

Cloudera Data Engineering 1.4.1

## Cloudera Data Engineering Release Notes

Date published: 2020-07-30

Date modified: 2022-11-18

# CLOUDERA

<https://docs.cloudera.com/>

# 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

<b>What's new in Cloudera Data Engineering Private Cloud.....</b>	<b>4</b>
November 18, 2022.....	4
September 15, 2022.....	5
June 21, 2022.....	5
May 25, 2022.....	5
April 13, 2022.....	5
January 14, 2022.....	6
November 10, 2021.....	6
October 4, 2021.....	6
 <b>Known issues and limitations in Cloudera Data Engineering on CDP</b>	
<b>Private Cloud.....</b>	<b>7</b>
 <b>Fixed issues in Cloudera Data Engineering on CDP Private Cloud.....</b>	<b>9</b>
 <b>Spark and Airflow versions for Cloudera Data Engineering Private</b>	
<b>Cloud.....</b>	<b>10</b>
 <b>How to make base cluster configuration changes.....</b>	<b>10</b>

# What's new in Cloudera Data Engineering Private Cloud

This section lists major features and updates for the Cloudera Data Engineering (CDE) service.

## November 18, 2022

This release of Cloudera Data Engineering (CDE) on CDP Private Cloud 1.4.1 includes the following features:

### Custom NFS Storage Class

You can now specify the name of the custom NFS storage class while creating a CDE service. By default, CDE uses the platform's in-built storage class discovered using the provisioner.

For more information, see [. Adding a Cloudera Data Engineering service.](#)

### Limit Resources (Technical Preview)

You can now set the maximum number of CPU cores and the maximum memory in gigabytes that can be used by this CDE service and virtual cluster. The cluster can utilize resources upto the set capacity to run the submitted Spark applications.

For more information, see [Adding a Cloudera Data Engineering service](#) and [Creating virtual clusters](#). For information about configuring resource pool and capacity, see [Managing cluster resources using Quota Management \(Technical Preview\)](#).

### Workload Secrets

CDE now provides a secure way to create and store workload secrets for Cloudera Data Engineering (CDE) Spark Jobs. This is a more secure alternative to storing credentials in plain text embedded in your application or job configuration.

For more information, see [Managing workload secrets with CDE Spark Jobs using the API.](#)

### Spark History Server

You can now use Spark history server to troubleshoot Spark jobs. The Spark history server is a monitoring tool that displays information about completed Spark applications. It provides information for debugging such as Spark configurations, DAG execution, driver and executor resource utilization, application logs, and job, stage and task-level details.

For more information, see [Using Spark history server to troubleshoot Spark job.](#)

### Loading example jobs and sample data using new VCs

CDE now provides an option to add in-product examples of data and jobs in new virtual clusters to facilitate smoother onboarding and learning for new customers.

For more information, see [CDE example jobs and sample data.](#)

### Set default values for the variables in CDE job specification

Using [--default-variable] flags you can now replace strings in job values.

For more information, see [Creating and updating Apache Spark jobs using the CLI.](#)

### Job email alerts

SLA miss and job failure conditions can be configured for email notifications during job submission.

For more information, see [Creating jobs](#) and [Automating data pipelines using Apache Airflow](#).

## September 15, 2022

There are no new features for the Cloudera Data Engineering (CDE) service in the CDP Data Services 1.4.0-H1 release.

For known issues and limitations, see [Known issues and limitations in Cloudera Data Engineering](#).

## June 21, 2022

There are no new features for the Cloudera Data Engineering (CDE) service in the CDP Data Services 1.4.0.

For known issues and limitations, see [Known issues and limitations in Cloudera Data Engineering](#).

## May 25, 2022

This release 1.15.2 of Cloudera Data Engineering (CDE) on CDP Private Cloud 1.4.0 includes the following features.

For known issues and limitations, see [Known issues and limitations in Cloudera Data Engineering](#).

