

Cloudera Data Engineering 1.23.0

Troubleshooting Cloudera Data Engineering

Date published: 2020-07-30

Date modified: 2024-11-12

The Cloudera logo is displayed in a bold, orange, sans-serif font. The word "CLOUDERA" is written in all caps, with the letter 'E' stylized as a horizontal bar with a small triangle in the center.

<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

Cloudera Data Engineering log files.....	4
Checking the node count on your Cloud Service Provider's website.....	4
Viewing Job run timeline.....	5
Accessing the Kubernetes dashboard.....	8
Connecting to Grafana dashboards in Cloudera Data Engineering Public Cloud.....	8
Cloudera Data Engineering diagnostic bundles and summary status.....	11
Downloading a diagnostic bundle for Cloudera Data Engineering.....	12
Downloading summary status for Cloudera Data Engineering.....	14
Preflight Checks.....	16
Cloudera Data Engineering CLI exit codes.....	26
Cloudera Data Engineering Deep Analysis.....	28
Enabling deep analysis on a Cloudera Data Engineering job from the CDE web UI.....	28
Enabling deep analysis on a Cloudera Data Engineering job run using the CLI.....	29
Running deep analysis on a Cloudera Data Engineering job run.....	29

Cloudera Data Engineering log files

You can view logs for Cloudera Data Engineering (CDE) using the web console, including CDE service logs, virtual cluster logs, and job logs.

To view logs for a CDE service or virtual cluster, click the three-dot menu for the service or virtual cluster, and then select View Logs. When the View Logs modal is displayed, you can download the logs or copy them to the clipboard by clicking the associated icon at the top right of the modal.



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 manage 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. 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.

To view logs for a job run:

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile and click Overview.
2. From the CDE Overview page, select the CDE service for the job you want to troubleshoot.
3. In the Virtual Clusters column, click the View Jobs icon for the cluster containing the job.
4. Select the job you want to troubleshoot.
5. In the Run History tab, click the Job ID for the job run you want to troubleshoot.
6. Go to the Logs tab.
7. Select the log you want to view using the Select log type drop-down menu and the log file tabs.
8. To download the logs, click the Download menu button. You can download a text file of the currently displayed log, or download a zip file containing all log files.

Checking the node count on your Cloud Service Provider's website

Learn about how to get real-time node count information on Amazon Web Services (AWS) and on Microsoft Azure.

Before you begin

To filter the nodes in the AWS Auto scaling groups (ASGs) or in Azure, obtain one of these identifiers:

- Provisioner ID
- Cluster name
- Cluster ID



Note: To filter using the Cluster name or Cluster ID:

- On AWS: you can filter the ASGs directly using the Cluster name or Cluster ID.
- On Azure, you must use the Add filter option on the Virtual machine scale sets page to query by the Cluster name or Cluster ID.

To obtain the provisioner ID (for example: *liftie-xyz*), in Cloudera Data Engineering, navigate to:

Administration Service-Details Logs and search for *provisionerID: liftie*.

Procedure

Navigate to your Cloud Service Provider's website.

For AWS

- a. In the AWS Management Console, log in as an IAM user, root user, or a similar role.
- b. At the top-right corner, select the region in which your cluster is located.



Note: Ensure that you select the correct region, because ASGs are region-specific.

- c. In the AWS Management Console Search bar, type EC2, and from the displayed list, select EC2.

The EC2 Dashboard for the selected region is displayed.

- d. In the EC2 Dashboard, in the left navigation pane, click Auto Scaling and select Auto Scaling Groups.

On the Auto Scaling Groups page, you can see a list of all ASGs in the selected region.

- e. In the search bar, filter the ASGs using one of the following:

- Cluster name
- Cluster ID
- provisioner ID

- f. To calculate the number of nodes in your Virtual Cluster (VC), add all the values displayed in the Instances column.



Note: The list includes infrastructure nodes also.

- g. Optional: If your cluster operates in multiple regions, switch to another region at the top-right corner of the AWS Management Console and repeat the steps described in this procedure.

For Azure

- a. Login to Azure Portal.
- b. In the top-right corner, select the Directory and the Subscription where your VC is hosted.



Note: Even though regions are not selected at this step, make sure that you know in which region your VC operates.

- c. At the top of the portal, in the Search bar, type Virtual machine scale sets and select it from the displayed list.

The Virtual machine scale sets page is displayed, where all Virtual Machine Scale Sets (VMSS) in your subscription are listed.

- d. To narrow down the list, perform one of these options:

- If your cluster resources are grouped within a specific Resource group, use the Resource group filter to narrow down the list.
- To search by the Cluster Name or Cluster ID tags associated with your Cluster or node pools, use the Add filter option. If you add a tag filter, the list of VMSS associated with the queried Cluster ID or Cluster Name is displayed.
- To filter by the Provisioner ID, enter the Provisioner ID in the Search field on the VMSS page.



Note: The list includes infrastructure nodes also.

- e. To calculate the number of nodes in your VC, add up all the values displayed in the Instances column.
- f. Optional: If your VC spans through multiple regions, make sure that you select the correct Resource group, or use the global search to find resources in the specific region.

Viewing Job run timeline

You can view the intermediate stages of the job run at every stage during its life cycle in real-time.

