Horizontal scaling for the Data Lake (Preview)

Date published: 2024-03-22
Date modified: 2024-12-04

Legal Notice

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided ‘as is’ without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.

Contents

Overview

Certain Data Lake services may require additional instances as the Data Lake grows, or performance for these services can suffer. With an enterprise Data Lake (EDL), several of these services can be scaled horizontally, meaning that you can add additional instances to dedicated host groups for some services.

Horizontal scaling for these services prevents you from having to scale a base host group like the gateway or core host groups. Vertically scaling the gateway or core host groups is not cost-efficient, as they may include more services than what you want to scale.

The enterprise Data Lake (EDL) contains specific host groups for horizontal scaling of the following services, which can all be scaled up individually:

- Solr
- Kafka
- Ranger Authorization Service (RAZ)
- Hive metastore (HMS)
- Storage

Note: Upscaling the Solr host group requires manual steps after the scaling action to rebalance the collection replicas. Contact Cloudera Customer Support for details.

The following hostgroups can also be scaled down:

- Solr
- Ranger Authorization Service (RAZ)
- Hive metastore (HMS)

The storage and Kafka host groups cannot be scaled down once they have been scaled up, as the action might result in data loss.

Horizontal scaling of Data Lake services is a technical preview feature under entitlement and is currently available only through the beta CDP CLI.

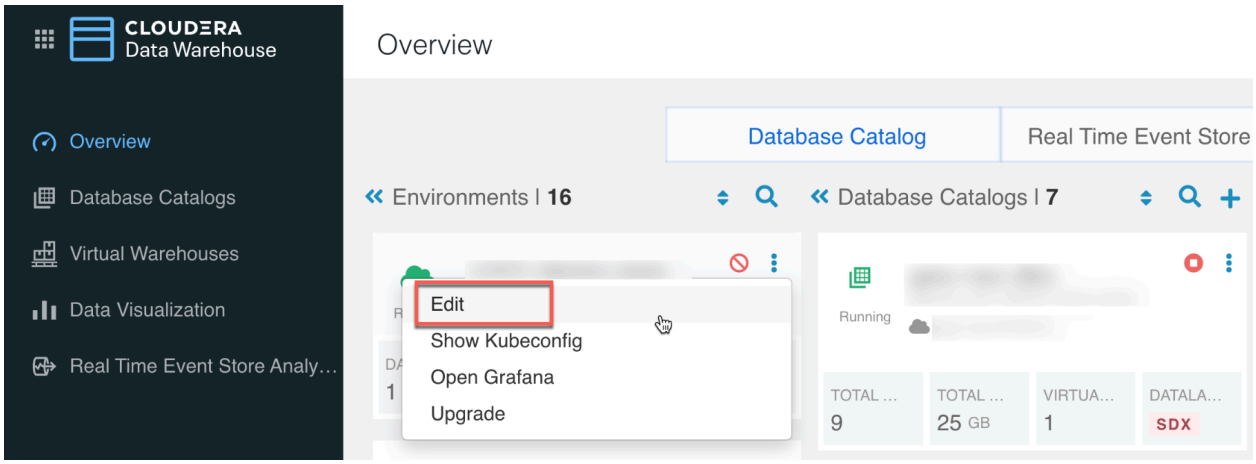
Limitations:

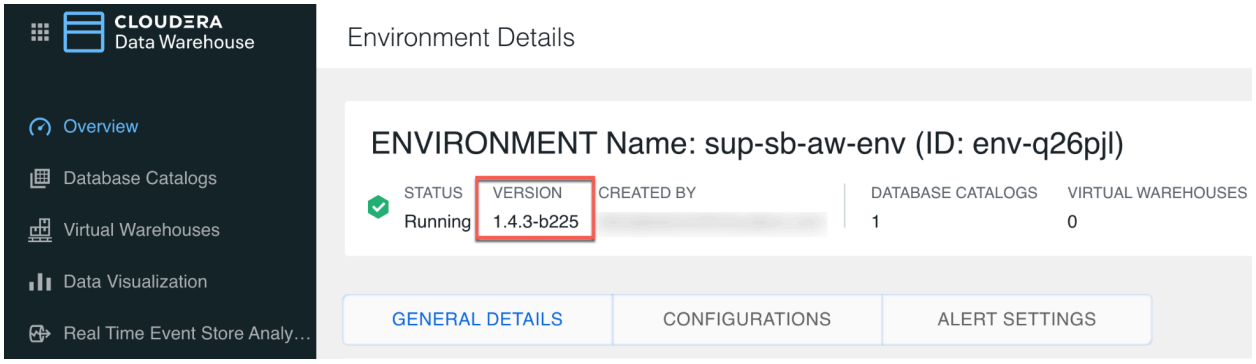
- Currently, Atlas is not supported for horizontal scaling despite having a dedicated host group in the EDL. Support for Atlas will be added in the future.
- The Kafka and Storage host groups cannot be scaled down through this feature.
- As the EDL can have a total of 40 nodes and 10 are used by default, the horizontal scaling host groups combined are limited to 30 nodes.
- Upscaling the Solr host group requires manual steps after the scaling action to rebalance the collection replicas.

