

Accessing the Cloudera Data Engineering service using the API

Date published: 2020-07-30

Date modified: 2022-11-18

CLOUDERA

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

Using the Cloudera Data Engineering API.....	4
Getting a Cloudera Data Engineering API access token.....	4
Using an access token in Cloudera Data Engineering API calls.....	6
Managing Cloudera Data Engineering job resources using the API.....	6
Creating a Cloudera Data Engineering resource using the API.....	7
Deleting a Cloudera Data Engineering resource using the API.....	8
Managing workload secrets with Cloudera Data Engineering Spark Jobs using the API.....	8
Creating a workload secret for Cloudera Data Engineering Spark Jobs using API.....	8
Listing an existing workload secret for Cloudera Data Engineering Spark Jobs using API.....	9
Deleting a workload secret for Cloudera Data Engineering Spark Jobs using API.....	10
Linking a workload secret to the Cloudera Data Engineering Spark Job definitions using API.....	10
Using the workload secret in the Spark application code.....	11
Creating a Cloudera Data Engineering job using the API.....	12
Listing Cloudera Data Engineering jobs using the API.....	13
Getting Cloudera Data Engineering job information using the API.....	13

Using the Cloudera Data Engineering API

Cloudera Data Engineering (CDE) provides a robust API for integration with your existing continuous integration/continuous delivery platforms.

The Cloudera Data Engineering service API is documented in Swagger. You can view the API documentation and try out individual API calls by accessing the API DOC link in any virtual cluster:

1. In the Data Engineering web console, select an environment.
2. Click the Cluster Details icon in any of the listed virtual clusters.
3. Click the link under API DOC.

Related Information

[Using CLI-API to Automate Access to Cloudera Data Engineering](#)

[Using Cloudera Data Engineering CLI](#)

Getting a Cloudera Data Engineering API access token

Cloudera Data Engineering uses JSON Web Tokens (JWT) for API authentication. To interact with a virtual cluster using the API, you must obtain an access token for that cluster.

Before you begin

Determine the authentication endpoint for your virtual cluster:

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 Services column, select the environment containing the virtual cluster you want to interact with.
3. In the Virtual Clusters column on the right, click the Cluster Details icon on the virtual cluster you want to interact with.
4. Click the link under GRAFANA CHARTS. The hostname of the URL in your browser is the base URL, and /gateway/<***authkn-OR-cdptkn***>/knoxtoken/api/v1/token is the endpoint.
 - Example: LDAP Authentication URL

```
https://service.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/gateway/authkn/knoxtoken/api/v1/token
```

- Example: Access Key Authentication URL

```
https://service.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/gateway/cdptkn/knoxtoken/api/v1/token
```

For Using LDAP authentication

1. From the client you want to use to access the API, run `curl -u <workload_user> <auth_endpoint>`. Enter your workload password when prompted.

For example:

```
curl -u csso_psherman https://service.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/gateway/authkn/knoxtoken/api/v1/token
```

The user account is your CDP workload user .

2. In the output, the `access_token` value is the JWT. For convenience, copy it and set it as an environment variable:

```
export CDE_TOKEN=<access_token>
```

Alternatively, you can set the token in a single step using `jq` to extract the token:

```
export CDE_TOKEN=$(curl -s -u <workload_user> <auth_endpoint> | jq -r '.access_token')
```

For Using CDP Access keys

- Create a `requirements.txt` file specifying the Python package and version dependencies required by your CDE job.

Steps

1. [Generate CDP Access Keys](#) in User Management Console.
2. Generate DE workload auth token using CDP IAM API.

The IAM API endpoint `<CDP_ENDPOINT>/api/v1/iam/generateWorkloadAuthToken` is called to generate a CDP Access Token. A CDP API call requires a request signature to be passed in the "x-altus-auth" header, along with a corresponding timestamp in the "x-altus-date" header. The `cdpcurl` library constructs the headers automatically. However, if you would rather use a different HTTP client, such as ordinary `curl`, you can use the `cdpv1sign` script within `cdpcurl` to generate these required headers.

