Cloudera DataFlow for Data Hub 4.0.0

Moving data from Cloudera Base on premises to Cloudera on cloud with NiFi site-to-site

Date published: 2019-12-16
Date modified: 2024-12-11

https://docs.cloudera.com/

Legal Notice

Contents

Moving data from Cloudera on premises to Cloudera on cloud with NiFi site-to-site 4
Understanding the use case 4
Preparing your clusters 4
Setting up your network configuration 5
Configuring your truststores 5
Defining your Cloudera on cloud data flow 7
Configuring Ranger policies for site-to-site communication 8
Defining your Cloudera Base on premises data flow 10

Moving data from Cloudera on premises to Cloudera on cloud with NiFi site-to-site

You can use Apache NiFi's site-to-site functionality to design a flow that moves data from your Cloudera Base on premises environment to Cloudera on cloud to ensure scalability and load balancing.

Understanding the use case

You can use the Apache NiFi site-to-site functionality to move data between a Cloudera on cloud and a Cloudera Base on premises environment. To do this, set up a cluster in each environment, prepare your network and truststore configurations, and then define your Cloudera Base on premises and Cloudera on cloud data flows and Apache Ranger configuration for site-to-site functionality.

Moving data between Cloudera on cloud and Cloudera Base on premises clusters is a common use case when there is a need for a lot of temporary compute resources that can be quickly provisioned in the cloud.

Imagine you have a large dataset on-premises and you wish to perform heavy computations on the dataset. You can use the following workflow to design a data flow that:

- Moves the dataset from your Cloudera Base on premises environment to your Cloudera on cloud environment
- Pushes the data to the appropriate destination
- Triggers the workload that processes the data while leveraging the auto-scaling capabilities that Cloudera on cloud provides
- Returns the results in your Cloudera Base on premises environment

All of this is powered by Cloudera Base on premises and Cloudera on cloud distributions, while ensuring consistent security policies at a fine-grained level with Apache Ranger, and data management and data lineage with Apache Atlas across the environments.

Preparing your clusters

The first step in preparing to move data from a Cloudera Base on premises cluster to a Cloudera on cloud cluster is to ensure that you have each cluster set up correctly.

Requirements for your Cloudera Base on premises cluster:

- Cloudera Flow Management running on Cloudera Base on premises
- Three-node NiFi compute cluster, secured with AutoTLS and configured with Apache Ranger

For details on deploying your Cloudera Flow Management cluster on Cloudera Base on premises, see *Cloudera Flow Management Deployment to Cloudera Base on premises.*

Requirements for your Cloudera on cloud Flow Management cluster:

- Flow Management clusters running on Cloudera on cloud
- Three-node NiFi compute cluster, secured with AutoTLS and configured with Apache Ranger

For details on deploying your Cloudera Flow Management cluster on Cloudera on cloud, see *Setting up your Flow Management cluster*.

Related Information

Cloudera Flow Management deployment to Cloudera Base on premises

Setting up your Flow Management cluster

Setting up your network configuration

You can use NiFi's site-to-site capabilities over a RAW TCP or over an HTTP network configuration. For this use case, you must configure site-to-site using HTTP over TLS. This has the advantage of using the NiFi port, which is also used to access the NiFi UI and APIs.

For the purpose of this use case, set up your site-to-site network configurations with the following assumptions:

- You are not using site-to-site through any proxy configuration
- You have a direct connection on port 8443 between NiFi nodes on your Cloudera Base on premises and Cloudera on cloud clusters

Set up your network configuration according to your architecture and requirements. For more information, see your Cloud provider documentation.

In this use case, NiFi on Cloudera Base on premises is responsible for initiating the site-to-site connection between the two environments to push and pull data to and from the NiFi cluster in Cloudera on cloud. The NiFi nodes in Cloudera on cloud must be reachable on port 8443 from the NiFi nodes in Cloudera Base on premises, but not necessarily the other way around.

Tip:

The site-to-site connection is bi-directional and depending on the cluster initiating the site-to-site connection, you will be in a push or pull model.

Configuring your truststores

You must configure your truststores so that each cluster is aware of and trusts the other cluster, to support the two-way TLS that is used to initiate the site-to-site communication between clusters. To do this, you need to download and merge the truststores for NiFi in Cloudera Base on premises and Cloudera on cloud.

Before you begin

- You have set up a Cloudera Base on premises and Cloudera on cloud cluster, and have the necessary network configurations established.
- You have the necessary administrative permissions to manipulate the truststore files. You require root access, and this is typically done by an Environment Administrator.
- You have the passwords for the Java Keystore (JKS) files available.

