Legal Notice

© Cloudera Inc. 2021. 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.

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 CLOUDEA, CLOUDEA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDEA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDEA DOES NOT WARRANT THAT CLOUDEA 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, CLOUDEA 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

## Concepts

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloudera Machine Learning Overview</td>
<td>8</td>
</tr>
<tr>
<td>Apache Spark 2 on CML</td>
<td>9</td>
</tr>
<tr>
<td>Key Differences between Cloudera Machine Learning and Cloudera Data Science Workbench</td>
<td>10</td>
</tr>
<tr>
<td>Challenges with Machine Learning in production</td>
<td>12</td>
</tr>
<tr>
<td>Challenges with model deployment and serving</td>
<td>12</td>
</tr>
<tr>
<td>Challenges with model monitoring</td>
<td>12</td>
</tr>
<tr>
<td>Challenges with model governance</td>
<td>13</td>
</tr>
<tr>
<td>Basic Concepts and Terminology</td>
<td>15</td>
</tr>
<tr>
<td>ML Runtimes versus the Legacy Engine</td>
<td>17</td>
</tr>
<tr>
<td>Engine Dependencies</td>
<td>17</td>
</tr>
<tr>
<td>Engines for Experiments and Models</td>
<td>19</td>
</tr>
<tr>
<td>Snapshot Code</td>
<td>19</td>
</tr>
<tr>
<td>Build Image</td>
<td>20</td>
</tr>
<tr>
<td>Run Experiment / Deploy Model</td>
<td>21</td>
</tr>
<tr>
<td>Environmental Variables</td>
<td>21</td>
</tr>
<tr>
<td>Model Training and Deployment Overview</td>
<td>22</td>
</tr>
<tr>
<td>Experiments</td>
<td>22</td>
</tr>
<tr>
<td>Experiments - Concepts and Terminology</td>
<td>23</td>
</tr>
<tr>
<td>Models</td>
<td>24</td>
</tr>
<tr>
<td>Models - Concepts and Terminology</td>
<td>24</td>
</tr>
<tr>
<td>Collaborating on Projects with Cloudera Machine Learning</td>
<td>26</td>
</tr>
<tr>
<td>Project Collaborators</td>
<td>26</td>
</tr>
<tr>
<td>Teams</td>
<td>27</td>
</tr>
<tr>
<td>Forking Projects</td>
<td>27</td>
</tr>
<tr>
<td>Collaborating with Git</td>
<td>27</td>
</tr>
<tr>
<td>Sharing Job and Session Console Outputs</td>
<td>27</td>
</tr>
<tr>
<td>Autoscaling Workloads with Kubernetes</td>
<td>28</td>
</tr>
<tr>
<td>Autoscale Groups</td>
<td>29</td>
</tr>
<tr>
<td>Critical and Non-critical Pods</td>
<td>29</td>
</tr>
</tbody>
</table>

## Planning

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to Private Cloud</td>
<td>29</td>
</tr>
<tr>
<td>Cloudera Machine Learning requirements</td>
<td>30</td>
</tr>
<tr>
<td>Get started with CML on Private Cloud</td>
<td>31</td>
</tr>
<tr>
<td>Test your connectivity to the CDP-DC cluster</td>
<td>32</td>
</tr>
<tr>
<td>Differences Between Public and Private Cloud</td>
<td>32</td>
</tr>
<tr>
<td>Limitations on Private Cloud</td>
<td>33</td>
</tr>
<tr>
<td>Network File System</td>
<td>33</td>
</tr>
<tr>
<td>Deploy an ML Workspace with Support for TLS</td>
<td>36</td>
</tr>
<tr>
<td>Replace a Certificate</td>
<td>36</td>
</tr>
</tbody>
</table>

## How To

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provision an ML Workspace</td>
<td>37</td>
</tr>
<tr>
<td>Monitoring ML Workspaces</td>
<td>38</td>
</tr>
<tr>
<td>Removing ML Workspaces</td>
<td>38</td>
</tr>
<tr>
<td>Managing your Personal Account</td>
<td>39</td>
</tr>
<tr>
<td>Contents</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>vi</td>
<td></td>
</tr>
</tbody>
</table>

**Troubleshooting**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Troubleshooting</td>
<td>152</td>
</tr>
<tr>
<td>Troubleshooting ML Workspaces on AWS</td>
<td>152</td>
</tr>
<tr>
<td>Common CML Errors and Solutions</td>
<td>154</td>
</tr>
<tr>
<td>Troubleshooting ML Workspaces on Azure</td>
<td>157</td>
</tr>
<tr>
<td>Logs for ML Workspaces</td>
<td>157</td>
</tr>
<tr>
<td>Downloading Diagnostic Bundles for a Workspace</td>
<td>158</td>
</tr>
<tr>
<td>Troubleshooting Issues with Workloads</td>
<td>159</td>
</tr>
<tr>
<td>Troubleshooting Kerberos Errors</td>
<td>159</td>
</tr>
</tbody>
</table>

**Reference**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Command Line Tools in CML</td>
<td>160</td>
</tr>
<tr>
<td>Cloudera Machine Learning Command Line Reference</td>
<td>160</td>
</tr>
<tr>
<td>cdswctl Command Line Interface Client</td>
<td>160</td>
</tr>
<tr>
<td>Download and Configure cdswctl</td>
<td>161</td>
</tr>
<tr>
<td>Initialize an SSH Endpoint</td>
<td>162</td>
</tr>
<tr>
<td>Log into cdswctl</td>
<td>163</td>
</tr>
<tr>
<td>Prepare to manage models using the model CLI</td>
<td>164</td>
</tr>
<tr>
<td>Create a model using the CLI</td>
<td>164</td>
</tr>
<tr>
<td>Build and deployment commands for models</td>
<td>166</td>
</tr>
<tr>
<td>Deploy a new model with updated resources</td>
<td>167</td>
</tr>
<tr>
<td>View replica logs for a model</td>
<td>167</td>
</tr>
<tr>
<td>Using ML Runtimes with cdswctl</td>
<td>168</td>
</tr>
<tr>
<td>cdswctl command reference</td>
<td>170</td>
</tr>
<tr>
<td>Azure NetApp Files Management with the CLI</td>
<td>171</td>
</tr>
<tr>
<td>Accessing Local Data from Your Computer</td>
<td>172</td>
</tr>
<tr>
<td>Accessing Data from HDFS</td>
<td>173</td>
</tr>
<tr>
<td>Accessing Data from Apache Hive</td>
<td>174</td>
</tr>
<tr>
<td>Accessing Data from Apache Impala</td>
<td>174</td>
</tr>
<tr>
<td>Loading CSV Data into an Impala Table</td>
<td>174</td>
</tr>
<tr>
<td>Running Queries on Impala Tables</td>
<td>175</td>
</tr>
<tr>
<td>Accessing Data in Amazon S3 Buckets</td>
<td>177</td>
</tr>
<tr>
<td>Accessing External SQL Databases</td>
<td>178</td>
</tr>
<tr>
<td>Accessing a Data Lake from CML</td>
<td>179</td>
</tr>
<tr>
<td>Example: Connect a Spark session to Hive Metastore in a Data Lake</td>
<td>184</td>
</tr>
<tr>
<td>Accessing Ozone from Spark</td>
<td>187</td>
</tr>
</tbody>
</table>
Built-in CML Visualizations............................................................................................................................ 187
  Simple Plots.................................................................................................................................................. 188
  Saved Images............................................................................................................................................... 188
  HTML Visualizations................................................................................................................................. 188
  IFrame Visualizations................................................................................................................................. 189
  Grid Displays............................................................................................................................................... 190
  Documenting Your Analysis....................................................................................................................... 191
Cloudera Data Visualization for ML............................................................................................................... 191
Jupyter Magic Commands................................................................................................................................ 191
  Python......................................................................................................................................................192
  Scala............................................................................................................................................................ 192

Release Notes................................................................................................................................................. 193
  What's New................................................................................................................................................... 193
  Known Issues and Limitations.................................................................................................................. 194
The Cloudera Machine Learning Experience encompasses a wide range of functions and features for enterprise-grade data science and machine learning applications. The Concepts section provides overviews of the main functional areas.

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

---

**Related Concepts**
- Apache Spark 2 on CML
- Key Differences between Cloudera Machine Learning and Cloudera Data Science Workbench

**Related Information**
- Management Console: Core Concepts
- Engines Overview

---

**Apache Spark 2 on CML**

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.

In Cloudera Machine Learning (CML), Spark and its dependencies are bundled directly into the CML engine Docker image.

CML supports fully-containerized execution of Spark workloads via Spark's support for the Kubernetes cluster backend. Users can interact with Spark both interactively and in batch mode.
Note: Spark is not currently supported for ML Runtimes.

Dependency Management: In both batch and interactive modes, dependency management, including for Spark executors, is transparently managed by CML and Kubernetes. No extra required configuration is required. In interactive mode, CML leverages your cloud provider for scalable project storage, and in batch mode, CML manages dependencies through container images.

Autoscaling: CML also supports native cloud autoscaling via Kubernetes. When clusters do not have the required capacity to run workloads, they can automatically scale up additional nodes. Administrators can configure autoscaling upper limits, which determine how large a compute cluster can grow. Since compute costs increase as cluster size increases, having a way to configure upper limits gives administrators a method to stay within a budget. Autoscaling policies can also account for heterogeneous node types such as GPU nodes.

Workload Isolation: In CML, each project is owned by a user or team. Users can launch multiple sessions in a project. Workloads are launched within a separate Kubernetes namespace for each user, thus ensuring isolation between users at the K8s level.

Related Concepts
Cloudera Machine Learning Overview

Key Differences between Cloudera Machine Learning and Cloudera Data Science Workbench

For existing Cloudera Machine Learning users, this topic will help highlight 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.
- **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

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

**Related Concepts**

Cloudera Machine Learning Overview

**Related Information**

Cloudera Machine Learning documentation
Challenges with Machine Learning in production

One of the hardest parts of Machine Learning (ML) is deploying and operating ML models in production applications. These challenges fall mainly into the following categories: model deployment and serving, model monitoring, and model governance.

Challenges with model deployment and serving

After models are trained and ready to deploy in a production environment, lack of consistency with model deployment and serving workflows can present challenges in terms of scaling your model deployments to meet the increasing numbers of ML usecases across your business.

Many model serving and deployment workflows have repeatable, boilerplate aspects which you can automate using modern DevOps techniques like high frequency deployment and microservices architectures. This approach can enable the ML engineers to focus on the model instead of the surrounding code and infrastructure.

Challenges with model monitoring

Machine Learning (ML) models predict the world around them which is constantly changing. The unique and complex nature of model behavior and model lifecycle present challenges after the models are deployed.

Cloudera Machine Learning provides you the capability to monitor the performance of the model on two levels: technical performance (latency, throughput, and so on similar to an Application Performance Management), and mathematical performance (is the model predicting correctly, is the model biased, and so on).

There are two types of metrics that are collected from the models:

• Time series metrics: Metrics measured in-line with model prediction. It can be useful to track the changes in these values over time. It is the finest granular data for the most recent measurement. To improve performance, older data is aggregated to reduce data records and storage.

• Post-prediction metrics: Metrics that are calculated after prediction time, based on ground truth and/or batches (aggregates) of time series metrics. To collect metrics from the models, the Python SDK has been extended to include the following functions that you can use to store different types of metrics:

To collect metrics from the models, the Python SDK has been extended to include the following functions that you can use to store different types of metrics:

• track_metrics: Tracks the metrics generated by experiments and models.
• read_metrics: Reads the metrics already tracked for a deployed model, within a given window of time.
• track_delayed_metrics: Tracks metrics that correspond to individual predictions, but aren’t known at the time the prediction is made. The most common instances are ground truth and metrics derived from ground truth such as error metrics.
• track_aggregate_metrics: Registers metrics that are not associated with any particular prediction. This function can be used to track metrics accumulated and/or calculated over a longer period of time.

The following two use-cases show how you can use these functions:

• Tracking accuracy of a model over time
• Tracking drift

Use-case 1: Tracking accuracy of a model over time

Consider the case of a large telco. When a customer service representative takes a call from a customer, a web application presents an estimate of the risk that the customer will churn. The service representative takes this risk into account when evaluating whether to offer promotions.

The web application obtains the risk of churn by calling into a model hosted on Cloudera Machine Learning (CML). For each prediction thus obtained, the web application records the UUID into a datastore alongside the customer ID. The prediction itself is tracked in CML using the track_metrics function.
At some point in the future, some customers do in fact churn. When a customer churns, they or another customer service representative close their account in a web application. That web application records the churn event, which is ground truth for this example, in a datastore.

An ML engineer who works at the telco wants to continuously evaluate the suitability of the risk model. To do this, they create a recurring CML job. At each run, the job uses the `read_metrics` function to read all the predictions that were tracked in the last interval. It also reads in recent churn events from the ground truth datastore. It joins the churn events to the predictions and customer ID’s using the recorded UUID’s, and computes an Receiver operating characteristic (ROC) metric for the risk model. The ROC is tracked in the metrics store using the `track_aggregate_metrics` function.

![Diagram](image)

**Note:** You can store the ground truth in an external datastore, such as Cloudera Data Warehouse or in the metrics store.

**Use-case 2: Tracking drift**

Instead of or in addition to computing ROC, the ML engineer may need to track various types of drift. Drift metrics are especially useful in cases where ground truth is unavailable or is difficult to obtain.

The definition of drift is broad and somewhat nebulous and practical approaches to handling it are evolving, but drift is always about changing distributions. The distribution of the input data seen by the model may change over time and deviate from the distribution in the training dataset, and/or the distribution of the output variable may change, and/or the relationship between input and output may change.

All drift metrics are computed by aggregating batches of predictions in some way. As in the use case above, batches of predictions can be read into recurring jobs using the `read_metrics` function, and the drift metrics computed by the job can be tracked using the `track_aggregate_metrics` function.

**Challenges with model governance**

Businesses implement ML models across their entire organization, spanning a large spectrum of usecases. When you start deploying more than just a couple models in production, a lot of complex governance and management challenges arise.
Almost all the governance needs for ML are associated with data and are tied directly to the data management practice in your organization. For example, what data can be used for certain applications, who should be able to access what data, and based on what data are models created.

Some of the other unique governance challenges that you could encounter are:

- How to gain visibility into the impact your models have on your customers?
- How can you ensure you are still compliant with both governmental and internal regulations?
- How does your organization’s security practices apply to the models in production?

Ultimately, the needs for ML governance can be distilled into the following key areas: model visibility, and model explainability, interpretability, and reproducibility.

**Model visibility**
A basic requirement for model governance is enabling teams to understand how machine learning is being applied in their organizations. This requires a canonical catalog of models in use. In the absence of such a catalog, many organizations are unaware of how their models work, where they are deployed, what they are being used for, and so on. This leads to repeated work, model inconsistencies, recomputing features, and other inefficiencies.

**Model explainability, interpretability, and reproducibility**
Models are often seen as a black box: data goes in, something happens, and a prediction comes out. This lack of transparency is challenging on a number of levels and is often represented in loosely related terms explainability, interpretability, and reproducibility.

- Explainability: Indicates the description of the internal mechanics of an Machine Learning (ML) model in human terms
- Interpretability: Indicates the ability to:
  - Understand the relationship between model inputs, features and outputs
  - Predict the response to changes in inputs
- Reproducibility: Indicates the ability to reproduce the output of a model in a consistent fashion for the same inputs

To solve these challenges, CML provides an end-to-end model governance and monitoring workflow that gives organizations increased visibility into their machine learning workflows and aims to eliminate the blackbox nature of most machine learning models.

The following image shows the end-to-end production ML workflow:

**Figure 1: Production ML Workflow**
Model governance using Apache Atlas
To address governance challenges, Cloudera Machine Learning uses Apache Atlas to automatically collect and visualize lineage information for data used in Machine Learning (ML) workflows — from training data to model deployments.

Lineage is a visual representation of the project. The lineage information includes visualization of the relationships between model entities such as code, model builds, deployments, and so on. and the processes that carry out transformations on the data involved, such as create project, build model, deploy model, and so on.

The Apache Atlas type system has pre-built governance features that can be used to define ML metadata objects. A type in Atlas is a definition of the metadata stored to describe a data asset or other object or process in an environment. For ML governance, Cloudera has designed new Atlas types that capture ML entities and processes as Atlas metadata objects.

In addition to the definition of the types, Atlas also captures the relationship between the entities and processes to visualize the end-to-end lineage flow, as shown in the following image. The blue hexagons represent an entity (also called the noun) and the green hexagons represent a process (also called the verb).

The ML metadata definition closely follows the actual machine learning workflow. Training data sets are the starting point for a model lineage flow. These data sets can be tables from a data warehouse or an embedded csv file. Once a data set has been identified, the lineage follows into training, building and deploying the model.

See ML operations entities created in Atlas for a list of metadata that Atlas collects from each CML workspace. Metadata is collected from machine learning projects, model builds, and model deployments, and the processes that create these entities.

Basic Concepts and Terminology

In the context of Cloudera Machine Learning, engines are responsible for running data science workloads and intermediating access to the underlying cluster. This topic walks you through some basic concepts and terminology related to engines.

Cloudera Machine Learning 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.
Cloudera Machine Learning engines are responsible for running R, Python, and Scala code written by users. You can think of an engine as a virtual machine, customized to have all the necessary dependencies while keeping each project’s environment entirely isolated.

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

**Base Engine Image**

The base engine image is a Docker image that contains all the building blocks needed to launch a Cloudera Machine Learning session and run a workload. It consists of kernels for Python, R, and Scala along with additional libraries that can be used to run common data analytics operations. When you launch a session to run a project, an engine is kicked off from a container of this image. The base image itself is built and shipped along with Cloudera Machine Learning.

Cloudera Machine Learning offers legacy engines and Machine Learning Runtimes. Both legacy engines and ML Runtimes are Docker images and contain OS, interpreters, and libraries to run user code in sessions, jobs, experiments, models, and applications. However, there are significant differences between these choices. See *ML Runtimes versus Legacy Engines* for a summary of these differences.

New versions of the base engine image are released periodically. However, existing projects are not automatically upgraded to use new engine images. Older images are retained to ensure you are able to test code compatibility with the new engine before upgrading to it manually.

**Engine**

The term engine refers to a virtual machine-style environment that is created when you run a project (via session or job) in Cloudera Machine Learning. You can use an engine to run R, Python, and Scala workloads on data stored in the underlying CDH cluster.

Cloudera Machine Learning allows you to run code using either a session or a job. A session is a way to interactively launch an engine and run code while a job lets you batch process those actions and schedule them to run recursively. Each session and job launches its own engine that lives as long as the workload is running (or until it times out).

A running engine includes the following components:

![Engine Environment Diagram](image)
• Kernel

Each engine runs a kernel with an R, Python or Scala process that can be used to run code within the engine. The kernel launched differs based on the option you select (either Python 2/3, PySpark, R, or Scala) when you launch the session or configure a job.

The Python kernel is based on the Jupyter IPython kernel; the R kernel is custom-made for CML; and the Scala kernel is based on the Apache Toree kernel.

• Project Filesystem Mount

Cloudera Machine Learning uses a persistent filesystem to store project files such as user code, installed libraries, or even small data files. Project files are stored on the master host at /var/lib/cdsw/current/projects.

Every time you launch a new session or run a job for a project, a new engine is created, and the project filesystem is mounted into the engine’s environment at /home/cdsw. Once the session/job ends, the only project artifacts that remain are a log of the workload you ran, and any files that were generated or modified, including libraries you might have installed. All of the installed dependencies persist through the lifetime of the project. The next time you launch a session/job for the same project, those dependencies will be mounted into the engine environment along with the rest of the project filesystem.

• Host Mounts

If there are any files on the hosts that should be mounted into the engines at launch time, use the Site Administration panel to include them.

For detailed instructions, see Configuring the Engine Environment.

Related Information
ML Runtimes versus Legacy Engines
Configuring the Engine Environment

ML Runtimes versus the Legacy Engine

Runtimes and the Legacy Engine serve the same basic goal: they are container images that contain a complete Linux OS, interpreter(s), and libraries. They are the environment in which your code runs.

There are many runtimes. Currently each runtime contains a single interpreter (for example, Python 3.8, R 4.0, etc.) and a set of UNIX tools including gcc. Each runtime supports a single UI for running code (for example, the Workbench or JupyterLab).

There is one Legacy Engine. The engine is monolithic. It contains the machinery necessary to run sessions using all four engine interpreter options that we currently support (Python 2, Python 3, R, and Scala) and a much larger set of UNIX tools including LaTeX.

Runtimes are the future of CML. Their design keeps the images small, which is good for performance, maintenance, and security.

Runtimes will soon exceed the capabilities of the Legacy Engine, but they do not yet cover the full engine feature set. In particular, you cannot deploy a custom runtime, R runtimes do not support Models, there are no Scala runtimes, and runtimes do not support Workers.

Engine Dependencies

This topic describes the options available to you for mounting a project’s dependencies into its engine environment. Depending on your projects or user preferences, one or more of these methods may be more appropriate for your deployment.
Important: Even though experiments and models are created within the scope of a project, the engines they use are completely isolated from those used by sessions or jobs launched within the same project. For details, see Engines for Experiments and Models.

Installing Packages Directly Within Projects

Creating a Customized Engine with the Required Package(s)

Directly installing a package to a project as described above might not always be feasible. For example, packages that require root access to be installed, or that must be installed to a path outside /home/cdsw (outside the project mount), cannot be installed directly from the workbench. For such circumstances, Cloudera recommends you extend the base Cloudera Machine Learning engine image to build a customized image with all the required packages installed to it.

This approach can also be used to accelerate project setup across the deployment. For example, if you want multiple projects on your deployment to have access to some common dependencies out of the box or if a package just has a complicated setup, it might be easier to simply provide users with an engine environment that has already been customized for their project(s).
Concepts

For detailed instructions with an example, see *Configuring the Engine Environment*.

**Managing Dependencies for Spark 2 Projects**

With Spark projects, you can add external packages to Spark executors on startup. To add external dependencies to Spark jobs, specify the libraries you want added by using the appropriate configuration parameters in a `spark-defaults.conf` file.

For a list of the relevant properties and examples, see *Spark Configuration Files*.

**Managing Dependencies for Experiments and Models**

To allow for versioned experiments and models, Cloudera Machine Learning executes each experiment and model in a completely isolated engine. Every time a model or experiment is kicked off, Cloudera Machine Learning creates a new isolated Docker image where the model or experiment is executed. These engines are built by extending the project's designated default engine image to include the code to be executed and any dependencies as specified.

For details on how this process works and how to configure these environments, see *Engines for Experiments and Models*.

**Related Information**

- Engines for Experiments and Models
- Installing Additional Packages
- Spark Configuration Files
- Configuring the Engine Environment

**Engines for Experiments and Models**

In Cloudera Machine Learning, models, experiments, jobs, and sessions are all created and executed within the context of a project. We've described the different ways in which you can customize a project's engine environment for sessions and jobs in *Environmental Variables*. However, engines for models and experiments are completely isolated from the rest of the project.

Every time a model or experiment is kicked off, Cloudera Machine Learning creates a new isolated Docker image where the model or experiment is executed. This isolation in build and execution makes it possible for Cloudera Machine Learning to keep track of input and output artifacts for every experiment you run. In case of models, versioned builds give you a way to retain build history for models and a reliable way to rollback to an older version of a model if needed.

The following topics describe the engine build process that occurs when you kick off a model or experiment.

**Related Information**

- Environmental Variables

**Snapshot Code**

When you first launch an experiment or model, Cloudera Machine Learning takes a Git snapshot of the project filesystem at that point in time. This Git server functions behind the scenes and is completely separate from any other Git version control system you might be using for the project as a whole.

However, this Git snapshot will recognize the `.gitignore` file defined in the project. This means if there are any artifacts (files, dependencies, etc.) larger than 50 MB stored directly in your project filesystem, make sure to add those files or folders to `.gitignore` so that they are not recorded as part of the snapshot. This ensures that the experiment/model environment is truly isolated and does not inherit dependencies that have been previously installed in the project workspace.
By default, each project is created with the following `.gitignore` file:

```bash
R
node_modules
*.pyc
*.pyc
!.gitignore
```

Augment this file to include any extra dependencies you have installed in your project workspace to ensure a truly isolated workspace for each model/experiment.

### Build Image

Once the code snapshot is available, Cloudera Machine Learning creates a new Docker image with a copy of the snapshot.

The new image is based off the project's designated default engine image (configured at **Project Settings > Engine**). The image environment can be customized by using environmental variables and a build script that specifies which packages should be included in the new image.

### Environmental Variables

Both models and experiments inherit environmental variables from their parent project. Furthermore, in case of models, you can specify environment variables for each model build. In case of conflicts, the variables specified per-build will override any values inherited from the project.

For more information, see **Engine Environment Variables**.

### Build Script - `cdsw-build.sh`

As part of the Docker build process, Cloudera Machine Learning runs a build script called `cdsw-build.sh` file. You can use this file to customize the image environment by specifying any dependencies to be installed for the code to run successfully. One advantage to this approach is that you now have the flexibility to use different tools and libraries in each consecutive training run. Just modify the build script as per your requirements each time you need to test a new library or even different versions of a library.

**Important:**

- The `cdsw-build.sh` script does not exist by default -- it has to be created by you within each project as needed.
- The name of the file is not customizable. It must be called `cdsw-build.sh`.

The following sections demonstrate how to specify dependencies in Python and R projects so that they are included in the build process for models and experiments.

#### Python

For Python, create a `requirements.txt` file in your project with a list of packages that must be installed. For example:

**Figure 2: requirements.txt**

```
beautifulsoup4==4.6.0
seaborn==0.7.1
```

Then, create a `cdsw-build.sh` file in your project and include the following command to install the dependencies listed in `requirements.txt`.

**Figure 3: cdsw-build.sh**

```
pip3 install -r requirements.txt
```
Now, when cdsw-build.sh is run as part of the build process, it will install the BeautifulSoup4 and seaborn packages to the new image built for the experiment/model.

R

For R, create a script called install.R with the list of packages that must be installed. For example:

**Figure 4: install.R**

```r
install.packages(repos="https://cloud.r-project.org", c("tidyr", "stringr"))
```

Then, create a cdsw-build.sh file in your project and include the following command to run install.R.

**Figure 5: cdsw-build.sh**

```bash
Rscript install.R
```

Now, when cdsw-build.sh is run as part of the build process, it will install the tidyr and stringr packages to the new image built for the experiment/model.

If you do not specify a build script, the build process will still run to completion, but the Docker image will not have any additional dependencies installed. At the end of the build process, the built image is then pushed to an internal Docker registry so that it can be made available to all the Cloudera Machine Learning hosts. This push is largely transparent to the end user.

**Note:** If you want to test your code in an interactive session before you run an experiment or deploy a model, run the cdsw-build.sh script directly in the workbench. This will allow you to test code in an engine environment that is similar to one that will eventually be built by the model/experiment build process.

**Run Experiment / Deploy Model**

Once the Docker image has been built and pushed to the internal registry, the experiment/model can now be executed within this isolated environment.

In case of experiments, you can track live progress as the experiment executes in the experiment’s Session tab.

Unlike experiments, models do not display live execution progress in a console. Behind the scenes, Cloudera Machine Learning will move on to deploying the model in a serving environment based on the computing resources and replicas you requested. Once deployed you can go to the model's Monitoring page to view statistics on the number of requests served/dropped and stderr/stdout logs for the model replicas.

**Environmental Variables**

This topic explains how environmental variables are propagated through an ML workspace.

Environmental variables help you customize engine environments, both globally and for individual projects/jobs. For example, if you need to configure a particular timezone for a project or increase the length of the session/job timeout windows, you can use environmental variables to do so. Environmental variables can also be used to assign variable names to secrets, such as passwords or authentication tokens, to avoid including these directly in the code.

For a list of the environmental variables you can configure and instructions on how to configure them, see Engine Environment Variables.

**Related Information**

Configuring Engine Environment Variables
Model Training and Deployment Overview

This section provides an overview of model training and deployment using Cloudera Machine Learning.

Machine learning is a discipline that uses computer algorithms to extract useful knowledge from data. There are many different types of machine learning algorithms, and each one works differently. In general however, machine learning algorithms begin with an initial hypothetical model, determine how well this model fits a set of data, and then work on improving the model iteratively. This training process continues until the algorithm can find no additional improvements, or until the user stops the process.

A typical machine learning project will include the following high-level steps that will transform a loose data hypothesis into a model that serves predictions.

1. Explore and experiment with and display findings of data
2. Deploy automated pipelines of analytics workloads
3. Train and evaluate models
4. Deploy models as REST APIs to serve predictions

With Cloudera Machine Learning, you can deploy the complete lifecycle of a machine learning project from research to deployment.

Experiments

This topic introduces you to experiments, and the challenge this features aims to solve.

Cloudera Machine Learning allows data scientists to run batch experiments that track different versions of code, input parameters, and output (both metrics and files).

Challenge

As data scientists iteratively develop models, they often experiment with datasets, features, libraries, algorithms, and parameters. Even small changes can significantly impact the resulting model. This means data scientists need the ability to iterate and repeat similar experiments in parallel and on demand, as they rely on differences in output and scores to tune parameters until they obtain the best fit for the problem at hand. Such a training workflow requires versioning of the file system, input parameters, and output of each training run.

Without versioned experiments you would need intense process rigor to consistently track training artifacts (data, parameters, code, etc.), and even then it might be impossible to reproduce and explain a given result. This can lead to wasted time/effort during collaboration, not to mention the compliance risks introduced.
Cloudera Machine Learning uses experiments to facilitate ad-hoc batch execution and model training. Experiments are batch executed workloads where the code, input parameters, and output artifacts are versioned. This feature also provides a lightweight ability to track output data, including files, metrics, and metadata for comparison.

**Experiments - Concepts and Terminology**

This topic walks you through some basic concepts and terminology related to experiments.

The term experiment refers to a non interactive batch execution script that is versioned across input parameters, project files, and output. Batch experiments are associated with a specific project (much like sessions or jobs) and have no notion of scheduling; they run at creation time. To support versioning of the project files and retain run-level artifacts and metadata, each experiment is executed in an isolated container.

**Lifecycle of an Experiment**

The rest of this section describes the different stages in the lifecycle of an experiment - from launch to completion.

1. **Launch Experiment**

   In this step you will select a script from your project that will be run as part of the experiment, and the resources (memory/GPU) needed to run the experiment. The engine kernel will be selected by default based on your script. For detailed instructions on how to launch an experiment, see *Getting Started with Cloudera Machine Learning*.

2. **Build**

   When you launch the experiment, Cloudera Machine Learning first builds a new versioned engine image where the experiment will be executed in isolation. This new engine includes:
   - the base engine image used by the project (check Project > Settings)
   - a snapshot of the project filesystem
   - environmental variables inherited from the project.
   - packages explicitly specified in the project's build script (`cdsw-build.sh`)

   It is your responsibility to provide the complete list of dependencies required for the experiment via the `cdsw-build.sh` file. As part of the engine's build process, Cloudera Machine Learning will run the `cdsw-build.sh` script and install the packages or libraries requested there on the new image.

   For details about the build process and examples on how to specify dependencies, see Engines for Experiments and Models.

3. **Schedule**

   Once the engine is built the experiment is scheduled for execution like any other job or session. Once the requested CPU/GPU and memory have been allocated to the experiment, it will move on to the execution stage.

   Note that if your deployment is running low on memory and CPU, your runs may spend some time in this stage.
4. Execute

This is the stage where the script you have selected will be run in the newly built engine environment. This is the same output you would see if you had executed the script in a session in the Workbench console.

You can watch the execution in progress in the individual run's Session tab.

You can also go to the project Overview > Experiments page to see a table of all the experiments launched within that project and their current status.

Run ID: A numeric ID that tracks all experiments launched on a Cloudera Machine Learning deployment. It is not limited to the scope of a single user or project.

Related Information
Running an Experiment with Cloudera Machine Learning

Models

Cloudera Machine Learning allows data scientists to build, deploy, and manage models as REST APIs to serve predictions.

Challenge

Data scientists often develop models using a variety of Python/R open source packages. The challenge lies in actually exposing those models to stakeholders who can test the model. In most organizations, the model deployment process will require assistance from a separate DevOps team who likely have their own policies about deploying new code.

For example, a model that has been developed in Python by data scientists might be rebuilt in another language by the devops team before it is actually deployed. This process can be slow and error-prone. It can take months to deploy new models, if at all. This also introduces compliance risks when you take into account the fact that the new re-developed model might not be even be an accurate reproduction of the original model.

Once a model has been deployed, you then need to ensure that the devops team has a way to rollback the model to a previous version if needed. This means the data science team also needs a reliable way to retain history of the models they build and ensure that they can rebuild a specific version if needed. At any time, data scientists (or any other stakeholders) must have a way to accurately identify which version of a model is/was deployed.

Solution

Cloudera Machine Learning allows data scientists to build and deploy their own models as REST APIs. Data scientists can now select a Python or R function within a project file, and Cloudera Machine Learning will:

- Create a snapshot of model code, model parameters, and dependencies.
- Package a trained model into an immutable artifact and provide basic serving code.
- Add a REST endpoint that automatically accepts input parameters matching the function, and that returns a data structure that matches the function’s return type.
- Save the model along with some metadata.
- Deploy a specified number of model API replicas, automatically load balanced.

Models - Concepts and Terminology

Model

Model is a high level abstract term that is used to describe several possible incarnations of objects created during the model deployment process. For the purpose of this discussion you should note that 'model' does not always refer to a specific artifact. More precise terms (as defined later in this section) should be used whenever possible.

Stages of the Model Deployment Process
The rest of this section contains supplemental information that describes the model deployment process in detail.

Create

- **File** - The R or Python file containing the function to be invoked when the model is started.
- **Function** - The function to be invoked inside the file. This function should take a single JSON-encoded object (for example, a python dictionary) as input and return a JSON-encodable object as output to ensure compatibility with any application accessing the model using the API. JSON decoding and encoding for model input/output is built into Cloudera Machine Learning.

The function will likely include the following components:

- **Model Implementation**
  
  The code for implementing the model (e.g. decision trees, k-means). This might originate with the data scientist or might be provided by the engineering team. This code implements the model's predict function, along with any setup and teardown that may be required.

- **Model Parameters**
  
  A set of parameters obtained as a result of model training/fitting (using experiments). For example, a specific decision tree or the specific centroids of a k-means clustering, to be used to make a prediction.

Build

This stage takes as input the file that calls the function and returns an artifact that implements a single concrete model, referred to as a model build.

- **Built Model**
  
  A built model is a static, immutable artifact that includes the model implementation, its parameters, any runtime dependencies, and its metadata. If any of these components need to be changed, for example, code changes to the implementation or its parameters need to be retrained, a new build must be created for the model. Model builds are versioned using build numbers.

To create the model build, Cloudera Machine Learning creates a Docker image based on the engine designated as the project's default engine. This image provides an isolated environment where the model implementation code will run.

To configure the image environment, you can specify a list of dependencies to be installed in a build script called `cdsw-build.sh`.

For details about the build process and examples on how to install dependencies, see *Engines for Experiments and Models*.

- **Build Number**
  
  Build numbers are used to track different versions of builds within the scope of a single model. They start at 1 and are incremented with each new build created for the model.

Deploy
This stage takes as input the memory/CPU resources required to power the model, the number of replicas needed, and deploys the model build created in the previous stage to a REST API.

- **Deployed Model**
  A deployed model is a model build in execution. A built model is deployed in a model serving environment, likely with multiple replicas.

- **Environmental Variable**
  You can set environmental variables each time you deploy a model. Note that models also inherit any environment variables set at the project and global level. (For more information see *Engine Environment Variables.*) However, in case of any conflicts, variables set per-model will take precedence.

  **Note:** If you are using any model-specific environmental variables, these must be specified every time you re-deploy a model. Models do not inherit environmental variables from previous deployments.

- **Model Replicas**
  The engines that serve incoming requests to the model. Note that each replica can only process one request at a time. Multiple replicas are essential for load-balancing, fault tolerance, and serving concurrent requests. Cloudera Machine Learning allows you to deploy a maximum of 9 replicas per model.

- **Deployment ID**
  Deployment IDs are numeric IDs used to track models deployed across Cloudera Machine Learning. They are not bound to a model or project.

**Related Information**
- Experiments - Concepts and Terminology
- Engines for Experiments and Models
- Engines Environment Variables

## Collaborating on Projects with Cloudera Machine Learning

This topic discusses all the collaboration strategies available to Cloudera Machine Learning users.

### Project Collaborators

If you want to work closely with trusted colleagues on a particular project, you can add them to the project as collaborators. This is recommended for collaboration over projects created under your personal account. Anyone who belongs to your organization can be added as a project collaborator.

Project Visibility Levels: When you create a project in your personal context, Cloudera Machine Learning asks you to assign one of the following visibility levels to the project - Private or Public. Public projects on Cloudera Machine Learning grant read-level access to everyone with access to the Cloudera Machine Learning application. For Private projects, you must explicitly add someone as a project collaborator to grant them access.

Project Collaborator Access Levels: You can grant project collaborators the following levels of access: Viewer, Operator, Contributor, Admin
Collaborating Securely on Projects

Before adding project collaborators, you must remember that assigning the Contributor or Admin role to a project collaborator is the same as giving them write access to your data in CDH. This is because project contributors and project administrators have write access to all your project code (including any library code that you might not be actively inspecting). For example, a contributor/admin could modify project file(s) to insert code that deletes some data on the cluster. The next time you launch a session and run the same code, it will appear as though you deleted the data yourself.

Additionally, project collaborators also have access to all actively running sessions and jobs by default. This means that a malicious user can easily impersonate you by accessing one of your active sessions. Therefore, it is extremely important to restrict project access to trusted collaborators only. Note that site administrators can restrict this ability by allowing only session creators to run commands within their own active sessions.

For these reasons, Cloudera recommends using Git to collaborate securely on shared projects.

Teams

Users who work together on more than one project and want to facilitate collaboration can create a Team. Teams allow streamlined administration of projects. Team projects are owned by the team, rather than an individual user. Only users that are already part of the team can be added as collaborators to projects created within the team context. Team administrators can add or remove members at any time, assigning each member different access permissions.

Team Member Access Levels: You can grant team members the following levels of access: Viewer, Operator, Contributor, Admin.

Forking Projects

You can fork another user's project by clicking Fork on the Project page. Forking creates a new project under your account that contains all the files, libraries, configuration, and jobs from the original project.

Creating sample projects that other users can fork helps to bootstrap new projects and encourage common conventions.

Collaborating with Git

Cloudera Machine Learning provides seamless access to Git projects. Whether you are working independently, or as part of a team, you can leverage all of benefits of version control and collaboration with Git from within Cloudera Machine Learning. Teams that already use Git for collaboration can continue to do so. Each team member will need to create a separate Cloudera Machine Learning project from the central Git repository.

For anything but simple projects, Cloudera recommends using Git for version control. You should work on Cloudera Machine Learning the same way you would work locally, and for most data scientists and developers that means using Git.

Sharing Job and Session Console Outputs

This topic describes how to share the results of your research (that is, output from sessions and jobs) with teammates and project stakeholders.

Cloudera Machine Learning lets you easily share the results of your analysis with one click. Using rich visualizations and documentation comments, you can arrange your console log so that it is a readable record of your analysis and results. This log continues to be available even after the session stops. This method of sharing allows you to show colleagues and collaborators your progress without your having to spend time creating a report.

To share results from an interactive session, click Share at the top of the console page. From here you can generate a link that includes a secret token that gives access to that particular console output. For jobs results, you can either share a link to the latest job result or a particular job run. To share the latest job result, click the Latest Run link for
a job on the Overview page. This link will always have the latest job results. To share a particular run, click on a job run in the job’s History page and share the corresponding link.

You can share console outputs with one of the following sets of users.

- All anonymous users with the link - By default, Cloudera Machine Learning allows anonymous access to shared consoles. However, site administrators can disable anonymous sharing at any time.

  Once anonymous sharing has been disabled, all existing publicly shared console outputs will be updated to be viewable only by authenticated users.

- All authenticated users with the link - This means any user with a Cloudera Machine Learning account will have access to the shared console.

- Specific users and teams - Click Change to search for users and teams to give access to the shared console. You can also come back to the session and revoke access from a user or team the same way.

Sharing Data Visualizations

If you want to share a single data visualization rather than an entire console, you can embed it in another web page. Click the small circular 'link' button located to the left of most rich visualizations to view the HTML snippet that you can use to embed the visualization.

Autoscaling Workloads with Kubernetes

Kubernetes dynamically resizes clusters by using the Kubernetes Cluster Autoscaler (on Amazon EKS) or cluster-autoscaler (on Azure). The cluster autoscaler changes the desired capacity of an autoscaling group to expand or contract a cluster based on pod resource requests.

Scaling Up

The primary trigger for scaling up (or expanding) an autoscaling group is failure by the Kubernetes pod scheduler to find a node that meets the pod’s resource requirements. In Cloudera Machine Learning (CML), if the scheduler cannot find a node to schedule an engine pod because of insufficient CPU or memory, the engine pod will be in “pending” state. When the autoscaler notices this situation, it will change the desired capacity of the autoscaling group (CPU or GPU) to provision a new node in the cluster. As soon as the new node is ready, the scheduler will place the session or engine pod there. In addition to the engine pod, certain CML daemonset pods will also be scheduled on the new node.

The time taken to schedule an engine pod on a new node depends on the amount of time the autoscaler takes to add a new node into the cluster, plus time taken to pull the engine’s Docker image to the new node.

Scaling Down

A cluster is scaled down by the autoscaler by removing a node, when the resource utilization on the given node is less than a pre-defined threshold, provided the node does not have any non-evictable pods running on it. This threshold is currently set to 20% CPU utilization. That is, a node is removed if the following criteria are met:

- The node does not have non-evictable pods
- The node’s CPU utilization is less than 20%
- The number of active nodes in the autoscaling group is more than the configured minimum capacity

It is possible that certain pods might be moved from the evicted node to some other node during the down-scaling process.

Limitations on Azure

On Azure, there are some specific limitations to how autoscaling works.

- CPU nodes cannot scale down to zero. You can only have one or more CPU nodes.
- Autoscaling down is sometimes blocked by Azure services. You can check the cluster autoscaler logs to see if this is occurring.

Autoscaling on Private Cloud
CML on Private Cloud supports application autoscaling on multiple fronts. Additional compute resources are utilized when users self-provision sessions, run jobs, and utilize other compute capabilities. Within a session, users can also leverage the worker API to launch resources necessary to host TensorFlow, pytorch, or other distributed applications. Spark on Kubernetes scales up to any number of executors as requested by the user at runtime.

**Autoscale Groups**

A Cloudera Machine Learning (CML) workspace or cluster consists of three different autoscaling groups: “infra”, “cpu” and “gpu”. These groups scale independently of one another.

**Infra Autoscaling Group**

The Infra autoscaling group is created automatically when a user provisions a CML cluster, and is not configurable from the UI. This group is meant to run the core CML services that are critical to the overall functioning of the workspace. This group is loosely analogous to the master node of legacy CDSW, however it can scale up or down if necessary. The instance count for this group ranges from 1 to 3, with the default set to 2. The instance type used for the group is **m5.2xlarge** on AWS, and **Standard DS4 v2** on Azure.

**CPU Autoscaling Group**

The CPU autoscaling group forms the main worker nodes of a CML cluster, and is somewhat configurable from the UI. The user can choose from three different instance types, and can also set the autoscaling range from 0 to 30 CPU worker nodes. This group is meant to run general CPU-only workloads.

**GPU Autoscaling Group (not available on Azure)**

The GPU autoscaling group consists of instances that have GPUs, and are meant for workloads that require GPU processing. Like the CPU group, this group is configurable from the UI. Unlike the CPU group, this group is meant exclusively for sessions that request > 0 GPUs, and are therefore fenced off from CPU-only workloads, in part because GPU instance types are much more expensive than regular instance types.

**Critical and Non-critical Pods**

The pods running various Cloudera Machine Learning (CML) services and jobs broadly fall into critical and non-critical types.

Critical pods are protected from preemption by autoscaling to avoid interrupting important services. Most of the pods running in the “infra” autoscaling group are critical. Pods that run user sessions, such as engine pods and Spark executor pods, are also considered critical, and are marked as not safe to evict. CML system services that are deployed as daemonsets (they run on all nodes in the cluster) are deemed important, but not critical. These pods are marked as “safe-to-evict” by autoscaling.

**Planning**

The information in this section will help you plan your Cloudera Machine Learning installation.

**Introduction to Private Cloud**

With the Cloudera Machine Learning (CML) service, data scientists and partners can build and run machine learning experiments and workloads in a secure environment. CML on Private Cloud provides an identical experience to CML on Public Cloud, but running in your own on-premises data center.

Cloudera Machine Learning enables you to:

- Easily onboard a new tenant and provision an ML workspace in a shared OpenShift environment.
- Enable data scientists to access shared data on CDP Private Cloud Base and CDW.
- Leverage Spark-on-K8s to spin up and down Spark clusters on demand.
Cloudera Machine Learning requirements

To launch the Cloudera Machine Learning service, the host must meet several requirements. Review the following CML-specific software, NFS server, and storage requirements.