About this task

In case of a job failure, you can view the specific event and component where the job run failed. This reduces turnaround time during the debugging process for job run failure. You can see the step-by-step advancement of the job run on the UI, including all the granular details instead of reviewing extensive logs to obtain the same insights.

Procedure

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile. The CDE Home page displays.
2. In the left navigation pane, click Jobs Runs. The Jobs Runs page displays.
3. Click on the Job Id for which you want to see the status.

4. Go to the Timeline tab. It displays the summary of the Job run progression in the reverse chronological order for both primary and subordinate stages.

The screenshot displays the Cloudera Data Engineering web interface. On the left is a dark sidebar with navigation links: Home, Jobs, Job Runs (highlighted with an orange box), Sessions, Repositories, Resources, and Administration. The main content area on the right shows the breadcrumb 'cluster-01 / Job Runs /' and a 'Status' section indicating 'Succeeded'. Below this, the 'Timeline' tab is selected and highlighted with an orange box, with a 'New' button next to it. The 'Summary' section, titled 'View your Job run progression i', lists three job stages: 'Succeeded' (Component: Runtime API Server), 'Running' (Component: Spark Container), and 'Starting' (Component: Runtime API Server). Each stage includes a 'Show Details' link.

Accessing the Kubernetes dashboard

As a user with the DEAdmin role, you can access the Kubernetes dashboard in Cloudera Data Engineering (CDE) for troubleshooting and debugging of the cluster deployment.

About this task

The Kubernetes dashboard is to be used for troubleshooting in coordination with Cloudera support. The Kubernetes dashboard allows you to view and download diagnostics and container logs. Additionally, the DEAdmin can initiate limited actions such as restarting pods, modifying configuration maps and deployments which is beneficial for a debugging session.



Note: If you prefer to use the kubectl tool for troubleshooting instead, see [Enabling kubectl for CDE](#) linked below. The kubectl tool is also to be used in coordination with Cloudera support.

Before you begin

Ensure that you have the DEAdmin role.



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.

Procedure

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile and click Overview.
2. In the CDE Services column, click Service Details.
3. Copy the Resource Scheduler URL and open a new browser window.
4. Paste the Resource Scheduler URL in your new browser window and edit the URL as follows:
 - a) In the URL, replace the word "YuniKorn" with "dashboard". For example, `https://yunikorn.` becomes `https://dashboard..`
 - b) Press Enter

Results

The Kubernetes Dashboard displays and provides an easy user experience for monitoring your diagnostics.

Related Information

[Deploy and Access the Kubernetes Dashboard](#)

[Enabling kubectl for CDE](#)

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.