Procedure

1. Download the truststore from the clusters.
   a) Create a temporary directory in which you can edit the truststore files.

```
$ mkdir s2s-temp && cd s2s-temp
```

   b) Download the JKS truststore file for CA used by NiFi in your Cloudera Base on premises cluster.

```
$ scp -i <key>
root@<nifi_node>:/var/lib/cloudera-scm-agent/agent-cert/cm-auto-global_
truststore.jks
privatecloud_cm-auto-global_truststore.jks
```

   c) Download the JKS truststore file for CA used by NiFi in your Cloudera on cloud cluster.

```
$ scp -i <key>
```

```
cloudbreak@<nifi_node_public_cloud>:/var/lib/cloudera-scm-agent/agent-
cert/cm-auto-global_truststore.jks
publiccloud_cm-auto-global_truststore.jks
```

2. Merge the truststores.
   a. Make a copy of the Cloudera Base on premises JKS:

```
$ cp privatecloud_cm-auto-global_truststore.jks
privatecloud_cm-auto-global_truststore.jks.bak
```

   b. Merge the Cloudera on cloud JKS into the Cloudera Private Cloud Base JKS and rename the entries alias to prevent conflict:

```
$ keytool
-importkeystore
-srckeystore publiccloud_cm-auto-global_truststore.jks
-destkeystore privatecloud_cm-auto-global_truststore.jks
```

   The result will be similar to:

```
Importing keystore publiccloud_cm-auto-global_truststore.jks to private
cloud_cm-auto-global_truststore.jks...
Enter destination keystore password:
Enter source keystore password:
Entry for alias imported-ca-b379e6601f5ecfbbee2fefc4eb2efd4a successfull
y imported.
[...]
Entry for alias imported-ca-5945bad341623ae14991e09ffe851725 successfu
lly imported.
Entry for alias cmrootca-1 successfully imported.
Existing entry alias cmrootca-0 exists, overwrite? [no]:  no
Enter new alias name (RETURN to cancel import for this entry):  cmrootc
a-1-bis
Entry for alias cmrootca-0 successfully imported.
Entry for alias imported-ca-10c56ecc972802e53d1b7287ac2d1c6c successfu
lly imported.
[...]
Entry for alias imported-ca-840644351dd523125493ff4c28e694f7 successfull
y imported.
Import command completed:  140 entries successfully imported, 0 entries
failed or cancelled
```

   c. Use the copy you made to merge the Cloudera Base on premises JKS into the Cloudera on cloud JKS and rename the entries alias to prevent conflict:

```
$ keytool
-importkeystore
-srckeystore privatecloud_cm-auto-global_truststore.jks.bak
-destkeystore publiccloud_cm-auto-global_truststore.jks
```

   The result will be similar to:

```
Importing keystore privatecloud_cm-auto-global_truststore.jks.bak to pu
bliccloud_cm-auto-global_truststore.jks...
Enter destination keystore password:
Enter source keystore password:
Existing entry alias cmrootca-0 exists, overwrite? [no]:  no
```

```
Enter new alias name (RETURN to cancel import for this entry):  cmrootc
a-0-bis
Entry for alias cmrootca-0 successfully imported.
Import command completed:  1 entries successfully imported, 0 entries
failed or cancelled
```

3. Deploy the truststores.

   Deploy the modified Cloudera Base on premises and Cloudera on cloud JKS files on each NiFi node of the respective clusters.

   Note:

   Do not change the permissions and owners of the file (chmod/chown).

4. Restart your Cloudera Base on premises and Cloudera on cloud clusters.

What to do next

After you have configured your truststores, proceed by defining your data flow in your Cloudera on cloud cluster.

Defining your Cloudera on cloud data flow

To move data between cloud environments using NiFi site-to-site communication, you require a data flow in Cloudera on cloud that can receive data from the Cloudera Base on premises data flow. To create this data flow, configure a process group, and both an input and output port.

Before you begin

You have prepared you clusters, set up your network configurations, and configured your truststores.

