

Lab 2

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

| Name |  |
| --- | --- |
| Date |  |
| Student No |  |
| Student Email |  |

### **Running a MapReduce Job on your local machine**

1. Go to the UCI Machine Learning Data Repository and explore the datasets available
2. Download a dataset of your choice and extract the files (I used the [Iris dataset](https://archive.ics.uci.edu/ml/datasets/iris))
3. Download the python [file here called MapReduceIris.py](https://github.com/marloftitsligo/ProgrammingForBigData/tree/main/MapReduceIris) and place in the same folder as your dataset
4. Open a Command Line (assuming you have Python installed)
5. Run the following command (replace iris.data with your data filename):  
   python MapReduceIris.py iris.data
6. You will likely get an error message as this python script uses a library called mrjob.   
   To install the library type:  
   pip install mrjob  
   You should get a confirmation message: Successfully installed mrjob-0.7.4
7. Try running the original command again (replace iris.data with your data filename):  
   python MapReduceIris.py iris.data
8. You should get the following in the output:   
   "setosa sepal width avg" 3.418
9. Post a screenshot of the output here:

|  |
| --- |

1. Try to adjust the Reducer calculation to find the average of other Species  
   Describe the rationales for your changes, your code and screenshot your output here:

|  |
| --- |

1. Find another dataset to use with this MapReduce code and make necessary changes to the code to perform an analysis of your choice.   
   Insert the name and a link to the dataset here and a short description of the analyses you performed:

|  |
| --- |

1. Push the code and screenshots from step 11 above to Github and post the link here:

|  |
| --- |

1. Brainstorm some ideas for your Programming for Big Data Project this semester

|  |
| --- |