Cloudera DataFlow

Developing Flow Definitions Using NiFi

Date published: 2021-04-06
Date modified: 2024-01-09

https://docs.cloudera.com/

Legal Notice

Contents

Best Practices for Developing Flow Definitions

Before you can deploy a flow definition in Cloudera DataFlow (CDF), you need to develop your data flow logic in a development environment using Apache NiFi. To make sure that your NiFi data flow can be deployed in CDF, follow the best practices outlined in this section.

What is a flow definition?

A flow definition represents data flow logic which was developed in Apache NiFi and exported by using the Download Flow Definition action on a NiFi process group or the root canvas. Flow definitions typically leverage parameterization to make them portable between different environments such as development or production NiFi environments.

To run one of your existing NiFi data flows in CDF you have to export it as a flow definition and upload it to the CDF catalog.

What is flow isolation?

CDF is the first cloud service allowing NiFi users to easily isolate data flows from each other and guarantee a set of resources to each one without requiring administrators to create additional NiFi clusters.

Flow isolation describes the ability to treat NiFi process groups which typically run on a shared cluster on shared resources as independent, deployable artifacts which can be exported as flow definitions from NiFi.

Flow isolation is useful when

- You want to guarantee a set of resources for a specific data flow.
- You want to isolate failure domains.

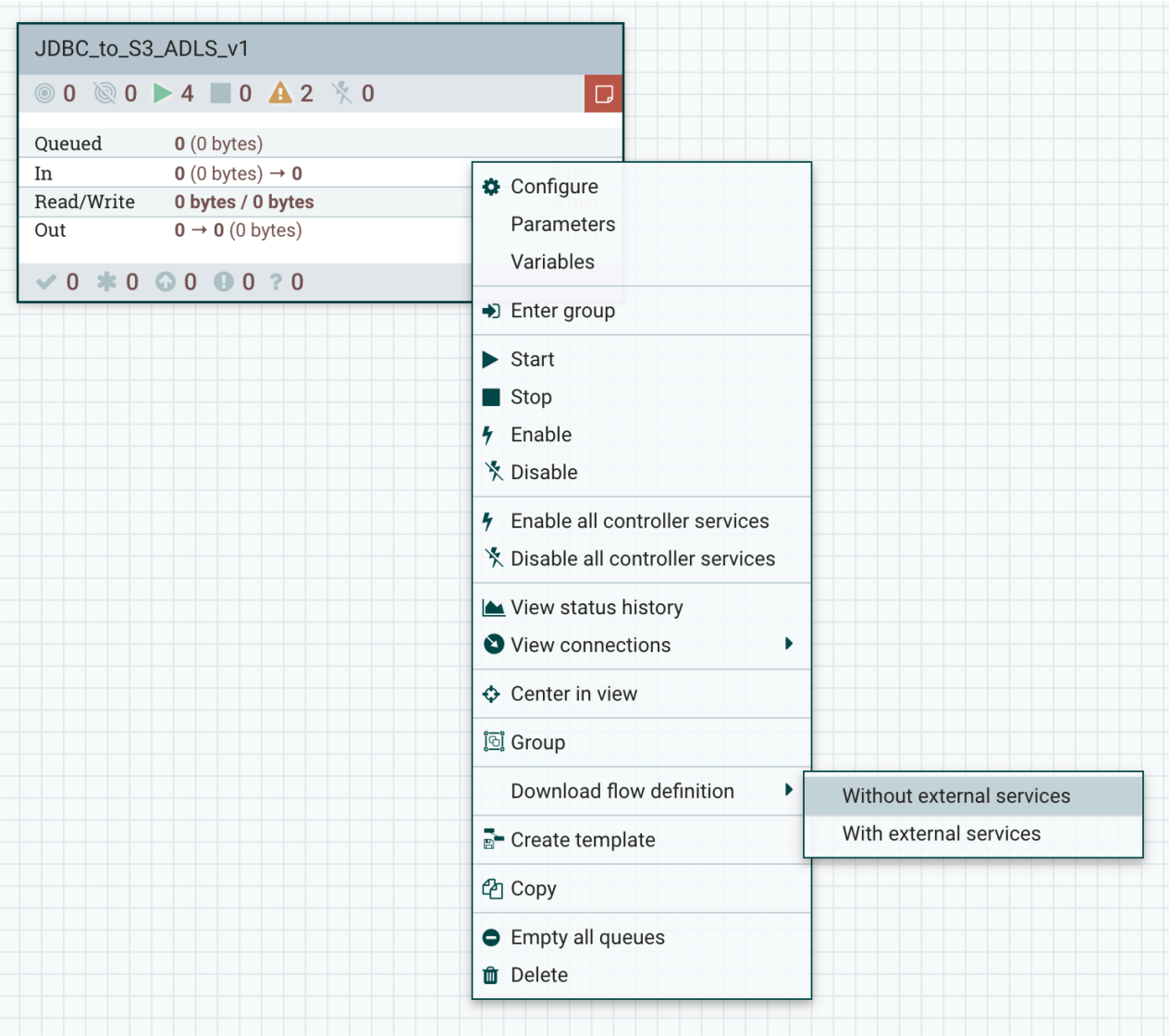
Considerations when developing flow definitions

As you are building your flow definition in your NiFi development environment, you should build the flow definition with ease of CDF export and isolation in mind.

