Cloudera Data Engineering Top Tasks

Date published: 2022-09-30
Date modified:

Legal Notice

Contents

Creating Sessions in Cloudera Data Engineering

A Cloudera Data Engineering (CDE) Session is an interactive short-lived development environment for running Spark commands to help you iterate upon and build your Spark workloads.

About this task

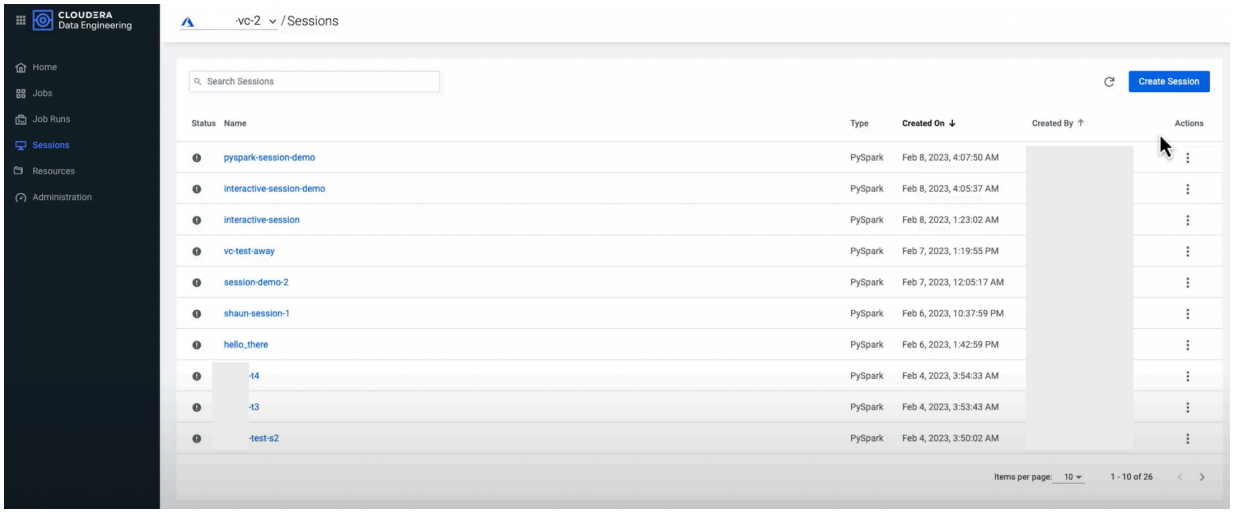
The commands that are run in a CDE Session are called Statements. You can submit the Statements through the connect CLI command or the Interact tab in the CDE UI for a Session. Python and Scala are the supported Session types. Learn how to use Cloudera Data Engineering (CDE) Sessions using the user interface.

Before you begin

Ensure that you are using a version of CDE 1.19 or higher for your Virtual Cluster.

Procedure

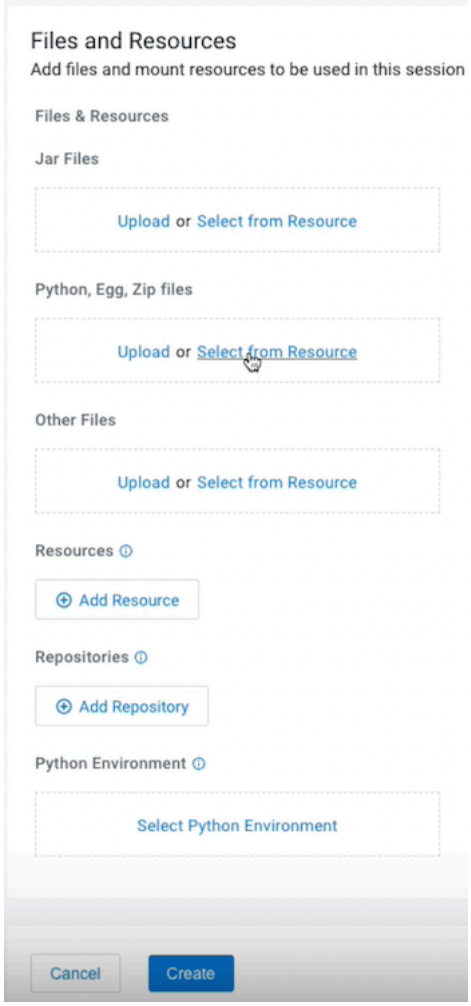
1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile. The Home page displays.
2. Click Sessions in the left navigation menu and then click Create Session.