The request body contains workload-name as DE and is authenticated using the CDP Access Key. This request must also be signed as per the specification is available here either manually or use the `cdpv1sign` library to generate these necessary headers through automation script.

```
curl -X POST '<CDP_ENDPOINT>/api/v1/iam/generateWorkloadAuthToken' \
-H "Content-Type: application/json" \
-H "x-altus-date: Tue, 15 Mar 2022 07:22:57 GMT" \
-H "x-altus-auth: <signature-string-as-per-the-specification>" \
-i --insecure \
-d '{
  "workloadName": "DE"
}'
```



Note: `CDP_ENDPOINT` corresponds to CDP Control Plane console url. For example, `https://console-test-cp.apps.shared-os-dev-01.kcloud.cloudera.com`

3. The response will include the CDP Access Token in the `token` field and expiry time in the `expireAt`.

```
{
  "token": "<token-string>",
  "expireAt": "2021-05-03T15:34:03.727000+00:00"
}
```

4. Export the CDP token to `CDP_TOKEN` from above output.

```
export CDP_TOKEN=<token-string>
```

5. Once you have a CDP access token `CDP_TOKEN` from the previous step, you can manually exchange it for a CDE access token.

```
curl https://service.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/gateway/
authtkn/knoxtoken/api/v1/token \
-XPOST \
-H "Accept: application/json" \
-H "Authorization: Bearer ${CDP_TOKEN}" \
```

```
-i --insecure
```

6. In the output, the `access_token` value is the JWT. For convenience, copy it and set it as an environment variable:

```
export CDE_TOKEN=<access_token>
```

Alternatively, you can set the token in a single step using `jq` to extract the token:

```
export CDE_TOKEN=$(curl <auth_endpoint> -XPOST \
-H "Accept: application/json" \
-H "Authorization: Bearer ${CDP_TOKEN}" \
-i --insecure | jq -r '.access_token')
```

See [Using an access token in Cloudera Data Engineering API calls](#) for instructions on using the token in API calls.

Using an access token in Cloudera Data Engineering API calls

Cloudera Data Engineering (CDE) uses JSON Web Tokens (JWT) for API authentication.

Before you begin

Get an access token and save it as an environment variable as described in [Getting a Cloudera Data Engineering API access token](#).

Procedure

1. Determine the API URL for the virtual cluster you want to access using the API. The API URL for managing jobs is different from the URL used to obtain the access token.
 - a) Navigate to the Cloudera Data Engineering Overview page.
 - b) In the CDE Services column, select the environment containing the virtual cluster you want to interact with using the API.
 - c) In the Virtual Clusters column on the right, click the Cluster Details icon for the virtual cluster you want to interact with.
 - d) Click JOBS API URL to copy the link to your clipboard.
For example: `https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/dex/api/v1`
2. When you make an API call, include the JWT as a bearer token. For example, to list all jobs associated with the virtual cluster, assuming you have saved the token as an environment variable named `CDE_TOKEN`:

```
curl -H "Authorization: Bearer ${CDE_TOKEN}" -H "Content-Type: application/json" -X GET "https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/dex/api/v1/jobs"
```

Managing Cloudera Data Engineering job resources using the API

A *resource* in Cloudera Data Engineering (CDE) is a named collection of files referenced by a job. The files can include application code, configuration files, and even Python virtual environment specifications (`requirements.txt`). These resources can be managed using the CDE API.

Creating a Cloudera Data Engineering resource using the API

A *resource* in Cloudera Data Engineering (CDE) is a named collection of files referenced by a job. The files can include application code, configuration files, and even Python virtual environment specifications (requirements.txt). You can create a resource using the CDE API.

Before you begin

As with all API calls, make sure you have a valid access token. For instructions, see [Getting a Cloudera Data Engineering API access token](#).