Controller Services

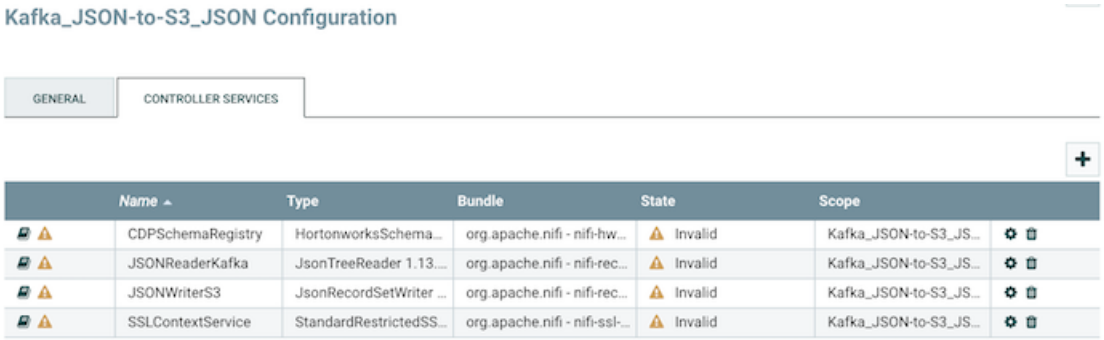
In traditional NiFi data flows, Controller Services are considered shared services and often used by multiple data flows, or by multiple Processors in different Process Groups to avoid redundancies and promote reusability. In CDF, you should plan for the possibility that you will run your process group in isolation and consider how you want to handle shared Controller Services.

When you download a flow definition from NiFi, you can choose whether you want to include external controller services (Requires NiFi 1.16 / 1.17.0 or newer). A controller service is considered external when it is referenced by components in the selected process group but is defined outside the process group scope (e.g. controller services in a parent group). Select “With external services” if you want to include referenced controller services in the resulting flow definition and select “without external services” if you only want to include controller services which are defined in the selected process group.



Note:

If use an older version of NiFi which does not support including external services in flow definitions, review the Controller Services defined outside your Process Group and if needed recreate them in the Process Group you are planning to download.



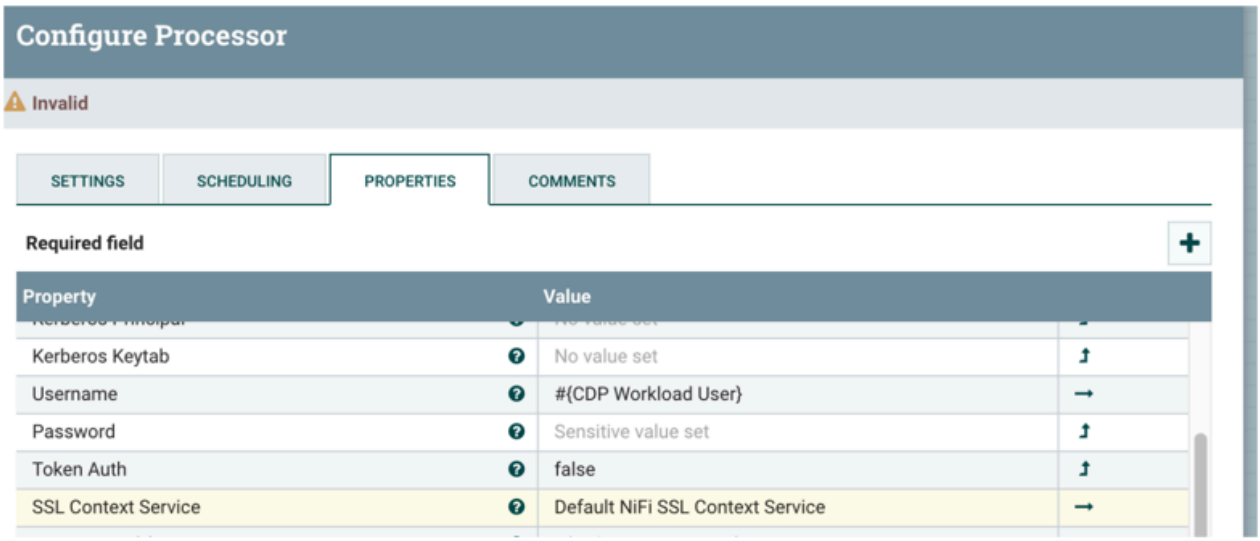
Tip:

An easy way to check whether a controller service is considered external is to compare its scope with the process group name you are planning to download. If they do not match, it is considered an external Controller Service.

Default NiFi SSL Context Service

For data flows that interact with other CDP Public Cloud experiences like Streams Messaging or Operational Database Data Hub clusters, you can reference an external NiFi controller service called Default NiFi SSL Context Service in your NiFi flow to automatically obtain the correct truststore configuration for your target CDP environment.