3. Enter a Name for the Session.
4. Select a Type, for example, PySpark or Scala.
5. Select a Timeout value.
   The Session will stop after the indicated time has passed.
6. Optionally, enter a Description for the Session.
7. Optionally, enter the Configurations.
8. Set the Compute options.

9. Under **Files and Resources**, you can upload Jar, Python, Egg, Zip, and other files. You can also add a resource, respositories, or a Python environment to be used in this session.

   Files that are uploaded to a session are stored in the app/mount directory.



10. Click Create.

    The Connect tab displays a list of connectivity options available to interact with the Session. The Interact tab allows you to interact with the Session, and becomes available once the Session is running.

11. To delete a Session, open the Session and click Delete.

    Note: If you delete a Session, doing so will result in the termination of an active session and the loss of any attached logs and details.

Creating jobs in Cloudera Data Engineering

A job in Cloudera Data Engineering (CDE) consists of defined configurations and resources (including application code). Jobs can be run on demand or scheduled.

Before you begin

In Cloudera Data Engineering (CDE), jobs are associated with virtual clusters. Before you can create a job, you must create a virtual cluster that can run it. For more information, see Creating virtual clusters.

Important: The user interface for CDE 1.17 and above has been updated. The left-hand menu was updated to provide easy access to commonly used pages. The steps below will vary slightly, for example, the Overview page has been replaced with the Home page. You can also create a job by clicking Jobs on the left-hand menu, then selecting your desired Virtual Cluster from a drop-down at the top of the Jobs page. To view CDE Services, click Administration on the left-hand menu. The new home page still displays Virtual Clusters, but now includes quick-access links located at the top for the following categories: Jobs, Resources, and Download & Docs.

Important: The CDE jobs API implicitly adds the default DataLake filesystem to the Spark configuration to save the user having to do that. If you need to reference other buckets, you can set the spark.yarn.access.hadoopFileSystems parameter with the extra comma-separated buckets needed. If you set this parameter in your application code before creating the session, it might override the default setting, leading to errors.

Procedure

1. In the Cloudera Data Platform (CDP) management console, click the Data Engineering tile and click Overview.
2. In the CDE Services column, select the service that contains the virtual cluster that you want to create a job for.
3. In the Virtual Clusters column on the right, locate the virtual cluster that you want to use and click the View Jobs icon.
4. In the left hand menu, click Jobs.
5. Click the Create Job button.
6. Provide the Job Details:
   a) Select Spark for the job type. For Airflow job types, see Automating data pipelines using Apache Airflow DAG files in Cloudera Data Engineering.
   b) Specify the Name.
   c) Select Resources, URL, or Repository for your application file, and provide or specify the file. You can upload a new file or select a file from an existing resource.

      If you select URL and specify an Amazon AWS S3 URL, add the following configuration to the job:

      config_key: spark.hadoop.fs.s3a.delegation.token.binding

      config_value: org.apache.knox.gateway.cloud.idbroker.s3a.IDBDelegationTokenBinding

      If you want to use files from a repository, you must first create a repository. Once the repository is created, you can select Repository, click Add from Repository and select the file. Then, click Select File.
   d) If your application code is a JAR file, specify the Main Class.
   e) Specify arguments if required. You can click the Add Argument button to add multiple command arguments as necessary.
   f) Enter Configurations if needed. You can click the Add Configuration button to add multiple configuration parameters as necessary.

      Important: For Spark jobs, setting the `spark.app.id` property at the Spark job level configuration or within the Spark application code is not supported in CDE.

   g) If your application code is a Python file, select the Python Version, and optionally select a Python Environment.
7. Click Advanced Configurations to display more customizations, such as additional files, initial executors, executor range, driver and executor cores and memory.

