Apache Hadoop and Spark Setup in CSB120

Installation Guide CS535 Big Data Spring 2023

STEP 1: Setup Passwordless SSH
If you are able to ssh between CS120 lab machines without using your password you can skip this step. Login into a CS120 lab machine and open up a terminal and run the following commands. You should be able to ssh into machines without using your password after running these commands.

```
~$ cd
~$ ssh-keygen
~$ ssh-copy-id -i ~/.ssh/id_rsa.pub $HOSTNAME
```

STEP 2: Setup Hadoop and Spark configuration files
Please refer to the CS535_SP23_nodes_ports.pdf from the PA1 page on Canvas for a list of your assigned nodes and ports. Pick one node from your assigned nodes as your master node and another, different node for your second node. Use the remaining 8 nodes for your workers. Download the setup_hadoop_spark.sh script from the PA1 page on canvas and run from the command line in your home folder.

```
~$ cd
~$ bash setup_hadoop_spark.sh
```

STEP 3: Load pa1 into your .bashrc
Edit .bashrc file, add *"source /etc/profile.d/modules.sh", "module purge", and "module load courses/cs535/pa1"* as the last 3 lines and save the file. If you have any other modules being loaded in your .bashrc remove/comment them out.

```
~$ vim ~/.bashrc
      press "i" for edit mode
  +   source /etc/profile.d/modules.sh
  +   module purge
  +   module load courses/cs535/pa1
      press escape
      type":wq"
      press enter
```

STEP 4: Update .bashrc
Reflect changes in .bashrc file. You only have to do this once after making any changes to your .bashrc file.

```
~$ source ~/.bashrc
```

STEP 5: Confirm module is loaded
Verify that module is loaded, should output "1) courses/cs535/pa1"

```
~$ module list
```

STEP 6: Start/Stop Hadoop/Spark Cluster
You should now be able to start your Hadoop/Spark cluster with the following commands "$HADOOP_HOME/sbin/start-dfs.sh", "$HADOOP_HOME/sbin/start-yarn.sh", and "*start-all.sh*".

```
~$ $HADOOP_HOME/sbin/start-dfs.sh
~$ $HADOOP_HOME/sbin/start-yarn.sh
~$ start-all.sh
```

You can stop the Hadoop/Spark cluster with the command "*stop-all.sh*", "$HADOOP_HOME/sbin/stop-yarn.sh", and "$HADOOP_HOME/sbin/stop-dfs.sh" commands

```
~$ stop-all.sh
~$ $HADOOP_HOME/sbin/stop-yarn.sh
~$ $HADOOP_HOME/sbin/stop-dfs.sh
```

STEP 7: Run Spark job

- Spark applications can be launched using spark-submit script.
- Change directory to your project folder.
- Run the following command with appropriate values.

```
~$ spark-submit --class <your Class> --deploy-mode cluster --supervise
<yourJar> <any_arguments>
```

You can refer http://spark.apache.org/docs/latest/submitting-applications.html for more information.