

Laboratory Assignment 1.2 - Putting your cluster to work!

- Due Sep 27 by 11:59pm
- Points 100
- Submitting a text entry box or a file upload
- Available Sep 15 at 12am - Sep 27 at 11:59pm

This assignment was locked Sep 27 at 11:59pm.

Laboratory Assignment 1.2 - Let's get to work!

Prerequisites:

- Functional Hadoop / Jupyter / Dask / pyspark etc. cluster. The procedure I provided for Docker will work best with this lab.
- Knowledge of the system, architecture, layout.
- Jupyter Notebooks knowledge.
- Some Python, experience with Python package includes.

Purpose:

The purpose of this lab is to demonstrate the use of your cluster. At this time, you likely have a "3 node system" composed of a master, worker, and edge node. While this isn't distributed per se, we can test our cluster functionality before expanding in a later lab. This is good practice as we want to ensure everything is functional before expanding.

Procedure:

For this lab, you'll need to install matplotlib. Matplotlib is used in conjunction with other packages like Numpy, to plot graphs, charts, and more. To install this, you will need to login / change user to root on your worker node and then install matplotlib with the following command: `conda install matplotlib` Give it a moment to install.





Okay, let's get to work!

Your cluster includes a few technologies which will make our lives as Data Engineers easier. In order to put the cluster to work, we must first login. Open a browser, navigate to `localhost:8888`, and login using the test accounts, i.e. "alice" with "testpass". Reminder, do not port-forward these ports to the internet or place this cluster in AWS unless you absolutely know the security implications.

Please note: I had to create text files as Canvas can interfere with the formatting. Download the file, copy the contents, and then run it in Jupyter.

1. Let's put it to work!
2. Starting small, the first test is to see if everything is working. In order for this to fully work, please install matplotlib on your worker node as it will be needed for the graphing parts of this lab. You can

install it as root using "conda install matplotlib"

3. Once installed, you are ready to continue.
4. Open your cluster homepage and login as Alice with testpass.
5. Navigate to the lab page with jupyter notebooks.
6. Click new notebook and you'll copy - paste the code into the code blocks, then click the play button to execute it.
7. Open this file: [ClusterTest_1.txt \(https://pacific.instructure.com/courses/123868/files/30479755?wrap=1\)](https://pacific.instructure.com/courses/123868/files/30479755?wrap=1)  [\(https://pacific.instructure.com/courses/123868/files/30479755/download?download_frd=1\)](https://pacific.instructure.com/courses/123868/files/30479755/download?download_frd=1)
Copy this code into your notebook, and run it. Note the results.
8. Open this file: [ClusterTest_2.txt \(https://pacific.instructure.com/courses/123868/files/30479757?wrap=1\)](https://pacific.instructure.com/courses/123868/files/30479757?wrap=1)  [\(https://pacific.instructure.com/courses/123868/files/30479757/download?download_frd=1\)](https://pacific.instructure.com/courses/123868/files/30479757/download?download_frd=1)
Copy the code into the notebook, run it, and note your results.
9. Now - let's test the cluster's performance. This first example is a word search in single thread mode. I.e. one core one worker.
10. Open this file: [ClusterTest_3.txt \(https://pacific.instructure.com/courses/123868/files/30479758?wrap=1\)](https://pacific.instructure.com/courses/123868/files/30479758?wrap=1)  [\(https://pacific.instructure.com/courses/123868/files/30479758/download?download_frd=1\)](https://pacific.instructure.com/courses/123868/files/30479758/download?download_frd=1)
Copy the code into the notebook, click run, and you will want to wait a moment. It can take up to 2-3 minutes if the system is running slowly. Note your findings and approximately how long it took.
11. Last, let's use pyspark to pseudo distribute this. Open this file: [ClusterTest_4.txt, \(https://pacific.instructure.com/courses/123868/files/30479759?wrap=1\)](https://pacific.instructure.com/courses/123868/files/30479759?wrap=1)  [\(https://pacific.instructure.com/courses/123868/files/30479759/download?download_frd=1\)](https://pacific.instructure.com/courses/123868/files/30479759/download?download_frd=1) copy the code into the notebook, and run it noting the results. This should be a bit faster.
12. You may want to run additional tests. Chapter 2 in Zybooks has many examples which should work with your cluster. I would recommend playing with this cluster to understand how it works. We will later expand this into a few more worker nodes to speed up the performance.

Deliverables:

You aren't required to submit any code / screenshots for this lab. However, I would like for you to write a one-page summary about the cluster and the time it takes to execute these tasks. In lab 2 (in a few weeks) we will expand this cluster and note the speed up as we add worker nodes. Ideally, each node we add should result in roughly a 2x speed increase. Though, this will vary with hardware differences, potentially different configurations, etc.

Do not discard this lab! Ensure you save the notebook and any relevant files. We will use these again in lab 1.3 and in our later lab 2. Also, if you made any changes to the Docker cluster, remember to commit your changes before you power off / discard the containerized machines. If you don't your work will be lost!