Cloudera Data Engineering 1.5.5

Cloudera Data Engineering Release Notes

Date published: 2020-07-30
Date modified: 2025-06-06

https://docs.cloudera.com/

Legal Notice

Contents

What's new in Cloudera Data Engineering on premises

Review the new features in the Cloudera Data Engineering service of the Cloudera Data Services on premises 1.5.5 release.

Integrating third-party certificate manager

Cert-manager is an open-source tool for Kubernetes that automates the provisioning, management, and renewal of TLS certificates. Its documentation at https://cert-manager.io/docs/ provides comprehensive guidance on installing, configuring, and using cert-manager to secure workloads with trusted X.509 certificates. Cloudera provides out-of-the-box support for Venafi TPP as part of the Cloudera Embedded Container Service installation. By integrating cert-manager, the Cloudera Data Services on premises achieve secure communication, reduced manual overhead, and compliance with security standards, leveraging its robust automation and flexibility. For more information on setting up and integrating Cert-manager using Venafi TPP, see Integrating third-party certificate manager.

User Access Management

Users must be assigned roles on Cloudera Data Engineering Services and Virtual Clusters to provide them with specific access to the Service or the Virtual Cluster. User Access Management allows you to assign the roles to manage and access the Cloudera Data Engineering Service and Virtual Clusters by defining the access levels for a particular user or user groups. This allows you to assign role-based access to individual users or user groups. For more information, see User Access Management.

Privacy Settings for Virtual Clusters

The new Privacy Settings option for Virtual Clusters allows administrators to define how the artifacts are shared between different users or user groups with VC User role. This allows you to keep all the artifacts private to the user or the user group. For more information, see Privacy Settings.

Artifact access management

By default, you can access and manage the artifacts that are either owned by you or if they are explicitly shared with you by another user. Only users or user groups with full access can share the artifacts with others. Users or user groups with read-only access cannot share artifacts. Users with full access can share an artifact that they own with another user or user group with either full access or read only access. For more information, see Artifact access management.

Hadoop authentication

The Hadoop Authentication tab allows you to authenticate the Kerberos Keytab file for a Cloudera Data Engineering Service or a Virtual Cluster. If the Kerberos Keytab file is not authenticated properly, then you cannot run the jobs or sessions. For more information, see Hadoop authentication.

Updating TLS certificate

Starting from Cloudera Data Engineering 1.5.5 release, you can update the TLS certificate for Cloudera Data Engineering Service or a Virtual Cluster in the Cloudera Data Engineering UI or API directly instead of running the cde-utils.sh script. For more information about how to upload a TLS certificate, see Updating the Control Plane certificates in Cloudera Data Engineering Services or Updating the Control Plane certificates in Cloudera Data Engineering Virtual Clusters.

Support for Ozone OFS data connector in Cloudera Data Engineering Sessions

You can now use data connectors to connect to Ozone in Cloudera Data Engineering Sessions which allows you to access data from Ozone directly. To configure Ozone OFS data connector in Cloudera Data Engineering Sessions, see Creating Sessions in Cloudera Data Engineering or Creating a Session using the CDE CLI.

Quota Management for multiple base cluster support

Quota management enables you to control how resources are allocated within your Cloudera Data Services on premises clusters. In order to prevent a single workload from consuming all available cluster resources, you can limit the number of CPUs, GPUs, and memory allocated by application, user, business units, or Data Service by defining resource pools that define resource limits. Pools are organized in a hierarchical manner by defining nodes in the hierarchy with resource limits, which can then be subdivided as needed to allocate resources for an organization and to allocate resources to cluster or environment wide services such as the monitoring service.

Fixed issues in Cloudera Data Engineering on Cloudera on premises

Review the list of issues that are resolved in the Cloudera Data Engineering service in the Cloudera Data Services 1.5.5 release.

DEX-16897: Mitigate envoy proxy out of memory issue

In a cluster with many services, the envoy proxy side may encounter out of memory issue due to the large number of services it has to keep track of. By using Sidecar CRD, the envoy proxy sidecar explicitly tracks only the namespaces or services that it is interested in, therefore reduces the number of services it has to track, and chances of having the out of memory issue. This issue is fixed.

Known issues and limitations in Cloudera Data Engineering on Cloudera on premises

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

Known Issues in Cloudera Data Services on premises 1.5.5

DEX-17361: Restore Cloudera Data Engineering job fails if the job is created with the schedule having start time or end time in format other than the pre-defined time format

Scheduled jobs created using the Jobs API do not restore after backup if their start and end times are not in RFC3339Milli or RFC3339Nano formats respectively.

