Cloudera Data Science Workbench

Running Distributed ML Workloads on YARN

Date published: 2020-02-28
Date modified: 2021-06-24

CLOUDERA

https://docs.cloudera.com/

Legal Notice

Contents

Running Distributed ML Workloads on YARN

Cloudera Data Science Workbench 1.6 (and higher) allows you to run distributed machine learning workloads on the CDH/HDP cluster with frameworks such as TensorFlowOnSpark, H2O, XGBoost, and so on. This is similar to what you can already do with Spark workloads that run on the attached CDH/HDP cluster.

To support this, Cloudera Data Science Workbench now forwards three extra ports from the host to each engine. The ports numbers for these ports are stored in the following environmental variables:

- CDSW_HOST_PORT_0
- CDSW_HOST_PORT_1
- CDSW_HOST_PORT_2

The engine's IP address is stored in CDSW_IP_ADDRESS and the host's IP address is stored in CDSW_HOST_IP_ADDRESS.

The information in these environmental variables can be used to make services running in the engine available to services running in the CDH cluster.

Example: H2O

The following shell script shows you how to use this new feature to run a distributed H2O workload.

You can run this script in any active session.

```
#!/bin/bash
wget https://h2o-release.s3.amazonaws.com/h2o/rel-yates/4/h2o-3.24.0.4-cdh6
.0.zip

unzip h2o-3.24.0.4-cdh6.0.zip

hadoop jar h2o-3.24.0.4-cdh6.0/h2odriver.jar \
-nodes 1 \
-mapperXmx 1g \
-extdriverif $CDSW_HOST_IP_ADDRESS \
-driverif $CDSW_IP_ADDRESS \
-driverport $CDSW_HOST_PORT_0 \
-disown

# Clean up
yarn application -kill \
$(yarn application -list 2>/dev/null | grep H2O | awk ' {print $1;}
')
```