Aims:

This exercise aims to get you to:

Install and configure the environment for programming in MapReduce:

1. Download and install Eclipse (with Maven) **(if you use your own computer/laptop only)**
2. Create a new Maven project for a MapReduce example
3. Run the program and inspect the results

Download and install Eclipse (if you use your own computer/laptop only)

1. Download Eclipse from this link https://www.eclipse.org/downloads
2. From the available version

   Click on the link (A newer package is available **here.)**

3. Get Eclipse IDE for Enterprise Java Developers
4. Save File (to the default location) and install the app

Create a new Maven Project

1. Open Eclipse: If you use the lab computer, you can find it in Applications Menu -> Development -> Eclipse.
2. Choose a workspace (which is a workspace for your project and all the related files will be located here)
3. (If Maven plugin is not installed in Eclipse, you can install it from using Eclipse Marketplace before continuing with the following steps)
4. Start a new project, choose a Maven project and call it wordcount
5. You need to create a group ID, for the purpose of this exercise we use the following groupID: au.edu.unsw.cse.bdmc ; in the artifact ID enter: wordcount
6. Click Finish
7. Open the project workspace and open the file pom.xml

Get the Maven dependencies

1. Go to https://mvnrepository.com/artifact/org.apache.hadoop/hadoop-core
2. From the list choose 1.2.1 (the latest version)
3. Copy the following code into the pom.xml file in your project between two dependencies tags as follow:

```
<dependencies>
    <dependency>
        <groupId>org.apache.hadoop</groupId>
        <artifactId>hadoop-core</artifactId>
        <version>1.2.1</version>
    </dependency>
</dependencies>
```

4. Save pom.xml

Sample code for Hadoop Map Reduce exercise:

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets). In this exercise we develop a MapReduce app which counts the number of appearances of words within a text file.

1. In wordcount project open App.java (under src/main/java → au.edu.unsw.cse.bdmc)
2. In the browser search for apache Hadoop MapReduce Tutorial
3. http://hadoop.apache.org/docs//stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
4. Look for Example: wordcount v1.0
5. Copy and paste the sample code into App.java file in your wordcount project

Leave the line package *au.edu.unsw.cse.bdmc.wordcount;*

After copy the whole code, you will need to rename the App.java file to WordCount.java and save it

6. Save your file
7. Now we need to define the files location for input and output
8. These two lines of the code expect path of the input and output files

   FileInputFormat.addInputPath(job, new Path(args[0]));

   FileOutputFormat.setOutputPath(job, new Path(args[1]));

9. Create a text file and call it input.txt, you will need the file location later on, so keep in mind where you create it
10. File content example:

hello world
hello students
hello unsw
I love data
big dataset

11. To define the files location for the input and output, and run the program, right click on the main folder → choose **Run Configuration**
12. In the arguments tab enter the directory of the input file *and leave a space between the first directory and then output*
    Example:
    /tmp_amd/cage/export/cage/3/zxxxxxx/Desktop/wordcount2/input.txt Space output
13. Save the configuration
14. Run the code: You should be able to see the result in output folder within the defined directory

Note:

Please note that for this exercise we use unsw lab's computer, which have some predefined configuration (e.g., it already have Java SDK installed). If you want to install Hadoop on your own computer, you will need to do more configurations on your own.

Set up VLab to access lab computers from your own computer:

1. For the virtual machine we use VLAB , you can get it from https://taggi.cse.unsw.edu.au/Vlab/
2. Download VLab 64-bit or 32-bit based on your laptop : https://www.realvnc.com/en/connect/download/viewer/
3. The VNC Server would be ***vlab.cse.unsw.edu.au:5919***
4. Login to the lab using your ZID, ZPass