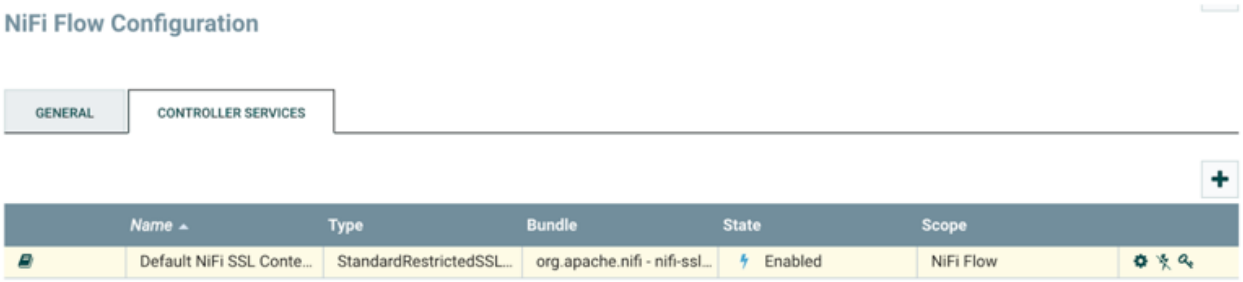
Figure 1: Using the Default NiFi SSL Context Service in a processor configuration



A Default NiFi SSL Context Service is already set up for you and can be used when you are developing NiFi flows on Flow Management clusters in Data Hub. Any flow that uses the Default NiFi SSL Context Service in Data Hub can be exported and deployed through Cloudera DataFlow.

If you are developing your NiFi flows outside of Data Hub, you can create a controller service called Default NiFi SSL Context Service yourself and reference it in any processor that requires an SSL Context Service configuration. You have to define the Default NiFi SSL Context Service as an external Controller Service outside of the process group that you plan to export and run in Cloudera DataFlow.

Figure 2: If you are creating the Default NiFi SSL Context Service yourself, make sure that the Scope it is created in is a parent to the process group that you plan to export



Review the following information if you developed a flow in a CDP Flow Management Data Hub cluster and the flow uses the SSLContextService that was defined outside the Process Group:

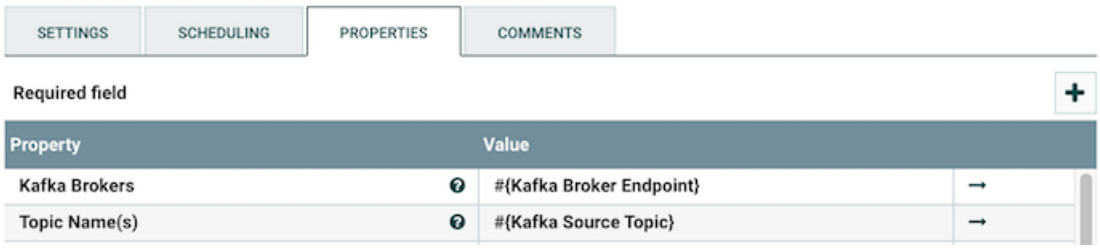
- If the name of the SSLContextService in the flow definition is Default NiFi SSL Context Service, then at the time of deployment, DataFlow automatically creates a new SSLContextService in the Root Process Group with the required information of the target environment.
- Specifically, DataFlow imports the Truststore certificates, creates the Truststore, puts the information on a shared space that all NiFi nodes can access, and updates the Truststore filename property with the correct file path.
- When the flow definition is deployed, the flow references the Default NiFi SSL Context Service in the Root Process Group and the deployment is successful.

If your flow interacts with non-CDP services that require a TLS/SSL connection, then you must do the following:

- Define a new SSLContext Service in the Process Group you are planning to export.
- Parameterize the Truststore Filename property so that DataFlow allows you to upload a custom Truststore when you deploy the flow definition using the Flow Deployment Wizard.

Parameterize Processor and Controller Service configurations

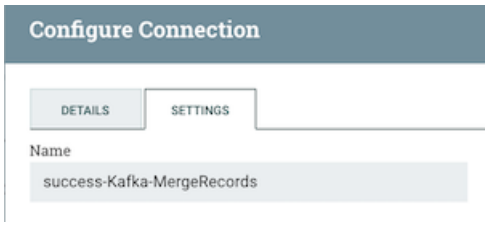
Ensure that your Process Group is portable by parameterizing your processor and controller services configurations. This allows you to deploy a flow in different environments without having to update Processor configuration details.



Customize your Processor and Connection names

To ensure that you are able to distinguish Processors and Connections when defining KPIs in your CDF flow deployment ensure that you specify a custom name for them when developing your data flow.

For example:



Review Reporting Tasks

Reporting tasks are not exported to CDF when you are exporting your flow definition. When you are designing your flow definition, you should be aware of your monitoring and report needs and plan to address these needs with KPIs and Alerts specified as part of your flow deployment.

Using the CDPEnvironment parameter to get Hadoop configuration resources

DataFlow makes it easy to use HDFS processors to read/write data from/to S3. Using HDFS processors to access data in S3 allows you to leverage CDP's IDBroker authentication method and doesn't require you to specify any S3 secrets in the processor configuration. HDFS processors however require Hadoop configuration files - specifically the core-site.xml file. CDF offers a special parameter called CDPEnvironment for you to use whenever you are working with processors that require Hadoop configuration files. Simply use the parameter #{CDPEnvironment} for the Hadoop configuration file property and CDF will automatically obtain the required files during the flow deployment process.

