Cloudera DataFlow for Data Hub 7.2.16

Ingesting Data into CDW using Iceberg Table Format

Date published: 2019-12-16
Date modified: 2023-01-11

https://docs.cloudera.com/

Legal Notice

Contents

Ingesting data into CDW using Iceberg table format

You can use an Apache NiFi data flow to ingest data in Iceberg table format into Cloudera Data Warehouse (CDW) in CDP Public Cloud by following these steps.

Understand the use case

Learn how to use NiFi to move data from a range of locations into an Iceberg table of a data warehouse in CDP Public Cloud.

Cloudera Data Warehouse (CDW) can be configured to use Iceberg as the table format for querying data with Impala and/or Hive. You can use Apache NiFi to move data from a range of locations into a Data Warehouse cluster running Apache Hive or Impala in CDP Public Cloud.

This use case walks you through creating a data flow that generates FlowFiles with random CSV data and writes this data into an Iceberg table in Hive or Impala.This gets you started with creating an Iceberg ingest data flow. If you want to use a different data source, see the other ingest documents for information on get and consume data processor options.

For more information on Iceberg table format, see *Apache Iceberg features*.

Related Information

Apache Iceberg features

Meet the prerequisites

Use this checklist to make sure that you meet all the requirements before you start building your data flow.

- You have a CDP Public Cloud environment.
- You have a workload username and password set to access Data Hub clusters. The predefined resource role of this user is at least "EnvironmentUser". This resource role provides the ability to view Data Hub clusters and set the FreeIPA password for the environment.
- Your user is synchronized to the CDP Public Cloud environment.
- You have a Flow Management cluster running Apache NiFi.
- You have a running Data Warehouse cluster in the same CDP environment properly configured to support Iceberg tables.

Create the Iceberg target table in CDW

Before you can ingest data into your Cloudera DataWarehouse (CDW) target in CDP Public Cloud, make sure that the table where you want to send data with NiFi is already created before you build your data flow.

About this task

The following example shows how to create a table with Impala:

```
CREATE TABLE customer (id int, name string, created_at timestamp)
PARTITIONED BY (country_code string)
STORED BY ICEBERG;
```

Build the data flow

Learn how you can create an ingest data flow to move data to an Iceberg table in Hive or Impala. This involves using NiFi in your Flow Management cluster. The example data flow generates FlowFiles with random CSV data and writes this data into an Iceberg table.

Procedure

1. Open NiFi in Data Hub.
   a) To access the NiFi service in your Flow Management Data Hub cluster, navigate to Management Console service Data Hub Clusters .
   b) Click the tile representing the Flow Management Data Hub cluster you want to work with.
   c) Click the NiFi icon in the Services section of the cluster overview page to access the NiFi UI.

      You will be logged into NiFi automatically with your CDP credentials.
2. Add the `GenerateFlowFile` processor for data input.

   This processor creates FlowFiles with random data or custom content.
   a) Drag and drop the processor icon into the canvas. This displays a dialog that allows you to choose the processor you want to add.
   b) Select the `GenerateFlowFile` processor from the list.
   c) Click Add or double-click the required processor type to add it to the canvas.

   You will configure the `GenerateFlowFile` processor to define how to create the sample data in *Configure the processor for your data source*.
3. Add the `PutIceberg` processor for data output.

   You will configure the `PutIceberg` processor in *Configure the processor for your data target*.
4. Connect the two processors to create a flow.
   a) Drag the connection icon from the first processor, and drop it on the second processor.

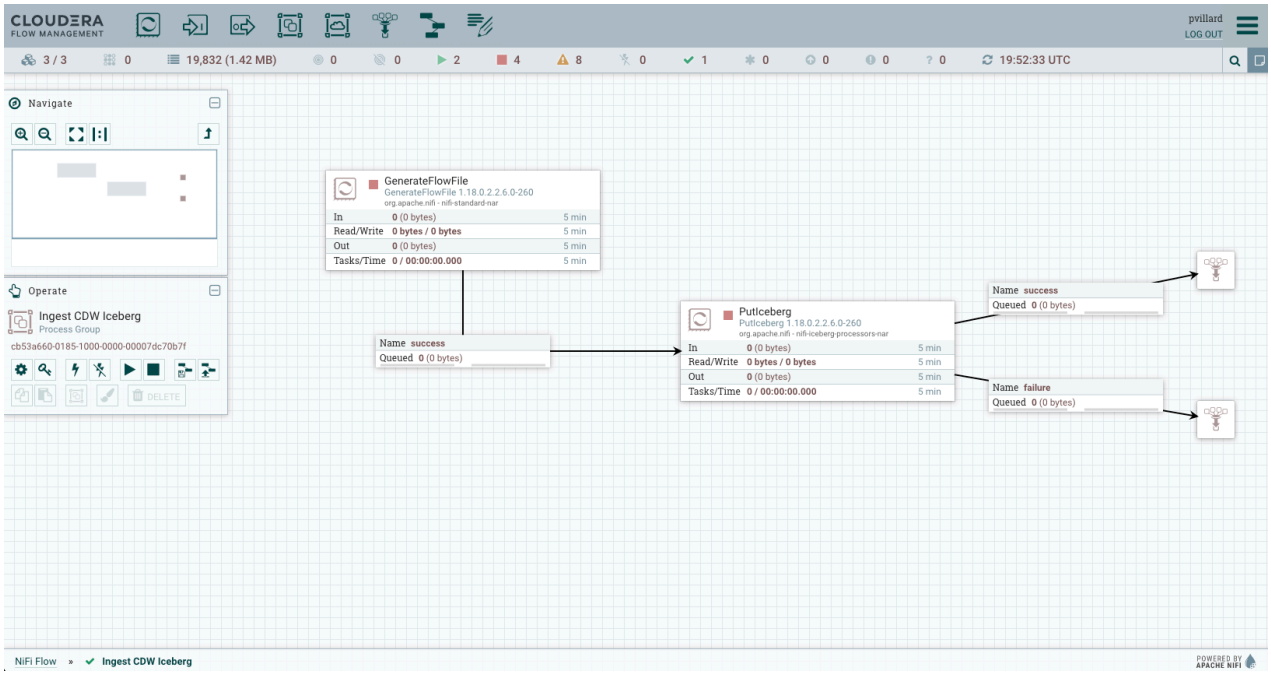
      A Create Connection dialog appears with two tabs: Details and Settings.
   b) Configure the connection.