Procedure

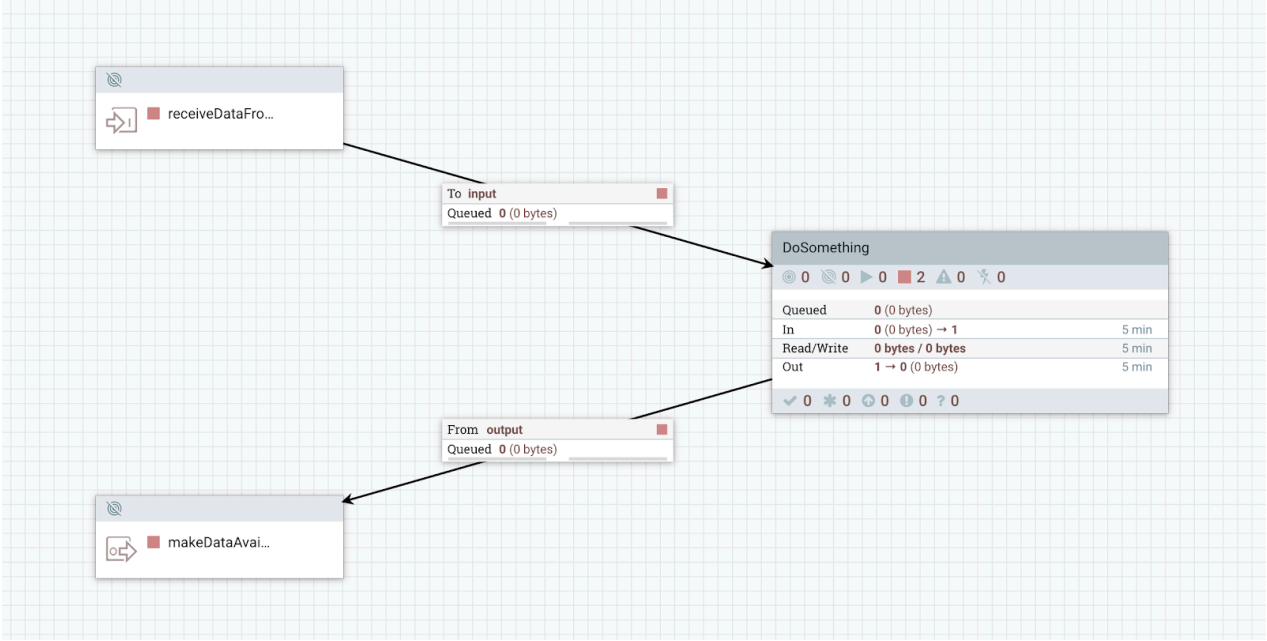
1. From your Cloudera on cloud NiFi cluster, create a Process Group to perform the operations you want to complete on the data received from and returned to the Cloudera Base on premises cluster.
2. Drag an Input Port onto the NiFi canvas.

   You must use this port for receiving data from NiFi's Cloudera Base on premises cluster.
3. Drag an Output Port onto the NiFi canvas.

   You must use this port to make data available for download to the Cloudera Base on premises cluster.
4. Connect your Cloudera on cloud data flow components.

   Ensure that you have specified the public endpoints of your NiFi nodes in the Cloudera on cloud cluster.
5. Start your data flow and ensure that both the input and output ports are running.

Example



What to do next

When you have completed your Cloudera on cloud data flow, proceed by configuring Apache Ranger to allow NiFi's site-to-site transmission.

Configuring Ranger policies for site-to-site communication

To allow NiFi's site-to-site communication between Cloudera on cloud and Cloudera Base on premises clusters, you need to configure Ranger authorization between the two clusters. To do this, create Ranger users in your Cloudera on cloud cluster that correspond to the Cloudera Base on premises NiFi nodes. Then create a new Ranger policy with site-to-site resources configured, and assign your Cloudera Base on premises NiFi node users to the policy.