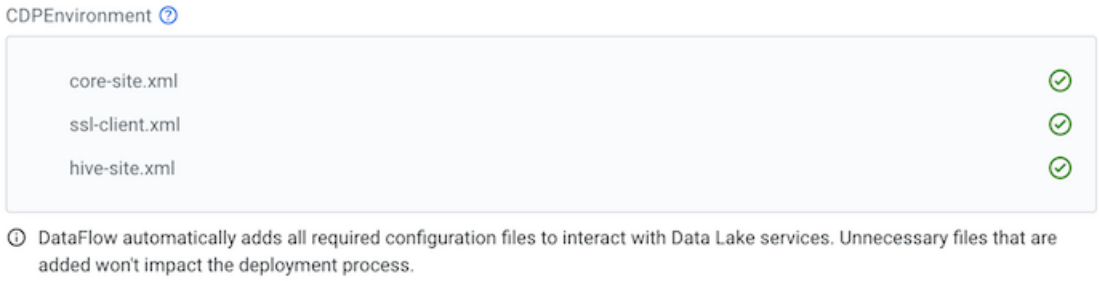
Note:

The value of the #{CDPEnvironment} parameter is not automatically provided when using it in your NiFi development environment. While developing your flow, make sure to add the CDPEnvironment parameter to your parameter context with the correct value for your environment.

For example, in DataHub set the value for this parameter to /etc/hadoop/conf/core-site.xml.

Providing the correct value for your environment allows you to successfully test the flow.

Using CDPEnvironment for the Hadoop Configuration Resources property.



The DataFlow Deployment Wizard detects usage of the CDPEnvironment parameter and automatically obtains the required Hadoop configuration files.

Resource consumption

When deciding how to export your process groups for deployment on DFX, review the data flow resource consumption to help you decide how to isolate your data flows within CDF.

Once you have developed your data flow in a NiFi development environment, you start to export your flow definition

- Root canvas
- Whole process group level
- Part of the process group

Inter process group communication

If you have process groups that exchange data between each other, you should treat them as one flow definition and therefore download their parent process group as a flow definition.

Configuring object store access

When you are developing flows for cloud provider object stores, `CDPObjectStore` are the preferred processors. Available processors are:

- `ListCDPObjectStore`
- `FetchCDPObjectStore`
- `PutCDPObjectStore`
- `DeleteCDPObjectStore`

`CDPObjectStore` processors are equipped with the latest best practices for accessing cloud provider object stores.

Configuring inbound connection support

For more information, see *Adding Inbound Connection support to a NiFi flow*.

Related Information

Adding Inbound Connection support to a NiFi flow

Creating parameter context for flow definitions

About this task

As you are building your data flow in your NiFi development environment, follow these steps to create a parameter context, add parameters to it, and assign the new context to your process group:

Procedure

1. Select Parameter Contexts from the Global menu of the NiFi UI.
2. In the NiFi Parameter Contexts dialog, click the (+) button to create a new parameter context.
3. Add a name for your parameter context.
4. Select the Parameters tab to add the parameters you need for configuring your data flow components.
5. Click Apply to save the parameter context.
6. To assign the parameter context to your process group, click Configure from the process group context menu and select the parameter context in the Process Group Parameter Context drop-down menu.

   Note: To add parameters to an existing parameter context, open the Parameter Contexts dialog from the Global menu and click the (Pencil) button in the row of the chosen parameter context. You can also add new parameters to the parameter context when you are configuring the components in your data flow. For more information about parameters and parameter contexts, see the Apache NiFi User Guide.

Creating controller service for flow definitions

About this task

As you are building your data flow in your NiFi development environment, follow these steps to create a controller service:

Procedure

1. Select your process group and click Configure from either the Operate Palette or the process group context menu of the NiFi UI.

   The Process Group Configuration dialog is displayed.
2. Select the Controller Services tab.
3. Click the (+) button to add a new controller service.

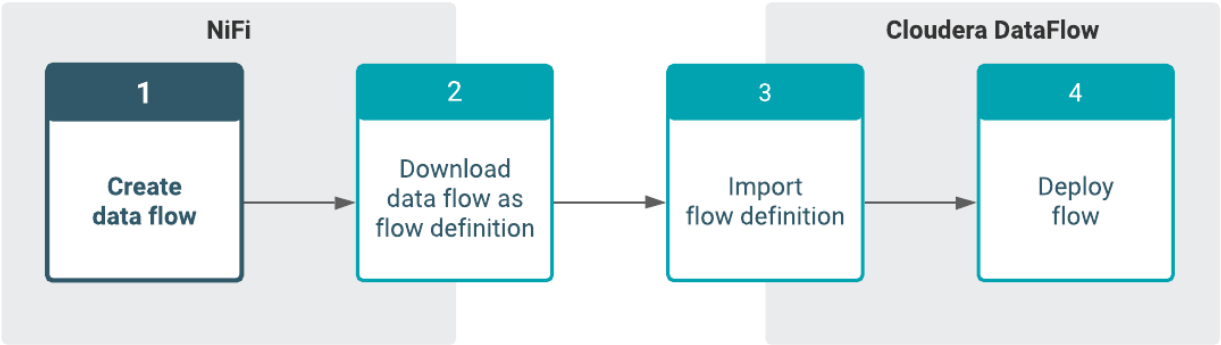
   The Add Controller Service dialog is displayed.
4. Select the required controller service and click Add.
5. Click the (Configure) icon in the right column and set the necessary options.
6. Click Apply to save the changes.