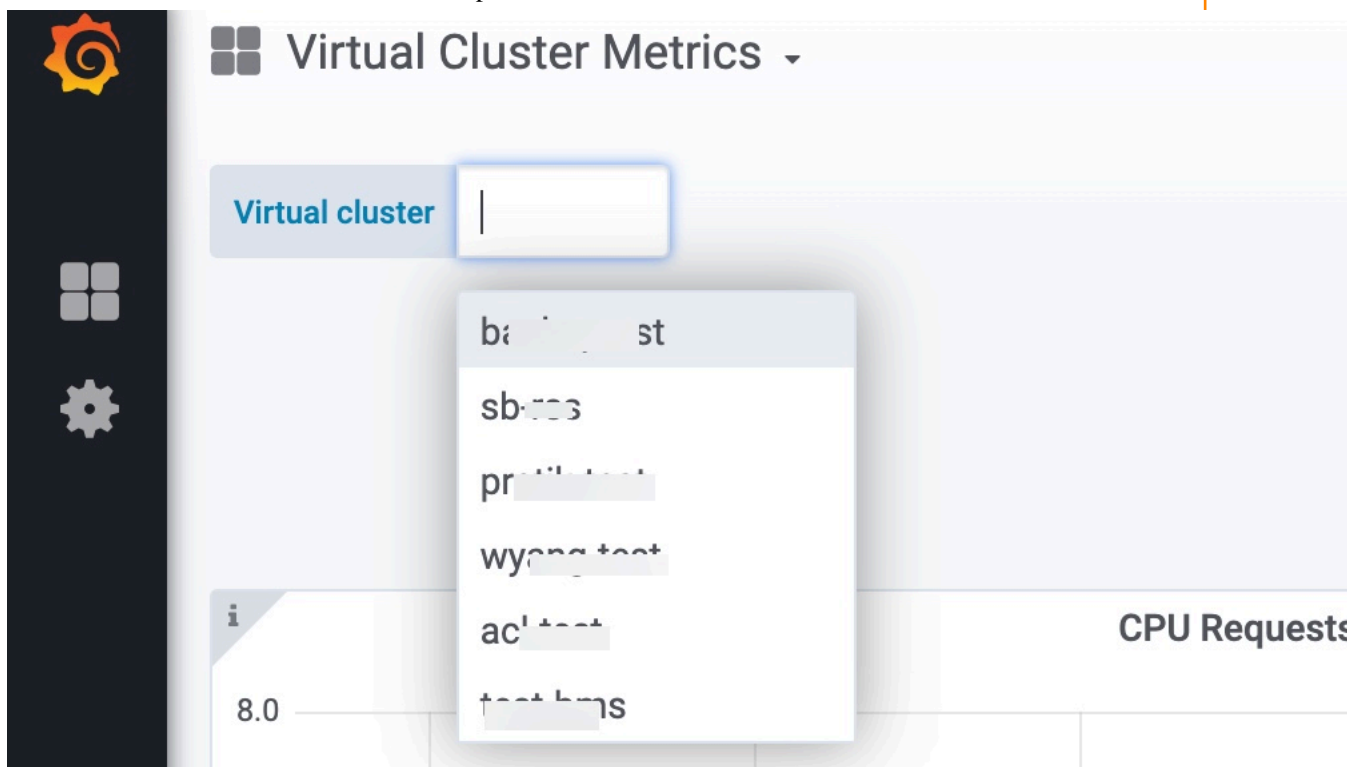
For CDE Service

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 Details button on the environment for which you want to see the Grafana dashboard.

The screenshot displays the Cloudera Data Engineering (CDE) Overview page. The left-hand navigation menu is dark blue with the Cloudera logo and 'Data Engineering' text. The 'Overview' link is highlighted. The main content area is light blue and titled 'Overview'. It shows 'CDE Services 2'. Two service cards are listed, both with a green checkmark and 'Enabled' status. The first card is for 'loadbalancer' and the second is for 'r-default-aws'. Both cards show 'NODES 0' and 'CPU 0'. A blue box highlights the second service card. At the bottom, there is a link 'Enable new CDE Service'.

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.

Overview / [\[redacted\]](#)

Running

[\[redacted\]](#) 23Mar

VERSION	VC ID	CREATED BY	CPU	MEMORY	JOBS
1.1	b26	dcy-ann-s00+77va	0	0 B	0 [external link icon]

[CLI TOOL](#) :
 [API DOC](#)
[JOBS API URL](#)
[GRAFANA CHARTS](#)

[Configuration](#)
[Charts](#)
[Logs](#)

CDE Service

[\[redacted\]](#)

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.

Cloudera Data Engineering diagnostic bundles and summary status

Cloudera Data Engineering provides the ability to download log files, diagnostic data, and a summary status for the running CDE services and virtual clusters. You can provide this data to Cloudera Support for assistance troubleshooting an issue.

The following section describes about CDE Diagnostic Bundles and Summary Status, and the information collected in each.



Note: A single click now generates one unified bundle containing both service logs and summary status.

Diagnostic Bundle

The diagnostic bundle is a collection of the logs from the CDE Services. These logs can be downloaded on-demand from CDE UI. CDE gives you the functionality to select the sources for which you want to download the logs and you can also select a predefined or custom time range for these logs.

Information Collected in Diagnostic Bundles

- Container logs for all running CDE service pods (excluding user compute pods).

Summary Status

The Status Summary shows the status of each service instance being managed by the CDE Control Plane. It is a package of JSON files consisting of information related to configuration, monitoring, logging, events and health test results of the service and its instances.

Information Collected in Summary Status

- Status of all cloud resources created during CDE provisioning
- Kubernetes resource status for critical infrastructure pods, deployments, pods, services and events for core service infrastructure pods or virtual cluster infrastructure pods

Downloading a diagnostic bundle for Cloudera Data Engineering

This section describes how to download a diagnostic bundle to troubleshoot a Cloudera Data Engineering (CDE) Service in Cloudera Data Platform (CDP) using CDE UI.

About this task

To troubleshoot issues with your CDE Service, download diagnostic bundles of log files. These diagnostic bundles, in the form of ZIP files, are downloaded to your local machine. This is available on both AWS and Azure environments.

Before you begin

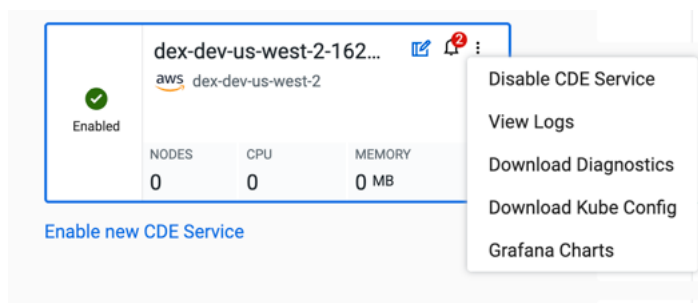
Required Role: DEAdmin

After granting or revoking the role on the environment, run the Sync Users to FreeIPA environment action.

Procedure

1. Log in to the CDP web interface and navigate to the CDE service
2. In the CDE service, click Overview in the left navigation panel, and select the CDE Service for which you want to download a diagnostic bundle.

3. In the selected CDE service, click on the three-dot menu, and select the Download Diagnostics option.



4. On the Diagnostics tab of the Service details, click the Generate Diagnostics Bundle button. The Diagnostic Bundle Options dialog box launches.
5. Set the time range and select the log sources that you want to include:

Generate Diagnostics Bundle

Select a time range for Logs

☒ Pre-defined Range ☐ Custom Range

Last 30 Min

Log Sources

All

Cancel

Generate

In the dialog box, you can choose or set the following options:

- **Pre-defined Range:** Select a specific time range of log files to generate from the drop-down list, or you can choose a custom interval in the next option.
 - **Custom Range:** Select the start and end time from the drop-down list to define the specific time interval for the log files in the diagnostic bundle.
 - **Log Sources:** Select All or any of the available categories of log files for the different Services and virtual clusters by clicking the adjacent checkbox.
6. After selecting which log sources you want to include in the diagnostic bundle, click Generate to generate the bundle.
 7. When the diagnostic bundle is created, click Download.