Requirements

If necessary, contact your Administrator to make sure the following requirements are satisfied:

1. The installed OpenShift Container Platform must be version 4.6.x.
2. CML assumes it has cluster-admin privileges on the OpenShift cluster.
3. Storage:
   a. 4 TB of persistent volume block storage per ML Workspace.
   b. 1 TB of NFS space recommended per Workspace (depending on user files).
   c. Access to NFS storage is routable from all pods running in the OpenShift cluster.
   d. For monitoring, recommended volume size is 60 GB.
4. A block storage class must be marked as default in the OpenShift cluster. This may be rook-ceph-block, Portworx, or another storage system. Confirm the storage class by listing the storage classes (run `oc get sc`) in the OpenShift cluster, and check that one of them is marked default.
5. If external NFS is used, the NFS directory and assumed permissions must be those of the cdsw user. For details see Using an External NFS Server in the Related information section at the bottom of this page.
6. If CML needs access to a database on the CDP Private Cloud Base cluster, then the user must be authenticated using Kerberos and must have Ranger policies set up to allow read/write operations to the default (or other specified) database.
7. Ensure that Kerberos is enabled for all services in the cluster. Custom Kerberos principals are not currently supported. For more information, see Enabling Kerberos for authentication.
8. Forward and reverse DNS must be working.
9. DNS lookups to sub-domains and the ML Workspace itself should work.
10. In DNS, wildcard subdomains (such as *.cml.yourcompany.com) must be set to resolve to the master domain (such as cml.yourcompany.com). The TLS certificate (if TLS is used) must also include the wildcard subdomains. When a session or job is started, an engine is created for it, and the engine is assigned to a random, unique subdomain.
11. The external load balancer server timeout needs to be set to 5 min. Without this, creating a project in an ML workspace with `git clone` or with the API may result in API timeout errors. For workarounds, see Known Issue DSE-11837.
12. If you intend to access a workspace over https, see Deploy an ML Workspace with Support for TLS.
13. For non-TLS ML workspaces, websockets need to be allowed for port 80 on the external load balancer.
14. Only a TLS-enabled custom Docker Registry is supported. Ensure that you use a TLS certificate to secure the custom Docker Registry. The TLS certificate can be self-signed, or signed by a private or public trusted Certificate Authority (CA).
15. Due to a Red Hat issue with OpenShift Container Platform 4.3.x, the image registry cluster operator configuration must be set to Managed.
16. Check if storage is set up in the cluster image registry operator. See Known Issues DSE-12778 for further information.

For more information on requirements, see CDP Private Cloud Base Installation Guide.

Hardware requirements

Storage

The OpenShift cluster must have persistent storage classes defined for both block and filesystem volumeModes of storage. Ensure that a block storage class is set up. The exact amount of storage classified as block or filesystem storage depends on the specific workload used.
• Machine Learning workload requirements for storage largely depend on the nature of your machine learning jobs. 4 TB of persistent volume block storage is required per Machine Learning Workspace instance for storing different kinds of metadata related to workspace configuration. Additionally, Machine Learning requires access to NFS storage routable from all pods running in the OpenShift cluster (see below).

• Monitoring uses a large Prometheus instance to scrape workloads. Disk usage depends on scale of workloads. Recommended volume size is 60 GB.

<table>
<thead>
<tr>
<th></th>
<th>Local Storage (for example, ext4)</th>
<th>Block PV (for example, Ceph or Portworx)</th>
<th>NFS (for ML user project files)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Plane</td>
<td>N/A</td>
<td>250 GB</td>
<td>N/A</td>
</tr>
<tr>
<td>CML</td>
<td>N/A</td>
<td>4 TB per workspace</td>
<td>1 TB per workspace (dependent on size of ML user files)</td>
</tr>
</tbody>
</table>

NFS

Cloudera Machine Learning (CML) requires NFS for storing project files and folders. An internal user-space NFS server can be deployed into the cluster which serves a block storage device (persistent volume) managed by the cluster’s software defined storage (SDS) system, such as Ceph or Portworx. This is the recommended option for CML in private cloud. Alternatively, the NFS server can be external to the cluster, such as a NetApp filer that is accessible from the private cloud cluster nodes. NFS storage is to be used only for storing project files and folders, and not for any other CML data, such as PostgreSQL database and LiveLog.

CML does not support shared volumes, such as Portworx shared volumes, for storing project files. A read-write-once (RWO) persistent volume must be allocated to the internal NFS server (for example, NFS server provisioner) as the persistence layer. The NFS server uses the volume to dynamically provision read-write-many (RWX) NFS volumes for the CML clients.

**Related Information**

CDP Private Cloud Base Installation Guide
Known Issues and Limitations
Deploy an ML Workspace with Support for TLS
Using an External NFS Server

**Get started with CML on Private Cloud**

To get started as a user with Cloudera Machine Learning on your Private Cloud, follow the steps described below. They will show you how to set up a Project and work on some data.

**Before you begin**

Make sure the Admin creates a new Workspace for you. If you are an Admin, see: Provision an ML Workspace.

*Note:* Make sure that an Admin user logs into the Workspace first.

**Procedure**

1. Log in to your workspace. On the Workspaces tab, click Launch Workspace.
2. Next, create a Project. See: Creating a Project.
3. Once you have a Project, run a Session to start your work. See: Launch a Session.
4. Test your access to the base cluster (Data Lake). See: CDP-DC cluster connectivity test.
5. You can then run a Model. Learn about Models here: Creating and Deploying a Model.
6. When you are finished with your workspace, your Admin can remove it, as described here: Removing ML Workspaces.
Test your connectivity to the CDP-DC cluster

Test that you can create a Project in your ML Workspace and access data that is stored in the data center cluster.

**Procedure**

1. Create a new Project, using the PySpark template.
2. Create a new file called testdata.txt (use this exact filename).
3. Add 2-3 lines of any text in the file to serve as sample data.
4. Run the following Spark commands to test the connection.

```python
from pyspark.sql import SparkSession

# Instantiate Spark-on-K8s Cluster
spark = SparkSession(
    builder
    .appName("Simple Spark Test")
    .config("spark.executor.memory","8g")
    .config("spark.executor.cores","2")
    .config("spark.driver.memory","2g")
    .config("spark.executor.instances","2")
    .getOrCreate()

# Validate Spark Connectivity
spark.sql("SHOW databases").show()
spark.sql('create table testcml (abc integer)').show()
spark.sql('insert into table testcml select t.* from (select 1) t').show()
spark.sql('select * from testcml').show()
# Stop Spark Session
spark.stop()
```

5. Run the following direct HDFS commands to test the connection.

```bash
# Run sample HDFS commands
# Requires an additional testdata.txt file to be created with sample data
in project home dir
!hdfs dfs -mkdir /tmp/testcml/
!hdfs dfs -copyFromLocal /home/cdsw/testdata.txt /tmp/testcml/
!hdfs dfs -cat /tmp/testcml/testdata.txt
```

**What to do next**

If you get errors, then check with your Admin to make sure that your user ID is set up in the Hadoop Authentication settings to access the CDP-DC cluster, and that the correct Ranger permissions have been applied.

### Differences Between Public and Private Cloud

There are some differences in Cloudera Machine Learning functionality between Public and Private Cloud.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Public Cloud</th>
<th>Private Cloud 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>CML application control plane (infrastructure containers and workload containers)</td>
<td>Control plane is hosted on public cloud servers.</td>
<td>Control plane is hosted on customer’s Openshift Container Platform.</td>
</tr>
<tr>
<td>Storage - CML internal state data (database, images, logs)</td>
<td>EBS on AWS, Azure Disks on Azure.</td>
<td>Software Defined Storage System, such as Ceph or Portworx.</td>
</tr>
<tr>
<td>Storage - User project files</td>
<td>EFS on AWS, external NFS on Azure.</td>
<td>Internal NFS storage is recommended.</td>
</tr>
<tr>
<td>Feature</td>
<td>Public Cloud</td>
<td>Private Cloud 1.0</td>
</tr>
<tr>
<td>--------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Autoscaling</td>
<td>CPU/GPU nodes scale up and down as needed.</td>
<td>Autoscaling concept is different; Private Cloud shares a pooled set of resources among workloads.</td>
</tr>
<tr>
<td>Logging</td>
<td>Per-workspace diagnostic bundles can be downloaded from the workspace.</td>
<td>Diagnostic bundles are not supported at Workspace level, but can be downloaded from the control plane at the cluster level.</td>
</tr>
<tr>
<td>Monitoring dashboards</td>
<td>Provides four dashboards.</td>
<td>Provides two dashboards, for K8s Container and K8s Cluster.</td>
</tr>
<tr>
<td>NFS support</td>
<td>AWS uses EFS; Azure requires external NFS.</td>
<td>Internal NFS is recommended, external NFS is supported.</td>
</tr>
<tr>
<td>TLS support</td>
<td>TLS access to workspaces is supported.</td>
<td>TLS access is supported, but requires manual setup of certificate and other steps.</td>
</tr>
<tr>
<td>Hadoop Authentication</td>
<td>Uses FreeIPA</td>
<td>User needs to provide credentials to communicate with the CDP Private Base cluster.</td>
</tr>
<tr>
<td>Remote Access</td>
<td>Available from each workspace.</td>
<td>Not available in the workspace. Instead, the environment's kubeconfig file may be downloaded from Environments using the Download Kubernetes configuration action for the specified environment.</td>
</tr>
<tr>
<td>Roles</td>
<td>MLAdmin, MLUser</td>
<td>The corresponding roles are: EnvironmentAdmin, EnvironmentUser</td>
</tr>
</tbody>
</table>

**Limitations on Private Cloud**

There are some limitations to keep in mind when you are working with Cloudera Machine Learning on Private Cloud. The following features are not yet supported in CML Private Cloud:

- Logging is limited, and diagnostic bundles for each workspace cannot be downloaded from the workspace UI. Instead, diagnostic bundles for the entire cluster can be downloaded from the control plane.
- Monitoring on Private Cloud does not support node-level resource metrics, hence only K8s Cluster and K8s Container dashboards are available.
- ML Runtimes are not supported.

**Network File System**

Network File System (NFS) is a protocol to access storage on a network that emulates accessing storage in a local file system. CML requires an NFS server for storing project files and folders, and the NFS export must be configured before you provision the first CML workspace in the cluster.

There are many different products or packages that can create an NFS in your private network. A Kubernetes cluster can host an internal NFS server, or an external NFS server can be installed on another cluster that is accessible by the private cloud cluster nodes. NFS storage is used only for storing project files and folders, and not for any other CML data, such as PostgreSQL database or livelog files.

CML does not support shared volumes, such as Portworx shared volumes, for storing project files. A read-write-once (RWO) persistent volume must be allocated to the internal NFS server (for example, NFS server provisioner) as the persistence layer. The NFS server uses the volume to dynamically provision read-write-many (RWX) NFS volumes for the CML clients.

**Note:** The internal NFS server option is currently the recommended option for Private Cloud. There are several options for setting up an internal NFS, described in the appendix. The Private Cloud Admin is responsible for setting up an NFS for use by your cluster.
### NFS Options for Private Cloud

The recommended approach is an internal NFS server which is deployed into the cluster. Solutions include NFS over Ceph or Portworx using NFS Server Provisioner (NFS Ganesha). The storage space for each workspace is transparently managed by the internal NFS server.

An alternative is to use an NFS server that is external to the cluster, such as a NetApp Filer appliance. In this case, you must manually create a directory for each workspace.

The NFS server must be configured before deploying the first CML workspace in the cluster. One important limitation is that CML does not support using shared volumes for storing project files. A read-write-once (RWO) persistent volume must be allocated to the internal NFS server (e.g., NFS server provisioner) as the persistence layer. The NFS server uses the volume to dynamically provision read-write-many (RWX) NFS volumes for the CML clients.

#### Deploying NFS Server Provisioner on Rook Ceph

As an example, you can deploy the NFS Server Provisioner using the Helm chart provided here.

- For the `nfs` storage class, set “allowVolumeExpansion=true”
- For the underlying persistent volume, specify a size of 1 TiB.
- On the block storage system class, `rook-ceph-block` in this case, set `allowVolumeExpansion=true`
- Download two `yaml` files here: `nfs-server-provisioner.yml` and `nfs-scc.yml`.

**Note:** If your read-write-once block pv storage class is not `rook-ceph-block`, then replace `rook-ceph-block` with your preferred read-write-once block pv storage class in the `nfs-server-provisioner.yml` file below.

**Note:** In this example, the size of the underlying persistent volume is specified as 1 TiB. You are recommended to increase the size by 1 TiB for each additional ML workspace that will be created. For example, increase the size by 5 TiB for 5 ML workspaces.

**Note:** The `yaml` file below has a line that will download the nfs server image from `quay.io/kubernetes_incubator/nfs-provisioner:v2.3.0`. If your installation is airgapped, download the file to a local repository and change the path accordingly.

1. **Install Path 1:** Installing using the `oc` command and `yaml` files:
   - If you do not have Tillerless Helm v2 set up, then you can simply apply the `nfs-server-provisioner.yml` file as follows:
     ```
     $ oc create -f nfs-server-provisioner.yml -n cml-nfs
     ```

2. **Install Path 2:** Installing using the `oc` command and Tillerless Helm v2:
   ```
   $ oc delete scc nfs-scc
   $ oc delete clusterrole cml-nfs-nfs-server-provisioner
   $ oc delete clusterrolebinding cml-nfs-nfs-server-provisioner
   $ oc delete namespace cml-nfs
   $ helm tiller run cml-nfs -- helm delete cml-nfs --purge
   $ oc delete scc nfs-scc securitycontextconstraints.security.openshift.io "nfs-scc" deleted
   ```

### Using an External NFS Server

As an alternative, you can install an NFS server that is external to the cluster. This is not the recommended approach. Currently, CML only works with NFS version 3 and 4.1. The NFS client within CML must be able to mount the NFS storage with default options, and also assumes these export options:

- `rw,sync,no_root_squash,no_all_squash,no_subtree_check`
The `no_root_squash` option has security implications, which is a strong reason to choose internal NFS instead.

Before creating a CML workspace, the storage administrator must create a directory that will be exported to the cluster for storing ML project files for that workspace. Either a dedicated NFS export path, or a subdirectory in an existing export must be specified for each workspace.

Each CML workspace needs a unique directory that does not have files in it from a different or previous workspace. For example, if 10 CML workspaces are expected, the storage administrator will need to create 10 unique directories. Either one NFS export and 10 subdirectories within it need to be created, or 10 unique exports need to be created.

For example, to use a dedicated NFS share for a workspace named “workspace1” from NFS server “nfs_server”, do the following:

1. Create NFS export directory “/workspace1”.
2. Change ownership for the exported directory
   a. CML accesses this directory as a user with a UID and GID of 8536. So run chown 8536:8536 /workspace1
   b. Make the export directory group-writeable and set GID: chmod g+srwx /workspace1
3. Provide the NFS export path nfs_server:/workspace1 when prompted by the CML Control Plane App while creating the workspace.
4. To use a subdirectory in an existing NFS share, say nfs_server:/export, do the following:
   a. Create a subdirectory /export/workspace1
   b. Change ownership: chown 8536:8536 /export/workspace1
   c. Set GID and make directory group writeable: chmod g+srwx /export/workspace1
   d. Provide the export path nfs_server:/export/workspace1 when prompted by the CML Control Plane App.

**Backing up Project Files and Folders**

The block device backing the NFS server data must be backed up to protect the CML project files and folders. The backup mechanism would vary depending on the underlying block storage system and backup policies in place.

1. Identify the underlying block storage to backup, first determine the NFS PV:

   ```bash
   $ echo `kubectl get pvc -n cml-nfs -o jsonpath='{.items[0].spec.volumeName}'`
   pvc-bec1de27-753d-11ea-a287-4cd98f578292
   ```

2. For Rook Ceph, the RBD volume/image name is the name of the dynamically created persistent volume (pvc-3d3d16b6-6cc7-11ea-828e-1418774847a1).

   Ensure this volume is backed up using an appropriate backup policy.

**Uninstalling the NFS server**

Uninstall the NFS server provisioner using either of the following commands.

Use this command if the NFS server provisioner was installed using `oc` and `yaml` files:

```bash
$ oc delete scc nfs-scc
$ oc delete clusterrole cml-nfs-nfs-server-provisioner
$ oc delete clusterrolebinding cml-nfs-nfs-server-provisioner
$ oc delete namespace cml-nfs
```

Use this command if the NFS server provisioner was installed using Helm:

```bash
$ helm tiller run cml-nfs -- helm delete cml-nfs --purge
$ oc delete scc nfs-scc securitycontextconstraints.security.openshift.io "nfs-scc" deleted
```
Deploy an ML Workspace with Support for TLS

You can provision an ML workspace with TLS enabled, so that it can be accessed via https.

**Before you begin**

You need to obtain a certificate from the Certificate Authority used by your organization. This may be an internal certificate authority.

Additionally, you need a computer with CLI access to the cluster, and with **kubectl** installed.

**Procedure**

1. Provision the ML Workspace. Follow the procedure Provisioning ML Workspaces.
   
   **Note:** Ensure you select **Enable TLS**.

2. Obtain the `.crt` and `.key` files for the certificate from your Certificate Authority.

   The certificate URL is generally of the form:
   
   `<workspaceid>.<apps.openshiftcluster>.<domain>.com`. Assuming an example URL for the certificate of `ml-30b43418-53c.apps.os4cluster.yourcompany.com`, check that the certificate correctly shows the corresponding Common Name (CN) and Subject Alternative Names (SAN):
   
   - CN: `ml-30b43418-53c.apps.os4cluster.yourcompany.com`
   - SAN: `*.ml-30b43418-53c.apps.os4cluster.yourcompany.com`
   - SAN: `ml-30b43418-53c.apps.os4cluster.yourcompany.com`

3. Create a Kubernetes secret inside the previously provisioned ML workspace namespace, and name the secret `cml-tls-secret`.

   On a machine with access to the `.crt` and `.key` files above, and access to the OpenShift cluster, run this command: `kubectl create secret tls cml-tls-secret --cert=<pathtocrt.crt> --key=<pathtokey.key> -o yaml --dry-run | kubectl -n <cml-workspace-namespace> create -f -`

   You can replace or update certificates in the secret at any time.

4. In **Admin > Security > Root CA configuration**, add the root CA certificate to the workspace.

   For example: `https://ml-def88113-acd.apps.nf-01.os4cluster.yourcompany.com/administration/security`

**Results**

The command creates routes to reflect the new state of ingress and secret, and enables TLS.

**Replace a Certificate**

You can replace a certificate in a deployed namespace.

**About this task**

**Procedure**

1. Obtain the new certificate `.crt` and `.key` files.

2. Run this command (example): `kubectl create secret tls cml-tls-secret --cert=<pathtocrt.crt> --key=<pathtokey.key> -o yaml --dry-run | kubectl -n <cml-workspace-namespace> replace -f -`
What to do next
The certificate of an existing session does not get renewed. The new certificate only applies to newly created sessions.

How To

you can learn about the various features and functions of Cloudera Machine Learning in the following sections.

Provision an ML Workspace

In CML on Private Cloud, the ML Workspace is where data scientists get their work done. After your Admin has created or given you access to an environment, you can set up a workspace. Only one workspace can be created per environment.

Before you begin
The first user to access the ML workspace after it is created must have the EnvironmentAdmin role assigned.

Procedure

1. Log in to the CDP Private Cloud web interface using your corporate credentials or other credentials that you received from your CDP administrator.
2. Click ML Workspaces.
3. Click Provision Workspace. The Provision Workspace panel displays.
4. In Provision Workspace, fill out the following fields.
   a) Workspace Name - Give the ML workspace a name. For example, test-cml.
   b) Select Environment - From the dropdown, select the environment where the ML workspace must be provisioned. If you do not have any environments available to you in the dropdown, contact your CDP admin to gain access.
   c) Namespace - Enter the namespace to use for the ML workspace.
   d) NFS Server - Select Internal to use an NFS server that is integrated into the Kubernetes cluster. This is the recommended selection at this time.
      The path to the internal NFS server is already set in the environment.
5. In Production Machine Learning, select to enable the following features.
   a) Enable Governance - Enables advanced lineage and governance features.
      Governance Principal Name - If Enable Governance is selected, you can use the default value of mlgov, or enter an alternative name. The alternative name must be present in your environment and be given permissions in Ranger to allow the MLGovernance service deliver events to Atlas.
   b) Enable Model Metrics - Enables exporting metrics for models to a PostgreSQL database.
6. In Other Settings, select to enable the following features.
   a) Enable TLS - Select this to enable https access to the workspace.
   b) Enable Monitoring - Administrators (users with the EnvironmentAdmin role) can use a Grafana dashboard to monitor resource usage in the provisioned workspace.
7. Click Provision Workspace. The new workspace provisioning process takes several minutes.

What to do next
After the workspace is provisioned, you can log in by clicking the workspace name on the Machine Learning Workspaces page. The first user to log in must be the administrator.

Related Information
Monitoring ML Workspaces
Monitoring ML Workspaces

This topic shows you how to monitor resource usage on your ML workspaces.

About this task
Cloudera Machine Learning leverages Prometheus and Grafana to provide a dashboard that allows you to monitor how CPU, memory, storage, and other resources are being consumed by ML workspaces. Prometheus is an internal data source that is auto-populated with resource consumption data for each workspace. Grafana is a monitoring dashboard that allows you to create visualizations for resource consumption data from Prometheus.

Each ML workspace has its own Grafana dashboard.

Before you begin
Required Role: MLAdmin
Without the MLAdmin role, you will not be able to view the Workspace details page.

Note: On Private Cloud, the corresponding role is EnvironmentAdmin.

Procedure
1. Log in to the CDP web interface.
2. Click ML Workspaces.
3. For the workspace you want to monitor, click Actions > Open Grafana.

CML provides you with CML Monitoring, a default Grafana dashboard which includes panels on CPU usage, memory usage, running processes, autoscaling, and network I/O on the workspace. You might choose to extend this dashboard or create more panels for other metrics. For more information, see the Grafana documentation.

Removing ML Workspaces

This topic describes how to remove an existing ML workspace and clean up any cloud resources associated with the workspace. Currently, only CDP users with both the MLAdmin role and the EnvironmentAdmin account role can remove workspaces.

Procedure
1. Log in to the CDP web interface.
2. Click ML Workspaces.
3. Click on the Actions icon and select Remove Workspace.
   a) Remove EFS Storage - This option is enabled by default. If you want to retain project files on EFS, disable this property.
   b) Force Delete - This property is not required by default. You should first attempt to remove your workspace with this property disabled.

Enabling this property will delete the workspace from CDP but does not guarantee that the underlying cloud resources used by the workspace will be cleaned up properly. Go to your cloud service provider account to make sure that the cloud resources have been successfully deleted.
4. Click OK to confirm.

Note: On Azure public cloud, you also need to delete NFS storage after removing the workspace, if the NFS service is no longer needed.

Managing your Personal Account

You can edit personal account settings such as email, SSH keys and Hadoop credentials.

About this task
You can also access your personal account settings by clicking Account settings in the upper right-hand corner drop-down menu. This option will always take you to your personal settings page, irrespective of the context you are currently in.

Procedure
2. From the upper right drop-down menu, switch context to your personal account.
3. Click Settings.
   
   Profile
   You can modify your name, email, and bio on this page.

   Teams
   This page lists the teams you are a part of and the role assigned to you for each team.

   SSH Keys
   Your public SSH key resides here. SSH keys provide a useful way to access to external resources such as databases or remote Git repositories. For instructions, see https://docs.cloudera.com/machine-learning/cloud/security/topics/ml-ssh-keys.html.

Creating a Team

Users who work together on more than one project and want to facilitate collaboration can create a Team. Teams allow streamlined administration of projects.

About this task
Team projects are owned by the team, rather than an individual user. Team administrators can add or remove members at any time, assigning each member different permissions.
How To

Procedure

1. Click the plus sign (+) in the title bar, to the right of the Search field.
2. Select Create Team.
3. Enter a Team Name.
4. Click Create Team.
5. Add or invite team members. Team members can have one of the following privilege levels:
   • Viewer - Read-only access to team projects. Cannot create new projects within the team but can be added to existing ones.
   • Operator - Read-only access to team projects. Additionally, Operators can start and stop existing jobs in the projects that they have access to.
   • Contributor - Write-level access to all team projects to all team projects with Team or Public visibility. Can create new projects within the team. They can also be added to existing team projects.
   • Admin - Has complete access to all team projects, can add new team members, and modify team account information.
6. Click Done.

Managing a Team Account

Team administrators can modify account information, add or invite new team members, and view/edit privileges of existing members.

Procedure

1. From the upper right drop-down menu, switch context to the team account.
2. Click Settings to open up the Account Settings dashboard.
3. Modify any of the following settings:
   
   **Profile**
   
   Modify the team description on this page.
   
   **Members**
   
   You can add new team members on this page, and modify privilege levels for existing members.
   
   **SSH Keys**
   
   The team's public SSH key resides here. Team SSH keys provide a useful way to give an entire team access to external resources such as databases. For instructions, see [https://docs.cloudera.com/machine-learning/cloud/security/topics/ml-ssh-keys.html](https://docs.cloudera.com/machine-learning/cloud/security/topics/ml-ssh-keys.html). Generally, team SSH keys should not be used to authenticate against Git repositories. Use your personal key instead.

### Projects in Cloudera Machine Learning

Projects form the heart of Cloudera Machine Learning. They hold all the code, configuration, and libraries needed to reproducibly run analyses. Each project is independent, ensuring users can work freely without interfering with one another or breaking existing workloads.

Access the **Projects** page by clicking **Projects** in the navigation panel. The **Projects** page gives you a quick summary of project information.

- **Active Workloads** - If there are active workloads running, this section describes the number of Sessions, Experiments, Models, Jobs, and Applications that are running.
- **Resource Usage Details** - A collapsible section that displays resource usage.
  - **Active Workloads** - If there are active workloads running, this section describes the number of Sessions, Experiments, Models, Jobs, and Applications that are running.
  - **User Resources and Workspace Resources**
    - Click on the **User Resources** tab to see the CPU and memory resource usage for the user. The maximum usage of the vCPU and GB is calculated based on whether or not you have a quota. If you have a quota, the maximum usage will be based on your quota. If you don't have a quota, the maximum usage will be what is available on the cluster. If you have a GPU, you'll also see the GPU usage.
    - Click on the **Workspace Resources** tab to see usage overall.
- **Search Projects** - Enter a term for keyword search across Project names.
- **Scope** - An additional filter only viewable by Administrators.
  - Selecting **My Projects** displays only the Projects that you have created or are a Collaborator of.
  - Selecting **All Projects** displays all Projects on the ML Workspace.
- **Creator** - An additional filter to only display Projects created by a specified user.
- **Projects View Selector** - A setting that enables you to display Projects in a summary card-based view or a detailed table-based view.
The following topics describe how to create and manage projects in Cloudera Machine Learning.

Creating a Project with Legacy Engine Variants
Projects create an independent working environment to hold your code, configuration, and libraries for your analysis. This topic describes how to create a project with Legacy Engine variants in Cloudera Machine Learning.

Procedure
1. Go to Cloudera Machine Learning and on the left sidebar, click Projects.
2. Click New Project.
3. If you are a member of a team, from the drop-down menu, select the Account under which you want to create this project. If there is only one account on the deployment, you will not see this option.
4. Enter a Project Name.
5. Select Project Visibility from one of the following options.
   - Private - Only project collaborators can view or edit the project.
   - Team - If the project is created under a team account, all members of the team can view the project. Only explicitly-added collaborators can edit the project.
   - Public - All authenticated users of Cloudera Machine Learning will be able to view the project. Collaborators will be able to edit the project.
6. Under Initial Setup, you can either create a blank project, or select one of the following sources for your project files.
   - Built-in Templates - Template projects contain example code that can help you get started with Cloudera Machine Learning. They are available in R, Python, PySpark, and Scala. Using a template project is not required, but it helps you start using Cloudera Machine Learning right away.
   - Custom Templates - Site administrators can add template projects that are customized for their organization's use-cases. For details, see Custom Template Projects.
   - Local - If you have an existing project on your local disk, use this option to upload compressed files or folders to Cloudera Machine Learning.
   - Git - If you already use Git for version control and collaboration, you can continue to do so with Cloudera Machine Learning. Specifying a Git URL will clone the project into Cloudera Machine Learning. If you use a Git SSH URL, your personal private SSH key will be used to clone the repository. This is the recommended approach. However, you must add the public SSH key from your personal Cloudera Machine Learning account to the remote Git hosting service before you can clone the project.
7. Click Create Project. After the project is created, you can see your project files and the list of jobs defined in your project.
   Note that as part of the project filesystem, Cloudera Machine Learning also creates the following .gitignore file.

   ```
   R
   node_modules
   *.pyc
   *
   .gitignore
   ```
8. (Optional) To work with team members on a project, add them as collaborators to the project.

Related Information
Custom Template Projects

Creating a Project with ML Runtimes Variants
Projects create an independent working environment to hold your code, configuration, and libraries for your analysis. This topic describes how to create a project with ML Runtimes variants in Cloudera Machine Learning.
**Procedure**

1. Go to Cloudera Machine Learning and on the left sidebar, click **Projects**.
2. Click **New Project**.
3. If you are a member of a team, from the drop-down menu, select the Account under which you want to create this project. If there is only one account on the deployment, you will not see this option.
4. Enter a **Project Name**.
5. Select Project Visibility from one of the following options.
   - Private - Only project collaborators can view or edit the project.
   - Team - If the project is created under a team account, all members of the team can view the project. Only explicitly-added collaborators can edit the project.
   - Public - All authenticated users of Cloudera Machine Learning will be able to view the project. Collaborators will be able to edit the project.
6. Under Initial Setup, you can either create a blank project, or select one of the following sources for your project files.
   - Blank - The project will contain no information from a template, local file, or Git.
   - Templates - Template projects contain example code that can help you get started with Cloudera Machine Learning. They are available in R, Python, PySpark, and Scala. Using a template project is not required, but it helps you start using Cloudera Machine Learning right away.
     
     **Note:** Templates are designed for Engines and might not work with ML Runtimes.

   - Local - If you have an existing project on your local disk, use this option to upload compressed files or folders to Cloudera Machine Learning.
   - Git - If you already use Git for version control and collaboration, you can continue to do so with Cloudera Machine Learning. Specifying a Git URL will clone the project into Cloudera Machine Learning. If you use a Git SSH URL, your personal private SSH key will be used to clone the repository. This is the recommended approach. However, you must add the public SSH key from your personal Cloudera Machine Learning account to the remote Git hosting service before you can clone the project.

7. Click **Create Project**. After the project is created, you can see your project files and the list of jobs defined in your project.

   Note that as part of the project filesystem, Cloudera Machine Learning also creates the following `.gitignore` file.

   ```
   R
   node_modules
   *.pyc
   *
   !.gitignore
   ```
8. Set or verify the ML Runtimes settings for the project.
   Within the selected project, you can modify the default engine configuration:
   a) In the left navigation bar, click **Project Settings**.
   b) Select the **Runtime/Engine** tab.
   c) Next to **Default Engine**, select **ML Runtime**.
   d) Click **Save Engine**.

   **Note:** The ability to switch the default engine is intended to support migration of projects from Engines to Runtimes. Cloudera does not recommend making this change frequently.

   Once switched to ML Runtimes, all modules (Sessions, Jobs, Models, etc.) for the project will configure using ML Runtimes instead of Legacy Engines. Setting Engine parameters for the project will no longer be possible.

   Existing and running instances (for example, Jobs, Models or Applications) previously configured with a particular Engine configuration will keep their configuration until you change the related settings. If you want to change the engine configuration for existing and running instances, you will need to update those based on the new, Runtime-based settings.

   **Note:** Note that site administrators now have the ability to select ML Runtimes as a default for newly created projects. This setting can be found within the Admin menu below the Runtime/Engine tab.

### Adding Project Collaborators

This topic shows you how to invite colleagues to collaborate on a project.

**About this task**

For a project created under your personal account, anyone who belongs to your organization can be added as a collaborator. For a project created under a team account, you can only add collaborators that already belong to the team. If you want to work on a project that requires collaborators from different teams, create a new team with the required members, and then create a project under that account. If your project was created from a Git repository, each collaborator must create the project from the same central Git repository.

You can grant project collaborators one of three levels of access:

- **Viewer** - Read-only access to code, data, and results.
- **Operator** - Read-only access to code, data, and results. Additionally, Operators can start and stop existing jobs in the projects that they have access to.
- **Contributor** - Can view, edit, create, and delete files and environmental variables, run sessions/experiments/jobs/models and run code in running jobs. Additionally, Contributors can set the default engine for the project.
- **Admin** - Has complete access to all aspects of the project. This includes the ability to add new collaborators, and delete the entire project.
Collaborating Securely on Projects

Before adding project collaborators, you must remember that assigning the Contributor or Admin role to a project collaborator is the same as giving them write access to your data in CDH. This is because project contributors and project administrators have write access to all your project code (including any library code that you might not be actively inspecting). For example, a contributor/admin could modify project file(s) to insert code that deletes some data on the CDH cluster. The next time you launch a session and run the same code, it will appear as though you deleted the data yourself.

Additionally, project collaborators also have access to all actively running sessions and jobs. This means that a malicious user can easily impersonate you by accessing one of your active sessions. Therefore, it is extremely important to restrict project access to trusted collaborators only. Note that site administrators can restrict this ability by allowing only session creators to run commands within their own active sessions.

For these reasons, Cloudera recommends using Git to collaborate securely on shared projects. This will also help avoid file modification conflicts when your team is working on more elaborate projects.

Procedure

1. In Cloudera Machine Learning, navigate to the project overview page.
2. Click Team to open the Collaborators page.
3. Search for collaborators by either name or email address and click Add.

Modifying Project Settings

Project contributors and administrators can modify aspects of the project environment such as the engine or ML Runtimes used to launch sessions, the environment variables, and to create SSH tunnels to access external resources.

Procedure

1. Switch context to the account where the project was created.
2. Click Projects.
3. From the list of projects, select the one to modify.
4. Click Project Settings to open the Project Settings dashboard.

   Options

   Modify the project name and its privacy settings on this page.

   Runtime/Engine

   Cloudera Machine Learning ensures that your code is always run with the specific engine version you selected. You can select to use either Legacy Engines or Machine Learning Runtimes. For legacy engines only, you can also select the engine version and add third-party editors here.

   Advanced

   • Environment Variables - If there are any environmental variables that should be injected into all the engines running this project, you can add them to this page. For more details, see Engine Environment Variables.
   • Shared Memory Limit - You can specify additional shared memory available to sessions running with the project.

   Tunnels

   In some environments, external databases and data sources reside behind restrictive firewalls. Cloudera Machine Learning provides a convenient way to connect to such resources using your SSH key. For instructions, see SSH Keys.

   Delete Project

   This page can only be accessed by project administrators. Remember that deleting a project is irreversible. All files, data, sessions, and jobs are removed.

Related Information

Managing Engines
Engine Environment Variables
SSH Keys

Managing Project Files

Cloudera Machine Learning allows you to move, rename, copy, and delete files within the scope of the project where they live. You can also upload new files to a project, or download project files. For use cases beyond simple projects, Cloudera strongly recommends using Git for Collaboration to manage your projects using version control.

Procedure

1. Switch context to the account where the project was created.
2. Click Projects.
3. From the list of projects, click on the project you want to modify. This will take you to the project overview.
4. Click Files.

Upload Files to a Project

Files can only be uploaded within the scope of a single project. Therefore, to access a script or data file from multiple projects, you will need to manually upload it to all the relevant projects.

Click Upload. Select Files or Folder from the dropdown, and choose the files or folder you want to upload from your local filesystem.

In addition to uploading files or a folder, you can upload a .tar file of multiple files and folders. After you select and upload the .tar file, you can use a terminal session to extract the contents:

   a. On the project overview page, click Open Workbench and select a running session or create a new one.
   b. Click Terminal access.
   c. In the terminal window, extract the contents of the .tar file:

   ```
   tar -xvf <file_name>.tar.gz
   ```

The extracted files are now available for the project.

Download Project Files

Click Download to download the entire project in a .zip file. To download only a specific file, select the checkbox next to the file(s) to be download and click Download.

5. You can also use the checkboxes to Move, Rename, or Delete files within the scope of this project.

Related Information

Git for Collaboration

Custom Template Projects

Site administrators can add template projects that have been customized for their organization's use-cases. These custom project templates can be added in the form of a Git repository.

Required Role: See User Role Authorization.

To add a new template project, go to Admin > Settings. Under the Project Templates section, provide a template name, the URL to the project's Git repository, and click Add.

The added templates will become available in the Template tab on the Create Project page. Site administrators can add, edit, or delete custom templates, but not the built-in ones. However, individual built-in templates can be disabled using a checkbox in the Project Templates table at Admin > Settings.

Deleting a Project

This topic demonstrates how to delete a project.

About this task

**Important:** Deleting a project is an irreversible action. All files, data, and history related to the project will be lost. This includes any jobs, sessions or models you created within the project.

Procedure

1. Go to the project Overview page.
2. On the left sidebar, click Settings.
3. Go to the Delete Project.
4. Click Delete Project and click OK to confirm.
Native Workbench Console and Editor

The workbench console provides an interactive environment tailored for data science, supporting R, Python and Scala. It currently supports R, Python, and Scala engines. You can use these engines in isolation, as you would on your laptop, or connect to your CDH cluster.

The workbench UI includes four primary components:

- An editor where you can edit your scripts.
- A console where you can track the results of your analysis.
- A command prompt where you can enter commands interactively.
- A terminal where you can use a Bash shell.

Typically, you would use the following steps to run a project in the workbench:

**Related Information**

**Managing Engines**

**Launch a Session**

Sessions allow you to perform actions such as run R or Python code. They also provide access to an interactive command prompt and terminal. This topic demonstrates how to launch a new session.

**Procedure**

1. Navigate to your project's Overview page.
2. Click *New Session*.

The information presented on this page will depend on which default engine you have chosen for your project: ML Runtimes or Legacy Engines. You can change the default engine later in this task.
3. Check the settings for your session:
   If your project is using ML Runtimes, you will see the following settings:
   **Editor**
   Selects the Editor; currently only Workbench is supported and therefore the selector is static.
   **Kernel**
   Selects the Kernel. Initially only Python Runtimes are supported.
   **Edition**
   Selects the Runtime Edition. Initially only Standard variants are supported
   **Version**
   Selects the Version for the selected Kernel. Initially version 3.6, 3.7 and 3.8 Python versions are supported.
   **Note:** The selector options only consider the configurations supported by the actual deployments and certain selections will automatically limit others. For example, certain versions are only relevant for Python or certain editors are supported only with certain kernels.

If your project is using Legacy Engines, you will see the following settings:
**Editor**
Selects the Editor; currently only Workbench is supported and therefore the selector is static.
**Kernel**
Selects the Kernel. Initially only Python Runtimes are supported.
**Engine Image**
Displays the Advanced tab in Project Settings and allows you to set environment variables and the shared memory limit.

4. If your project is using Legacy Engines, you can modify the engine image used by this session:
   a) By **Engine Image**, click **Configure**.
      Cloudera Machine Learning displays the Project Settings page.
   b) Select the **Runtime/Engine** tab.
   c) Next to **Default Engine**, select **ML Runtime** or **Legacy Engine**.
   d) Click **Save Engine**.

5. Specify your Resource Profile.
   The minimum configuration is 1vCPU and 2 GB memory.

6. Click **Start 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.

**Run Code**
This topic shows you how to enter and run code in the interactive Workbench command prompt or the editor after you launch a session.
The editor is best for code you want to keep, while the command prompt is best for quick interactive exploration.
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.
If you created your project from a template, you should see project files in the editor. You can open a file in the editor by clicking the file name in the file navigation bar on the left.

Editor - To run code from the editor:

1. Select a script from the project files on the left sidebar.
2. 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).

When doing real analysis, writing and executing your code from the editor rather than the command prompt makes it easy to iteratively develop your code and save it along the way.

If you require more space for your editor, you can collapse the file list by double-clicking between the file list pane and the editor pane. You can hide the editor using editor's View menu.

**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.

**Project Code Files**

All project files are stored to persistent storage within the respective project directory at /var/lib/cdsw/current/projects. They can be accessed within the project just as you would in a typical directory structure. For example, you can import functions from one file to another within the same project.

**Access the Terminal**

Cloudera Machine Learning provides full terminal access to running engines from the web console. This topic show you how to access the Terminal from a running Workbench session.

You can use the terminal to move files around, run Git commands, access the YARN and Hadoop CLIs, or install libraries that cannot be installed directly from the engine. To access the Terminal from a running session, click Terminal Access above the session log pane.
The terminal's default working directory is `/home/cdsw`, which is where all your project files are stored. Any modifications you make to this folder will persist across runs, while modifications to other folders are discarded.

If you are using Kerberos authentication, you can run `klist` to see your Kerberos principal. If you run `hdfs dfs -ls` you will see the files stored in your HDFS home directory.

Note that the terminal does not provide root or sudo access to the container. To install packages that require root access, see `Customized Engine Images`.

### Related Information
- `Customized Engine Images`

---

**Stop a Session**

This topic demonstrates how to stop a session to free up resources for other users when you are finished.

When you are done with the session, click **Stop** in the menu bar above the console, or use code to exit by typing the following command:

**R**

```r
quit()
```

**Python**

```python
exit
```

**Scala**

```scala
quit()
```

Sessions automatically stop after an hour of inactivity.

---

**Third-Party Editors**

In addition to the built-in Cloudera Machine Learning editor, you can configure Cloudera Machine Learning to work with third-party, browser-based IDEs such as Jupyter and also certain local IDEs that run on your machine, such as PyCharm.

**Note:** Custom editors run inside CML sessions. If the CML session is stopped, this may cause unexpected behavior in the editor UI and, in some cases, may result in data loss. You should, therefore, use the custom editor's UI to shut the editor down first. This will automatically end the CML session too.

In JupyterLab you do that by clicking "Shut Down" in the JupyterLab "File" menu." This applies to both engines and Runtimes, and all versions of CML.

When you bring your own editor, you still get many of the benefits Cloudera Machine Learning behind an editor interface you are familiar with:

- Dependency management that lets you share code with confidence
- CDH client configurations
- Automatic Kerberos authentication through Cloudera Machine Learning
- Reuse code in other Cloudera Machine Learning features such as experiments and jobs
- Collaboration features such as teams
- Compliance with IT rules for where compute, data, and/or code must reside. For example, compute occurs within the Cloudera Machine Learning deployment, not the local machine. Browser IDEs run within a Cloudera Machine Learning session and follow all the same compliance rules. Local IDEs, on the other hand, can bring data or code to a user's machine. Therefore, Site Administrators can opt to disable local IDEs to balance user productivity with compliance concerns.
In the Cloudera Machine Learning documentation, browser-based IDEs like Jupyter will be referred to as "browser IDEs". IDEs such as PyCharm that run on your machine outside of your browser will be referred to as "local IDEs" because they run on your local machine. You can use the browser or local IDE of your choice to edit and run code interactively.

Note that you can only edit and run code interactively with the IDEs. Tasks such as creating a project or deploying a model require the Cloudera Machine Learning web UI and cannot be completed through an editor.

Modes of Configuring Third-Party Editors

The configuration for an IDE depends on which type of editor you want to use.

In addition to the native Cloudera Machine Learning editor, you can configure Cloudera Machine Learning to work with third-party, browser-based IDEs, such as Jupyter, and also certain local IDEs that run on your machine, such as PyCharm.

Workbench editor

The Workbench editor is the built-in editor for Cloudera Machine Learning. No additional configuration is required to use it. When you launch a session, select the Workbench editor.

Third-party, browser-based IDEs

Browser IDEs are editors such as Jupyter or RStudio. When you use a browser IDE, it runs within a session and allows you to edit and run code interactively. Changes that you make in the editor are propagated to the Cloudera Machine Learning project. Base Engine Image v8 and higher ships with Jupyter preconfigured as a browser IDE. You can select it when you start a session or add a different browser IDE. For more information, see Configure a Browser IDE as an Editor.

Keep the following in mind when using browser IDEs:

- Engine Version Requirements
  - Browser-based IDEs require Base Engine Image v8 or higher.
  - When you are finished using a browser IDE, you must exit the IDE properly, including saving your work if necessary. Do not just stop the Cloudera Machine Learning session. Doing so will cause you to lose your session state. For example, if you want RStudio to save your state, including variables, to ~/.RData, exit the RStudio workspace using the power button in the top right of the RStudio UI.
  - Depending on the behavior of the browser IDE, multiple users within a project may overwrite each other’s state. For example, RStudio state is persisted in /home/cdsw/.RData that is shared by all users within a project.
  - Browser IDEs do not adhere to the timeout set in IDLE_MAXIMUM_MINUTES. Instead, they use the timeout set in SESSION_MAXIMUM_MINUTES, which is 7 days by default. Cloudera recommends that users stop their session manually after using a browser-based editor. Running sessions continue to consume resources and may impact other users.
  - Logs for browser IDEs are available on the Logs tab of the session window. This includes information that the IDE may generate, such as error messages, in addition to any Cloudera Machine Learning logs.

Local IDE Editors on your machine that can use SSH-based remote editing

These editors, referred to as Local IDEs in the documentation, are editors such as PyCharm that run on your local machine. They connect to Cloudera Machine Learning with an SSH endpoint and allow you to edit and run code interactively. You must manually configure some sort of file sync and ignore list between your local machine and Cloudera Machine Learning. You can use functionality within the local IDE, such as PyCharm's sync, or external tools that can sync via the SSH endpoint, such as Mutagen.