7. Click the (Flash) button to enable the controller service.

Important: You must define the controller services for your data flow in the configuration of the process group where they will be used. This ensures that the controller services are included in the process group flow definition.

Creating a flow definition in NiFi

Before you can run a data flow in Cloudera DataFlow, you need to (1) create the flow in Apache NiFi, (2) download the NiFi flow as a flow definition, (3) import it to Cloudera DataFlow and finally, (4) deploy the flow. The flow definition acts as a configuration logic for your flow deployments. It enables you to deploy your data flow without the need to maintain cluster infrastructure. Also, you can deploy the same flow to multiple environments in Cloudera DataFlow.



Before you begin

When you want to develop a NiFi flow that you intend to use in Cloudera DataFlow, review and adjust your traditional NiFi flow development process to make sure that you can create portable data flows that will work in the Cloudera DataFlow environment. Before you get started with flow development, it is useful to understand where you need special attention and what adjustments you have to make in your development workflow.

You can create and download flow definitions starting with version 1.11 of Apache NiFi. Cloudera provides the following Apache NiFi based products:

- Cloudera DataFlow for Data Hub
- CFM 1.1.0 and higher
- HDF 3.5

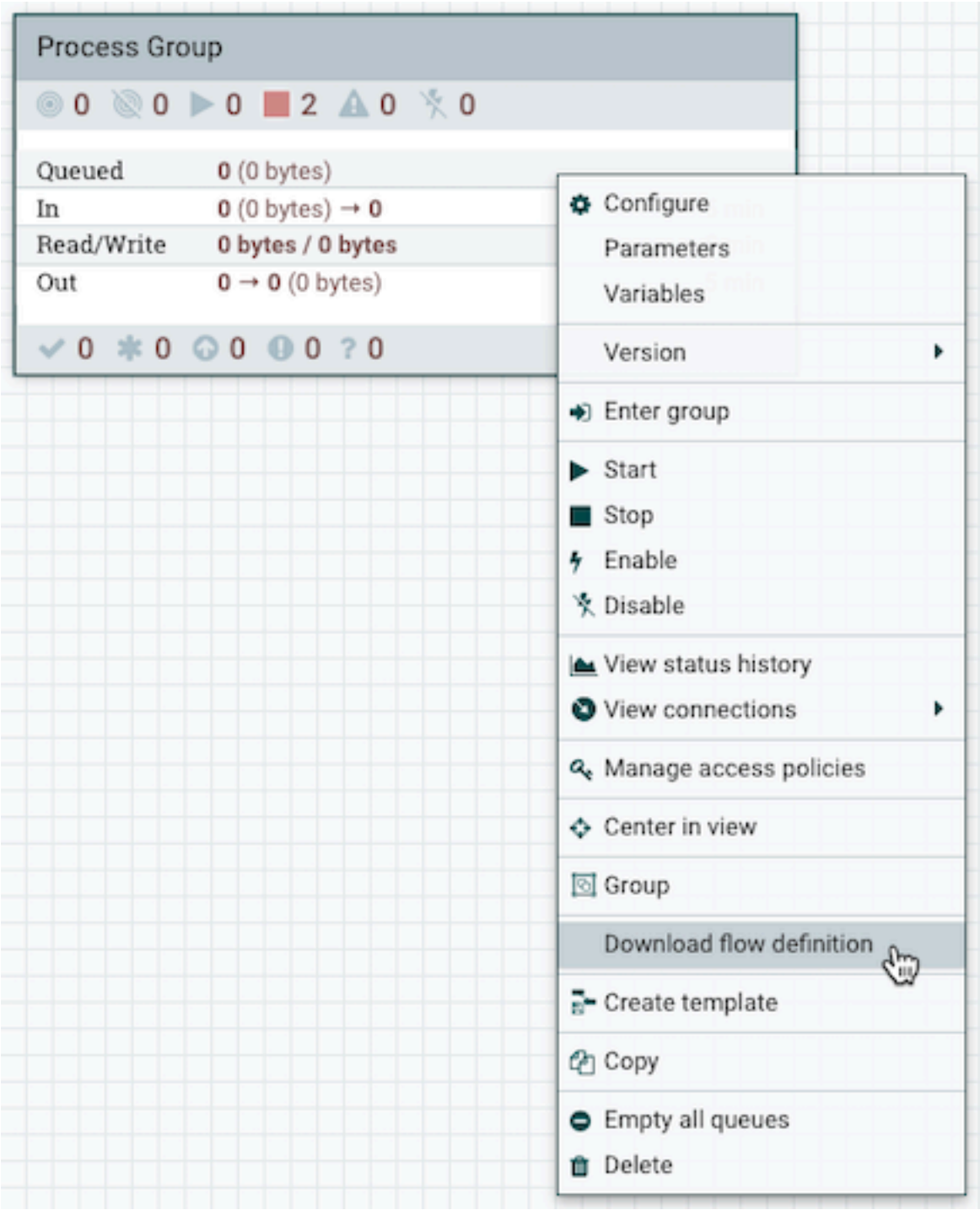
Cloudera recommends that you develop your flow definitions using CDP Data Hub Flow Management clusters. For more information on how to set up a managed and secured Flow Management cluster in CDP Public Cloud, see *Setting up your Flow Management cluster*.