For the job's schedule, make sure that the start time uses the RFC3339Milli format (for example, 2006-01-02T15:04:05.000Z) and the end time uses RFC3339Nano format (for example, 2006-01-02T15:04:05.999999999Z07:00).

DEX-17197: Cloudera Data Engineering job and DAG runs - Different components or log types for the same job are recorded in different time zones

When job run submitter event logs is stored in the DB, it is converted to UTC time. This causes inconsistency where every other log and pod log are in a custom timezone, but the submitter logs are in UTC.

DEX-11277: Spark 3 job shows "HS2 delegation token" related error message as warning

Spark 3 jobs show the following warning message in the driver logs even if the job is successful and has no issues:

```
WARN HiveServer2CredentialProvider: Failed to get HS2 delegation
token java.util.NoSuchElementException: spark.sql.hive.hiveser
ver2.jdbc.url
```

This specific warning message must be ignored as it doesn't impact the working of a job.

DEX-17286: Wrong error messages are displayed for non-existing jobs and sessions

The following error messages are displayed while running non-existing jobs or sessions in the Cloudera Data Engineering installed using Cloudera Embedded Container Service:

- For admin users:
  - Jobs: could not get run from storage: job run not found
  - Sessions: could not get session from storage: session not found
- For non-admin users:
  - Jobs: job not found
  - Sessions: session not found

DEX-17195: Cloudera Data Engineering job is going to "succeeded" status directly from "Started" status

When more than 100 jobs are running in Cloudera Data Engineering, the job status might sometimes change directly from "Started" to "Succeeded".

DEX-17187: Cloudera Data Engineering Spark jobs fail at Cloudera Runtime when memory is set using gb instead of g through Jobs API

Spark jobs are allowed to be created with unsupported memory units through API.

User must use supported spark memory units (k, m, g, or t) for driver and executor memory in payload. For more information about supported memory units, see Spark Application Properties.

DEX-17037: Jobs with python script referencing resources on airflow editor page failed to save

In Cloudera Data Engineering installed using OpenShift Container Platform, user would not be able to use resources with DAG Editor UI. For Embedded Container Service, this works fine.

DEX-16604: Disable settting acl fields in cli and api for OpenShift Container Platform

The artifact level acls is not supported in OpenShift Container Platform for Cloudera Data Engineering 1.5.5. The API and CLI does not show error, even if you try to do that. Refrain from doing this, since it stores the metadata though, which could create a problem when upgraded to future version which starts supporting the artifact acls, since users might suddenly start seeing the artifacts acls in action for that particular artifacts.

DEX-17028: Cloudera Data Engineering UI does not automatically remove VC roles if roles are revoked from Cloudera Data Engineering Service

Role assignment and unassignment can be done at the Service or Virtual Cluster level, and such action only impacts the assignments at respective the Service or Virtual Cluster only. Adding or removing a role assignment from a Service does not implicitly add or remove role assignments from the underlying Virtual Clusters.

DEX-17241: Unable to submit pipeline job

When you click on the Run button on Jobs Editor page while creating a DAG, following error comes up:

```
Pipeline submit failed with error: Exception while fetching data
 (/runPipeline) : Unauthorized
```

You can save the job and run from the Jobs page in the UI by clicking the Run Now button.

DEX-17291: Job or pyenv restore is failing due to timeout occurring during a restore operation

During a job restore or upgrade, operations may time out and fail if there are over 1,000 jobs and 1,000 resources, especially if the process exceeds 10 minutes. Any proxy ahead of the ECS or OCP cluster with a shorter timeout will cause the operation to fail even sooner.

If this issue occurs, increase the ingress layer timeout settings for Embedded Container Service and OpenShift Container Platform, as well as the timeout for any proxy present. This change must be performed with the Support and Engineering teams' supervision.

DEX-17345: Without clicking the + sign, adding email id to notifiers list doest work

Typing the email ids and then clicking create or update job does not save the email in alerts.

Make sure that you click the + icon after entering each email id to add it to the input dialog box and then click on create or update job button.

DEX-17344: The form submit on jobs page must validite the fields in the alter section

The job create or update form submit does not do any validation right now in UI. For example, if you have enabled alerts but has not added any email, it will not show any error, only once you click submit it shows error.

Known issues from previous releases carried in Cloudera Data Services on premises 1.5.5

DEX-5444: Cloudera Data Engineering on premises is not able to distinguish between stdout and stderr when forwarding logs

Entire Spark job driver and executor logs stderr and stdout are all redirected to the stderr log file.

Refer the driver/executor stderr log file which contains both stderr and stdout content.