Keep the following in mind before setting up local IDEs:

- Local IDEs do not require a specific engine image, but Cloudera always recommends you use the latest engine image.
• Site Administrators should work with IT to determine the data access policies for your organization. For example, your data policy may not allow users to sync certain files to their machines from Cloudera Machine Learning. Verify that users understand the requirements and adhere to them when configuring their file sync behavior.

• Users should ensure that any IDEs that the IDEs they want to use support SSH. For example, VS Code supports "remote development over SSH," and PyCharm supports using a "remote interpreter over SSH."

**Related Information**
Configure a Browser IDE as an Editor
Configure a Local IDE using an SSH Gateway

**Configure a Browser IDE as an Editor**

When you use a browser IDE, changes that you make in the editor are propagated to the Cloudera Machine Learning project.

**About this task**

For example, if you create a new .py file or modify an existing one with the third-party editor, the changes are propagated to Cloudera Machine Learning. When you run the code from the IDE, execution is pushed from the IDE to Cloudera Machine Learning.

Base Engine Image v8 and higher for Cloudera Machine Learning comes preconfigured with Jupyter, and any browser IDEs you want to add must be added to Base Engine Image v8 or higher. Jupyter can be selected in place of the built-in Workbench editor when you launch a session, and no additional configuration is required. You can configure additional IDEs to be available from the dropdown.

You have two configuration options:

• **Project Level:** You can configure an editor at the project level so that any session launched within that project can use the editor configured. Other projects across the deployment will not be able to use any editors configured in such a manner. For steps, see Configure a Browser IDE at the Project Level.

• **Engine Level:** You can create a custom engine configured with the editor so that any project across the deployment that uses this custom engine can also use the editor configured. This might be the only option in case of certain browser IDEs (such as RStudio) that require root permission to install and therefore cannot be directly installed within the project. For steps, see Configure a Browser IDE at the Engine Level.

Cloudera recommends you first test the browser IDE you intend to install in a session before you install it to the project or build a custom engine with it. For steps, see Test a Browser IDE in a Session Before Installation.

**Test a Browser IDE in a Session Before Installation**

This process can be used to ensure that a browser IDE works as expected before you install it to a project or to a customized engine image. This process is not meant for browser IDEs that require root permission to install, such as RStudio.

**About this task**

These steps are only required if you want to use an editor that does not come preinstalled as part of the default engine image. Perform the following steps to configure an editor for your session:

**Procedure**

1. Ensure that your browser accepts pop-up windows and cookies from Cloudera Machine Learning web UI.
2. Open the Cloudera Machine Learning web UI.
3. Go to your project and launch a session with the kernel of your choice and the **Workbench** editor. Alternatively, open an existing session.
4. In the interactive command prompt or terminal for the session, install the editor you want to use. See the documentation for your editor for specific instructions.

   For example:

   **Jupyter Lab**

   **Python 3**

   The following example command installs Jupyter Lab for Python 3:

   ```bash
   !pip3 install jupyterlab
   ```

5. After the installation completes, enter the command to start the server for the notebook on the port specified in the `CDSW_APP_PORT` environment variable on IP address `127.0.0.1`.

   For example, the following command starts the server for Jupyter Lab on the port specified in the `CDSW_APP_PORT` environment variable:

   ```bash
   /home/cdsw/.local/bin/jupyter-lab --no-browser --ip=127.0.0.1 --port=${CDSW_APP_PORT} --NotebookApp.token= --NotebookApp.allow_remote_access=True --log-level=ERROR
   ```

6. Click on the grid icon in the top right. You should see the editor in the drop-down menu. If you select the editor, it opens in a new browser tab.

**Configure a Browser IDE at the Project Level**

The following steps are only required if you want to use an editor that does not come preinstalled as part of the default engine image that Cloudera Machine Learning ships with.

**Before you begin**

Before you start, verify that you have installed the IDE of your choice to the project. For information about how to install additional packages to a project, see *[Installing Additional Packages]*.

**About this task**

Perform the following steps to add an editor to a project:

**Procedure**

1. Open the Cloudera Machine Learning web UI.
2. Go to the project you want to configure an editor for.
3. Go to **Settings > Editors** and click **New Editor**.
4. Complete the fields:
   - **Name**: Provide a name for the editor. This is the name that appears in the dropdown menu for **Editors** when you start a new session.
   - **Command**: Enter the command to start the server for the editor on the Cloudera Machine Learning public port specified in the `CDSW_APP_PORT` environment variable (default 8081).

   For example, the following command starts Jupyter Lab on the port specified by the `CDSW_APP_PORT` environment variable:

   ```bash
   /home/cdsw/.local/bin/jupyter-lab --no-browser --ip=127.0.0.1 --port=${CDSW_APP_PORT} --NotebookApp.token= --NotebookApp.allow_remote_access=True --log-level=ERROR
   ```

   This is the same command you used to start the IDE to test it in a session.

5. Save the changes.

   When a user starts a new session, the editor you added is available in the list of editors. Browsers must be configured to accept cookies and allow pop-up windows from the Cloudera Machine Learning web UI.
**Related Information**

**Installing Additional Packages**

**Configure a Browser IDE at the Engine Level**

You can make a browser IDE available to any project within a Cloudera Machine Learning deployment by creating a customized engine image, installing the editor to it, and whitelisting it for a project. Additionally, browser IDEs that require root permission to install, such as RStudio, can only be used as part of a customized engine image.

**About this task**

When a user launches a session, they can select the customized engine with the editors available. The following steps describe how to make a customized engine image for RStudio:

**Procedure**

1. Create a Dockerfile for the new custom image. Note that the Base Engine Image uses Ubuntu, and you must use Base Engine Image v9 or higher.
   
   The following sample Dockerfile is for RStudio:
   
   ```
   #Dockerfile
   FROM docker.repository.cloudera.com/cdsw/engine:9-cml1.1
   WORKDIR /tmp

   # The RUN commands that install an editor
   # For example: RUN apt-get install myeditor
   RUN apt-get update && apt-get dist-upgrade -y && 
        apt-get install -y --no-install-recommends 
        libapparmor1 
        libclang-dev 
        lsb-release 
        psmisc 
        sudo

   # The command that follows RUN is the same command you used to install the IDE to test it in a the session.
   RUN wget https://download2.rstudio.org/server/trusty/amd64/rstudio-server-1.2.1335-amd64.deb && 
      dpkg -i rstudio-server-1.2.1335-amd64.deb

   COPY rserver.conf /etc/rstudio/rserver.conf
   COPY rstudio-cdsw /usr/local/bin/rstudio-cdsw
   RUN chmod +x /usr/local/bin/rstudio-cdsw
   ```

2. Create **rserver.conf**:

   ```
   # Must match CDSW_APP_PORT
   www-port=8090
   server-app-armor-enabled=0
   server-daemonize=0
   www-address=127.0.0.1
   auth-none=1
   auth-validate-users=0
   ```

   Make sure that the **www-port** property matches the port set in the **CDSW_APP_PORT** environment variable (default 8090).

3. Create **rstudio-cdsw**:

   ```
   #!/bin/bash
   ```
# This saves RStudio's user runtime information to /tmp, which ensures several
# RStudio sessions can run in the same project simultaneously
mkdir -p /tmp/rstudio/sessions/active
mkdir -p /home/cdsw/.rstudio/sessions
if [ -d /home/cdsw/.rstudio/sessions/active ]; then rm -rf /home/cdsw/.rstudio/sessions/active; fi
ln -s /tmp/rstudio/sessions/active /home/cdsw/.rstudio/sessions/active
# This ensures RStudio picks up the environment. This may not be necessary if
# you are installing RStudio Professional. See
# https://docs.rstudio.com/ide/server-pro/r-sessions.html#customizing-sessions-launches.
# SPARK_DIST_CLASSPATH is treated as a special case to workaround a bug in
# R
# with very long environment variables.
env | grep -v ^SPARK_DIST_CLASSPATH >> /usr/local/lib/R/etc/Renviron.site
echo "Sys.setenv("SPARK_DIST_CLASSPATH"="${SPARK_DIST_CLASSPATH}\")" >> /usr/local/lib/R/etc/Rprofile.site

# Now start RStudio
/usr/sbin/rstudio-server start

4. Build the Dockerfile:

docker build -t <image-name>:<tag> . -f Dockerfile

If you want to build your image on a Cloudera Machine Learning workspace, you must add the --network=host option to the build command:

docker build --network=host -t <image-name>:<tag> . -f Dockerfile

5. Distribute the image:

• Push the image to a public registry such as DockerHub.
  For instructions, refer the Docker documentation.
• Push the image to your company's Docker registry.
  When using this method, make sure to tag your image with the following schema:

  docker tag <image-name> <company-registry>/<user-name>/<image-name>:<tag>

Once the image has been tagged properly, use the following command to push the image:

docker push <company-registry>/<user-name>/<image-name>:<tag>

6. Whitelist the image in Cloudera Machine Learning:

  a) Log in to the Cloudera Machine Learning web UI as a site administrator.
  b) Click Admin > Engines.
  c) Add <company-registry>/<user-name>/<image-name>:<tag> to the list of whitelisted engine images.

7. Whitelist the new engine for a project:

  a) Go to the project Settings page.
  b) Click Engines.
  c) Select the new customized engine from the dropdown list of available Docker images. Sessions and jobs you run in your project will have access to this engine.
8. **Configure RStudio for the project.** When this is done, you will be able to select RStudio from the dropdown list of editors on the Launch New Session page.
   a) Go to Settings > Editors and click New Editor.
   b) Complete the fields:
      - Name: Provide a name for the editor. For example, RStudio. This is the name that appears in the dropdown menu for Editors when you start a new session.
      - Command: Enter the command to start the server for the editor. For example, the following command will start RStudio:
        ```bash
        /usr/local/bin/rstudio-cdsw
        ```
   c) Save the changes.

**Related Information**

**Docker push**

**Limitations**

**Configure a Local IDE using an SSH Gateway**

The specifics for how to configure a local IDE to work with Cloudera Machine Learning are dependent on the local IDE you want to use.

Cloudera Machine Learning relies on the SSH functionality of the editors to connect to the SSH endpoint on your local machine created with the `cdswctl` client. Users establish an SSH endpoint on their machine with the `cdswctl` client. This endpoint acts as the bridge that connects the editor on your machine and the Cloudera Machine Learning deployment.

The following steps are a high-level description of the steps a user must complete:

1. Establish an SSH endpoint with the CML CLI client. See [Initialize an SSH Endpoint](#).
2. Configure the local IDE to use Cloudera Machine Learning as the remote interpreter.
3. Optionally, sync files with tools (like mutagen, SSHFS, or the functionality built into your IDE) from Cloudera Machine Learning to your local machine. Ensure that you adhere to IT policies.
4. Edit the code in the local IDE and run the code interactively on Cloudera Machine Learning.
5. Sync the files you edited locally to Cloudera Machine Learning.
6. Use the Cloudera Machine Learning web UI to perform actions such as deploying a model that uses the code you edited.

You can see an end-to-end example for PyCharm configuration in the [CML Editors Pycharm](#).

**Configure PyCharm as a Local IDE**

Cloudera Machine Learning supports using editors on your machine that allow remote execution and/or file sync over SSH, such as PyCharm.

**About this task**

This topic describes the tasks you need to perform to configure Cloudera Machine Learning to act as a remote SSH interpreter for PyCharm. Once finished, you can use PyCharm to edit and sync the changes to Cloudera Machine Learning. To perform actions such as deploying a model, use the Cloudera Machine Learning web UI.

**Note:** These instructions were written for the Professional Edition of PyCharm Version 2019.1. See the documentation for your version of PyCharm for specific instructions.

Before you begin, ensure that the following prerequisites are met:

- You have an edition of PyCharm that supports SSH, such as the Professional Edition.
- You have an SSH public/private key pair for your local machine.
• You have Contributor permissions for an existing Cloudera Machine Learning project. Alternatively, create a new project you have access to.

**Add Cloudera Machine Learning as an Interpreter for PyCharm**
In PyCharm, you can configure an SSH interpreter. Cloudera Machine Learning uses this method to connect to PyCharm and act as its interpreter.

**About this task**
Before you begin, ensure that the SSH endpoint for Cloudera Machine Learning is running on your local machine. These instructions were written for the Professional Edition of PyCharm Version 2019.1 and are meant as a starting point. If additional information is required, see the documentation for your version of PyCharm for specific instructions.

**Procedure**
1. Open PyCharm.
2. Create a new project.
3. Expand **Project Interpreter** and select **Existing interpreter**.
4. Click on ... and select **SSH Interpreter**
5. Select **New server configuration** and complete the fields:
   - **Host**: localhost
   - **Port**: 2222
   - **Username**: cdsw
6. Select **Key pair** and complete the fields using the RSA private key that corresponds to the public key you added to the **Remote Editing** tab in the Cloudera Machine Learning web UI.
   For macOS users, you must add your RSA private key to your keychain. In a terminal window, run the following command:
   ```bash
   ssh-add -K <path to your private key>/<private_key>
   ```
7. Complete the wizard. Based on the Python version you want to use, enter one of the following parameters:
   - `/usr/local/bin/python2`
   - `/usr/local/bin/python3`
   You are returned to the **New Project** window. **Existing interpreter** is selected, and you should see the connection to Cloudera Machine Learning in the **Interpreter** field.
8. In the **Remote project location** field, specify the following directory:
   `/home/cdsw`
9. Create the project.

**Configure PyCharm to use Cloudera Machine Learning as the Remote Console**

**Procedure**
1. In your project, go to **Settings** and search for **Project Interpreter**.
   Depending on your operating system, **Settings** may be called **Preferences**.
2. Click the gear icon and select **Show All**.
3. Select the Remote Python editor that you added, which is connected to the Cloudera Machine Learning deployment.
4. Add the following interpreter path by clicking on the folder icon:

```
/usr/local/bin/python2.7/site-packages
```

### (Optional) Configure the Sync Between Cloudera Machine Learning and PyCharm
Configuring what files PyCharm ignores can help you adhere to IT policies.

#### About this task
Before you configure syncing behavior between the remote editor and Cloudera Machine Learning, ensure that you understand the policies set forth by IT and the Site Administrator. For example, a policy might require that data remains within the Cloudera Machine Learning deployment but allow you to download and edit code.

#### Procedure
1. In your project, go to **Settings** and search for **Project Interpreter**. Depending on your operating system, **Settings** may be called **Preferences**.
2. Search for **Deployment**.
3. On the **Connection** tab, add the following path to the **Root path** field:

```
/home/cdsw
```
4. Optionally, add a **Deployment path** on the **Mappings** tab if the code for your Cloudera Machine Learning project lives in a subdirectory of the root path.
5. Expand **Deployment** in the left navigation and go to **Options > Upload changed files automatically to the default server** and set the behavior to adhere to the policies set forth by IT and the Site Administrator.
   
   Cloudera recommends setting the behavior to **Automatic upload** because the data remains on the cluster while your changes get uploaded.
6. Sync for the project file(s) to your machine and begin editing.

### Git for Collaboration
Cloudera Machine Learning provides seamless access to Git projects. Whether you are working independently, or as part of a team, you can leverage all of benefits of version control and collaboration with Git from within Cloudera Machine Learning.

Teams that already use Git for collaboration can continue to do so. Each team member will need to create a separate Cloudera Machine Learning project from the central Git repository. For anything but simple projects, Cloudera recommends using Git for version control. You should work on Cloudera Machine Learning the same way you would work locally, and for most data scientists and developers that means using Git.

Cloudera Machine Learning does not include significant UI support for Git, but instead allows you to use the full power of the command line. If you launch a session and open a Terminal, you can run any Git command, including `init`, `add`, `commit`, `branch`, `merge` and `rebase`. Everything should work exactly as it does locally.

When you create a project, you can optionally supply an HTTPS or SSH Git URL that points to a remote repository. The new project is a clone of that remote repository. You can commit, push and pull your code by running a console and opening a Terminal. Note that if you want to use SSH to clone the repo, you will need to first add your personal Cloudera Machine Learning SSH key to your GitHub account. For instructions, see *Adding SSH Key to GitHub*.

If you see Git commands hanging indefinitely, check with your cluster administrators to make sure that the SSH ports on the Cloudera Machine Learning hosts are not blocked.

#### Related Information
- Adding an SSH Key to GitHub
- Creating a Project
Linking an Existing Project to a Git Remote

If you did not create your project from a Git repository, you can link an existing project to a Git remote (for example, `git@github.com:username/repo.git`) so that you can push and pull your code.

### Procedure

1. Launch a new session.
2. Open a terminal.
3. Enter the following commands:

   ```shell
   git init
   git add *
   git commit -a -m 'Initial commit'
   git remote add origin git@github.com:username/repo.git
   ```

   You can run `git status` after `git init` to make sure your `.gitignore` includes a folder for libraries and other non-code artifacts.

Web Applications Embedded in Sessions

This topic describes how Cloudera Machine Learning allows you to embed web applications for frameworks such as Spark 2, TensorFlow, Shiny, and so on within sessions and jobs.

Many data science libraries and processing frameworks include user interfaces to help track progress of your jobs and break down workflows. These are instrumental in debugging and using the platforms themselves. For example, TensorFlow visualizations can be run on TensorBoard. Other web application frameworks such as Shiny and Flask are popular ways for data scientists to display additional interactive analysis in the languages they already know.

Cloudera Machine Learning allows you to access these web UIs directly from sessions and jobs. This feature is particularly helpful when you want to monitor and track progress for batch jobs. Even though jobs don't give you access to the interactive workbench console, you can still track long running jobs through the UI. However, note that the UI is only active so long as the job or session is active. If your session times out after 60 minutes (default timeout value), so will the UI.

**Important:** If you want to share your web application as a long-running standalone application that other business users can access, Cloudera recommends you now use the Applications feature to support long-running web applications on ML workspaces.

If you are only running a server-backed visualization as part of your own analysis, then you can continue to keep embedding web applications in sessions as described in this topic. Note that running web applications in sessions is also the recommended way to develop, test, and debug analytical apps before deployment.

`CDSW_APP_PORT` and `CDSW_READONLY_PORT` are environment variables that point to general purpose public ports. Any HTTP services running in containers that bind to `CDSW_APP_PORT` or `CDSW_READONLY_PORT` are available in browsers at: `http://<$CDSW_ENGINE_ID>.<$CDSW_DOMAIN>`. Therefore, TensorBoard, Shiny, Flask or any other web framework accompanying a project can be accessed directly from within a session or job, as long as it is run on `CDSW_APP_PORT` or `CDSW_READONLY_PORT`.

`CDSW_APP_PORT` is meant for applications that grant some level of control to the project, such as access to the active session or terminal. `CDSW_READONLY_PORT` must be used for applications that grant read-only access to project results.

To access the UI while you are in an active session, click the grid icon in the upper right hand corner of the Cloudera Machine Learning web application, and select the UI from the dropdown. For a job, navigate to the job overview page and click the History tab. Click on a job run to open the session output for the job. You can now click the grid icon in the upper right hand corner of the Cloudera Machine Learning web application to access the UI for this session.
Limitations with port availability

Cloudera Machine Learning exposes only one port per-access level. This means, in version 1.6.0, you can run a maximum of 3 web applications simultaneously:

- one on `CDSW_APP_PORT`, which can be used for applications that grant some level of control over the project to Contributors and Admins,
- one on `CDSW_READONLY_PORT`, which can be used for applications that only need to give read-only access to project collaborators,
- and, one on the now-deprecated `CDSW_PUBLIC_PORT`, which is accessible by all users.

However, by default the editors feature runs third-party browser-based editors on `CDSW_APP_PORT`. Therefore, for projects that are already using browser-based third-party editors, you are left with only 2 other ports to run applications on: `CDSW_READONLY_PORT` and `CDSW_PUBLIC_PORT`. Keep in mind the level of access you want to grant users when you are selecting one of these ports for a web application.

Example: A Shiny Application

This example demonstrates how to create and run a Shiny application and view the associated UI while in an active session.

Create a new, blank project and run an R console. Create the files, `ui.R` and `server.R`, in the project, and copy the contents of the following example files provided by Shiny by RStudio:

R

```r
# ui.R
library(shiny)

# Define UI for application that draws a histogram
shinyUI(fluidPage(

  # Application title
  titlePanel("Hello Shiny!")

  # Sidebar with a slider input for the number of bins
  sidebarLayout(
    sidebarPanel(
      sliderInput("bins", "Number of bins:", min = 1, max = 50, value = 30)
    ),

    # Show a plot of the generated distribution
    mainPanel(
      plotOutput("distPlot")
    )
  )
))
```

R

```r
# server.R
library(shiny)

# Define server logic required to draw a histogram
shinyServer(function(input, output) {

  # Expression that generates a histogram. The expression is
  # wrapped in a call to renderPlot to indicate that:
```
Run the following code in the interactive workbench prompt to install the Shiny package, load the library into the engine, and run the Shiny application.

```
R
install.packages('shiny')
library('shiny')
runApp(port=as.numeric(Sys.getenv("CDSW_READONLY_PORT")), host="127.0.0.1", launch.browser="FALSE")
```

Finally, click the grid icon in the upper right hand corner of the Cloudera Machine Learning web application, and select the Shiny UI, **Hello Shiny!**, from the dropdown. The UI will be active as long as the session is still running.

### Managing Engines

This topic describes how to manage engines and configure engine environments to meet your project requirements.

**Required Role:** Site Administrator

**Note:** On Private Cloud, the corresponding role is **EnvironmentAdmin**.

Site administrators and project administrators are responsible for making sure that all projects on the deployment have access to the engines they need. Site admins can create engine profiles, determine the default engine version to be used across the deployment, and white-list any custom engines that teams require. As a site administrator, you can also customize engine environments by setting global environmental variables and configuring any files/folders that need to be mounted into project environments on run time.

By default, Cloudera Machine Learning ships a base engine image that includes kernels for Python, R, and Scala, along with some additional libraries (see *Configuring Cloudera Machine Learning Engines* for more information) that can be used to run common data analytics operations. Occasionally, new engine versions are released and shipped with Cloudera Machine Learning releases.

Engine images are available in the Site Administrator panel at **Admin > Engines**, under the Engine Images section. As a site administrator, you can select which engine version is used by default for new projects. Furthermore, project administrators can explicitly select which engine image should be used as the default image for a project. To do so, go to the project's Overview page and click **Settings** on the left navigation bar.

If a user publishes a new custom Docker image, site administrators are responsible for white-listing such images for use across the deployment. For more information on creating and managing custom Docker images, see *Configuring the Engine Environment*.

**Related Information**

- Configuring the Engine Environment
- Installing Additional Packages
Managing Resource Profiles

Resource profiles define how many vCPUs and how much memory will reserve for a particular workload (for example, session, job, model).

About this task

As a site administrator you can create several different vCPU, GPU, and memory configurations which will be available when launching a session/job. When launching a new session, users will be able to select one of the available resource profiles depending on their project's requirements.

Resource profiles available when launching a session

![Screenshot of resource profiles](image)

Procedure

1. To create resource profiles, go to the Admin > Runtime/Engines page.
2. Add a new profile under Resource.
   - Cloudera recommends that all profiles include at least 2 GB of RAM to avoid out of memory errors for common user operations.
   - You will see the option to add GPUs to the resource profiles only if your hosts are equipped with GPUs, and you have enabled them for use by setting the relevant properties in `cdsw.conf`.

Configuring the Engine Environment

This section describes some of the ways you can configure engine environments to meet the requirements of your projects.

Install Additional Packages

For information on how to install any additional required packages and dependencies to your engine, see *Installing Additional Packages*.  

![Start A New Session](image)
Environmental Variables

For information on how environmental variables can be used to configure engine environments in Cloudera Machine Learning, see Engine Environment Variables.

Configuring Host Mounts

By default, Cloudera Machine Learning will automatically mount the CDH parcel directory and client configuration for required services such as HDFS, Spark, and YARN into each project's engine. However, if users want to reference any additional files/folders on the host, site administrators will need to configure them here so that they are loaded into engine containers at runtime. Note that the directories specified here will be available to all projects across the deployment.

To configure additional mounts, go to **Admin > Engines** and add the paths to be mounted from the host to the Mounts section.

By default, mount points are loaded into engine containers with read-only permissions. CDSW 1.4.2 (and higher) also include a Write Access checkbox (see image) that you can use to enable read-write access for individual mounted directories. Note that these permissions will apply to all projects across the deployment.

Points to Remember:

- When adding host mounts, try to be as generic as possible without mounting common system files. For example, if you want to add several files under `/etc/spark2-conf`, you can simplify and mount the `/etc/spark2-conf` directory; but adding the parent `/etc` might prevent the engine from running.

  As a general rule, do not mount full system directories that are already in use; such as `/var`, `/tmp`, or `/etc`. This also serves to avoid accidentally exposing confidential information in running sessions.

- Do not add duplicate mount points. This will lead to sessions crashing in the workbench.

Configuring Shared Memory Limit for Docker Images

You can increase the shared memory size for the sessions, experiments, and jobs running within an Engine container within your project. For Docker, the default size of the available shared memory is 64 MB.

To increase the shared memory limit:

1. From the web UI, go to **Projects > Project Settings > Engine > Advanced Settings**
2. Specify the shared memory size in the **Shared Memory Limit** field.
3. Click **Save Advanced Settings** to save the configuration and exit.

This mounts a volume with the `tmpfs` file system to `/dev/shm` and Kubernetes will enforce the given limit. The maximum size of this volume is the half of your physical RAM in the node without the swap.

Related Information

- Engine Environment Variables
- Installing Additional Packages
Set up a custom repository location

You can set up a custom default location for Python and R code package repositories. This is especially useful for air-gapped clusters that are isolated from the PIP and CRAN repositories on the public internet.

Python PIP repository

Custom PIP repositories can be set as default for all engines at a site or project level. The environmental variables can be set at the Project or Site level. If the values are set at the Site level, they can be overridden at the Project level.

1. Set the environmental variables at the appropriate level.
   - For Site level, go to: Site Administration > Engine
   - For Project level, go to: Project Settings > Engine

2. To set a new default URL for the PIP index, enter:
   - PIP_INDEX_URL = <new url>
   - PIP_EXTRA_INDEX_URL = <new url>

CRAN repository

Custom CRAN repositories must be set in a session or as part of a custom engine. To set a new default URL for a CRAN repository, set the following in the /home/cdsw/.Rprofile file:

   options(repos=structure(c(CRAN="<mirror URL>")))

Installing Additional Packages

Cloudera Machine Learning engines are preloaded with a few common packages and libraries for R, Python, and Scala. However, a key feature of Cloudera Machine Learning is the ability of different projects to install and use libraries pinned to specific versions, just as you would on your local computer.

Generally, Cloudera recommends you install all required packages locally into your project. This will ensure you have the exact versions you want and that these libraries will not be upgraded when Cloudera upgrades the base engine image. You only need to install libraries and packages once per project. From then on, they are available to any new engine you spawn throughout the lifetime of the project.

You can install additional libraries and packages from the workbench, using either the command prompt or the terminal.

**Note:**

Cloudera Machine Learning does not currently support installation of packages that require root access to the hosts. For such use-cases, you will need to create a new custom engine that extends the base engine image to include the required packages. For instructions, see Creating a Customized Engine Image.

(Python and R) Install Packages Using Workbench Command Prompt

To install a package from the command prompt:

1. Navigate to your project's Overview page. Click Open Workbench and launch a session.

2. At the command prompt (see Native Workbench Console and Editor) in the bottom right, enter the command to install the package. Some examples using Python and R have been provided.

R

```r
# Install from CRAN
install.packages("ggplot2")

# Install using devtools
install.packages('devtools')
library(devtools)
```
Install with GitHub

```python
install_github("hadley/ggplot2")
```

### Python 2

# Installing from console using ! shell operator and pip:
!pip install beautifulsoup

# Installing from terminal
pip install beautifulsoup

### Python 3

# Installing from console using ! shell operator and pip3:
!pip3 install beautifulsoup4

# Installing from terminal
pip3 install beautifulsoup4

#### (Python Only) Using a Requirements File

For a Python project, you can specify a list of the packages you want in a requirements.txt file that lives in your project. The packages can be installed all at once using pip/pip3.

1. Create a new file called requirements.txt file within your project:

   ```
   beautifulsoup4==4.6.0
   seaborn==0.7.1
   ```

2. To install the packages in a Python 3 engine, run the following command in the workbench command prompt.

   ```
   !pip3 install -r requirements.txt
   ```

   For Python 2 engines, use pip.

   ```
   !pip install -r requirements.txt
   ```

#### Related Information

- Conda

#### Using Conda to Manage Dependencies

You can install additional libraries and packages from the workbench, using either the command prompt or the terminal. Alternatively, you might choose to use a package manager such as Conda to install and maintain packages and their dependencies. This topic describes some basic usage guidelines for Conda.

Cloudera Machine Learning recommends using pip for package management along with a requirements.txt file (as described in the previous section). However, for users that prefer Conda, the default engine in Cloudera Machine Learning includes two environments called python2.7 and python3.6. These environments are added to sys.path, depending on the version of Python selected when you launch a new session.

In Python 2 and Python 3 sessions and attached terminals, Cloudera Machine Learning automatically sets the CONDA_DEFAULT_ENV and CONDA_PREFIX environment variables to point to Conda environments under /home/cdsw/.conda.

However, Cloudera Machine Learning does not automatically configure Conda to pin the actual Python version. Therefore, if you are using Conda to install a package, you must specify the version of Python. For example, to use Conda to install the feather-format package into the python3.6 environment, run the following command in the Workbench command prompt:

```
!conda install -y -c conda-forge python=3.6.9 feather-format
```
To install a package into the python2.7 environment, run:

```
!conda install -y -c conda-forge python=2.7.17 feather-format
```

Note that on sys.path, pip packages have precedence over conda packages.

**Note:**

- Cloudera Machine Learning does not automatically configure a Conda environment for R and Scala sessions and attached terminals. If you want to use Conda to install packages from an R or Scala session or terminal, you must manually configure Conda to install packages into the desired environment.

**Creating an Extensible Engine With Conda**

Cloudera Machine Learning also allows you to configure the Engine Environment to include packages of your choice using Conda. To create an extended engine:

1. Add the following lines to a Dockerfile to extend the base engine, push the engine image to your Docker registry, and include the new engine in the allowlist, for your project. For more details on this step, see Configuring the Engine Environment.

   **Python 2**
   ```
   RUN mkdir -p /opt/conda/envs/python2.7
   RUN conda install -y nbconvert python=2.7.17 -n python2.7
   ```

   **Python 3**
   ```
   RUN mkdir -p /opt/conda/envs/python3.6
   RUN conda install -y nbconvert python=3.6.9 -n python3.6
   ```

2. Set the PYTHONPATH environmental variable as shown below. You can set this either globally in the site administrator dashboard, or for a specific project by going to the project's Settings > Engine page.

   **Python 2**
   ```
   PYTHONPATH=$PYTHONPATH:/opt/conda/envs/python2.7/lib/python2.7/site-packages
   ```

   **Python 3**
   ```
   PYTHONPATH=$PYTHONPATH:/opt/conda/envs/python3.6/lib/python3.6/site-packages
   ```

**Related Information**

- Conda
- Configuring the Engine Environment

**Engine Environment Variables**

This topic describes how engine environmental variables work. It also lists the different scopes at which they can be set and the order of precedence that will be followed in case of conflicts.

Environmental variables allow you to customize engine environments for projects. For example, if you need to configure a particular timezone for a project, or increase the length of the session/job timeout windows, you can use environmental variables to do so. Environmental variables can also be used to assign variable names to secrets such as passwords or authentication tokens to avoid including these directly in the code.

In general, Cloudera recommends that you do not include passwords, tokens, or any other secrets directly in your code because anyone with read access to your project will be able to view this information. A better place to store...
secrets is in your project's environment variables, where only project collaborators and admins have view access. They can therefore be used to securely store confidential information such as your AWS keys or database credentials.

Cloudera Machine Learning allows you to define environmental variables for the following scopes:

Global

A site administrator for your Cloudera Machine Learning deployment can set environmental variables on a global level. These values will apply to every project on the deployment.

To set global environmental variables, go to Admin > Runtime/Engines.

Project

Project administrators can set project-specific environmental variables to customize the engines launched for a project. Variables set here will override the global values set in the site administration panel.

To set environmental variables for a project, go to the project's Overview page and click Settings > Advanced.

Job

Environments for individual jobs within a project can be customized while creating the job. Variables set per-job will override the project-level and global settings.

To set environmental variables for a job, go to the job's Overview page and click Settings > Set Environmental Variables.

Experiments

Engines created for execution of experiments are completely isolated from the project. However, these engines inherit values from environmental variables set at the project-level and/or global level. Variables set at the project-level will override the global values set in the site administration panel.

Models

Model environments are completely isolated from the project. Environmental variables for these engines can be configured during the build stage of the model deployment process. Models will also inherit any environment variables set at the project and global level. However, variables set per-model build will override other settings.

Related Information

Basic Concepts and Terminology

Engine Environment Variables

The following table lists Cloudera Machine Learning environment variables that you can use to customize your project environments. These can be set either as a site administrator or within the scope of a project or a job.

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAX_TEXT_LENGTH</td>
<td>Maximum number of characters that can be displayed in a single text cell. By default, this value is set to 800,000 and any more characters will be truncated. Default: 800,000</td>
</tr>
<tr>
<td>SESSION_MAXIMUM_MINUTES</td>
<td>Maximum number of minutes a session can run before it times out. Default: 60<em>24</em>7 minutes (7 days) Maximum Value: 35,000 minutes</td>
</tr>
<tr>
<td>JOB_MAXIMUM_MINUTES</td>
<td>Maximum number of minutes a job can run before it times out. Default: 60<em>24</em>7 minutes (7 days) Maximum Value: 35,000 minutes</td>
</tr>
</tbody>
</table>
### How To

#### Environment Variable Description

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IDLE_MAXIMUM_MINUTES</strong></td>
<td>Maximum number of minutes a session can remain idle before it exits.</td>
</tr>
<tr>
<td>Default: 60 minutes</td>
<td>Maximum Value: 35,000 minutes</td>
</tr>
<tr>
<td><strong>CONDA_DEFAULT_ENV</strong></td>
<td>Points to the default Conda environment so you can use Conda to install/manage packages in the Workbench. For more details on when to use this variable, see Installing Additional Packages.</td>
</tr>
</tbody>
</table>

**Per-Engine Environmental Variables:** In addition to the previous table, there are some more built-in environmental variables that are set by the Cloudera Machine Learning application itself and do not need to be modified by users. These variables are set per-engine launched by Cloudera Machine Learning and only apply within the scope of each engine.

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CDSW_PROJECT</strong></td>
<td>The project to which this engine belongs.</td>
</tr>
<tr>
<td><strong>CDSW_ENGINE_ID</strong></td>
<td>The ID of this engine. For sessions, this appears in your browser's URL bar.</td>
</tr>
<tr>
<td><strong>CDSW_MASTER_ID</strong></td>
<td>If this engine is a worker, this is the CDSW_ENGINE_ID of its master.</td>
</tr>
<tr>
<td><strong>CDSW_MASTER_IP</strong></td>
<td>If this engine is a worker, this is the IP address of its master.</td>
</tr>
<tr>
<td><strong>CDSW_PUBLIC_PORT</strong></td>
<td><strong>Note:</strong> This property is deprecated. See CDSW_APP_PORT and CDSW_READONLY_PORT for alternatives.</td>
</tr>
<tr>
<td>A port on which you can expose HTTP services in the engine to browsers. HTTP services that bind CDSW_PUBLIC_PORT will be available in browsers at: http(s)://&lt;CDSW_ENGINE_ID&gt;.&lt;CDSW_DOMAIN&gt;. By default, CDSW_PUBLIC_PORT is set to 8080. A direct link to these web services will be available from the grid icon in the upper right corner of the Cloudera Machine Learning web application, as long as the job or session is still running. For more details, see Accessing Web User Interfaces from Cloudera Machine Learning. In Cloudera Machine Learning, setting CDSW_PUBLIC_PORT to a non-default port number is not supported.</td>
<td></td>
</tr>
</tbody>
</table>

| **CDSW_APP_PORT**     | A port on which you can expose HTTP services in the engine to browsers. HTTP services that bind CDSW_APP_PORT will be available in browsers at: http(s)://<CDSW_ENGINE_ID>.<CDSW_DOMAIN>. Use this port for applications that grant some control to the project, such as access to the session or terminal. A direct link to these web services will be available from the grid icon in the upper right corner of the Cloudera Machine Learning web application as long as the job or session runs. Even if the web UI does not have authentication, only Contributors and those with more access to the project can access it. For more details, see Accessing Web User Interfaces from Cloudera Machine Learning. Note that if the Site Administrator has enabled Allow only session creators to run commands on active sessions, then the UI is only available to the session creator. Other users will not be able to access it. Use 127.0.0.1 as the IP. |

| **CDSW_READONLY_PORT**| A port on which you can expose HTTP services in the engine to browsers. HTTP services that bind CDSW_READONLY_PORT will be available in browsers at: http(s)://<CDSW_ENGINE_ID>.<CDSW_DOMAIN>. Use this port for applications that grant read-only access to project results. A direct link to these web services will be available to users with from the grid icon in the upper right corner of the Cloudera Machine Learning web application as long as the job or session runs. Even if the web UI does not have authentication, Viewers and those with more access to the project can access it. For more details, see Accessing Web User Interfaces from Cloudera Machine Learning. Use 127.0.0.1 as the IP. |
| **CDSW_DOMAIN**       | The domain on which Cloudera Machine Learning is being served. This can be useful for iframing services, as demonstrated in Accessing Web User Interfaces from Cloudera Machine Learning. |
### How To

#### Environment Variable

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDSW_CPU_MILLICORES</td>
<td>The number of CPU cores allocated to this engine, expressed in thousandths of a core.</td>
</tr>
<tr>
<td>CDSW_MEMORY_MB</td>
<td>The number of megabytes of memory allocated to this engine.</td>
</tr>
<tr>
<td>CDSW_IP_ADDRESS</td>
<td>Other engines in the Cloudera Machine Learning cluster can contact this engine on this IP address.</td>
</tr>
</tbody>
</table>

**Related Information**

- Installing Additional Packages

#### Accessing Environmental Variables from Projects

This topic shows you how to access environmental variables from your code.

Environmental variables are injected into every engine launched for a project, contingent on the scope at which the variable was set (global, project, etc.). The following code samples show how to access a sample environment variable called `DATABASE_PASSWORD` from your project code.

**R**

```
database.password <- Sys.getenv("DATABASE_PASSWORD")
```

**Python**

```
import os
database_password = os.environ['DATABASE_PASSWORD']
```

**Scala**

```
System.getenv("DATABASE_PASSWORD")
```

**Appending Values to Environment Variables:**

You can also set environment variables to append to existing values instead of replacing them. For example, when setting the `LD_LIBRARY_PATH` variable, you can set the value to `LD_LIBRARY_PATH:/path/to/set`.

**Customized Engine Images**

This topic explains how custom engines work and when they should be used.

By default, Cloudera Machine Learning engines are preloaded with a few common packages and libraries for R, Python, and Scala. In addition to these, Cloudera Machine Learning also allows you to install any other packages or libraries that are required by your projects. However, directly installing a package to a project as described above might not always be feasible. For example, packages that require root access to be installed, or that must be installed to a path outside `/home/cdsw` (outside the project mount), cannot be installed directly from the workbench.

For such circumstances, Cloudera Machine Learning allows you to extend the base Docker image and create a new Docker image with all the libraries and packages you require. Site administrators can then include this new image in the allowlist for use in projects, and project administrators set the new white-listed image to be used as the default engine image for their projects. For an end-to-end example of this process, see *End-to-End Example: MeCab*.

**Note:** You will need to remove any unnecessary Cloudera sources or repositories that are inaccessible because of the paywall.

Note that this approach can also be used to accelerate project setup across the deployment. For example, if you want multiple projects on your deployment to have access to some common dependencies (package or software or driver) out of the box, or even if a package just has a complicated setup, it might be easier to simply provide users with an engine that has already been customized for their project(s).
Related Resources

- The Cloudera Engineering Blog post on *Customizing Docker Images in Cloudera Machine Learning* describes an end-to-end example on how to build and publish a customized Docker image and use it as an engine in Cloudera Machine Learning.
- For an example of how to extend the base engine image to include Conda, see *Installing Additional Packages*.

**Related Information**
- End-to-End Example: MeCab
- Installing Additional Packages
- Customizing Docker Images in Cloudera Machine Learning

### Creating a Customized Engine Image

This section walks you through the steps required to create your own custom engine based on the Cloudera Machine Learning base image.

For a complete example, see *End-to-End Example: MeCab*.

#### Create a Dockerfile for the Custom Image

This topic shows you how to create a Dockerfile for a custom image.

The first step when building a customized image is to create a Dockerfile that specifies which packages you would like to install in addition to the base image.

For example, the following Dockerfile installs the `beautifulsoup4` package on top of the base Ubuntu image that ships with Cloudera Machine Learning.

```shell
# Dockerfile

# Specify a Cloudera Machine Learning base image
FROM docker.repository.cloudera.com/cloudera/cdsw/engine:13-cml-2021.02-1

# Update packages on the base image and install beautifulsoup4
RUN apt-get update
RUN pip install beautifulsoup4 && pip3 install beautifulsoup4
```

#### Build the New Docker Image

This topic shows you how to use Docker to build a custom image.

A new custom Docker image can be built on any host where Docker binaries are installed. To install these binaries, run the following command on the host where you want to build the new image:

```shell
docker build -t <image-name>:<tag> . -f Dockerfile
```

If you want to build your image on the Cloudera Machine Learning workspace, you must add the `--network=host` option to the build command:

```shell
docker build --network=host -t <image-name>:<tag> . -f Dockerfile
```

#### Distribute the Image

This topic explains the different methods that can be used to distribute a custom engine to all the hosts.

Once you have built a new custom engine, use one of the following ways to distribute the new image to all your Cloudera Machine Learning hosts:

- **Push the image to a public registry such as DockerHub**
  
  For instructions, refer the Docker documentation `docker push` and `Push images to Docker Cloud`.

- **Push the image to your company’s Docker registry**
When using this method, make sure to tag your image with the following schema:

```
docker tag <image-name> <company-registry>/<user-name>/<image-name>:<tag>
```

Once the image has been tagged properly, use the following command to push the image:

```
docker push <company-registry>/<user-name>/<image-name>:<tag>
```

The MeCab example at the end of this topic uses this method.

**Related Information**

- `docker push`

**Including Images in allowlist for Cloudera Machine Learning projects**

This topic describes how to include custom images in the allowlist so that they can be used in projects.

Including a customized image in Cloudera Machine Learning is a two-step process.

1. Include the image in the allowlist for the whole deployment.
   
   First, a site administrator will need to clear the new image for use on the deployment.
   
   a. Log in as a site administrator.
   b. Click Admin > Engines.
   c. Add `<company-registry>/<user-name>/<image-name>:<tag>` to the allowlist of engine images.

2. Include the image in the allowlist for a specific project
   
   If you want to start using the image in a project, the project administrator will need to set this image as the default image for the project.
   
   a. Go to the project Settings page.
   b. Click Engines.
   c. Select the new customized engine from the drop-down list of available Docker images. Sessions and jobs you run in your project will now have access to this engine.

**Add Docker registry credentials**

To enable CML to fetch custom engines from a secure repository, as Administrator you need to add Docker registry credentials.

Create a `kubectl` secret named `regcred` for your secured Docker registry. The following command creates the secret in your Kubernetes cluster:

```
kubectl create secret docker-registry regcred
  --docker-server=<server host>
  --docker-username=<username>
  --docker-password=<password>
  -n <compute namespace eg. mlx>
```

The next time the engine image is pulled, the new secret will be picked up.

**Limitations**

This topic lists some limitations associated with custom engines.

- Cloudera Machine Learning only supports customized engines that are based on the Cloudera Machine Learning base image.
- Cloudera Machine Learning does not support creation of custom engines larger than 10 GB.

  Cloudera Bug: DSE-4420
Cloudera Machine Learning does not support pulling images from registries that require Docker credentials.

Cloudera Bug: DSE-1521

The contents of certain pre-existing standard directories such as /home/cdsw, /tmp, and so on, cannot be modified while creating customized engines. This means any files saved in these directories will not be accessible from sessions that are running on customized engines.

Workaround: Create a new custom directory in the Dockerfile used to create the customized engine, and save your files to that directory.

### Add Docker registry credentials

To enable CML to fetch custom engines from a secure repository, as Administrator you need to add Docker registry credentials.

Create a kubectl secret named regcred for your secured Docker registry. The following command creates the secret in your Kubernetes cluster:

```
kubectl create secret docker-registry regcred
--docker-server=<server host>
--docker-username=<username>
--docker-password=<password>
-n <compute namespace eg. mlx>
```

The next time the engine image is pulled, the new secret will be picked up.

### End-to-End Example: MeCab

This topic walks you through a simple end-to-end example on how to build and use custom engines.

This section demonstrates how to customize the Cloudera Machine Learning base engine image to include the MeCab (a Japanese text tokenizer) library.