For more information on planning and preparing your NiFi flows for Cloudera DataFlow, see *Best Practices for Developing Flow Definitions*.

Procedure

1. Create a process group that will contain your NiFi flow.
   a) Drag and drop the process group icon onto the canvas.
   b) Add a name for the process group.

   Once you have the new process group available on the canvas, you can interact with it by right-clicking it and selecting an option from the context menu. The available options vary depending on the privileges assigned to you.



2. Enter the process group by double-clicking it.

   Alternatively, you can select the Enter group option from the context menu.

3. Add the appropriate flow components to the NiFi canvas.

   You can add processors and other components to build your data flow. To add a processor to your flow, drag the processor icon to the canvas and select the name of the processor from the list.

4. Configure the components in your data flow.

   Make sure that you externalize the component properties where values change depending on the environment in which the data flow is running. For more information on parameterizing your processor configurations, see *Best Practices for Flow Definition Development*.

   a) Create a parameter context for your data flow and add parameters to it.

      Using parameters for certain properties (for example, connection information, truststores, or drivers) makes the flow portable. You can download your data flow from NiFi and then import it to Cloudera DataFlow as a flow definition. When you deploy the flow definition, you can specify values or upload files for these parameters in the deployment wizard to adjust them to your needs.

      For instructions on how to set up your parameters, see *Best Practices for Flow Definition Development*.

   b) Create controller services for your flow.

      If you want to add controller services to your flow that you will later use in Cloudera Dataflow, you must define the controller services in the configuration of the process group that you will download as a flow definition. In this case the services will be available to all processors in that process group and will be available in Cloudera DataFlow as well, when you import your flow definition.

      Important:

      The only exception to the rule that controller services must be defined in the process group that you plan to export, is when you use the Default NiFi SSL Context Service controller service in Data Hub Flow Management clusters. If a property references a controller service with that exact name, Default NiFi SSL Context Service, DataFlow creates a matching controller service when deploying the flow. The controller service points to a truststore that is setup with the environment's FreeIPA root certificate. This allows you to reference this controller service when interacting with CDP services that require TLS and are running in the same environment.

      For more information and instructions on how to set up your controller services, see *Best Practices for Flow Definition Development*.

   c) Configure each processor (and any other components) in your flow with the required values by double-clicking it.

      Alternatively, you can right-click the processor and select the Configure option from the processor's context menu.

      Parameterize component properties and use controller services in the configuration where needed. You can use the parameters you previously created or you can also create new parameters as you configure the components in your flow. To create a new parameter for a property, select the (Convert to Parameter) icon in the property's row.

      Note: To reference a parameter, the process group must first be assigned a parameter context. Processors and controller services of a process group can only reference parameters within the parameter context assigned to the process group.

      Important: A process group can only be assigned one parameter context, while a given parameter context can be assigned to multiple process groups.

   d) After configuring the processors and other data flow components, click Apply.

5. Connect the components in the data flow and configure the connections.
6. Check your data flow to make sure that configuration of all components and connections is valid.

Results

Your NiFi flow is ready to be downloaded as a JSON file.

What to do next

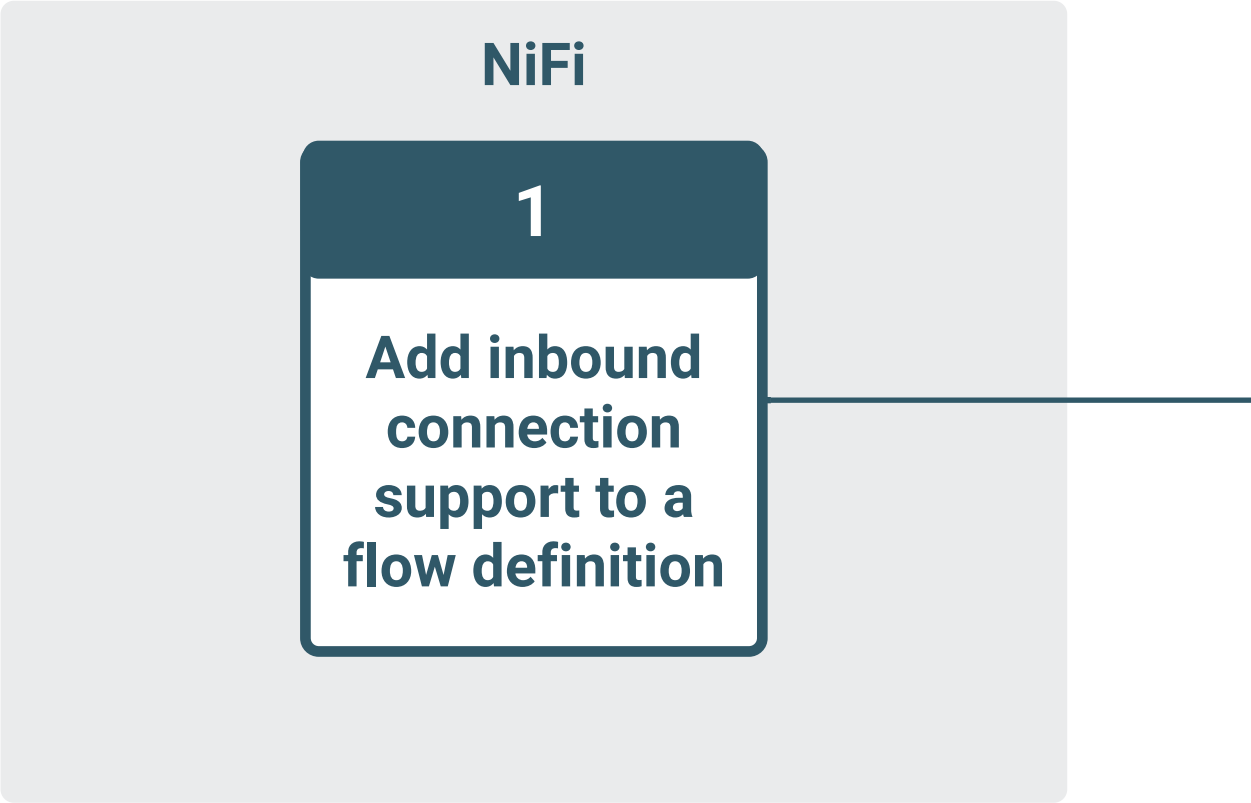
Download the data flow as a flow definition from NiFi and import it to Cloudera DataFlow.