Before you begin

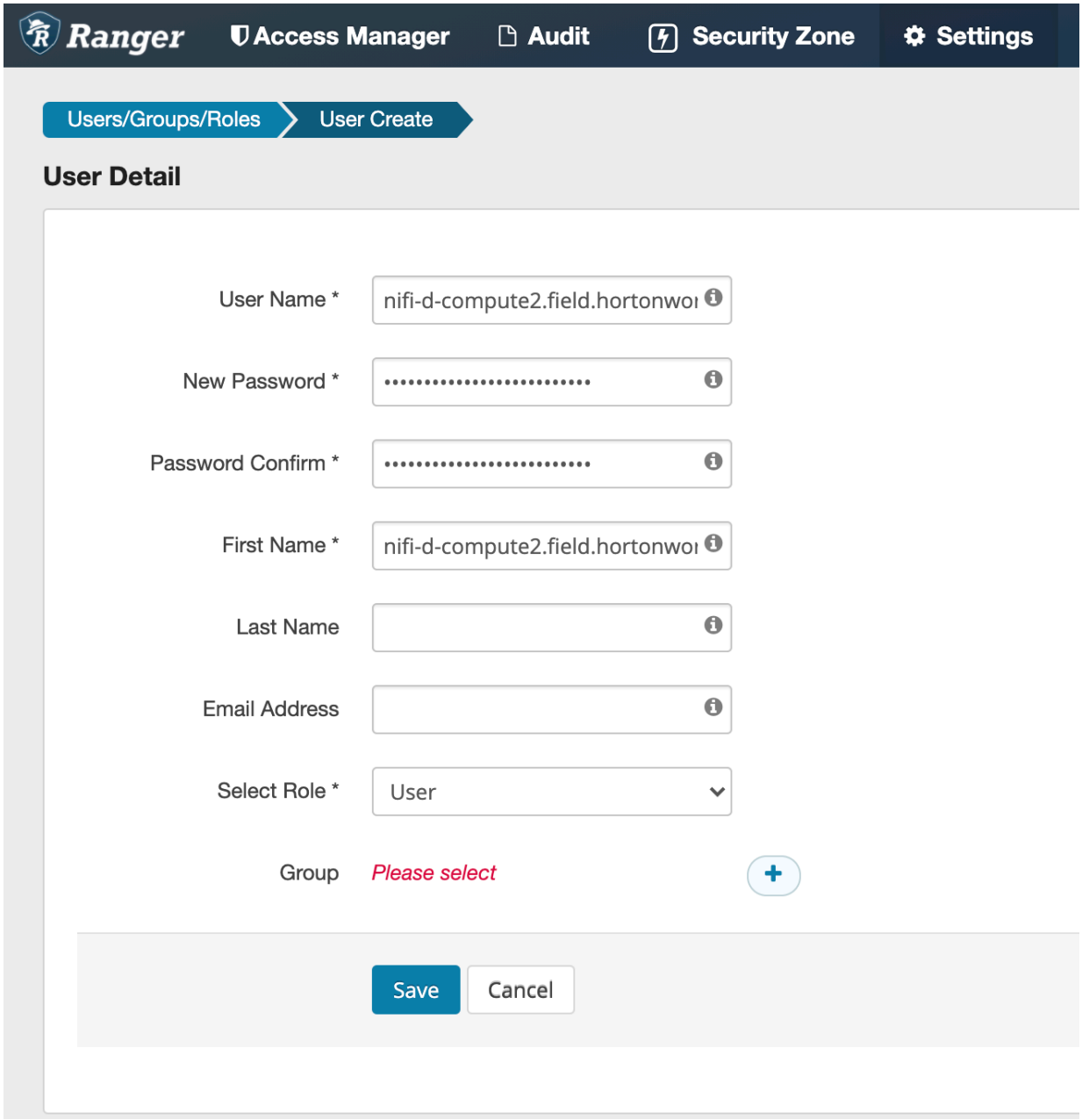
- You have defined your Cloudera on cloud data flow.
- You have a list of your FQDN Cloudera Base on premises host names. You need the host names to create the Ranger policies in Cloudera on cloud.

Procedure

1. In you Cloudera on cloud environment, launch the Ranger UI, click Settings Users/Groups/Roles User Create to add the users corresponding to the nodes of the Cloudera Base on premises cluster.

2. Click User Create to create one user per NiFi node running your Cloudera Base on premises environment.

   You create this user to make Ranger aware of the Cloudera Base on premises nodes, so that you can create policies by including them. Because this user is not used to authenticate on the Ranger UI, the password can be random.