This is a sample Dockerfile that adds MeCab to the Cloudera Machine Learning base image.

```
# Dockerfile
FROM docker.repository.cloudera.com/cloudera/cdsw/engine:13-cml-2021.02-1
RUN rm /etc/apt/sources.list.d/*
RUN apt-get update && \
    apt-get install -y -q mecab \ 
    libmecab-dev \ 
    mecab-ipadic-utf8 && \
    apt-get clean && \
    rm -rf /var/lib/apt/lists/*
RUN cd /tmp && \
    git clone --depth 1 https://github.com/neologd/mecab-ipadic-neologd.git && \
    /tmp/mecab-ipadic-neologd/bin/install-mecab-ipadic-neologd -y -n -p /var/lib/mecab/dic/neologd && \
    rm -rf /tmp/mecab-ipadic-neologd
RUN pip install --upgrade pip
RUN pip install mecab-python==0.996
```

To use this image on your Cloudera Machine Learning project, perform the following steps.

1. Build a new image with the Dockerfile.

```
docker build --network=host -t <company-registry>/user/cdsw-mecab:latest . -f Dockerfile
```
2. Push the image to your company’s Docker registry.

   ```bash
docker push <your-company-registry>/user/cdsw-mecab:latest
   ```

3. Whitelist the image, `<your-company-registry>/user/cdsw-mecab:latest`. Only a site administrator can do this.

   Go to Admin > Engines and add `<company-registry>/user/cdsw-mecab:latest` to the list of whitelisted engine images.

4. Ask a project administrator to set the new image as the default for your project. Go to the project Settings, click Engines, and select `<company-registry>/user/cdsw-mecab:latest` from the dropdown.

You should now be able to run this project on the customized MeCab engine.

**Pre-Installed Packages in Engines**

Cloudera Machine Learning ships with several base engine images that include Python and R kernels, and frequently used libraries.

**Base Engine 13-cml-2020.08-1**

Engine 13 ships Python versions 2.7.18 and 3.6.10, and R version 3.6.3.

Items in bold indicate a new version since the last release.

**Note:** This is the only engine available on CML Private Cloud 1.0.

**Table 2: Python 3 Libraries**

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
</tbody>
</table>
### Table 3: Python 2 Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.16.10</td>
</tr>
<tr>
<td>numpy</td>
<td>1.16.5</td>
</tr>
<tr>
<td>pandas</td>
<td>0.24.2</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.8.0</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.8.1</td>
</tr>
<tr>
<td>matplotlib</td>
<td>3.1.2</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.9.0</td>
</tr>
<tr>
<td>cython</td>
<td>0.29.13</td>
</tr>
<tr>
<td>six</td>
<td>1.15.0</td>
</tr>
</tbody>
</table>

### Table 4: R Libraries

<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCurl</td>
<td>1.98.1.2</td>
</tr>
<tr>
<td>caTools</td>
<td>1.18.0</td>
</tr>
<tr>
<td>svTools</td>
<td>0.9.5</td>
</tr>
<tr>
<td>png</td>
<td>0.1.7</td>
</tr>
<tr>
<td>RJSONIO</td>
<td>1.3.1.4</td>
</tr>
<tr>
<td>ggplot2</td>
<td>3.3.1</td>
</tr>
<tr>
<td>cluster</td>
<td>2.1.0</td>
</tr>
<tr>
<td>codetools</td>
<td>0.2.16</td>
</tr>
<tr>
<td>foreign</td>
<td>0.8.69</td>
</tr>
<tr>
<td>dplyr</td>
<td>1.0.0</td>
</tr>
<tr>
<td>htttr</td>
<td>1.4.1</td>
</tr>
<tr>
<td>httpuv</td>
<td>1.5.4</td>
</tr>
<tr>
<td>jsonlite</td>
<td>1.6.1</td>
</tr>
</tbody>
</table>
Package | Version
---|---
magrittr | 1.5
knitr | 1.28
purr | 0.3.4
tm | 0.7.7
proxy | 0.4.24
data.table | 1.12.8
stringr | 1.4.0
Rook | 1.1.1
rJava | 0.9.12
devtools | 2.3.0

### Related Information

**Base Engine 9**
**Base Engine 10**
**Base Engine 11**
**Base Engine 12**

**Base Engine 12-cml-2020.06-2**

Engine 12 ships Python versions 2.7.18 and 3.6.10, and R version 3.6.3.

### Table 5: Python 3 Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.16.0</td>
</tr>
<tr>
<td>numpy</td>
<td>1.17.2</td>
</tr>
<tr>
<td>pandas</td>
<td>0.25.1</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.8.1</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.8.1</td>
</tr>
<tr>
<td>matplotliblib</td>
<td>3.1.2</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.9.0</td>
</tr>
<tr>
<td>Cython</td>
<td>0.29.13</td>
</tr>
</tbody>
</table>

### Table 6: Python 2 Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.16.0</td>
</tr>
<tr>
<td>numpy</td>
<td>1.16.5</td>
</tr>
<tr>
<td>pandas</td>
<td>0.24.2</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.8.0</td>
</tr>
<tr>
<td>Library</td>
<td>Version</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.8.1</td>
</tr>
<tr>
<td>futures</td>
<td>3.3.0</td>
</tr>
<tr>
<td>matplotlib</td>
<td>2.2.4</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.9.0</td>
</tr>
<tr>
<td>Cython</td>
<td>0.29.13</td>
</tr>
</tbody>
</table>

**Table 7: R Libraries**

<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCurl</td>
<td>1.98.1.2</td>
</tr>
<tr>
<td>caTools</td>
<td>1.18.0</td>
</tr>
<tr>
<td>svTools</td>
<td>0.9.5</td>
</tr>
<tr>
<td>png</td>
<td>0.1.7</td>
</tr>
<tr>
<td>RJSONIO</td>
<td>1.3.1.4</td>
</tr>
<tr>
<td>ggplot2</td>
<td>3.3.1</td>
</tr>
<tr>
<td>cluster</td>
<td>2.1.0</td>
</tr>
<tr>
<td>codetools</td>
<td>0.2.16</td>
</tr>
<tr>
<td>foreign</td>
<td>0.8.69</td>
</tr>
<tr>
<td>dplyr</td>
<td>1.0.0</td>
</tr>
<tr>
<td>httr</td>
<td>1.4.1</td>
</tr>
<tr>
<td>httpuv</td>
<td>1.5.4</td>
</tr>
<tr>
<td>jsonlite</td>
<td>1.6.1</td>
</tr>
<tr>
<td>magrittr</td>
<td>1.5</td>
</tr>
<tr>
<td>knitr</td>
<td>1.28</td>
</tr>
<tr>
<td>purrr</td>
<td>0.3.4</td>
</tr>
<tr>
<td>tm</td>
<td>0.7.7</td>
</tr>
<tr>
<td>proxy</td>
<td>0.4.24</td>
</tr>
<tr>
<td>data.table</td>
<td>1.12.8</td>
</tr>
<tr>
<td>stringr</td>
<td>1.4.0</td>
</tr>
<tr>
<td>Rook</td>
<td>1.1.1</td>
</tr>
<tr>
<td>rJava</td>
<td>0.9.12</td>
</tr>
<tr>
<td>devtools</td>
<td>2.3.0</td>
</tr>
</tbody>
</table>

**Related Information**

Base Engine 9  
Base Engine 10  
Base Engine 11  
Base Engine 13  

**Base Engine 11-cml1.4**

Engine 11 ships Python versions 2.7.17 and 3.6.9, and R version 3.6.2.
### Table 8: Python 3 Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.16.0</td>
</tr>
<tr>
<td>numpy</td>
<td>1.17.2</td>
</tr>
<tr>
<td>pandas</td>
<td>0.25.1</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.8.1</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.8.1</td>
</tr>
<tr>
<td>matplotlib</td>
<td>2.0.0</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.9.0</td>
</tr>
<tr>
<td>Cython</td>
<td>0.29.13</td>
</tr>
</tbody>
</table>

### Table 9: Python 2 Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.16.0</td>
</tr>
<tr>
<td>numpy</td>
<td>1.16.5</td>
</tr>
<tr>
<td>pandas</td>
<td>0.24.2</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.8.0</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.8.1</td>
</tr>
<tr>
<td>futures</td>
<td>3.3.0</td>
</tr>
<tr>
<td>matplotlib</td>
<td>2.0.0</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.9.0</td>
</tr>
<tr>
<td>Cython</td>
<td>0.29.13</td>
</tr>
</tbody>
</table>

### Table 10: R Libraries

<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCurl</td>
<td>1.95.4.12</td>
</tr>
<tr>
<td>caTools</td>
<td>1.17.1.3</td>
</tr>
<tr>
<td>svTools</td>
<td>0.9.5</td>
</tr>
<tr>
<td>png</td>
<td>0.1.7</td>
</tr>
<tr>
<td>RJSONIO</td>
<td>1.3.1.3</td>
</tr>
<tr>
<td>ggplot2</td>
<td>3.2.1</td>
</tr>
<tr>
<td>cluster</td>
<td>2.1.0</td>
</tr>
<tr>
<td>codetools</td>
<td>0.2.16</td>
</tr>
<tr>
<td>foreign</td>
<td>0.8.73</td>
</tr>
<tr>
<td>dplyr</td>
<td>0.8.3</td>
</tr>
<tr>
<td>htttr</td>
<td>1.4.1</td>
</tr>
</tbody>
</table>
### Related Information

| Base Engine 9 |
| Base Engine 10 |
| Base Engine 12 |
| Base Engine 13 |

#### Base Engine 10-cml1.3

Engine 10 ships Python versions 2.7.17 and 3.6.9, and R version 3.5.1.

#### Table 11: Python 3 Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.16.0</td>
</tr>
<tr>
<td>numpy</td>
<td>1.17.2</td>
</tr>
<tr>
<td>pandas</td>
<td>0.25.1</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.8.1</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.8.1</td>
</tr>
<tr>
<td>matplotlib</td>
<td>2.0.0</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.9.0</td>
</tr>
<tr>
<td>Cython</td>
<td>0.29.13</td>
</tr>
</tbody>
</table>

#### Table 12: Python 2 Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.22.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.16.0</td>
</tr>
<tr>
<td>numpy</td>
<td>1.16.5</td>
</tr>
<tr>
<td>Library</td>
<td>Version</td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>pandas</td>
<td>0.24.2</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.8.0</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.8.1</td>
</tr>
<tr>
<td>futures</td>
<td>3.3.0</td>
</tr>
<tr>
<td>matplotlib</td>
<td>2.0.0</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.9.0</td>
</tr>
<tr>
<td>Cython</td>
<td>0.29.13</td>
</tr>
</tbody>
</table>

**Table 13: R Libraries**

<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCurl</td>
<td>1.95.4.12</td>
</tr>
<tr>
<td>caTools</td>
<td>1.17.1.3</td>
</tr>
<tr>
<td>svTools</td>
<td>0.9.5</td>
</tr>
<tr>
<td>png</td>
<td>0.1.7</td>
</tr>
<tr>
<td>RJSONIO</td>
<td>1.3.1.3</td>
</tr>
<tr>
<td>ggplot2</td>
<td>3.2.1</td>
</tr>
<tr>
<td>cluster</td>
<td>2.1.0</td>
</tr>
<tr>
<td>codetools</td>
<td>0.2.16</td>
</tr>
<tr>
<td>foreign</td>
<td>0.8.73</td>
</tr>
<tr>
<td>dplyr</td>
<td>0.8.3</td>
</tr>
<tr>
<td>httr</td>
<td>1.4.1</td>
</tr>
<tr>
<td>httpuv</td>
<td>1.5.2</td>
</tr>
<tr>
<td>jsonlite</td>
<td>1.6</td>
</tr>
<tr>
<td>magrittr</td>
<td>1.5</td>
</tr>
<tr>
<td>knitr</td>
<td>1.26</td>
</tr>
<tr>
<td>purrr</td>
<td>0.3.3</td>
</tr>
<tr>
<td>tm</td>
<td>0.7.7</td>
</tr>
<tr>
<td>proxy</td>
<td>0.4.23</td>
</tr>
<tr>
<td>data.table</td>
<td>1.12.8</td>
</tr>
<tr>
<td>stringr</td>
<td>1.4.0</td>
</tr>
<tr>
<td>Rook</td>
<td>1.1.1</td>
</tr>
<tr>
<td>rJava</td>
<td>0.9.11</td>
</tr>
<tr>
<td>devtools</td>
<td>2.2.1</td>
</tr>
</tbody>
</table>

**Related Information**

- Base Engine 9
- Base Engine 11
- Base Engine 12
- Base Engine 13
Base Engine 9-cml1.2

Engine 9 ships Python 2.7.11 and 3.6.8, and R version 3.5.1.

Table 14: Python Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipython</td>
<td>5.1.0</td>
</tr>
<tr>
<td>requests</td>
<td>2.13.0</td>
</tr>
<tr>
<td>Flask</td>
<td>0.12.0</td>
</tr>
<tr>
<td>simplejson</td>
<td>3.10.0</td>
</tr>
<tr>
<td>numpy</td>
<td>1.13.3</td>
</tr>
<tr>
<td>pandas</td>
<td>0.20.1</td>
</tr>
<tr>
<td>pandas-datareader</td>
<td>0.2.1</td>
</tr>
<tr>
<td>py4j</td>
<td>0.10.7</td>
</tr>
<tr>
<td>futures</td>
<td>2.1.4</td>
</tr>
<tr>
<td>matplotlib</td>
<td>2.0.0</td>
</tr>
<tr>
<td>seaborn</td>
<td>0.8.0</td>
</tr>
<tr>
<td>Cython</td>
<td>0.25.2</td>
</tr>
<tr>
<td>kudu-python</td>
<td>1.2.0</td>
</tr>
</tbody>
</table>

Table 15: R Libraries

<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCurl</td>
<td>1.95.4.12</td>
</tr>
<tr>
<td>caTools</td>
<td>1.17.1.2</td>
</tr>
<tr>
<td>svTools</td>
<td>0.9.5</td>
</tr>
<tr>
<td>png</td>
<td>0.1.7</td>
</tr>
<tr>
<td>RJSONIO</td>
<td>1.3.1.2</td>
</tr>
<tr>
<td>ggplot2</td>
<td>3.1.1</td>
</tr>
<tr>
<td>cluster</td>
<td>2.0.9</td>
</tr>
<tr>
<td>codetools</td>
<td>0.2.16</td>
</tr>
<tr>
<td>foreign</td>
<td>0.8.71</td>
</tr>
<tr>
<td>dplyr</td>
<td>0.8.1</td>
</tr>
<tr>
<td>httr</td>
<td>1.4.0</td>
</tr>
<tr>
<td>httpuv</td>
<td>1.5.1</td>
</tr>
<tr>
<td>jsonlite</td>
<td>1.6</td>
</tr>
<tr>
<td>magrittr</td>
<td>1.5</td>
</tr>
<tr>
<td>knitr</td>
<td>1.23</td>
</tr>
<tr>
<td>purrr</td>
<td>0.3.2</td>
</tr>
<tr>
<td>tm</td>
<td>0.7.6</td>
</tr>
<tr>
<td>proxy</td>
<td>0.4.23</td>
</tr>
<tr>
<td>data.table</td>
<td>1.12.2</td>
</tr>
<tr>
<td>stringr</td>
<td>1.4.0</td>
</tr>
</tbody>
</table>
Sparks Configuration Files

Cloudera Machine Learning supports configuring Spark 2 properties on a per project basis with the spark-defaults.conf file. If there is a file called spark-defaults.conf in your project root, this will be automatically be added to the global Spark defaults.

**Note:** Spark is not supported for ML Runtimes.

To specify an alternate file location, set the environmental variable, SPARK_CONFIG, to the path of the file relative to your project. If you’re accustomed to submitting a Spark job with key-values pairs following a --conf flag, these can also be set in a spark-defaults.conf file instead. For a list of valid key-value pairs, refer to Spark Configuration.

Administrators can set environment variable paths in the /etc/spark2/conf/spark-env.sh file.

Related Information
Spark Configuration

Managing Memory Available for Spark Drivers

By default, the amount of memory allocated to Spark driver processes is set to a 0.8 fraction of the total memory allocated for the engine container. If you want to allocate more or less memory to the Spark driver process, you can override this default by setting the spark.driver.memory property in spark-defaults.conf (as described above).

Managing Dependencies for Spark 2 Jobs

As with any Spark job, you can add external packages to the executor on startup. To add external dependencies to Spark jobs, specify the libraries you want added by using the appropriate configuration parameter in a spark-defaults.conf file.

The following table lists the most commonly used configuration parameters for adding dependencies and how they can be used:

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>spark.files</td>
<td>Comma-separated list of files to be placed in the working directory of each Spark executor.</td>
</tr>
<tr>
<td>spark.submit.pyFiles</td>
<td>Comma-separated list of .zip, .egg, or .py files to place on PYTHONPATH for Python applications.</td>
</tr>
<tr>
<td>spark.jars</td>
<td>Comma-separated list of local jars to include on the Spark driver and Spark executor classpaths.</td>
</tr>
</tbody>
</table>
How To

Property Description

spark.jars.packages Comma-separated list of Maven coordinates of jars to include on the Spark driver and Spark executor classpaths. When configured, Spark will search the local Maven repo, and then Maven central and any additional remote repositories configured by spark.jars.ivy. The format for the coordinates are groupId:artifactId:version.

spark.jars.ivy Comma-separated list of additional remote repositories to search for the coordinates given with spark.jars.packages.

Example spark-defaults.conf

Here is a sample spark-defaults.conf file that uses some of the Spark configuration parameters discussed in the previous section to add external packages on startup.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>spark.jars.packages</td>
<td>org.scalaj:scalaj-http_2.11:2.3.0</td>
</tr>
<tr>
<td></td>
<td>spark.jars my_sample.jar</td>
</tr>
<tr>
<td></td>
<td>spark.files data/test_data_1.csv,data/test_data_2.csv</td>
</tr>
</tbody>
</table>

**spark.jars.packages**

The scalaj package will be downloaded from Maven central and included on the Spark driver and executor classpaths.

**spark.jars**

The pre-existing jar, my_sample.jar, residing in the root of this project will be included on the Spark driver and executor classpaths.

**spark.files**

The two sample data sets, test_data_1.csv and test_data_2.csv, from the /data directory of this project will be distributed to the working directory of each Spark executor.

For more advanced configuration options, visit the Apache 2 reference documentation.

**Related Information**

- Spark Configuration
- LOG4J Configuration
- Natural Language Toolkit
- Making Python on Apache Hadoop Easier with Anaconda and CDH

**Spark Log4j Configuration**

Cloudera Machine Learning allows you to update Spark’s internal logging configuration on a per-project basis.

Spark 2 uses Apache Log4j, which can be configured through a properties file. By default, a log4j.properties file found in the root of your project will be appended to the existing Spark logging properties for every session and job. To specify a custom location, set the environmental variable LOG4J_CONFIG to the file location relative to your project.

The Log4j documentation has more details on logging options.

Increasing the log level or pushing logs to an alternate location for troublesome jobs can be very helpful for debugging. For example, this is a log4j.properties file in the root of a project that sets the logging level to INFO for Spark jobs.

shell.log.level=INFO
PySpark logging levels should be set as follows:

```
log4j.logger.org.apache.spark.api.python.PythonGatewayServer=<LOG_LEVEL>
```

And Scala logging levels should be set as:

```
log4j.logger.org.apache.spark.repl.Main=<LOG_LEVEL>
```

### Setting Up an HTTP Proxy for Spark 2

If you are using an HTTP proxy, you must set the Spark configuration parameter `extraJavaOptions` at runtime to be able to support web-related actions in Spark.

```
spark.driver.extraJavaOptions=\
-Dhttp.proxyHost=<YOUR HTTP PROXY HOST> \\n-Dhttp.proxyPort=<HTTP PORT> \\n-Dhttps.proxyHost=<YOUR HTTPS PROXY HOST> \\n-Dhttps.proxyPort=<HTTPS PORT>
```

### Spark Web UIs

This topic describes how to access Spark web UIs from the CML UI.

Spark 2 exposes one web UI for each Spark application driver running in Cloudera Machine Learning. The UI will be running within the container, on the port specified by the environmental variable `CDSW_SPARK_PORT`. By default, `CDSW_SPARK_PORT` is set to 20049. The web UI will exist only as long as a SparkContext is active within a session. The port is freed up when the SparkContext is shutdown.

Spark 2 web UIs are available in browsers at: https://spark-<CDSW_ENGINE_ID>.<CDSW_DOMAIN>. To access the UI while you are in an active session, click the grid icon in the upper right hand corner of the Cloudera Machine Learning web application, and select Spark UI from the dropdown. Alternatively, the Spark UI is also available as a tab in the session itself. For a job, navigate to the job overview page and click the History tab. Click on a job run to open the session output for the job.

### Using Spark 2 from Python

Cloudera Machine Learning supports using Spark 2 from Python via PySpark. This topic describes how to set up and test a PySpark project.

#### PySpark Environment Variables

The default Cloudera Machine Learning engine currently includes Python 2.7.17 and Python 3.6.9. To use PySpark with lambda functions that run within the CDH cluster, the Spark executors must have access to a matching version of Python. For many common operating systems, the default system Python will not match the minor release of Python included in Machine Learning.

To ensure that the Python versions match, Python can either be installed on every CDH host or made available per job run using Spark’s ability to distribute dependencies. Given the size of a typical isolated Python environment, Cloudera recommends installing Python 2.7 and 3.6 on the cluster if you are using PySpark with lambda functions.

You can install Python 2.7 and 3.6 on the cluster using any method and set the corresponding `PYSPARK_PYTHON` environment variable in your project. Cloudera Machine Learning includes a separate environment variable for Python 3 sessions called `PYSPARK3_PYTHON`. Python 2 sessions continue to use the default `PYSPARK_PYTHON` variable. This will allow you to run Python 2 and Python 3 sessions in parallel without either variable being overridden by the other.
Creating and Running a PySpark Project

To get started quickly, use the PySpark template project to create a new project. For instructions, see Create a Project from a Built-in Template.

To run a PySpark project, navigate to the project's overview page, open the workbench console and launch a Python session. For detailed instructions, see Native Workbench Console and Editor.

Testing a PySpark Project in Spark Local Mode

Spark's local mode is often useful for testing and debugging purposes. Use the following sample code snippet to start a PySpark session in local mode.

```python
from pyspark.sql import SparkSession
spark = SparkSession.builder
   .appName("LocalSparkSession")
   .master("local")
   .getOrCreate()

for more details, refer to the Spark documentation: Running Spark Application.

Related Information
Native Workbench Console and Editor

Example: Montecarlo Estimation

Within the template PySpark project, pi.py is a classic example that calculates Pi using the Montecarlo Estimation.

What follows is the full, annotated code sample that can be saved to the pi.py file.

```python
# # Estimating $\pi$
#
# This PySpark example shows you how to estimate $\pi$ in parallel
# using Monte Carlo integration.

from __future__ import print_function
import sys
from random import random
from operator import add
# Connect to Spark by creating a Spark session
from pyspark.sql import SparkSession
spark = SparkSession
   .appName("PythonPi")
   .getOrCreate()

partitions = int(sys.argv[1]) if len(sys.argv) > 1 else 2
n = 100000 * partitions

def f(_):
    x = random() * 2 - 1
    y = random() * 2 - 1
    return 1 if x ** 2 + y ** 2 < 1 else 0

# To access the associated SparkContext
count = spark.sparkContext.parallelize(range(1, n + 1), partitions).map(f)
    .reduce(add)
print("Pi is roughly %f" % (4.0 * count / n))

spark.stop()
```
Example: Locating and Adding JARs to Spark 2 Configuration

This example shows how to discover the location of JAR files installed with Spark 2, and add them to the Spark 2 configuration.

```python
# # Using Avro data
# # This example shows how to use a JAR file on the local filesystem on Spark on Yarn.

from __future__ import print_function
import os, sys
import os.path
from functools import reduce
from pyspark.sql import SparkSession
from pyspark.files import SparkFiles

# Add the data file to HDFS for consumption by the Spark executors.
!hdfs dfs -put resources/users.avro /tmp

# Find the example JARs provided by the Spark parcel. This parcel is available on both the driver, which runs in Cloudera Machine Learning, and the executors, which run on Yarn.
exampleDir = os.path.join(os.environ["SPARK_HOME"], "examples/jars")
exampleJars = [os.path.join(exampleDir, x) for x in os.listdir(exampleDir)]
# Add the Spark JARs to the Spark configuration to make them available for use.
spark = SparkSession.builder.
    .config("spark.jars", ",".join(exampleJars))
    .appName("AvroKeyInputFormat")
    .getOrCreate()
sc = spark.sparkContext

# Read the schema.
schema = open("resources/user.avsc").read()
conf = {"avro.schema.input.key": schema }
avro_rdd = sc.newAPIHadoopFile(
    "/tmp/users.avro", # This is an HDFS path!
    "org.apache.avro.mapreduce.AvroKeyInputFormat",
    "org.apache.avro.mapred.AvroKey",
    "org.apache.hadoop.io.NullWritable",
    keyConverter="org.apache.spark.examples.pythonconverters.AvroWrapperT oJavaConverter",
    conf=conf)

output = avro_rdd.map(lambda x: x[0]).collect()
for k in output:
    print(k)
spark.stop()
```

Using Spark 2 from R

R users can access Spark 2 using sparklyr. Although Cloudera does not ship or support sparklyr, we do recommend using sparklyr as the R interface for Cloudera Machine Learning.

Before you begin

The `spark_apply()` function requires the R Runtime environment to be pre-installed on your cluster. This will likely require intervention from your cluster administrator. For details, refer the RStudio documentation.
Procedure

1. Install the latest version of sparklyr:

   ```r
   install.packages("sparklyr")
   ```

2. Optionally, connect to a local or remote Spark 2 cluster:

   ```r
   ## Connecting to Spark 2
   # Connect to an existing Spark 2 cluster in YARN client mode using the
   # spark_connect function.
   library(sparklyr)
   system.time(sc <- spark_connect(master = "yarn-client"))
   # The returned Spark 2 connection (sc) provides a remote dplyr data source
   # to the Spark 2 cluster.
   ```

   For a complete example, see Importing Data into Cloudera Machine Learning.

Related Information

- sparklyr: R interface for Apache Spark
- sparklyr Requirements

Using Spark 2 from Scala

This topic describes how to set up a Scala project for CDS 2.x Powered by Apache Spark along with a few associated tasks. Cloudera Machine Learning provides an interface to the Spark 2 shell (v 2.0+) that works with Scala 2.11.

Unlike PySpark or Sparklyr, you can access a SparkContext assigned to the spark (SparkSession) and sc (SparkContext) objects on console startup, just as when using the Spark shell.

By default, the application name will be set to CML_<sessionID>, where sessionId is the id of the session running your Spark code. To customize this, set the spark.app.name property to the desired application name in a spark-defaults.conf file.

Pi.scala is a classic starting point for calculating Pi using the Montecarlo Estimation.

This is the full, annotated code sample.

```scala
//Calculate pi with Monte Carlo estimation
import scala.math.random
//make a very large unique set of 1 -> n
val partitions = 2
val n = math.min(100000L * partitions, Int.MaxValue).toInt
val xs = 1 until n

//split up n into the number of partitions we can use
val rdd = sc.parallelize(xs, partitions).setName("N values rdd")

//generate a random set of points within a 2x2 square
val sample = rdd.map { i =>
  val x = random * 2 - 1
  val y = random * 2 - 1
  (x, y)
}.setName("Random points rdd")

//points w/in the square also w/in the center circle of r=1
val inside = sample.filter { case (x, y) => (x * x + y * y < 1) }.setName("Random points inside circle")
val count = inside.count()

//Area(circle)/Area(square) = inside/n => pi=4*inside/n
```

87
println("Pi is roughly " + 4.0 * count / n)

Key points to note:

- import scala.math.random
  Importing included packages works just as in the shell, and need only be done once.
- Spark context (sc).
  You can access a SparkContext assigned to the variable sc on console startup.

val rdd = sc.parallelize(xs, partitions).setName("'N values rdd'")

Managing Dependencies for Spark 2 and Scala

This topic demonstrates how to manage dependencies on local and external files or packages.

Example: Read Files from the Cluster Local Filesystem

Use the following command in the terminal to read text from the local filesystem. The file must exist on all hosts, and the same path for the driver and executors. In this example you are reading the file ebay-xbox.csv.

sc.textFile("file:///tmp/ebay-xbox.csv")

Adding Remote Packages

External libraries are handled through line magics. Line magics in the Toree kernel are prefixed with %. You can use Apache Toree's AddDeps magic to add dependencies from Maven central. You must specify the company name, artifact ID, and version. To resolve any transitive dependencies, you must explicitly specify the --transitive flag.

%AddDeps org.scalaj scalaj-http_2.11 2.3.0
import scalaj.http._
response.body
response.code
response.headers
response.cookies

Adding Remote or Local Jars

You can use the AddJars magic to distribute local or remote JARs to the kernel and the cluster. Using the -f option ignores cached JARs and reloads.

%AddJar http://example.com/some_lib.jar -f
%AddJar file:/path/to/some/lib.jar

Using GPUs for Cloudera Machine Learning projects

A GPU is a specialized processor that can be used to accelerate highly parallelized computationally-intensive workloads. Because of their computational power, GPUs have been found to be particularly well-suited to deep learning workloads. Ideally, CPUs and GPUs should be used in tandem for data engineering and data science workloads. A typical machine learning workflow involves data preparation, model training, model scoring, and model fitting. You can use existing general-purpose CPUs for each stage of the workflow, and optionally accelerate the math-intensive steps with the selective application of special-purpose GPUs. For example, GPUs allow you to accelerate model fitting using frameworks such as Tensorflow, PyTorch, and Keras.
By enabling GPU support, data scientists can share GPU resources available on Cloudera Machine Learning workspaces. Users can request a specific number of GPU instances, up to the total number available, which are then allocated to the running session or job for the duration of the run.

Enabling GPUs on ML Workspaces

If you are using ML Runtimes, you must use the ML Runtimes version for your Python library Nvidia GPU Edition.

Note: Nvidia GPU Edition comes with CUDA 11.1 preinstalled.

If you are using a Legacy Engine, to enable GPU usage on Cloudera Machine Learning, select GPUs when you are provisioning the workspace. If your existing workspace does not have GPUs provisioned, contact your ML administrator to provision a new one for you. For instructions, see Provisioning ML Workspaces.

Important: Review your cloud service provider account limits

For example, AWS imposes certain default limits for AWS services, and you might not have default access to GPU instances at all. Make sure you review your account’s current usage status and resource limits before you start provisioning GPU resources for CML.

Related Information
Custom CUDA-capable Engine Image
Site Admins: Add the Custom CUDA Engine to your Cloudera Machine Learning Deployment
Project Admins: Enable the CUDA Engine for your Project
Testing GPU Setup

Using GPUs with Legacy Engines

To use GPUs with legacy engines, you must create a custom CUDA-capable engine image.

Custom CUDA-capable Engine Image

Note: Before proceeding with creating a custom CUDA-capable engine, the Administrator needs to install the Nvidia plugin.

The base engine image (docker.repository.cloudera.com/CML/engine:<version>) that ships with Cloudera Machine Learning will need to be extended with CUDA libraries to make it possible to use GPUs in jobs and sessions.

The following sample Dockerfile illustrates an engine on top of which machine learning frameworks such as Tensorflow and PyTorch can be used. This Dockerfile uses a deep learning library from NVIDIA called NVIDIA CUDA Deep Neural Network (cuDNN). For detailed information about compatibility between NVIDIA driver versions and CUDA, refer the cuDNN installation guide (prerequisites).

When creating the Dockerfile for the custom image, you must delete the Cloudera repository that is inaccessible because of the paywall by running the following:

```
RUN rm /etc/apt/sources.list.d/*
```

Make sure you also check with the machine learning framework that you intend to use in order to know which version of cuDNN is needed. As an example, Tensorflow's NVIDIA hardware and software requirements for GPU support are listed in the Tensorflow documentation here. Additionally, the Tensorflow version compatibility matrix for CUDA and cuDNN is documented here.

The following sample Dockerfile uses NVIDIA's official Dockerfiles for CUDA and cuDNN images.

cuda.Dockerfile

```
FROM docker.repository.cloudera.com/CML/engine:13-cml
```
RUN rm /etc/apt/sources.list.d/*
RUN apt-get update && apt-get install -y --no-install-recommends \
gnupg2 curl ca-certificates && \
curl -fsSL https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64/7fa2af80.pub | apt-key add - && \
echo "deb https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64 /" > /etc/apt/sources.list.d/cuda.list && \
echo "deb https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64 /" > /etc/apt/sources.list.d/nvidia-ml.list && \
apt-get purge --autoremove -y curl && \
rm -rf /var/lib/apt/lists/*

ENV CUDA_VERSION 10.1.243
LABEL com.nvidia.cuda.version="${CUDA_VERSION}"

ENV CUDA_PKG_VERSION 10-1=${CUDA_VERSION}-1
RUN apt-get update && apt-get install -y --no-install-recommends \
cuda-cudart=${CUDA_PKG_VERSION} && \
cuda-libraries=${CUDA_PKG_VERSION} && \
ln -s cuda-10.1 /usr/local/cuda && \
rm -rf /var/lib/apt/lists/*

RUN echo "/usr/local/cuda/lib64" >> /etc/ld.so.conf.d/cuda.conf && \
ldconfig

RUN echo "/usr/local/nvidia/lib" >> /etc/ld.so.conf.d/nvidia.conf && \
echo "/usr/local/nvidia/lib64" >> /etc/ld.so.conf.d/nvidia.conf

ENV PATH /usr/local/nvidia/bin:/usr/local/cuda/bin:${PATH}
ENV LD_LIBRARY_PATH /usr/local/nvidia/lib:/usr/local/cuda-10.2/targets/x86_64-linux/lib/

ENV CUDNN_VERSION 7.6.5.32
LABEL com.nvidia.cudnn.version="${CUDNN_VERSION}"

RUN apt-get update && apt-get install -y --no-install-recommends \
libcudnn7=${CUDNN_VERSION}-1+cuda10.1 && \
apt-mark hold libcudnn7 && \
rm -rf /var/lib/apt/lists/*

Use the following example command to build the custom engine image using the cuda.Dockerfile command:

docker build --network host -t <company-registry>/CML-cuda:13 . -f cuda.Dockerfile

Push this new engine image to a public Docker registry so that it can be made available for Cloudera Machine Learning workloads. For example:

docker push <company-registry>/CML-cuda:13

Site Admins: Add the Custom CUDA Engine to your Cloudera Machine Learning Deployment

After you create a custom CUDA-capable engine image, you must add the new engine to Cloudera Machine Learning.
About this task
You must have the Site Administrator role to perform this task.

Procedure
2. Click Admin.
3. Go to the Engines tab.
4. Under Engine Images, add the custom CUDA-capable engine image created in the previous step.
   This allows project administrators across the deployment to start using this engine in their jobs and sessions.
5. Site administrators can also set a limit on the maximum number of GPUs that can be allocated per session or job.
   From the Maximum GPUs per Session/Job dropdown, select the maximum number of GPUs that can be used by an engine.
6. Click Update.

Project Admins: Enable the CUDA Engine for your Project
You can make the CUDA-capable engine the default engine for workloads within a particular project.

Before you begin
You must be a Project administrator to specify the default engine used for workloads within a particular project.

Procedure
1. Navigate to your project's Overview page.
2. Click Settings.
3. Go to the Engines tab.
4. Under Engine Image, select the CUDA-capable engine image from the dropdown.

Testing GPU Setup
Use these code samples to test that your GPU setup works with several common deep learning libraries. The specific versions of libraries depend on the particular GPU used and the GPU driver version. You can use this testing for GPU setup using ML Runtimes or Legacy Engines.

1. Go to a project that is using the CUDA engine and click Open Workbench.
2. Launch a new session with GPUs.
3. Run the following command in the workbench command prompt to verify that the driver was installed correctly:

   ```bash
   ! /usr/bin/nvidia-smi
   ```
4. Use any of the following code samples to confirm that the new engine works with common deep learning libraries.

Pytorch

```bash
!pip3 install torch==1.4.0
from torch import cuda
assert cuda.is_available()
assert cuda.device_count() > 0
print(cuda.get_device_name(cuda.current_device()))
```

Tensorflow

```bash
!pip3 install tensorflow-gpu==2.1.0
from tensorflow.python.client import device_lib
```
assert 'GPU' in str(device_lib.list_local_devices())
device_lib.list_local_devices()

Keras

!pip3 install keras
from keras import backend
assert len(backend.tensorflow_backend._get_available_gpus()) > 0
print(backend.tensorflow_backend._get_available_gpus())

Testing ML Runtime GPU Setup

You can use the following simple examples to test whether the new ML Runtime is able to leverage GPUs as expected.

1. Go to a project that is using the ML Runtimes Nvidia GPU edition and click Open Workbench.
2. Launch a new session with GPUs.
3. Run the following command in the workbench command prompt to verify that the driver was installed correctly:

   ! /usr/bin/nvidia-smi

4. Use any of the following code samples to confirm that the new engine works with common deep learning libraries.

Pytorch

!pip3 install torch==1.4.0
from torch import cuda
assert cuda.is_available()
assert cuda.device_count() > 0
print(cuda.get_device_name(cuda.current_device()))

Tensorflow

!pip3 install tensorflow-gpu==2.1.0
from tensorflow.python.client import device_lib
assert 'GPU' in str(device_lib.list_local_devices())
device_lib.list_local_devices()

Keras

!pip3 install keras
from keras import backend
assert len(backend.tensorflow_backend._get_available_gpus()) > 0
print(backend.tensorflow_backend._get_available_gpus())

Running an Experiment (QuickStart)

This topic walks you through a simple example to help you get started with experiments in Cloudera Machine Learning.

The following steps describe how to launch an experiment from the Workbench console. In this example we are going to run a simple script that adds all the numbers passed as arguments to the experiment.

1. Go to the project Overview page.
2. Click Open Workbench.
3. Create/modify any project code as needed. You can also launch a session to simultaneously test code changes on the interactive console as you launch new experiments.

As an example, you can run this Python script that accepts a series of numbers as command-line arguments and prints their sum.

```
import sys
import cdsw

args = len(sys.argv) - 1
sum = 0
x = 1

while (args >= x):
    print ("Argument %i: %s" % (x, sys.argv[x]))
    sum = sum + int(sys.argv[x])
    x = x + 1

print ("Sum of the numbers is: %i." % sum)
```

To test the script, launch a Python session and run the following command from the workbench command prompt:

```
!python add.py 1 2 3 4
```
4. Click **Run Experiment**. If you're already in an active session, click **Run > Run Experiment**. Fill out the following fields:

- **Script** - Select the file that will be executed for this experiment.
- **Arguments** - If your script requires any command line arguments, enter them here.
  
  **Note:** Arguments are not supported with Scala experiments.

- **Engine Kernel and Resource Profile** - Select the kernel and computing resources needed for this experiment.
  
  **Note:** The list of options here is specific to the default engine you have specified in your Project Settings: ML Runtimes or Legacy Engines. Engines allow kernel selection, while ML Runtimes allow Editor, Kernel, Variant, and Version selection. Resource Profile list is applicable for both ML Runtimes and Legacy Engines.

For this example we will run the `add.py` script and pass some numbers as arguments.
5. Click **Start Run**.
6. To track progress for the run, go back to the project Overview. On the left navigation bar click Experiments. You should see the experiment you've just run at the top of the list. Click on the Run ID to view an overview for each individual run. Then click Build.

On this Build tab you can see realtime progress as Cloudera Machine Learning builds the Docker image for this experiment. This allows you to debug any errors that might occur during the build stage.

7. Once the Docker image is ready, the run will begin execution. You can track progress for this stage by going to the Session tab.

For example, the Session pane output from running add.py is:

![Session output]

8. (Optional) The cdsw library that is bundled with Cloudera Machine Learning includes some built-in functions that you can use to compare experiments and save any files from your experiments.

For example, to track the sum for each run, add the following line to the end of the add.py script.

```python
cdsw.track_metric("Sum", sum)
```

This will be tracked in the Experiments table:
Limitations

This topic lists some of the known issues and limitations associated with experiments.

- Experiments do not store snapshots of project files. You cannot automatically restore code that was run as part of an experiment.
- Experiments will fail if your project filesystem is too large for the Git snapshot process. As a general rule, any project files (code, generated model artifacts, dependencies, etc.) larger than 50 MB must be part of your project's .gitignore file so that they are not included in snapshots for experiment builds.
- Experiments cannot be deleted. As a result, be conscious of how you use the track_metrics and track_file functions.
- Do not track files larger than 50MB.
- Do not track more than 100 metrics per experiment. Excessive metric calls from an experiment may cause Cloudera Machine Learning to stop responding.
- The Experiments table will allow you to display only three metrics at a time. You can select which metrics are displayed from the metrics dropdown. If you are tracking a large number of metrics (100 or more), you might notice some performance lag in the UI.
- Arguments are not supported with Scala experiments.
- The track_metrics and track_file functions are not supported with Scala experiments.
- The UI does not display a confirmation when you start an experiment or any alerts when experiments fail.

Related Information

Engines for Experiments and Models

Tracking Metrics

This topic teaches you how to use the track_metric function to log metrics associated with experiments.

The cdsw library includes a track_metric function that can be used to log up to 50 metrics associated with a run, thus allowing accuracy and scores to be tracked over time.

The function accepts input in the form of key value pairs.

