Laboratory Assignment 1.1 - Designing, Building, and Using an Apache Hadoop Cluster + Jupyter

- Due Sep 27 by 11:59pm
- Points 100
- Submitting a text entry box, a website url, or a file upload
- Available Sep 8 at 12am - Sep 27 at 11:59pm

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

Laboratory Assignment #1.1 - Designing, Building, and Using an Apache Hadoop Cluster

This is a Group Assignment. Please use the Canvas Groups feature to sign up as groups. There will be seven (7) groups of five (5). Ideally, your groups should be composed in the following way: A Group Leader, group communicator, group engineer / technical leads, group story teller / analyst. At a minimum, one person in your group should have a computer capable of the procedure below. If this is a challenge, please let me know as soon as possible.

+++++++++

For this assignment, you will need to follow the procedures outlined below. While this works for most systems, I am unable to test each version of Windows or hardware platform. However, the technologies mentioned below are well supported in a Windows environment.

Important Note for Apple / Mac users: There are challenges with the M1 / M2 chipsets for part of this lab. The problem is the virtualization of any operating systems is blocked in tools like VM Ware, Virtual Box, and in Docker. You are welcome to use an Apple computer, but I cannot offer much support due to this limitation.

Prerequisites:

- Ideally, a Windows 10+ computer with at least 8Gb of RAM, 256Gb storage NVME / SSD, and modern Intel CPU.
- Virtualization will need to be enabled for your specific platform. This is set in the BIOS.
- Windows Subsystem for Linux or WSL installed. Use Version 2 for full support.
- Installed / Working Virtual Box or Docker Desktop.
    - Note: Virtual Box will require a substantial amount of resources to run. I recommend using Docker Desktop.
- Download of Docker Images, Ubuntu OS ISO, and any related tools you wish to use - i.e. Putty, Win SCP, etc.
- I recommend using GIT for managing your work, code, downloads, and more.

Procedure:

This procedure may change slightly due to bugs, errors, or other necessary changes. Due to the nature of computing and the challenges that may arise, the procedure will be adjusted to match any workarounds or known issues. Please be sure to check often to see if any new changes are made. I will be very clear what changes are new.

1. Start off by reading the procedure found here: **https://jupyterhub-on-hadoop.readthedocs.io/en/latest/demo.html** (https://jupyterhub-on-hadoop.readthedocs.io/en/latest/demo.html) At the time of this writing, the procedure works well and I was able to get JupyterHub running with a working notebook.
2. When you are ready to proceed, using Git, clone the repository to your computer. You may wish to keep this in a directory called "ANLT214" or similar. This way you can keep track of all files and this will make it easier when we port data into our system.
3. Once the repo is cloned, you should review the docker-compose.yaml file. While this is not a Docker class, you should become familiar with the format should you need to make any changes.
4. In Docker at the bottom, click Terminal and "cd" to your directory. I.e. cd "C:\ANLT214\jupyter-on-hadoop\docker-demo\" directory.
5. Next, issue the following command: "docker-compose up -d" this will launch the cluster. Note: You will need to be signed into Docker in order to access the repositories. Issue this command to login: "docker login". This will alleviate any "Access Denied" errors.
6. The cluster will build, on slower internet connections, this may take a minute or two to download the images. Once the downloads are complete, the cluster will be brought up.
7. You should see a successful start with no errors. If you see errors, troubleshoot these in order. Normally, it's typos in the commands or you might be in the wrong directory.
8. When you are ready, open a web browser, navigate to localhost:8888 and you'll be presented with the Jupyter login. The login I used was "alice" with the "testpass" password.
9. If you are successfully logged in, you may proceed to the next part.
10. If / when you need to shutdown the cluster, issue the following command: "docker-compose down" IMPORTANT! Save all work before doing this! When you shutdown the cluster, the data may / may not persist. You can export your work from the cluster for future work. Save this procedure to recreate the cluster as we will use it again.

Navigation - Your new Jupyter Cluster and Hadoop

1. You will want to browse around the cluster. Two URLs are important: **http://localhost:8888/user/alice/tree** (http://localhost:8888/user/alice/tree) (this is your overall view) and **http://localhost:8888/user/alice/lab** (http://localhost:8888/user/alice/lab) (this is where you will work with your Jupyter notebooks / projects)
2. At this point, you should have a working cluster and can easily navigate. In the next part of the lab, we will analyze data and perform a few operations. I recommend reviewing Chapter 2, Section 1 in ZyBooks for how to use Jupyter Notebooks.

Deliverables:

None at this time as this lab is to configure the cluster and ensure it is working. Lab 1.2 and on will have deliverables.

This is all for now. Expect Part 1.2 of this lab to be posted the week of September 16th. We will analyze a few datasets using this cluster and Jupyter Notebooks. Practice use and become familiar with this cluster. The author provides great detail in the write-ups and YouTube videos. Thank you to Jim Crist for this demo and making it public! This expedites the deployment of this cluster and can be modified for many different purposes.