3. Create a new policy in the NiFi Service in Ranger.

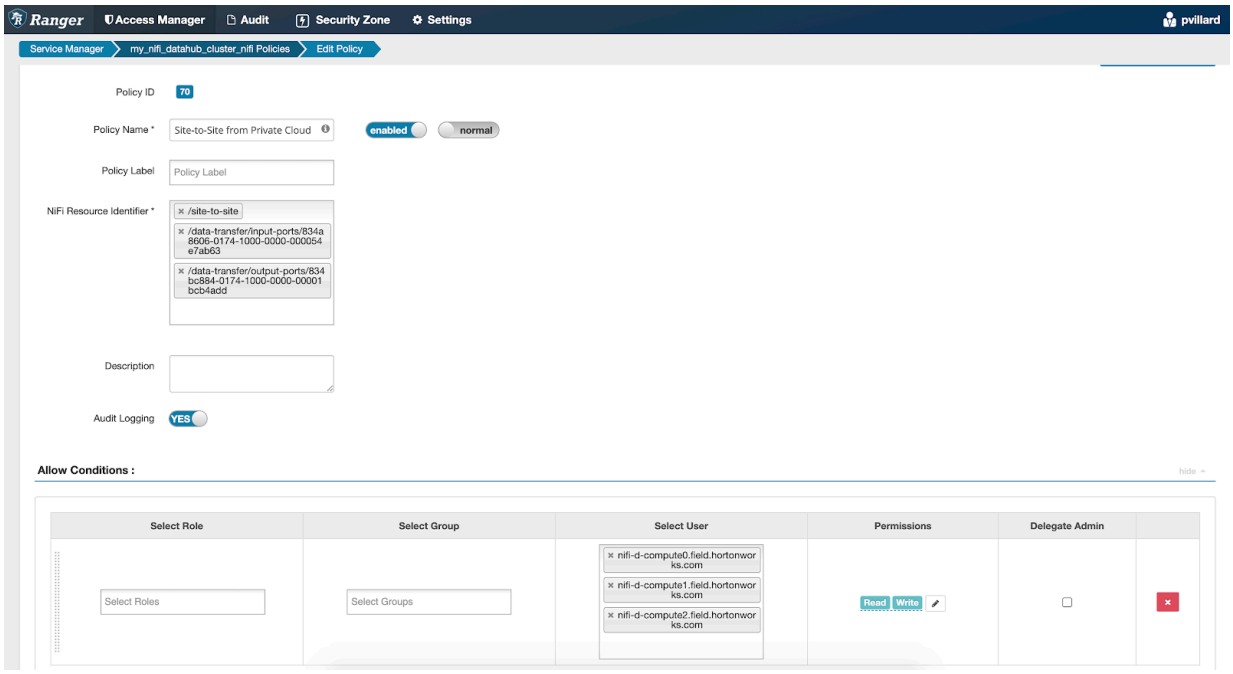
   You need to enter the following NiFi Resources:

   - /site-to-site
   - /data-transfer/input-ports/<ID of the Input Port>
   - /data-transfer/output-ports/<ID of the Output Port>

   Note:

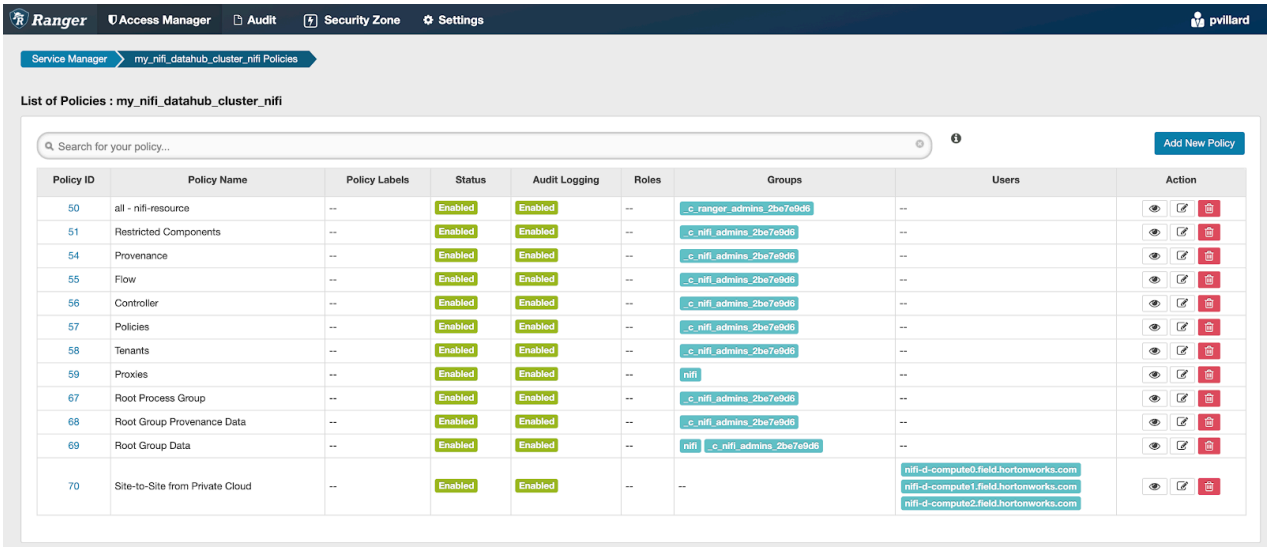
   You can retrieve the input and output port IDs by right-clicking the component and reviewing the configuration view.

4. Add the Cloudera Base on premises users you created in Step 2, and assign Read and Write permissions:



Results

Your policies are now listed.



