Cloudera Flow Management 2.1.7

Moving data to Ozone

Date published: 2019-06-26
Date modified: 2024-06-03

https://docs.cloudera.com/

Legal Notice

Contents

Pushing data to Ozone using Apache NiFi

Integrating Apache NiFi with Ozone enables you to transfer data to Ozone storage within a secure CDP cluster. You can acquire data from various sources, including local file systems, databases, or other storage systems, and push this data to Ozone. Depending on your specific data source, you will need to incorporate the appropriate processors into your dataflow.

To extract data from the source system, you can use processors such as GetFile or GetDatabaseTable. Once you have acquired the data, Apache NiFi offers the PutHDFS and PutCDPObjectStore processors that help you to transfer the data to Ozone.

If you need to move data from Ozone, you can use the FetchHDFS / FetchCDPObjectStore or the ListHDFS / ListCDPObjectStore processors for the transfer.

The following example shows you a dataflow that generates sample data with the GenerateFlowFile processor and writes this data to Ozone.

Before you begin

Before developing the NiFi dataflow to push data to Ozone, you must meet the following prerequisites.

- You have set up an Ozone cluster in CDP, with Kerberos authentication enabled.
- Ranger permissions have been configured to grant your user the ability to create volumes in Ozone.
- You have created a volume and a bucket in Ozone.
- You have Cloudera Flow Management (CFM) installed on your system, as NiFi is delivered through CFM.
- You have added and configured the NiFi service.

Note: CFM is a separate parcel from the base Cloudera Runtime. For more information, see:

- CFM installation workflow
- Download locations

Building your dataflow

Learn how you can set up your NiFi dataflow that will enable you to move data to Ozone. This involves adding processors and other dataflow elements to the NiFi canvas, configuring them, and connecting the elements to create the dataflow.

Before you begin

You must have reviewed and met the prerequisites.

Procedure

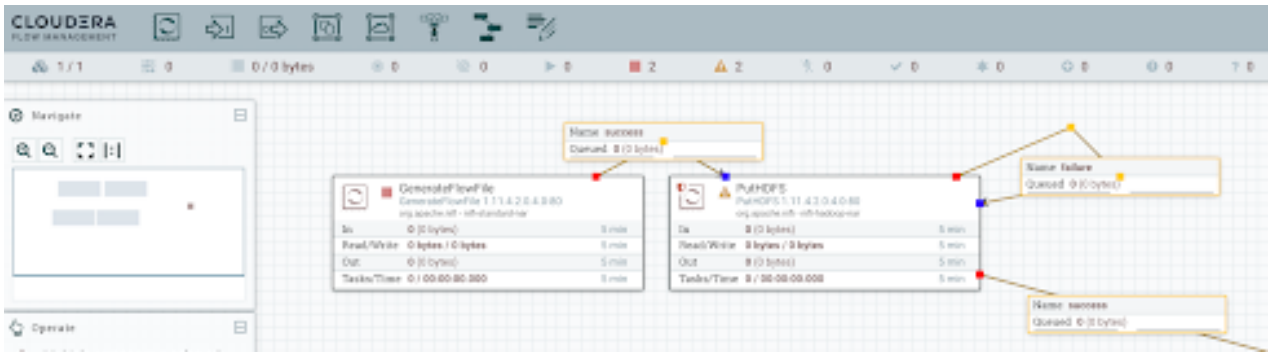
1. Launch NiFi from your CDP Public Cloud or CDP Private Cloud Base cluster.
2. Add the NiFi processors to your canvas.
   a. Select the Processor icon from the Cloudera Flow Management Actions pane, and drag a processor to the canvas.
   b. Use the Add Processor filter box to search for the processor you want to add, and then click Add.
   c. Add the following processors on the canvas:
      - GenerateFlowFile as your data source
      - PutHDFS or PutCDPObjectStore as your data ingest tool

3. Connect the two processors to create your basic dataflow.
   a) Click the Connection icon in the first processor, and drag it to the second processor.

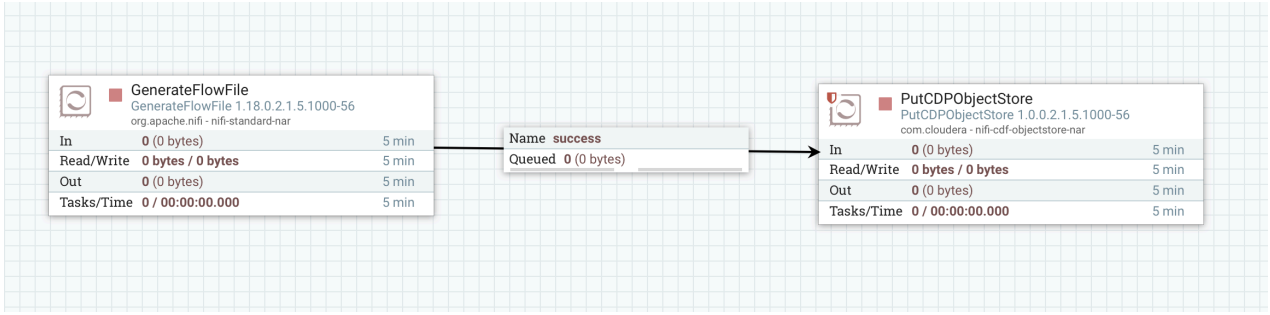
      A Create Connection dialog displays. It has two tabs: Details and Settings where you can configure the connection's name, flow file expiration time period, thresholds for back pressure, load balance strategy, and prioritization.
   b) Click Add to close the dialog box and add the connection to your flow.
   c) You can add success and failure funnels to your dataflow, which help you see where flow files are routed when your dataflow is running.

Results

Using PutHDFS



Using PutCDPObjectStore



What to do next

Once you have finished building the dataflow, move on to the following steps:

- Configure your source processor.
- Configure your target processor.

Configuring your source processor

You can set up a dataflow to push data into Ozone from many different locations.

About this task

In this example, you can see the configuration for the `GenerateFlowFile` processor of the Ozone ingest dataflow. You can use the `GenerateFlowFile` processor to create random data. It is escially useful during initial flow testing or when developing proof of concept dataflows. When you have validated that your dataflow aligns with your specific business use case, you can replace the source processor with one that retrieves data from your actual data source. For more information on this Apache NiFi processor, see the *Apache NiFi documentation*.

Before you begin
You must have built the dataflow.

Procedure
1. Launch the Configure Processor window, by right-clicking the `GenerateFlowFile` processor and selecting Configure.

   A configuration dialog box with the following tabs is displayed: Settings, Scheduling, Properties, and Comments.
2. Configure the processor according to the behavior you expect in your dataflow.

   See the *Example* section below for recommended configuration to satisfy this example use case.
3. Save the changes by clicking Apply.

Example
The following settings and properties are used in this example:

Table 1: GenerateFlowFile processor scheduling

Scheduling	Description	Example value for ingest data flow
Run Schedule	Run schedule dictates how often the processor should be scheduled to run. The valid values for this field depend on the selected scheduling strategy.	60 s

Table 2: GenerateFlowFile processor properties

Title	Description	Example value for ingest data flow

Custom text	If the value of Data Format is text and if Unique FlowFiles is set to false, you can provide custom to be used as the content of the generated FlowFiles.
	The expression statement in the example value generates a random ID between 1 and 10 000, with random last names assigned.
		`R_REGIONKEY, R_NAME, R_COMMENT`
		`100,foo1, blablabla`
		`101, foo2, blabla`
		`102, foo3, bla`

What to do next
Once you have configured your source processor, proceed to configuring your target processor.

Related Information
Apache NiFi documentation

Configuring your target processor
You can use the PutHDFS or the PutCDPObjectStore processor to move your data to Ozone, and you can select and configure the processor that best suits your use case.

About this task
The following examples step-by-step instructions for configuring each processor. They serve as a foundation for your setup and can be customized further. For more information on these Apache NiFi processors, see the *Apache NiFi documentation*.

PutHDFS
1. Right-click PutHDFS and click Configure Processor.

2. In the Hadoop Configuration Resources field, specify the Hadoop file system configuration.

   The path depends on where your configuration files are located. In this example, /etc/ozone/conf/core-site.xml,/etc/ozone/conf/ozone-site.xml is used.

   Note:

   Check /etc/ozone/conf/core-site.xml and make sure fs.defaultFS uses the same protocol as "Directory". For example, ofs://ozone1/ (Recommended), or o3fs://buck1.vol1.ozone1/ (Deprecated).

   You can provide a separate set of core-site.xml and ozone-site.xml just for NiFi PutHDFS processors so that the change to fs.defaultFS would not affect other components that act as an HCFS client.
3. Provide your Kerberos credentials:
   - Kerberos Principal
   - Kerberos Keytab or Kerberos Password
4. Set the Directory field to your output directory.

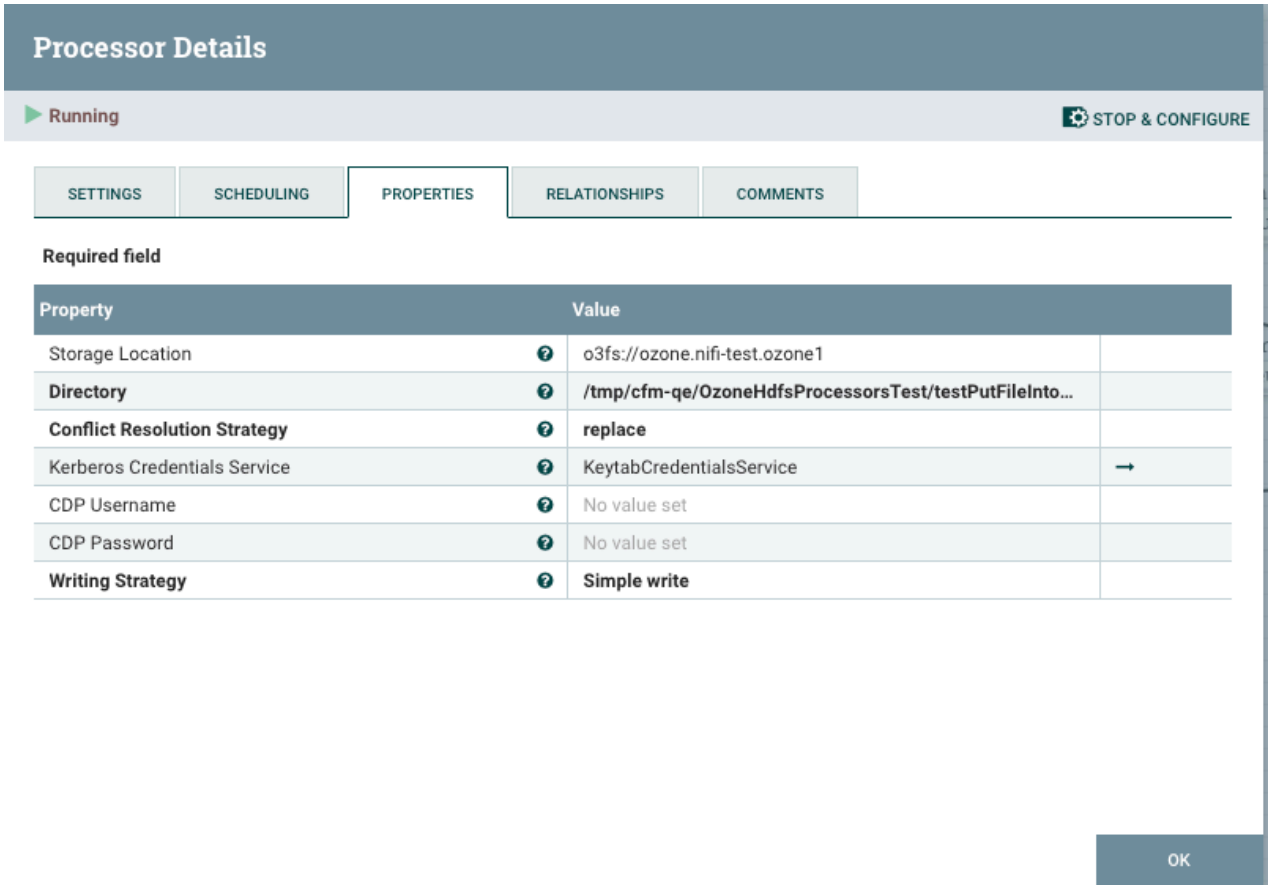
   For example: ofs://ozone1/vol1/buck1

   Note: Ensure to match the protocol used by fs.defaultFS.
5. Click Apply.

PutCDPObjectStore

1. Right-click PutCDPObjectStore and click Configure Processor.
2. Specify the Storage Location property. You can use this to set the desired fs.defaultFS value. For example: ofs://ozone1
3. Set the Directory property field to your output directory. For example: /vol1/buck1
4. Set the Conflict Resolution Policy. With this property, you define what should happen when a file with the same name already exists in the target directory. Possible values are fail, ignore, and replace.
5. For the credentials, you have two options:
   - Provide a Kerberos Principal and its Keytab using a Kerberos Credentials Service.
   - Provide a principal / username and password using the CDP Username and Password properties.
6. Define the Writing Strategy. For example: simple write
7. Click Apply.

Sample PutCDPObjectStore processor configuration:



What to do next

After you have configured your processors, start your dataflow and confirm success.

Related Information

Apache NiFi documentation

Starting your dataflow

After you start the dataflow, the GenerateFlowFile processor will continuously generate flow files based on the configured schedule and content settings, creating random data. You can confirm that you have successfully built a dataflow that can push data into Ozone by starting your dataflow and verifying that the data is moving through it.

Procedure

1. Select the processors that you want to start.
2. Click the Start icon in the Actions toolbar.

   Alternatively, right-click a single processor and choose Start from the context menu.

   The `Generate Flow File` processor should generate 1 MB data every minute, and the generated data will be written in the "Directory" you configured earlier.
3. You can verify that the files have indeed been written to the target directory by running the following command: ozone fs -ls ofs://ozone1/vol1/buck1/.

```
ozone fs -ls ofs://ozone1/vol1/buck1/

Found 4 items
```

```
-rw-rw-rw-   3 systest systest    1048576 2020-10-27 18:05 ofs://ozone1/
vol1/buck1/02e3bf6a-e419-4a12-9354-90e33f80f598
-rw-rw-rw-   3 systest systest    1048576 2020-10-27 18:05 ofs://ozone1/
vol1/buck1/a17b240a-6239-4262-b265-f1fc6af77882
-rw-rw-rw-   3 systest systest    1048576 2020-10-27 18:05 ofs://ozone1/
vol1/buck1/a6147d8c-91f6-41d6-a99d-7afb985dc96d
-rw-rw-rw-   3 systest systest    1048576 2020-10-27 18:05 ofs://ozone1/
vol1/buck1/f5610e4f-7b80-4a28-92ac-1646916a2324
```

4. Make sure to adjust the configuration of the GenerateFlowFile processor, the data ingest processor, and any other connected processors to match your specific use case and dataflow requirements.