Procedure

1. Create a JSON file describing the resource, using the structure for the type of resource you want to create.

The JSON payload for a resource is structured as follows:

files resource type

```
{
  "name": "<resourceName>",
  "type": "files",
  "retentionPolicy": "keep_indefinitely"
}
```

python-env resource type

```
{
  "name": "<resourceName>",
  "type": "python-env",
  "retentionPolicy": "keep_indefinitely",
  "pythonEnv": {
    "pythonVersion": "python3"
  }
}
```

custom-runtime-image resource type

```
{
  "customRuntimeImage": {
    "credential": "string",
    "engine": "string",
    "image": "string",
    "modified": "string"
  },
  "name": "<resourceName>",
  "retentionPolicy": "keep_indefinitely",
  "type": "custom-runtime-image"
}
```

2. Create the resource by submitting a POST request to the resources endpoint. The JSON filename is referenced using the @/path/to/filename.json convention. In this example, the JSON filename is fileResource.json, and describes a files type resource named example-job-files.

```
curl <jobs_api_url>/resources
-H "Authorization: Bearer ${CDE_TOKEN}" \
-H "Content-Type: application/json" \
-X POST -d @${HOME}/fileResource.json
```

3. Verify that the resource was created by requesting the resource details from the `/resources/<resourceName>` endpoint:

```
curl <jobs_api_url>/example-job-files \
-H "Authorization: Bearer ${CDE_TOKEN}" \
-H "Content-Type: application/json" \
-X GET
```

Deleting a Cloudera Data Engineering resource using the API

A *resource* in Cloudera Data Engineering (CDE) is a named collection of files referenced by a job. The files can include application code, configuration files, and even Python virtual environment specifications (requirements.txt). You can delete a resource using the CDE API.

Before you begin

As with all API calls, make sure you have a valid access token. For instructions, see [Getting a Cloudera Data Engineering API access token](#).

Procedure

1. Make sure that you no longer need the resource before deleting it.
2. Delete the resource by sending a DELETE request to the `/resources/<resourceName>` endpoint.
For example, to delete a resource named example-job-files:

```
curl <jobs_api_url>/example-job-files \
-H "Authorization: Bearer ${CDE_TOKEN}" \
-H "Content-Type: application/json" \
-X DELETE
```

Managing workload secrets with Cloudera Data Engineering Spark Jobs using the API

This API 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.

Creating a workload secret for Cloudera Data Engineering Spark Jobs using API

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

Procedure

1. Create Workload Credentials with name workload-cred:-

```
curl -k -X "POST" 'https://<dex-vc-host>/dex/api/v1/credentials' \
-H 'Accept: application/json' \
-H 'Connection: keep-alive' \
-H 'Content-Type: application/json' \
```

```
-H "Authorization: Bearer ${CDE_TOKEN}" \
--data '{
  "workloadCred": {
    "aws-secret": "secret123",
    "db-pass": "dbpass123"
  },
  "name": "workload-cred-1",
  "type": "workload-credential",
  "description": "workload credential description"
}'
```

2. Edit the existing workload credentials with new data. For example, edit workload-cred-1:

```
curl -k -X "PATCH" 'https://<dex-vc-host>/dex/api/v1/credentials/workload-cred-1' \
-H 'Accept: application/json' \
-H 'Connection: keep-alive' \
-H 'Content-Type: application/json' \
-H "Authorization: Bearer ${CDE_TOKEN}" \
--data '{
  "workloadCred": {
    "aws-secret": "newsecret123",
    "db-pass": "newdbpass123"
  }
}'
```

Listing an existing workload secret for Cloudera Data Engineering Spark Jobs using API

You can list an existing secret for Cloudera Data Engineering (CDE) Spark Jobs using API.