Defining your Cloudera Base on premises data flow

To move data between cloud environments using NiFi site-to-site communication, you require a data flow in your Cloudera Base on premises cluster that can send and receive data from the Cloudera on cloud cluster. To create this data flow, connect a processor to a Remote Process Group configured with HTTP and enable transmission.

Before you begin

- You have defined your Cloudera on cloud data flow and configured Ranger policies for site-to site communication.

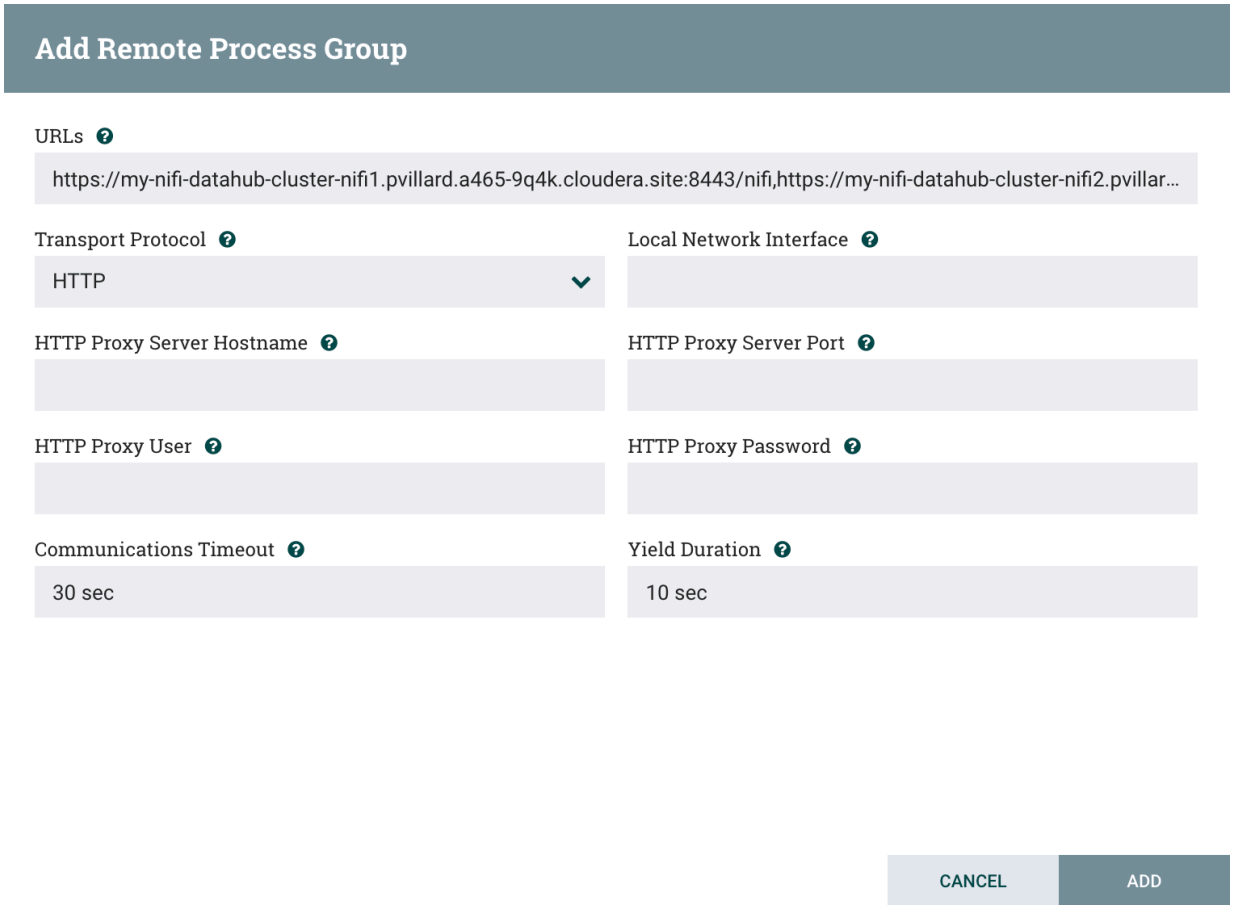
- You have the public FQDNs for your Cloudera on cloud cluster nodes.

Procedure

1. In your Cloudera Base on premises cluster, launch the NiFi UI and drag a `GenerateFlowFile` processor onto the canvas.

