**Hadoop Lab**

**Starting Hadoop VM**

i) Enter the following URL in **Google Chrome** (only Chrome is supported currently):

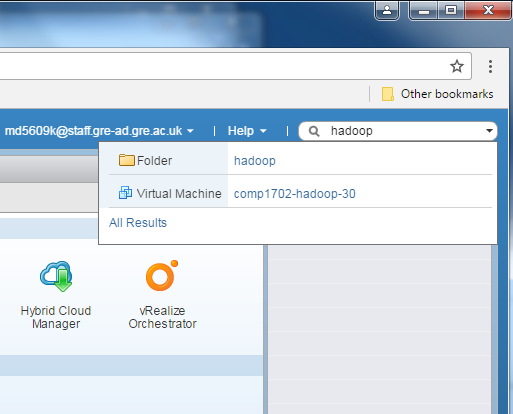
[https://classroom-vc.gre.ac.uk](https://classroom-vc.gre.ac.uk/)

ii) If you see a security message, click proceed/continue and then login with your University username and password.

If you see an error message about the Adobe Flash Player version, or get a white screen after you enter your username and password, follow the instructions in this document to fix it.

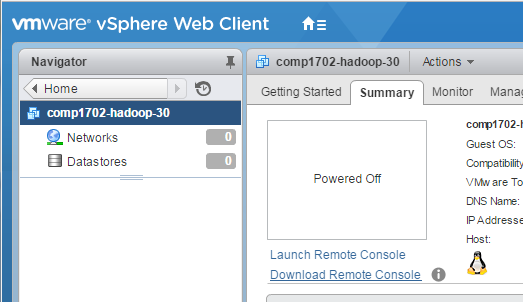
<http://ach-support.gre.ac.uk/vce/Chrome-Flash.pdf>

iii) In the VMware vSphere web client enter ‘hadoop’ in the search box on the top right hand side of the screen.

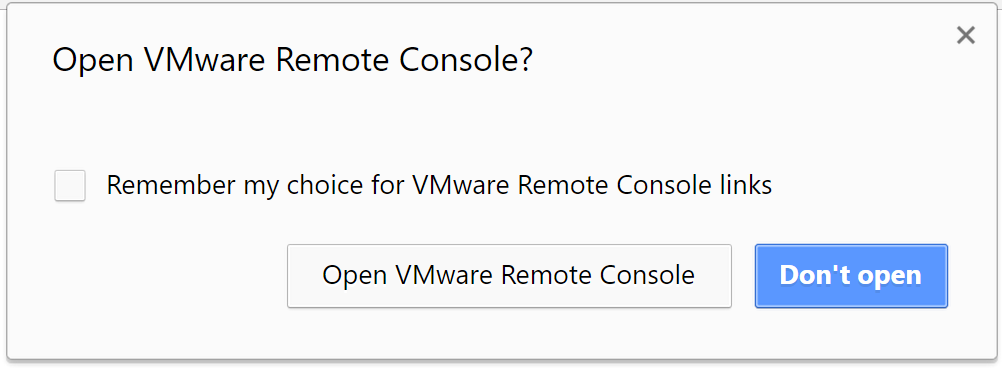


Select the Virtual Machine ‘**comp1702-hadoop-lec**’

iv) In the tab bar select ‘**Summary**’ and select ‘**Launch Remote Console**’