Prerequisites for Data Services

Prior to horizontally scaling the Data Lake, ensure that the following are in place:

1. If you are using Cloudera Data Warehouse, you must upgrade to version 1.4.1 or higher before you can resize the Data Lake. Determine the Cloudera Data Warehouse version you are on by clicking edit on the environment:





2. If you are using Cloudera Data Warehouse, stop the virtual warehouses and data catalogs associated with the environment.
3. If you are using Cloudera Data Engineering, do one of the following:
    1. Upgrade to Cloudera Data Engineering 1.15, or
    2. Create a new service.
        1. Take a backup of your jobs following Backing up Cloudera Data Engineering jobs.
        2. Create a new Cloudera Data Engineering service and virtual cluster.
        3. Restore the jobs following the instructions in Restoring Cloudera Data Engineering jobs from backup.
4. If you are using Cloudera AI:
    1. Backup Cloudera AI workbenches (AWS only). If backup is not supported, then proceed to the next step.

2. Suspend Cloudera AI workspaces. If the suspend capability is not available, follow the steps in Refreshing Cloudera AI governance pods after resizing the Data Lake.

Horizontally scaling Data Lake services through the CDP CLI

If you have not already done so, download the beta CDP CLI. You do not need to stop the Data Lake, and both Cloudera Data Hubs and Data Services can continue to run during this operation.

To horizontally scale a service, run the following command in the CLI:

```
cdp datalake scale-horizontally --datalake-name <DATALAKE_NAME>
--instance-group-name <INSTANCE_GROUP_NAME>
--instance-group-desired-count <DESIRED_COUNT>
```

Option	Description
`-datalake-name`	Name or CRN of the Data Lake to be horizontally scaled.
`-instance-group-name`	Name of the service host group that you want to scale up or down. Valid inputs are: ● Solr_scale_out ● Storage_scale_out ● Kafka_scale_out ● Raz_scale_out ● Hms_scale_out Note that the Storage and Kafka groups are not available for downscale.
`-instance-group-desired-count`	Total number of instances that the service host group should contain after the scaling action finishes. For instance, if the host group already contains one instance, and you provide “3” for `-instance-group-desired-count`, the scaling action will add two instances to the host group. If you provide a number that is less than the current number of instances for the host

	group, the action will downscale the host group by removing nodes to the desired count.

You can monitor the progress of the scaling operation from the Data Lake *Event History* tab.

Important: If the horizontal scaling fails, due to issues on the cloud provider side or other issues, use the cdp datalake retry-datalake CLI command once you have fixed the issue that is causing the failure. The **`retry-datalake`** command will resume the horizontal scaling flow.

Important: To use the new instances added after an upscale, you must refresh the Cloudera Data Hub clusters that run these services after the horizontal scaling action completes. You can either stop and start the Cloudera Data Hub, or you can use the `cdp datalake refresh-datahub` CLI command:

`cdp datalake refresh-datahub –datalake` `<datalake_name>` **`–datahub`** `<datahub_name>`

Option	Description
`–datalake`	Name or CRN of the Data Lake.
`–datahub`	Name or CRN of the Cloudera Data Hub to refresh. If none is provided, all Cloudera Data Hub clusters associated with the Data Lake are refreshed.

After the refresh, the Cloudera Data Hub clusters will use the new instances that were added to the Data Lake during the upscale.

Horizontal scaling of Solr instances

After you scale the Solr host group horizontally, you may need to perform some manual steps so that the new instances are available for Solr operations. These steps are required to rebalance the collection replicas. For assistance with scaling Solr, contact Cloudera Customer Support.

Post-requisites for Data Services

After horizontally scaling the Data Lake, complete the following tasks if you use the noted Data Services.

Cloudera AI:

If you did not suspend your Cloudera AI workspaces, refresh the Cloudera AI governance pods after the horizontal scaling is complete.

Cloudera Data Warehouse: Start the Data Catalogs and Virtual Warehouses. For each virtual warehouse, Cloudera recommends that you start, stop, and start again. This will completely refresh the Data Lake details for the virtual warehouse.