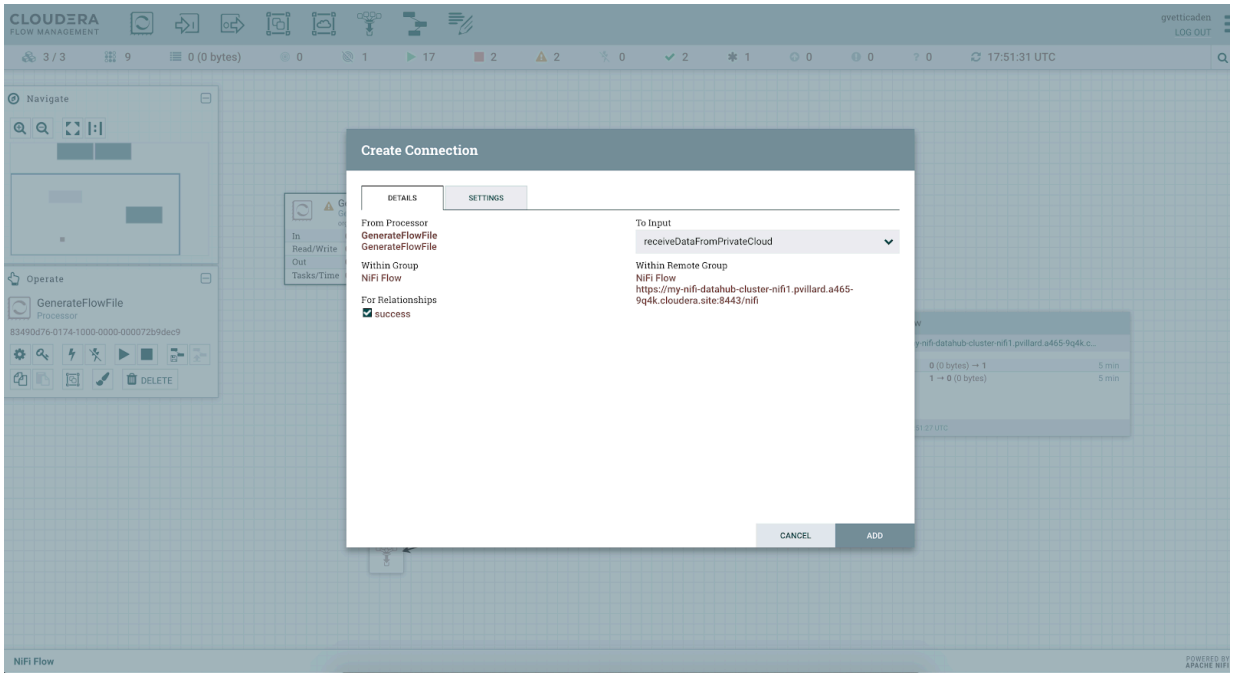
   For this use case, `GenerateFlowFile` creates 1MB files every 10 seconds.
2. Drag a Remote Process Group onto the NiFi canvas, configure HTTP protocol, and specify one or more of the NiFi nodes running on your Cloudera on cloud cluster.

   After the site-to-site connection is initiated, the source NiFi cluster is aware of the topology of the remote NiFi cluster and of any increase or decrease of the size of the remote cluster. However, it is recommended that you specify at least 2 nodes to ensure higher availability when the site-to-site connection is initiated.

