

Getting Started with Cloudera Data Science Workbench

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

Getting Started with Cloudera Data Science Workbench.....	4
Sign up.....	4
Create a Project from a Built-in Template.....	4
Launch a Session to Run the Project.....	6
Export Session List.....	9
Next Steps.....	9

Getting Started with Cloudera Data Science Workbench

This topic provides a suggested method for quickly getting started with data science projects on Cloudera Data Science Workbench.

Watch the following video for a quick demo of the steps described in this topic: [CDSW Quickstart Demo](#)

Figure 1: Cloudera Data Science Workbench Quickstart Demo

Sign up

To sign up, open the Cloudera Data Science Workbench web application in a browser.

The application is typically hosted on the master host at `http://cdsw.<your_domain>.com`. The first time you log in, you will be prompted to create a username and password. Note that the first account created will receive site administrator privileges.

If your site administrator has configured your deployment to require invitations, you will need an invitation link to sign up.

Create a Project from a Built-in Template

Cloudera Data Science Workbench is organized around projects. Projects hold all the code, configuration, and libraries needed to reproducibly run analyses.

About this task

To help you get started, Cloudera Data Science Workbench includes sample template projects in R, Python, PySpark, and Scala. Using a template project gives you the impetus to start using the Cloudera Data Science Workbench right away.

Create a Template Project

Create a New Project

Account

Documentation

Project Name

Python 2 Template Test

Project Visibility

☒ **Private** - Only added collaborators can view the project.

☐ **Team** - All members of your team can view this project.

☐ **Public** - All authenticated users can view this project.

Initial Setup

Blank Template Local Git

Python

Templates include example code to help you get started.

Create Project

To create a template project:

Procedure

1. Sign in to Cloudera Data Science Workbench.
2. Click New Project.
3. Enter the account and project name.
4. Under the Template tab, you can choose one of the programming languages to create a project from one of the built-in templates. Alternatively, if your site administrator has added any custom template projects, those will also be available in this dropdown list.
5. Click Create Project.

What to do next

After creating your project, you see your project files and the list of jobs defined in your project. These project files are stored on an internal NFS server, and are available to all your project sessions and jobs, regardless of the gateway hosts they run on. Any changes you make to the code or libraries you install into your project will be immediately available when running an engine.

Launch a Session to Run the Project

Cloudera Data Science Workbench provides an interactive environment tailored for data science called the workbench. It supports R, Python, and Scala engines, one of which we will use to run the template project.

