

Creating flow deployments

Date published: 2021-04-06

Date modified: 2024-06-03

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.

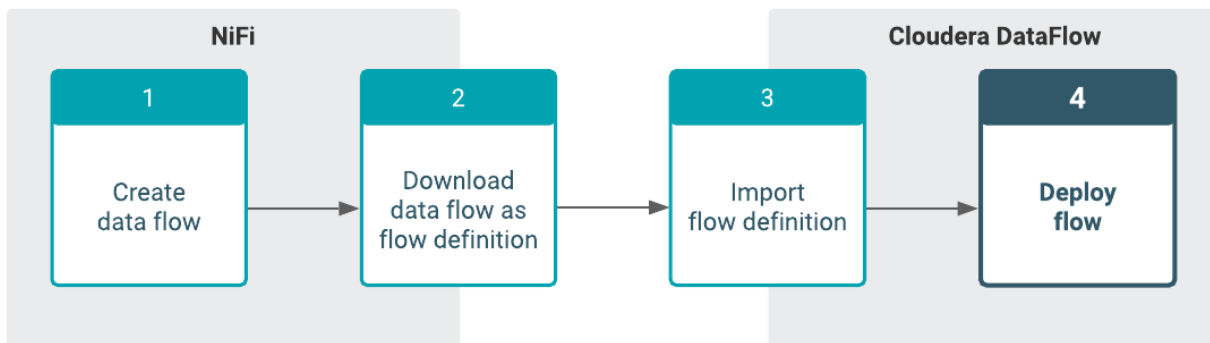
Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES 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

Deploying a flow definition using the wizard.....	4
Select the flow definition version you want to deploy from the catalog.....	4
Launch the deployment wizard.....	4
Name your flow deployment and assign it to a project.....	5
Configure NiFi.....	6
Provide parameter values.....	7
Configure sizing and scaling.....	7
Set Key performance indicators.....	7
Verify your settings and initiate deployment.....	8
 Deploying a flow definition using the CLI.....	 8

Deploying a flow definition using the wizard

Deploy a flow definition to run Apache NiFi flows as flow deployments in Cloudera DataFlow. To do this, launch the Deployment wizard and specify your environment, parameters, sizing, and KPIs.



Before you begin

- You have an enabled and healthy Cloudera DataFlow environment.
- You have been assigned the DFCatalogAdmin or DFCatalogViewer role granting you access to the Catalog.
- The flow definition you want to deploy has been added to the Catalog by someone with DFCatalogAdmin role.
- You have been assigned the DFFlowAdmin role for the environment to which you want to deploy the flow definition.
- You have been assigned DFProjectMember role for the Project where you want to deploy the flow definition.
- If you are deploying custom processors or controller services, you may need to meet additional prerequisites.

Select the flow definition version you want to deploy from the catalog

The Catalog is where you manage the flow definition lifecycle, from initial import, to versioning, to deploying a flow definition.

Procedure

1. In Cloudera DataFlow, select Catalog from the left navigation pane.
Flow definitions available for you to deploy are displayed, one definition per row.
2. Select a row to display the flow definition details and available versions.
The flow details pane opens on the right.

Launch the deployment wizard

After selecting a flow definition version from the catalog, you need to select an environment, provide a deployment name and assign it to a project using the deployment wizard.

Procedure

1. Click Deploy to launch the Deployment wizard.

The screenshot shows the Cloudera DataFlow Dashboard. On the left is a sidebar with navigation links: Dashboard, Catalog, ReadyFlow Gallery, and Environments. The main area is titled 'Dashboard' and contains a 'Filter By' section with dropdowns for STATUS (All - 10), ENVIRONMENTS (All - 3), DEPLOYMENTS (All - 8), and PROCESSOR TYPES (All - 42). There is a 'Reset' button and a 'METRICS WINDOW' dropdown set to '30 Minutes'. Below the filters is a table of flows:

Status	Name ↑	Current Received	Current Sent	Data Throughput (Received/Sent)
Concerning Health	Aptar Kafka Kudu dataflow-demo-new	0 B/s	0 B/s	0
Concerning Health	Kafka filter Kafka dataflow-demo-new	0 B/s	0 B/s	0
Good Health	Kafka to COD dataflow-demo-new	0 B/s	0 B/s	0
Good Health	Kafka to COD docs dataflow-demo-new	0 B/s	0 B/s	0
Good Health	Kafka to Kafka dataflow-demo-new	0 B/s	0 B/s	0
Good Health	Kafka to S3 dataflow-demo-new	0 B/s	0 B/s	0

Each row includes a status icon (yellow triangle for 'Concerning Health', green checkmark for 'Good Health'), a name, and throughput metrics. The 'Data Throughput' column shows a bar chart and a numerical value (0). A '30 Minutes' label and a 'Current' label are also present for each row. A 'REFRESHED: 24 seconds ago' indicator is at the top right of the table area.

