

Running Distributed ML Workloads on YARN

Date published: 2020-02-28

Date modified: 2021-06-24



Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Running Distributed ML Workloads on YARN.....	4
Example: H2O.....	4

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;}
')
```