# Cloudera Data Science Workbench Requirements and Supported Platforms

**Date published: 2020-02-28**
**Date modified:**

**https://docs.cloudera.com/**

# Legal Notice

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.

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

# Cloudera Manager and CDH Requirements

Cloudera Data Science Workbench is supported on several versions of CDH and Cloudera Manager.

Cloudera Data Science Workbench is supported on the following versions of CDH and Cloudera Manager:

**Note:** For requirements and supported platforms specific to Hortonworks Data Platform, see Deploying Cloudera Data Science Workbench 1.10.0 on Hortonworks Data Platform .

| Type | CDH | Cloudera Manager |
|---|---|---|
| CSD Deployments | • CDH 6.1.x or higher<br>• Cloudera Runtime 7.0.3 or higher (CDP Private Cloud Base) | • Cloudera Manager 6.1.x or higher<br>• Cloudera Manager 6.2.3 or higher<br>• Cloudera Manager 6.3.x: 6.3.3 or higher |
| RPM Deployments | • CDH 6.1.x or higher | • Cloudera Manager 6.1.x or higher |

All cluster hosts must be managed by Cloudera Manager. Note that all Cloudera Data Science Workbench administrative tasks require root access to the cluster's gateway hosts where Cloudera Data Science Workbench is installed. Therefore, Cloudera Data Science Workbench does not support single-user mode installations.

## Apache Spark Requirements

| CDH Version | Spark 2 Compatibility |
|---|---|
| CDH 6 | On CDH 6 clusters, Apache Spark 2 is packaged with CDH and can no longer be installed separately.<br>To find out which version of Spark 2 ships with your version of CDH 6, refer the CDH 6 Packaging Information guide. |
| CDP Private Cloud Base | On CDP Private Cloud Base clusters, Apache Spark 2 is packaged with Cloudera Runtime and can no longer be installed separately.<br>To find out which version of Spark 2 ships with your version of Cloudera Runtime, refer the Cloudera Runtime Component Versions guide. |

**Note:** Apache Spark 3.x is not supported.

For a list of supported CDSW versions, see Support Lifecycle Policy

# Operating System Requirements

Cloudera Data Science Workbench is supported on the following operating systems. A gateway host that is dedicated to running Cloudera Data Science Workbench must use one of the following supported versions. Cloudera does not support multiple Operating Systems running on the same cluster.

| Operating System | Versions |
|---|---|
| RHEL | 8.6, 8.4, 7.9, 7.8, 7.7, 7.6 |
| CentOS / Oracle Linux RHCK | 8.6, 8.4, 7.9, 7.8, 7.7, 7.6 |
| SLES | 12 SP5 |

**Note:** Before upgrading the host's operating system or installing a patch, you must stop all CDSW services.

**Note:** Support for CDSW running on the listed OS versions is subject to OS vendor support. Please refer to your OS vendor for end of support dates.

### Additional OS-level Settings

- CDSW requires root access.
- Enable memory cgroups on your operating system.
- Disable swap for optimum stability and performance. For instructions, see Setting the vm.swappiness Linux Kernel Parameter.
- Cloudera Data Science Workbench uses uid 8536 and uid 28536 for an internal service account. Make sure that these user IDs is not assigned to any other service or user account.
- Cloudera recommends that all users have the max-user-processes ulimit set to at least 65536.
- Cloudera recommends that all users have the max-open-files ulimit set to 1048576.
- Although CDSW is not supported on Ubuntu, it is possible for CDSW to run on hosts using RHEL/CentOS/Oracle/SLES while the rest of the CDH/HDP/CDP cluster runs on hosts using Ubuntu.
- All CDSW hosts should use the same OS flavor and version.

## JDK Requirements

The entire CDH cluster, including the hosts used for Cloudera Data Science Workbench, should use the same version of JDK.