### Security Improvements

CVE-2021-44228 (Apache Log4j 2 vulnerability) has been addressed in CDE on CDP Private Cloud 1.4.0 by upgrading Apache Log4j 2 to version 2.17.

## April 13, 2022

This release 1.15.1 of Cloudera Data Engineering (CDE) on CDP Private Cloud 1.3.4 includes the following features.

For known issues and limitations, see [Known issues and limitations in Cloudera Data Engineering](#).

### CDE diagnostic bundles

You can now generate and download diagnostic data of Cloudera Data Engineering for troubleshooting purposes. You can specify the time duration, the associated environments and services, and limit the size of the file to include in the diagnostic data.

For more information, see [Working with CDP Private Cloud diagnostic data](#).

### Embedded Grafana dashboards for monitoring virtual clusters

You can now access Grafana dashboards for advanced visualization of Virtual Cluster's metrics such as memory and CPU usage in Cloudera Data Engineering (CDE) Private Cloud.

For more information, see [Connecting to Grafana dashboards](#).

### Access key based authentication

Cloudera Data Engineering (CDE) now supports transparent authentication using a password file, Cloudera Data Platform (CDP) access keys, and CDP credentials file.

For more information, see [Getting an API access token](#) and [CLI authentication](#).

### Custom Docker runtime images

Support for custom docker runtime images is now GA. Custom dependencies and packages can be included in a docker image built on top of the default CDE Spark runtime. Jobs using custom runtime images can be deployed using the API or the CLI.

For more information, see [Using Custom Spark Runtime Docker Images](#).

## January 14, 2022

This release (1.13.1) of Cloudera Data Engineering (CDE) on CDP Private Cloud 1.3.3 includes the following features and fixes.

For known issues and limitations, see [Known issues and limitations in Cloudera Data Engineerings](#).

### New features

#### Support for HDFS transparent encryption

Encryption at rest (HDFS transparent encryption) is now supported.

#### Custom Docker container images

Custom Docker container images are now supported.

#### Apache Spark 3

You can now create Spark 3 virtual clusters and run Spark 3 jobs.

#### Apache Airflow 2

The embedded Apache Airflow deployment has been upgraded to Airflow 2.

#### Default external volume size increased to 500 GB

The default external volume size is now 500 GB for each virtual cluster (100 GB each for the 5 components in a virtual cluster).

### Fixed issues

#### DEX-4860 : Add support for cross-VC Apache Airflow jobs

Airflow workflows that run jobs in a separate virtual cluster (VC) now work.

#### CVE-2021-44228 (Apache Log4j 2 vulnerability)

CVE-2021-44228 has been addressed in CDE on CDP Private Cloud 1.3.3 by upgrading Apache Log4j 2 to version 2.16.

## November 10, 2021

There are no new features for the Cloudera Data Engineering (CDE) service in the CDP Data Services 1.3.2

For known issues and limitations, see [Known issues and limitations in Cloudera Data Engineering](#).

## October 4, 2021

This is the initial release of CDE in CDP Private Cloud Experiences. For known issues and limitations, see [Known issues and limitations in Cloudera Data Engineering](#).

## Known issues and limitations in Cloudera Data Engineering on CDP Private Cloud

This page lists the current known issues and limitations that you might run into while using the Cloudera Data Engineering (CDE) service.

### **DEX-14676: Deep Analysis is not working in CDE PvC under analysis tab**

If you are using Spark version 2.x for running your jobs, then the Run Deep Analysis feature present under the Analysis tab is not supported on Cloudera Data Engineering Private Cloud.

### **DEX-6743: CDE CLI command execution sometimes displays End of File (EOF) error message in the end.**

CDE CLI command execution sometimes displays an EOF error message in the end even though the command executes successfully. This generally happens due to error message or delay in response due to network issues or timeout error.

No workaround. But you can check for the error message in the pod logs or is it due to slowness in resources over kubernetes clusters.