DEX-8542: Newly created Iceberg tables are owned by "sparkuser"

The Iceberg tables created in Cloudera Data Engineering using Spark 3.2.3 are being displayed as owned by the "sparkuser" user. The Iceberg tables must be owned by the user who created them. For example,

```
hive=> SELECT "TBL_NAME", "OWNER" FROM "TBLS" WHERE "TBL_NAME"='
iceberg_test';
   TBL_NAME   |   OWNER
------------{}+{}---------
 iceberg_test | sparkuser
```

Spark 3.2.3 uses Iceberg version 0.14, which is causing this issue. Create and use a Cloudera Data Engineering Virtual Cluster with Spark version 3.3.2 which is not affected by this.

DEX-14676: Deep Analysis is not working in Cloudera Data Engineering on premises 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 on premises.

DEX-12150: Recursive search for a file in resources is not working

If you search for any file using the Search field in the Resources page, the result does not display any files present with that name inside the resources.

Navigate to the relevant resource and then locate the file in that resource.

DEX-8540: Job Analysis tab is not working

When you access the Jobs Runs Analysis tab through the Cloudera Data Engineering UI, the Analysis tab fails to load data for Spark 2.

To view the data in the Job Analysis tab, open the JOBS API URL from the Virtual Cluster details page and access the Analysis tab.

DEX-12426: Data Connector UI does not load

Data Connector UI does not load for self-signed certificates if the browser certificates are not trusted.

When utilising the self-signed certificates, it is crucial to trust the certificates associated with the Cloudera Data Engineering URLs. Perform the following:

1. Using the Cloudera Data Engineering UI, install a virtual cluster (VC) on the cluster.
2. Initialize the installed VC using the `cde-utils.sh` script.
3. Open the **Jobs UI** in a new tab, which prompts the acknowledgement and trust of the certificates, accept it.

4. Refresh the **Data Connectors** page.

You can now access Grafana and the Data Connectors UI.

DEX-11300: Editing the configuration of a job created using a Git repository shows Resources instead of Repository

Jobs which use application file from *REPOSITORIES* when edited, shows Resources as a source under Select application file. This issue does not affect the functionality of the job but could confuse as it displays the source as a Resource for the application even if the selected file is from a repository. Though it would show Resource in this case, in the backend it is selected from the chosen repository.

DEX-11340: Sessions go to unknown state if you start the Cloudera Data Engineering upgrade process before killing live Sessions

If spark sessions are running during the Cloudera Data Engineering upgrade then they are not be automatically killed which can leave them in an unknown state during and after the upgrade.

You must kill the running Spark Sessions before you start the Cloudera Data Engineering upgrade.

DEX-10939: Running the `prepare-for-upgrade` command puts the workload side database into read-only mode

Running the `prepare-for-upgrade` command puts the workload side database into read-only mode. If you try to edit any resources or jobs or run jobs in any virtual cluster under the Cloudera Data Engineering service for which the `prepare-for-upgrade` command was executed, The MySQL server is running with the --read-only option so it cannot execute this statement error is displayed.

This means that all the APIs that perform `write` operations will fail for all virtual clusters. This is done to ensure that no changes are done to the data in the cluster after the `prepare-for-upgrade` command is executed, so that the new restored cluster is consistent with the old version.

You must ensure that you have sufficient time to complete the entire upgrade process before running the `prepare-for-upgrade` command.

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.

DOCS-18585: Changes to the log retention configuration in the existing virtual cluster do not reflect the new configuration

When you edit the log retention policy configuration for an existing virtual cluster, the configuration changes are not applied.

When you edit the log retention policy configuration, you must restart the runtime-api-server pod using the `kubectl rollout restart deployment/<deployment-name> -n <namespace>` command to apply the changes.

For example:

```
kubectl rollout restart deployment/dex-app-fww6lrgm-api -n dex-a
pp-fww6lrgm
```

DEX-11231: In OpenShift, the Spark 3.3 virtual cluster creation fails due to Airflow pods crashing

This is an intermittent issue during virtual cluster installation in the OCP cluster where the airflow-scheduler and airflow-webserver pods are stuck in the *CRASHLOOPBACKOFF* state. This leads to virtual cluster installation failure.

Retry the virtual cluster installation because the issue is intermittent.

DEX-10576: Builder job does not start automatically when the resource is restored from an archive

For the airflow python environment resource, the restoration does not work as intended. Though the resource is restored, the build process is not triggered. Even if the resource was activated during backup, it is not reactivated automatically. This leads to job failure during restoration or creation, if there is a dependency on this resource.

