Cloudera Data Science Workbench Architecture Overview

Date published: 2020-02-28
Date modified: 2020-12-15

https://docs.cloudera.com/

Legal Notice

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

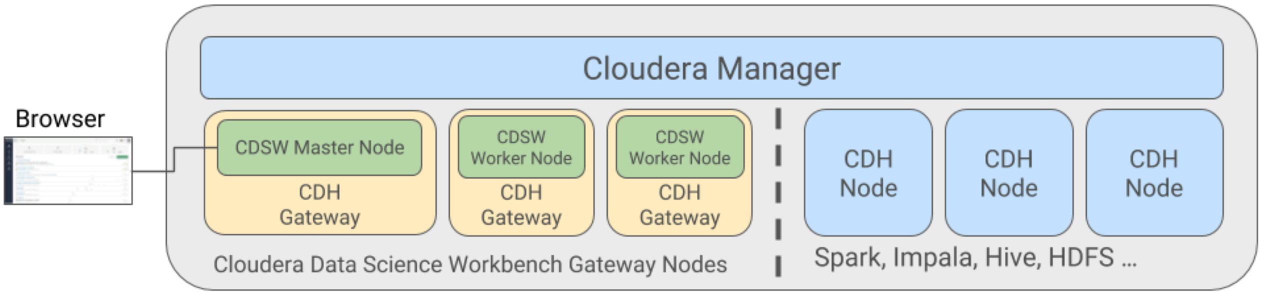
Architecture Overview

Important: The rest of this documentation assumes you are familiar with CDH and Cloudera Manager. If not, make sure you read the documentation for CDH and Cloudera Manager before you proceed.

Cloudera Manager

Cloudera Manager is an end-to-end application used for managing CDH clusters. When a CDH service (such as Impala, Spark, etc.) is added to the cluster, Cloudera Manager configures cluster hosts with one or more functions, called roles.

In a Cloudera Manager cluster, a gateway role is one that designates that a host should receive client configuration for a CDH service even though the host does not have any role instances for that service running on it. Gateway roles provide the configuration required for clients that want to access the CDH cluster. Hosts that are designated with gateway roles for CDH services are referred to as gateway hosts.



Cloudera Data Science Workbench runs on one or more dedicated gateway hosts on CDH clusters. Each of these hosts has the Cloudera Manager Agent installed on them. The Cloudera Management Agent ensures that Cloudera Data Science Workbench has the libraries and configuration necessary to securely access the CDH cluster.

Cloudera Data Science Workbench does not support running any other services on these gateway hosts. Each gateway host must be dedicated solely to Cloudera Data Science Workbench. This is because user workloads require dedicated CPU and memory, which might conflict with other services running on these hosts. Any workloads that you run on Cloudera Data Science Workbench hosts will have immediate secure access to the CDH cluster.

From the assigned gateway hosts, one will serve as the master host while others will serve as worker hosts.

Master Host

The master host keeps track of all critical persistent and stateful data within Cloudera Data Science Workbench. This data is stored at /var/lib/cdsw.

- **Project Files**

  Cloudera Data Science Workbench uses an NFS server to store project files. Project files can include user code, any libraries you install, and small data files. The master host provides a persistent filesystem which is exported to worker hosts using NFS. This filesystem allows users to install packages interactively and have their dependencies and code available on all Cloudera Data Science Workbench nodes without any need for synchronization. The files for all the projects are stored on the master host at /var/lib/cdsw/current/projects. When a job or session is launched, the project’s filesystem is mounted into an isolated Docker container at /home/cdsw.
- **Relational Database**

  The Cloudera Data Science Workbench uses a PostgreSQL database that runs within a container on the master host at /var/lib/cdsw/current/postgres-data.

- **Livelog**

  Cloudera Data Science Workbench allows users to work interactively with R, Python, and Scala from their browser and display results in realtime. This realtime state is stored in an internal database called Livelog, which stores data on the master host at /var/lib/cdsw/current/livelog. Users do not need to be connected to the server for results to be tracked or jobs to run.

Worker Hosts

While the master host stores the stateful components of the Cloudera Data Science Workbench, the worker hosts are transient. These can be added or removed as needed, which gives you flexibility with scaling the deployment. As the number of users and workloads increases, you can add more worker hosts to Cloudera Data Science Workbench over time.

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).

Even on multi-host deployments, the Master host doubles up to perform both functions: those of the Master, and those of a worker. Starting with version 1.4.3, multi-host 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.