Procedure

1. List the existing workload credentials in the virtual cluster by creating a filter with type that is equal to workload-credential

```
curl -k -X "GET" 'https://<dex-vc-host>/dex/api/v1/credentials?filter=type%5Breq%5Dworkload-credential' \
-H 'Accept: application/json' \
-H 'Connection: keep-alive' \
-H 'Content-Type: application/json' \
-H "Authorization: Bearer ${CDE_TOKEN}"
```

Response:

```
{
  "credentials": [
    {
      "name": "workload-cred-1",
      "type": "workload-credential",
      "description": "workload credential description",
      "created": "2022-10-18T07:26:41Z",
      "modified": "2022-10-18T07:26:41Z"
    }
  ],
  "meta": {
    "hasNext": false,
    "limit": 20,
    "offset": 0,
    "count": 1
  }
}
```

```
}
}
```

2. Get the existing specific workload credential. For example, workload-cred-1:

```
curl -k -X "GET" 'https://<dex-vc-host>/dex/api/v1/credentials/workload-cred-1' \
-H 'Accept: application/json' \
-H 'Connection: keep-alive' \
-H 'Content-Type: application/json' \
-H "Authorization: Bearer ${CDE_TOKEN}"
```

Response:

```
{
  "name": "workload-cred-1",
  "type": "workload-credential",
  "description": "workload credential description",
  "created": "2022-10-18T07:26:41Z",
  "modified": "2022-10-18T07:26:41Z"
}
```

Deleting a workload secret for Cloudera Data Engineering Spark Jobs using API

You can delete an existing secret for Cloudera Data Engineering (CDE) Spark Jobs using API.

Procedure

Delete the workload-credential. For example, delete workload-cred-1:

```
curl -k -X "DELETE" 'https://<dex-vc-host>/dex/api/v1/credentials/workload-cred-1' \
-H 'Accept: application/json' \
-H 'Connection: keep-alive' \
-H 'Content-Type: application/json' \
-H "Authorization: Bearer ${CDE_TOKEN}"
```

Linking a workload secret to the Cloudera Data Engineering Spark Job definitions using API

Once you've created a workload secret, you can link it to the Cloudera Data Engineering (CDE) Spark Job definitions that you created using API.

Procedure

1. Use 1 or more workload-credentials in a new Spark job. For example, workload-cred-1 and workload-cred-2 in the job below:

```
curl -k -X "POST" 'https://<dex-vc-host>/dex/api/v1/jobs' \
-H 'Accept: application/json' \
-H 'Connection: keep-alive' \
-H 'Content-Type: application/json' \
-H "Authorization: Bearer ${CDE_TOKEN}" \
--data '{
  "name": "test-workload-job",
  "spark": {
```

```

      "className": "org.apache.spark.examples.SparkPi",
      "file": "spark-examples_2.11-2.4.8.7.2.12.0-195.jar"
    },
    "mounts": [
      {
        "resourceName": "spark-jar",
        "dirPrefix": ""
      }
    ],
    "type": "spark",
    "workloadCredentials": ["workload-cred-1", "workload-cred-2"]
  },
  '

```

2. Use 1 or more workload-credentials in an existing Spark job. For example, workload-cred-1 and workload-cred-2 in the job below:

```

curl -k -X "PATCH" 'https://<dex-vc-host>/dex/api/v1/jobs/test-workload-
job' \
  -H 'Accept: application/json' \
  -H 'Connection: keep-alive' \
  -H 'Content-Type: application/json' \
  -H 'Authorization: Bearer ${CDE_TOKEN}' \
  --data '{
    "workloadCredentials": ["workload-cred-1", "workload-cred-2"]
  },
  '

```

Using the workload secret in the Spark application code

To use the workload secret credentials, you can read the credentials that are mounted into the Spark drivers and executors as read-only files.

Procedure

The workload secrets are mounted into the Spark drivers and executors in this path: `/etc/dex/secrets/<workload-credential-name>/<credential-key-1>/etc/dex/secrets/<workload-credential-name>/<credential-key-2>`

Example workload credentials:

The workload-credential was created with the payload below.

```

{
  "workloadCred": {
    "aws-secret": "secret123",
    "db-pass": "dbpass123"
  },
  "name": "workload-cred-1",
  "type": "workload-credential",
  "description": "workload credential description"
}

```

The secrets can be read as local files from the paths below within the Spark drivers and executors:

`/etc/dex/secrets/workload-cred-1/aws-secret`

`/etc/dex/secrets/workload-cred-1/db-pass`

Example of a Pyspark application code to read a secret:

```

from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \

```

```
.appName("Sample DB Connection") \
.getOrCreate()

# read the password from the local file
dbPass=open("/etc/dex/secrets/workload-cred-1/db-pass").read()

# use the password in a jdbc connection
jdbcDF= spark.read \
```

Creating a Cloudera Data Engineering job using the API

You can create a job in Cloudera Data Engineering (CDE) using the CDE jobs API endpoint.

