

Cloudera Data Science Workbench Overview

Date published: 2020-02-28

Date modified: 2020-12-15

CLOUDERA

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

Cloudera Machine Learning Overview.....	4
Key Differences between Cloudera Machine Learning and Cloudera Data Science Workbench.....	5

Cloudera Machine Learning Overview

Machine learning has become one of the most critical capabilities for modern businesses to grow and stay competitive today. From automating internal processes to optimizing the design, creation, and marketing processes behind virtually every product consumed, ML models have permeated almost every aspect of our work and personal lives.

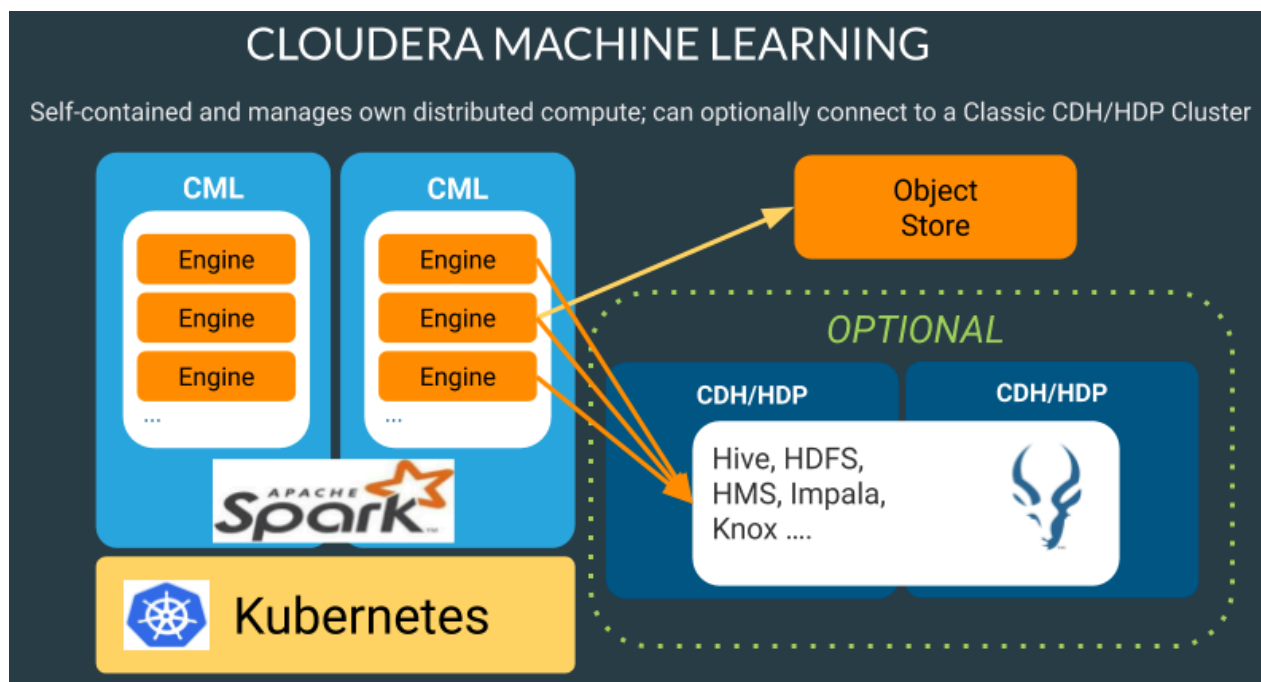
Cloudera Machine Learning (CML) is Cloudera's new cloud-native machine learning service, built for CDP. The CML service provisions clusters, also known as ML workspaces, that run natively on Kubernetes.

Each ML workspace enable teams of data scientists to develop, test, train, and ultimately deploy machine learning models for building predictive applications all on the data under management within the enterprise data cloud.

ML workspaces are ephemeral, allowing you to create and delete them on-demand. ML workspaces support fully-containerized execution of Python, R, Scala, and Spark workloads through flexible and extensible engines.

Core Capabilities:

- Seamless portability across private cloud, public cloud, and hybrid cloud powered by Kubernetes
- Rapid cloud provisioning and autoscaling
- Fully containerized workloads - including Python, R, and Spark-on-Kubernetes - for scale-out data engineering and machine learning with seamless distributed dependency management
- High performance deep learning with distributed GPU scheduling and training
- Secure data access across HDFS, cloud object stores, and external databases