You can use the Cloudera Data Engineering API or CLI to download the requirements.txt file and upload it to the resource. You can activate the environment if required.

```
# cde resource download --name <python-environment-name> --resou
rce-path requirements.txt
# cde resource upload --name <python-environment-name> --local-
path requirements.txt
```

DEX-10147: Grafana issue if the same VC name is used under different Cloudera Data Engineering services which share same environment

In Cloudera Data Engineering 1.5.1, when you have two different Cloudera Data Engineering services with the same name under the same environment, and you click the Grafana charts for the second Cloudera Data Engineering service, metrics for the Virtual Cluster in the first Cloudera Data Engineering service will display.

After you have upgraded Cloudera Data Engineering, you must verify other things in the upgraded Cloudera Data Engineering cluster except the data shown in Grafana. After you verified that everything in the new upgraded Cloudera Data Engineering service, the old Cloudera Data Engineering service must be deleted and the Grafana issue will be fixed.

DEX-10116: Virtual Cluster installation fails when Ozone S3 Gateway proxy is enabled

Virtual Cluster installation fails when Ozone S3 gateway proxy is enabled. Ozone s3 gateway proxy gets enabled when more than one Ozone S3 Gateway is configured in the Cloudera Base on premises cluster.

Add the `127.0.0.1 s3proxy-<environment-name>.<private-cloud-control-plane-name>-services.svc.cluster.local` entry in the /etc/hosts of all nodes in the Cloudera Base on premises cluster where the Ozone S3 gateway is installed. For example, if the on premises environment name is cdp-env-1 and on premises control plane name is cdp, then add the `127.0.0.1 s3proxy-cdp-env-1.cdp-services.svc.cluster.local` entry in `/etc/hosts`.

DEX-10052: Logs are not available for python environment resource builder in Cloudera on premises

When creating a python environment resource and uploading the requirements.txt file, the python environment is built using a k8s job that runs in the cluster. These logs cannot be viewed currently for debugging purposes using CDE CLI or UI. However, you can view the events of the job.

None

DEX-10051: Spark sessions is hung at the Preparing state if started without running the cde-utils.sh script

You might run into an issue when creating a spark session without initialising the Cloudera Data Engineering virtual cluster and the UI might hang in a Preparing state.

Run the cde-utils.sh to initialise the virtual cluster as well as the user in the virtual cluster before creating a Spark long-running session.

DEX-9783: While creating the new VC, it shows wrong CPU and Memory values

When clicking on the Virtual Cluster details for a Virtual Cluster that is in the Installing state, the configured CPU and Memory values that are displayed are inaccurate for until the VC is created.

Refresh the Virtual Cluster details page to get the correct values, five minutes after the Virtual Cluster installation has started.

DEX-9916: Cloudera Data Engineering Service installation is failing when retrieving `aws_key_id`

Cloudera Data Engineering Service installation is failing when retrieving `aws_key_id` with the Could not add shared cluster overrides, error: unable to retrieve aws_key_id from the env se rvice error.

1. Restart the Ozone service on the Cloudera Base cluster and make sure all the components are healthy.
2. Create a new environment in Cloudera on premises using the Management Console.
3. Use the same environement for creating the Cloudera Data Engineering Service.

DEX-8996: Cloudera Data Engineering service stuck at the initialising state when a user who does not have correct permission tries to create it

When a Cloudera Data Engineering user tries to create a Cloudera Data Engineering service, it gets stuck at the initializing state and does not fail. Additionally, cleanup cannot be done from the UI and must be done on the backend.

Only the user who has the correct permission should create a Cloudera Data Engineering service. If you experience any issue, delete the stuck Cloudera Data Engineering service from the database.

DEX-8226: Grafana Charts of new virtual clusters will not be accessible on upgraded clusters if virtual clusters are created on existing Cloudera Data Engineering service

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

Create a new Cloudera Data Engineering 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 causes airflow jobs to fail intermittently, when parallel Airflow tasks using CDEJobRunOperator are triggered at the exact same time in an Airflow DAG.

Using the below steps, create a workaround bashoperator job which will prevent this error from occurring. This job will keep running indefinitely as part of the workaround and should not be killed.