Results

When you extract the diagnostic bundle .gz file that you downloaded from CDE UI, you find the directories that contain service log files and summary status.

Downloading summary status for Cloudera Data Engineering

This section describes how to download a summary status of a Cloudera Data Engineering (CDE) Service in Cloudera Data Platform (CDP) using CDE UI.

About this task


To get a current snapshot of the current CDE infrastructure status (cloud resources, Kubernetes pod statuses, pod events and service metadata). This summary status, in the form of a ZIP file, is downloaded to your computer. This is available on both AWS and Azure environments.

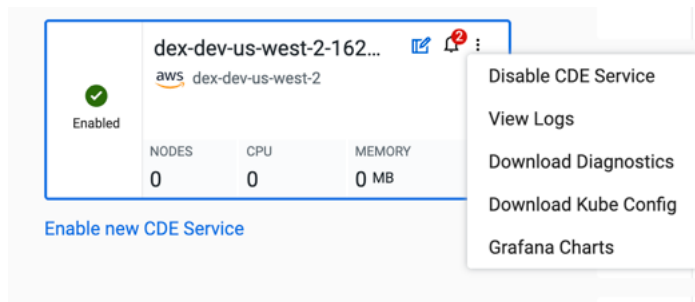
Before you begin

Required Role: DEAdmin

After granting or revoking the role on the environment, run the Sync Users to FreeIPA environment action.

Procedure

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile. The CDE Home page displays.
2. Click Administration in the left navigation menu and locate the service in the Services column, and click Service Details on the environment for which you want to download a summary status.
3. In the selected CDE service, click  to see a drop down menu, and select Download Diagnostics.



4. It takes you to the Diagnostics tab of the Service details. Click the Download Summary Status button.

✓ Enabled

cde-XXXXXXXXXX

VERSION	CLUSTER ID	CREATED BY	NODES	CPU	MEMORY	DATA LAKE	ENVIRONMENT
1	XXXXXXXXXX	XXXXXXXXXX	1 / 50	0 / 800	0 MB / 3200 GB	XXXXXXXXXX	XXXXXXXXXX

[GRAFANA CHARTS](#) [RESOURCE SCHEDULER](#)

Configuration Charts Logs **Diagnostics**

[Generate Diagnostics Bundle](#)

Status	Requested Time ↓	Requested Range	Time Range	Actions
✓	Feb XXXXXXXXXX AM, 11:24:11 AM	30 minutes	Download Bundle location	

Items per page: 10 1 – 1 of 1

[Switch to Old Diagnostics UI](#)

5. Summary Status gets downloaded on your local machine in a Zip file.

Results

When you extract the Summary Status ZIP file that you downloaded from CDE UI, you can see the directories that contain files in JSON format.

Below is an example of format of the Zip file for Summary Status:



Note: In the extracted Summary Status file, cluster.json refers to the CDE service and 'instances' refer to CDE virtual clusters.

```
cde-service-diagnostics-{clusterID}-{timestamp}/
cluster.json           # output from cluster describe
cluster-events.json    # output from get cluster events
cloud-diagnostics.json # cloud resource status (RDS, EFS, EKS, LB, ...)
instances/
  {id}.json            # output from instance id describe
  {id}-events.json     # output from get instance id events
kubernetes/
  namespaces/
    nodes.json
    {namespace}/
      namespace.json
      pods.json
```

```

events.json
deployments.json
replicasets.json
services.json
daemonsets.json
{namespace}/
namespace.json
pods.json
events.json
deployments.json
replicasets.json
services.json
daemonsets.json
...

```

Preflight Checks

Part of provisioning and managing a Kubernetes cluster in the cloud is ensuring that your account and your environment is properly configured before beginning. The Cloudera Data Platform (CDP) automatically runs a series of preflight checks which can help in determining if there is a problem before creating new resources or adjusting existing ones.

When creating a brand new cluster, preflight validation checks are used to ensure that the resources you are requesting as well as your environment are ready and configured correctly. For existing clusters, many of the infrastructure validations are skipped since they are not needed anymore. Instead, the validations which concern the resources being adjusted are executed.

Results

The result of running a collection of preflight checks is represented as a single aggregated value that lets you know whether it is safe to proceed with your actions. Each individual validation which was run is included in the response.

A preflight validation contains information concerning what it was checking for and what the result of that check was. For example, this is a very simple check to ensure that the type of node pool image is valid in the region:

Name	Instance type
Description	Instance groups must have an instance type that exists in the region in which they will be created. For EKS, there is additional verification for EKS support and usage class.
Category	COMMON
Status	FAILED
Message	The instance type validation failed.
Detailed Message	Instance type validation failed for Standard_B2s1 in west us2. Check to ensure that this instance type is valid in that region.
Duration	659 ms

PASSED

The validation successfully passed all criteria.

WARNING

The validation was either unable to fully check all of its criteria or it found a potential issue which could affect the success of the operation. This type of failure will not stop a cluster from being created or modified, but it does require further investigation.

