

## Auto-scaling

Date published: 2021-04-06

Date modified: 2025-07-17

# CLOUdera

# Legal Notice

© Cloudera Inc. 2025. 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

- Auto-scaling flow deployments.....4**
  - CPU based auto-scaling.....4
  - Configure CPU based auto-scaling.....4
  - Flow metrics based auto-scaling.....5
  - Enable Flow metrics based auto-scaling.....6

## Auto-scaling flow deployments

Flow deployments can be configured to automatically scale up or down the number of NiFi nodes depending on the resource utilization in the cluster.

With Auto-scaling, you get the following benefits:

- You can choose a minimum and maximum number of nodes to ensure that the required resources are available while controlling cost.
- Flow deployments support CPU based auto-scaling as well as flow metrics based auto-scaling.
- When a flow deployment scales down, the affected NiFi nodes are being offloaded before they are removed from the cluster to ensure that no data is lost.
- After new nodes have been added to the cluster, new data is automatically distributed to all cluster nodes including the recently provisioned ones.

### CPU based auto-scaling

When you turn this feature on, Cloudera Data Flow uses the Kubernetes Horizontal Pod Autoscaler (HPA) to increase or decrease the number of NiFi pods in a deployment based on resource utilization.

HPA monitors CPU utilization across all NiFi pods of a flow deployment. When CPU utilization reaches 75%, a new NiFi pod gets created and is added to the NiFi cluster. HPA continues to add additional NiFi pods until the aggregated CPU utilization for a deployment becomes stable around the 75% threshold. Once the new NiFi pod has joined the NiFi cluster, it starts receiving and processing data.



**Note:** Data that has already been received by the NiFi cluster and is being processed is not redistributed to new NiFi pods.

A similar logic is applied to automatically scale down and decrease the number of NiFi pods. HPA monitors CPU utilization and when it falls below 75% for five minutes it starts removing NiFi pods. Before the NiFi pod is terminated, Cloudera Data Flow offloads all existing data from the NiFi pod that has been selected for termination. Offloading data distributes data that is currently being processed by the affected node to the remaining nodes in the cluster. This offloading procedure ensures that NiFi pods are terminated gracefully and avoids data loss.

#### Use cases

CPU based auto-scaling is great for data flows that require CPU intensive transformations like compressing or decompressing data and have to deal with bursts of data every now and then.

### Configure CPU based auto-scaling

#### For UI

Before you begin





When deploying a flow definition, continue to Step 4: Sizing & Scaling and pick your NiFi Node size Steps

1. Activate the Auto Scaling toggle to turn on CPU based auto-scaling.

## Sizing & Scaling

Select the NiFi node size and the number of nodes provisioned for your flow.

### NiFi Node Sizing ?

			
<input checked="" type="radio"/> Extra Small	<input type="radio"/> Small	<input type="radio"/> Medium	<input type="radio"/> Large
2 vCores Per Node 4 GB Per Node	3 vCores Per Node 6 GB Per Node	6 vCores Per Node 12 GB Per Node	12 vCores Per Node 24 GB Per Node

### Number of NiFi Nodes

Auto Scaling ?

☒ Enabled

1
32

☐ Flow Metrics Scaling

Min. Nodes

1

Max. Nodes

3

2. Use the slider or text fields to adjust the minimum and maximum number of nodes. Cloudera Data Flow automatically determines the best number of nodes within the specified boundaries.

### For CLI

To configure auto-scaling using the CDP CLI, execute the `cdp df create-deployment` command with the following configuration:

```
cdp df create-deployment \
  --service-crn <your DFService CRN> \
  --flow-version-crn "<your flow version CRN>" \
  --deployment-name "<your flow deployment name>" \
  --cfm-nifi-version <your desired NiFi version> \
  --auto-start-flow \
  --cluster-size-name <your desired NiFi Node Size> \
  --auto-scaling-enabled \
  --auto-scale-min-nodes [***MINIMUM NUMBER OF NODES***] \
  --auto-scale-max-nodes [***MAXIMUM NUMBER OF NODES***]
```

Replace `[***MINIMUM NUMBER OF NODES***]` and `[***MAXIMUM NUMBER OF NODES***]` with the minimum and maximum number of nodes you want to allow for this deployment, respectively.

## Flow metrics based auto-scaling

Flow metrics based auto-scaling predicts future load on a connection to trigger scaling decisions.

In contrast to CPU based auto-scaling where scaling decisions are made based on infrastructure utilization, flow metrics based auto-scaling works by predicting backpressure on a source connection queue. Source connections are connections attached to processors where data is first introduced into the flow. Backpressure occurs when such

a connection becomes 100% full. Source connections are automatically detected through static analysis of the flow definition. Queue percentage of each source connection is tracked and a linear regression is performed on the histogram for each to predict the fullness of the queue in 20 minutes. Scaling up occurs when that prediction is greater than 80% for 5 minutes.

Flow metrics based auto-scaling can be used to detect the need to scale based on flow performance metrics. In cases where queues fill up without CPU utilization going over the threshold, flow metrics based auto-scaling still triggers a scaling operation.

## Enable Flow metrics based auto-scaling

Flow metrics based scaling is an optional setting which can be enabled in addition to CPU based auto-scaling.

### For UI

Before you begin

When deploying a flow definition, continue to Step 4: Sizing & Scaling and pick your NiFi Node size

Steps

1. Activate the Auto Scaling toggle to turn on CPU based auto-scaling.

2. Use the slider or text fields to adjust the minimum and maximum number of nodes. Cloudera Data Flow automatically determines the best number of nodes within the specified boundaries.
3. Select Flow Metrics Scaling to enable flow metrics based auto-scaling for your deployment.

### For CLI

To configure auto-scaling using the CDP CLI, execute the `cdp df create-deployment` command with the following configuration:

```
cdp df create-deployment \
  --service-crn <your DFService CRN> \
  --flow-version-crn "<your flow version CRN>" \
  --deployment-name "<your flow deployment name>" \
  --cfm-nifi-version <your desired NiFi version> \
  --auto-start-flow \
  --cluster-size-name <your desired NiFi Node Size> \
  --auto-scaling-enabled \
```

```
--auto-scale-min-nodes 1 \  
--auto-scale-max-nodes 3  
--flow-metrics-scaling-enabled
```