### **DOCS-17844: Logs are lost if the log lines are longer than 50000 characters in fluentd**

This issue occurs when the `Buffer_Chunk_Size` parameter for the `fluent-bit` is set to a value that is lesser than the size of the log line.

The values that are currently set are:

```
Buffer_Chunk_Size=50000
Buffer_Max_Size=50000
```

When required, you can set higher values for these parameters in the `fluent-bit` configuration map which is present in the `dex-app-xxxx` namespace.

### **DEX-8659 A non-functional Authoring UI field is displayed in the Airflow job creation page.**

If you are using the Default virtual cluster in CDP 1.4.1, you might see a new Authoring UI field on the airflow job creation page but it is not functional.

Do not use the default virtual cluster option in your CDE clusters, but use the non-default virtual clusters.

### **OPSAPS-65424: Embedded Container Service (ECS) 1.3.4 to 1.4.1 control plane upgrade looping forever in error state**

Upgrading the ECS version while CDE service is enabled, can cause Control Pane upgrade looping forever in error state.

Back up CDE jobs in the CDE virtual cluster, and then delete the CDE service and CDE virtual cluster. Restore it after the upgrade. For more information about backup and restore CDE jobs, see [Backing up and restoring CDE jobs](#).

### **DEX-8226: Grafana Charts of new virtual clusters will not be accessible on upgraded clusters if virtual clusters are created on existing CDE service.**

If you upgrade the cluster from 1.3.4 to 1.4.x and create a new virtual clusters on the existing CDE Service, Grafana Charts will not be displayed. This is due to broken APIs.

Create a new CDE Service and a new virtual cluster on that service. Grafana Charts of the virtual cluster will be displayed.

### **DEX-7000: Parallel Airflow tasks triggered at exactly same time by the user throws the 401:Unauthorized error.**

Error 401:Unauthorized is displayed when parallel Airflow tasks in an airflow job are triggered or launched exactly at the same time by the user.

1. Navigate to the Cloudera Data Engineering Overview page by clicking the Data Engineering tile in the Cloudera Data Platform (CDP) management console.
2. In the Environments column, select the environment containing the virtual cluster where you want to create the job.
3. In the Virtual Clusters column on the right, click the View Jobs icon on the virtual cluster where you want to create the application.
4. In the left hand menu, click Jobs.
5. Click the Create Job button.
6. Provide the job details:
  - a. Select Airflow for the job type.
  - b. Specify the job name as bashoperator-job.
  - c. Save the following python script to attach it as a DAG file.

```
from dateutil import parser
from airflow import DAG
from airflow.utils import timezone
from airflow.operators.bash_operator import BashOperator
default_args = {
    'depends_on_past': False,
}
dag = DAG(
    'bashoperator-job',
    default_args = default_args,
    start_date = parser.isoparse('2022-06-17T23:52:00.123Z'
).replace(tzinfo=timezone.utc),
    schedule_interval = None,
    is_paused_upon_creation = False
)
task1 = BashOperator(
    task_id = 'task1',
    dag = dag,
    bash_command = 'sleep 600'
)
task2 = BashOperator(
    task_id = 'task2',
    dag = dag,
    bash_command = 'sleep 600'
)
task3 = BashOperator(
    task_id = 'task3',
    dag = dag,
    bash_command = 'sleep 600'
)
[task1, task2] >> task3
```

- d. Select File, click Select a file to upload the above python, and select a file from an existing resource.
7. Select the Python Version, and optionally select a Python Environment.
8. Click Create and Run.

**DEX-7001: When Airflow jobs are run, the privileges of the user who created the job is applied and not the user who submitted the job.**

Irrespective of who submits the Airflow job, the Airflow job is run with the user privileges who created the job. This causes issues when the job submitter has lesser privileges than the job owner who has higher privileges.

Spark and Airflow jobs must be created and run by the same user.