Cloudera Data Science Workbench Engines

Cloudera Data Science Workbench engines are responsible for running R, Python, and Scala code written by users and intermediating access to the CDH cluster.

You can think of an engine as a virtual machine, customized to have all the necessary dependencies to access the CDH cluster while keeping each project’s environment entirely isolated. To ensure that every engine has access to the parcels and client configuration managed by the Cloudera Manager Agent, a number of folders are mounted from the host into the container environment. This includes the parcel path -/opt/cloudera, client configuration, as well as the host’s JAVA_HOME. For more details on basic concepts and terminology related to engines in Cloudera Data Science Workbench, see Cloudera Data Science Workbench Engines.

Known Issues

- Apache Phoenix requires additional configuration to run commands successfully from within Cloudera Data Science Workbench engines (sessions, jobs, experiments, models).

  Workaround

  Explicitly set HBASE_CONF_PATH to a valid path before running Phoenix commands from engines.

  ```
  export HBASE_CONF_PATH=/usr/hdp/hbase/<hdp_version>/0/
  ```

Docker and Kubernetes

Cloudera Data Science Workbench uses Docker containers to deliver application components and run isolated user workloads. On a per project basis, users can run R, Python, and Scala workloads with different versions of libraries and system packages. CPU and memory are also isolated, ensuring reliable, scalable execution in a multi-tenant setting.

Each Docker container running user workloads, also referred to as an engine, provides a visualized gateway with secure access to CDH cluster services such as HDFS, Spark 2, Hive, and Impala. CDH dependencies and client configuration, managed by Cloudera Manager, are mounted from the underlying gateway host. Workloads that leverage CDH services such as HDFS, Spark, Hive, and Impala are executed across the full CDH cluster.

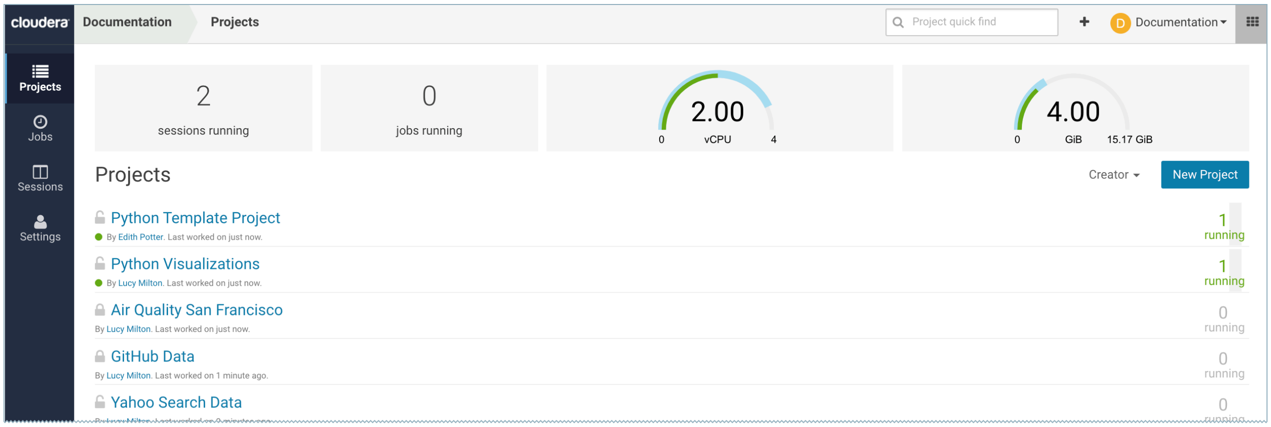
To enable multiple users and concurrent access, Cloudera Data Science Workbench transparently subdivides and schedules containers across multiple hosts dedicated as gateway hosts. This scheduling is done using Kubernetes, a container orchestration system used internally by Cloudera Data Science Workbench. Neither Docker nor Kubernetes are directly exposed to end users, with users interacting with Cloudera Data Science Workbench through a web application.

Cloudera Data Science Workbench Web Application

The Cloudera Data Science Workbench web application is typically hosted on the master host, at http://cdsw.<*your_domain*>.com.

The web application provides a rich GUI that allows you to create projects, collaborate with your team, run data science workloads, and easily share the results with your team. For a quick demonstration, either watch this video or read the Quickstart Guide.

You can log in to the web application either as a site administrator or a regular user. See the Administration and User Guides respectively for more details on what you can accomplish using the web application.