### Supported CDSW configurations with JDK

- The entire CDH cluster, including the hosts used for Cloudera Data Science Workbench, should use the same version of JDK. If required, the hosts used for CDSW can have two versions of JDK installed, such that the entire CDH cluster can use JDK 11 for all CDH roles, including gateway roles, and the hosts used for CDSW can additionally have JDK 8 installed, with JAVA_HOME set to JDK 8, to meet the requirements of CDSW.
- Oracle JDK 7 is supported across all versions of Cloudera Manager 5 and CDH 5. Oracle JDK 8 is supported in Cloudera Enterprise 5.3.x and higher. Note the JDK 8 Requirement for Spark 2.2 (or higher) on page 5.
- OpenJDK 8 is supported in Cloudera Enterprise 5.16.1 and higher. OpenJDK 7 is not supported.
- For Red Hat/CentOS deployments in particular, Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction must be enabled on the Cloudera Data Science Workbench gateway hosts.

For more specifics on the versions of Oracle JDK and OpenJDK recommended for CDH and Cloudera Manager clusters, and instructions on how to install the Java Cryptography Extension, see the Cloudera Product Compatibility Matrix.

### JDK 8 Requirement for Spark 2.2 (or higher)

CSD-based deployments:

To upgrade your entire CDH cluster to JDK 1.8, see Upgrading to Oracle JDK 1.8.

Package-based deployments:

Set JAVA_HOME to the JDK 8 path in cdsw.conf during the installation process. If you need to modify JAVA _HOME after the fact, restart the master and worker hosts to have the changes go into effect.

# Networking and Security Requirements

It is important to note networking and security requirements for Cloudera Data Science Workbench.

**Note:** Make sure that your networking/security settings on Cloudera Data Science Workbench gateway hosts are not being overwritten behind-the-scenes, either by any automated scripts, or by other high-priority configuration such as /etc/sysctl.conf, /etc/krb5.conf, or /etc/hosts.deny.

- All Cloudera Data Science Workbench gateway hosts must be part of the same datacenter and use the same network. Hosts from different data-centers or networks can result in unreliable performance.
- A wildcard subdomain such as *.cdsw.*company*.com must be configured. Wildcard subdomains are used to provide isolation for user-generated content.

  The wildcard DNS hostname configured for Cloudera Data Science Workbench must be resolvable from both, the CDSW cluster, and your browser.
- Disable all pre-existing iptables rules. While Kubernetes makes extensive use of iptables, it’s difficult to predict how pre-existing iptables rules will interact with the rules inserted by Kubernetes. Therefore, Cloudera recommends you to disable all pre-existing rules before you proceed with the installation.

  It is recommended to save the iptables and check whether the changes have been written to the /etc/sysconfig/iptables file before you disable them. If you disable the iptables without saving, then the settings can get erased upon system reboot.

  1. Save the iptables by running the following command:

     ```
     service iptables save
     ```
  2. Verify whether the changes have been written to the file by running the following command:

     ```
     ls -l /etc/sysconfig/iptables
     ```
  3. Disable the iptables by running the following commands:

     ```
     sudo iptables -P INPUT ACCEPT
     sudo iptables -P FORWARD ACCEPT
     sudo iptables -P OUTPUT ACCEPT
     sudo iptables -t nat -F
     sudo iptables -t mangle -F
     sudo iptables -F
     sudo iptables -X
     ```
- Cloudera Data Science Workbench sets the following sysctl options in /etc/sysctl.d/k8s.conf:

  - net.bridge.bridge-nf-call-iptables=1
  - net.bridge.bridge-nf-call-ip6tables=1
  - net.ipv4.ip_forward=1
  - net.ipv4.conf.default.forwarding=1

  Underlying components of Cloudera Data Science Workbench (Docker, Kubernetes, and NFS) require these options to work correctly. Make sure they are not overridden by high-priority configuration such as /etc/sysctl.conf.
- SELinux must either be disabled or run in permissive mode.
- Multi-homed networks are supported with Cloudera Data Science Workbench 1.2.2 (and higher). However, you will need to explicitly configure the private IP address of the worker nodes in the kubelet start script as follows:

  ```
  # vi /opt/cloudera/parcels/CDSW/scripts/start-kubelet-worker-standalone-
  core.sh
  88 kubelet_opts+=(--v=2)
  ```

```
kubelet_opts+=(--node-ip=172.x.x.x)
```