```
cdsw.track_metric(key, value)
```

Python

```
cdsw.track_metric("R_squared", 0.79)
```

R

```
cdsw::track.metric("R_squared", 0.62)
```

These metrics will be available on the project's Experiments tab where you can view, sort, and filter experiments on the values. The table on the Experiments page will allow you to display only three metrics at a time. You can select which metrics are displayed from the metrics dropdown.

Note: This function is not supported with Scala experiments.
Saving Files

This topic teaches you how to use the `track_file` function to save files associated with experiments.

Cloudera Machine Learning allows you to select which artifacts you'd like to access and evaluate after an experiment is complete. These artifacts could be anything from a text file to an image or a model that you have built through the run.

The `cdsw` library includes a `track_file` function that can be used to specify which artifacts should be retained after the experiment is complete.

**Python**

```python
cdsw.track_file('model.pkl')
```

**R**

```r
cdsw::track.file('model.pkl')
```

Specified artifacts can be accessed from the run's Overview page. These files can also be saved to the top-level project filesystem and downloaded from there.

**Note:** This function is not supported with Scala experiments.

---

Debugging Issues with Experiments

This topic lists some common issues to watch out for during an experiment's build and execution process:

**Experiment spends too long in Scheduling/Built stage**

If your experiments are spending too long in any particular stage, check the resource consumption statistics for the cluster. When the cluster starts to run out of resources, often experiments (and other entities like jobs, models) will spend too long in the queue before they can be executed.

Resource consumption by experiments (and jobs, sessions) can be tracked by site administrators on the **Admin > Activity** page.

**Experiment fails in the Build stage**

During the build stage Cloudera Machine Learning creates a new Docker image for the experiment. You can track progress for this stage on each experiment's **Build** page. The build logs on this page should help point you in the right direction.

Common issues that might cause failures at this stage include:

- Lack of execute permissions on the build script itself.
- Inability to reach the Python package index or R mirror when installing packages.
- Typo in the name of the build script (`cdsw-build.sh`). Note that the build process will only run a script called `cdsw-build.sh`; not any other bash scripts from your project.
- Using `pip3` to install packages in `cdsw-build.sh`, but selecting a Python 2 kernel when you actually launch the experiment. Or vice versa.

**Experiment fails in the Execute stage**

Each experiment includes a **Session** page where you can track the output of the experiment as it executes. This is similar to the output you would see if you test the experiment in the workbench console. Any runtime errors will display on the **Session** page just as they would in an interactive session.
Creating and Deploying a Model

This topic describes a simple example of how to create and deploy a model using Cloudera Machine Learning.

Using Cloudera Machine Learning, you can create any function within a script and deploy it to a REST API. In a machine learning project, this will typically be a predict function that will accept an input and return a prediction based on the model's parameters.

For the purpose of this quick start demo we are going to create a very simple function that adds two numbers and deploy it as a model that returns the sum of the numbers. This function will accept two numbers in JSON format as input and return the sum.

1. Create a new project. Note that models are always created within the context of a project.
2. Click **Open Workbench** and launch a new Python 3 session.
3. Create a new file within the project called `add_numbers.py`. This is the file where we define the function that will be called when the model is run. For example:

   ```python
   def add(args):
       result = args["a"] + args["b"]
       return result
   ```

   **Note:** In practice, do not assume that users calling the model will provide input in the correct format or enter good values. Always perform input validation.

4. Before deploying the model, test it by running the `add_numbers.py` script, and then calling the `add` function directly from the interactive workbench session. For example:

   ```python
   add(\"a\": 3, \"b\": 5)  
   ```
How To

5. Deploy the `add` function to a REST endpoint.
   a. Go to the project Overview page.
   b. Click Models > New Model.
   c. Give the model a Name and Description.
   d. Enter details about the model that you want to build. In this case:
      - File: add_numbers.py
      - Function: add
      - Example Input: `{"a": 3, "b": 5}`
      - Example Output: 8
   e. Select the resources needed to run this model, including any replicas for load balancing.
      
      **Note:** The list of options here is specific to the default engine you have specified in your Project Settings: ML Runtimes or Legacy Engines. Engines allow kernel selection, while ML Runtimes allow Editor, Kernel, Variant, and Version selection. Resource Profile list is applicable for both ML Runtimes and Legacy Engines.
   f. Click Deploy Model.
6. Click on the model to go to its Overview page. Click Builds to track realtime progress as the model is built and deployed. This process essentially creates a Docker container where the model will live and serve requests.

![Model Overview](image)

7. Once the model has been deployed, go back to the model Overview page and use the Test Model widget to make sure the model works as expected.

If you entered example input when creating the model, the Input field will be pre-populated with those values. Click Test. The result returned includes the output response from the model, as well as the ID of the replica that served the request.

Model response times depend largely on your model code. That is, how long it takes the model function to perform the computation needed to return a prediction. It is worth noting that model replicas can only process one request at a time. Concurrent requests will be queued until the model can process them.

### Usage Guidelines

This section calls out some important guidelines you should keep in mind when you start deploying models with Cloudera Machine Learning.

#### Model Code

Models in Cloudera Machine Learning are designed to run any code that is wrapped into a function. This means you can potentially deploy a model that returns the result of a `SELECT *` query on a very large table. However, Cloudera strongly recommends against using the models feature for such use cases.

As a best practice, your models should be returning simple JSON responses in near-real time speeds (within a fraction of a second). If you have a long-running operation that requires extensive computing and takes more than 15 seconds to complete, consider using batch jobs instead.

#### Model Artifacts

Once you start building larger models, make sure you are storing these model artifacts in HDFS, S3, or any other external storage. Do not use the project filesystem to store large output artifacts.

In general, any project files larger than 50 MB must be part of your project's .gitignore file so that they are not included in Engines for Experiments and Models for future experiments/model builds. Note that in case your models require resources that are stored outside the model itself, it is up to you to ensure that these resources are available and immutable as model replicas may be restarted at any time.

#### Resource Consumption and Scaling
Models should be treated as any other long-running applications that are continuously consuming memory and computing resources. If you are unsure about your resource requirements when you first deploy the model, start with a single replica, monitor its usage, and scale as needed.

If you notice that your models are getting stuck in various stages of the deployment process, check the Monitoring Active Models page to make sure that the cluster has sufficient resources to complete the deployment operation.

**Security Considerations**

As stated previously, models do not impose any limitations on the code they can run. Additionally, models run with the permissions of the user that creates the model (same as sessions and jobs). Therefore, be conscious of potential data leaks especially when querying underlying data sets to serve predictions.

Cloudera Machine Learning models are not public by default. Each model has an access key associated with it. Only users/applications who have this key can make calls to the model. Be careful with who has permission to view this key.

Cloudera Machine Learning also prints `stderr/stdout` logs from models to an output pane in the UI. Make sure you are not writing any sensitive information to these logs.

**Deployment Considerations**

Cloudera Machine Learning does not currently support high availability for models. Additionally, there can only be one active deployment per model at any given time. This means you should plan for model downtime if you want to deploy a new build of the model or re-deploy with more/less replicas.

Keep in mind that models that have been developed and trained using Cloudera Machine Learning are essentially Python/R code that can easily be persisted and exported to external environments using popular serialization formats such as Pickle, PMML, ONNX, and so on.

**Related Information**

- Engines for Experiments and Models
- Monitoring Active Models

**Known Issues and Limitations**

- **Known Issues with Model Builds and Deployed Models**
  - Re-deploying or re-building models results in model downtime (usually brief).
  - Re-starting Cloudera Machine Learning does not automatically restart active models. These models must be manually restarted so they can serve requests again.

  Cloudera Bug: DSE-4950

- **Model deployment will fail if your project filesystem is too large for the Git Engines for Experiments and Models process.** As a general rule, any project files (code, generated model artifacts, dependencies, etc.) larger than...
than 50 MB must be part of your project's .gitignore file so that they are not included in snapshots for model builds.

- Model builds will fail if your project filesystem includes a .git directory (likely hidden or nested). Typical build stage errors include:

  Error: 2 UNKNOWN: Unable to schedule build: [Unable to create a checkpoint of current source: [Unable to push sources to git server: ...

To work around this, rename the .git directory (for example, NO.git) and re-build the model.

Cloudera Bug: DSE-4657

- JSON requests made to active models should not be more than 5 MB in size. This is because JSON is not suitable for very large requests and has high overhead for binary objects such as images or video. Call the model with a reference to the image or video, such as a URL, instead of the object itself.

- Any external connections, for example, a database connection or a Spark context, must be managed by the model's code. Models that require such connections are responsible for their own setup, teardown, and refresh.

- Model logs and statistics are only preserved so long as the individual replica is active. Cloudera Machine Learning may restart a replica at any time it is deemed necessary (such as bad input to the model).

- (MLLib) The MLLib model.save() function fails with the following sample error. This occurs because the Spark executors on CML all share a mount of /home/cdsw which results in a race condition as multiple executors attempt to write to it at the same time.

  Caused by:
  java.io.IOException: Mkdirs failed to create file:/home/cdsw/model.mllib/metadata/_temporary ....

  Recommended workarounds:
  • Save the model to /tmp, then move it into /home/cdsw on the driver/session.
  • Save the model to either an S3 URL or any other explicit external URL.

- Limitations
  • Scala models are not supported.
  • Spawning worker threads are not supported with models.
  • Models deployed using Cloudera Machine Learning are highly available subject to the following limitations:
    • Model high availability is dependent on the high availability of the cloud provider's Kubernetes service. Please refer to your chosen cloud provider for precise SLAs.
    • Model high availability is dependent on the high availability of the cloud provider's load balancer service. Please refer to your chosen cloud provider for precise SLAs.
    • In the event that the Kubernetes pod running the model proxy service becomes unavailable, the Model may be unavailable for multiple seconds during failover.
    • Dynamic scaling and auto-scaling are not currently supported. To change the number of replicas in service, you will have to re-deploy the build.

  Related Information
  Engines for Experiments and Models
  Distributed Computing with Workers

Model Request and Response Formats

Every model function in Cloudera Machine Learning takes a single argument in the form of a JSON-encoded object, and returns another JSON-encoded object as output. This format ensures compatibility with any application accessing the model using the API, and gives you the flexibility to define how JSON data types map to your model's datatypes.

Model Requests
When making calls to a model, keep in mind that JSON is not suitable for very large requests and has high overhead for binary objects such as images or video. Consider calling the model with a reference to the image or video such as a URL instead of the object itself. Requests to models should not be more than 5 MB in size. Performance may degrade and memory usage increase for larger requests.

Ensure that the JSON request represents all objects in the request or response of a model call. For example, JSON does not natively support dates. In such cases consider passing dates as strings, for example in ISO-8601 format, instead.

For a simple example of how to pass JSON arguments to the model function and make calls to deployed model, see *Creating and Deploying a Model*.

**Model Responses**

Models return responses in the form of a JSON-encoded object. Model response times depend on how long it takes the model function to perform the computation needed to return a prediction. Model replicas can only process one request at a time. Concurrent requests are queued until a replica is available to process them.

When Cloudera Machine Learning receives a call request for a model, it attempts to find a free replica that can answer the call. If the first arbitrarily selected replica is busy, Cloudera Machine Learning will keep trying to contact a free replica for 30 seconds. If no replica is available, Cloudera Machine Learning will return a `model.busy` error with HTTP status code 429 (Too Many Requests). If you see such errors, re-deploy the model build with a higher number of replicas.

**Related Information**

- Creating and Deploying a Model
- Workflows for Active Models

**Testing Calls to a Model**

Cloudera Machine Learning provides two ways to test calls to a model:
• Test Model Widget

On each model's Overview page, Cloudera Machine Learning provides a widget that makes a sample call to the deployed model to ensure it is receiving input and returning results as expected.

```json
{
  "a": 3,
  "b": 5
}
```

![Test Model](image)

**Result**

<table>
<thead>
<tr>
<th>Status</th>
<th>✔ success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>8</td>
</tr>
<tr>
<td>Replica ID</td>
<td>add-two-numbers-1-1-86b9b58b7b-g6s8r</td>
</tr>
</tbody>
</table>

• Sample Request Strings

On the model Overview page, Cloudera Machine Learning also provides sample curl and POST request strings that you can use to test calls to the model. Copy/paste the curl request directly into a Terminal to test the call.

Note that these sample requests already include the example input values you entered while building the model, and the access key required to query the model.

```
curl -H "Content-Type: application/json" -X POST http://cloudera.com/api/altus-ds-1/models/call-model -d '{"accessKey":"mfgflyfjc5esdy8t6jyold9cfzwq_5678","request":{"a":1,"b":2}}'
```

**Securing Models**
Access Keys for Models

Each model in Cloudera Machine Learning has a unique access key associated with it. This access key is a unique identifier for the model.

Models deployed using Cloudera Machine Learning are not public. In order to call an active model your request must include the model's access key for authentication (as demonstrated in the sample calls above).

To locate the access key for a model, go to the model Overview page and click Settings.

Important:

Only one access key per model is active at any time. If you regenerate the access key, you will need to re-distribute this access key to users/applications using the model.

Alternatively, you can use this mechanism to revoke access to a model by regenerating the access key. Anyone with an older version of the key will not be able to make calls to the model.

Model API Key for Models

You can prevent unauthorized access to your models by specifying a Model API Key in the “Authorization” header of your model HTTP request. This topic covers how to create, test, and use a Model API Key in Cloudera Machine Learning.

The Model API key governs the authentication part of the process and the authorization is based on what privileges the users already have in terms of the project that they is a part of. For example, if a user/application has read-only access to a project, then the authorization is based on their current access-level to the project, which is “read-only”. If the users have been authenticated to a project, then they can make a request to a model with the Model API Key. This is different to the previously described Access Key, which is only used to identify which model should serve a request.

Enabling authentication

Restricting access using Model API Keys is an optional feature. By default, the “Enable Authentication” option is turned on. However, it is turned off by default for the existing models for backward compatibility. You can enable authentication for all your existing models.

To enable authentication, go to Projects > Models > Settings and check the Enable Authentication option.

Note: It can take up to five minutes by the system to update.

Generating a Model API Key

If you have enabled authentication, then you need a Model API Key to call a model. If you are not a collaborator on a particular project, then you cannot access the models within that project using the Model API Key that you generate. You need to be added as a collaborator by the admin or the owner of the project to use the Model API Key to access a model.
About this task

There are two types of API keys used in Cloudera Machine Learning:

- **API Key**: This is used in the CDSW-specific internal APIs for CLI automation. This can’t be deleted and neither does it expire. This API Key is not required when sending requests to a model.
- **Model API Key**: These are used to authenticate requests to a model. You can choose the expiration period and delete them when no longer needed.

You can generate more than one Model API Keys to use with your model, depending on the number of clients that you are using to call the models.

Procedure

2. Click **Settings** from the left navigation pane.
3. On the **User Settings** page, click the **API Keys** tab.
4. Select an expiry date for the Model API Key, and click **Create API keys**.
   An API key is generated along with a Key ID.
   If you do not specify an expiry date, then the generated key is active for one year from the current date, or for the duration set by the Administrator. If you specify an expiration date that exceeds the duration value set by the Administrator, you will get an error. The Administrator can set the default duration value at **Admin > Security > Default Model API keys expiration in days**
   
   **Note:**
   - The Model API Key is private and ephemeral. Copy the key and the corresponding key ID on to a secure location for future use before refreshing or leaving the page. If you miss storing the key, then you can generate another key.
   - You can delete the Model API Keys that have expired or no longer in use. It can take up to five minutes by the system to take effect.
5. To test the Model API Key:
   a) Navigate to your project and click **Models** from the left navigation pane.
   b) On the **Overview** page, paste the Model API Key in the Model API Key field that you had generated in the previous step and click **Test**.
      The test results, along with the HTTP response code and the Replica ID are displayed in the Results table.
      If the test fails and you see the following message, then you must get added as a collaborator on the respective project by the admin or the creator of the project:

      "User APIkey not authorized to access model": "Check APIKEY permissions or model authentication permissions"

Managing Model API Keys

The admin user can access the list of all the users who are accessing the workspace and can delete the Model API Keys for a user.

About this task

To manage users and their keys:

Procedure

1. Sign in to Cloudera Machine Learning as an admin user.
2. From the left navigation pane, click **Admin**.
   The **Site Administration** page is displayed.
3. On the **Site Administration** page, click on the **Users** tab.
   All the users signed under this workspace are displayed.
   The **Model API Keys** column displays the number of Model API Keys granted to a user.
4. To delete a Model API Key for a particular user:
   a) Select the user for which you want to delete the Model API Key.
      A page containing the user’s information is displayed.
   b) To delete a key, click **Delete** under the Action column corresponding to the Key ID.
   c) Click **Delete all keys** to delete all the keys for that user.

   **Note:** It can take up to five minutes by the system to take effect.

As a non-admin user, you can delete your own Model API Key by navigating to **Settings > User Settings > API Keys**.

### Workflows for Active Models

This topic walks you through some nuances between the different workflows available for re-deploying and re-building models.

**Active Model** - A model that is in the Deploying, Deployed, or Stopping stages.

You can make changes to a model even after it has been deployed and is actively serving requests. Depending on business factors and changing resource requirements, such changes will likely range from changes to the model code itself, to simply modifying the number of CPU/GPUs requested for the model. In addition, you can also stop and restart active models.

Depending on your requirement, you can perform one of the following actions:

#### Re-deploy an Existing Build

Re-deploying a model involves re-publishing a previously-deployed model in a new serving environment - this is, with an updated number of replicas or memory/CPU/GPU allocation. For example, circumstances that require a re-deployment might include:

- An active model that previously requested a large number of CPUs/GPUs that are not being used efficiently.
- An active model that is dropping requests because it is falling short of replicas.
- An active model needs to be rolled back to one of its previous versions.

**Warning:** Currently, Cloudera Machine Learning only allows one active deployment per model. This means when you re-deploy a build, the current active deployment will go offline until the re-deployment process is complete and the new deployment is ready to receive requests. Prepare for model downtime accordingly.

To re-deploy an existing model:
1. Go to the model **Overview page**.
2. Click **Deployments**.
3. Select the version you want to deploy and click **Re-deploy this Build**.

   **Note:**

<table>
<thead>
<tr>
<th>Add Two Numbers</th>
<th>Deployments</th>
<th>Builds</th>
<th>Monitoring</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>Deployments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id</td>
<td>Build</td>
<td>Status</td>
<td>Deployed At</td>
<td>Stopped At</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Deployed</td>
<td>Jun 7, 2018, 1:49 PM</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Stopped</td>
<td>Jun 5, 2018, 11:16 AM</td>
<td>Jun 7, 2018, 1:46 PM</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Stopped</td>
<td>Jun 5, 2018, 6:45 PM</td>
<td>Jun 7, 2018, 11:13 AM</td>
</tr>
</tbody>
</table>

4. Modify the model serving environment as needed.

5. Click **Deploy Model**.

### Deploy a New Build for a Model

Deploying a new build for a model involves both, re-building the Docker image for the model, and deploying this new build. Note that this is not required if you only need to update the resources allocated to the model. As an example, changes that require a new build might include:

- Code changes to the model implementation.
- Renaming the function that is used to invoke the model.

**Warning:** Currently, Cloudera Machine Learning does not allow you to create a new build for a model without also deploying it. This combined with the fact that you can only have one active deployment per model means that once the new model is built, the current active deployment will go offline so that the new build can be deployed. Prepare for model downtime accordingly.

To create a new build and deploy it:

1. Go to the model Overview page.
2. Click **Deploy New Build**.

<table>
<thead>
<tr>
<th>Add Two Numbers</th>
<th>Deployments</th>
<th>Builds</th>
<th>Monitoring</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>Deployments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Code</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| curl -H 'Content-Type: application/json' -X POST http:="" xg.cloudera.com/api/altus-ds-1/models/call-model -d '{"accessKey..."}"

3. Complete the form and click **Deploy Model**.

### Stop a Model

To stop a model (all replicas), go to the model **Overview** page and click **Stop**. Click **OK** to confirm.

### Restart a Model

To restart a model (all replicas), go to the model **Overview** page and click **Restart**. Click **OK** to confirm.

Restarting a model does not let you make any code changes to the model. It should primarily be used as a way to quickly re-initialize or re-connect to resources.
**Monitoring Active Models**

This topic describes how to monitor your active models and access any logs related to them.

**What is an Active Model?**

A model that is in the Deploying, Deployed, or Stopping stages is referred to as an active model.

Cloudera Machine Learning provides two ways to monitor active models:

**Monitoring Individual Models**

When a model is deployed, Cloudera Machine Learning allows you to specify a number of replicas that will be deployed to serve requests. For each active model, you can monitor its replicas by going to the model's Monitoring page. On this page you can track the number of requests being served by each replica, success and failure rates, and their associated stderr and stdout logs. Depending on future resource requirements, you can increase or decrease the number of replicas by re-deploying the model.

The most recent logs are at the top of the pane (see image). stderr logs are displayed next to a red bar while stdout logs are by a green bar. Note that model logs and statistics are only preserved so long as the individual replica is active. When a replica restarts (for example, in case of bad input) the logs also start with a clean slate.

### Related Information

- **Workflows for Active Models**

**Debugging Issues with Models**

This topic describes some common issues to watch out for during different stages of the model build and deployment process.

As a general rule, if your model spends too long in any of the afore-mentioned stages, check the resource consumption statistics for the cluster. When the cluster starts to run out of resources, often models will spend some time in a queue before they can be executed.

Resource consumption by active models on a deployment can be tracked by site administrators on the Admin > Models page.

**Building**

Live progress for this stage can be tracked on the model's Build tab. It shows the details of the build process that creates a new Docker image for the model. Potential issues:

- If you specified a custom build script (cdsw-build.sh), ensure that the commands inside the script complete successfully.
• If you are in an environment with restricted network connectivity, you might need to manually upload dependencies to your project and install them from local files.

Pushing
Once the model has been built, it is copied to an internal Docker registry to make it available to all the Cloudera Machine Learning hosts. Depending on network speeds, your model may spend some time in this stage.

Deploying
If you see issues occurring when Cloudera Machine Learning is attempting to start the model, use the following guidelines to begin troubleshooting:

• Make sure your model code works in a workbench session. To do this, launch a new session, run your model file, and then interactively call your target function with the input object. For a simple example, see the Creating and Deploying a Model.
• Ensure that you do not have any syntax errors. For Python, make sure you have the kernel with the appropriate Python version (Python 2 or Python 3) selected for the syntax you have used.
• Make sure that your cdsw-build.sh file provides a complete set of dependencies. Dependencies manually installed during a session on the workbench are not carried over to your model. This is to ensure a clean, isolated, build for each model.
• If your model accesses resources such as data on the CDH cluster or an external database make sure that those resources can accept the load your model may exert on them.

Deployed
Once a model is up and running, you can track some basic logs and statistics on the model's Monitoring page. In case issues arise:

• Check that you are handling bad input from users. If your function throws an exception, Cloudera Machine Learning will restart your model to attempt to get back to a known good state. The user will see an unexpected model shutdown error.
  
  For most transient issues, model replicas will respond by restarting on their own before they actually crash. This auto-restart behavior should help keep the model online as you attempt to debug runtime issues.
• Make runtime troubleshooting easier by printing errors and output to stderr and stdout. You can catch these on each model's Monitoring tab. Be careful not to log sensitive data here.
• The Monitoring tab also displays the status of each replica and will show if the replica cannot be scheduled due to a lack of cluster resources. It will also display how many requests have been served/dropped by each replica.

Related Information
Engines for Experiments and Models
Creating and Deploying a Model
Monitoring Active Models

Deleting a Model

Before you begin

Important:

• You must stop all active deployments before you delete a model. If not stopped, active models will continue serving requests and consuming resources even though they do not show up in Cloudera Machine Learning UI.
• Deleted models are not actually removed from disk. That is, this operation will not free up storage space.
Procedure

1. Go to the model Overview > Settings.
2. Click Delete Model.

Deleting a model removes all of the model's builds and its deployment history from Cloudera Machine Learning.

You can also delete specific builds from a model's history by going to the model's Overview > Build page.

Example - Model Training and Deployment (Iris)

This topic uses Cloudera Machine Learning's built-in Python template project to walk you through an end-to-end example where we use experiments to develop and train a model, and then deploy it using Cloudera Machine Learning.

This example uses the canonical Iris dataset from Fisher and Anderson to build a model that predicts the width of a flower’s petal based on the petal's length.

The scripts for this example are available in the Python template project that ships with Cloudera Machine Learning. First, create a new project from the Python template:

Once you've created the project, go to the project's Files page. The following files are used for the demo:
• `cdsw-build.sh` - A custom build script used for models and experiments. Pip installs our dependencies, primarily the `scikit-learn` library.

• `fit.py` - A model training example to be run as an experiment. Generates the `model.pkl` file that contains the fitted parameters of our model.

• `predict.py` - A sample function to be deployed as a model. Uses `model.pkl` produced by `fit.py` to make predictions about petal width.

**Related Information**

**Engines for Experiments and Models**

**Train the Model**

This topic shows you how to run experiments and develop a model using the `fit.py` file.

**About this task**

The `fit.py` script tracks metrics, mean squared error (MSE) and R², to help compare the results of different experiments. It also writes the fitted model to a `model.pkl` file.

**Procedure**

1. Navigate to the Iris project's Overview > Experiments page.
2. Click Run Experiment.
3. Fill out the form as follows and click Start Run. Make sure you use the Python 3 kernel.
4. The new experiment should now show up on the Experiments table. Click on the Run ID to go to the experiment's Overview page. The Build and Session tabs display realtime progress as the experiment builds and executes.

5. Once the experiment has completed successfully, go back to its Overview page. The tracked metrics show us that our test set had an MSE of ~0.0078 and an R² of ~0.0493. For the purpose of this demo, let's consider this an accurate enough model to deploy and use for predictions.

6. Once you have finished training and comparing metrics from different experiments, go to the experiment that generated the best model. From the experiment's Overview page, select the model.pkl file and click Add to Project. This saves the model to the project filesystem, available on the project's Files page. We will now deploy this model as a REST API that can serve predictions.

**Deploy the Model**

This topic shows you how to deploy the model using the predict.py script from the Python template project.

**About this task**

The predict.py script contains the predict function that accepts petal length as input and uses the model built in the previous step to predict petal width.

**Procedure**

1. Navigate to the Iris project's Overview > Models page.
2. Click **New Model** and fill out the fields. Make sure you use the Python 3 kernel. For example:
Create a Model

General

Name *
Predict Petal Width

Description *
This model uses petal length to predict petal width.

Build

File *
predict.py

Function *
predict

Example Input ️
```python
{
    "petal_length": 5.4
}
```

Example Output ️
```json
{ "result": "value" }
```

Kernel

- Python 2
- Python 3
- R

Comment
Using Python 3 for this build

Deployment

Engine Profile
1 vCPU / 2 GiB Memory

Replicas
3

Set Environmental Variables

Deploy Model  Cancel
3. Deploy the model.
4. Click on the model to go to its **Overview** page. As the model builds you can track progress on the **Build** page. Once deployed, you can see the replicas deployed on the **Monitoring** page.
5. To test the model, use the Test Model widget on the model’s **Overview** page.

### Enabling model governance

You must enable governance to capture and view information about your ML projects, models, and builds centrally from Apache Atlas (Data Catalog) for a given environment. If you do not select this option while provisioning workspaces, then integration with Atlas won’t work.

#### About this task

**Procedure**

1. Go to Cloudera Machine Learning and click **Provision Workspace** on the top-right corner.
2. Enter the workspace name and other details.
3. Click **Advanced Options**.
4. Select **Enable Governance**.

### Registering training data lineage using a linking file

The Machine Learning (ML) projects, model builds, model deployments, and associated metadata are tracked in Apache Atlas, which is available in the environment's SDX cluster. You can also specify additional metadata to be tracked for a given model build. For example, you can specify metadata that links training data to a project through a special file called the linking file (**lineage.yaml**).

The **lineage.yaml** file describes additional metadata and the lineage relationships between the project’s models and training data. You can use a single **lineage.yaml** file for all the models within the project.
Note: Your lineage file should be present in your project before you create a model build. The lineage file is parsed and metadata is attached during the model build process.

1. Create a YAML file in your ML project called lineage.yaml.

   If you have used a template to create your project, a lineage.yaml file should already exist in your project.

2. Insert statements in the file that describe the relationships you want to track between a model and the training data. You can include additional descriptive metadata through key-value pairs in a `metadata` section.

<table>
<thead>
<tr>
<th>YAML</th>
<th>YAML Structure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model name</td>
<td>Top-level entry</td>
<td>A ML model name associated with the current project. There can be more than one model per linking file.</td>
</tr>
<tr>
<td>hive_table_qualified_names</td>
<td>Second-level entry</td>
<td>This pre-defined key introduces sequence items that list the names of Hive tables used as training data.</td>
</tr>
<tr>
<td>Table names</td>
<td>Sequence items</td>
<td>The qualified names of Hive tables used as training data enclosed in double quotation marks. Qualified names are of the format <code>db-name.table-name@cluster-name</code></td>
</tr>
<tr>
<td>metadata</td>
<td>Second-level entry</td>
<td>This pre-defined key introduces additional metadata to be included in the Atlas representation of the relationship between the model and the training data.</td>
</tr>
<tr>
<td>key: value</td>
<td>Third-level entries</td>
<td>Key-value pairs that describe information about how this data is used in the model. For example, consider including the query text that is used to extract training data or the name of the training file used.</td>
</tr>
</tbody>
</table>

The following example linking file shows entries for two models in your project: `modelName1` and `modelName2`:

```
modelName1:
  hive_table_qualified_names:
    - "db.table1@namespace"
    - "db.table2@ns"
  metadata:
    key1: value1
    key2: value2
    query: "select id, name from table"
    training_file: "fit.py"
modelName2:
  hive_table_qualified_names:
    - "db.table2@ns"
```

Viewing lineage for a model deployment in Atlas

You can view the lineage information for a particular model deployment and trace it back to the specific data that was used to train the model through the Atlas' Management Console.

Procedure

1. Navigate to Data Lake > Atlas from the Management Console.
2. Search for `ml_model_deployment`. Click the model deployment of your interest.
3. Click the **Lineage** tab to see a visualization of lineage information for the particular model deployment and trace it back to the specific data that was used to train the model.

   You can also search for a specific table, click through to its **Lineage** tab and see if the table has been used in any model deployments.

### Enabling model metrics

Metrics are used to track the performance of the models. When you enable model metrics while creating a workspace, the metrics are stored in a scalable metrics store. You can track individual model predictions and analyze metrics using custom code.

**About this task**

**Procedure**

1. Go to Cloudera Machine Learning and click **Provision Workspace** on the top-right corner.
2. Enter the workspace name and other details.
3. Click **Advanced Options**.
4. Select **Enable Model Metrics**.

   If you want to connect to an external (custom) Postgres database, then specify the details in the additional optional arguments that are displayed. If you do not specify these details, a managed Postgres database will be used to store the metrics.

### Tracking model metrics without deploying a model

Cloudera recommends that you develop and test model metrics in a workbench session before actually deploying the model. This workflow avoids the need to rebuild and redeploy a model to test every change.

Metrics tracked in this way are stored in a local, in-memory datastore instead of the metrics database, and are forgotten when the session exits. You can access these metrics in the same session using the regular metrics API in the cdsw.py file.

The following example demonstrates how to track metrics locally within a session, and use the `read_metrics` function to read the metrics in the same session by querying by the time window.

To try this feature in the local development mode, use the following files from the Python template project:

- `use_model_metrics.py`
- `predict_with_metrics.py`

The predict function from the `predict_with_metrics.py` file shown in the following example is similar to the function with the same name in the `predict.py` file. It takes input and returns output, and can be deployed as a model. But unlike the function in the `predict.py` file, the predict function from the `predict_with_metrics.py` file tracks mathematical metrics. These metrics can include information such as input, output, feature values, convergence metrics, and error estimates. In this simple example, only input and output are tracked. The function is equipped to track metrics by applying the decorator `cdsw.model_metrics`.

```python
@cdsw.model_metrics
def predict(args):
    # Track the input.
    cdsw.track_metric("input", args)

    # If this model involved features, ie transformations of the
    # raw input, they could be tracked as well.
    # cdsw.track_metric("feature_vars", {"a":1,"b":23})

    petal_length = float(args.get('petal_length'))
```

---
result = model.predict([[petal_length]])
# Track the output.
cdsw.track_metric("predict_result", result[0][0])
return result[0][0]

You can directly call this function in a workbench session, as shown in the following example:

```
predict(
    
"petal_length": 3
)
```

You can fetch the metrics from the local, in-memory datastore by using the regular metrics API. To fetch the metrics, set the dev keyword argument to True in the use_model_metrics.py file. You can query the metrics by model, model build, or model deployment using the variables cdsw.dev_model_crn and cdsw.dev_model_build_crn or cdsw.dev_model_deploy_crn respectively.

For example:

```
end_timestamp_ms=int(round(time.time() * 1000))
cdsw.read_metrics(model_deployment_crn=cdsw.dev_model_deployment_crn,
                   start_timestamp_ms=0,
                   end_timestamp_ms=end_timestamp_ms,
                   dev=True)
```

where CRN denotes Cloudera Resource Name, which is a unique identifier from CDP, analogous to Amazon's ARN.

### Tracking metrics for deployed models

When you have finished developing your metrics tracking code and the code that consumes the metrics, simply deploy the predict function from predict_with_metrics.py as a model. No code changes are necessary.

Calls to read_metrics, track_delayed_metrics, and track_aggregate_metrics need to be changed to take the CRN of the deployed model, build or deployment. These CRNs can be found in the model’s Overview page.

Calls to the call_model function also requires the model’s access key (model_access_key in use_model_metrics.py) from the model’s Settings page. If authentication has been enabled for the model (the default), a model API key for the user (model_api_token in use_model_metrics.py) is also required. This can be obtained from the user’s Settings page.

### Analytical Applications

This topic describes how to use an ML workspace to create long-running web applications.

#### About this task

This feature gives data scientists a way to create ML web applications/dashboards and easily share them with other business stakeholders. Applications can range from single visualizations embedded in reports, to rich dashboard solutions such as Tableau. They can be interactive or non-interactive.

Applications stand alongside other existing forms of workloads in CML (sessions, jobs, experiments, models). Like all other workloads, applications must be created within the scope of a project. Each application is launched within its own isolated engine. Additionally, like models, engines launched for applications do not time out automatically. They will run as long as the web application needs to be accessible by any users and must be stopped manually when needed.
Before you begin

Testing applications before you deploy

Before you deploy an application using the steps described here, make sure your application has been thoroughly tested. You can use sessions to develop, test, and debug your applications. You can test web apps by embedding them in sessions as described here: https://docs.cloudera.com/machine-learning/cloud/projects/topics/ml-embedded-web-apps.html.

Procedure

1. Go to a project's Overview page.
2. Click Applications.
3. Click New Application.
4. Fill out the following fields.
   - Name: Enter a name for the application.
   - Subdomain: Enter a subdomain that will be used to construct the URL for the web application. For example, if you use test-app as the subdomain, the application will be accessible at test-app.<ml-workspace-domain-name>.
     Subdomains should be valid DNS hostname characters: letters from a to z, digits from 0 to 9, and the hyphen.
   - Description: Enter a description for the application.
   - Script: Select a script that hosts a web application on either CDSW_READONLY_PORT or CDSW_APP_PORT. Applications running on either of these ports are available to any users with at least read access to the project. The Python template project includes an entry.py script that you can use to test this out.
     Note: CML does not prevent you from running an application that allows a read-only user (i.e. Viewers) to modify files belonging to the project. It is up to you to make the application truly read-only in terms of files, models, and other resources belonging to the project.
   - Engine Kernel and Resource Profile: Select the kernel and computing resources needed for this application.
     Note: The list of options here is specific to the default engine you have specified in your Project Settings: ML Runtimes or Legacy Engines. Engines allow kernel selection, while ML Runtimes allow Editor, Kernel, Variant, and Version selection. Resource Profile list is applicable for both ML Runtimes and Legacy Engines.
   - Set Environment Variables: Click Set Environment Variables, enter the name and value for the new application variable, and click Add.
     If there is a conflict between the project-level and application-level environment variables, the application-level environment variables override the project-level environment variables.
5. Click Create Application.

Results

In a few minutes, you should see the application status change to Running on the Applications page. Click on the name of the application to access the web application interface.

What to do next

You can Stop, Restart, or Delete an application from the Applications page.

If you want to make changes to an existing application, click Overview under the application name. Then go to the Settings tab to make any changes and update the application.

Securing Applications

You can provide access to Applications via either the CDSW_APP_PORT or the CDSW_READONLY_PORT. Any user with read or higher permissions to the project is able to access an application served through either port.
• Securing project resources

CML applications are accessible by any user with read-only or higher permissions to the project. The creator of the application is responsible for managing the level of permissions the application users have on the project through the application. CML does not actively prevent you from running an application that allows a read-only user (i.e. Viewers) to modify files belonging to the project.

• Public Applications

By default, authentication for applications is enforced on all ports and users cannot create public applications. If desired, the Admin user can allow users to create public applications that can be accessed by unauthenticated users.

To allow users to create public applications on an ML workspace:

1. As an Admin user, turn on the feature flag in Admin > Security by selecting Allow applications to be configured with unauthenticated access.
2. When creating a new application, select Enable Unauthenticated Access.
3. For an existing application, in Settings select Enable Unauthenticated Access.

To prevent all users from creating public applications, go to Admin > Security and deselect Allow applications to be configured with unauthenticated access. After one minute, all existing public applications stop being publicly accessible.

• Transparent Authentication

CML can pass user authentication to an Application, if the Application expects an authorized request. The October 2020 release (and earlier) accomplishes this authentication by setting the REMOTE_USER field of the HTTP header to the username. The November 2020 release (and after) uses the REMOTE_USER field for this task.

Limitations with Analytical Applications

This topic lists all the limitations associated with the Applications feature.

• Port availability

Cloudera Machine Learning exposes only 2 ports per project. Therefore, you can run a maximum of 2 web applications simultaneously, on these ports:

• CDSW_APP_PORT
• CDSW_READONLY_PORT

By default, third-party browser-based editors run on CDSW_APP_PORT. Therefore, for projects that are already using browser-based editors, you are left with only one other port to run applications on: CDSW_READONLY_PORT.

Creating a Job

This topic describes how to automate analytics workloads with a built-in job and pipeline scheduling system that supports real-time monitoring, job history, and email alerts.

About this task

A job automates the action of launching an engine, running a script, and tracking the results, all in one batch process. Jobs are created within the purview of a single project and can be configured to run on a recurring schedule. You can customize the engine environment for a job, set up email alerts for successful or failed job runs, and email the output of the job to yourself or a colleague.

Jobs are created within the scope of a project. When you create a job, you will be asked to select a script to run as part of the job, and create a schedule for when the job should run. Optionally, you can configure a job to be dependent on another existing job, thus creating a pipeline of tasks to be accomplished in a sequence. Note that the script files and any other job dependencies must exist within the scope of the same project.
Procedure

1. Navigate to the project for which you want to create a job.
2. On the left-hand sidebar, click Jobs.
3. Click New Job.
4. Enter a Name for the job.
5. In Script, select a script to run for this job by clicking on the folder icon. You will be able to select a script from a list of files that are already part of the project. To upload more files to the project, see Managing Project Files.
6. In Arguments, enter command-line arguments to provide to the script.
   This feature only works with R or Python engines.
7. Depending on the code you are running, select an Engine Kernel for the job from one of the following option:
   Python 3.
   The resources you can choose are dependent on the default engine you have chosen: ML Runtimes or Legacy Engines. For ML Runtimes, you can also choose a Kernel Edition and Version.
8. Select a Schedule for the job runs from one of the following options.
   - Manual - Select this option if you plan to run the job manually each time.
   - Recurring - Select this option if you want the job to run in a recurring pattern every X minutes, or on an hourly, daily, weekly or monthly schedule.
   - Dependent - Use this option when you are building a pipeline of jobs to run in a predefined sequence. From a dropdown list of existing jobs in this project, select the job that this one should depend on. Once you have configured a dependency, this job will run only after the preceding job in the pipeline has completed a successful run.
9. Select an Resource Profile to specify the number of cores and memory available for each session.
   Note: The list of options here is specific to the default engine you have specified in your Project Settings: ML Runtimes or Legacy Engines. Engines allow kernel selection, while ML Runtimes allow Editor, Kernel, Variant, and Version selection. Resource Profile list is applicable for both ML Runtimes and Legacy Engines.
10. Enter an optional timeout value in minutes.
11. Click Set environment variables if you want to set any values to override the overall project environment variables.
12. Specify a list of Job Report Recipients to whom you can send email notifications with detailed job reports for job success, failure, or timeout. You can send these reports to yourself, your team (if the project was created under a team account), or any other external email addresses.
13. Add any Attachments such as the console log to the job reports that will be emailed.
14. Click Create Job.
   You can use the Jobs API to schedule jobs from third partly workflow tools. For details, see Cloudera Machine Learning Jobs API.

Related Information
Managing Project Files
Cloudera Machine Learning Jobs API

Creating a Pipeline

This topic describes how to create a scheduled pipeline of jobs within a project.

About this task
As data science projects mature beyond ad hoc scripts, you might want to break them up into multiple steps. For example, a project may include one or more data acquisition, data cleansing, and finally, data analytics steps. For such
projects, Cloudera Machine Learning allows you to schedule multiple jobs to run one after another in what is called a pipeline, where each job is dependent on the output of the one preceding it.

The **Jobs** overview presents a list of all existing jobs created for a project along with a dependency graph to display any pipelines you’ve created. Job dependencies do not need to be configured at the time of job creation. Pipelines can be created after the fact by modifying the jobs to establish dependencies between them. From the job overview, you can modify the settings of a job, access the history of all job runs, and view the session output for individual job runs.

Let's take an example of a project that has two jobs, Read Weblogs and Write Weblogs. Given that you must read the data before you can run analyses and write to it, the Write Weblogs job should only be triggered after the Read Weblogs job completes a successful run. To create such a two-step pipeline:

**Procedure**

1. Navigate to the project where the Read Weblogs and Write Weblogs jobs were created.
2. Click **Jobs**.
3. From the list of jobs, select Write Weblogs.
4. Click the **Settings** tab.
5. Click on the Schedule dropdown and select Dependent. Select Read Weblogs from the dropdown list of existing jobs in the project.
6. Click **Update Job**.

**Viewing Job History**

This topics shows you how to view the history for jobs run within a project.

**Procedure**

1. Navigate to the project where the job was created.
2. Click **Jobs**.
3. Select the relevant job.
4. Click the **History** tab. You will see a list of all the job runs with some basic information such as who created the job, run duration, and status. Click individual runs to see the session output for each run.

**Cloudera Machine Learning Jobs API**

This topic demonstrates how to use the API to launch jobs.

Cloudera Machine Learning exposes a REST API that allows you to schedule jobs from third-party workflow tools. You must authenticate yourself before you can use the API to submit a job run request. The Jobs API supports HTTP Basic Authentication, accepting the same users and credentials as Cloudera Machine Learning.

**Note:** The Jobs API will be deprecated in the future. See [CML API v2 (Beta)](https://example.com) for the successor API.

**API Key Authentication**

Cloudera recommends using your API key for requests instead of your actual username/password so as to avoid storing and sending your credentials in plaintext. The API key is a randomly generated token that is unique to each user. It must be treated as highly sensitive information because it can be used to start jobs via the API. To look up your Cloudera Machine Learning API key:

2. From the upper right drop-down menu, switch context to your personal account.
3. Click Settings.
4. Select the API Key tab.

The following example demonstrates how to construct an HTTP request using the standard basic authentication technique. Most tools and libraries, such as Curl and Python Requests, support basic authentication and can set the required Authorization header for you. For example, with curl you can pass the API Key to the --user flag and leave the password field blank.

```
curl -v -XPOST http://cdsw.example.com/api/v1/<path_to_job> --user "<API_KEY>:
```

To access the API using a library that does not provide Basic Authentication convenience methods, set the request's Authorization header to Basic <API_KEY_encoded_in_base64>. For example, if your API key is uysgxtj7jzkps96njext.nxxm05usp0b, set Authorization to Basic dXlzZ3h0ajdqemtwczk2bmpleHRueHhtcTA1dXNwMGI6.

**Starting a Job Run Using the API**

Once a job has been created and configured through the Cloudera Machine Learning web application, you can start a run of the job through the API. This will constitute sending a POST request to a job start URL of the form: http://cdsw.example.com/api/v1/projects/<$USERNAME>/<$PROJECT_NAME>/jobs/<$JOB_ID>/start.

To construct a request, use the following steps to derive the username, project name, and job ID from the job's URL in the web application.

1. Log in to the Cloudera Machine Learning web application.
2. Switch context to the team/personal account where the parent project lives.
3. Select the project from the list.
4. From the project's Overview, select the job you want to run. This will take you to the job Overview page. The URL for this page is of the form: http://cdsw.example.com/<$USERNAME>/<$PROJECT_NAME>/jobs/<$JOB_ID>.
5. Use the $USERNAME, $PROJECT_NAME, and $JOB_ID parameters from the job Overview URL to create the following job start URL: http://cdsw.example.com/api/v1/projects/<$USERNAME>/<$PROJECT_NAME>/jobs/<$JOB_ID>/start.

For example, if your job Overview page has the URL http://cdsw.example.com/alice/sample-project/jobs/123, then a sample POST request would be of the form:

```
curl -v -XPOST http://cdsw.example.com/api/v1/projects/alice/sample-project/jobs/123/start    
--user "<API_KEY>:" --header "Content-type: application/json"
```

Note that the request must have the Content-Type header set to application/json, even if the request body is empty.

**Setting Environment Variables**

You can set environment variables for a job run by passing parameters in the API request body in a JSON-encoded object with the following format.

```
{
    "environment": {
        "ENV_VARIABLE": "value 1",
        "ANOTHER_ENV_VARIABLE": "value 2"
    }
}
```
The values set here will override the defaults set for the project and the job in the web application. This request body is optional and can be left blank.

Be aware of potential conflicts with existing defaults for environment variables that are crucial to your job, such as PATH and the CML variables.

Sample Job Run

As an example, let’s assume user Alice has created a project titled Risk Analysis. Under the Risk Analysis project, Alice has created a job with the ID, 208. Using curl, Alice can use her API Key (uysgxtj7jzkps96njextnnxmq05usp0b) to create an API request as follows:

```bash
curl -v -XPOST http://cdsw.example.com/api/v1/projects/alice/risk-analysis/jobs/208/start      \
   --user "uysgxtj7jzkps96njextnnxmq05usp0b:" --header "Content-type: application/json"           
   --data '{"environment": {"START_DATE": "2017-01-01", "END_DATE": "2017-01-31"}}'
```

In this example, START_DATE and END_DATE are environment variables that are passed as parameters to the API request in a JSON object.

In the resulting HTTP request, curl automatically encodes the Authorization request header in base64 format.

