Run a YARN Job and Observe

|  |  |
| --- | --- |
| **Exercise Path** | ~/workspace/yarn |
| **HDFS Dataset** | lanier/html\_site |

**In this lab you will submit an application to the YARN cluster, and monitor the application using both the Hue Job Browser and the YARN Web UI.**

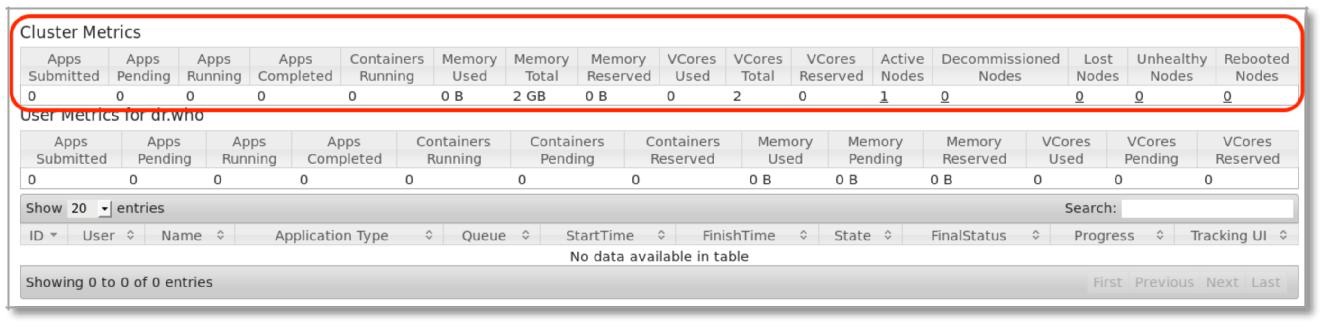
The application you will run is provided for you. The focus of this exercise is on how YARN distributes tasks in a job across a cluster, and how to monitor an application and view its log files.

**Exploring the YARN Cluster**

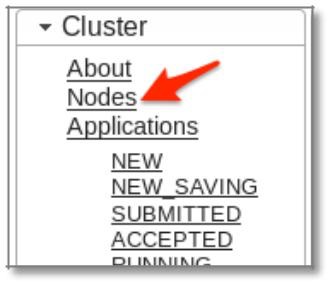
The YARN Resource Manager (RM) UI can be found by visiting UR http://localhost:8088 in Firefox.

NOTE: No jobs are currently running so the current view shows the cluster “at rest.”

Take note of the values in the Cluster Metrics section, which displays information about applications such as the number of applications running currently, previously run or waiting to run; the amount of memory used and available; and how many worker nodes are in the cluster.

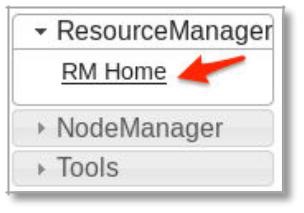


Click on the Nodes link in the Cluster menu on the left. The bottom section will display a list of worker nodes in the cluster. The pseudo-distributed cluster used for this class has only a single node, which is running on the local machine. In the real world, this list would show multiple worker nodes.



Click on the Node HTTP Address to open the Node Manager UI on that specific node. This displays statistics about the selected node, including amount of available memory, currently running applications (none, currently) and so on.

To return to the Resource Manager, expand ResourceManager > RM Home on the left.



**Submitting an Application to the YARN Cluster**

1. In a terminal window, change to the exercise directory

$ cd ~/workspace/yarn

1. Run the example wordcount.py program on the YARN cluster to count the frequency of words in the knowledge dataset:

$ pyspark --master local wordcount.py lanier/ html\_site/

**END**