About this task

Workbench

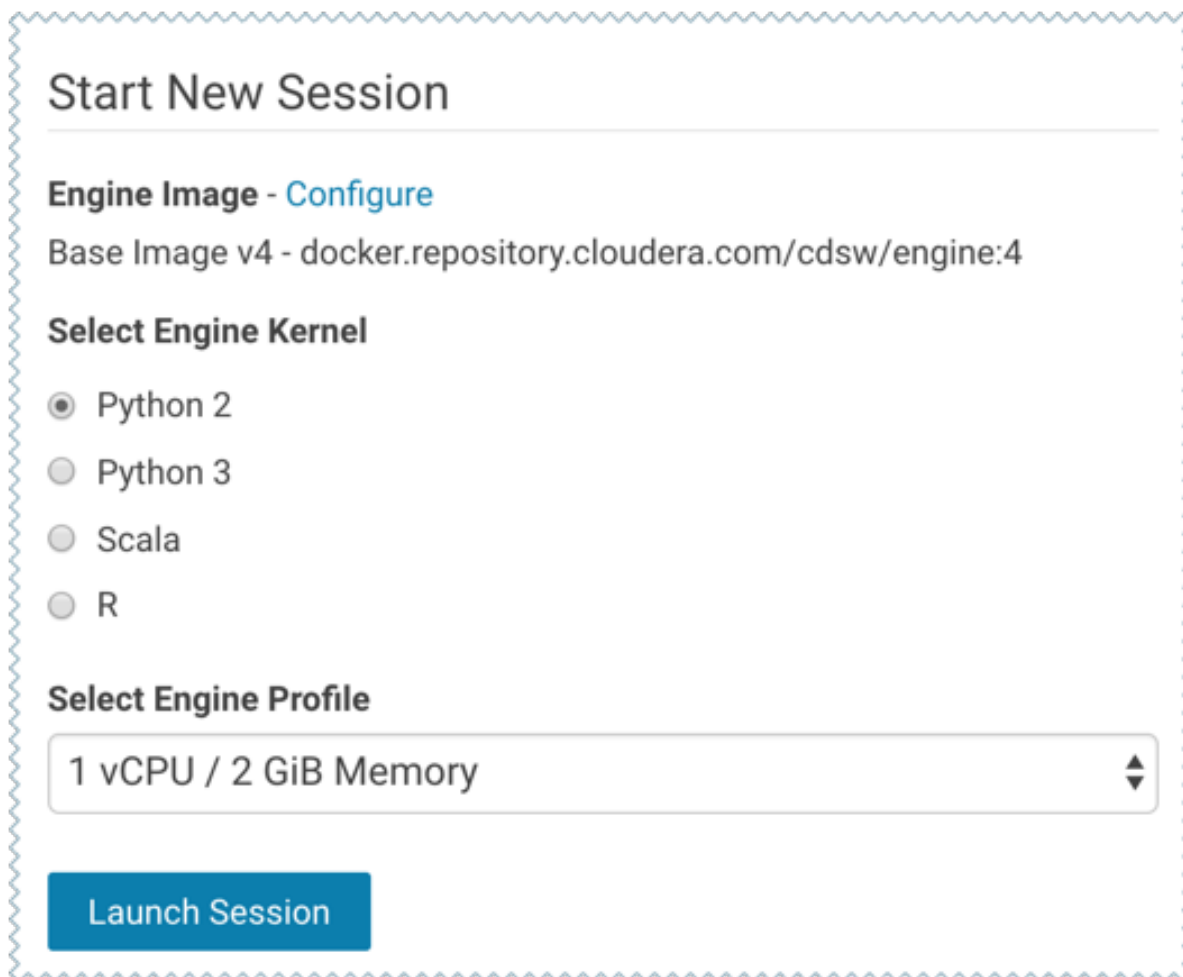
The screenshot displays the Cloudera Data Science Workbench interface. On the left, the 'Project file system' pane shows a file tree with 'analysis.py', 'README.md', and 'seaborn-data'. The main editor displays the contents of 'analysis.py', which includes code for loading the 'tips' dataset, creating a joint plot, and displaying it. On the right, the 'Python Template Session' pane shows the session status as 'Running'. Below the session header, the 'Setup' section lists the installed packages: 'pandas' and 'seaborn'. The 'Basic Data Manipulation' section contains the code executed in the session, which matches the code in the editor. Below the code, a scatter plot titled 'Tips Regression' is displayed, showing a positive correlation between 'total_bill' and 'tip'. The plot includes a regression line and a shaded confidence interval. The Pearson correlation coefficient is shown as 'pearsonr = 0.68; p = 6.7e-34'. An arrow points from the 'Terminal access to running engine' label to the session header. Another arrow points from the 'Interactive command prompt' label to the bottom of the session pane.

Perform the following steps to run the project:

To run the project code, open the workbench and launch a new session.

Procedure

1. Open the Workbench to Launch a Session.
 - a) Navigate to the new project's Overview page.
 - b) Click Open Workbench.
 - c) Launch a New Session



Start New Session

Engine Image - Configure

Base Image v4 - docker.repository.cloudera.com/cdsw/engine:4

Select Engine Kernel

- ☒ Python 2
- ☐ Python 3
- ☐ Scala
- ☐ R

Select Engine Profile

1 vCPU / 2 GiB Memory

Launch Session

1. Use Select Engine Kernel to choose the programming language that your project uses.
2. Use Select Engine Profile to select the number of [CPU cores](#) and memory to be used.
3. Click Launch Session.

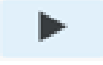
The command prompt at the bottom right of your browser window will turn green when the engine is ready. Sessions typically take between 10 and 20 seconds to start.

2. Run project code.

You can enter and run code using either the editor or the command prompt. The editor is best used for code you want to keep, while the command prompt is best for quick interactive exploration.

Editor - To run code in the editor:

- a) Select a script from the project files on the left sidebar.
- b)

To run the whole script click  on the top navigation bar, or, highlight the code you want to run and press Ctrl+Enter (Windows/Linux) or cmd+Enter (macOS).

Command Prompt - The command prompt functions largely like any other. Enter a command and press Enter to run it. If you want to enter more than one line of code, use Shift+Enter to move to the next line. The output of your code, including plots, appears in the console.



```
> ls
analysis.py  README.md  seaborn-data/

> !pip install beautifulsoup4
```

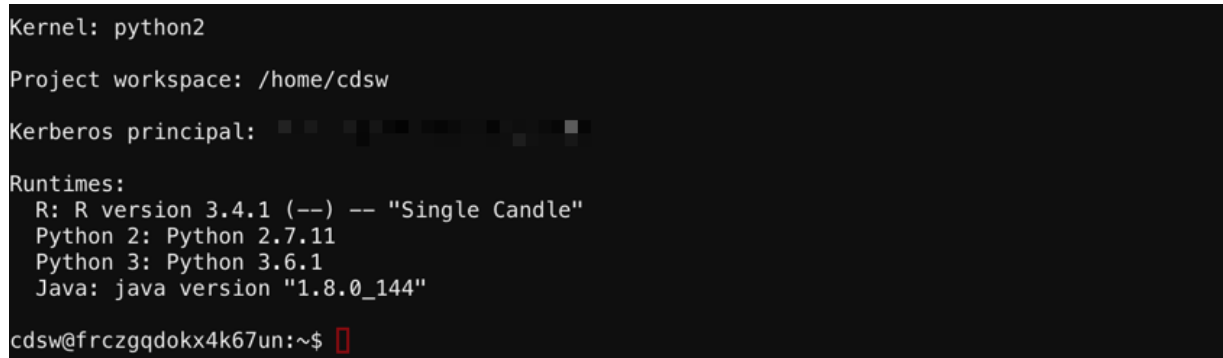
Enter a command and press Enter

Code Autocomplete - The Python and R kernels include support for automatic code completion, both in the editor and the command prompt. Use single tab to display suggestions and double tab for autocomplete.

3. Test terminal access.

Cloudera Data Science Workbench provides terminal access to the running engines from the web console. You can use the terminal to move files around, run Git commands, and understand what resources are already available to you in the project environment.

To access the Terminal from a running session, click Terminal Access above the console pane. The terminal's default working directory is /home/cdsw, which is a temporary directory where all your project files are stored for this session.



```
Kernel: python2
Project workspace: /home/cdsw
Kerberos principal: ■ ■ ■ ■ ■ ■ ■ ■ ■ ■
Runtimes:
R: R version 3.4.1 (--) -- "Single Candle"
Python 2: Python 2.7.11
Python 3: Python 3.6.1
Java: java version "1.8.0_144"
cdsw@frczgqdkx4k67un:~$
```



Note: The terminal does not provide root or sudo access to the container.

4. Logs.

The logs tab displays the engine logs for the running session and, if applicable, the Spark logs.

5. Stop the session.

6. When you are done with the session, click Stop in the menu bar above the console.

Export Session List

You can export a list of sessions. You can either download a complete list of sessions or you can filter the sessions by date to download a more concise list.

Procedure

1. Click Admin in the navigation bar to display the Site Administration page.
2. Click the Usage tab.
3. If you want a list of sessions specific to a date range, you can filter the list of sessions by setting the Date Range.
4. Click Export Session List to download your list of sessions.

Next Steps

As a site administrator, you can invite new users, monitor resource utilization, secure the deployment, and upload a license key for the product.

Depending on the size of your deployment, you might also want to customize how Cloudera Data Science Workbench schedules workloads on your gateway hosts. For more details on these tasks, see:

- [Site Administration](#)
- [Customize Workload Scheduling](#)
- [Security](#)

You can also start using the product by configuring your personal account and creating a new project. For a quickstart that walks you through creating and running a simple template project, see [Getting Started with Cloudera Data Science Workbench](#). For more details on collaborating with teams, working on projects, and sharing results, see the [Managing Cloudera Data Science Workbench Users](#).