

Cloudera Data Engineering Overview

Date published: 2020-07-30

Date modified: 2023-06-13

The Cloudera logo is displayed in a bold, orange, sans-serif font. The word "CLOUDERA" is written in all caps. The letter 'E' is stylized with three horizontal bars, giving it a unique, geometric appearance.

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

| | |
|---|----------|
| Cloudera Data Engineering service..... | 4 |
| Cloudera Data Engineering resources..... | 4 |

Cloudera Data Engineering service

Cloudera Data Engineering (CDE) is a service for CDP Private Cloud Data Services that allows you to submit jobs to auto-scaling virtual clusters.

Cloudera Data Engineering allows you to create, manage, and schedule Apache Spark jobs without the overhead of creating and maintaining Spark clusters. With Cloudera Data Engineering, you define virtual clusters with a range of CPU and memory resources, and the virtual cluster scales up and down as needed to run your Spark workloads.

The CDE service involves several components:

Environment

A logical subset of your private cloud deployment, including a datalake and multiple compute resources. For more information, see [Environments](#).

CDE Service

A logical subset of the long-running Kubernetes cluster and services that manage the virtual clusters. The CDE service must be enabled on an environment before you can create any virtual clusters.

Virtual Cluster

An individual auto-scaling cluster with defined CPU and memory ranges. Virtual Clusters in CDE can be created and deleted on demand. Jobs are associated with clusters.

Job

Application code along with defined configurations and resources. Jobs can be run on demand or scheduled. An individual job execution is called a job run.

Resource

A defined collection of files such as a Python file or application JAR, dependencies, and any other reference files required for a job.

Job run

An individual job run.

Cloudera Data Engineering resources

A *resource* in Cloudera Data Engineering (CDE) is a named collection of files used by a job. Resources can include application code, configuration files, custom Docker images, and Python virtual environment specifications (requirements.txt).

Resources are associated with virtual clusters. A resource can be used by multiple jobs, and jobs can use multiple resources. The resource types supported by CDE are files and python-env.

files

An arbitrary collection of files that a job can reference. The application code for the job, including any necessary configuration files or supporting libraries, can be stored in a files resource. Files can be uploaded to and removed from a resource as needed.

python-env

A defined virtual Python environment that a job runs in. The only file that can be uploaded to a python-env resource is a [requirements.txt](#) file. When you associate a python-env resource with a job, the job runs within a Python virtual environment built according to the requirements.txt specification.

custom-runtime-image

A Docker container image. When you run a job using a custom-runtime-image resource, the executors that are launched use your custom image.