SKIPPED

The validation was skipped because it does not apply to the current request. This can happen for many reasons, such as using a cloud provider that is not applicable to the preflight check.

FAILED

The validation failed its expected criteria and provisioning or updating of a cluster cannot proceed. In this case, there's an identified problem that needs resolution before continuing.

List of Preflight Checks**Node / Agent Pools**

The following preflight validation checks pertain to the node / agent instance groups which are created to launch new nodes inside of the Kubernetes cluster.

Name	Description	Cloud	Type
Instance Count	<p>The number of nodes in each instance group must not exceed a predefined threshold.</p> <p>Remediation: Change the requested size of an instance group to be within the boundaries specified by the error message.</p>	All	Create Update
Group Count	<p>The number of distinct node/agent pool groups must not exceed a predefined threshold.</p> <p>Remediation: Change the number of distinct instance groups to be within the boundaries specified by the error message.</p>	All	Create Update
Instance Naming	<p>The name of each instance group is restricted by cloud providers. There are differences in the size and characters allowed by each provider.</p> <p>Remediation: Change the name of the instance group to conform to the cloud provider's requirements. This could include changing the overall length of the name</p>	All	Create Update

Name	Description	Cloud	Type
	or only using certain approved characters.		
Instance Type	<p>Instance image types are not universal across all regions. Some providers further restrict this to service level, and whether they can be used for Kubernetes.</p> <p>Remediation: Choose a different region or service type for the desired instance type. If this is not possible, then a different instance type must be chosen.</p>	All	Create Update
Kubernetes Version	<p>Each cloud provider supports different versions of Kubernetes. CDP has also only been certified to work with particular versions.</p> <p>Remediation: Choose a different version of Kubernetes that is supported by your cloud provider in the region you are deploying in.</p> <p>https://docs.aws.amazon.com/eks/latest/userguide/kubernetes-versions.html</p>	All	Create
Placement Rules	<p>Some instance types are not allowed to be grouped within a single availability zone.</p> <p>Remediation: Remove the restriction on single availability zone, or choose a different instance type.</p>	AWS	Create Update

Infrastructure

The following preflight validation checks pertain to Cloudera's control plane infrastructure and your specific account within that control plane.

Name	Description	Cloud	Type
Restricted IAM Policies	<p>If your account is trying to provision with restricted IAM policies, then it needs to have those policies defined before deploying the cluster.</p> <p>Remediation: Check Cloudera's documentation on restricted IAM policies to ensure that you have the correctly named policies defined and are accessible.</p>	AWS	Create
Proxy Connectivity	<p>When provisioning a private cluster, your environment must have the cluster proxy enabled and it must be healthy.</p> <p>Remediation: Create a new environment which has the cluster proxy service enabled or check that your existing FreeIPA Virtual Machine is running and healthy.</p> <p>https://docs.cloudera.com/management-console/test/connection-to-private-subnets/topics/mc-ccm-overview.html</p>	All	Create Update
Data Lake Connectivity	<p>A healthy data lake with a functioning FreeIPA DNS server is required in order to provision a new cluster.</p> <p>Remediation: Check to ensure that the FreeIPA DNS server is running and healthy inside of your network.</p> <p>https://docs.cloudera.com/management-console/cloud/data-lakes/topics/mc-data-lake.html</p> <p>https://docs.cloudera.com/management-console/cloud/identity-management/topics/mc-identity-management.html</p>	All	Create

Networking

The following preflight validation checks pertain to specific network configurations inside of the cloud provider.

Name	Description	Cloud	Type
Shared VPC	<p>When a Virtual Private Cloud is shared between multiple subscriptions, access to modify this VPC needs to be granted.</p> <p>Remediation: Check the permission for which roles can make modifications to the VPC</p> <p>https://docs.aws.amazon.com/vpc/latest/userguide/vpc-sharing.html</p> <p>https://docs.cloudera.com/cdp-public-cloud/cloud/requirements-aws/topics/mc-aws-req-vpc.html</p>	AWS	Create
Subnet Availability Zone	<p>All subnets which are part of the environment must be located in at least 2 different Availability Zones.</p> <p>Remediation: Recreate your environment and choose subnets that satisfy the requirement of being in at least 2 different Availability Zones.</p>	AWS	Create
Subnet Load Balancer Tagging	<p>In order for load balancers to choose subnets correctly, a subnet needs to have either the public or private ELB tags defined.</p> <p>Remediation: Tag subnets with either <code>kubernetes.io/role/elb</code> or <code>kubernetes.io/role/internal-elb</code> based on whether they are public or private.</p> <p>https://docs.cloudera.com/cdp-public-cloud/cloud/requirements-aws/topics/mc-aws-req-vpc.html</p>	AWS	Create