Before you begin

Request an access token and save it as an environment variable to use in API calls. For instructions, see [Getting a Cloudera Data Engineering API access token](#).

Procedure

1. Determine the API URL for the virtual cluster you want to access using the API:
 - a) Navigate to the Cloudera Data Engineering Overview page.
 - b) In the CDE Services column, select the environment containing the virtual cluster you want to interact with using the API.
 - c) In the Virtual Clusters column on the right, click the Cluster Details icon for the virtual cluster you want to interact with.
 - d) Copy the URL under JOBS API URL.
For example: `https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/dex/api/v1`
2. Submit the job creation request using the API to the `/jobs` endpoint.

The JSON payload to create a job is structured as follows:

```
{
  "name": "demoJob",
  "spark": {
    "className": "com.example.demoJobMainClass",
    "file": "local:/path/to/demoJobJar"
  },
  "type": "spark"
}
```

```
curl -H "Authorization: Bearer ${CDE_TOKEN}" <jobs_api_url>/jobs \
-H "Content-Type: application/json" \
-X POST -d '{"name":"demoJob","spark":{"className":"com.example.d
emoJobMainClass","file":"local:/path/to/demoJobJar"},"type":"spark
"}'
```

3. Verify the job was created. You can view job details using the `/jobs/<jobName>` endpoint:

```
curl -H "Authorization: Bearer ${CDE_TOKEN}" -H "Content-Type: applicati
on/json" -X GET "https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.
site/dex/api/v1/jobs/demoJob"
```

Listing Cloudera Data Engineering jobs using the API

You can list Cloudera Data Engineering (CDE) jobs using the API by issuing a GET request to the /jobs endpoint.

Before you begin

Get an access token and save it as an environment variable as described in [Getting a Cloudera Data Engineering API access token](#).

Procedure

1. Determine the API URL for the virtual cluster you want to access using the API:
 - a) Navigate to the Cloudera Data Engineering Overview page.
 - b) In the CDE Services column, select the environment containing the virtual cluster you want to interact with using the API.
 - c) In the Virtual Clusters column on the right, click the Cluster Details icon for the virtual cluster you want to interact with.
 - d) Copy the URL under JOBS API URL.
For example: `https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/dex/api/v1`
2. Issue a GET request to the /jobs endpoint:

```
curl -H "Authorization: Bearer ${CDE_TOKEN}" -X GET "https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/dex/api/v1/jobs"
```

Getting Cloudera Data Engineering job information using the API

You can get Cloudera Data Engineering (CDE) job information using the API by issuing a GET request to the /jobs/<job_name> endpoint.

Before you begin

Get an access token and save it as an environment variable as described in [Getting a Cloudera Data Engineering API access token](#).

Procedure

1. Determine the API URL for the virtual cluster you want to access using the API:
 - a) Navigate to the Cloudera Data Engineering Overview page.
 - b) In the CDE Services column, select the environment containing the virtual cluster you want to interact with using the API.
 - c) In the Virtual Clusters column on the right, click the Cluster Details icon for the virtual cluster you want to interact with.
 - d) Copy the URL under JOBS API URL.
For example: `https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/dex/api/v1`
2. Issue a GET request to the /jobs/<jobname> endpoint:

```
curl -H "Authorization: Bearer ${CDE_TOKEN}" -X GET "https://pmjkrn5.cde-czlmkz4y.na-01.xvp2-7p8o.cloudera.site/dex/api/v1/jobs/SparkPi"
```