- Firewall restrictions must be disabled across Cloudera Data Science Workbench and CDH/HDP cluster hosts. For more details on cluster communication, see Ports Required by Cloudera Data Science Workbench.
- Untrusted (non-sudo) SSH access to Cloudera Data Science Workbench hosts must be disabled to ensure a secure deployment.

  Cloudera Data Science Workbench assumes that users only access the gateway hosts through the web application. Untrusted users with SSH access to a Cloudera Data Science Workbench host can gain full access to the cluster, including access to other users' workloads.
- localhost must resolve to 127.0.0.1.
- Forward and reverse DNS lookup must be enabled for the Cloudera Data Science Workbench domain name and IP address (CDSW master host).
- Cloudera Data Science Workbench does not support DNS servers running on 127.0.0.1:53. This IP address resolves to the container localhost within Cloudera Data Science Workbench containers. As a workaround, use either a non-loopback address or a remote DNS server.
- All third-party security software (such as McAfee, Tanium, Symantec, etc.) must be disabled on CDSW hosts. Failure to do so can result in Cloudera Data Science Workbench failing randomly. After CDSW is started, you should be able to re-enable the security software.

Cloudera Data Science Workbench does not support hosts or clusters that do not conform to these restrictions.

## Ports Required by Cloudera Data Science Workbench

Cloudera Data Science Workbench runs on gateway hosts in a CDH/HDP cluster. As such, Cloudera Data Science Workbench acts as a gateway and requires full connectivity to cluster services such as Impala, Spark 2, etc. Additionally, in the case of Spark 2, cluster hosts will require access to the Spark driver running on a set of random ports (20050-32767) on Cloudera Data Science Workbench hosts.

Firewall restrictions must be disabled across Cloudera Data Science Workbench and CDH/HDP cluster hosts. Internally, the Cloudera Data Science Workbench master and worker hosts require full connectivity with no firewalls. Externally, end users connect to Cloudera Data Science Workbench exclusively through a web server running on the master host, and therefore do not need direct access to any other internal Cloudera Data Science Workbench or CDH services.

This information has been summarized in the following table.

| Components | Details |
|---|---|
| Communication with the CDH / HDP cluster | CDH / HDP -> Cloudera Data Science Workbench<br>The CDH/HDP cluster must have access to the Spark driver that runs on Cloudera Data Science Workbench hosts, on a set of randomized ports in the range, 20050-32767. |
| | Cloudera Data Science Workbench -> CDH / HDP<br>As a gateway service, Cloudera Data Science Workbench must have access to all the ports used by CDH and Cloudera Manager. |
| Communication with the Web Browser | The Cloudera Data Science Workbench web application is available at port 80. HTTPS access is available over port 443. |

| Components | Details |
|---|---|
| Communication with the CDH / HDP cluster | CDH / HDP -> Cloudera Data Science Workbench<br>The CDH/HDP cluster must have access to the Spark driver that runs on Cloudera Data Science Workbench hosts, on a set of randomized ports in the range, 20050-32767. |

| Components | Details |
|---|---|
| | Cloudera Data Science Workbench -> CDH / HDP<br>As a gateway service, Cloudera Data Science Workbench must have access to all the ports used by CDH and Cloudera Manager. |
| Communication with the Web Browser | The Cloudera Data Science Workbench web application is available at port 80. HTTPS access is available over port 443. |