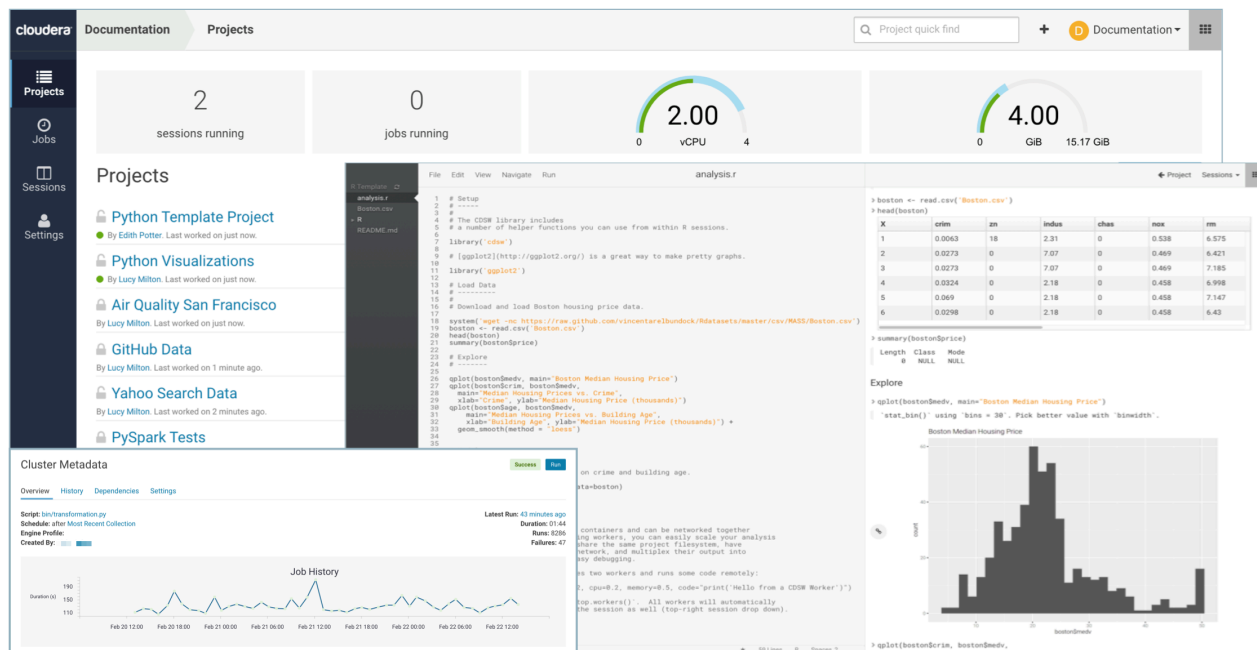
Benefits:

The cloud offers many advantages for unpredictable and heterogeneous workloads, but there are two challenges: 1) data is often spread across multiple clouds and on-premises systems, and 2) existing products only cover parts of the machine learning lifecycle. Cloudera Machine Learning directly addresses both these issues. It's built for the agility and power of cloud computing, but isn't limited to any one provider or data source. And it is a comprehensive platform to collaboratively build and deploy machine learning capabilities at scale. CML gives you the power to transform your business with machine learning and AI.

CML users are:

- Data management and data science executives at large enterprises who want to empower teams to develop and deploy machine learning at scale.
- Data scientist developers (use open source languages like Python, R, Scala) who want fast access to compute and corporate data, the ability to work collaboratively and share, and an agile path to production model deployment.
- IT architects and administrators who need a scalable platform to enable data scientists in the face of shifting cloud strategies while maintaining security, governance and compliance. They can easily provision environments and enable resource scaling so they - and the teams they support - can spend less time on infrastructure and more time on innovation.

ML Workspace Web Application



Key Differences between Cloudera Machine Learning and Cloudera Data Science Workbench

This topic highlights some key differences between Cloudera Data Science Workbench and its cloud-native counterpart, Cloudera Machine Learning.

How is Cloudera Machine Learning (CML) related to Cloudera Data Science Workbench (CDSW)?

CML expands the end-to-end workflow of Cloudera Data Science Workbench (CDSW) with cloud-native benefits like rapid provisioning, elastic autoscaling, distributed dependency isolation, and distributed GPU training.

It can run its own native distributed computing workloads without requiring a separate CDH cluster for scale-out compute. It is designed to run on CDP in existing Kubernetes environments, such as managed cloud Kubernetes services (EKS, AKS, GKE) or Red Hat OpenShift, reducing operational costs for some customers while delivering multi-cloud portability.

Both products help data engineers and data science teams be more productive on shared data and compute, with strong security and governance. They share extensive code.

There is one primary difference:

- CDSW extends an existing CDH cluster, by running on gateway nodes and pushing distributed compute workloads to the cluster. CDSW requires and supports a single CDH cluster for its distributed compute, including Apache Spark.

- In contrast, CML is self-contained and manages its own distributed compute, natively running workloads - including but not limited to Apache Spark - in containers on Kubernetes.

Note: It can still connect to an existing cluster to leverage its distributed compute, data, or metadata (SDX).

Table 1: Key Differences

	CDSW	CML
Architecture	CDSW can run on a CDP-DC, CDH (5 or 6), and HDP cluster and runs on one or more dedicated gateway nodes on the cluster.	CML is self-contained and does not require an attached CDH/HDP cluster.
	Notion of 1 master and multiple worker hosts.	No designated master and worker hosts; all nodes are ephemeral.
Security	Kerberos authentication integrated via the CDH/HDP cluster	Centralised identity management using FreeIPA via the Cloudera Data Platform (CDP).
	External authentication via LDAP/SAML.	
App Storage	Project files, internal postgresDB, and Livelog, are all stored persistently on the Master host.	All required persistent storage is on cloud-managed block store, NFS, and a relational data store. For example, for AWS, this is managed via EFS.
Compute	Python/R/Scala workloads execute on the CDSW gateway nodes of the cluster.	Python/R/Scala workloads run on the CDP/cloud-provider-managed K8s cluster.
	CDSW pushes distributed compute workloads, such as Spark-on-YARN, to the CDH/HDP cluster.	Spark-on-YARN is not supported; Spark-on-K8s instead. Workloads will run on a dedicated K8s cluster provisioned within the customer environment.
	No autoscaling.	Autoscaling via your cloud service provider. Kubernetes/node-level autoscaling will be used to expand/contract the cluster size based on demand.
Packaging	Available as a downloadable RPM and CSD.	Available as a managed service on CDP.
	Spark is packaged with CDH.	Spark on K8s is packaged with CML - no dependency on an external cluster.
Data Access	Data usually resides on the attached CDH/HDP cluster in HDFS, Hive, HBase, and so on.	Data can reside on object storage such as S3 or any pre-existing workload clusters registered with CDP.