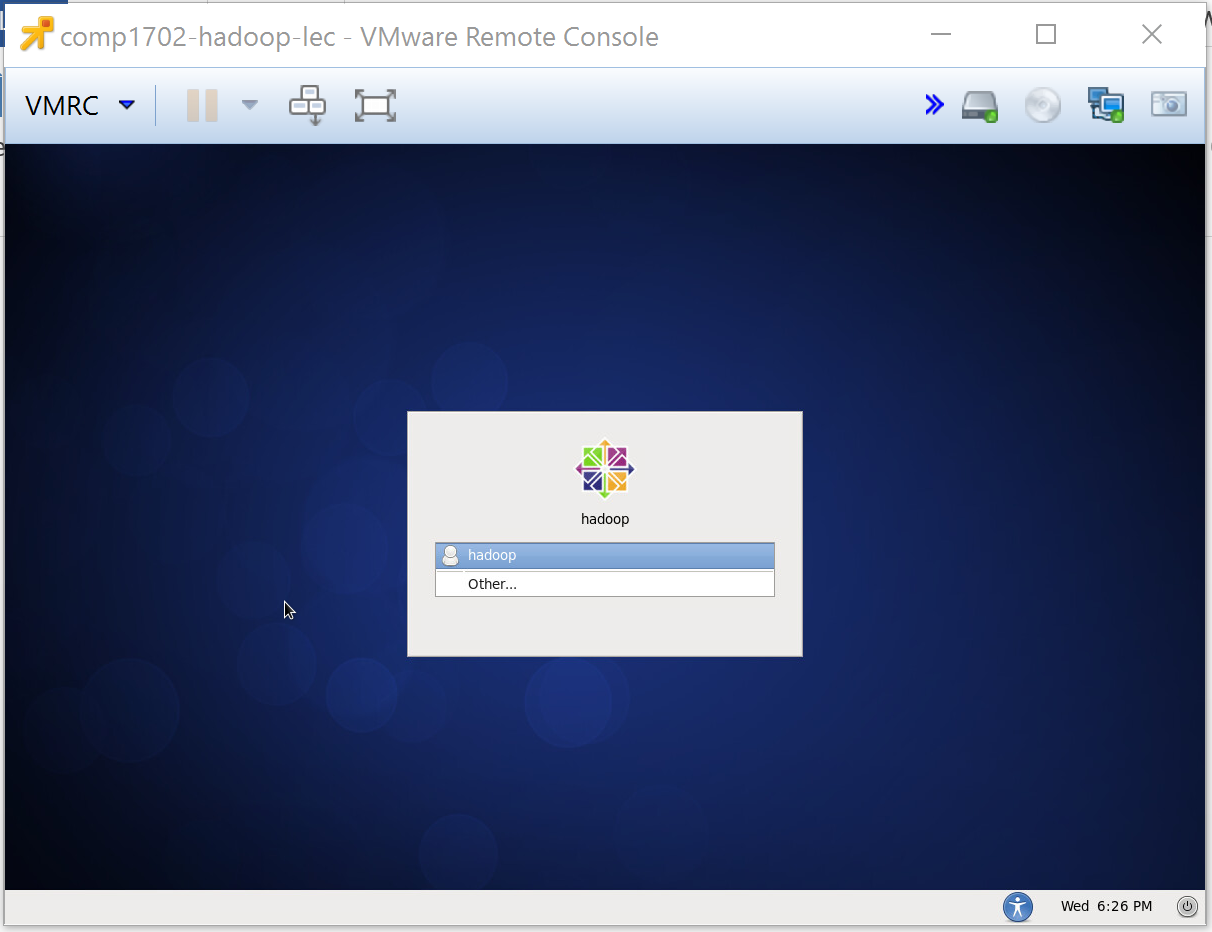


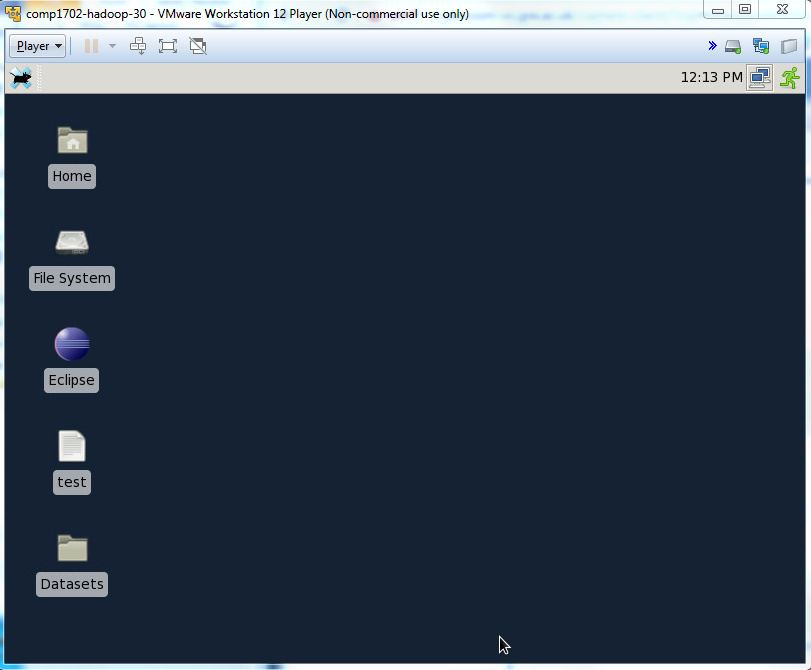
If an External Protocol Request dialog pops up select ‘Open VMWare Remote Console’



If an ‘Invalid Security Certificate’ pops up select ‘Connect Anyway’

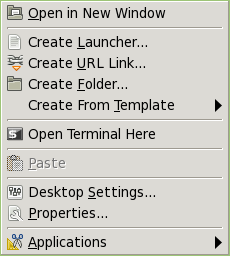
v) Login as user ‘**hadoop**’ using password ‘**hadoop**’



You should now see the hadoop desktop

**Tasks**

1. Right click on the VM’s desktop and start a terminal.



1. Start up Hadoop and the other processess, type:

**start-dfs.sh**

**start-yarn.sh**

1. Using the JDK's jps utility to see which Java processes are running, type in the terminal:

**jps**

What is shown?

Explain the difference between Local filesystem and Distributed filesystem (see lecture notes)

1. Inspect the file system. Then type:

**mkdir -p ~/Data/Books**

1. Type:

**echo "To check this out, this is a test. This is only a test." >> ~/Data/Books/test.txt**

1. Next type:

**cat ~/Data/Books/test.txt**

What has happened?

1. Upload your your data to Hadoop using the **put** command.

**hdfs dfs -put ~/Data/Books/test.txt**

Check that the data is in the HDFS

**hdfs dfs –ls**

**hdfs dfs -cat test.txt**

1. Hadoop comes with a set of demonstration programs. One of them is [**WordCount.java**](http://wiki.apache.org/hadoop/WordCount) which will automatically compute the word frequency of all text files found in the HDFS directory you ask it to process. The program has several sections:
   * The Map class takes lines of text that are fed to it (the text files are automatically broken down into lines by Hadoop), and breaks them into words. Outputs a datagram for each word that is a (String, int) tuple, of the form (“some-word", 1), since each tuple corresponds to the first occurrence of each word, so the initial frequency for each word is 1.
   * The reduce section gets collections of datagrams of the form [( word, n1 ), (word, n2)...] where all the words are the same, but with different numbers. These collections are the result of a sorting process that is integral to Hadoop and which gathers all the datagrams with the same word together. The reduce process gathers the datagrams inside a data node, and also gathers datagrams from the different data nodes into a final collection of datagrams where all the words are now unique, with their total frequency (number of occurrences).
2. Run the wordcount and put the results in to a folder called **Results**:

**hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.4.1.jar wordcount test.txt Results**

To observe the total execution time (real), include the word **time:**

**time hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.4.1.jar wordcount test.txt Results**

The time will be shown for **real, user** and **sys**, what does this refer to?

1. When mapreduce has finished, to get the Results directory type:

**hdfs dfs –get Results**

Where has the **Results** directory been moved to?