### Table 1: Ports used for communication with Unsecure Master

| Port | Process | Mandatory | Note |
|---|---|---|---|
| 22/tcp | sshd | yes | secure shell server (mandatory for CM managed host provisioning) |
| 80/tcp | ingress-controller | yes | CDSW web interface |
| | | | |
| 2049/tcp | nfs | yes | shared filesystem |
| 2379/tcp | etcd-client | yes | k8s shared data store client |
| 2380/tcp | etcd-server | yes | k8s shared data store server |
| 3306/tcp | mysql | | for CM Agent |
| 6443/tcp | kube-apiserver | yes | k8s API endpoint |
| 6783/tcp | weaver | yes | virtual network for docker containers |
| 7191/tcp | CM Agent | yes | for CM Agent |
| 9000/tcp | CM Agent | yes | CM Agent status server |
| 9100/tcp | node_exporter | | Prometheus node monitoring service |
| 10250/tcp | kubelet | yes | k8s the primary "node agent" |
| 10256/tcp | kube-proxy | yes | network proxy that implements part of the k8s Service concept |
| 20048/tcp | rpc.mountd | yes | server side of the NFS MOUNT protocol |

### Table 2: Ports used for communication with Secure Master

| Port | Process | Mandatory | Note |
|---|---|---|---|
| 22/tcp | sshd | yes | secure shell server (mandatory for CM managed host provisioning) |
| 80/tcp | ingress-controller | | CDSW web interface |
| 111/tcp | NFS portmapper | yes | |
| 443/tcp | secure ingress-controller | yes | CDSW web interface |
| 2049/tcp | nfs | yes | shared filesystem |
| 2379/tcp | etcd-client | yes | k8s shared data store client |
| 2380/tcp | etcd-server | yes | k8s shared data store server |
| 3306/tcp | mysql | | for CM Agent |
| 6443/tcp | kube-apiserver | yes | k8s API endpoint |
| 6783/tcp | weaver | yes | virtual network for docker containers |
| 7191/tcp | CM Agent | yes | for CM Agent |

| Port | Process | Mandatory | Note |
|---|---|---|---|
| 9000/tcp | CM Agent | yes | CM Agent status server |
| 9100/tcp | node_exporter | | Prometheus node monitoring service |
| 10053/tcp | Kub-DNS | yes | |
| 10250/tcp | kubelet | yes | k8s the primary "node agent" |
| 10251/tcp | kube-scheduler | yes | |
| 10252/tcp | kube-control-manager | yes | |
| 10256/tcp | kube-proxy | yes | network proxy that implements part of the k8s Service concept |
| 20048/tcp | rpc.mountd | yes | server side of the NFS MOUNT protocol |

### Table 3: Ports used for communication with Secure/Unsecure Worker

| Port | Process | Mandatory | Note |
|---|---|---|---|
| 22/tcp | sshd | yes | secure shell server (mandatory for CM managed host provisioning) |
| 3306/tcp | mysql | | for CM Agent |
| 6783/tcp | weaver | yes | virtual network for docker containers |
| 7191/tcp | CM Agent | yes | for CM Agent |
| 9000/tcp | CM Agent | yes | CM Agent status server |
| 9100/tcp | node_exporter | | Prometheus node monitoring service |
| 10250/tcp | kubelet | yes | k8s the primary "node agent" |
| 10256/tcp | kube-proxy | yes | network proxy that implements part of the k8s Service concept |

# GPU Support

Known issues with GPU support.

## Only CUDA-enabled NVIDIA GPU hardware is supported

Cloudera Data Science Workbench only supports CUDA-enabled NVIDIA GPU cards.

## Heterogeneous GPU hardware is not supported

CDSW nodes that have GPUs must all use the same GPU make and model.

## Known Issues with Tensorflow 2.4 and CUDA 11.2

During GPU set up, the dynamic library libculsolver.so.10 is not read. For more information, see https://github.com/tensorflow/tensorflow/issues/44777.

Workaround: Enter the following commands when starting a session:

```
!ln -s /usr/local/cuda-11.1/targets/x86_64-linux/lib/libcusolver.so.11.0.1.1
05 /usr/local/cuda-11.1/targets/x86_64-linux/lib/libcusolver.so.10
```

```
!ln -s /usr/lib/x86_64-linux-gnu/libcuda.so.460.73.01 /usr/lib/x86_64-linux-
gnu/libcuda.so.1
```

### Multi-Instance GPU (MIG) Support

The NVIDIA Multi-Instance GPU (MIG) feature is not supported.