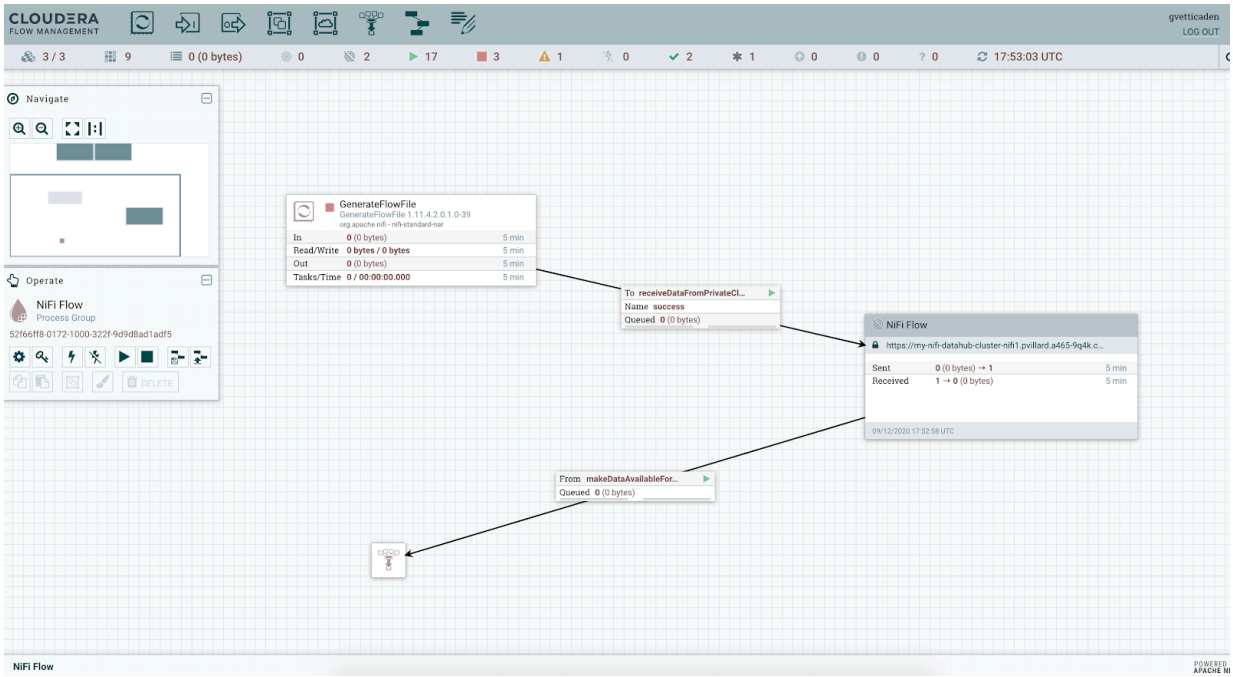


3. Right-click the Remote Process group and select Enable transmission.

4. Connect the `GenerateFlowFile` processor to the Remote Process Group and select the Input Port that you created and started on the remote cluster in Cloudera on cloud:



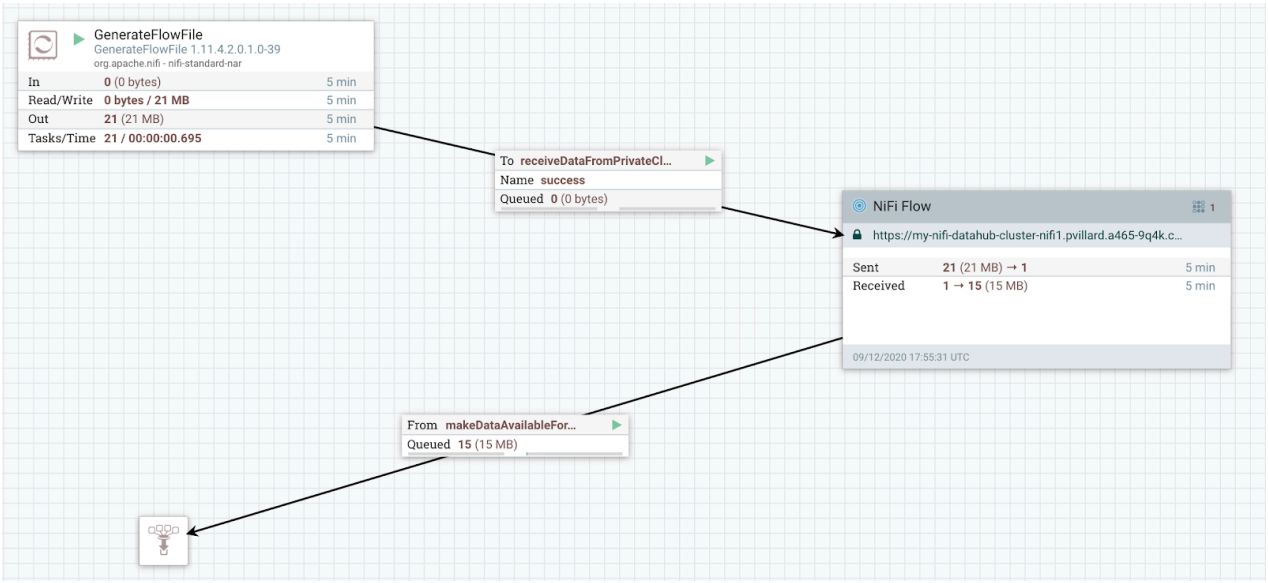
5. You can also define a connection from the Remote Process Group to another component to download data made available by the remote cluster running in the Cloudera on cloud environment. In the this example, the Remote Process Group is connected to a funnel.



Results

After you have defined the data flow for your Cloudera Base on premises cluster, start the Cloudera Base on premises data flow and confirm that the data is moving back and forth between the environments:

In the Cloudera Base on premises environment, your data flow looks similar to the following:



In the Cloudera on cloud environment, your data flow will look similar to the following: