

Lab 1

[Click here to Register Attendance](https://forms.gle/pesShFDe1ibVfS3L7)

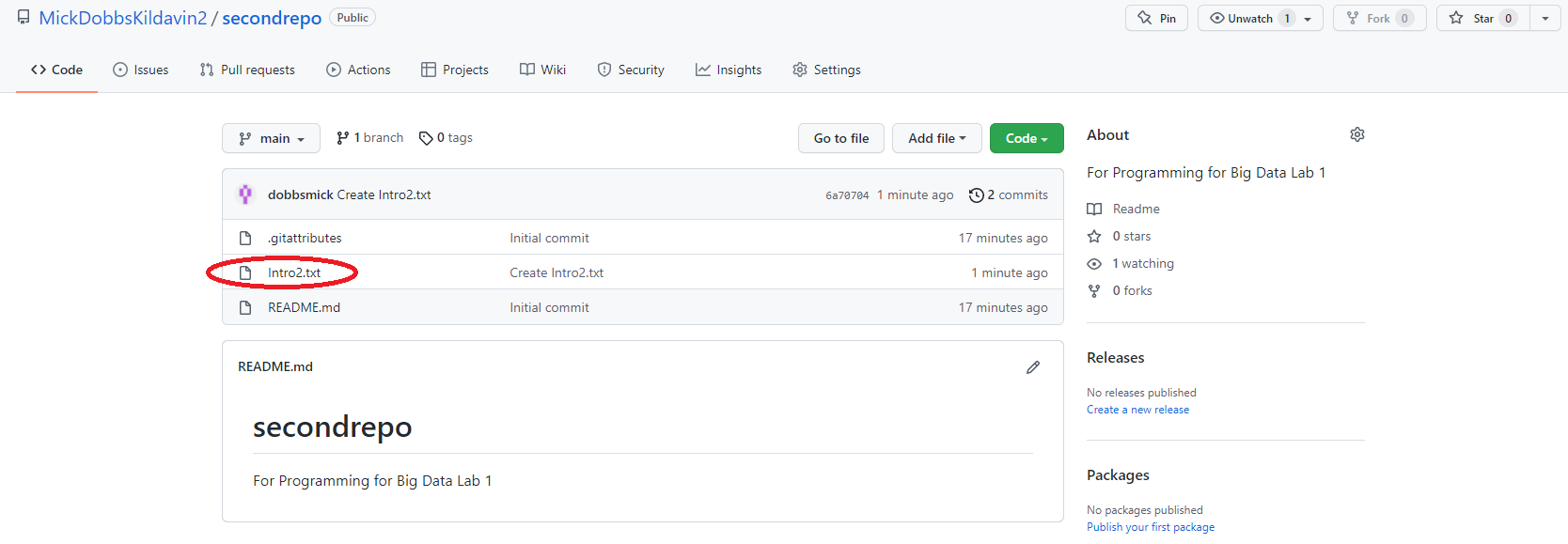
|  |  |
| --- | --- |
| Name | Mick Dobbs |
| Date | 3rd February 2022 |
| Student No | S00217208 |
| Student Email | S00217208.mail.itsligo.ie |
| When completed, please upload this Lab document to the Assessments page - Lab 1 | |

### **Part A - GitHub**

1. Go to <http://github.com>
2. Sign up for a GitHub Account (make sure to use your IT Sligo student email address when you signup as this will entitle you to a free Student that includes 5 free private repositories.
3. Enter your GitHub username here and also enter it into this [form](https://forms.gle/2WXr81guCMnnAb989):

|  |
| --- |
| MickDobbsKildavin2 |

1. Once your account has been setup, try creating a new Repository called **firstrepo** using the Add Repository button below your profile picture.
2. Use the Add File button to add a new file to your repository, give your new file a name, type some text and save the file.
3. On your Laptop/Desktop, install GitHub Desktop from here: <https://desktop.github.com/>
4. Using GitHub Desktop, Clone a repository from your GitHub account to your laptop/computer
5. On your Laptop/Desktop, create a repository called **secondrepo**
6. In your secondrepo folder add a file and save it, then **commit** the change and **push** it to your GitHub account
7. Paste a screenshot here of your pushed **secondrepo** repository on GitHub.com



1. Part B - Choose one of B1 or B2 (next page)

### 

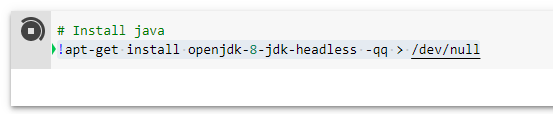
### **B2 - Installing Hadoop on Google Colab & Running a MapReduce op**

1. Create a new Notebook in [Google Colab](https://www.google.ie/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwj36-WQ1ND1AhXTVsAKHet9AWUQFnoECAUQAQ&url=https%3A%2F%2Fcolab.research.google.com%2F&usg=AOvVaw3A5aPK2kLFzKOzb6sOckVw) and call it Hadoop - Getting Started
2. Add and run the following cells:

|  |
| --- |
| # Install java  !apt-get install openjdk-8-jdk-headless -qq > /dev/null |
| #create java home variable  import os  os.environ["JAVA\_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"  os.environ["SPARK\_HOME"] = "/content/spark-3.0.0-bin-hadoop3.2" |
| #download hadoop  !wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz |
| #we’ll use the tar command with the -x flag to extract, -z to uncompress,  #-v for verbose output, and -f to specify that we’re extracting from a file  !tar -xzvf hadoop-3.3.0.tar.gz |
| #copying the hadoop file to user/local  !cp -r hadoop-3.3.0/ /usr/local/ |
| #finding the default Java path  !readlink -f /usr/bin/java | sed "s:bin/java::" |
| #Running Hadoop  !/usr/local/hadoop-3.3.0/bin/hadoop |
| !mkdir ~/input |
| !cp /usr/local/hadoop-3.3.0/etc/hadoop/\*.xml ~/input |
| !ls ~/input |
| !/usr/local/hadoop-3.3.0/bin/hadoop jar /usr/local/hadoop-3.3.0/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar grep ~/input ~/grep\_example 'allowed[.]\*' |
| Adapted from <https://www.analyticsvidhya.com/blog/2021/05/integration-of-python-with-hadoop-and-spark/> |

1. Write a short description of what you understand is going on in each cell
2. Add a screenshots of each cell executing

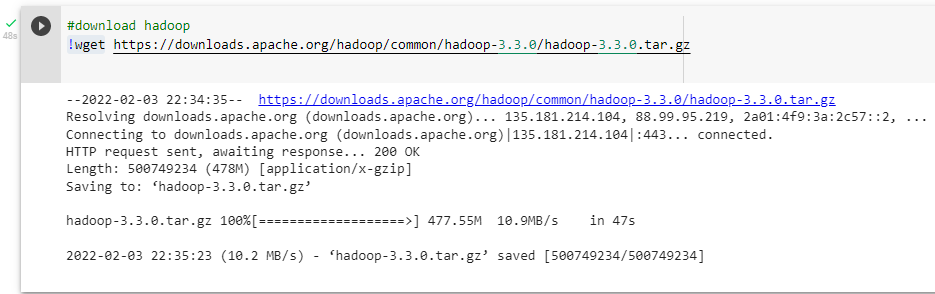
* **Installing Java Dev Toolkit.**



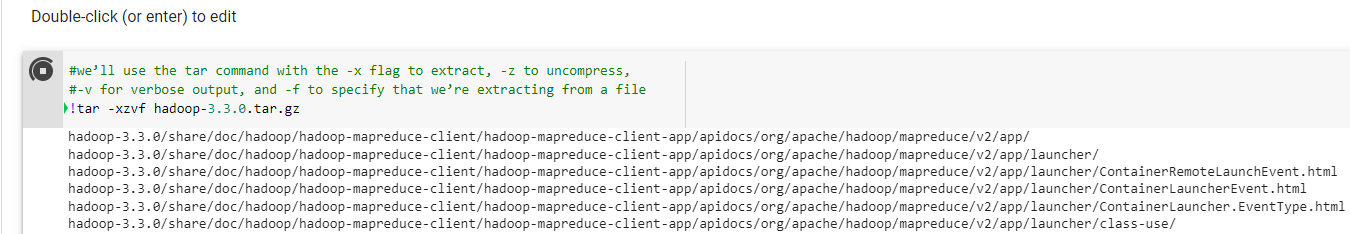
* **Import python OS module**
* **Set home directories for Java and Apache Spark**



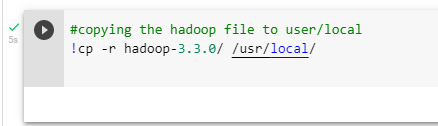
* **Use wget to retrieve a tar file for Hadoop install**



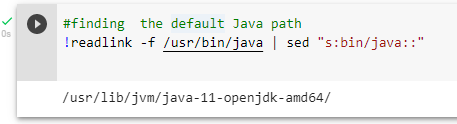
* **Extract and uncompress the install files from the tar files**

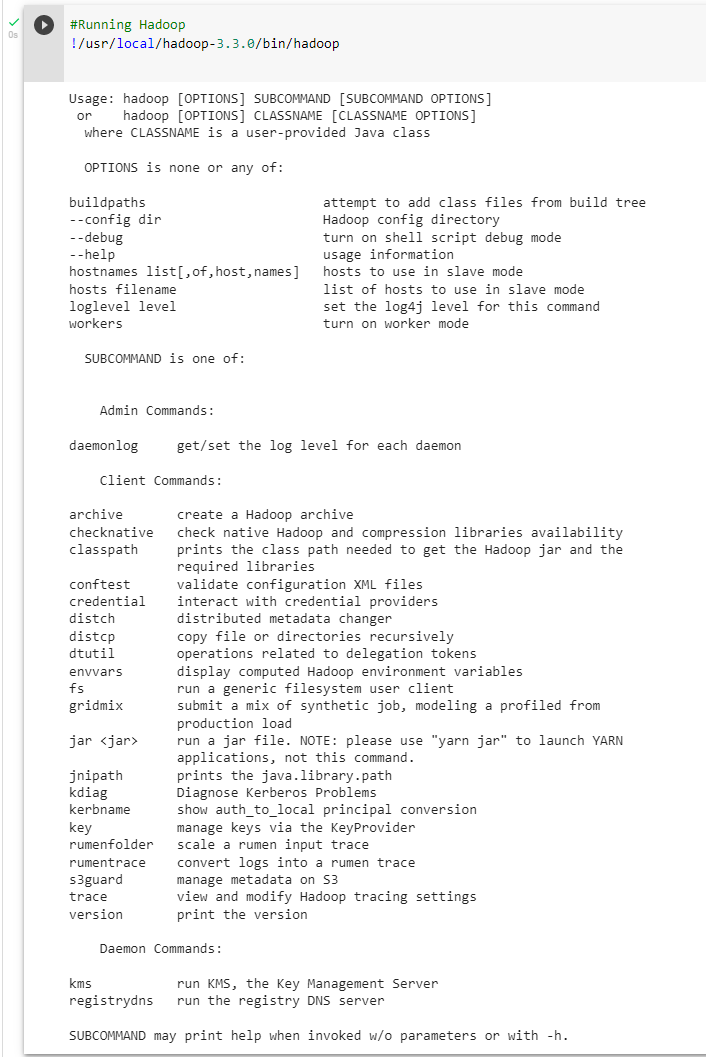


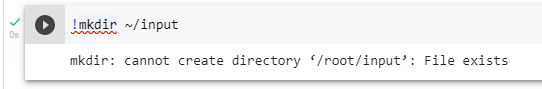
* **Copy the extracted Hadoop files to the local path**



* **Get the java path**

****

* **Run Hadoop in Colab**
* **Creating an input folder, failing because I ran it already**



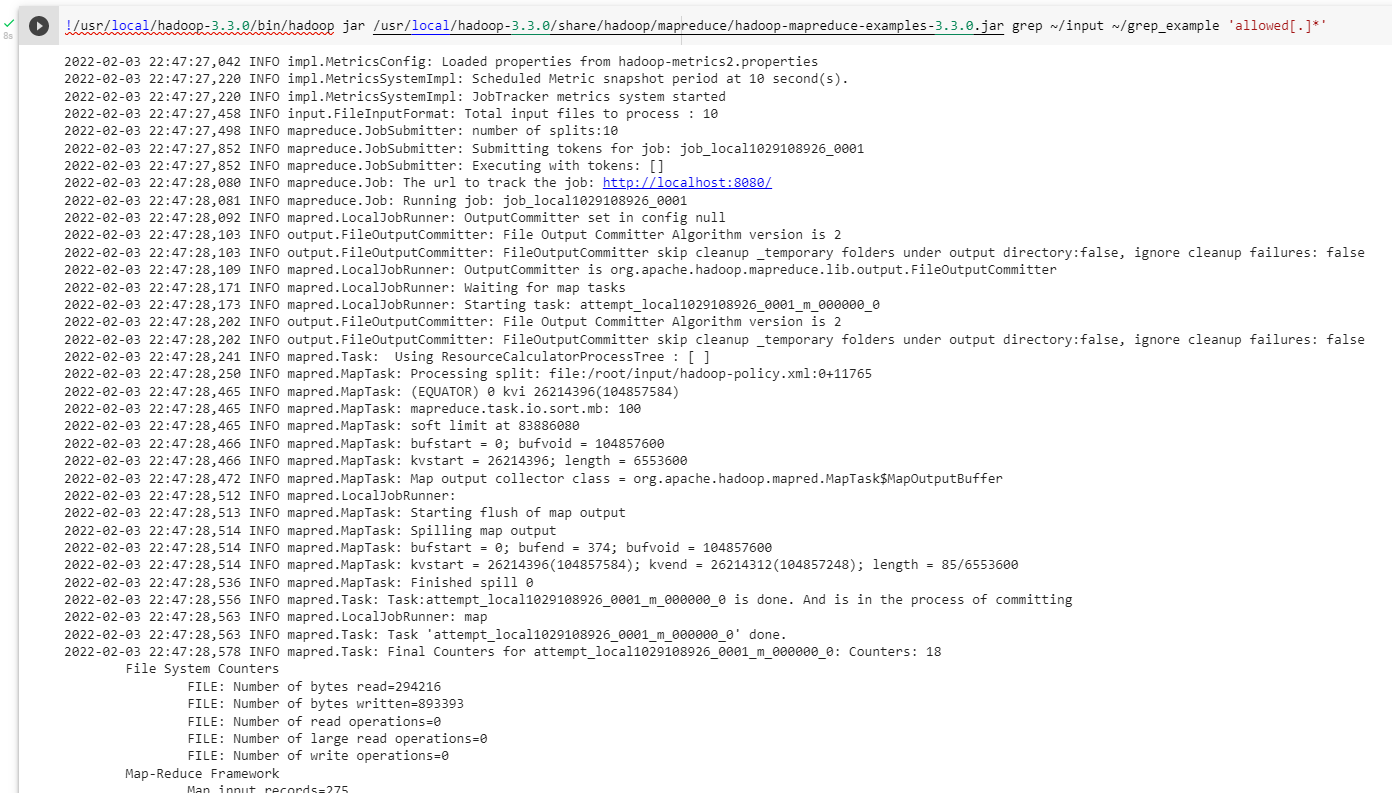
* **Copying all Hadoop xml files to the new input folder, already copied**



* **List all the copied xml files**



* **Use grep to print all files matching the given pattern**



1. Include a URL to your working Google Colab notebook here:

|  |
| --- |
| [Untitled4.ipynb - Colaboratory (google.com)](https://colab.research.google.com/github/MickDobbsKildavin2/firstrepo/blob/main/Hadoop%20-%20Getting%20Started.ipynb) |