   By default, the executor range is set to match the range of CPU cores configured for the virtual cluster. This improves resource utilization and efficiency by allowing jobs to scale up to the maximum virtual cluster resources available, without manually tuning and optimizing the number of executors per job.

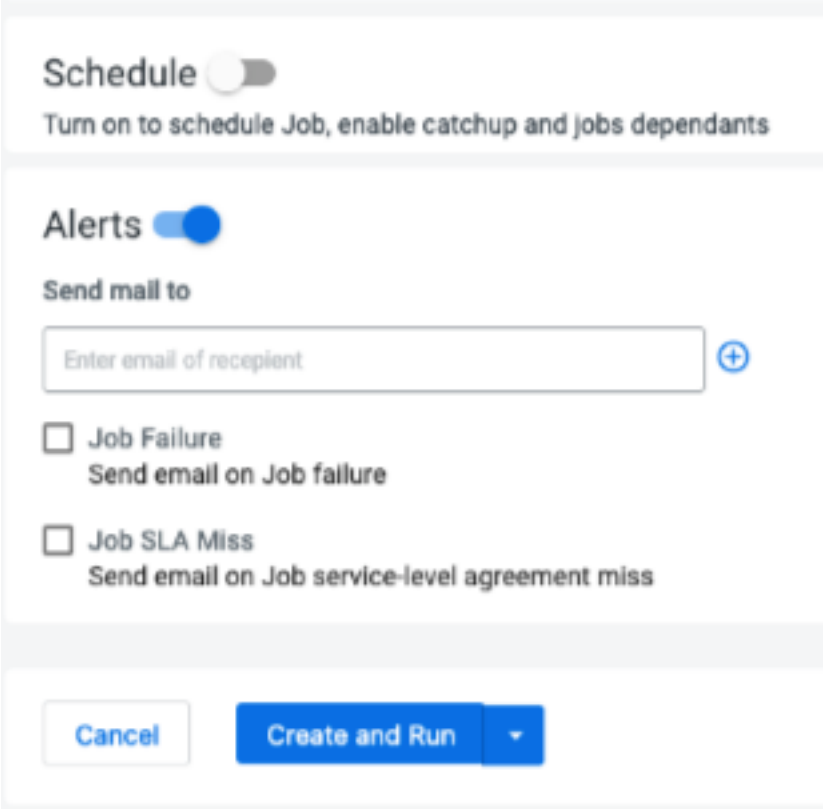
8. Click Schedule to display scheduling options.

   You can schedule the application to run periodically using the Basic controls or by specifying a Cron Expression.

   Note: Scheduled job runs start at the end of the first full schedule interval after the start date, at the end of the scheduled period. For example, if you schedule a job with a daily interval with a start_date of 14:00, the first scheduled run is triggered at the end of the next day, after 23:59:59. However if the start_date is set to 00:00, it is triggered at the end of the same day, after 23:59:59.

9. If you provided a schedule, click Schedule to create the job. If you did not specify a schedule, and you do not want the job to run immediately, click the drop-down arrow on Create and Run and select Create. Otherwise, click Create and Run to run the job immediately.

10. Optional: Toggle Alerts to send mail to the email address that you choose. You have the option to select Job Failure to send an email upon job failure, and Job SLE Miss to send an email on a Job service-level agreement miss.



Note: The Alerts option will only display if you have selected Configure Email Alerting during Virtual Cluster creation.

Creating an Airflow DAG using the Pipeline UI

With the CDE Pipeline UI, you can create multi-step pipelines with a combination of available operators.

About this task

This feature is available starting in CDE 1.16 in new Virtual Cluster installations.

Note: Cloudera supports all major browsers (Google Chrome, Firefox and Safari) for this feature. If you are using a browser in incognito mode, you have to allow all cookies in your browser settings so that you can view Pipelines, Spark, and Airflow pages.

Procedure

1. Go to Jobs Create Job .

   Under Job details, select Airflow.