# Recommended Hardware Configuration

Cloudera Data Science Workbench hosts are added to your CDP cluster as gateway hosts.

**Note:** For proof-of-concept deployments, you can deploy a 1-host cluster with just a Master host. The Master host can run user workloads just as a worker host can when required for demonstration purposes. For production deployments, you must have a reserved, dedicated master host and separate worker host(s).

The recommended minimum hardware configuration for Cloudera Data Science Workbench gateway hosts is:

**Note:**

- Allocate separate CDP gateway hosts for Cloudera Data Science Workbench. Do not reuse existing hosts that are already running other CDP services. Doing this can lead to port conflicts, unreliable execution of user workloads, and out-of-memory errors.
- Starting with version 1.4.3, multi-host CDSW deployments can be customized to reserve the Master only for internal processes while user workloads are run exclusively on workers. For details, see Reserving the Master Host for Internal CDSW Components.

| Resource Type | Master | Workers | Notes |
|---|---|---|---|
| Supported architecture | | | Intel 64 bit, AMD 64 |
| CPU | Minimum 16+ CPU (vCPU) cores | Minimum 16+ CPU (vCPU) cores | |
| RAM | Minimum 32+ GB | Minimum 32+ GB | See the following Scaling Guidelines for more information. |
| Disk Space | | | |
| Root Volume | 200+ GB | 200+ GB | If you are going to partition the root volume, make sure you allocate at least 200 GB to /var/lib/cdsw/docker-tmp so that the installer can proceed without running out of space. |
| Application Block Device | 1 TB | (Not required) | The Application Block Device is only required on the Master where it is mounted to /var/lib/cdsw.<br>You will be asked to create a /var/lib/cdsw directory on all the Worker hosts during the installation process. However, they do not need to be mounted to a block device. It is only used to store client configuration for HDP cluster services on Workers. |
| Docker Block Device | 1 TB | 1 TB | A raw device is expected on all Master and Worker hosts. Docker will create the necessary volumes on these hosts. |
| Parcel directory | 50 GB | 50 GB | If using Cloudera Manager parcels to distribute CDSW, all hosts inside Cloudera Manager must have at least 50 GB of free space inside the parcel directory, which defaults to /opt/cloudera/parcels/. |

Scaling Guidelines

New hosts can be added and removed from a Cloudera Data Science Workbench deployment without interrupting any jobs already scheduled on existing hosts. Therefore, it is rather straightforward to increase capacity based on observed usage. At a minimum, Cloudera recommends you allocate at least 1 CPU core and 2 GB of RAM per concurrent session or job. CPU can burst above a 1 CPU core share when spare resources are available. Therefore, a 1 CPU core allocation is often adequate for light workloads. Allocating less than 2 GB of RAM can lead to out-of-memory errors for many applications.

As a general guideline, Cloudera recommends hosts with RAM between 60GB and 256GB, and between 16 and 48 cores. This provides a useful range of options for end users. Note that SSDs are strongly recommended for application data storage. Using standard HDDs can sometimes result in poor application performance.

For some data science and machine learning applications, users can collect a significant amount of data in memory within a single R or Python process, or use a significant amount of CPU resources that cannot be easily distributed into the CDH cluster. If individual users frequently run larger workloads or run workloads in parallel over long durations, increase the total resources accordingly. Understanding your users' concurrent workload requirements or observing actual usage is the best approach to scaling Cloudera Data Science Workbench.

# Docker and Kubernetes Support

Cloudera Data Science Workbench only supports the versions of Docker and Kubernetes that are shipped with each release.

Cloudera Data Science Workbench only supports the versions of Docker and Kubernetes that are shipped with each release. Upgrading Docker or Kubernetes, or running on third-party Kubernetes clusters is not supported.