2. Select the environment where you want to deploy the flow.



Note:

Only environments which have been enabled for Cloudera DataFlow, are in a healthy state, and to which you have access show up in the dropdown. Once you have selected the target environment, you cannot change it later in the flow deployment process without starting over.

3. Click Deploy.

Name your flow deployment and assign it to a project

After selecting the flow version and an environment, the deployment wizard takes you to the Overview page. Here you need to provide a name for your flow deployment and assign it to a project. At this point you can also import a previously exported deployment configuration, auto-filling configuration values and thus speeding up deployment.

Procedure

1. Give your flow a unique Deployment Name.

You can use this name to distinguish between different versions of a flow definition, flow definitions deployed to different environments, and similar.



Note:

Flow Deployment names need to be unique within an Environment. The Deployment wizard indicates whether a name is valid by displaying a green check below the Deployment Name text box.

2. Select a Target Project for your flow deployment from the list of Projects available to you.

- If you do not want to assign the deployment to any of the available Projects, select Unassigned. Unassigned deployments are accessible to every user with DFFlowUser role in the environment.
- This field is automatically populated if you import a configuration and the Project referenced there exists in your environment, and you have access to it.

3. If you have previously exported a deployment configuration that closely aligns with the one you are about to deploy, you can import it under Import Configuration to auto-fill as much of the wizard as possible.
You can later manually modify auto-filled configuration values during deployment.
4. Click Next.

Configure NiFi

After selecting the target environment, project, and naming your flow, you need to set Apache NiFi version, possible inbound connections, and custom processors. Depending on the flow definition, you may also need to provide values for a number of configuration parameters. Finally, you need to set the capacity of the NiFi cluster servicing your deployment.

Procedure

1. Pick a NiFi Runtime Version for your flow deployment.

Select if you want to use Apache NiFi 1.x or 2.x with your deployment.



Important: NiFi 2.x is currently provided as a technical preview feature, do not use it for deployments in production environments.

Cloudera recommends that you always use the latest available version within the 1.x and 2.x lines, if possible.

2. Specify whether you want the flow deployment to auto-start once deployed.
3. Specify whether you want to use Inbound Connections that allow your flow deployment receiving data from an external data source.

If yes, specify the endpoint host name and listening port(s) where your flow deployment listens to incoming data.

See *Creating an inbound connection endpoint* for complete information on endpoint configuration options.

4. Specify whether you want to use NiFi Archives (NARs) to deploy custom NiFi processors or controller services.

If yes, specify the CDP Workload Username, password, and cloud storage location you used when preparing to deploy custom processors.



Tip: If you want to provide a machine user as CDP Workload Username during flow deployment, make sure to note the full workload user name including the `srv_` prefix.

Make sure that you click the Apply button specific to Custom NAR Configuration before proceeding.

5. If you selected to run your flow with NiFi 2.x [Technical Preview], specify whether you want to use custom Python processors with your flow deployment.

If yes, specify the CDP Workload Username, password, and cloud storage location where the processors are stored.



Tip: Create a dedicated directory in your object store where you keep all your Python processors. Create one Python script per processor and store it in this directory.



Tip: If you want to provide a machine user as CDP Workload Username during flow deployment, make sure to note the full workload user name including the `srv_` prefix.

Make sure that you click the Apply button specific to Custom Python Processors before proceeding.

6. Click Next.

Related Information

[Inbound connections](#)

Provide parameter values

Depending on the flow you deploy, you may need to specify parameter values like connection strings, usernames and similar, and upload files like truststores, JARs, and similar.

Procedure

1. Provide values to parameters required for your flow deployment.

You have to provide values for all parameters. You can filter for the still empty fields by selecting the No value checkbox.



Tip: If you are deploying a ReadyFlow, you can learn about required parameters and instructions on how to obtain parameter values by checking *Prerequisites* and *Required parameters* in the documentation of the respective ReadyFlow.

2. When you finished setting configuration parameters, click Next.

Configure sizing and scaling

Set the size and number of Apache NiFi nodes, auto-scaling, and the type of storage to be used.