**DEX-7022: Virtual Cluster does not accept spark or airflow jobs if the tzinfo library is used as the start date.**



If you use the tzinfo library for start\_date, then the Virtual Cluster may not complete execution of spark or airflow jobs launched later. For example:

```
example_dag = DAG(
    'bashoperator-parameter-job',
    default_args=default_args,
    start_date=parser.isoparse("2020-11-11T20:20:04.268Z").replace(tzinfo=timezone.utc),
    schedule_interval='@once',
    is_paused_upon_creation=False
)
```

Use start\_date as start\_date=pendulum.datetime(2017, 1, 1, tz="UTC") instead of code like the tzinfo library. For more information about time zones, see [Airflow time zone aware DAGs documentation](#).

### Changing LDAP configuration after installing CDE breaks authentication

If you change the LDAP configuration after installing CDE, as described in [Configuring LDAP authentication for CDP Private Cloud](#), authentication no longer works.

Re-install CDE after making any necessary changes to the LDAP configuration.

### Gang scheduling is not supported

Gang scheduling is not currently supported for CDE on CDP Private Cloud.

### HDFS is the default filesystem for all resource mounts

For any jobs that use local filesystem paths as arguments to a Spark job, explicitly specify file:// as the scheme. For example, if your job uses a mounted resource called test-resource.txt, in the job definition, you would typically refer to it as /app/mount/test-resource.txt. In CDP Private Cloud, this should be specified as file:///app/mount/test-resource.txt.

### Apache Ozone is supported only for log files

Apache Ozone is supported only for log files. It is not supported for job configurations, resources, and so on.

### Scheduling jobs with URL references does not work

Scheduling a job that specifies a URL reference does not work.

Use a file reference or create a resource and specify it

### Limitations

#### Access key-based authentication will not be enabled in upgraded clusters prior to CDP PVC 1.3.4 release.

After you upgrade to PVC 1.3.4 version from earlier versions, you must create the CDE Base service and Virtual Cluster again to use the new Access Key feature. Otherwise, the Access Key feature will not be supported in the CDE Base service created prior to the 1.3.4 upgrade.

## Fixed issues in Cloudera Data Engineering on CDP Private Cloud

No existing known issues have been fixed in this release. For information about new features and improvements, see the What's new topic.

# Spark and Airflow versions for Cloudera Data Engineering Private Cloud

Cloudera Data Engineering (CDE) uses Spark and Airflow as its components. The following table lists the versions of Airflow and Spark in this release:

**Table 1: Spark and Airflow versions**

Component	Version
Spark	2.4.7 and 3.2.1
Airflow	2.2.5

## How to make base cluster configuration changes

In general, as Administrator you perform the following steps:

1. Make the necessary configuration changes in the base cluster.
2. Restart the base cluster.
3. In the Private Cloud compute cluster, run the specific kubernetes commands to restart the pods for CDE after identifying the correct CDE Service namespace.

### Identifying the CDE Namespace

1. Navigate to the Cloudera Data Engineering Overview page by clicking the Data Engineering tile in the Cloudera Data Platform (CDP) management console.
2. In the CDE Services column, click the Service Details for the CDE service.
3. Note the Cluster ID shown in the page. For example, if the Cluster ID is cluster-abcd1234, then the CDE Namespace is dex-base-abcd1234.
4. Use this CDE Namespace (in the above example, it is dex-base-abcd1234) in the following instructions to run kubernetes commands.

### Embedded Container Service

1. Access Cloudera Manager.
2. Navigate to the Experiences Cluster ECS Web UI: Clusters Your embedded Cluster ECS Web UI ECS Web UI .
3. Select the CDE namespace obtained previously on the top left dropdown.
4. Navigate to Workloads Deployments .
5. Locate dex-base-configs-manager in the list and click Restart from the breadcrumbs on the right.

### OpenShift Container Platform

Access the openshift cluster with oc or kubectl, and scale the deployment of dex-base-configs-manager down and back up. Use the following commands:

```
oc scale deployment/dex-base-configs-manager --namespace <CDE Namespace> --replicas 0
```

```
oc scale deployment/dex-base-configs-manager --namespace <CDE Namespace> --replicas 1
```