   The UI refreshes, only Airflow-specific options remain.
2. Specify a name for the job.
3. Under DAG File select the Editor option.
4. Click Create.
   You are redirected to the job Editor tab.
5. Build your Airflow pipeline.
   - Drag and drop operators to the canvas from the left hand pane.
   - When selecting an operator, you can configure it in the editor pane that opens up.

     On the Configure tab you can provide operator-specific settings. The Advanced tab allows you to make generic settings that are common to all operators, for example execution timeout or retries.
   - Create dependencies between tasks by selecting them and drawing an arrow from one of the four nodes on their edges to another task. If the dependency is valid the task is highlighted in green. If invalid, it is highlighted in red.
   - To modify DAG-level configuration, select Configurations on the upper right.
6. When you are done with building your pipeline, click Save.

Scheduling jobs in Cloudera Data Engineering

Jobs in Cloudera Data Engineering (CDE) can be run on demand, or scheduled to run on an ongoing basis. The following instructions demonstrate how to create or modify a schedule for an existing job.

Before you begin

Important: The user interface for CDE 1.17 and above has been updated. The left-hand menu was updated to provide easy access to commonly used pages. The steps below will vary slightly, for example, the Overview page has been replaced with the Home page. You can also schedule a job from the new Home page by clicking Schedule a Job, or by clicking Jobs on the left-hand menu, then selecting your desired Virtual Cluster from a drop-down at the top of the Jobs page. To view CDE Services, click Administration on the left-hand menu. The new home page still displays Virtual Clusters, but now includes quick-access links located at the top for the following categories: Jobs, Resources, and Download & Docs.

Procedure

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile and click Overview.
2. In the CDE Services column, select the environment containing the virtual cluster where you want to schedule the job.
3. In the Virtual Clusters column on the right, click the View Jobs icon on the virtual cluster containing the job you want to schedule.
4. Click the Configure.
5. Click the Advanced Configurations link at the bottom of the page to view additional configuration parameters.
6. Click the Actions menu next to the application, and then click Configuration.

7. Select the Schedule toggle, and then set the Start time, End time, and Cron expression.

   The start and end times designate the time frame for which the schedule is active. The Cron expression uses the cron scheduling syntax to specify when the application should run within the start and end times. For information and examples of the cron syntax, see the Cron entry on Wikipedia.

   Note: Timestamps must be specified in ISO-8601 UTC format ('yyyy-MM-ddTHH:mm:ssZ'). UTC offsets are not supported.

   Note: Scheduled job runs start at the end of the first full schedule interval after the start date, at the end of the scheduled period. For example, if you schedule a job with a daily interval with a start_date of 14:00, the first scheduled run is triggered at the end of the next day, after 23:59:59. However if the start_date is set to 00:00, it is triggered at the end of the same day, after 23:59:59.
8. If you want to start a job immediately, check the Start job box.
9. Click Update to save your changes.
10. Select optional scheduling configurations:
    a) Select Enable Catchup to kick off job runs for any data interval that has not been run since the last data interval. If this option is not selected, only the runs that start after the time that the job was created will be included.
    b) Select Depends on Previous to ensure that each job run is preceeded by a successful job run.
11. Click Schedule.

