The software packages used are:

• Eclipse (www.eclipse.org) is a software development environment that consists of an integrated

development environment (IDE) and an extensible plug-in system. It is written mostly in Java and

can be used to develop applications in Java and, by means of various plug-ins, in C, C++, Perl, PHP,

Python, R, Ruby, and several other languages. The IDE is often called Eclipse CDT for C/C++,

Eclipse JDT for Java, and Eclipse PDT for PHP.

• Apache Hadoop is a software framework that supports data-intensive distributed applications under

a free license. Hadoop was inspired by Google’s MapReduce. See Section 4.6 for a discussion of

MapReduce and Section 4.7 for an application using Hadoop.

• Cygwin is a Unix-like environment for MicrosoftWindows. It is open-source software released under

the GNU General Public License version 2. The cygwin environment consists of (1) a dynamic-link

library (DLL) as an API compatibility layer providing a substantial part of the POSIX API functionality;

and (2) an extensive collection of software tools and applications that provide a Unix-like

look and feel.

A. Prerequisites

• Java 1.6; set JAVA\_Home = path where JDK is installed

• Eclipse Europa 3.3.2

Note: the Hadoop plugin was specially designed for Europa and newer releases of Eclipse might

have some issues with the Hadoop plugin.

B. SSH Installation

1. Install cygwin using the installer downloaded from www.cygwin.com. From the Select Packages

window, select the openssh and openssl under Net.

Note: Create a desktop icon when asked during installation.

2. Display the “Environment Variables” panel:

Computer -> System Properties -> Advanced System Settings

-> Environment Variables

Click on the variable named Path and press Edit; append the following value to the path

variable:

;c:ÄcygwinÄbin;c:ÄcygwinÄusrÄbin

3. Configure the ssh daemon using cygwin. Left-click on the cygwin icon on the desktop and click

“Run as Administrator.” Type in the command window of cygwin:

ssh-host-config.

4. Answer “Yes” when prompted with sshd should be installed as a service; answer “No” to all

other questions.

5. Start the cygwin service by navigating to:

Control Panel -> Administrative Tools -> Services

Look for cygwin sshd and start the service.

6. Open the cygwin command prompt and execute the following command to generate keys:

ssh-keygen

7. When prompted for filenames and passphrases, press Enter to accept default values. After the

command has finished generating keys, enter the following command to change into your .ssh

directory:

cd˜.ssh

8. Check to see whether the keys were indeed generated:

ls -l

9. The two files id\_rsa.pub and id\_rsa with recent creation dates contain authorization keys.

10. To register the new authorization keys, enter the following command (note: the sharply-angled

double brackets are very important):

cat id\_rsa.pub   authorized\_keys

11. Check to see whether the keys were set up correctly:

ssh localhost

12. Since it is a new ssh installation, you will be warned that authenticity of the host could not be

established and will be asked whether you really want to connect. Answer Yes and press Enter.

You should see the cygwin prompt again, which means that you have successfully connected.

13. Now execute again the command:

ssh localhost

This time no prompt should appear.

C. Download Hadoop

1. Download Hadoop 0.20.1 and place it in a directory such as:

C:Java

2. Open the cygwin command prompt and execute:

cd

3. Enable the home directory folder to be shown in the Windows Explorer window:

explorer

4. Open another Windows Explorer window and navigate to the folder that contains the downloaded

Hadoop archive.

5. Copy the Hadoop archive into the home directory folder.

D. Unpack Hadoop

1. Open a new cygwin window and execute:

tar -xzf hadoop-0.20.1.tar.gz

2. List the contents of the home directory:

ls -l

You should see a newly created directory called Hadoop-0.20.1. Execute:

cd hadoop-0.20.1

ls -l

You should see the files listed in Figure 11.9.

E. Set properties in configuration file

1. Open a new cygwin window and execute the following commands:

cd hadoop-0.20.1

cd conf

explorer

2. The last command will cause the Explorer window for the conf directory to pop up. Minimize it

for now or move it to the side.

3. Launch Eclipse or a text editor such as Notepad ++ and navigate to the conf directory. Open the

Hadoop-site file to insert the following lines between the <configuration> and </configuration>

tags:

<property>

<name>fs.default.name</name>

<value>hdfs://localhost:9100</value>

</property> <property>

<name>mapred.job.tracker</name>

<value>localhost:9101</value>

</property> <property>

<name>dfs.replication</name>

<value>1</value>

</property>

F. Format the Namenode

Format the namenode to create a Hadoop Distributed File System (HDFS). Open a new cygwin

window and execute the following commands:

cd hadoop-0.20.1

mkdir logs

bin/hadoop namenode -format

When the formatting of the namenode is finished, the message in Figure 11.10 appears.