Related Information

Setting up your Flow Management cluster
Best Practices for Developing Flow Definitions
Downloading a Flow Definition
Importing Flow Definitions
Deploying Flows
Parameters in Apache NiFi User Guide

Adding Inbound Connection support to a NiFi flow

You can enable a dataflow to listen on a public endpoint to data sources that are outside your CDP environment by adding inbound connection support to your flow



definition.

To enable your dataflow to use listen processors with inbound connection endpoint support, make the following addition to your flow definition:

1. In NiFi, open the flow definition where you want to enable inbound connection support.
2. Add the required listen processor to your flow definintion.

   CDF supports all listen processors, including Custom Processors.

3. Configure the processor to enable inbound connections.

   Port

   Provide a port number where the flow deployment listens for incoming data. You can either parameterize it and set the actual port number during flow deployment, or you can set an explicit port number when designing the flow.

   SSL Context Service

   Create an external StandardRestrictedSSLContextService for your processor. You must name this context service Inbound SSL Context Service. No other configuration is required. The SSL context service will be created during cluster deployment and all other properties will be populated with values generated for that NiFi cluster.

   Client Auth

   Set to “REQUIRED” to use mTLS

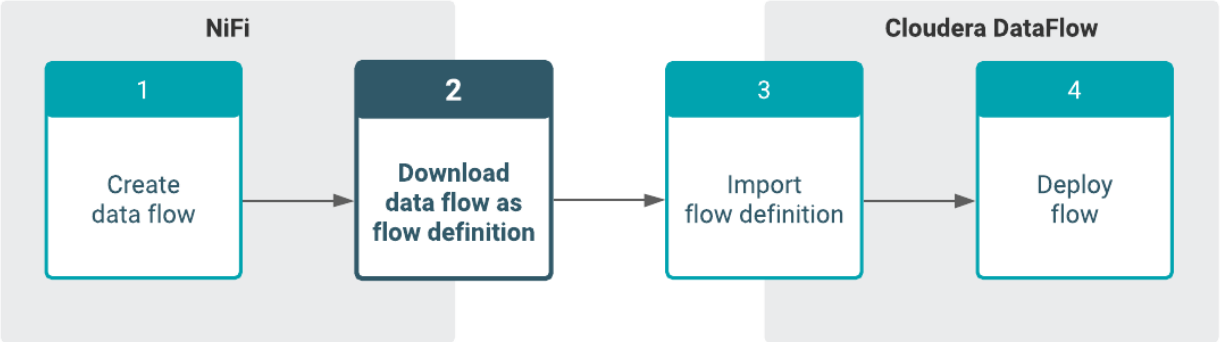
   Note: Most listen processors only support TCP. ListenSyslog supports both TCP and UDP. As dataflow deployments do not support mixed protocols for listening ports, and UDP does not support SSL Context Service and Client authentication, Cloudera advises to configure your ListenSyslog processor to use TCP protocol. This allows your CDF dataflow to use SSL Context Service for authentication and to listen to different data sources on different ports.

   If this is not possible, create separate flows that listen on UDP and TCP respectively.

4. Save and download the updated flow definition.

Download a flow definition from NiFi

Before you can run a NiFi flow in Cloudera DataFlow, you need to download the NiFi flow as a flow definition. The Download flow definition option in NiFi allows you to download the flow definition of a process group as a JSON file. The file can be imported into Cloudera DataFlow using the Import Flow Definition option.



Before you begin

- Review *Best Practices for Flow Definition Development* to ensure that you have a good understanding of flow isolation best practices.
- Develop your data flow in NiFi.