Connecting to Grafana dashboards in Cloudera Data Engineering Public Cloud

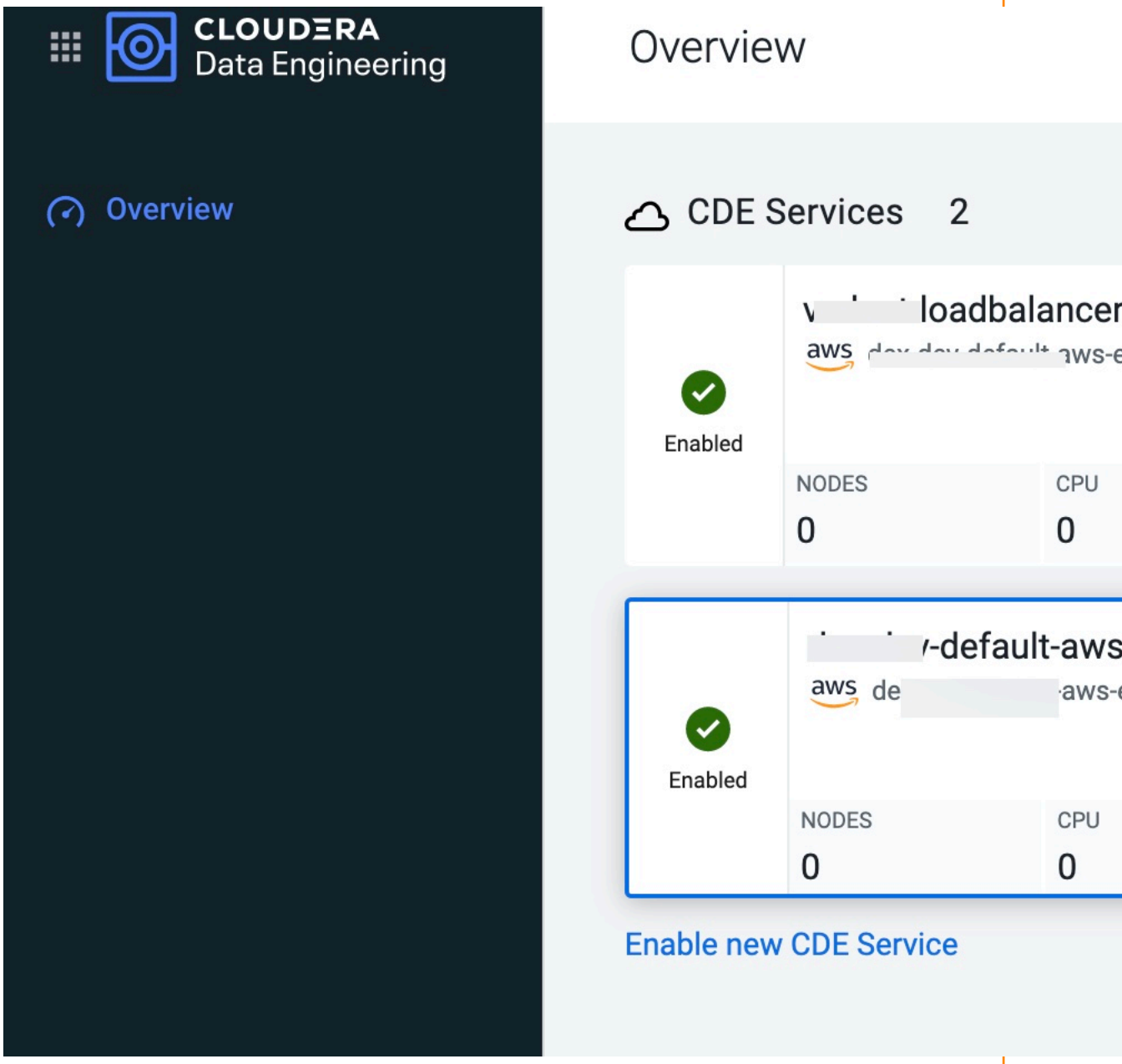
This topic describes how to access Grafana dashboards for advanced visualization of Virtual Cluster's metrics such as memory and CPU usage in Cloudera Data Engineering (CDE) Public Cloud.

Important: The user interface for CDE 1.17 and above has been updated. The left-hand menu was updated to provide easy access to commonly used pages. The steps below will vary slightly, for example, the Overview page has been replaced with the Home page. To view CDE Services, click Administration in the left-hand menu. The new home page still displays Virtual Clusters, but now includes quick-access links located at the top for the following categories: Jobs, Resources, and Download & Docs.