Procedure

1. Specify NiFi node size.

Select one of the following options:

- Extra Small: 2 vCores per Node, 4 GB per Node
- Small: 3 vCores per Node, 6 GB per Node
- Medium: 6 vCores per Node, 12 GB per Node
- Large: 12 vCores per Node, 24 GB per Node

2. Set the number of NiFi nodes and auto-scaling.

- You can set whether you want to automatically scale your cluster according to flow deployment capacity requirements. When you enable auto-scaling, the minimum number of NiFi nodes are used for initial size and the workload scales up or down depending on resource demands.
- You can set the number of nodes between 1 and 32.
- You can set whether you want to enable Flow Metrics Scaling.

3. Select storage type.

Select whether you want your deployment to use storage optimized for cost or for performance.

- Standard: 512 GB Content Repo Size, 512 GB Provenance Repo Size, 256 GB Flow File Repo Size, 2300 IOPS, 150 MB/s Max Throughput
- Performance: 1024 GB Content Repo Size, 1024 GB Provenance Repo Size, 256 GB Flow File Repo Size, 5000 IOPS, 200 MB/s Max Throughput

4. Click Next.

Related Information

[Auto-scaling](#)

Set Key performance indicators

Optionally add key performance indicators to help you track the performance of your flow deployment then review your settings and launch the deployment process.

Procedure

1. From KPIs, you may choose to identify key performance indicators (KPIs), the metrics to track those KPIs, and when and how to receive alerts about the KPI metrics tracking.

**Tip:**

You can reorder the KPIs by dragging them up or down the list. The order you configure here will be reflected in the detailed monitoring view of the resulting deployment.

See *Working with KPIs* for complete information about the KPIs available to you and how to monitor them.

2. Click Next.

Related Information

[Working with KPIs](#)

Verify your settings and initiate deployment

Review deployment settings, make any necessary changes, and start deployment.

Procedure

1. Review a summary of the information provided and make any necessary edits by clicking Previous.
2. When you are finished, complete your flow deployment by clicking Deploy.

**Tip:**

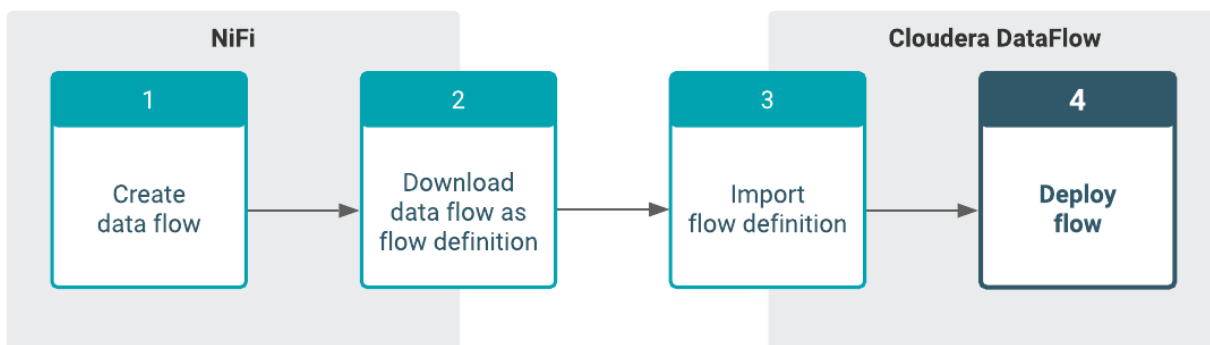
Click View CLI Command to see the equivalent Cloudera CLI syntax in a help pane.

Results

After you click Deploy, you are redirected to the **Alerts** tab in the **Flow Details** where you can track how the deployment progresses.

Deploying a flow definition using the CLI

Deploy a flow definition to run NiFi flows as flow deployments in Cloudera DataFlow. To do this, launch the Cloudera CLI and specify your environment, parameters, sizing, and KPIs.



Before you begin

- You have installed Cloudera CLI.
- Run `cdp df list-services` to get the service-crn value.
- Run `cdp df list-flows` to get the flow-version-crn value.

Procedure

To deploy a flow, enter:

```
cdp df create-deployment
  --service-crn [***service-crn-value***]
  --flow-version-crn [***flow-version-crn-value***]
  --deployment-name [***flow-deployment-name***]
  [--cluster-size-name [EXTRA_SMALL|SMALL|MEDIUM|LARGE]]
  [--static-node-count [***value***]]
  [--auto-scaling-enabled | --no-auto-scaling-enabled]
  [--auto-scale-min-nodes [***number-min-nodes***]]
  [--auto-scale-max-nodes [***number-max-nodes***]]
  [--cfm-nifi-version [***nifi-version***]]
  [--auto-start-flow | --no-auto-start-flow]
  [--parameter-groups [***json-file-location***]]
  [--kpis [***json-file-location***]]
  [---node-storage-profile-name [STANDARD_AWS|STANDARD_AZURE|PERFORMANCE_AWS|PERFORMANCE_AZURE]]
```

Where:

--service-crn

Specifies the service-crn value you obtained when completing the prerequisites.

