Cloudera Data Science Workbench

Glossary

Date published: 2020-02-28
Date modified: 2021-02-25

https://docs.cloudera.com/

Legal Notice

Contents

Cloudera Data Science Workbench Glossary

Terms related to Cloudera Data Science Workbench.

Cloudera Data Science Workbench

Cloudera Data Science Workbench is a product that enables fast, easy, and secure self-service data science for the enterprise. It allows data scientists to bring their existing skills and tools, such as R, Python, and Scala, to securely run computations on data in Hadoop clusters.

site administrator

A Cloudera Data Science Workbench user with all-access permissions. Site administrators can add or disable users/teams, monitor and manage resource usage, secure access to the deployment, and more. The Site Administration dashboard is only accessible to site administrators.

API

Cloudera Data Science Workbench exposes a limited REST API that allows you to schedule existing jobs from third-party workflow tools.

cluster

Refers to the CDH cluster managed by Cloudera Manager, including the gateway hosts that are running Cloudera Data Science Workbench.

context

Cloudera Data Science Workbench uses the notion of contexts to separate your personal account from any team accounts you belong to. This gives you leave to run experiments in your own personal context, while you can simultaneously collaborate with others in your organization within a team context.

engine

In Cloudera Data Science Workbench, engines are responsible for running R, Python, and Scala code written by users and for facilitating access to the CDH cluster. Each engine functions as an isolated virtual machine, customized to have all the necessary dependencies to access the CDH cluster while keeping each project's environment entirely isolated. The only artifacts that remain after an engine runs is a log of the analysis and any files that were generated or modified inside the project's filesystem, which is mounted to each engine at /home/cdsw.

experiment

Experiments are batch executed workloads that help facilitate model training in Cloudera Data Science Workbench.

gateway host

On a Cloudera Manager cluster, a gateway host is one that has been assigned a gateway role for a CDH service. Such a host will receive client configuration for that CDH service even though the host does not have any role instances for that service running on it.

Cloudera Data Science Workbench runs on dedicated gateway hosts on a CDH cluster. These hosts are assigned gateway roles for the Spark and HDFS services so that Cloudera Data Science Workbench has the client configuration required to access the CDH cluster.

job

Jobs are sessions that can be scheduled in advance and do not need to be launched manually each time.

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.

master

A typical Cloudera Data Science Workbench deployment consists of 1 master host and zero or more worker hosts. The master host keeps track of all critical, persistent, and stateful application data within Cloudera Data Science Workbench.

model

Model is a high level abstract term that is used to describe several possible incarnations of objects created during the model deployment process in Cloudera Data Science Workbench. You should note that 'model' does not always refer to a specific artifact. More precise terms (as defined in the documentation) should be used whenever possible.

pipeline

A series of jobs that depend on each other and must therefore be executed in a specific pre-defined sequence.

project

Projects hold the code, configuration, and libraries needed to reproducibly run data analytics workloads. Each project is independent, ensuring users can work freely without interfering with one another or breaking existing workloads.

runtimes

As an alternative to the existing Engines, ML Runtimes are more lightweight than the current monolithic Engines. By specifying the desired Editor, Kernel, Edition, and Version, a streamlined Runtime will be used to run the user’s code in Sessions, Jobs, Experiments, Models, or Applications.

session

A session is an interactive environment where you can run exploratory analysis in R, Python, and Scala.

team

A group of trusted users who are collaborating on a project in Cloudera Data Science Workbench.

terminal

Cloudera Data Science Workbench allows terminal access to actively running engines. The terminal can be used to move project files around, run Git commands, access the YARN and Hadoop CLIs, or install libraries that cannot be installed directly from the engine.

web application

Refers to the Cloudera Data Science Workbench web application running at cdsw.<*your_domain*>.com.

workbench

The console in the web application that is used to launch interactive sessions and run exploratory data analytic workloads. It consists of two panes, a navigable filesystem and editor on the left, and an interactive command prompt on the right.

worker

Worker hosts are transient hosts that can be added or removed from a Cloudera Data Science Workbench deployment depending on the number of users and workloads you are running.