Python Template Project
By Edith Potter. Last worked on just now.
1 running

Python Visualizations
By Lucy Milton. Last worked on just now.
1 running

Air Quality San Francisco
By Lucy Milton. Last worked on just now.
0 running

GitHub Data
By Lucy Milton. Last worked on 1 minute ago.
0 running

Yahoo Search Data
0

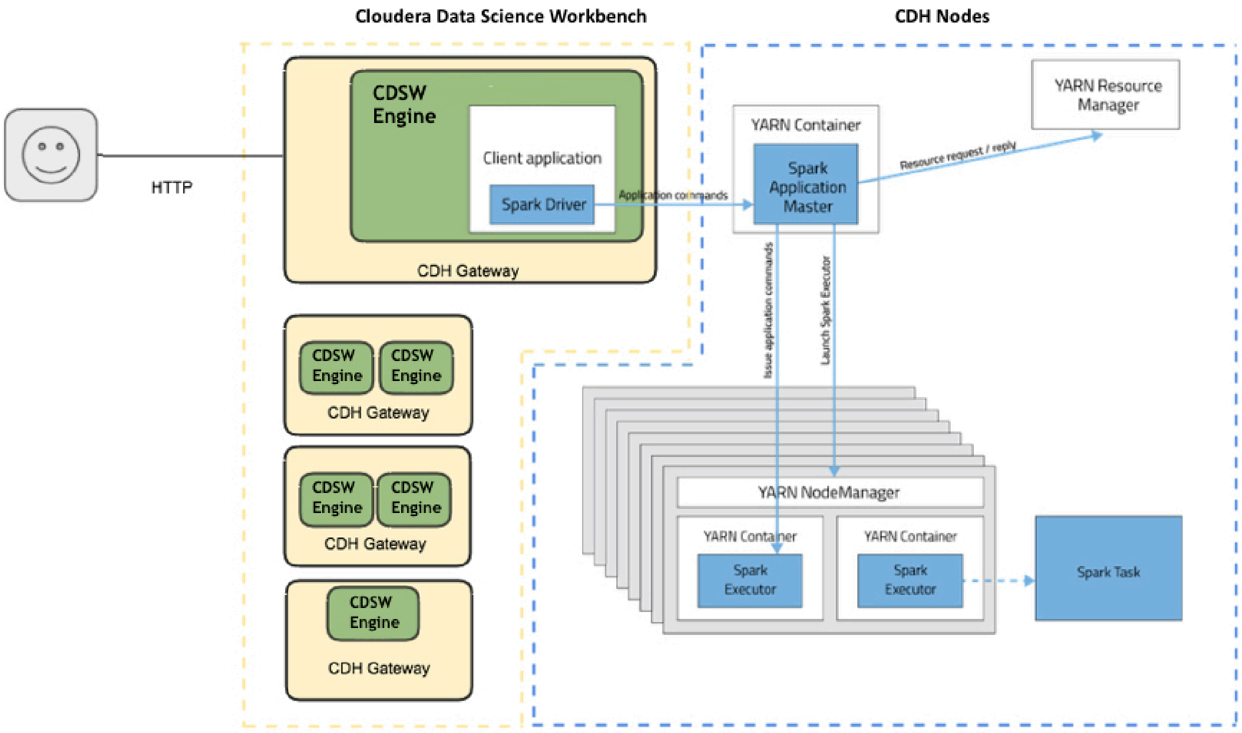
CDS 2.x Powered by Apache Spark

Apache Spark is a general purpose framework for distributed computing that offers high performance for both batch and stream processing. It exposes APIs for Java, Python, R, and Scala, as well as an interactive shell for you to run jobs.

Important: The rest of this topic assumes you are familiar with Apache Spark and CDS 2.x Powered by Apache Spark. If not, make sure you read the CDS 2.x documentation before you proceed.

Cloudera Data Science Workbench provides interactive and batch access to Spark 2. Connections are fully secure without additional configuration, with each user accessing Spark using their Kerberos principal. With a few extra lines of code, you can do anything in Cloudera Data Science Workbench that you might do in the Spark shell, as well as leverage all the benefits of the workbench. Your Spark applications will run in an isolated project workspace.

Cloudera Data Science Workbench's interactive mode allows you to launch a Spark application and work iteratively in R, Python, or Scala, rather than the standard workflow of launching an application and waiting for it to complete to view the results. Because of its interactive nature, Cloudera Data Science Workbench works with Spark on YARN's client mode, where the driver persists through the lifetime of the job and runs executors with full access to the CDH cluster resources. This architecture is illustrated the following figure:



More resources:

- Documentation for CDS 2.x Powered by Apache Spark
- Apache Spark 2 upstream documentation