--flow-version-crn

Specifies the flow-version-crn value you obtained when completing the prerequisites.

--deployment-name

Specifies a unique name for your flow deployment.

[--cluster-size-name]

Specifies the cluster size. Valid values are:

- EXTRA_SMALL
- SMALL
- MEDIUM
- LARGE

The default value is EXTRA_SMALL.

[--static-node-count]

Specifies the number of NiFi nodes when autoscaling is not enabled. You can select between 1 and 32 nodes. The default value is 1.

[--auto-scaling-enabled | --no-auto-scaling-enabled]

Specifies whether you want to enable autoscaling. The default is to disable autoscaling.

[--auto-scale-min-nodes]

Specifies the minimum nodes when you have autoscaling enabled. If you have autoscaling enabled, this parameter is required.

[--auto-scale-max-nodes]

Specifies the maximum nodes when autoscaling is enabled. If you have autoscaling enabled, this parameter is required.

[--cfm-nifi-version]

Specifies the NiFi runtime version. The default is the latest version.

[--auto-start-flow | --no-auto-start-flow]

Specifies whether you want to automatically start your flow once it has been deployed. The default is to enable the automatic start.

[--parameter-groups]

Specifies the location of the parameter group JSON file, if you are using one for this flow deployment.

[--kpis]

Specifies the location of the KPIs JSON file, if you are providing KPIs for this flow.

--node-storage-profile-name

Specifies the storage profile. Valid values are:

- STANDARD_AWS
- STANDARD_AZURE
- PERFORMANCE_AWS
- PERFORMANCE_AZURE

The default values are STANDARD_AWS and STANDARD_AZURE, depending on the cloud provider.

Example parameter group file:



Note: The JSON file you develop for parameter group will be different depending on your flow objectives and requirements. This is an example of the parameter group file format.

```
[
  {
    "name": "kafka-filter-to-kafka",
    "parameters": [
      {
        "name": "CDP Workload User",
        "assetReferences": [],
        "value": "srv_nifi-machine-ingest"
      },
      {
        "name": "CDP Workload User Password",
        "assetReferences": [],
        "value": "<<CDP_MISSING_SENSITIVE_VALUE>>"
      },
      {
        "name": "CSV Delimiter",
        "assetReferences": [],
        "value": ","
      },
      {
        "name": "Data Input Format",
        "assetReferences": [],
        "value": "CSV"
      },
      {
        "name": "Data Output Format",
        "assetReferences": [],
        "value": "JSON"
      },
      {
        "name": "Filter Rule",
        "assetReferences": [],
        "value": "SELECT * FROM FLOWFILE"
      },
      {
        "name": "Kafka Broker Endpoint",
        "assetReferences": [],
        "value": "streams-messaging-broker0.pm-sandb.a465-9q4k.cloudera.sit
e:9093"
```

```

    },
    {
      "name": "Kafka Consumer Group ID",
      "assetReferences": [],
      "value": "cdf"
    },
    {
      "name": "Kafka Destination Topic",
      "assetReferences": [],
      "value": "MachineDataJSON"
    },
    {
      "name": "Kafka Producer ID",
      "assetReferences": [],
      "value": "cdf"
    },
    {
      "name": "Kafka Source Topic",
      "assetReferences": [],
      "value": "MachineDataCSV"
    },
    {
      "name": "Schema Name",
      "assetReferences": [],
      "value": "SensorReading"
    },
    {
      "name": "Schema Registry Hostname",
      "assetReferences": [],
      "value": "streams-messaging-master0.pm-sandb.a465-9q4k.cloudera.site"
    }
  ]
}
]

```

Example KPI file:



Note: The JSON file you develop for KPIs will be different depending on your flow objectives and requirements. This is an example of the KPI file format.

```

[
  {
    "metricId": "rateBytesReceived",
    "alert": {
      "thresholdLessThan": {
        "unitId": "kilobytesPerSecond",
        "value": 150
      },
      "frequencyTolerance": {
        "unit": {
          "id": "MINUTES"
        },
        "value": 1
      }
    }
  },
  {
    "metricId": "processorAmountBytesSent",
    "alert": {},
    "componentId": "a7f7df1c-a32d-3c25-9b09-e1d1036dcc04;a33a1b48-005b-32dd-bb88-b63230bb8525"
  }
]

```

```
]
```

Results

Successfully deploying a flow results in output similar to:

```
{  
  "crn": "deployment-crn"  
}
```