Important:

When you download your data flow from NiFi:

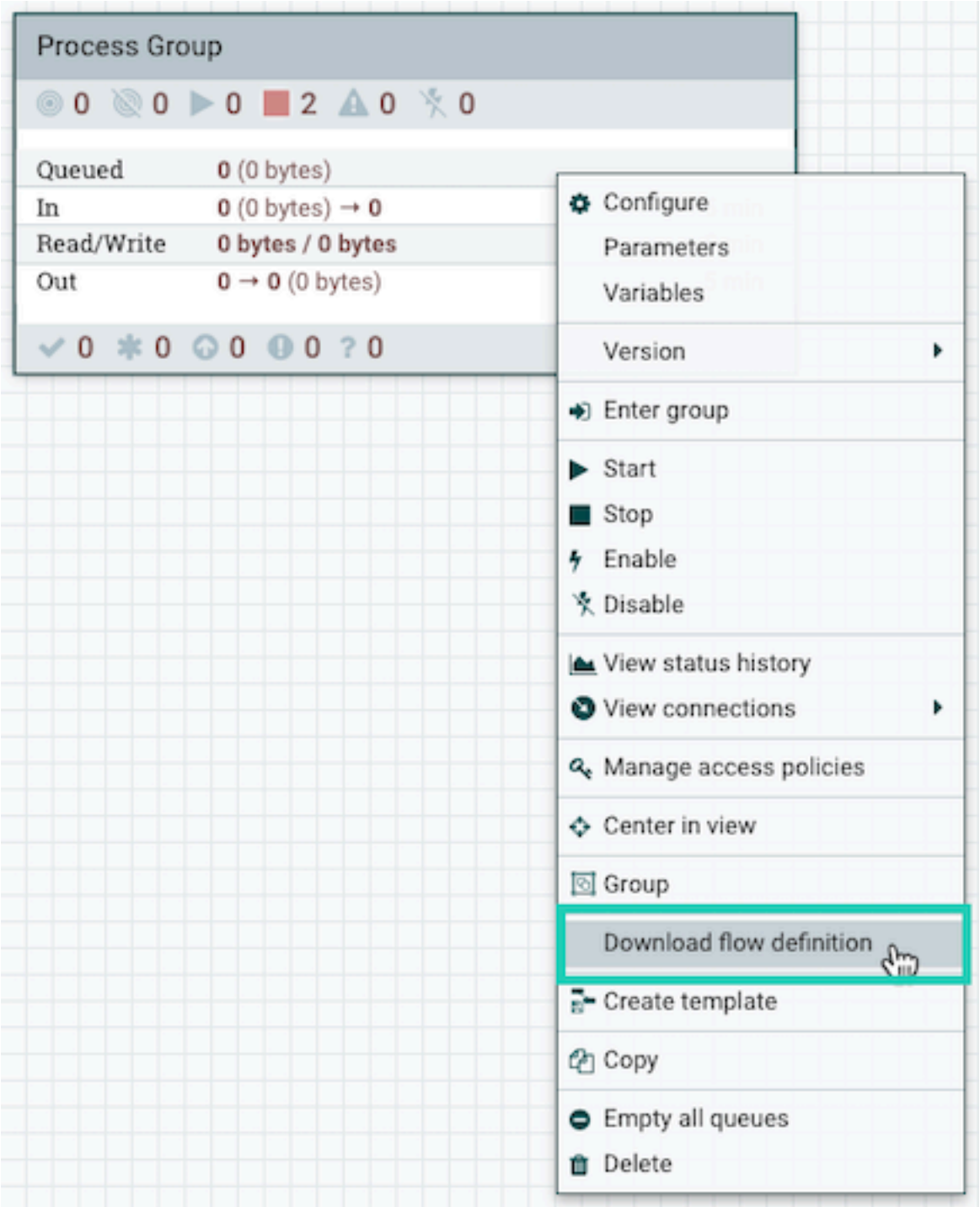
- Default values are exported with the flow definition. When you import this flow definition to Cloudera DataFlow and want to deploy your flow, the default values are carried over into the deployment wizard.
- Sensitive values are removed and you have to set them again during flow deployment in Cloudera DataFlow.
- Empty strings remain empty.

Procedure

1. Select the process group that you want to transfer to Cloudera DataFlow.

   For recommendations on how to export your process groups for deployment in Cloudera Dataflow, see *Best Practices for Flow Definition Development*.
2. Right-click the process group.

3. Select Download flow definition.



Note:

You can download flow definitions from NiFi at any level of your process group hierarchy. If you download one particular process group, your package contains the flow(s) included in the selected process group. When you download your flow definition from the root canvas level, you basically export all your data flows running on a cluster in one package, as the root canvas is the top level in the process group hierarchy.

If you download a versioned process group, the versioning information is not included in the downloaded JSON file.

After the download, you can upload your flow to Cloudera DataFlow.

For more information on how to plan and organize your NiFi data flows, see *Best Practices for Flow Definition Development*.

Results

The data flow is downloaded as a JSON file. You have a portable NiFi flow ready to be imported into Cloudera DataFlow.

What to do next

Import the NiFi flow to Cloudera DataFlow as a flow definition.

Related Information

Creating a Flow Definition

Best Practices for Flow Definition Development

Importing Flow Definitions

Deploying Flows