session

1. Create a text file which will be input to our program
2. Create a folder in hdfs
3. Put the input file inside this folder in hdfs

$hadoop hdfs -mkdir /WordCountInput

$hadoop hdfs -put input_file.txt /WordCountInput

From the source code provided, generate the jar file and run the following command:

$HADOOP_HOME/bin/hadoop jar /home/user/se256/target/se256-1.0-jar-with-dependencies.jar in.dreamlab.iisc.se256.Driver /WordCountInput **/output**

where /WordCountInput is path to input folder in hdfs which has input text files, output is directory in hdfs where result is generated<should not be present before running the job>.