Name	Description	Cloud	Type
API Server Access	<p>Validates that there are no conflicting requests between Kubernetes API CIDR ranges and private AKS clusters.</p> <p>Remediation: When using a private AKS cluster, Kubernetes API CIDR ranges are not supported,</p>	All	Create Update
Available Subnets	<p>Subnets cannot be shared when provisioning Kubernetes clusters on Azure. At least one available subnet must exist that is not being used by another AKS cluster and must not have an existing route table with conflicting pod CIDRs.</p> <p>Remediation: Create a new subnet to satisfy this requirement or delete an old and unused cluster to free an existing subnet.</p>	Azure	Create
Delegated Subnet	<p>A subnet which has been delegated for a particular service cannot be used to provision an Azure AKS cluster.</p> <p>Remediation: Choose a different subnet or remove the delegated service from at least one subnet in the environment.</p>	All	Create
Kubernetes API Server Security	<p>Validates that the supplied IP CIDR ranges are valid and do not overlap any reserved IP ranges. Each cloud provider has a limit set on the maximum number of allowed CIDRs.</p> <p>Remediation: Use valid CIDR formats and ranges when limiting access to the Kubernetes API server and limit the number of ranges specified.</p>	All	Create

Name	Description	Cloud	Type
Kubernetes Service CIDR Validation	<p>Validates that the specified service CIDR for Kubernetes services does not overlap any restricted CIDR ranges and is a valid CIDR format.</p> <p>Remediation: Change the service CIDR so that it doesn't conflict with any pod CIDRs or other routes on the subnet.</p>	All	Create
Autoscale Parameters	<p>Azure's built-in autoscaler has limitations on the ranges of values for scale-up and scale-down operations.</p> <p>Remediation: Adjust the specified parameters from the error message which are not within the required ranges.</p>	Azure	Create Update

Example

```
{
  "result": "PASSED",
  "summary": {
    "passed": 8,
    "warning": 0,
    "failed": 0,
    "skipped": 10,
    "total": 18
  },
  "message": "The cluster validation has passed, but some checks were skipped",
  "validations": [
    {
      "name": "Instance Count",
      "description": "Each instance count must be between minInstance and maxInstance inclusively. The minInstance and maxInstance of infrastructure group should comply with minimum number of infra nodes and maximum number of infra nodes.",
      "category": "COMMON",
      "status": "PASSED",
      "message": "The minimum and maximum instance counts are correct for all instance groups.",
      "detailedMessage": "The minimum and maximum instance counts are correct for all instance groups.",
      "duration": "1µs"
    },
    {
      "name": "Instance Group Count",
      "description": "Total instance group count must be less than or equal to maximum instance group limit.",
      "category": "COMMON",
```

```

        "status": "PASSED",
        "message": "The number of instance groups in the request is less
than or equal to the maximum allowed.",
        "detailedMessage": "The total instance group count of 2 is within
the limit.",
        "duration": "3µs"
    },
    {
        "name": "Instance Group Naming",
        "description": "Each instance group name must conform the rest
rixtions of the cloud provider. This includes using valid characters and adh
ering to length restrictions.",
        "category": "COMMON",
        "status": "PASSED",
        "message": "All instance groups meet the naming restrictions for
Azure.",
        "detailedMessage": "All instance groups meet the naming restrict
ions for Azure.",
        "duration": "12µs"
    },
    {
        "name": "Instance Type",
        "description": "Instance groups must have an instance type that
exists in the region in which they will be created. For EKS, there is addi
tional verification for EKS support and usage class.",
        "category": "COMMON",
        "status": "PASSED",
        "message": "All instance groups have valid instance types.",
        "detailedMessage": "The following instance types were validated
for westus2: Standard_B2s",
        "duration": "599ms"
    },
    {
        "name": "Kubernetes Version",
        "description": "Each cloud provider (Amazon, Azure, Google, etc)
supports different versions of Kubernetes.",
        "category": "COMMON",
        "status": "PASSED",
        "message": "The specified Kubernetes version 1.18 has been resol
ved to 1.18.17 and is valid on Azure",
        "detailedMessage": "The specified Kubernetes version 1.18 has
been resolved to 1.18.17 and is valid on Azure",
        "duration": "388ms"
    },
    {
        "name": "Placement Rule",
        "description": "Instance Types must be allowed by the placement r
ule.",
        "category": "COMMON",
        "status": "SKIPPED",
        "message": "Skipping validation since the cloud platform is Azure
.",
        "detailedMessage": "Skipping validation since the cloud platform
is Azure.",
        "duration": "6µs"
    },
    {
        "name": "Entitlement Check",
        "description": "When the entitlement LIFTIE_USE_PRECREATED_IAM_RE
SOURCES is enabled, the expected profile (cdp-liftie-instance-profile) shoul
d exist and it needs to have the necessary roles attached to it. ",
        "category": "ENTITLEMENTS",
        "status": "SKIPPED",

