Cloudera Data Engineering 1.5.0

Cloudera Data Engineering Release Notes

Date published: 2020-07-30
Date modified: 2023-01-24

CLOUDERA

https://docs.cloudera.com/

Legal Notice

Contents

What's new in Cloudera Data Engineering Private Cloud

This release of Cloudera Data Engineering (CDE) on CDP Private Cloud 1.5.0 includes the following features:

Using spark-submit drop-in migration tool

Cloudera Data Engineering (CDE) now provides a command line tool cde-env to help migrate your Cloudera Data Platform (CDP) Spark workloads running on CDP Private Cloud Base (spark-on-YARN) to CDE without having to completely rewrite your existing spark-submit command-lines.

For more information, see Using spark-submit drop-in migration tool.

Support for Apache Iceberg (Technical Preview)

Apache Iceberg tables are now supported with Spark 3 virtual clusters.

- Use tables at petabyte scale without impacting query planning, while benefiting from efficient metadata management, snapshotting, and time-travel.
- CDE supports row level updates via copy-on-write MERGE / UPDATES/ DELETES operations. Copy-on-write is helpful in bulk updates in read heavy use-cases.

Note: Support for Apache Iceberg V1 is in technical preview in the Private Cloud 1.5.0 release and is not recommended for production deployments. Cloudera recommends that you try this feature in test or development environments.

For more information, see Using Apache Iceberg.

Updated CDE user interface

The user interface has been updated with easy access to commonly used pages, a new Home page, and a Virtual Cluster drop-down menu that allows you to view relevant content related to each Virtual Cluster that you select. The following user interface changes were made:

Left-hand menu displays the following:

- **Home**: New landing page that displays Virtual Clusters and convenient quick-access links.
- **Jobs**: Displays jobs for the Virtual Cluster that you select from the drop-down menu in the upper left-hand corner.
- **Job Runs**: Displays the run history of all jobs within a selected Virtual Cluster.
- **Resources**: Displays resources created within a selected Virtual Cluster.
- **Administration**: Displays services and Virtual Clusters that can be customized (previously known as the Overview page.

Note: If you are using a browser in incognito mode, you have to allow all cookies in your browser settings so that you can view the following CDE pages: Pipelines, Spark, and Airflow.

Support for workload secrets using CLI

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 CDE CLI.

Using multiple profiles

You can now create and use multiple profiles using CDE CLI. You can add a collection of CDE CLI configurations grouped together as profiles, to the config.yaml file. You can use these profiles while running commands. You can set the configurations either at a profile level or at a global level.

For more information, see Creating and using multiple profiles using CDE CLI.

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-8682: CDE PvC 1.5.0 : CDP upgrade to 1.5.0 with OCP upgrade (4.8 to 4.10) Jobs UI is not opening

Upgrading the OCP version from 4.8 to 4.10 while CDE service is enabled, causes the Jobs UI to not open. This is due to OCP 4.10 upgrading to the Kubernetes version 1.23 which removes the old ingress APIs used.

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.

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-8614: Sometimes Spark job is not getting killed even though its parent Airflow job gets killed

Sometimes if an issue is encountered while sending the request to kill a spark batch to the Livy API and the error is logged but not propagated properly to the Airflow job. In such cases, the underlying spark job might still be running, though the airflow job considers that the job is killed successfully.

Kill the spark job manually using CDE user interface, CLI, or API.

DEX-9237: Job fails with the “Permission Denied” error after updating the virtual cluster resource quota

Whenever the virtual cluster resource quota is updated, the newly launched jobs on the virtual cluster fail with the Permission Denied error. This error can be seen in various stages of the job life cycle, in submitters, drivers or executors, and Airflow workers.

Restart all the virtual cluster pods manually every time you update the virtual cluster resource quota.

DEX-8601: ECS 1.4.x to 1.5.0 Upgrade: jobs fail after upgrade

Upgrading the ECS version while CDE service is enabled, causes the jobs launched in the old CDE virtual cluster fail. This is due to ECS upgrading to the kubernetes version 1.23 which removes the old ingress APIs used.

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-8600: ECS 1.4.x to 1.5.0 Upgrade: Virtual cluster creation and deletion fails

Upgrading the ECS version while CDE service is enabled, causes the old CDE service and virtual cluster creation and deletion to fail. This is due to ECS upgrading to the kubernetes version 1.23 which removes the old ingress APIs used.

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 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 Data Platform (CDP) console.
2. In the CDE 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 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.

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

Review the list of issues that are resolved in the Cloudera Data Engineering (CDE) service in the CDP Data Services 1.4.0-H1 release..

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

Previously, CDE CLI command execution sometimes displayed an EOF error message in the end even though the command executes successfully. This generally happened due to error message or delay in response due to network issues or timeout error. With this fix, this issue is resolved.

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

Component	Version
Airflow	2.3.4