For CDE Service

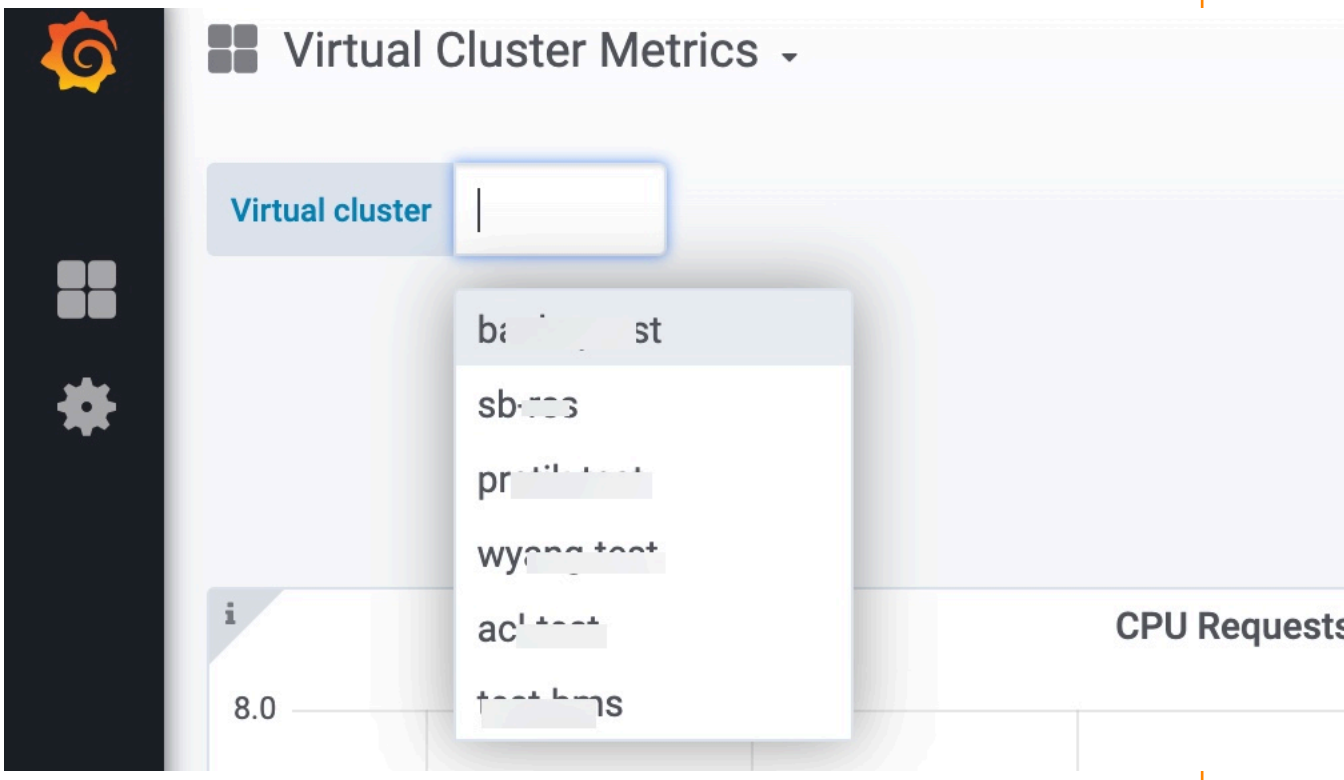
1. In the Cloudera Data Platform (CDP) management console, click the Data Engineering tile and click Overview.

2. In the CDE Services column, click the Service Detailsbutton on the environment for which you want to see the Grafana dashboard.



3. In the Service details page, click Grafana Charts in the hamburger menu. A read-only version of the Grafana interface opens in a new tab in your browser.
4. Select Virtual Cluster Metrics under the Dashboards pane.

5. Click on a virtual cluster name from the dropdown list to view the Grafana charts.



about CPU requests, memory requests, jobs, and other information related to the virtual cluster is displayed.