```

```

    "message": "IAM Resource Entitlement validation skipped for Cl
oud Provider azure.",
    "detailedMessage": "IAM Resource Entitlement validation skipped
for Cloud Provider azure.",
    "duration": "14µs"
  },
  {
    "name": "Cluster Proxy Connectivity",
    "description": "Verifies connectivity to the cluster proxy ser
vice which is used to register private cluster endpoints.",
    "category": "CONTROL_PLANE",
    "status": "SKIPPED",
    "message": "Connectivity to the cluster connectivity manager will
be skipped since this is not a private cluster.",
    "detailedMessage": "The cluster being provisioned is not marked
as private in the provisioning request.",
    "duration": "1µs"
  },
  {
    "name": "Cluster Proxy Enabled",
    "description": "Verifies that the environment was created with th
e cluster proxy service enabled.",
    "category": "CONTROL_PLANE",
    "status": "SKIPPED",
    "message": "Skipping the environment check for cluster proxy c
onnectivity since the cluster is public.",
    "detailedMessage": "Skipping the environment check for cluster
proxy connectivity since the cluster is public.",
    "duration": "1µs"
  },
  {
    "name": "DataLake Connectivity",
    "description": "Validates whether DataLake connection is reach
able and if FreeIPA is available.",
    "category": "CONTROL_PLANE",
    "status": "PASSED",
    "message": "DataLake validation succeeded.",
    "detailedMessage": "Data lake is healthy and reachable. Service
Discovery Feature is enabled, verified DNS entries retrieved for Data Lake
s. Datalake URL : localhost:8081 Service discovery URL : localhost:8082 "
  },
  {
    "name": "AWS Shared VPC Access",
    "description": "When a shared VPC is used, proper access should
be granted.",
    "category": "NETWORK",
    "status": "SKIPPED",
    "message": "Skipping validation since the cloud platform is Azur
e.",
    "detailedMessage": "Skipping validation since the cloud platform
is Azure.",
    "duration": "3µs"
  },
  {
    "name": "AWS Subnet Availability Zones",
    "description": "When existing AWS subnets are provided for provi
sioning an EKS cluster, the subnets must be in at least 2 different Availabi
lity Zones.",
    "category": "NETWORK",
    "status": "SKIPPED",
    "message": "Skipping validation since the cloud platform is Azure
.",
    "detailedMessage": "Skipping validation since the cloud platform
is Azure.",

```



```

        "duration": "2µs"
      },
      {
        "name": "AWS Subnet Tagging",
        "description": "In order for load balancers to choose subnets co
rrectly a subnet needs to have either the public or private ELB tags defined
.",
        "category": "NETWORK",
        "status": "SKIPPED",
        "message": "Skipping validation since the cloud platform is Azu
re.",
        "detailedMessage": "Skipping validation since the cloud platform
is Azure.",
        "duration": "26µs"
      },
      {
        "name": "Azure API Access Parameters",
        "description": "Verifies that the security parameters for locking
down access to the Azure Kubernetes API Server are correct.",
        "category": "NETWORK",
        "status": "PASSED",
        "message": "The API server access parameters specified in the cl
uster request are valid.",
        "detailedMessage": "The cluster will be provisioned as public
with the following whitelist CIDRs: ",
        "duration": "48µs"
      },
      {
        "name": "Azure Available Subnets",
        "description": "When an existing Azure subnet is chosen for pro
visioning an AKS cluster, the subnet must not be in use by any other cluster
. This is a restriction of Kubenet, which is the CNI used on the new cluster
. Although the subnet may have a routing table, it may not have any existing
IP address associations.",
        "category": "NETWORK",
        "status": "PASSED",
        "message": "At least 1 valid subnet was found and can be used for
cluster creation.",
        "detailedMessage": "The cluster can be provisioned using subnet
liftie-dev.internal.2.westus2 in virtual network liftie-dev and resource gr
oup liftie-test",
        "duration": "917ms"
      },
      {
        "name": "Kubernetes API Server CIDR Security",
        "description": "CIDR blocks for whitelisting access to the Kube
rnetes API Server must not overlap restricted IP ranges.",
        "category": "NETWORK",
        "status": "SKIPPED",
        "message": "Skipping CIDR validation for whitelisting because it
is not enabled.",
        "detailedMessage": "The ability to secure access the Kubernetes
API server via a list of allowed CIDRs is not enabled. This can be enabled
either in the controlplane (currently false) or via the provisioning reques
t (currently false).",
        "duration": "11µs"
      },
      {
        "name": "Service CIDR Validation",
        "description": "CIDR blocks that Kubernetes assigns service IP
addresses from should not overlap with any other networks that are peered or
connected to existing VPC.",
        "category": "NETWORK",
        "status": "SKIPPED",

```

```

      "message": "Service CIDR is missing in Network Profile.",
      "detailedMessage": "VPC validation is executed only if the VPC already exists and a service CIDR is specified in the network profile.",
      "duration": "426µs"
    },
    {
      "name": "Azure Autoscale Parameters",
      "description": "The following autoscale parameters for Azure, which are specified during provisioning and update, need to be in multiples of 60s. Autoscale parameters: scaleDownDelayAfterAdd, scaleDownDelayAfterFailure, scaleDownUnneededTime, scaleDownUnreadyTime.",
      "category": "DEPLOYMENT",
      "status": "SKIPPED",
      "message": "There are no autoscale parameters specified in the request.",
      "detailedMessage": "The request did not contain an Autoscaler structure.",
      "duration": "0s"
    }
  ]
}

```

Cloudera Data Engineering CLI exit codes

When a script that has been executed from the Cloudera Data Engineering (CDE) CLI ends, a number is displayed in the command prompt window. This number is an exit code. If the script ends unexpectedly, this exit code can help you identify the error. These exit codes are applicable to CDE 1.19 or higher.