# Supported Browsers

It is important to note the browsers supported by Cloudera Data Science Workbench.

- Chrome (latest stable version)
- Firefox (latest released version and latest ESR version)
- Safari 9+
- Microsoft Edge (latest, Chromium-based)

# Cloudera Altus Director Support (AWS and Azure Only)

Altus Director support for Cloudera Data Science Workbench is available for a variety of platforms.

Altus Director support for Cloudera Data Science Workbench is available for the following platforms:

- Amazon Web Services (AWS) - Cloudera Altus Director 2.6.0 (and higher)

  Microsoft Azure - Cloudera Altus Director 2.7 (and higher)
- Cloudera Manager 5.13.1 (and higher)
- CSD-based Cloudera Data Science Workbench 1.2.x (and higher)

## Deploying Cloudera Data Science Workbench with Altus Director

Points to note when using Altus Director to install Cloudera Data Science Workbench:

- (Required for Director 2.6) Before you run the command to bootstrap a new cluster, set the lp.normalization.mountAllUnmountedDisksRequired property to false in the Altus Director server's application.properties file, and then restart Altus Director.

  Higher versions of Altus Director do not require this step. Altus Director 2.7 (and higher) include an instance-level setting called mountAllUnmountedDisks that must be set to false as demonstrated in the following sample configuration files.

- Depending on your cloud platform, you can use one of the following sample configuration files to deploy a Cloudera Manager cluster with Cloudera Data Science Workbench.
    - AWS - aws.cdsw.conf
    - Azure - azure.cdsw.conf

  Note that these sample files are tailored to Altus Director 2.7 (and higher) and they install a very limited CDH cluster with just the following services: HDFS, YARN, and Spark 2. You can extend them as needed to match your use case.

Related Topics:

- Using Products outside CDH with Altus Director
- Altus Director CLI
- The Altus Director Configuration File
- Setting Altus Director Properties

# Recommended Configuration on Amazon Web Services (AWS)

On AWS, Cloudera Data Science Workbench must be used with persistent/long-running Apache Hadoop clusters only.

CDH and Cloudera Manager Hosts

- For instructions on deploying CDH and Cloudera Manager on AWS, refer the Cloudera Reference Architecture for AWS deployments.

Cloudera Data Science Workbench Hosts

- Operations
    - Use Cloudera Director to orchestrate operations. Use Cloudera Manager to monitor the cluster.
- Networking
    - No security group or network restrictions between hosts.
    - HTTP connectivity to the corporate network for browser access. Do not use proxies or manual SSH tunnels.
- Recommended Instance Types
    - m4.4xlarge–m4.16xlarge

      In this case, bigger is better. That is, one m4.16large is better than four m4.4xlarge hosts. AWS pricing scales linearly, and larger instances have more EBS bandwidth.
- Storage
    - 100 GB root volume block device (gp2) on all hosts
    - 500 GB Docker block devices (gp2) on all hosts
    - 1 TB Application block device (io1) on master host

# Recommended Configuration on Microsoft Azure

When setting up Cloudera Data Science Workbench on Microsoft Azure, you should consider the recommended configuration.

CDH and Cloudera Manager Hosts

- For instructions on deploying CDH and Cloudera Manager on Azure, refer the Cloudera Reference Architecture for Azure deployments.

Cloudera Data Science Workbench Hosts

- Operations
  - Use Cloudera Director to orchestrate operations. Use Cloudera Manager to monitor the cluster.
- Networking
  - No security group or network restrictions between hosts.
  - HTTP connectivity to the corporate network for browser access. Do not use proxies or manual SSH tunnels.
- Recommended Instance Types
  - DS13-DS14 v2 instances on all hosts.
- Storage
  - P30 premium storage for the Application and Docker block devices.

    Cloudera Data Science Workbench requires premium disks for its block devices on Azure. Standard disks can lead to unacceptable performance even on small clusters.