```
* Connected to cdsw.example.com (10.0.0.3) port 80 (#0)
* Server auth using Basic with user 'uysgxtj7jzkps96njextnnxmq05usp0b'
> POST /api/v1/projects/alice/risk-analysis/jobs/21/start HTTP/1.1
> Host: cdsw.example.com
> Authorization: Basic dXlzZ3h0ajdqemtwczk2bmpleHRueHhtcTA1dXNwMG16
> User-Agent: curl/7.51.0
> Accept: */*
> Content-type: application/json
> HTTP/1.1 200 OK
> Access-Control-Allow-Origin: *
> Content-Type: application/json; charset=utf-8
> Date: Mon, 10 Jul 2017 12:00:00 GMT
> Vary: Accept-Encoding
> Transfer-Encoding: chunked
<
{
   "engine_id": "cwg6wclmg0x482u0"
}
```

You can confirm that the job was started by going to the Cloudera Machine Learning web application.

Starting a Job Run Using Python

To start a job run using Python, Cloudera recommends using Requests, an HTTP library for Python; it comes with a convenient API that makes it easy to submit job run requests to Cloudera Machine Learning. Extending the Risk Analysis example from the previous section, the following sample Python code creates an HTTP request to run the job with the job ID, 208.

Python 2

```python
# example.py

import requests
import json

HOST = "http://cdsw.example.com"
USERNAME = "alice"
```
How To

API_KEY = "uygxt7jzkps96njeztxmq05usp0b"
PROJECT_NAME = "risk-analysis"
JOB_ID = "208"

def run_job():
    url = "/".join([HOST, "api/v1/projects", USERNAME, PROJECT_NAME, "jobs",
                   JOB_ID, "start"])
    job_params = {"START_DATE": "2017-01-01", "END_DATE": "2017-01-31"}
    res = requests.post(url,
                         headers = {
                                      "Content-Type": "application/json"},
                         auth = (API_KEY,""),
                         data = json.dumps({"environment": job_params})
                        )
    print "URL", url
    print "HTTP status code", res.status_code
    print "Engine ID", res.json().get('engine_id')

When you run the code, you should see output of the form:

```python example.py

URL http://cdsw.example.com/api/v1/projects/alice/risk-analysis/jobs/208/start
HTTP status code 200
Engine ID rllw5q3q589ryg9o
```

Limitations

- Cloudera Machine Learning does not support changing your API key, or having multiple API keys.
- Currently, you cannot create a job, stop a job, or get the status of a job using the Jobs API.

Related Information

- Basic Access Authentication
- Creating a Pipeline
- Environment Variables
- CML API v2 (Beta)

Distributed Computing with Workers

Cloudera Machine Learning provides basic support for launching multiple engine instances, known as workers, from a single interactive session. Any R or Python session can be used to spawn workers. These workers can be configured to run a script (e.g. a Python file) or a command when they start up.

Workers can be launched using the `launch_workers` function. Other supported functions are `list_workers` and `stop_workers`. Output from all the workers is displayed in the workbench console of the session that launched them. These workers are terminated when the session exits.

Using Workers for Machine Learning

The simplest example of using this feature would involve launching multiple workers from a session, where each one prints 'hello world' and then terminates right after. To extend this example, you can remove the print command and configure the workers to run a more elaborate script instead. For example, you can set up a queue of parameters (inputs to a function) in your main interactive session, and then configure the workers to run a script that pulls parameters off the queue, applies a function, and keeps doing this until the parameter queue is empty. This generic idea can be applied to multiple real-world use-cases. For example, if the queue is a list of URLs and the workers apply a function that scrapes a URL and saves it to a database, CML can easily be used to do parallelized web crawling.
Hyperparameter optimization is a common task in machine learning, and workers can use the same parameter queue pattern described above to perform this task. In this case, the parameter queue would be a list of possible values of the hyperparameters of a machine learning model. Each worker would apply a function that trains a machine learning model. The workers run until the queue is empty, and save snapshots of the model and its performance.

**Workers API**

This section lists the functions available as part of the workers API.

**Launch Workers**

Launches worker engines into the cluster.

**Syntax**

```
launch_workers(n, cpu, memory, nvidia_gpu=0, kernel="python3", script="", code="", env={})
```

**Parameters**

- **n (int)** - The number of engines to launch.
- **cpu (float)** - The number of CPU cores to allocate to the engine.
- **memory (float)** - The number of gigabytes of memory to allocate to the engine.
- **nvidia_gpu (int, optional)** - The number of GPU’s to allocate to the engine.
- **kernel (str, optional)** - The kernel. Can be "r", "python2", "python3" or "scala".
- **script (str, optional)** - The name of a Python source file the worker should run as soon as it starts up.
- **code (str, optional)** - Python code the engine should run as soon as it starts up. If a script is specified, code will be ignored.
- **env (dict, optional)** - Environment variables to set in the engine.

**Example Usage**

**Python**

```
import cdsw
workers = cdsw.launch_workers(n=2, cpu=0.2, memory=0.5, code="print('Hello from a CDSW Worker')")
```

**R**

```
library("cdsw")
workers <- launch.workers(n=2, cpu=0.2, memory=0.5, env="", code="print('Hello from a CML Worker')")
```

*Note:* The env parameter has been defined due to a bug that appears when parsing the `launch.workers` function. When not defined, the env parameter is serialized internally into a format that is incompatible with Cloudera Machine Learning. This bug does not affect the Python engine.

**List Workers**

Returns all information on all the workers in the cluster.

**Syntax**

```
list_workers()
```

**Stop Workers**

Stops worker engines.
Syntax

\[ \text{stop_workers(*worker_id)} \]

Parameter

- \text{worker_id} (int, optional) - The ID numbers of the worker engines that must be stopped. If an ID is not provided, all the worker engines on the cluster will be stopped.

Worker Network Communication

This section demonstrates some trivial examples of how two worker engines communicate with the master engine.

Workers are a low-level feature to help use higher level libraries that can operate across multiple hosts. As such, you will generally want to use workers only to launch the backends for these libraries.

To help you get your workers or distributed computing framework components talking to one another, every worker engine run includes an environmental variable \text{CML_MASTER_IP} with the fully addressable IP of the master engine.

Every engine has a dedicated IP access with no possibility of port conflicts.

For instance, the following are trivial examples of two worker engines talking to the master engine.

R

From the master engine, the following \text{master.r} script will launch two workers and accept incoming connections from them.

```r
# master.r
library("cdsw")
# Launch two CML workers. These are engines that will run in
# the same project, run a given code or script, and exit.
workers <- launch.workers(n=2, cpu=0.2, memory=0.5, env="", script="worker.r")
# Accept two connections, one from each worker. Workers will
# run worker.r.
for(i in c(1,2)) {
  # Receive a message from each worker and return a response.
  con <- socketConnection(host="0.0.0.0", port = 6000, blocking=TRUE, server=TRUE, open="r+")
  data <- readLines(con, 1)
  print(paste("Server received:" , data))
  writeLines("Hello from master!", con)
  close(con)
}
```

The workers will run the following \text{worker.r} script and respond to the master.

```r
# worker.r
print(Sys.getenv("CML_MASTER_IP"))
con <- socketConnection(host=Sys.getenv("CML_MASTER_IP"), port = 6000, blocking=TRUE, server=FALSE, open="r+")
write_resp <- writeLines("Hello from Worker", con)
server_resp <- readLines(con, 1)
print(paste("Worker received: ", server_resp))
close(con)
```
Python

From the master engine, the following master.py script will launch two workers and accept incoming connections from them.

```python
# master.py
import cdsw, socket

# Launch two CDSW workers. These are engines that will run in
# the same project, run a given code or script, and exit.
workers = cdsw.launch_workers(n=2, cpu=0.2, memory=0.5, script="worker.py")

# Listen on TCP port 6000
s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
s.bind(('0.0.0.0', 6000))
s.listen(1)

# Accept two connections, one from each worker. Workers will
# run worker.py.
conn, addr = s.accept()
for i in range(2):
    # Receive a message from each worker and return a response.
    data = conn.recv(20)
    if not data: break
    print("Master received:", data)
    conn.send("Hello From Server!").encode())
conn.close()
```

The workers will run the following worker.py script and respond to the master.

```python
# worker.py
import os, socket

# Open a TCP connection to the master.
s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
s.connect((os.environ['CDSW_MASTER_IP'], 6000))

# Send some data and receive a response.
s.send("Hello From Worker!").encode())
data = s.recv(1024)
s.close()
print("Worker received:", data)
```

Managing Users

This topic describes how to manage an ML workspace as a site administrator. Site administrators can monitor and manage all user activity across a workspace, add new custom engines, and configure certain security settings.

By default, the first user that logs in to a workspace must always be a site administrator. That is, they should have been granted the MLAdmin role by a CDP PowerUser.

**Note:** On Private Cloud, the corresponding role is **EnvironmentAdmin**.

**Important:** Site administrators have complete access to all activity on the deployment. This includes access to all teams and projects on the deployment, even if they have not been explicitly added as team members or collaborators.

Only site administrators have access to an Admin dashboard that can be used to manage the workspace. To access the site administrator dashboard:
1. Go to the Cloudera Machine Learning web application and log in as a site administrator.
2. On the left sidebar, click Admin. You will see an array of tabs for all the tasks you can perform as a site administrator.

**Monitoring Users**

The Users tab on the admin dashboard displays the complete list of users. You can see which users are currently active, and when a user last logged in to Cloudera Machine Learning. To modify a user's username, email or permissions, click the Edit button under the Action column.

**Related Information**

Cloudera Machine Learning Email Notifications

### Configuring Quotas

This topic describes how to configure CPU, GPU, and memory quotas for users of an ML workspace.

**Before you begin**

Required Role: MLAdmin

**Note:** On Private Cloud, the corresponding role is EnvironmentAdmin.

Make sure you are assigned the MLAdmin role in CDP. Only users with the MLAdmin role will be logged into ML workspaces with Site Administrator privileges.

There are two types of quota: Default and Custom. Default quotas apply to all users of the workspace. Custom quotas apply to individual users in the workspace, and take precedence over the default quota.

**Procedure**

1. Log in to the CDP web interface.
2. Click ML Workspaces, then open the Workspace for which you want to set quotas.
3. Click Admin > Quotas.
4. Switch the Default Quotas toggle to ON.

This applies a default quota of 2 vCPU and 8 GB memory to each user in the workspace.

If your workspace was provisioned with GPUs, a default quota of 0 GPU per user applies. If you want users to have access to GPUs, you must modify the default quotas as described in the next step.
5. If you want to change the default quotas, click on **Default (per user)**.
   CML displays the **Edit default quota** dialog box.
6. Enter the CPU, Memory, and GPU quota values that should apply to all users of the workspace.
7. Click **Update**.
8. To add a custom quota for a specific user, click **Add User**.
9. Enter the user name, and enter the quotas for CPU, Memory, and GPU.
10. Click **Add**.

**Results**
Enabling and modifying quotas will only affect new workloads. If users have already scheduled workloads that exceed the new quota limits, those will continue to run uninterrupted. If a user is over their limit, they will not be able to schedule any more workloads.

**Assigning the Site Administrator Role**
Site Administrators manage the Cloudera Machine Learning deployment, including security and whether certain features are enabled. Only existing site administrators can assign the site administrator role to an existing user.

**About this task**
If you have set up centralized identity management with CDP, you do not need to follow these steps to set the Site Admin role for each user individually. Set up user groups in CDP to assign Site Admin or regular user access to provisioned ML workspaces: Configuring User Access to CML.

**Procedure**
1. Sign in to Cloudera Machine Learning as a site administrator.
2. Click **Admin**.
3. Click the **Users** tab.
4. Click on the username of the user who you want to make a site administrator.
5. Select the **Site Administrator** checkbox.
6. Click **Update**.

**Disabling User Accounts**
Disabled users cannot login and do not count towards named users for licensing.

**Procedure**
1. Sign in to Cloudera Machine Learning as a site administrator.
2. Click **Admin**.
3. Click the **Users** tab.
4. Click on the username of the user who you want to disable.
5. Select the **Disabled** checkbox.
6. Click **Update**.

**Monitoring Cloudera Machine Learning Activity**
This topic describes how to monitor user activity on an ML workspace.
Required Role: Site Administrator

The **Admin > Overview** tab displays basic information about your deployment, such as the number of users signed up, the number of teams and projects created, memory used, and some average job scheduling and run times. You can also see the version of Cloudera Machine Learning you are currently running.

The **Admin > Activity** tab of the dashboard displays the following time series charts. These graphs should help site administrators identify basic usage patterns, understand how cluster resources are being utilized over time, and how they are being distributed among teams and users.

**Important:** The graphs and numbers on the **Admin > Activity** page do not account for any resources used by active models on the deployment. For that information, go to **Admin > Models** page.

- **CPU** - Total number of CPUs requested by sessions running at this time.
  
  Note that code running inside an n-CPU session, job, experiment or model replica can access at least n CPUs worth of CPU time. Each user pod can utilize all of its host's CPU resources except the amount requested by other user workloads or Cloudera Machine Learning application components. For example, a 1-core Python session can use more than 1 core if other cores have not been requested by other user workloads or CML application components.

- **Memory** - Total memory (in GiB) requested by sessions running at this time.

- **GPU** - Total number of GPUs requested by sessions running at this time.

- **Runs** - Total number of sessions and jobs running at this time.

- **Lag** - Depicts session scheduling and startup times.
  
  - **Scheduling Duration:** The amount of time it took for a session pod to be scheduled on the cluster.
  - **Starting Duration:** The amount of time it took for a session to be ready for user input. This is the amount of time since a pod was scheduled on the cluster until code could be executed.

The **Export Sessions List** provides a CSV export file of the columns listed in the table. It is important to note that the exported duration column is in seconds for a more detailed output.

**Tracked User Events**

The tables on this page describe the user events that are logged by Cloudera Machine Learning.

**Table 16: Database Columns**

When you query the `user_events` table, the following information can be returned:
How To

Description

**id**
The ID assigned to the event.

**user_id**
The UUID of the user who triggered the event.

**ipaddr**
The IP address of the user or component that triggered the event. 127.0.0.1 indicates an internal component.

**user agent**
The user agent for this action, such as the web browser. For example:

```
Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36
```

**event_name**
The event that was logged. The tables on this page list possible events.

**description**
This field contains the model name and ID, the user type (NORMAL or ADMIN), and the username.

**created_at**
The date (YYYY-MM-DD format) and time (24-hour clock) the event occurred.

<table>
<thead>
<tr>
<th>Table 17: Events Related to Engines</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event</strong></td>
</tr>
<tr>
<td>engine environment vars updated</td>
</tr>
<tr>
<td>engine mount created</td>
</tr>
<tr>
<td>engine mount deleted</td>
</tr>
<tr>
<td>engine mount updated</td>
</tr>
<tr>
<td>engine profile created</td>
</tr>
<tr>
<td>engine profile deleted</td>
</tr>
<tr>
<td>engine profile updated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 18: Events Related to Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event</strong></td>
</tr>
<tr>
<td>experiment run created</td>
</tr>
<tr>
<td>experiment run repeated</td>
</tr>
<tr>
<td>experiment run cancelled</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 19: Events Related to Files</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event</strong></td>
</tr>
<tr>
<td>file downloaded</td>
</tr>
<tr>
<td>file updated</td>
</tr>
<tr>
<td>file deleted</td>
</tr>
<tr>
<td>file copied</td>
</tr>
<tr>
<td>file renamed</td>
</tr>
<tr>
<td>file linked</td>
</tr>
<tr>
<td>directory uploaded</td>
</tr>
</tbody>
</table>
### Table 20: Events Related to Models

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>model created</td>
<td></td>
</tr>
<tr>
<td>model deleted</td>
<td></td>
</tr>
</tbody>
</table>

### Table 21: Events Related to Jobs

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>job created</td>
<td></td>
</tr>
<tr>
<td>job started</td>
<td></td>
</tr>
<tr>
<td>stopped all runs for job</td>
<td></td>
</tr>
<tr>
<td>job shared with user</td>
<td></td>
</tr>
<tr>
<td>job unshared with user</td>
<td></td>
</tr>
<tr>
<td>job sharing updated</td>
<td>The logged event indicates when the sharing status for a job is changed</td>
</tr>
<tr>
<td></td>
<td>from one of the following options to another:</td>
</tr>
<tr>
<td></td>
<td>• All anonymous users with the link</td>
</tr>
<tr>
<td></td>
<td>• All authenticated users with the link</td>
</tr>
<tr>
<td></td>
<td>• Specific users and teams</td>
</tr>
</tbody>
</table>

### Table 22: Events Related to Licenses

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>license created</td>
<td></td>
</tr>
<tr>
<td>license deleted</td>
<td></td>
</tr>
</tbody>
</table>

### Table 23: Events Related to Projects

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>project created</td>
<td></td>
</tr>
<tr>
<td>project updated</td>
<td></td>
</tr>
<tr>
<td>project deleted</td>
<td></td>
</tr>
<tr>
<td>collaborator added</td>
<td></td>
</tr>
<tr>
<td>collaborator removed</td>
<td></td>
</tr>
<tr>
<td>collaborator invited</td>
<td></td>
</tr>
</tbody>
</table>

### Table 24: Events Related to Sessions

<table>
<thead>
<tr>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>session launched</td>
<td></td>
</tr>
<tr>
<td>session terminated</td>
<td></td>
</tr>
<tr>
<td>session stopped</td>
<td></td>
</tr>
<tr>
<td>session shared with user</td>
<td></td>
</tr>
<tr>
<td>session unshared with user</td>
<td></td>
</tr>
</tbody>
</table>
**How To**

### Event Description

#### update session sharing status

The logged event indicates when the sharing status for a session is changed from one of the following options to another:
- All anonymous users with the link
- All authenticated users with the link
- Specific users and teams

<table>
<thead>
<tr>
<th>Table 25: Events Related to Admin Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event</strong></td>
</tr>
<tr>
<td>site config updated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 26: Events Related to Teams</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event</strong></td>
</tr>
<tr>
<td>add member to team</td>
</tr>
<tr>
<td>delete team member</td>
</tr>
<tr>
<td>update team member</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 27: Events Related to Users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event</strong></td>
</tr>
<tr>
<td>forgot password</td>
</tr>
<tr>
<td>password reset</td>
</tr>
<tr>
<td>update user</td>
</tr>
<tr>
<td>user signup</td>
</tr>
<tr>
<td>user login</td>
</tr>
<tr>
<td>user logout</td>
</tr>
<tr>
<td>ldap/saml user creation</td>
</tr>
</tbody>
</table>

### Monitoring User Events

This topic shows you how to query the PostgreSQL database that is embedded within the Cloudera Machine Learning deployment to monitor or audit user events.

#### About this task

Querying the PostgreSQL database that is embedded within the Cloudera Machine Learning deployment requires root access to the Cloudera Machine Learning Master host.

#### Procedure

1. SSH to the Cloudera Machine Learning Master host and log in as root.

   For example, the following command connects to `cdsw-master-host`, as root:

   `ssh root@cdsw-master-host.yourcdswdomain.com`
2. Get the name of the database pod:

```
kubectl get pods -l role=db
```

The command returns information similar to the following example:

```
NAME                  READY   STATUS    RESTARTS   AGE
db-86bbb69b54-d5q88   1/1     Running   0          4h46m
```

3. Enter the following command to log into the database as the `sense` user:

```
kubectl exec <database pod> -ti -- psql -U sense
```

For example, the following command logs in to the database on pod `db-86bbb69b54-d5q88`:

```
kubectl exec db-86bbb69b54-d5q88 -ti -- psql -U sense
```

You are logged into the database as the `sense` user.

4. Run queries against the `user_events` table.

Run the following query to view the most recent user event:

```
select * from user_events order by created_at DESC LIMIT 1
```

The command returns information similar to the following:

```
id          | 3658
user_id     | 273
ipaddr      | ::ffff:127.0.0.1
user_agent  | node-superagent/2.3.0
event_name  | model created
description | {"model":"Simple Model 1559154287-ex5yn","modelId":"50","userType":"NORMAL","username":"DonaldBatz"}
created_at  | 2019-05-29 18:24:47.65449
```
5. Optionally, you can export the user events to a CSV file for further analysis:

a) Copy the `user_events` table to a CSV file:

```bash
copy user_events to '/tmp/user_events.csv' DELIMITER ',' CSV HEADER;
```

b) Find the container that the database runs on:

```bash
docker ps | grep db-86bbb
```

The command returns output similar to the following:

```
c56d04bbd58 c230b2f564da "docker-entrypoint..." 7 days ago Up 7 days k8s _db_db-86bbb69b54-fcfm6_default_8b2dd23d-88b9-11e9-bc34-0245eb679f96_0
```

The first entry is the container ID.

c) Copy the `user_events.csv` file out of the container into a temporary directory on the Master host:

```bash
docker cp <container ID>:/tmp/user_events.csv /tmp/user_events.csv
```

For example:

```bash
docker cp 8c56d04bbd58:/tmp/user_events.csv /tmp/user_events.csv
```

d) Use SCP to copy `/tmp/user_events.csv` from the Cloudera Machine Learning Master host to a destination of your choice.

For example, run the following command on your local machine to copy `user_events.csv` to a local directory named `events`:

```bash
scp root@cdsw-master-host.yourcdswdomain.com:/tmp/user_events.csv /local/directory/events/
```

**What to do next**

For information about the different user events, see *Tracked User Events*.

**Related Information**

- Tracked User Events

---

### Monitoring Active Models Across the Workspace

This topic describes how to monitor all active models currently deployed on your workspace.

**What is an Active Model?**

A model that is in the Deploying, Deployed, or Stopping stages is referred to as an active model.

Cloudera Machine Learning provides two ways to monitor active models:

**Monitoring All Active Models Across the Workspace**

Required Role: Site Administrator

To see a complete list of all the models that have been deployed on a deployment, and review resource usage across the deployment by models alone, go to **Admin > Models**. On this page, site administrators can also **Stop/Restart/Rebuild** any of the currently deployed models.
Monitoring and Alerts

Cloudera Machine Learning leverages CDP Monitoring based on Prometheus and Grafana to provide dashboards that allow you to monitor how CPU, memory, storage, and other resources are being consumed by your ML workspaces.

Prometheus is an internal data source that is auto-populated with resource consumption data for each deployment. Grafana is the monitoring dashboard that allows you to create visualizations for resource consumption data from Prometheus. By default, CML provides three Grafana dashboards: K8 Cluster, K8s Containers, and K8s Node. You can extend these dashboards or create more panels for other metrics. For more information, see the Grafana documentation.

Related Information

Grafana documentation

Choosing Default Engine

This topic describes how to choose a default engine for creating projects.

Before you begin

Required Role: MLAdmin

Note: On Private Cloud, the corresponding role is EnvironmentAdmin.

Make sure you are assigned the MLAdmin role in CDP. Only users with the MLAdmin role will be logged into ML workspaces with Site Administrator privileges.

There are two types of default engines: ML Runtime and Legacy Engines. Legacy engines contain the machinery necessary to run sessions using all four interpreter options that CML currently supports (Python 2, Python 3, R and Scala) and other support utilities (C and Fortran compilers, LaTeX, etc.). ML Runtimes are thinner and more lightweight than legacy engines. Rather than supporting multiple programming languages in a single engine, each Runtime variant supports a single interpreter version and a subset of utilities and libraries to run the user’s code in Sessions, Jobs, Experiments, Models, or Applications.

Procedure

1. Log in to the CDP web interface.
2. Click ML Workspaces, then open the Workspace for which you want to set Default Engine.
3. Click Admin > Runtime/Engine.
4. Choose the Default Engine you would like to use as the default for all newly created projects in this workspace.
5. Modify the remaining information on the page:
   - Resource Profiles listed in the table are selectable resource options for both legacy Engines and ML Runtimes (for example, when starting a Session or Job)
   - The remaining information on the page applies to site-level settings specific for legacy Engines.

Related Information
ML Runtimes versus Legacy Engines

Controlling User Access to Features
Cloudera Machine Learning provides Site Administrators with the ability to restrict or hide specific functionality that non-Site Administrator users have access to in the UI. For example, a site administrator can hide the models and experiments features from the ML workspace UI.

The settings on this page can be configured through the Security and Settings tabs on the Admin page.

Table 28: Security Tab

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allow remote editing</td>
<td>Disable this property to prevent users from connecting to the Cloudera Machine Learning deployment with <code>cdswctl</code> and using local IDEs, such as PyCharm.</td>
</tr>
<tr>
<td>Allow only session creators to run commands on active sessions</td>
<td>By default, a user's permission to active sessions in a project is the same as the user's permission to that project, which is determined by the combination of the user's permission as a project collaborator, the user's permission in the team if this is a team project, and whether the user is a Site Administrator. By checking this checkbox, only the user that created the active session will be able to run commands in that session. No other users, regardless of their permissions in the team or as project collaborators, will be able to run commands on active sessions that are not created by them. Even Site Administrators will not be able to run commands in other users' active sessions.</td>
</tr>
<tr>
<td>Allow console output sharing</td>
<td>Disable this property to remove the Share button from the project workspace and workbench UI as well as disable access to all shared console outputs across the deployment. Note that re-enabling this property does not automatically grant access to previously shared consoles. You will need to manually share each console again</td>
</tr>
<tr>
<td>Allow anonymous access to shared console outputs</td>
<td>Disable this property to require users to be logged in to access shared console outputs.</td>
</tr>
<tr>
<td>Allow file upload/download through UI</td>
<td>Use this checkbox to show/hide file upload/download UI in the project workspace. When disabled, Cloudera Machine Learning API will forbid request of downloading file(s) as attachment. Note that the backend API to upload/edit/read the project files are intact regardless of this change in order to support basic Cloudera Machine Learning functionality such as file edit/read.</td>
</tr>
</tbody>
</table>

Table 29: Settings Tab

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Require invitation to sign up</td>
<td>Enable this property to send email invitations to users when you add them to a group. To send email, an SMTP server must first be configured in Settings &gt; Email.</td>
</tr>
<tr>
<td>Allow users to create public projects</td>
<td>Disable this property to restrict users from creating new public projects. Site Administrators will have to create any new public projects.</td>
</tr>
<tr>
<td>Allow Legacy Engine users to use the Python 2 kernel</td>
<td>Enable this property to allow Legacy Engine users to select the Python 2 kernel when creating a job. Python 2 is disabled by default.</td>
</tr>
<tr>
<td>Property</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Allow users to create projects</td>
<td>Disable this property to restrict users from creating new projects. Site Administrators will have to create any new projects.</td>
</tr>
<tr>
<td>Allow users to create teams</td>
<td>Disable this property to restrict users from creating new teams. Site Administrators will have to create any new teams.</td>
</tr>
<tr>
<td>Allow users to run experiments</td>
<td>Disable this property to hide the Experiments feature in the UI. Note that this property does not affect any active experiments. It will also not stop any experiments that have already been queued for execution.</td>
</tr>
<tr>
<td>Allow users to create models</td>
<td>Disable this property to hide the Models feature in the UI. Note that this property does not affect any active models. In particular, if you do not stop active models before hiding the Models feature, they continue to serve requests and consume computing resources in the background.</td>
</tr>
<tr>
<td>Allow users to create applications</td>
<td>Disable this property to hide the Applications feature in the UI. Note that this property does not affect any active applications. In particular, if you do not stop active applications before hiding the feature, they continue to serve requests and consume computing resources in the background.</td>
</tr>
</tbody>
</table>

**Cloudera Machine Learning Email Notifications**

Cloudera Machine Learning allows you to send email notifications when you add collaborators to a project, share a project with a colleague, and for job status updates (email recipients are configured per-job). This topic shows you how to specify email address for such outbound communications.

Note that email notifications are not currently enabled by default. Emails are not sent when you create a new project. Email preferences cannot currently be configured at an individual user level.

Option 1: If your existing corporate SMTP server is accessible from the VPC where your ML workspace is running, you can continue to use that server. Go to the **Admin > Settings** tab to specify an email address for outbound invitations and job notifications.

Option 2: If your existing SMTP solution cannot be used, consider using an email service provided by your cloud provider service. For example, Amazon provides Amazon Simple Email Service (Amazon SES).

**Downloading Diagnostic Bundles for a Workspace**

This topic describes how to download diagnostic bundles for an ML workspace.

Before you begin

**Note:** This topic applies only to public cloud releases.

Required Role: MLAdmin

Make sure you are assigned the MLAdmin role in CDP. Only users with the MLAdmin role will be logged into ML workspaces with Site Administrator privileges.

Procedure

1. Log in to the CDP web interface.
2. Click **ML Workspaces**.
3. Click **Admin > Support**.
4. Select Include Engine Logs if needed.
5. Select the time period from the dropdown and click **Download Logs**.
How To

Results
This will download a .zip file to your machine. The data in these bundles may be incomplete. If it does not contain logs for time period you are looking for, there are a number of possible reasons:

- There is a delay between the time the logs are initially generated by a workload and the time they are visible in cloud storage. This may be approximately 1 minute due to buffering during streaming, but can be significantly longer due to eventual consistency in the cloud storage.
- Another user or process may have deleted data from your bucket; this is beyond the control of Cloudera Machine Learning.
- There may be a misconfiguration or an invalid parameter in your request. Retrieving logs requires a valid cloud storage location to be configured for logging, as well as authentication for Cloudera Machine Learning to be set up properly for it. Requests must pertain to a valid engine in a valid project.

Configuring External Authentication with LDAP and SAML

Important: Cloudera recommends you leverage Single Sign-On for users via the CDP Management Console. For instructions on how to configure this, see Configuring User Access to CML. If you cannot do this, we recommend contacting Cloudera Support before attempting to use the LDAP or SAML instructions provided in this section.

Cloudera Machine Learning supports user authentication against its internal local database, and against external services such as Active Directory, OpenLDAP-compatible directory services, and SAML 2.0 Identity Providers. By default, Cloudera Machine Learning performs user authentication against its internal local database. This topic describes the signup process for the first user, how to configure authentication using LDAP, Active Directory or SAML 2.0, and an optional workaround that allows site administrators to bypass external authentication by logging in using the local database in case of misconfiguration.

User Signup Process

The first time you visit the Cloudera Machine Learning web console, the first account that you sign up with is a local administrator account. If in the future you intend to use external services for authentication, Cloudera recommends you use exclusive username & email combinations, rather than site administrators' work email addresses, for both the first site administrator account, and any other local accounts created before switching to external authentication. If the username/email combinations are not unique, an email address might end up being associated with different usernames, one for the external authentication service provider and one for a local Cloudera Machine Learning account. This will prevent the user from logging into Cloudera Machine Learning with their credentials for the external authentication service.

The link to the signup page is only visible on the login page when the authentication type is set to local. When you enable external services for authentication, signing up through the local database is disabled, and user accounts are automatically created upon their first successful login.

Optionally, site administrators can use a Require invitation to sign up flag under the Admin > Settings tab to require invitation tokens for account creation. When this flag is enabled, only users that are invited by site administrators can login to create an account, regardless of the authentication type.

Important: If you forget the original credentials, or make a mistake with LDAP or SAML configuration, you can use the workaround described in #unique_179.

When you switch to using external authentication methods such as LDAP or SAML 2.0, user accounts will be automatically created upon each user's first successful login. Cloudera Machine Learning will extract user attributes such as username, email address and full name from the authentication responses received from the LDAP server or SAML 2.0 Identity Provider and use them to create the user accounts.

Configuring LDAP/Active Directory Authentication

This topic describes how to set up LDAP authentication for a workspace.
Important: This is not the recommended method to set up LDAP authentication. Cloudera recommends you use the CDP management console to set this up: Configuring User Access to CML.

Cloudera Machine Learning supports both search bind and direct bind operations to authenticate against an LDAP or Active Directory directory service. The search bind authentication mechanism performs an ldapsearch against the directory service, and binds using the found Distinguished Name (DN) and password provided. The direct bind authentication mechanism binds to the LDAP server using a username and password provided at login.

You can configure Cloudera Machine Learning to use external authentication methods by clicking the Admin link on the left sidebar and selecting the Security tab. Select LDAP from the list to start configuring LDAP properties.

**LDAP General Settings**

Lists the general settings required to configure LDAP authentication.

- **LDAP Server URI**: Required. The URI of the LDAP/Active Directory server against which Cloudera Machine Learning should authenticate. For example, ldaps://ldap.company.com:636.
- **Use Direct Bind**: If checked, the username and password provided at login are used with the LDAP Username Pattern for binding to the LDAP server. If unchecked, Cloudera Machine Learning uses the search bind mechanism and two configurations, LDAP Bind DN and LDAP Bind Password, are required to perform the ldapsearch against the LDAP server.
- **LDAP Bind DN**: Required when using search bind. The DN to bind to for performing ldapsearch. For example, cn=admin,dc=company,dc=com.
- **LDAP Bind Password**: Required when using search bind. This is the password for the LDAP Bind DN.
- **LDAP Search Base**: Required. The base DN from which to search for the provided LDAP credentials. For example, ou=Engineering,dc=company,dc=com.
- **LDAP User Filter**: Required. The LDAP filter for searching for users. For example, (&(sAMAccountName={0})(objectclass=person)). The {0} placeholder will be replaced with the username provided at login.
- **LDAP User Username Attribute**: Required. The case-sensitive username attribute of the LDAP directory service. This is used by Cloudera Machine Learning to perform the bind operation and extract the username from the response. Common values are uid, sAMAccountName, or userPrincipalName.

When you select Use Direct Bind, Cloudera Machine Learning performs a direct bind to the LDAP server using the LDAP Username Pattern with the credentials provided on login (not LDAP Bind DN and LDAP Bind Password).

By default, Cloudera Machine Learning performs an LDAP search using the bind DN and credentials specified for the LDAP Bind DN and LDAP Bind Password configurations. It searches the subtree, starting from the base...
DN specified for the **LDAP Search Base** field, for an entry whose attribute specified in **LDAP User Username Attribute**, has the same value as the username provided on login. Cloudera Machine Learning then validates the user-provided password against the DN found as a result of the search.

**LDAP Group Settings**
In addition to the general LDAP settings, you can use group settings to restrict the access to Cloudera Machine Learning to certain groups in LDAP.

- **LDAP Group Search Base**: The base distinguished name (DN) where Cloudera Machine Learning will search for groups.
- **LDAP Group Search Filter**: The LDAP filter that Cloudera Machine Learning will use to determine whether a user is affiliated to a group.

  A group object in LDAP or Active Directory typically has one or more member attributes that stores the DN of users in the group. If LDAP Group Search Filter is set to `member={0}`, Cloudera Machine Learning will automatically substitute the `{0}` placeholder for the DN of the authenticated user.

- **LDAP User Groups**: A list of LDAP groups whose users have access to Cloudera Machine Learning. When this property is set, only users that successfully authenticate themselves AND are affiliated to at least one of the groups listed here, will be able to access Cloudera Machine Learning.

  If this property is left empty, all users that can successfully authenticate themselves to LDAP will be able to access Cloudera Machine Learning.

- **LDAP Full Administrator Groups**: A list of LDAP groups whose users are automatically granted the site administrator role on Cloudera Machine Learning.

  The groups listed under LDAP Full Administrator Groups do not need to be listed again under the LDAP User Groups property.

**Figure 6: Example**

If you want to restrict access to Cloudera Machine Learning to members of a group whose DN is:

- `CN=CMLUsers,OU=Groups,DC=company,DC=com`

And automatically grant site administrator privileges to members of a group whose DN is:

- `CN=CMLAdmins,OU=Groups,DC=company,DC=com`

Add the CNs of both groups to the following settings in Cloudera Machine Learning:

- **LDAP User Groups**: CMLUsers
- **LDAP Full Administrator Groups**: CMLAdmins

**How Login Works with LDAP Group Settings Enabled**

With LDAP Group settings enabled, the login process in Cloudera Machine Learning works as follows:

1. **Authentication with LDAP**

   When an unauthenticated user first accesses Cloudera Machine Learning, they are sent to the login page where they can login by providing a username and password.

   Cloudera Machine Learning will search for the user by binding to the LDAP Bind DN and verify the username/password credentials provided by the user.
2. Authorization Check for Access to Cloudera Machine Learning

If the user is authenticated successfully, Cloudera Machine Learning will then use the LDAP Group Search Filter to search for all groups the user is affiliated to, in the DN provided by LDAP Group Search Base.

The list of LDAP groups the user belongs to is then compared to the pre-authorized lists of groups specified in the LDAP User Groups and LDAP Full Administrator Groups properties.

If there is a match with a group listed under LDAP User Groups, this user will be allowed to access Cloudera Machine Learning as a regular user.

If there is a match with a group listed under LDAP Full Administrator Groups, this user will be allowed to access Cloudera Machine Learning as a site administrator.

Test LDAP Configuration

Use the Test LDAP Configuration form to test your settings.

You can test your LDAP/Active Directory configuration by entering your username and password in the Test LDAP Configuration section. This form simulates the user login process and allows you to verify the validity of your LDAP/Active Directory configuration without opening a new window.

Before using this form, make sure you click Update to save the LDAP configuration you want to test.

Configuring SAML Authentication

This topic describes how to set up SAML for Single Sign-on authentication for a workspace.

⚠️ Important: This is not the recommended method to set up SSO. Cloudera recommends you use the CDP management console to set this up: Configuring User Access to CML.

Cloudera Machine Learning supports the Security Assertion Markup Language (SAML) for Single Sign-on (SSO) authentication; in particular, between an identity provider (IDP) and a service provider (SP). The SAML specification defines three roles: the principal (typically a user), the IDP, and the SP. In the use case addressed by SAML, the principal (user agent) requests a service from the service provider. The service provider requests and obtains an identity assertion from the IDP. On the basis of this assertion, the SP can make an access control decision—in other words it can decide whether to perform some service for the connected principal.

The primary SAML use case is called web browser single sign-on (SSO). A user with a user agent (usually a web browser) requests a web resource protected by a SAML SP. The SP, wanting to know the identity of the requesting user, issues an authentication request to a SAML IDP through the user agent. In the context of this terminology, Cloudera Machine Learning operates as a SP.

Cloudera Machine Learning supports both SP- and IDP-initiated SAML 2.0-based SSO. Its Assertion Consumer Service (ACS) API endpoint is for consuming assertions received from the Identity Provider. If your Cloudera Machine Learning domain root were cdsw.company.com, then this endpoint would be available at http://cdsw.company.com/api/v1/saml/acs. SAML 2.0 metadata is available at http://cdsw.company.com/api/v1/saml/metadata for IDP-initiated SSO. Cloudera Machine Learning uses HTTP Redirect Binding for authentication requests and expects to receive responses from HTTP POST Binding.

When Cloudera Machine Learning receives the SAML responses from the Identity Provider, it expects to see at least the following user attributes in the SAML responses:

- The unique identifier or username. Valid attributes are:
  - uid
  - urn:oid:0.9.2342.19200300.100.1.1
- The email address. Valid attributes are:
  - mail
  - email
  - urn:oid:0.9.2342.19200300.100.1.3
• The common name or full name of the user. Valid attributes are:
  • cn
  • urn:oid:2.5.4.3

In the absence of the cn attribute, Cloudera Machine Learning will attempt to use the following user attributes, if they exist, as the full name of the user:

• The first name of the user. Valid attributes are:
  • givenName
  • urn:oid:2.5.4.42

• The last name of the user. Valid attributes are:
  • sn
  • urn:oid:2.5.4.4

**Configuration Options**
List of properties to configure SAML authentication and authorization in Cloudera Machine Learning.

**Cloudera Machine Learning Settings**

• **Entity ID**: Required. A globally unique name for Cloudera Machine Learning as a Service Provider. This is typically the URI.

• **NameID Format**: Optional. The name identifier format for both Cloudera Machine Learning and Identity Provider to communicate with each other regarding a user. Default: urn:oasis:names:tc:SAML:1.1:nameid-format:emailAddress.

• **Authentication Context**: Optional. SAML authentication context classes are URIs that specify authentication methods used in SAML authentication requests and authentication statements. Default: urn:oasis:names:tc:SAML:2.0:ac:classes:PasswordProtectedTransport.

**Signing SAML Authentication Requests**

• **CDSW Private Key for Signing Authentication Requests**: Optional. If you upload a private key, you must upload a corresponding certificate as well so that the Identity Provider can use the certificate to verify the authentication requests sent by Cloudera Machine Learning. You can upload the private key used for both signing authentication requests sent to Identity Provider and decrypting assertions received from the Identity Provider.

• **CML Certificate for Signature Validation**: Required if the Cloudera Machine Learning Private Key is set, otherwise optional. You can upload a certificate in the PEM format for the Identity Provider to verify the authenticity of the authentication requests generated by Cloudera Machine Learning. The uploaded certificate is made available at the http://cdsw.company.com/api/v1/saml/metadata endpoint.

**SAML Assertion Decryption**

Cloudera Machine Learning uses the following properties to support SAML assertion encryption & decryption.

• **CML Certificate for Encrypting SAML Assertions**: Must be configured on the Identity Provider so that Identity Provider can use it for encrypting SAML assertions for Cloudera Machine Learning

• **CML Private Key for Decrypting SAML Assertions**: Used to decrypt the encrypted SAML assertions.

**Identity Provider**

• **Identity Provider SSO URL**: Required. The entry point of the Identity Provider in the form of URI.

• **Identity Provider Signing Certificate**: Optional. Administrators can upload the X.509 certificate of the Identity Provider for Cloudera Machine Learning to validate the incoming SAML responses.

Cloudera Machine Learning extracts the Identity Provider SSO URL and Identity Provider Signing Certificate information from the uploaded Identity Provider Metadata file. Cloudera Machine Learning also expects all
Identity Provider metadata to be defined in a `<md:EntityDescriptor>` XML element with the namespace "urn:oasis:names:tc:SAML:2.0:metadata", as defined in the SAMLMeta-xsd schema.

For on-premises deployments, you must provide a certificate and private key, generated and signed with your trusted Certificate Authority, for Cloudera Machine Learning to establish secure communication with the Identity Provider.

**Authorization**

When you're using SAML 2.0 authentication, you can use the following properties to restrict the access to Cloudera Machine Learning to certain groups of users:

- **SAML Attribute Identifier for User Role**: The Object Identifier (OID) of the user attribute that will be provided by your identity provider for identifying a user’s role/affiliation. You can use this field in combination with the following SAML User Groups property to restrict access to Cloudera Machine Learning to only members of certain groups.

  For example, if your identity provider returns the OrganizationalUnitName user attribute, you would specify the OID of the OrganizationalUnitName, which is urn:oid:2.5.4.11, as the value for this property.

- **SAML User Groups**: A list of groups whose users have access to Cloudera Machine Learning. When this property is set, only users that are successfully authenticated AND are affiliated to at least one of the groups listed here, will be able to access Cloudera Machine Learning.

  For example, if your identity provider returns the OrganizationalUnitName user attribute, add the value of this attribute to the SAML User Groups list to restrict access to Cloudera Machine Learning to that group.

  If this property is left empty, all users that can successfully authenticate themselves will be able to access Cloudera Machine Learning.

- **SAML Full Administrator Groups**: A list of groups whose users are automatically granted the site administrator role on Cloudera Machine Learning.

  The groups listed under SAML Full Administrator Groups do not need to be listed again under the SAML User Groups property.

**How Login Works with SAML Group Settings Enabled**

With SAML Group settings enabled, the login process in Cloudera Machine Learning works as follows:

1. **Authentication by Identity Provider**

   When an unauthenticated user accesses Cloudera Machine Learning, they are first sent to the identity provider’s login page, where the user can login as usual.

   Once successfully authenticated by the identity provider, the user is sent back to Cloudera Machine Learning along with a SAML assertion that includes, amongst other things, a list of the user's attributes.

2. **Authorization Check for Access to Cloudera Machine Learning**

   Cloudera Machine Learning will attempt to look up the value of the SAML Attribute Identifier for User Role in the SAML assertion and check to see whether that value, which could be one or more group names, exists in the SAML User Groups and SAML Full Administrator Groups whitelists.

   If there is a match with a group listed under SAML User Groups, this user will be allowed to access Cloudera Machine Learning as a regular user.

   If there is a match with a group listed under SAML Full Administrator Groups, this user will be allowed to access Cloudera Machine Learning as a site administrator.

**Debug Login URL**

When using external authentication, such as LDAP, Active Directory or SAML 2.0, even a small mistake in authentication configurations in either Cloudera Machine Learning or the Identity Provider could potentially block all users from logging in.
Cloudera Machine Learning provides an optional fallback debug login URL for site administrators to log in against the local database with their username/password created during the signup process before changing the external authentication method. The debug login URL is `http://cml-workspace-domain.com/login?debug=1`. If you do not remember the original password, you can reset it by going directly to `http://cml-workspace-domain.com/forgot-password`. When configured to use external authentication, the link to the forgot password page is disabled on the login page for security reasons.

**Disabling the Debug Login Route**

Optionally, the debug login route can be disabled to prevent users from accessing Cloudera Machine Learning via local database when using external authentication. In case of external authentication failures, when the debug login route is disabled, root access to the master host is required to re-enable the debug login route.

Contact Cloudera Support for more information.

**Configuring HTTP Headers for Cloudera Machine Learning**

This topic explains how to customize the HTTP headers that are accepted by Cloudera Machine Learning.

Required Role: Site Administrator

These properties are available under the site administrator panel at Admin > Security.

⚠️ **Important:** Any changes to the following properties require a full restart of Cloudera Machine Learning. To do so, run `cdsw restart` on the master host.

**Enable Cross-Origin Resource Sharing (CORS)**

Most modern browsers implement the Same-Origin Policy, which restricts how a document or a script loaded from one origin can interact with a resource from another origin. When the Enable cross-origin resource sharing property is enabled on Cloudera Machine Learning, web servers will include the `Access-Control-Allow-Origin: *` HTTP header in their HTTP responses. This gives web applications on different domains permission to access the Cloudera Machine Learning API through browsers.

This property is disabled by default.

If this property is disabled, web applications from different domains will not be able to programmatically communicate with the Cloudera Machine Learning API through browsers.

**Enable HTTP Security Headers**

When Enable HTTP security headers is enabled, the following HTTP headers will be included in HTTP responses from servers:

- X-XSS-Protection
- X-DNS-Prefetch-Control
- X-Frame-Options
- X-Download-Options
- X-Content-Type-Options

This property is enabled by default.

Disabling this property could leave your Cloudera Machine Learning deployment vulnerable to clickjacking, cross-site scripting (XSS), or any other injection attacks.
Enable HTTP Strict Transport Security (HSTS)

**Note:** Without TLS/SSL enabled, configuring this property will have no effect on your browser.

When both TLS/SSL and this property (Enable HTTP Strict Transport Security (HSTS)) are enabled, Cloudera Machine Learning will inform your browser that it should never load the site using HTTP. Additionally, all attempts to access Cloudera Machine Learning using HTTP will automatically be converted to HTTPS.

This property is disabled by default.

If you ever need to downgrade to back to HTTP, use the following sequence of steps: First, deactivate this checkbox to disable HSTS and restart Cloudera Machine Learning. Then, load the Cloudera Machine Learning web application in each browser to clear the respective browser's HSTS setting. Finally, disable TLS/SSL across the cluster. Following this sequence should help avoid a situation where users get locked out of their accounts due to browser caching.

Enable HTTP Security Headers

When Enable HTTP security headers is enabled, the following HTTP headers will be included in HTTP responses from servers:

- X-XSS-Protection
- X-DNS-Prefetch-Control
- X-Frame-Options
- X-Download-Options
- X-Content-Type-Options

This property is enabled by default.

Disabling this property could leave your Cloudera Machine Learning deployment vulnerable to clickjacking, cross-site scripting (XSS), or any other injection attacks.

Enable Cross-Origin Resource Sharing (CORS)

Most modern browsers implement the Same-Origin Policy, which restricts how a document or a script loaded from one origin can interact with a resource from another origin. When the Enable cross-origin resource sharing property is enabled on Cloudera Machine Learning, web servers will include the Access-Control-Allow-Origin: * HTTP header in their HTTP responses. This gives web applications on different domains permission to access the Cloudera Machine Learning API through browsers.

This property is disabled by default.
If this property is disabled, web applications from different domains will not be able to programmatically communicate with the Cloudera Machine Learning API through browsers.

## SSH Keys

This topic describes the different types of SSH keys used by Cloudera Machine Learning, and how you can use those keys to authenticate to an external service such as GitHub.

### Personal Key

Cloudera Machine Learning automatically generates an SSH key pair for your user account. You can rotate the key pair and view your public key on your user settings page. It is not possible for anyone to view your private key.

Every console you run has your account's private key loaded into its SSH-agent. Your consoles can use the private key to authenticate to external services, such as GitHub. For instructions, see #unique_195.

### Team Key

Team SSH keys provide a useful way to give an entire team access to external resources such as databases or GitHub repositories (as described in the next section).

Like Cloudera Machine Learning users, each Cloudera Machine Learning team has an associated SSH key. You can access the public key from the team's account settings. Click Account, then select the team from the drop-down menu at the upper right corner of the page.

When you launch a console in a project owned by a team, you can use that team's SSH key from within the console.

### Adding an SSH Key to GitHub

Add your Cloudera Machine Learning SSH public key to your GitHub account if you want to use GitHub repositories to create new projects or collaborate on projects.

**Procedure**

2. Go to the upper right drop-down menu and switch context to the account whose key you want to add.
3. On the left sidebar, click Settings.
4. Go to the SSH Keys tab and copy your public SSH key.
5. Sign in to your GitHub account and add the Cloudera Machine Learning key copied in the previous step to your GitHub account. For instructions, refer the GitHub documentation on adding SSH keys to GitHub.

### Creating an SSH Tunnel

You can use your SSH key to connect Cloudera Machine Learning to an external database or cluster by creating an SSH tunnel.

**About this task**

In some environments, external databases and data sources reside behind restrictive firewalls. A common pattern is to provide access to these services using a bastion host with only the SSH port open. Cloudera Machine Learning provides a convenient way to connect to such resources using an SSH tunnel.

If you create an SSH tunnel to an external server in one of your projects, then all engines that you run in that project are able to connect securely to a port on that server by connecting to a local port. The encrypted tunnel is completely transparent to the user and code.

**Procedure**

1. Open the Project Settings page.
2. Open the Tunnels tab.
3. Click New Tunnel.
4. Enter the server IP Address or DNS hostname.
5. Enter your username on the server.
6. Enter the local port that should be proxied, and to which remote port on the server.

**What to do next**
On the remote server, configure SSH to accept password-less logins using your individual or team SSH key. Often, you can do so by appending the SSH key to the file `/home/username/.ssh/authorized_keys`.

---

**Restricting User-Controlled Kubernetes Pods**

Cloudera Machine Learning 1.6.0 (and higher) includes three properties that allow you to control the permissions granted to user-controlled Kubernetes pods.

**Required Role:** Site Administrator

An example of a user-controlled pod is the engine pod, which provides the environment for sessions, jobs, etc. These pods are launched in a per-user Kubernetes namespace. Since the user has the ability to launch arbitrary pods, these settings restrict what those pods can do.

They are available under the site administrator panel at Admin > Security under the Control of User-Created Kubernetes Pods section.

Do not modify these settings unless you need to run pods that require special privileges. Enabling any of these properties puts CML user data at risk.

**Allow privileged pod containers**

Pod containers that are "privileged" are extraordinarily powerful. Processes within such containers get almost the same privileges that are available to processes outside the container.

If this property is enabled, a privileged container could potentially access all data on the host.

This property is disabled by default.

**Allow pod containers to mount unsupported volume types**

The volumes that can be mounted inside a container in a Kubernetes pod are already heavily restricted. Access is normally denied to volume types that are unfamiliar, such as GlusterFS, Cinder, Fibre Channel, etc. If this property is enabled, pods will be able to mount all unsupported volume types.

This property is disabled by default.

---

**CML and outbound network access**

Cloudera Machine Learning expects access to certain external networks. See the related information Outbound network access destinations for further information.

**Note:** The outbound network access destinations listed in Outbound internet access and proxy are only the minimal set required for CDP installation and operation. For environments with limited outbound internet access due to using a firewall or proxy, access to Python or R package repositories such as Python Package Index or CRAN may need to be whitelisted if your use cases require installing packages from those repositories. Alternatively, you may consider creating mirrors of those repositories within your environment.

**Related Information**

Outbound internet access and proxy
Transparent Proxy and Egress Trusted List

Cloudera Machine Learning, when used on AWS public cloud, supports transparent proxies. Transparent proxy enables CML to proxy web requests without requiring any particular browser setup.

Egress Trusted List

In normal operation, CML requires the ability to reach several external domains. See Outbound network access destinations for more information.

Related Information

Outbound internet access and proxy

Troubleshooting

Troubleshooting tips may help you out of some situations with Cloudera Machine Learning.

Troubleshooting

This topic describes a recommended series of steps to help you start diagnosing issues with a Cloudera Machine Learning workspace.

• Issues with Provisioning ML Workspaces: If provisioning an ML workspace fails, first go to your cloud provider account and make sure that you have all the resources required to provision an ML workspace. If failures persist, start debugging by reviewing the error messages on the screen. Check the workspace logs to see what went wrong. For more details on the troubleshooting resources available to you, see Troubleshooting ML Workspaces on AWS on page 152.

• Issues with Accessing ML Workspaces: If your ML Admin has already provisioned a workspace for you but attempting to access the workspace fails, confirm with your ML Admin that they have completed all the steps required to grant you access: Configuring User Access to CML.

• Issues with Running Workloads: If you have access to a workspace but are having trouble running sessions/jobs/experiments, and so on, see if your error is already listed here: Troubleshooting Issues with Workloads on page 159.

Cloudera Support

If you need assistance, contact Cloudera Support. Cloudera customers can register for an account to create a support ticket at the support portal. For CDP issues in particular, make sure you include the Request ID associated with your error message in the support case you create.

Troubleshooting ML Workspaces on AWS

This topic describes the ML workspace provisioning workflow and tells you how to start debugging issues with ML workspaces on AWS.

ML Workspace Provisioning Workflow

When you provision a new ML Workspace on AWS, CML performs the following actions:

1. Communicates with the CDP Management Console to check your AWS credentials. It will also enable Single Sign-On so that authorized CDP users are automatically logged in to the workspace that will be created.

2. Provisions an NFS filesystem for the workspace on your cloud service provider. On AWS, CML will provision storage on EFS.

3. Provisions a Kubernetes cluster on your cloud service provider. This cluster runs the workspace infrastructure and compute resources. On AWS, CML provisions an EKS cluster.
4. Mounts the provisioned NFS filesystem to the Kubernetes cluster.
6. Registers the workspace with the cloud provider's DNS service. On AWS, this is Route53.
7. Installs Cloudera Machine Learning onto the EKS cluster.

Troubleshooting Resources

Any of the steps listed above can experience failures. To start debugging, you will require access to one or more of the following resources.

- **Workspace > Details Page**
  
  Each workspace has an associated Details page that lists important information about the workspace. To access this page, sign in to CDP, go to ML Workspaces and click on the workspace name.

  This page lists basic information about the workspace such as who created it and when. More importantly, it includes a link to the environment where the workspace was created, a link to the underlying EKS cluster on AWS, a list of tags associated with the workspace, and the computing resources in use. The rest of this topic explains how to use these resources.

- **Workspace > Events Page**
  
  Each workspace also has an associated Events page that captures every action performed on the workspace. This includes creating, upgrading, and removing the workspace, among other actions. To access this page, sign in to CDP, go to ML Workspaces, click on the workspace name, and then click **Events**.

  Click the **View Logs** button associated with an action to see a high-level overview of all the steps performed by CML to complete the action.

  The Request ID associated with each action is especially useful in case of a failure as it allows Cloudera Support to efficiently track the series of operations that led to the failure.

- **Environment > Summary Page**
  
  CML workspaces depend quite heavily on the environment in which they are provisioned. Each environment's Summary page lists useful information that can help you debug issues with the CML service. You can access the environment directly from the workspace Details page.

  This page includes important information such as:

  - **Credential Setup** - Tells you how security has been configured for the environment. Your cloud credential gives CDP access to the region and virtual network that make up the environment thus allowing CDP to provision resources within that environment.
  
  - **Region** - The AWS region where the environment is provisioned. This is especially important because it tells you which region's AWS console you might need to access for further debugging.
  
  - **Network** - The VPC and subnets that were created for the environment. Each CML workspace requires a set of unique IP addresses to run all of its associated Kubernetes services. If you begin to run out of IP addresses, you will need these VPC and subnet IDs to debug further in the AWS console.
  
  - **Logs** - When you create a CDP environment, you are asked to specify an S3 bucket in that environment that will be used to store logs. All CML operational logs and Spark logs are also written to this bucket.

  You can use the AWS console to access these logs. Alternatively, Site administrators can download these logs directly from their workspace Site Admin panel (**Admin > Support**).

  **Note:** If you file a support case, Cloudera Support will not automatically have access to these logs because they live in your environment.
• AWS Management Console

If you have all the relevant information about the environment and the workspace, you can go to the AWS console (for the region where your environment was created) to investigate further. The AWS Management Console has links to dashboards for all the services used by CML.

• EC2

You can use the EC2 service dashboards to check the instance-type (CPU, GPU), VPC, subnet, and security group limits imposed on your AWS account. For example, there is typically a limit of 5 VPCs per region.

If you need more resources, submit a request to Amazon to raise the limit of a resource.

• EKS

EKS will give you more information such as the version of Kubernetes CML is using, network information, and the status of the cluster. The workspace Details page gives you a direct link to the provisioned EKS cluster on the AWS console.

  **Note:** By default, users do not have Kubernetes-level access to the EKS cluster. If a user wants to use `kubectl` to debug issues with the EKS cluster directly, an MLAdmin must explicitly grant access using the instructions provided here: [Granting Remote Access to ML Workspaces on EKS](#).

  **Note:** By default, users do not have Kubernetes-level access to the EKS cluster. If a user wants to use `kubectl` to debug issues with the EKS cluster directly, an MLAdmin must explicitly grant access using the instructions provided here: [Granting Remote Access to ML Workspaces on EKS](#).

• VPC

Use the VPC ID obtained from the CDP environment Summary page to search for the relevant VPC where you have provisioned or are trying to provision an ML workspace. Each CML workspace requires a set of unique IP addresses to run all of its associated Kubernetes services. You can use this service to see how many IP addresses are available for each subnet.

• S3

Use the S3 bucket configured for the environment to check/download logs for more debugging.

• Tags

When provisioning an ML workspace, you will have the option to assign one or more tags to the workspace. These tags are then applied to all the underlying AWS resources used by the workspace. If failures occur during provisioning or de-provisioning, it can be very useful to simply query the tags associated with the workspace to see if any resources need to be cleaned up manually. Tags associated with a workspace are available on the workspace Details page.

You can search by tags in the EC2 and VPC services. You can also use the AWS CLI to search for specific tags: `resourcegroupstaggingapi`

• Trusted Advisor (available with AWS Support)

Use the Trusted Advisor dashboard for a high-level view of how you are doing with your AWS account. The dashboard displays security risks, service limits, and possible areas to optimize resource usage. If you have access to AWS Support, it's a good idea to review your current account status with Trusted Advisor before you start provisioning ML workspaces.

**Cloudera Support**

When creating a case for Cloudera Support, make sure you have the following information:

• Request ID

**Common CML Errors and Solutions**

The following sections describe recommended steps to start debugging common error messages you might see in the workspace logs (found under **Events > View Logs**).
Before you begin

Make sure you have reviewed the list of resources available to you for debugging on CML and AWS:
Troubleshooting ML Workspaces on AWS on page 152

**AWS Account Resource Limits Exceeded (Compute, VPC, etc.)**

ML workspace provisioning fails because CDP could not get access to all the AWS resources needed to deploy a CML workspace. This is likely because your AWS account either does not have access to those resources or is hitting the resource limits imposed on it.

Sample errors include (from **Events > View Logs**):

| Failed to provision cluster. Reason: Failed to wait for provisioner: Wait for status failed with status CREATE_FAILED: error creating eks cluster (cause: InvalidParameterException: Provided subnets subnet-0a648a0cc5976b7a9 Free IPs: 0 , need at least 3 IPs in each subnet to be free for this operation |
| Failed to mount storage. Reason: Failed to create mount target: NoFreeAddressesInSubnet: The specified subnet does not have enough free addresses to satisfy the request. |

AWS accounts have certain hard and soft resource limits imposed on them by default. For example, certain CPU/GPU instances that CML allows you to provision might even have an initial default limit of 0 (set by AWS). This means if you attempt to provision a cluster with those instance types, your request will fail.

Aside from the CPU and GPU compute resource limits, there are other types of limits you can run into. For example, the second error shows that the subnets in your VPC don't have any more free IP addresses for the workspace (and each of the underlying Kubernetes pods). This occurs if the CIDR range mentioned while registering the environment was not large enough for your current needs.

You can use the AWS console to request an increase in limits as needed. Go to the AWS console for the region where the environment was provisioned and then navigate to **EC2 > Limits**.

For networking failures, navigate to **EC2 > VPC**. Search for the environment's VPC ID (available on environment Summary page) to see the list of available IP addresses for each subnet. Request more resources as needed.

Related AWS documentation: AWS Service Limits, Amazon EC2 Resource Limits, EKS Cluster VPC Considerations, AWS CNI Custom Networking.

**Access denied to AWS credential used for authentication**

The cloud credential used to give CDP access to your AWS account failed authentication. Therefore, CDP could not provision the resources required to deploy a CML workspace.

Sample error (from **Events > View Logs**):

| Failed to provision storage. Reason: Failed to create new file system: AccessDenied: User: arn:aws:iam::1234567890:user/cross-account-trust-user is not authorized to perform: |

Your cloud credential gives CDP access to the region and virtual network that make up the environment thus allowing CDP to provision resources within that environment. If authentication fails, go to your environment to see how the cloud credentials were set up and confirm whether your account has the permission to perform these actions.

**CML Installation Failures**

While the steps to provision resources on AWS were completed successfully, the CML workspace installation on EKS failed.
Sample error (from Events > View Logs):

Failed to install ML workspace. Reason: Error: release mlx-mlx failed: timed out waiting for the condition

If you are an advanced user, you can log in to the underlying EKS cluster and use kubectl to investigate further into which pods are failing.

**Note:** This error might be an indication that DNS has been turned off for the VPC. Go to the AWS console for the region where the environment was provisioned and then navigate to EC2 > Load Balancers to confirm that DNS is configured properly for the environment's VPC.

Related AWS documentation: EKS and kubectl

**Failures due to API Throttling**

These errors can be harder to prepare for due to their seemingly random nature. Occasionally, AWS will block API calls if it receives too many requests at the same time. For example, this can occur when multiple users are attempting to provision/delete/upgrade clusters at the same time.

Sample error (from Events > View Logs):

Failed to delete cluster. Reason: Failed to wait for deletion: Wait for status DELETE_FAILED: Throttling: Rate exceeded

Currently, if you see a 'Throttling: Rate exceeded' error, our recommendation is that you simply try again later.

Related AWS documentation: AWS API Request Throttling

**De-provisioning Failures**

De-provisioning operations can sometimes fail if AWS resources are not terminated in the right order. This is usually due to timing issues where certain resources might take too long to terminate. This can result in a cascading set of failures where AWS cannot delete the next set of resources because they still have active dependencies on the previous set.

Sample error (from Events > View Logs):

Failed to delete cluster. Reason: Failed to wait for deletion: DELETE_FAILED: msg: failed to delete aws stack

Cloudformation says resource xyz has a dependent object (Service: AmazonEC2; Status Code: 400; Error Code: DependencyViolation; Request ID: 815928e2-277e-4b8b-9fed-4b89716a205b) EKS - cluster still existed, was blocking CF delete

CML includes a Force Delete option now that will remove the workspace from the CML service. However, this not mean all the underlying resources have been cleaned up. This is where tags are very useful.

If you assigned tags to the workspace at the time of provisioning, you can use the AWS console or the CLI to query the tags associated with the workspace to see if any resources need to be cleaned up manually. Tags associated with a workspace are available on the workspace Details page.

You can search by tags in the EC2 and VPC services. You can also use the AWS CLI to search for specific tags: resourcegroupstaggngapi

**Users unable to access provisioned ML workspaces**

If you have provisioned a workspace but your colleagues cannot automatically access the workspace using CDP Single-Sign on, make sure that you have completed all the steps required to grant users access to workspaces: Configuring User Access to CML. All CML users must have CDP accounts.
Environment Issues

Need to figure out SDX config between different services on the environment - HMS IDBroker FreeIPA - if data lake is in a bad state you go there and figure out what went wrong.


TLS Certificate Creation Fails

Still under investigation by dev (CML failure or LetsEncrypt or AWS failure unknown) - no cert can be obtained, then no more progress in provisioning - try again.

Failed to install ML workspace. Reason: Unable to get TLS cert and key: Unable to poll certificate creation

Troubleshooting ML Workspaces on Azure

You can collect logs to troubleshoot issues that occur in ML Workspaces with Azure.

How to access Azure logs

Logs from the AKS control plane can be found in the Logs blade of the liftie-xxxxxxx resource group (not to be confused with the Logs blade of the AKS cluster itself or the Log Analytics Workspace in that resource group). The logs can be looked up using a query language developed by Microsoft.

Cluster fails to scale down

If a worker node is idle but is not being scaled down, check the cluster autoscaler logs.
Use this example to look up the logs:

```
AzureDiagnostics | where Category == "cluster-autoscaler"
```

The logs list the pods that are scheduled on a given node that are preventing it from being scaled down, or other reasons for its scaling decisions. Services running in the kube-system namespace (such as tunnelfront, or metrics-server) have been known to delay scale-down when scheduled on an otherwise idle node.

Delete ML Workspace fails

If you delete a workspace, and the delete operation fails, you can use **Force delete** to remove the workspace.
In this case, CML attempts to delete associated cloud resources for the workspace including metadata files. However, users should check that all such resources have been deleted, and delete manually if necessary.

Logs for ML Workspaces

You can access logs to troubleshoot issues with the CML service and your workloads on ML workspaces.

Access to logs

When you create a CDP environment, you specify an S3 bucket (on AWS) or an Azure Storage container (on Azure) in that environment for storing logs. If you have access to the log storage, you can use the AWS or Azure console to access certain CML and Spark logs directly. You can get the details of the specific bucket or container from the environment's Summary page.
Note: If you file a support case, Cloudera Support will not automatically have access to these logs because they live in your environment.

ML Workspace access to logs

CML workspace users also have access to these logs depending on their authorization level:

- Site Administrators
  Site administrators can download the same logs directly from their workspace Site Admin panel (Admin > Support). For more details, see Downloading Diagnostic Bundles for a Workspace on page 141.

- Data Scientists
  While data scientists don't have access to the full set of workspace logs, they do have access to engine logs for their own workloads (sessions/jobs/experiments). While in an interactive session or on a job/experiment's Overview page, click Download Logs at any time to review the full set of logs for that workload's engine. In the case of Spark workloads, Spark executor and event logs are also downloaded as part of this bundle.

Related Information
Configure lifecycle management for logs on AWS
Configure lifecycle management for logs on Azure

Downloading Diagnostic Bundles for a Workspace

This topic describes how to download diagnostic bundles for an ML workspace.

Before you begin

Note: This topic applies only to public cloud releases.

Required Role: MLAdmin

Make sure you are assigned the MLAdmin role in CDP. Only users with the MLAdmin role will be logged into ML workspaces with Site Administrator privileges.

Procedure

1. Log in to the CDP web interface.
2. Click ML Workspaces.
3. Click Admin > Support.
4. Select Include Engine Logs if needed.
5. Select the time period from the dropdown and click Download Logs.

Results

This will download a .zip file to your machine. The data in these bundles may be incomplete. If it does not contain logs for time period you are looking for, there are a number of possible reasons:

- There is a delay between the time the logs are initially generated by a workload and the time they are visible in cloud storage. This may be approximately 1 minute due to buffering during streaming, but can be significantly longer due to eventual consistency in the cloud storage.
- Another user or process may have deleted data from your bucket; this is beyond the control of Cloudera Machine Learning.
- There may be a misconfiguration or an invalid parameter in your request. Retrieving logs requires a valid cloud storage location to be configured for logging, as well as authentication for Cloudera Machine Learning to be set up properly for it. Requests must pertain to a valid engine in a valid project.
Troubleshooting Issues with Workloads

This section describes some potential issues data scientists might encounter once the ML workspace is running workloads.

Engines cannot be scheduled due to lack of CPU or memory

A symptom of this is the following error message in the Workbench: "Unschedulable: No node in the cluster currently has enough CPU or memory to run the engine."

Either shut down some running sessions or jobs or provision more hosts for Cloudera Machine Learning.

Workbench prompt flashes red and does not take input

The Workbench prompt flashing red indicates that the session is not currently ready to take input.

Cloudera Machine Learning does not currently support non-REPL interaction. One workaround is to skip the prompt using appropriate command-line arguments. Otherwise, consider using the terminal to answer interactive prompts.

PySpark jobs fail due to Python version mismatch

Exception: Python in worker has different version 2.6 than that in driver 2.7, PySpark cannot run with different minor versions

One solution is to install the matching Python 2.7 version on all the cluster hosts. A better solution is to install the Anaconda parcel on all CDH cluster hosts. Cloudera Machine Learning Python engines will use the version of Python included in the Anaconda parcel which ensures Python versions between driver and workers will always match. Any library paths in workloads sent from drivers to workers will also match because Anaconda is present in the same location across all hosts. Once the parcel has been installed, set the PYSPARK_PYTHON environment variable in the Cloudera Machine Learning Admin dashboard.

Troubleshooting Kerberos Errors

This topic describes some common Kerberos issues and their recommended solutions.

HDFS commands fail with Kerberos errors even though Kerberos authentication is successful in the web application

If Kerberos authentication is successful in the web application, and the output of klist in the engine reveals a valid-looking TGT, but commands such as hdfs dfs -ls / still fail with a Kerberos error, it is possible that your cluster is missing the Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction Policy File. The JCE policy file is required when Red Hat uses AES-256 encryption. This library should be installed on each cluster host and will live under $JAVA_HOME. For more information, see Using AES-256 Encryption.

Cannot find renewable Kerberos TGT

Cloudera Machine Learning runs its own Kerberos TGT renewer which produces non-renewable TGT. However, this confuses Hadoop's renewer which looks for renewable TGTs. If the Spark 2 logging level is set to WARN or lower, you may see exceptions such as:

16/12/24 16:38:40 WARN security.UserGroupInformation: Exception encountered while running the renewal command. Aborting renew thread. ExitCodeException
exitCode=1: kinit: Resource temporarily unavailable while renewing credentials

16/12/24 16:41:23 WARN security.UserGroupInformation: PrivilegedActionException as:user@CLOUDERA.LOCAL (auth:KERBEROS) cause:javax.security.sasl.SaslEx
ception: GSS initiate failed [Caused by GSSException: No valid credentials provided (Mechanism level: Failed to find any Kerberos tgt)]

This is not a bug. Spark 2 workloads will not be affected by this. Access to Kerberized resources should also work as expected.

Reference

Command Line Tools in CML

Cloudera Machine Learning ships with the following command line tools. The purpose of each tool differs.

- CDP CLI for Cloudera Machine Learning - If you prefer to work in a terminal window, you can download and configure the CDP client that gives you access to the CDP CLI tool. The CDP CLI allows you to perform the same actions as can be performed from the management console. Use this CLI to create, delete, upgrade, and manage ML workspaces on CDP.

  To view all the available commands, run:

  cdp ml help

  To view help for a specific command, run:

  cdp ml <operation> help

If you don't already have the CDP CLI set up, see Installing the CDP CLI Client.

- cdswctl - Cloudera Machine Learning also ships with a CLI client that you can download from the Cloudera Machine Learning web UI. This is also referred to as the Model CLI client. The cdswctl client allows you to log in, create an SSH endpoint, launch new sessions, automate model deployment, model updates, and so on.

Cloudera Machine Learning Command Line Reference

This topic describes the commands available to the cdsw command line utility that exists within a Cloudera Machine Learning workspace. This utility is meant to manage your Cloudera Machine Learning workspace. Running cdsw without any arguments will print a brief description of each command.

In addition, there is a cdswctl CLI client that offers different functionality that is meant for use by data scientists to manage their sessions. For information about that, see cdswctl Command Line Interface Client.

<table>
<thead>
<tr>
<th>Command</th>
<th>Description and Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>cdsw start</td>
<td>Initializes and bootstraps the master host. Use this command to start Cloudera Machine Learning.</td>
</tr>
<tr>
<td>cdsw stop</td>
<td>De-registers, resets, and stops a host.</td>
</tr>
<tr>
<td></td>
<td>On a worker host, this command will remove the worker from the cluster.</td>
</tr>
<tr>
<td></td>
<td>On the master host, this command will bring down the application and effectively tear down the Cloudera Machine Learning deployment.</td>
</tr>
<tr>
<td>cdsw restart</td>
<td>Run on the master host to restart application components.</td>
</tr>
<tr>
<td></td>
<td>To restart a worker host, use cdsw stop, followed by cdsw join. These commands have been explained further in this topic.</td>
</tr>
</tbody>
</table>
### Command Line Interface Client

Cloudera Machine Learning ships with a CLI client that you can download from the Cloudera Machine Learning web UI. The `cdswctl` client allows you to perform the following tasks:

- Logging in
- Creating an SSH endpoint
- Listing sessions that are starting or running
- Starting or stopping a session
- Creating a model
- Building and deploying models
- Listing model builds and model deployments
- Checking the status of a deployment
- Redeploying a model with updated resources
- Viewing the replica logs for a model

Other actions, such as creating a project, require you to use the Cloudera Machine Learning web UI. For information about the available commands, run the following command:

```bash
cdswctl --help
```

### Download and Configure cdswctl

This topic describes how to download the `cdswctl` CLI client and configure your SSH public key to authenticate CLI access to sessions.

### About this task

Before you begin, ensure that the following prerequisites are met:

- You have an SSH public/private key pair for your local machine.
- You have Contributor permissions for an existing project. Alternatively, create a new project you have access to.
- If you want to configure a third-party editor, make sure the Site Administrator has not disabled remote editing for Cloudera Machine Learning.
(Optional) Generate an SSH Public/Private Key

About this task
This task is optional. If you already have an SSH public/private key pair, skip this task. The steps to create an SSH public/private key pair differ based on your operating system. The following instructions are meant to be an example and are written for macOS using `ssh-keygen`.

Procedure
1. Open Terminal.
2. Run the following command and complete the fields:
   ```
   ssh-keygen -t rsa -f ~/.ssh/id_rsa
   ```

   Keep the following guidelines in mind:
   - Make sure that the SSH key you generate meets the requirements for the local IDE you want to use. For example, PyCharm requires the `-m PEM` option because PyCharm does not support modern (RFC 4716) OpenSSH keys.
   - Provide a passphrase when you generate the key pair. Use this passphrase when prompted for the SSH key passphrase.
   - Save the SSH key to the default `~/.ssh` location.

Download cdswctl and Add an SSH Key

Procedure
1. Open the Cloudera Machine Learning web UI and go to Settings > Remote Editing for your user account.
2. Download `cdswctl` client for your operating system.
   Unpack it, and optionally, you can add it to the `PATH` environment variable on your system.
3. Add your SSH public key to SSH public keys for session access.
   Cloudera Machine Learning uses the SSH public key to authenticate your CLI client session, including the SSH endpoint connection to the Cloudera Machine Learning deployment.
   Any SSH endpoints that are running when you add an SSH public key must also be restarted.

Initialize an SSH Endpoint

This topic describes how to establish an SSH endpoint for Cloudera Machine Learning.

About this task
Creating an SSH endpoint is also the first step to configuring a remote editor for an ML workspace.

Procedure
1. Create a configuration file at: `$HOME/.cdsw/config.yaml`. The contents of `config.yaml` should be:
   ```
   username: <username>
   url: <ML_workspace_url>
   auth:
     authtype: 1
     basic: null
     apikey: <your_api_key>
   ```

   To collect the values for these fields, first log in to your CML workspace using SSO:
   - `username`: The username with which you are logged into the CML workspace. Found in the top right corner of your ML workspace.
Reference

- **url**: The complete URL used to access the CML workspace. For example: https://ml-<randomly-generated-cluster-name>
- **apikey**: Go to Settings > API Keys. Copy the value to this field.

2. Create a local SSH endpoint to Cloudera Machine Learning. Run the following command:

```bash
cdsctl ssh-endpoint -p <username>/<project_name> [-c <CPU_cores>] [-m < memory_in_GB>] [-g <number_of_GPUs>] [-r <runtime ID>] 
```

If the project is configured to use ML runtimes, the `-r` parameter must be specified, otherwise it must be omitted. See *Using ML runtimes with cdsctl* documentation page for more information.

The command uses the following defaults for optional parameters:
- CPU cores: 1
- Memory: 1 GB
- GPUs: 0

For example, the following command starts a session for the user milton under the customerchurn project with .5 cores, .75 GB of memory, 0 GPUs, and the Python3 kernel:

```bash
cdsctl ssh-endpoint -p customerchurn -c 0.5 -m 0.75
```

To create an SSH endpoint in a project owned by another user or a team, for example finance, prepend the username to the project and separate them with a forward slash:

```bash
cdsctl ssh-endpoint -p finance/customerchurn -c 0.5 -m 0.75
```

This command creates session in the project customerchurn that belongs to the team finance.

Information for the SSH endpoint appears in the output:

```
... 
You can SSH to it using
    ssh -p <some_port> cdsw@localhost
... 
```

3. Open a new command prompt and run the outputted command from the previous step:

```bash
ssh -p <some_port> cdsw@localhost
```

For example:

```bash
ssh -p 9750 cdsw@localhost
```

You will be prompted for the passphrase for the SSH key you entered in the ML workspace web UI.

Once you are connected to the endpoint, you are logged in as the cdsw user and can perform actions as though you are accessing the terminal through the web UI.

4. Test the connection.
   - If you run `ls`, the project files associated with the session you created are shown. If you run `whoami`, the command returns the cdsw user.

5. Leave the SSH endpoint running as long as you want to use a local IDE.

**Related Information**
- [Using ML Runtimes with cdsctl](#)
- [ML Runtimes versus Legacy Engines](#)

**Log into cdsctl**

This topic describes how to log into cdsctl.
Procedure
1. Open the Model CLI client.
2. Run the following command while specifying the actual values for the variables:

   ```bash
cdswctl login -u <workspace_url> -n <username> -y <api_key>
   ```

   where
   - `workspace_url` is the workspace URL including the protocol (http(s)://domain.com)
   - `username` is your user name on the workspace
   - `api_key` is the API key that you can obtain from the Cloudera Machine Learning UI. Go to Settings > API Keys and copy the API Key (and not the Model API Key).

   To see more information about the login command parameters, run

   ```bash
cdswctl login --help
   ```

   If all goes well, then you'll see "Login succeeded".

Prepare to manage models using the model CLI

Before you can start using the model CLI to automate model deployment or to perform any other tasks, you must install the scikit-learn machine learning library for Python through the Cloudera Machine Learning web UI.

About this task
You must perform this task through the Cloudera Machine Learning web UI.

Procedure
1. Create a new project with Python through the web UI.
   Python provides sample files that you can use to create models using CLI.
2. To start a new session, go to the Sessions page from the left navigation panel and click new session.
   The Start the new session page is displayed.
3. On Start the new session page, select Python 3 from the Engine Kernel drop-down menu, and click Launch Session.
   A new “Untitled Session” is created.
4. From the input prompt, install the scikit-learn machine learning library for Python by running the following command:

   ```bash
   !pip3 install sklearn
   ```
5. Open the fit.py file available within your project from the left navigation panel.
   You can use the fit.py file to create a fitted model which creates a model.pkl file that you can use to deploy the actual model.
6. Run the fit.py file by clicking Run > Run all.
   The model.pkl directory is created that you can see within your project on the left navigation pane.
7. Close the session by clicking Stop.

Create a model using the CLI

This topic describes how to create models using the model CLI.

Procedure
1. Open a terminal window and log into cdswctl.
2. Obtain the project ID as described in the following steps:
   a) Run the following command:

   ```bash
cdswctl projects list
   ```

   The project ID, your username, and the project name are displayed. For example:
   ```plaintext
   1: john-smith/petal-length-predictor
   ```
   b) Note the project ID, which is a number in front of your project name.
   In this case, it is "1".

3. Run the following command while specifying the project name and note the engine image ID:

   **Note:** The following examples are specific to projects configured to use legacy engines and projects configured to use runtimes. Be sure to use the commands appropriate to your project configuration.

   For projects configured to use legacy engines:

   ```bash
cdswctl engine-images list -p <project-name>
   ```

   For example,

   ```bash
cdswctl engine-images list -p john-smith/petal-length-predictor
   ```

   For projects configured to use runtimes:

   ```bash
cdswctl runtimes list
   ```

   Depending on your local setup, you may get a more readable output by post-processing the result with the following command:

   ```bash
cdswctl runtimes list | python3 -m json.tool
   ```

   For this example you should pick a runtime with a Python kernel and Workbench editor. Depending on your local setup, you may filter the results using the following command:

   ```bash
cdswctl runtimes list | jq '.runtimes[] | select((.editor == "Workbench") and (.kernel | contains("Python")))'
   ```

4. Create a model by using the following command:

   **Note:** The following examples are specific to projects configured to use legacy engines and projects configured to use runtimes. Be sure to use the commands appropriate to your project configuration.

   For projects configured to use legacy engines:

   ```bash
cdswctl models create
   --kernel="python3"
   --targetFilePath="predict.py"
   --targetFunctionName="predict"
   --name="Petal Length Predictor"
   --cpuMillicores=1000
   --memoryMb=2000
   --description="Model of the Iris dataset"
   --replicationType=fixed
   --numReplicas=1
   --visibility="private"
   --autoBuildModel
   --autoDeployModel
   --projectId=<project ID>
   --examples='{"request":{"petal_length":1}}'
   ```
For projects configured to use runtimes:

```bash
cdswctl models create
--targetFilePath="predict.py"
--targetFunctionName="predict"
--name="Petal Length Predictor"
--cpuMillicores=1000
--memoryMb=2000
--description="Model of the Iris dataset"
--replicationType=fixed
--numReplicas=1
--visibility=private
--autoBuildModel
--autoDeployModel
--projectId=<project ID>
--runtimeId=<runtime ID obtained above>
```

If the command runs successfully, the system displays the model details in a JSON format.

5. For more information about the `models create` command parameters, run the following command:

```bash
cdswctl models create --help
```

### Build and deployment commands for models

Models have separate parameters for builds and deployments. When a model is built, an image is created. Whereas, the deployment is the actual instance of the model. You can list model builds and deployment, and monitor their state using from model CLI client (cdswctl).

#### Listing a model

To list the models, run the following command:

```bash
cdswctl models list
```

#### Monitoring the status of the model

To monitor the status of the build for a particular model, use the following command:

```bash
cdswctl models listBuild --modelId <model_ID> --projectId <project_ID>
```

You can use the `--latestModelDeployment` flag to get the build for the latest deployment.

#### Listing a deployment

To list the deployment for a particular model, run the following command:

```bash
cdswctl models listDeployments --modelId <model_ID>
```

#### Checking the status of a deployment

To check the status of your deployment, run the following command:

```bash
cdswctl models listDeployments --statusSet=deployed
```

Following is a list of arguments for the statusSet parameter:
• deployed
• deploying
• failed
• pending
• stopping
• stopped

Note: You can use the parameter more than once in a command to check multiple statuses of your deployed models. For example,

cdswctl models listDeployments --statusSet=deployed --statusSet=stopped --statusSet=failed

Deploy a new model with updated resources

You can republish a previously-deployed model in a new serving environment with an updated number of replicas or memory/CPU/GPU allocation by providing the model build ID of the model you want to rebuild.

To deploy a new model, use the following command:

cdswctl models deploy --modelBuildId=<build_ID> --cpuMilli cores=<num_of_cpu_cores> --memoryMb=<memory_in_mb> --numRepli cas=<num_of_replicas> --replicationType=<replication_type>

For example:

cdswctl models deploy --modelBuildId=<build_ID> --cpuMillicores=1200 --memoryMb=2200 --numReplicas=2 --replicationType=fixed

Note: You must specify values for all the non-zero resources, even if you do not wish to update their values. For example, in your existing deployment, if you set the cpuMillicores capacity to 1200 and you do not wish to increase or decrease it, you must still specify cpuMillicores=1200 in the command.

View replica logs for a model

When a model is deployed, Cloudera Machine Learning enables you to specify the number of replicas that must be deployed to serve requests. If a replica crashes or fails to come up, you can diagnose it by viewing the logs for every replica using the model CLI.

Procedure

1. Obtain the modelReplicaId by using the following command:

   cdswctl models listReplicas --modelDeploymentId=<model_deployment_ID>

   where the model_deployment_ID is the ID of a successfully deployed model.

2. To view the replica logs, run the following command:

   cdswctl models getReplicaLogs --modelDeploymentId=<model_deployment_ID> --modelReplicaId="<replica_ID>" --streams=stdout

   For example:

   cdswctl models getReplicaLogs --modelDeploymentId=2 --modelReplicaId="petal-length-predictor-1-2-6d6496b467-hp6tz" --streams=stdout

   The valid values for the streams parameter are stdout and stderr.
Using ML Runtimes with cdswctl

If a project is configured to use Runtimes, `cdswctl` workflows for starting sessions or models are slightly different.

**Querying the engine type**

You can query whether a project is configured using ML Runtimes or Legacy Engine.

**Procedure**

To determine if a project is configured to use either ML Runtimes or Legacy Engines, use the `cdswctl projects getEngineType` command and specify the project with the `-p` parameter.

For example, to determine if configured to use ML Runtimes:

```
cdswctl projects getEngineType -p demouser/runtimeproject
ml_runtime
```

```
cdswctl projects getEngineType -p demouser/legacyproject
legacy_engine
```

**Listing runtimes**

The first step to working with projects using runtimes is to query the available runtimes using the `cdswctl runtimes list` command.

**About this task**

The `cdswctl runtimes list` command returns all runtimes in a large JSON result. For easier consumption, you can post-process this result with some 3rd-party tool, such as `jq` or Python's `json.tool`.

**Procedure**

To query the available runtimes, use the `cdswctl runtimes list` command.

⚠️ **Note:** The following examples are for presentation purposes only. Neither Python's `json.tool` nor `jq` are supported directly by Cloudera.

The following example pipes the `cdswctl runtimes list` result through Python's `json.tool` to produce a more readable output:

```bash
user@host:~ $ cdswctl runtimes list | python3 -m json.tool
{
   "runtimes": [
   {
      "id": 1,
      "imageIdentifier": "docker.repository.cloudera.com/cdsw/ml-runtime-workbench-python3.6-standard:2020.11.1-b6",
      "editor": "Workbench",
      "kernel": "Python 3.6",
      "edition": "Standard",
      "shortVersion": "2020.11",
      "fullVersion": "2020.11.1-b6",
      "maintenanceVersion": 1,
      "description": "Standard edition Python runtime provided by Cloudera"
   },
   {
      "id": 2,
      "imageIdentifier": "docker.repository.cloudera.com/cdsw/ml-runtime-jupyterlab-python3.7-standard:2020.11.1-b6",
      "editor": "JupyterLab",
      "kernel": "Python 3.7",
      "edition": "Technical Preview",
```

```
The following example pipes the `cdswctl runtimes list` result through `jq` to transform the JSON output into arbitrary formats:

```
user@host:~ $ cdswctl runtimes list | jq -r '.runtimes[] | 
  (.id) (; .imageIdentifier)''
```

```
1 docker.repository.cloudera.com/cdsw/ml-runtime-workbench-python3.6-standard:2020.11.1-b6
2 docker.repository.cloudera.com/cdsw/ml-runtime-jupyterlab-python3.7-standard:2020.11.1-b6
```

The following example filters the `cdswctl runtimes list` result using `jq` to only show runtimes with specific editors and kernels:

```
user@host:~ $ cdswctl runtimes list | jq -r '.runtimes[] | select((.editor == "Workbench") and (.kernel | contains("Python")))''
```

```
{
  "id": 1,
  "imageIdentifier": "docker.repository.cloudera.com/cdsw/ml-runtime-workbench-python3.6-standard:2020.11.1-b6",
  "editor": "Workbench",
  "kernel": "Python 3.6",
  "edition": "Standard",
  "shortVersion": "2020.11",
  "fullVersion": "2020.11.1-b6",
  "maintenanceVersion": 1,
  "description": "Standard edition Python runtime provided by Cloudera"
}
```

## Starting sessions and creating SSH endpoints

Once you choose a runtime, you can start a session using the `cdswctl sessions start` command and create SSH endpoints using the `cdswctl ssh-endpoint` command.

### About this task

The runtime ID used in the following steps is obtained using the steps outlined in *Listing runtimes*.

### Procedure

1. To start a session with a runtime, use the `cdswctl sessions start` command, specifying the runtime ID with the `-r` parameter and the project with the `-p` parameter.

   For example:

```
    cdswctl sessions start -r 2 -p demouser/runtimeproject
```

2. To specify SSH endpoints for the runtimes sessions, use the `cdswctl ssh-endpoint` command and specify the runtime ID using the `-r` parameter, and the project with the `-p` parameter. For example:

```
  cdswctl ssh-endpoint -r 1 -p demouser/runtimeproject
```

**Creating a model**

Creating a model in a project that uses runtimes is similar to model creation with an legacy engine, but you must use a different parameter to specify the runtime ID.

**About this task**

To create a model in a project that uses runtimes you must use the `--runtimeId=` parameter to specify a runtime ID (instead of using the `--engineImageId=` and `--kernel=` parameters used for a legacy engine).

**Procedure**

To create a model in a project that uses runtimes use the `--runtimeId=` parameter to specify a runtime ID. For example:

```
  cdswctl models create --targetFilePath=predict.py --targetFunctionName=predict --projectId=4 --name=created-using-cdswctl --description=created-using-cdswctl --memoryMb=1024 --authEnabled --cpuMilliCores=250 --autoBuildModel --autoDeployModel --examples='{"request":{"petal_length":1}}' --runtimeId=1
```

**cdswctl command reference**

You can manage your Cloudera Machine Learning Workbench cluster with the CLI client (cdswctl) that exists within the Cloudera Machine Learning Workbench. Running `cdswctl` without any arguments prints a brief description of each command.

**Table 30: Model CLI Command Reference**

<table>
<thead>
<tr>
<th>Command</th>
<th>Description and usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>cdswctl login</td>
<td>Enables you to log into the model CLI client</td>
</tr>
<tr>
<td>cdswctl projects list</td>
<td>Lists the projects</td>
</tr>
<tr>
<td>cdswctl models create</td>
<td>Creates a model with the specified parameters</td>
</tr>
<tr>
<td>cdswctl models list</td>
<td>Lists all models</td>
</tr>
<tr>
<td></td>
<td>You can refine the search by specifying the <code>modelId</code></td>
</tr>
<tr>
<td>cdswctl models listBuild</td>
<td>Lists the builds for a model</td>
</tr>
<tr>
<td></td>
<td>You can monitor the status of the build by specifying the <code>modelId</code> and the <code>projectId</code></td>
</tr>
<tr>
<td>cdswctl models listDeployments</td>
<td>List the deployments for a model</td>
</tr>
<tr>
<td></td>
<td>You can refine the search by specifying the <code>modelId</code></td>
</tr>
<tr>
<td></td>
<td>Use the <code>statusSet</code> parameter to check the status of the model being deployed</td>
</tr>
</tbody>
</table>
### Reference

#### Command Description and usage

<table>
<thead>
<tr>
<th>Command</th>
<th>Description and usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>cdswctl models deploy</code></td>
<td>Deploys a model with the specified parameters</td>
</tr>
<tr>
<td><code>cdswctl models listReplicas</code></td>
<td>Enables you to view the list of model replicas</td>
</tr>
<tr>
<td></td>
<td>You also need this information to obtain replica logs</td>
</tr>
<tr>
<td><code>cdswctl models getReplicaLogs</code></td>
<td>Enables you to view the logs for a model replica</td>
</tr>
<tr>
<td><code>cdswctl models restart</code></td>
<td>Restarts a model</td>
</tr>
<tr>
<td></td>
<td>Usage:</td>
</tr>
<tr>
<td></td>
<td><code>cdswctl models restart --modelDeploymentId=&lt;deployment_ID&gt;</code></td>
</tr>
<tr>
<td></td>
<td>Note: Running this command does not change the resources if you previously ran the</td>
</tr>
<tr>
<td></td>
<td><code>cdswctl models update</code></td>
</tr>
<tr>
<td><code>cdswctl models update</code></td>
<td>Changes the name, description, or visibility of the model</td>
</tr>
<tr>
<td></td>
<td>To change a model’s resources, use the <code>cdswctl models deploy</code> command</td>
</tr>
<tr>
<td><code>cdswctl models delete</code></td>
<td>Deletes a model</td>
</tr>
<tr>
<td></td>
<td>Usage:</td>
</tr>
<tr>
<td></td>
<td><code>cdswctl models delete --id=&lt;model_ID&gt;</code></td>
</tr>
</tbody>
</table>

### Azure NetApp Files Management with the CLI

You can manage the NetApp Files setup using the CLI. This can be helpful for automating setup and teardown of workspaces as ML project needs change.