Exit code	Description
0	Success
1	Unknown Non-Retriable Error
2	Redirection Happened (HTTP 3xx status code received)
3	Bad Request (HTTP Status Code 400)
4	Authorization Error
5	Forbidden (HTTP Status Code 403)
6	Not Found (HTTP Status Code 404)
7	Method Not Allowed (HTTP Status Code 405)
8	Not Acceptable (HTTP Status Code 406)
9	Conflict (HTTP Status Code 409)
10	Gone (HTTP Status Code 410)
11	Length Required (HTTP Status Code 411)
12	Precondition Failed (HTTP Status Code 412)

Exit code	Description
13	Request Entity Too Large (HTTP Status Code 413)
14	URI Too Long (HTTP Status Code 414)
15	Unsupported Media Type (HTTP Status Code 415)
16	Range Not Satisfiable (HTTP Status Code 416)
17	Expectation Failed (HTTP Status Code 417)
18	Unprocessable Entity (HTTP Status Code 422)
19	Failed Dependency (HTTP Status Code 424)
20	Upgrade Required (HTTP Status Code 426)
21	Precondition Required (HTTP Status Code 428)
22	Request Header Fields Too Large (HTTP Status Code 431)
23	Unavailable For Legal Reasons (HTTP Status Code 451)
24	Internal Server Error (HTTP Status Code 500)
25	Not Implemented (HTTP Status Code 501)
26	HTTP Version Not Supported (HTTP Status Code 505)
27	Not Extended (HTTP Status Code 510)
28	Network Authentication Required (HTTP Status Code 511)
70	Unknown Retriable Error
71	Timeout during API call
72	Bad Gateway (HTTP Status Code 502)
73	Service Unavailable (HTTP Status Code 503)
74	Gateway Timeout (HTTP Status Code 504)
75	Request Timeout (HTTP Status Code 408)
76	Too Early (HTTP Status Code 425)
77	Too Many Requests (HTTP Status Code 429)

Cloudera Data Engineering Deep Analysis

You can run an on-demand *deep analysis* on a Cloudera Data Engineering (CDE) job run. Deep analysis analyzes job logs and generates detailed information for a given job run, including memory utilization and stage-level analysis.

Because deep analysis consumes cluster resources by running an internal CDE job, you must run it manually for any job run you want to analyze or troubleshoot.

Enabling deep analysis on a Cloudera Data Engineering job from the CDE web UI

Using the Cloudera Data Engineering (CDE) web UI you can enable an on-demand *deep analysis* on a CDE job run to generate detailed information, including memory utilization and stage-level analysis.

Before you begin



Important: Deep analysis is currently supported only for Spark 2 jobs. For Spark 3 jobs, the Spark Analysis toggle is greyed out.



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.

About this task

Because deep analysis consumes cluster resources by running an internal CDE job, you must enable and run it manually for any job run you want to analyze or troubleshoot.

Procedure

1. From the CDE Overview page, select the CDE service for the job you want to troubleshoot or analyze.
2. In the Virtual Clusters column, click the View Jobs icon for the cluster containing the job.
3. Select the job you want to analyze.
4. Select the Configuration tab and click Edit.
5. Select the Spark Analysis option.
This enables collecting metrics during future runs of the job that you want to investigate.
6. Click Update.
7. Click Actions Run Now to run the job.
Metrics are collected for deep analysis.

What to do next



Important: Do not forget to disable Spark Analysis on the Configuration tab once the deep analysis job completes. It adds unnecessary overhead and can have a negative impact on performance.

Enabling deep analysis on a Cloudera Data Engineering job run using the CLI

Using the Cloudera Data Engineering (CDE) CLI you can enable an on-demand *deep analysis* on a CDE job run to generate detailed information, including memory utilization and stage-level analysis.

About this task



Important: Deep analysis is currently supported only for Spark 2 jobs.

Procedure

- You can enable deep analysis for a job run by using the `cde job run` or `cde spark submit` commands with the `--enable-analysis` flag.

`cde job run`:

```
cde job run --name [***JOB NAME***] --enable-analysis
```

For example:

```
cde job run --name test_job --enable-analysis
```

`cde spark submit`:

```
cde spark submit [***JAR/PY FILE***] --enable-analysis
```

For example:

```
cde spark submit test_job.jar --enable-analysis
```

Metrics are collected for deep analysis.

Related Information

[Running a Spark job using the CLI](#)

[Creating and updating Apache Spark jobs using the CLI](#)

[Submitting a Spark job using the CLI](#)

Running deep analysis on a Cloudera Data Engineering job run

You can run an on-demand *deep analysis* on a Cloudera Data Engineering (CDE) job run to generate detailed information, including memory utilization and stage-level analysis.

Before you begin

- You must enable deep analysis for the job you want to analyze and run the job once to collect data.



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.

About this task

You can only analyze job runs that took place after Spark Analysis was enabled for the job.

Procedure

1. From the CDE Overview page, select the CDE service for the job you want to troubleshoot or analyze.
2. In the Virtual Clusters column, click the View Jobs icon for the cluster containing the job.
3. In the Run History tab, click the Job ID for the job run you want to analyze.
4. Go to the Analysis tab.
5. Click Run Deep Analysis.

Results

After the deep analysis job completes, you can view additional job run information in the Analysis tab for the job run.

What to do next



Important: Do not forget to disable Spark Analysis on the Configuration tab once the deep analysis job completes. It adds unnecessary overhead and can have a negative impact on performance.