1. To read the file that has the results, cd to the **Results** directory and then type in the terminal :

**geany part-r-00000**

1. From the terminal window delete test.txt and the **Results** directory.

**hdfs dfs –rmr Results**

**hdfs dfs –rm test.txt**

1. On the desktop go to the Datasets/Lab5 directory. There are a number of books downloaded from the Gutenberg repository:

The Outline of Science, Vol. 1 (of 4) by J. Arthur Thomson (20417.txt)

The Notebooks of Leonardo Da Vinci (5000.txt)

The Art of War by 6th cent. B.C. Sunzi (132.txt)

The Adventures of Sherlock Holmes by Sir Arthur Conan Doyle (48320.txt)

1. Copy the files to the Booksdirectory you created earlier
2. Choose one of the files and load it onto the HDFS
3. Check that the data has been loaded.
4. Run the word count example as before (Hadoop will create the output folder (Results), so if you supply a folder that already exists Hadoop will fail).
5. Repeat steps 14-17 with a different book.

**Shutting down the VM**

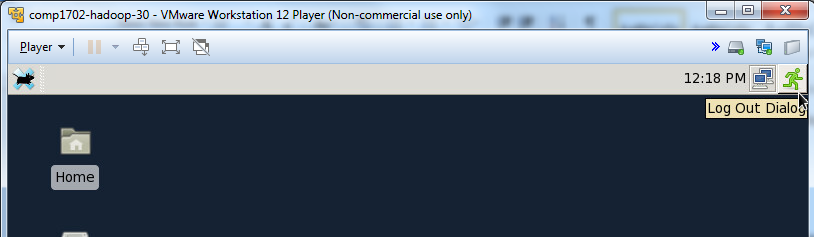
i) Stop hadoop as follows:

**stop-yarn.sh**

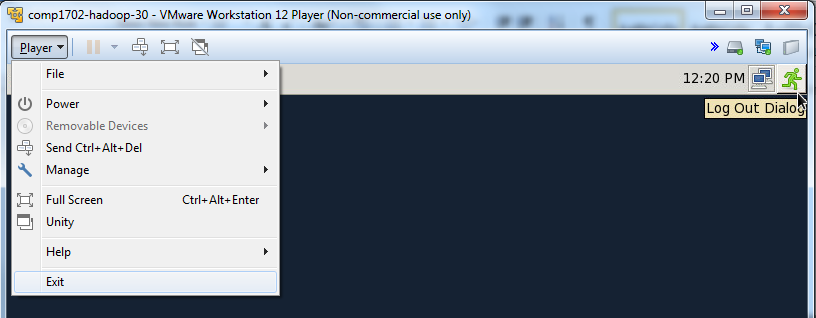
**stop-dfs.sh**

Wait for the hadoop processes to stop.

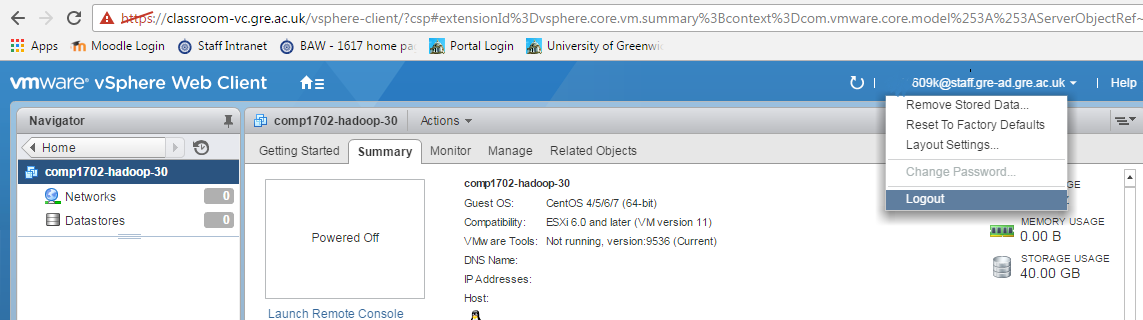
ii) Log out



iii) Exit the VMPlayer console



iv) Log off the VMware vSphere web client



Full instructions are on our support site at <https://blogs.gre.ac.uk/cmssupport/systems/classroom-environment>.