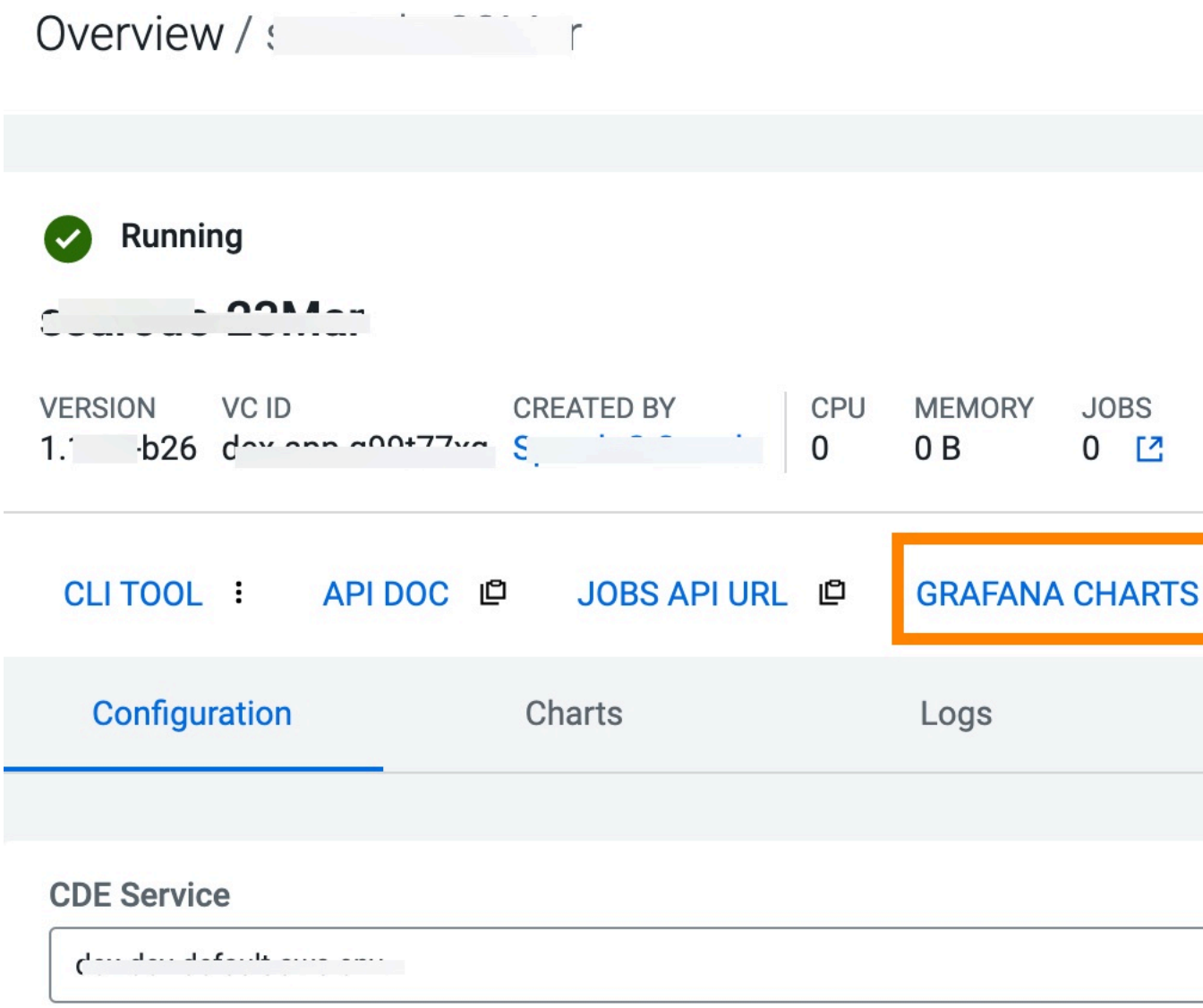
For Virtual Cluster

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 Service details column, select the environment containing the virtual cluster for which you want to see the Grafana dashboard.
3. In the Virtual Clusters column on the right, click the Cluster Details icon of the virtual cluster.

   The virtual cluster's Overview page is displayed.

4. In the Overview page, click Grafana Charts.

   A read-only version of the Grafana interface opens in a new tab in your browser.



   Information about CPU requests, memory requests, jobs, and other information related to the virtual cluster is displayed.

5. In the Virtual Cluster Metrics page, click on a virtual cluster name from the Virtual Cluster dropdown list to view the Grafana charts of that virtual cluster.