1. Navigate to the Cloudera Data Engineering Overview page by clicking the Data Engineering tile in the Cloudera console.
2. In the Cloudera Data Engineering Services column, select the service 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 job.
4. In the left hand menu, click Jobs.
5. Click Create Job.
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,
}
with 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
   ) as dag:
    [ BashOperator(task_id = 'task1', bash_command = 'sleep
infinity'),
    BashOperator(task_id = 'task2', bash_command = 'sleep in
finity') ]
```

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.

Changing LDAP configuration after installing Cloudera Data Engineering breaks authentication

If you change the LDAP configuration after installing Cloudera Data Engineering, as described in Configuring LDAP authentication for Cloudera on premises, authentication no longer works.

Re-install Cloudera Data Engineering after making any necessary changes to the LDAP configuration.

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 Cloudera on premises, this should be specified as file:///app/mount/test-resource.txt.

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

DEX-13775: The synchronization operation fails when using a non-default branch from the Git repository with Cloudera Data Engineering Git repositories

When you use a non-default branch from a Git repository with the Cloudera Data Engineering Git repositories, the synchronization operation fails.

Clone the Git repository from the non-default branch again after the latest commit.

Creating Cloudera Data Engineering Virtual Cluster without installing Atlas in your Cloudera Base cluster

If the Cloudera Data Engineering Virtual Cluster creation fails because Atlas is not installed, you must identify the Cloudera Data Engineering Namespace and set an environment variable prior to creating the Virtual Cluster.

Procedure

1. Identify the Cloudera Data Engineering Namespace
   a. In the Cloudera console, click the Data Engineering tile. The Cloudera Data Engineering Home page displays.
   b. In the Cloudera Data Engineering Services column, click  for the Cloudera Data Engineering service you want to create a VC.
   c. Note the Cluster ID shown on the page and identify the Cloudera Data Engineering Namespace. For example, if the Cluster ID is cluster-sales8098, then the Cloudera Data Engineering Namespace is *DEX-BASE-SALES8098*.
2. Use this Cloudera Data Engineering Namespace (*DEX-BASE-SALES8098*) to run Kubernetes commands using kubectl or OpenShift's command line oc.

   `kubectl`

   ```
   kubectl set env deployment/dex-base-configs-manager -c dex-base-configs-
   manager ATLAS_CONFIGS_DISABLED=true --namespace <CDE Namespace>
   ```

   `oc`

   ```
   oc set env deployment/dex-base-configs-manager -c dex-base-configs-manager
    ATLAS_CONFIGS_DISABLED=true --namespace <CDE Namespace>
   ```

Compatibility for Cloudera Data Engineering and Cloudera Runtime components

Learn about Cloudera Data Engineering and compatibility for Cloudera Runtime components across different versions.

Table 1: Cloudera Data Engineering compatibility with Cloudera Runtime component details for Security Hardened

Cloudera Runtime Version	Spark 3.2.x	Spark 3.3.x	Airflow	Iceberg	Kubernetes
7.1.7 SP 3	• Spark 3.2.3 • Scala 2.12.15 • Python 3.9.22 • Java 11.0.27	NA	• Airflow 2.10.4 • Python 3.11.12 • Java 17.0.15	Iceberg 0.14.1	1.30
7.1.9 SP1	• Spark 3.2.3 • Scala 2.12.15 • Python 3.9.22 • Java 11.0.27	• Spark 3.3.2 • Scala 2.12.15 • Python 3.10.17 • Java 17.0.15	• Airflow 2.10.4 • Python 3.11.12 • Java 17.0.15	Iceberg 1.3.0	1.30

Table 2: Cloudera Data Engineering compatibility with Cloudera Runtime component details for Redhat insecure

Cloudera Runtime Version	Spark 2.4.x	Spark 3.2.x	Spark 3.3.x	Airflow	Iceberg	Kubernetes
7.1.7 SP3	• Spark 2.4.7 • Scala 2.11.12 • Python 2.7.18 • Python 3.6.8 • Java 1.8.0_452	• Spark 3.2.3 • Scala 2.12.15 • Python 2.7.18 • Python 3.6.8 • Java 11.0.25	NA	• Airflow 2.10.4 • Python 3.11.12 • Java 17.0.15	Iceberg 0.14.1	1.30
7.1.9 SP1	• Spark 2.4.7 • Spark 2.4.8 • Scala 2.11.12 • Python 2.7.18 • Python 3.6.8 • Java 1.8.0_452	• Spark 3.2.3 • Scala 2.12.15 • Python 2.7.18 • Python 3.6.8 • Java 11.0.25	• Spark 3.3.2 • Scala 2.12.15 • Python 2.7.18 • Python 3.8.17 • Java 11.0.25	• Airflow 2.10.4 • Python 3.11.12 • Java 17.0.15	Iceberg 1.3.0	1.30

Important:

- Starting from Cloudera Data Services on premises 1.5.5, Spark 3.4 is no longer available.
- Hive Warehouse Connector (HWC) is supported on all Spark 2.x versions. However, on Spark 3.x versions, HWC is supported only on Spark 3.3.2.