#### Create an Azure NetApp Files account

The following code sample creates an Azure NetApp Files account.

```bash
az netappfiles account create \
  --account-name my-anf-account \
  --resource-group my-cdp-resource-group \
  --location westus2
```

#### Create a capacity pool

A capacity pool is a storage container for volumes, which are accessed directly by CML. The minimum size for an Azure NetApp Files capacity pool is 4 TiB.

```bash
MINIMUM_POOL_SIZE=4 # 4 TiB is the minimum az netappfiles pool create \
  --account-name my-anf-account \
  --pool-name my-anf-pool \
  --resource-group my-cdp-resource-group \
  --service-level Standard \
  --location westus2 \
  --size ${MINIMUM_POOL_SIZE}
```

#### Create a volume

Create one or more volumes in the capacity pool. The "Usage threshold" is referred to as the "quota" in the Azure web portal. It is measured in GiB. The volume must support the NFSv3 protocol (which is the default).

```bash
az netappfiles volume create \
  --account-name my-anf-account \
  --pool-name my-anf-account \
  --volume-name my-anf-volume \
  --resource-group my-cdp-resource-group \
  --location westus2 \
  --file-path my-anf-volume \
  --usage-threshold 1000 \
  --vnet my-cdp-vnet
```
The mount path for this volume, or a dedicated, empty subdirectory inside that volume, must be provided for the "Existing NFS" field when provisioning CML workspaces. It can be found in the "Mount Instructions" blade of the volume in the Azure portal.

Since each capacity pool has a large minimum, and each volume requires a dedicated subnet, users may wish to have a single volume that is shared between workspaces. This can be managed by having a VM that has the Azure volume mounted (instructions for doing this are also in the "Mount Instructions" blade of the volume in the Azure portal). This VM can then be used to quickly manage directories for individual workspaces on a single, shared volume. For instance:

```
USER= # username for accessing management VM
VM= # IP address or hostname for accessing management VM
VOLUME= # NFS volume name
WORKSPACE= # CML workspace name (or other unique directory name)
ssh ${USER}@${VM} "sudo mkdir ${VOLUME}/${WORKSPACE}; sudo chown 8536:8 536 ${VOLUME}/${WORKSPACE}"
# ...
ssh ${USER}@${VM} "sudo rm -r ${VOLUME}/${WORKSPACE}"
```

### Accessing Local Data from Your Computer

This topic includes code samples that demonstrate how to access local data for CML workloads.

If you want to perform analytics operations on existing data files (.csv, .txt, etc.) from your computer, you can upload these files directly to your Cloudera Machine Learning project. Go to the project's Overview page. Under the Files section, click Upload and select the relevant data files to be uploaded.

The following sections use the `tips.csv` dataset to demonstrate how to work with local data stored within your project. Upload this dataset to the `data` folder in your project before you run these examples.

#### Pandas (Python)

```python
import pandas as pd
tips = pd.read_csv('data/tips.csv')
tips |
     .query('sex == "Female"') |
     .groupby('day') |
     .agg({"tip": 'mean'}) |
     .rename(columns={"tip": 'avg_tip_dinner'}) |
     .sort_values('avg_tip_dinner', ascending=False)
```

#### dplyr (R)

```r
library(readr)
library(dplyr)

# load data from .csv file in project
tips <- read_csv("data/tips.csv")

# query using dplyr
tips %>%
     filter(sex == "Female") %>%
```

172
Accessing Data from HDFS

There are many ways to access HDFS data from R, Python, and Scala libraries. The following code samples demonstrate how to count the number of occurrences of each word in a simple text file in HDFS.

Navigate to your project and click Open Workbench. Create a file called sample_text_file.txt and save it to your project in the data folder. Now write this file to HDFS. You can do this in one of the following ways:

- Click Terminal above the Cloudera Machine Learning console and enter the following command to write the file to HDFS:

  ```bash
  hdfs dfs -put data/sample_text_file.txt /tmp
  ```

- Use the workbench command prompt:

  **Python Session**

  ```bash
  !hdfs dfs -put data/sample_text_file.txt /tmp
  ```

  **R Session**

  ```r
  system("hdfs dfs -put data/tips.csv /user/hive/warehouse/tips/")
  ```

The following examples use Python and Scala to read sample_text_file.txt from HDFS (written above) and perform the count operation on it.

**Python**

```python
from __future__ import print_function
import sys, re
from operator import add
from pyspark.sql import SparkSession

spark = SparkSession\
    .builder\
    .appName("PythonWordCount")\
    .getOrCreate()

# Access the file
lines = spark.read.text("/tmp/sample_text_file.txt").rdd.map(lambda r: r[0])
counts = lines.flatMap(lambda x: x.split(' ')) \
    .map(lambda x: (x, 1)) \
    .reduceByKey(add) \
    .sortBy(lambda x: x[1], False)
output = counts.collect()
for (word, count) in output:
    print("%s: %i" % (word, count))

spark.stop()
```

**Scala**

```scala
//count lower bound
```
val threshold = 2

// read the file added to hdfs
val tokenized = sc.textFile("/tmp/sample_text_file.txt").flatMap(_.split(" "))

// count the occurrence of each word
val wordCounts = tokenized.map((_, 1)).reduceByKey(_ + _)

// filter out words with fewer than threshold occurrences
val filtered = wordCounts.filter(_._2 >= threshold)
System.out.println(filtered.collect().mkString(","))

---

**Accessing Data from Apache Hive**

The following code sample demonstrates how to establish a connection with the Hive metastore and access data from tables in Hive.

**Python**

```python
import os
import pandas
from impala.dbapi import connect
from impala.util import as_pandas

# Specify HIVE_HMS_HOST as an environment variable in your project settings
HIVE_HMS_HOST = os.getenv('HIVE_HS2_HOST', '<hiveserver2_hostname>

# This connection string depends on your cluster setup and authentication mechanism
conn = connect(host=HIVE_HS2_HOST,
                port='10000',
                auth_mechanism='GSSAPI',
                kerberos_service_name='hive')

cursor = conn.cursor()
cursor.execute('SHOW TABLES')
tables = as_pandas(cursor)
tables
```

---

**Accessing Data from Apache Impala**

In this section, we take some sample data in the form of a CSV file, save the contents of this file to a table in Impala, and then use some common Python and R libraries to run simple queries on this data.

**Loading CSV Data into an Impala Table**

For this demonstration, we will be using the `tips.csv` dataset. Use the following steps to save this file to a project in Cloudera Machine Learning, and then load it into a table in Apache Impala.

1. Create a new Cloudera Machine Learning project.
2. Create a folder called `data` and upload `tips.csv` to this folder. For detailed instructions, see `Managing Project Files`.
3. The next steps require access to services on the CDH cluster. If Kerberos has been enabled on the cluster, enter your credentials (username, password/keytab) in Cloudera Machine Learning to enable access.
4. Navigate back to the project Overview page and click `Open Workbench`.
5. Launch a new session (Python or R).
6. Open the Terminal.

   a. Run the following command to create an empty table in Impala called tips. Replace `<impala_daemon_hostname>` with the hostname for your Impala daemon.

      ```sql
      impala-shell -i <impala_daemon_hostname>:21000 -q 'CREATE TABLE default.tips (
      `total_bill` FLOAT,
      `tip` FLOAT,
      `sex` STRING,
      `smoker` STRING,
      `day` STRING,
      `time` STRING,
      `size` TINYINT
      ) ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LOCATION "hdfs:///user/hive/warehouse/tips/;"
      '"
      
   b. Run the following command to load data from the /data/tips.csv file into the Impala table.

      ```bash
      hdfs dfs -put data/tips.csv /user/hive/warehouse/tips/
      ```

Running Queries on Impala Tables

This section demonstrates how to run queries on the tips table created in the previous section using some common Python and R libraries such as Pandas, Impyla, Sparklyr and so on. All the examples in this section run the same query, but use different libraries to do so.

**PySpark (Python)**

```python
from pyspark.sql import SparkSession

spark = SparkSession.builder.master('yarn').getOrCreate()
# load data from .csv file in HDFS
# tips = spark.read.csv("/user/hive/warehouse/tips/", header=True, inferSchema=True)
# OR load data from table in Hive metastore
tips = spark.table('tips')

from pyspark.sql.functions import col, lit, mean

# query using DataFrame API
tips \
  .filter(col('sex').like("%Female%")) \
  .groupBy('day') \
  .agg(mean('tip').alias('avg_tip')) \
  .orderBy('avg_tip',ascending=False) \
  .show()

# query using SQL
spark.sql('''
SELECT day,AVG(tip) AS avg_tip 
FROM tips 
WHERE sex LIKE "%Female%" 
GROUP BY day 
ORDER BY avg_tip DESC''').show()

spark.stop()
```
**Impyla (Python)**

Due to an incompatibility with the `thrift_sasl` package, Impyla has been known to fail with Python 3.

**Python 2**

```python
# (Required) Install the impyla package
# !pip install impyla
# !pip install thrift_sasl
import os
import pandas
from impala.dbapi import connect
from impala.util import as_pandas

# Connect to Impala using Impyla
# Secure clusters will require additional parameters to connect to Impala.
# Recommended: Specify IMPALA_HOST as an environment variable in your project settings
IMPALA_HOST = os.getenv('IMPALA_HOST', '<impala_daemon_hostname>')
conn = connect(host=IMPALA_HOST, port=21050)
# Execute using SQL
cursor = conn.cursor()
cursor.execute('SELECT day,AVG(tip) AS avg_tip 
    FROM tips 
    WHERE sex ILIKE "%Female%" 
    GROUP BY day 
    ORDER BY avg_tip DESC')

# Pretty output using Pandas
tables = as_pandas(cursor)
tables
```

**Ibis (Python)**

```python
# (Required) Install the ibis-framework[impala] package
# !pip3 install ibis-framework[impala]
import ibis
import os
ibis.options.interactive = True
ibis.options.verbose = True

# Connection to Impala
# Secure clusters will require additional parameters to connect to Impala.
# Recommended: Specify IMPALA_HOST as an environment variable in your project settings
IMPALA_HOST = os.getenv('IMPALA_HOST', '<impala_daemon_hostname>')
con = ibis.impala.connect(host=IMPALA_HOST, port=21050, database='default')
con.list_tables()
tips = con.table('tips')
tips 
    .filter(tips.sex.like(['%Female%'])) 
    .group_by('day') 
    .aggregate(avg_tip=tips.tip.mean()) 
    .sort_by(ibis.desc('avg_tip')) 
    .execute()
```
Sparklyr (R)

```r
# (Required) Install the sparklyr package
# install.packages("sparklyr")

library(stringr)
library(sparklyr)
library(dplyr)

spark <- spark_connect(master = "yarn")

# load data from file in HDFS
tips <- spark_read_csv(
  sc = spark,
  name = "tips",
  path = "/user/hive/warehouse/tips/"
)

# OR load data from table
tips <- tbl(spark, "tips")

# query using dplyr
tips %>%
  filter(sex %like% "%Female%") %>%
  group_by(day) %>%
  summarise(
    avg_tip = mean(tip, na.rm = TRUE)
  ) %>%
  arrange(desc(avg_tip))

# query using SQL
tbl(spark, sql("SELECT day,AVG(tip) AS avg_tip \nFROM tips \nWHERE sex LIKE '%Female%' \nGROUP BY day \nORDER BY avg_tip DESC")

spark_disconnect(spark)
```

Accessing Data in Amazon S3 Buckets

Every language in Cloudera Machine Learning has libraries available for uploading to and downloading from Amazon S3.

To work with S3:

1. Add your Amazon Web Services access keys to your project's environment variables as AWS_ACCESS_KEY_ID and AWS_SECRET_ACCESS_KEY.
2. Pick your favorite language from the code samples below. Each one downloads the R 'Old Faithful' dataset from S3.

```r
library("devtools")
install_github("armstrtw/AWS.tools")

Sys.setenv("AWSACCESSKEY"=Sys.getenv("AWS_ACCESS_KEY_ID"))
Sys.setenv("AWSSECRETKEY"=Sys.getenv("AWS_SECRET_ACCESS_KEY"))
library("AWS.tools")
```
s3.get("s3://sense-files/faithful.csv")

Python

# Install Boto to the project
!pip install boto
# Create the Boto S3 connection object.
from boto.s3.connection import S3Connection
aws_connection = S3Connection()

# Download the dataset to file 'faithful.csv'.
bucket = aws_connection.get_bucket('sense-files')
key = bucket.get_key('faithful.csv')
key.get_contents_to_filename('/home/cdsw/faithful.csv')

Accessing External SQL Databases

Every language in Cloudera Machine Learning has multiple client libraries available for SQL databases.

If your database is behind a firewall or on a secure server, you can connect to it by creating an SSH tunnel to the server, then connecting to the database on localhost.

If the database is password-protected, consider storing the password in an environmental variable to avoid displaying it in your code or in consoles. The examples below show how to retrieve the password from an environment variable and use it to connect.

Python

You can access data using pyodbc or SQLAlchemy

# pyodbc lets you make direct SQL queries.
!wget https://pyodbc.googlecode.com/files/pyodbc-3.0.7.zip
!unzip pyodbc-3.0.7.zip
!cd pyodbc-3.0.7;python setup.py install --prefix /home/cdsw
import os

# See http://www.connectionstrings.com/ for information on how to construct ODBC connection strings.
db = pyodbc.connect("DRIVER={PostgreSQL Unicode};SERVER=localhost;PORT=5432;DATABASE=test_db;USER=cdswuser;OPTION=3;PASSWORD=%s" % os.environ["POSTGRES_PASSWORD"])
cursor = cnxn.cursor()
cursor.execute("select user_id, user_name from users")

# sqlalchemy is an object relational database client that lets you make database queries in a more Pythonic way.
!pip install sqlalchemy
import os

import sqlalchemy
from sqlalchemy.orm import sessionmaker
from sqlalchemy import create_engine
db = create_engine("postgresql://cdswuser:%s@localhost:5432/test_db" % os.environ["POSTGRES_PASSWORD"])
session = sessionmaker(bind=db)
user = session.query(User).filter_by(name='ed').first()
R

# dplyr lets you program the same way with local data frames and remote SQL databases.
install.packages("dplyr")
library("dplyr")
db <- src_postgres(dbname="test_db", host="localhost", port=5432, user="cdswuser", password=Sys.getenv("POSTGRESQL_PASSWORD"))
flights_table <- tbl(db, "flights")
select(flights_table, year:day, dep_delay, arr_delay)

Accessing a Data Lake from CML

CML can access data tables stored in an AWS or Microsoft Azure Data Lake. As a CML Admin, follow this procedure to set up the necessary permissions.

About this task

The instructions apply to Data Lakes on both AWS and Microsoft Azure. Follow the instructions that apply to your environment.

Procedure

1. Cloud Provider Setup

   Make sure the prerequisites for AWS or Azure are satisfied (see the Related Topics for AWS environments and Azure environments). Then, create a CDP environment as follows.

   a) For environment logs, create an S3 bucket or ADLS Gen2 container.

      Note: For ADLS Gen2, create a dedicated container for logs, and a dedicated container for data, within the same account.

   b) For environment storage, create an S3 bucket or ADLS Gen2 container.

   c) For AWS, create AWS policies for each S3 bucket, and create IAM roles (simple and instance profiles) for these policies.

   d) For Azure, create managed identities for each of the personas, and create roles to map the identities to the ADLS permissions.

   For detailed information on S3 or ADLS, see Related information.
2. Environment Setup

In CDP, set up paths for logs and native data access to the S3 bucket or ADLS Gen2 container.

In the Environment Creation wizard, set the following:

## Logs Storage and Audits

Provide an existing location where log files will be stored.

**Select an Instance Profile**

*Click here* to refresh instance profiles from the cloud provider.

**Log Location Base**

```
MLX_DEV_DATALAKE_LOG_ROLE
```

**Ranger Audit Role**

```
arn:aws:iam::886883559913:role/mlx-dev-prod-env_RANGER_AUD
```

### a) Logs Storage and Audits

1. **Instance Profile** - The IAM role or Azure identity that is attached to the master node of the Data Lake cluster. The Instance Profile enables unauthenticated access to the S3 bucket or ADLS container for logs.
2. **Log Location Base** - The location in S3 or ADLS where environment logs are saved.

**Note:** The instance profile or Azure identity must refer to the same logs location base in S3 or ADLS.

3. **Ranger Audit Role** - The IAM role or Azure identity that has S3 or ADLS access to write Ranger audit events. Ranger uses Hadoop authentication, therefore it uses IDBroker to access the S3 bucket or ADLS container, rather than using Instance profiles or Azure identities directly.

### b) Data Access
1. **Instance Profile** - The IAM role or Azure identity that is attached to the IDBroker node of the Data Lake cluster. IDBroker uses this profile to assume roles on behalf of users and get temporary credentials to access S3 buckets or ADLS containers.

2. **Storage Location Base** - The S3 or ADLS location where data pertaining to the environment is saved.

3. **Data Access Role** - The IAM role or Azure identity that has access to read or write environment data. For example, Hive creates external tables by default in the CDP environments, where metadata is stored in HMS running in the Data Lake. The data itself is stored in S3 or ADLS. As Hive uses Hadoop authentication, it uses IDBroker to access S3 or ADLS, rather than using Instance profiles or Azure identities. Hive uses the data access role for storage access.

   **Note:** The data access role must have permissions to access the S3 or ADLS storage location.

4. **ID Broker Mappings** - These specify the mappings between the CDP user or groups to the AWS IAM roles or Azure roles that have appropriate S3 or ADLS access. This setting enables IDBroker to get appropriate S3 or ADLS credentials for the users based on the role mappings defined.

   **Note:** There is no limit to the number of mappings that one can define but each user can only be assigned to one of the role mappings.

This completes installation of the environment.
3. User Group Mappings

In CDP, you can assign users to groups to simplify permissions management. For example, you could create a group called ml-data-scientists, and assign two individual users to it, as shown here. For instructions, see link.

![User Management Console](image)

<table>
<thead>
<tr>
<th>Groups</th>
<th>ml-data-scientists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>ml-data-scientists</td>
</tr>
<tr>
<td>CRN</td>
<td>ormsuitas iam.us-west-1.mdl4kw9c3a86m021-dk21-f244a69da74e57-group-1</td>
</tr>
<tr>
<td>Creation Date</td>
<td>06/04/2019 11:50 AM PDT</td>
</tr>
<tr>
<td>Sync Membership On User Login</td>
<td>Yes</td>
</tr>
<tr>
<td>Members</td>
<td>2</td>
</tr>
</tbody>
</table>

### Members

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nagare Prasad Adhikary</td>
<td><a href="mailto:ram@cloudera.com">ram@cloudera.com</a></td>
</tr>
<tr>
<td></td>
<td>Vansh Verkadee</td>
<td><a href="mailto:vansh@cloudera.com">vansh@cloudera.com</a></td>
</tr>
</tbody>
</table>

### a. Sync users

Whenever you make changes to user and group mappings, make sure to sync the mappings with the authentication layer. In User Management > Actions, click Sync Users, and select the environment.

![User Management Console](image)

### 4. IDBroker

IDBroker allows an authenticated and authorized user to exchange a set of credentials or a token for cloud vendor access tokens. You can also view and update the IDBroker mappings at this location. IDBroker mappings can be accessed through Environments > Manage Access. Click on the IDBroker Mappings tab. Click Edit to edit or add...
mappings. When finished, sync the mappings to push the settings from CDP to the IDBroker instance running inside the Data Lake of the environment.

At this point, CDP resources can access the AWS S3 buckets or Azure ADLS storage.

5. Ranger

To get admin access to Ranger, users need the EnvironmentAdmin role, and that role must be synced with the environment.

a. Click Environments > Env > Actions > Manage Access > Add User
b. Select EnvironmentAdmin resource role.
c. Click Update Roles
d. On the Environments page for the environment, in Actions, select Synchronize Users to FreeIPA.

The permissions are now synchronized to the Data Lake, and you have admin access to Ranger.

Update permissions in Ranger

a. In Environments > Env > Data Lake Cluster, click Ranger.
b. Select the Hadoop SQL service, and check that the users and groups have sufficient permissions to access databases, tables, columns, and urls.

For example, a user can be part of these policies:

- all - database, table, column
- all - url

This completes all configuration needed for CML to communicate with the Data Lake.

6. CML User Setup

Now, CML is able to communicate with the Data Lake. There are two steps to get the user ready to work.

a. In Environments > Environment name > Actions > Manage Access > Add user, the Admin selects MLUser resource role for the user.
b. The User logs into the workspace in ML Workspaces > Workspace name, click Launch Workspace.

The user can now access the workspace.

Related Information
AWS environments
Azure environments
Example: Connect a Spark session to Hive Metastore in a Data Lake

After the Admin sets up the correct permissions, you can access the Data Lake from your project, as this example shows.

Before you begin

Make sure you have access to a Data Lake containing your data.

Procedure

1. Create a project in your ML Workspace.
2. Create a file named spark-defaults.conf, or update the existing file with the property:
   - For S3: `spark.yarn.access.hadoopFileSystems=s3a://STORAGE LOCATION OF ENV>
   - For ADLS: `spark.yarn.access.hadoopFileSystems=abfs://STORAGE CONTAINER OF ENV>@STORAGE ACCOUNT OF ENV`

   Use the same location you defined in Data Access.
3. Start a session (Python or Spark) and start a Spark session.

Results

Setting up the project looks like this:

```
CML / SDX Interaction ➤ Running
By Varnsee Yarlagadda — Python 3 Session — 1 vCPU / 2 GiB Memory — 6 minutes ago

Session Logs Spark UI
> from pyspark.sql import SparkSession
> spark = SparkSession(
    .builder()
    .appName('PythonPi')
    .getOrCreate())

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/etc/spark/jars/slf4j-log4j12-1.7.16.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Setting spark.hadoop.yarn.resourcemanager.principal to csso_vamsee

> spark.sql("show databases").show()
+-------------+---------------+--------------+-------------------+--------+-------------------+--------------------------------------------------+
| Hive Session ID | 76b46c9f-6497-971b-4ebf-9a62-6b37ebc7192d |
| databaseName | +-------------------+-------------------+-------------------+--------+-------------------+
| | +-------------------+-------------------+-------------------+--------+-------------------+
| | default | information_schema | sys | test_alpha | tpcds_bin_parti... | vamsee |
+-------------------+-------------------+-------------------+--------+-------------------+
```

Now you can run Spark SQL commands. For example, you can:
• Create a database `foodb`.

```sql
spark.sql("create database foodb").show()
```

• List databases and tables.

```sql
spark.sql("use foodb").show()
```

```sql
spark.sql("show tables").show()
```

<table>
<thead>
<tr>
<th>database</th>
<th>tableName</th>
<th>isTemporary</th>
</tr>
</thead>
<tbody>
<tr>
<td>----------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
• Create a table bartable.

CML / SDX Interaction  Running
By Vamsee Yarlagadda — Python 3 Session — 1 vCPU / 2 GiB Memory — 6 minutes ago

<table>
<thead>
<tr>
<th>Session</th>
<th>Logs</th>
<th>Spark UI</th>
</tr>
</thead>
</table>

```python
spark.sql("create table bartable(id int)").show()
```

```text
++
||
++
```

```python
spark.sql("show tables").show()
```

```text
+------------+-----------------+---------------------+
| database   | tableName       | isTemporary         |
+------------+-----------------+---------------------+
| foodb      | bartable        | false              |
+------------+-----------------+---------------------+
```

• Insert data into the table.

CML / SDX Interaction  Running
By Vamsee Yarlagadda — Python 3 Session — 1 vCPU / 2 GiB Memory — just now

<table>
<thead>
<tr>
<th>Session</th>
<th>Logs</th>
<th>Spark UI</th>
</tr>
</thead>
</table>

```python
spark.sql("insert into bartable values(1)").show()
```

```text
++
||
++
```
Accessing Ozone from Spark

In CML, you can connect Spark to the Ozone object store with a script. The following example demonstrates how to do this.

This script, in Scala, counts the number of word occurrences in a text file. The key point in this example is to use the following string to refer to the text file: o3fs://hivetest.s3v.o3service1/spark/jedi_wisdom.txt

Word counting example in Scala

```scala
import sys.process._

// Put the input file into Ozone
"hdfs dfs -put data/jedi_wisdom.txt o3fs://hivetest.s3v.o3service1/spark"

// Set the following spark setting in the file "spark-defaults.conf" on the CML session using terminal
spark.yarn.access.hadoopFileSystems=o3fs://hivetest.s3v.o3service1.neptune.e01.olympus.cloudera.com:9862

//count lower bound
val threshold = 2
// this file must already exist in hdfs, add a
// local version by dropping into the terminal.
val tokenized = sc.textFile("o3fs://hivetest.s3v.o3service1/spark/jedi_wisdom.txt").flatMap(_.split(" "))

// count the occurrence of each word
val wordCounts = tokenized.map(_ , 1).reduceByKey(_ + _)

// filter out words with fewer than threshold occurrences
val filtered = wordCounts.filter(_._2 >= threshold)

System.out.println(filtered.collect().mkString(","))
```

Built-in CML Visualizations

You can use built-in CML tools to create data visualizations including simple plots, saved images, HTML and iFrame visualizations, and grid displays.
Simple Plots

Cloudera Machine Learning supports using simple plot to create data visualizations.

To create a simple plot, run a console in your favorite language and paste in the following code sample:

R

```r
# A standard R plot
plot(rnorm(1000))
# A ggplot2 plot
library("ggplot2")
qplot(hp, mpg, data=mtcars, color=am,
    facets=gear~cyl, size=I(3),
    xlab="Horsepower", ylab="Miles per Gallon")
```

Python

```python
import matplotlib.pyplot as plt
import random
plt.plot([random.normalvariate(0,1) for i in xrange(1,1000)])
```

Cloudera Machine Learning processes each line of code individually (unlike notebooks that process code per-cell). This means if your plot requires multiple commands, you will see incomplete plots in the workbench as each line is processed.

To get around this behavior, wrap all your plotting commands in one Python function. Cloudera Machine Learning will then process the function as a whole, and not as individual lines. You should then see your plots as expected.

Saved Images

You can display images within your reports.

Use the following commands:

R

```r
library("cdsw")
download.file("https://upload.wikimedia.org/wikipedia/commons/2/29/Minard.png", "/cdn/Minard.png")
image("Minard.png")
```

Python

```python
import urllib
from IPython.display import Image
urllib.urlretrieve("http://upload.wikimedia.org/wikipedia/commons/2/29/Minard.png", "Minard.png")
Image(filename="Minard.png")
```

HTML Visualizations

Your code can generate and display HTML in Cloudera Machine Learning.

To create an HTML widget, paste in the following:

R

```r
library("cdsw")
html('<svg><circle cx="50" cy="50" r="50" fill="red" /></svg>')
```
Python

```python
from IPython.display import HTML
HTML('<svg><circle cx="50" cy="50" r="50" fill="red" /></svg>')
```

Scala

Cloudera Machine Learning allows you to build visualization libraries for Scala using `jvm-repr`. The following example demonstrates how to register a custom HTML representation with the "text/html" mimetype in Cloudera Machine Learning. This output will render as HTML in your workbench session.

```scala
// HTML representation
case class HTML(html: String)
// Register a displayer to render html
Displayers.register(classOf[HTML],
  new Displayer[HTML] {
    override def display(html: HTML): java.util.Map[String, String] = {
      Map("text/html" -> html.html
    ).asJava
  })
}
val helloHTML = HTML("<h1> <em> Hello World </em> </h1>"

display(helloHTML)
```

**IFrame Visualizations**

Most visualizations require more than basic HTML. Embedding HTML directly in your console also risks conflicts between different parts of your code. The most flexible way to embed a web resource is using an IFrame.

**Note:**

Cloudera Machine Learning versions 1.4.2 (and higher) added a new feature that allowed users to HTTP security headers for responses to Cloudera Machine Learning. This setting is enabled by default. However, the X-Frame-Options header added as part of this feature blocks rendering of iFrames injected by third-party data visualization libraries.

To work around this issue, a site administrator can go to the Admin > Security page and disable the Enable HTTP security headers property. Restart Cloudera Machine Learning for this change to take effect.

R

```r
library("cdsw")
iframe(src="https://www.youtube.com/embed/8pHzROP1D-w", width="854px", height="510px")
```

Python

```python
from IPython.display import HTML
HTML('<iframe width="854" height="510" src="https://www.youtube.com/embed/8pHzROP1D-w"></iframe>')
```

You can generate HTML files within your console and display them in IFrames using the `/cdn` folder. The `cdn` folder persists and services static assets generated by your engine runs. For instance, you can embed a full HTML file with IFrames.

R

```r
library("cdsw")
f <- file("/cdn/index.html")
```
Cloudera Machine Learning uses this feature to support many rich plotting libraries such as htmlwidgets, Bokeh, and Plotly.

**Grid Displays**

Cloudera Machine Learning supports built-in grid displays of DataFrames across several languages.

**Python**

Using DataFrames with the pandas package requires per-session activation:

```python
import pandas as pd
pd.DataFrame(data=[range(1,100)])
```

For PySpark DataFrames, use pandas and run `df.toPandas()` on a PySpark DataFrame. This will bring the DataFrame into local memory as a pandas DataFrame.

**Note:**

A Python project originally created with engine 1 will be running pandas version 0.19, and will not auto-upgrade to version 0.20 by simply selecting engine 2 in the project's Settings > Engine page.

The pandas data grid setting only exists starting in version 0.20.1. To upgrade, manually install version 0.20.1 at the session prompt.

```
!pip install pandas==0.20.1
```

**R**

In R, DataFrames will display as grids by default. For example, to view the Iris data set, you would just use:

```
iris
```

Similar to PySpark, bringing Sparklyr data into local memory with `as.data.frame` will output a grid display.

```
sparkly_df %>% as.data.frame
```

**Scala**

Calling the `display()` function on an existing dataframe will trigger a collect, much like `df.show()`.

```
val df = sc.parallelize(1 to 100).toDF()
display(df)
```
Documenting Your Analysis

Cloudera Machine Learning supports Markdown documentation of your code written in comments.

This allows you to generate reports directly from valid Python and R code that runs anywhere, even outside Cloudera Machine Learning. To add documentation to your analysis, create comments in Markdown format:

R

```r
# Heading
# -------
# This documentation is **important.**
# Inline math: \( e^x \)
# Display math: \( y = \Sigma x + \epsilon \)
print("Now the code!")
```

Python

```python
# Heading
# -------
# This documentation is **important.**
# Inline math: \( e^x \)
# Display math: \( y = \Sigma x + \epsilon \)
print("Now the code!")
```

Cloudera Data Visualization for ML

Cloudera Data Visualization enables you to explore data and communicate insights across the whole data lifecycle by using visual objects. The fast and easy self-service data visualization streamlines collaboration in data analytics through the common language of visuals.

Using this rich visualization layer enables you to accelerate advanced data analysis. The web-based, no-code, drag-and-drop user interface is highly intuitive and enables you to build customized visualizations on top of your datasets, build dashboards and applications, and publish them anywhere across the data lifecycle. This solution allows for customization and collaboration, and it provides you with a dynamic and data-driven insight into your business.

In CDP Public Cloud, Data Visualization is integrated with Cloudera Machine Learning (CML). For on prem use, Data Visualization is integrated with Cloudera Data Science Workbench (CDSW) workflows. You can use the same visualization tool for structured, unstructured/text, and ML analytics, which means deeper insights and more advanced dashboard applications. You can create native data visualizations to provide easy predictive insights for business users and accelerate production ML workflows from raw data to business impact.

For more information, see the Cloudera Data Visualization documentation.

Related Information
Cloudera Data Visualization in CDP Public Cloud
Cloudera Data Visualization in CDSW

Jupyter Magic Commands
Cloudera Machine Learning’s Scala and Python kernels are based on Jupyter kernels. Jupyter kernels support varying magic commands that extend the core language with useful shortcuts. This section details the magic commands (magics) supported by Cloudera Machine Learning.

Line magics begin with a single %: for example, `%timeit`. Cell magics begin with a double %%: for example, `%%bash`.

**Python**

In the default Python engine, Cloudera Machine Learning supports most line magics, but no cell magics.

Cloudera Machine Learning supports the shell magic `!`: for example, `!ls -alh /home/cdsw`.

Cloudera Machine Learning supports the help magics `?` and `??`: for example, `?numpy` and `??numpy`. `?` displays the docstring for its argument. `??` attempts to print the source code. You can get help on magics using the `?` prefix: for example, `?%timeit`.

Cloudera Machine Learning supports the line magics listed at [https://ipython.org/ipython-doc/3/interactive/magics.html#line-magics](https://ipython.org/ipython-doc/3/interactive/magics.html#line-magics), with the following exceptions:

- `%colors`
- `%debug`
- `%edit`
- `%gui`
- `%history`
- `%install_default_config`
- `%install_profiles`
- `%lsmagic`
- `%macro`
- `%matplotlib`
- `%notebook`
- `%page`
- `%pastebin`
- `%pdb`
- `%prun`
- `%pylab`
- `%recall`
- `%rerun`
- `%save`
- `%sc`

**Related reference**

Scala

Cloudera Machine Learning’s Scala kernel is based on Apache Toree. It supports the line magics documented in the Apache Toree magic tutorial.

**Related reference**

Python
Release Notes

Find out about the latest features of each release here. Also, descriptions and workarounds for some known issues are described here.

What's New

Major features and updates for the Cloudera Machine Learning service on Private Cloud.

April 27, 2021

CML on Private Cloud, version 1.2, has the following new features and updates.

New features and updates

- Support for OCP 4.6 and upgrading from PVC 1.1.
- Improved non-transparent proxy support for air-gapped environments.
- Introduced Applied ML Prototypes (AMPs).
- Added NFS support:
  - NFS versions v3 and v4.x are supported.
  - External NFS security improvements - no_root_squash export option has been removed.
- Support added for custom service principals (Beta).
- API v2 (Beta). This is the beta release of a supported API for CML. As it is in beta, it is subject to change without warning before becoming GA.
- Monitoring now uses CDP centralized Grafana. Added database metrics and improved alerts.

Bug fixes

- DSE-12037 - Fixed an issue with the seamless login for Grafana.
- DSE-14891 - Fixed an issue with broken Engine and Session log links.
- Various security fixes.

December 16, 2020

CML on Private Cloud, version 1.1, has the following new features and updates.

- MLOPS-216 - Production ML Support
  Model Metrics track machine model serving performance metrics. Model Governance use Apache Atlas to track builds, experiments and deployment of machine learning models.
- DSE-10777 - UMS Integration
  MLUser and MLAdmin resource roles are now available and assignable through Environment settings.
- DSE-12955 - Self Signed Private CA certs For custom container registries
  Customers can now use Container registries that are using self signed or private CA signed certificates. There is an option to upload the self signed or private CA signed certificates certificate during Private Cloud installation.
- DSE-10759 - GPU support
  The OpenShift Nvidia operator is now supported for use with CML workloads.

August 17, 2020

This is the first release of CML on Private Cloud, version 1.0.

CML on Private Cloud lets you:
• Run Machine Learning workloads on OpenShift clusters in your own data center.
• Easily onboard a new tenant and provision an ML workspace in a shared OpenShift environment.
• Enable data scientists to access shared data on CDP Private Cloud Base and CDW.
• Leverage Spark-on-K8s to spin up and down Spark clusters on demand.
• Take advantage of most CML features on public cloud, including Teams, Projects, Experiments, Models, and Applications.

**Related Information**

**Known Issues and Limitations**

You might run into some known issues while using Cloudera Machine Learning on Private Cloud.

**DSE-13117: Container Image Registries assuming mutual TLS for authentication are not supported**

If Private Cloud images are hosted in an image registry assuming mutual TLS for authentication, this will cause Model deployments and Experiments to fail. Mutual TLS registries are not supported.

**DSE-12778: Unconfigured Storage in Cluster Image Registry Operator causes Certificate failures**

If storage is not configured in the OpenShift cluster image registry operator, this will prevent S2I certificates from being written to each node of the cluster, and therefore the nodes will fail to pull model or experiment images from the s2i-registry. The error seen by running `kubectl describe pod <model/experiment pod name> -n <namespace-user-<userId>>` will look similar to: `Failed to pull image x509: certificate signed by unknown authority`.

1. To confirm that storage for cluster image registry operator is configured, run `oc get configs.imageregistry.operator.openshift.io -o yaml` to check the output of the command:

```yaml
apiVersion: imageregistry.operator.openshift.io/v1
kind: Config
 [...] Config
 spec:
 [...] Spec
  managementState: Managed
 [...] ManagementState
  status:
 [...] Status
  storage:{}
 [...] Storage

If the result includes `storage: {}`, then storage is not configured.

2. Configure the storage as `emptyDir: {}` as follows.

   a. Run

```bash
oc edit configs.imageregistry.operator.openshift.io
```

   b. Navigate to the storage section under `spec:` and set it to `emptyDir: {}`, as shown here:

```yaml
apiVersion: imageregistry.operator.openshift.io/v1
kind: Config
 [...] Config
 spec:
 [...] Spec
  managementState: Managed
 [...] ManagementState
```
Release Notes

Storage:
   emptyDir: {}

Note: There is a space between emptyDir: and {}.

3. Confirm that storage is properly configured by checking the status section of the output from the command in step 1. The status section should show that Storage is configured to emptyDir.

```yaml
apiVersion: imageregistry.operator.openshift.io/v1
kind: Config
spec:
  managementState: Managed
  Storage:
    emptyDir: {}
status:

- lastTransitionTime: "2020-07-22T00:34:51Z"
  message: EmptyDir storage successfully created
  reason: Creation Successful
  status: "True"
  type: StorageExists
```

DSE-12541: Self Signed Certificates for Container Registry cause Models and Experiments to fail

If you are using self-signed or Private CA signed certificates for Container image registry authentication, model deployments and experiments will fail with an error similar to: Error initializing source docker://<registryIP>:5000/alpine:latest: error pinging docker registry <registryIP>:5000: Get https://<registryIP>:5000/v2/: x509: certificate signed by unknown authority

As a workaround, create a ConfigMap in the namespace where the CML workspace is installed.

1. Create a ConfigMap as shown in this example. Here, <namespace> indicates the workspace where the CML workspace is installed.

```yaml
kind: ConfigMap
apiVersion: v1
metadata:
  name: <external-registry-name>
  Namespace: <namespace>
data:
  registry.crt: |
    -----BEGIN CERTIFICATE-----
    < certificate content goes here >
    -----END CERTIFICATE-----
```

2. Mount the ConfigMap to the s2i-builder deployment as shown here. Add the following mountPath in the volumeMounts section for the s2i-builder pod:

```yaml
- mountPath: /etc/docker/certs.d/<registry>[:portnum]
```
name: external-registry

Under the volumes section, add the ConfigMap reference:

- configMap:
  
  - name: <external-registry-name>
    name: external-registry

3. Run the following command and check the output. Note in particular the mountPath and configMap specifications at the end.

```bash
# kubectl get deployment s2i-builder -n <namespace> -o yaml
```

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  [...]
name: s2i-builder
namespace: <namespace/workspacename>
[...]
spec:
  [...]
template:
    [...]
spec:
  [...]
containers:
  - name: s2i-builder
    [...]
    volumeMounts:
      - mountPath:/etc/docker/certs.d/<registry>[:portnum]
        name: external-registry
    volumes:
      [...]
      - configMap:
          defaultMode: 420
            name: <external-registry-name>
            name: external-registry
status:
[...]
```

**DSE-12367: s2i-queue pod goes into CrashLoop Failure causing ML workspace installation to fail**

CML workspace install can fail because the s2i-queue pod may be stuck in a CrashLoop Failure. The error in the logs might look similar to: Failed to create thread: Resource temporarily unavailable

(11) /usr/lib/rabbitmq/bin/rabbitmq-server: line 182: 45 Aborted (core dumped)
start_rabbitmq_server "$@" Only root or rabbitmq can run rabbitmq-server

To fix this, apply the following workaround:

```
# kubectl set env statefulset/s2i-queue
RABBITMQ_IO_THREAD_POOL_SIZE="50" -n <namespace>
```

**DSE-12329: Email invitation feature**

The feature to invite new users by email does not work in Public or Private cloud, but it still appears in the UI.
DSE-12289: Airgap support: Proxies are not supported in CML Private Cloud 1.0

Use of a proxy server, for example for external internet connectivity for an airgap cluster, is not supported. Transparent proxies, however, should work normally.

DSE-12238: Create Project request takes longer than timeout

If a Create Project request takes longer than a certain timeout, a second request might be submitted. If this happens, multiple projects with similar names might be created.

As a workaround, create an empty project, create a session inside the project, then `git clone` your project inside a workbench terminal. Additionally, you can upload a zip file or a folder using the file preview table.

If multiple forks are created, delete the extra ones.

DSE-12090: User displays as unknown in Event History

In the **Event History** on the workspace **Events** tab, a user may display as **unknown** if they are authenticated by LDAP.

Fix: The user needs to be assigned the **IamViewer** role to view these details.

DSE-11979: Certificate failure when pulling images from the S2I container registry

During Model or Experiment deployment, a certificate failure similar to **Failed to pull image x509: certificate signed by unknown authority** can occur. This is due to a Red Hat issue with OpenShift Container Platform 4.3.x where the image registry cluster operator configuration must be set to **Managed**.

To set the configuration, first apply a patch using this command:

```
# oc patch configs.imageregistry.operator.openshift.io cluster --patch '{"spec":
{"managementState":"Managed"}}'
```

Next, run the following command:

```
# oc get config cluster -o yaml
```

The **managementState** is now set to **Managed**.

DSE-11870: Hung File, Stale File, and Fork issues with NFS

Hung File Operations: Certain file operations, such as `stat(2)` or `stat(1)` might stop responding, and if the file operation was performed through the CML web UI, the web operation might timeout. This indicates an NFS server that is not reachable for some reason. The error might manifest itself on the web UI when you try to open an ML project as an HTTP error, code 500. Check the logs for error messages similar to the following:

```
{"rpc":"1","service":"2","reqId":"3","err":"4"},"stat","VFS","18a07980-c55a-11ea-9bb9-a35829b422d9","message":"5","stack":"6","code":4,"metadata":"7","details":"8","futureStack ":"6"},
"DEADLINE_EXCEEDED: Deadline Exceeded","Error: 4 DEADLINE_EXCEEDED: Deadline Exceeded\n at Object.exports.createStatusError (/home/cdswint/services/web/node_modules/grpc/src/common.js:91:15)\n at Object.onReceiveStatus ((/home/cdswint/services/web/node_modules/grpc/src/client_interceptors.js:568:42)\n at InterceptingListener._callNext ((/home/cdswint/services/web/node_modules/grpc/src/client_interceptors.js:568:42)\n at InterceptingListener.onReceiveStatus ((/home/cdswint/services/web/node_modules/grpc/src/client_interceptors.js:618:8)\n at callback (/home/cdswint/services/web/n
```
Solution: Check your NFS server and make sure it is running. You will need to restart the NFS clients in your ML workspace’s namespace. These are the “ds-vfs” and “s2i-client” pods. Simply delete the Kubernetes pods whose names start with “ds-vfs” and “s2i-client”.

Stale File Handles: When opening a project from the ML web UI, an error message like “NFS: Stale file handle” shows up on the UI.

Solution: This is indicative of an NFS server and a client being out of sync, probably caused by a server restart along with file system content change on the server that the client is not aware of. You should restart NFS client pods in your ML workspace’s namespace. The are the “ds-vfs”, “s2i-client”, and any user sessions that are affected by the “Stale file handle” error.

Project Fork Creating Multiple Copies: When creating a new project from an existing project using the “Fork” feature, you might see the operation seemingly fail on the UI, but it still ends up creating multiple copies of the source project.

Solution: This issue happens when forking a project takes longer than the idle connection timeout set on the external load balancer, as well as in HA Proxy policy settings on OpenShift. Increase the idle connection timeout to at least 5 minutes. Depending on the performance of the NFS server, a higher timeout may be necessary.

DSE-11837: Timeout limitation for Project API

If you create a project in the UI using `git clone`, you may get the error message `Whoops, there was an unexpected error`. If you create a project using the API, a timeout may occur.

Prerequisites for CML in Private Cloud:

- Set any external load balancer server timeout to 5 min.

For a TLS Enabled Workspace:

- Set the annotation `haproxy.router.openshift.io/timeout=300` on each route in a deployed CML workspace namespace:

  ```bash
  oc annotate route --all=true --overwrite=true -n <cml-namespace> haproxy.router.openshift.io/timeout=300s
  ```

For non-TLS Enabled Workspaces, this setting is made automatically.

Workaround: Even though an error message displays, project creation still occurs. Check the Projects page after a few minutes; project creation should be complete.

DSE-9549: TLS enabled workspaces require manual configuration

To provision a TLS-enabled workspace, the customer needs to perform several manual steps. This procedure is described in Deploy an ML Workspace with Support for TLS.

OPSAPS-58019: CML workspace installation failure due to `includedir` in `krb5.conf` file

If the `/etc/krb5.conf` on the Cloudera Manager host contains `include` or `includedir` directives, Kerberos-related failures may occur.

As a workaround, comment out the `include` and `includedir` lines in `/etc/krb5.conf` on the Cloudera Manager host. If configuration in those files and directories are needed, add them directly to `/etc/krb5.conf`.

Related Information

Known Issues and Limitations