      You can configure the connection's name, flowfile expiration time period, thresholds for back pressure, load balance strategy and prioritization.
   c) Click Add to close the dialog box and add the connection to your data flow.
5. Optionally, you can add success and failure funnels to your data flow, which help you see where flow files are being routed when your flow is running. Connect the `PutIceberg` to these funnels with success and failure connections.

   If you want to know more about working with funnels, see the *Apache NiFi User Guide*.

Results

Your data flow will look similar to the following:



What to do next

Create controller services for your data flow. You will need these services when configuring the `PutIceberg` processor.

Create and configure controller services for your data flow

Learn how to create and configure controller services for the CDW Iceberg ingest data flow.

About this task

You can add controller services to provide shared services to be used by the processors in your data flow. You will use these controller services later when you configure your processors.

Procedure

1. To add a controller service to your flow, right-click on the canvas and select Configure from the pop-up menu. This displays the Controller Services Configuration window.
2. Select the Controller Services tab.
3. Click the + button to display the Add Controller Service dialog.
4. Select the required controller service and click Add.
5. Click the Configure icon in the right-hand column and configure the options that you need.
6. When you have finished the configuration, click Apply to save the changes.
7. Enable the controller service by clicking the Enable button (flash) in the far-right column of the Controller Services tab.

Example

The following controller services are used in this CDW Iceberg ingest example:

- Hive Catalog Controller Service
- Kerberos User Service

See below for property details.

Configure the Hive Catalog Controller Service

In the context of CDW, the Hive Metastore URI is not exposed publicly. Instead, you need to provide the core-site.xml and hive-site.xml files of the CDP environment. In Flow Management DataHub clusters, these files are made available automatically on every node but the paths must be retrieved. If you use the Cloudera DataFlow data service and its Flow Designer, you can just use the #{CDPEnvironment} parameter.

Note: If you want to push data into Iceberg tables in Data Engineering DataHub clusters, you can use the Hive Metastore URI instead.

Table 1: Hive Catalog Controller Service properties

Property	Description	Example value for ingest data flow
Hive Metastore URI	Provide the URI of the metastore location.	
Default Warehouse Location	Provide the default warehouse location in the HDFS file system.	
Hadoop Configuration Resources	Add a comma-separated list of Hadoop Configuration files, such as hive-site.xml and core-site.xml for kerberos. Include full paths to the files so that NiFi can refer to those configuration resources from your specified path.	/etc/hive/conf.cloudera.data_context_connector-975b/hive-site.xml,/etc/hadoop/conf.cloudera.stub_dfs/core-site.xml

Configure the Kerberos User Service

Usethe Kerberos Password User Service so that you do not need to distribute a keytab file across the NiFi nodes of the cluster.

It is best practice to have a dedicated Machine User created in the control plane for your specific use case so that you can configure specific policies in Ranger and have better control in case of multi-tenancy with many use cases.

Table 2: Kerberos User Service properties

Property	Description	Example value for ingest data flow
Kerberos Principal	Specify the user name that should be used for authenticating with Kerberos. Use your CDP workload username to set this Authentication property.	srv_nifi_to_iceberg
Kerberos Password	Provide the password that should be used for authenticating with Kerberos. Use your CDP workload password to set this Authentication property.	password (sensitive value)

What to do next

Configure the processor for your data source.

Configure the processor for your data source

Learn how to configure a data source processor for the Kafka ingest data flow.

About this task

You can set up a data flow to move data in Iceberg table format into Cloudera Data Warehouse (CDW) from many different locations. This example assumes that you are using sample data generated by the GenerateFlowFile processor.

Procedure

1. Launch the Configure Processor window, by right clicking the `GenerateFlowFile` processor and selecting Configure. A configuration dialog with the following tabs is displayed: Settings, Scheduling, Properties, and Comments.
2. Configure the processor according to the behavior you expect in your data flow.

   The `GenerateFlowFile` processor can create many FlowFiles very quickly. Setting the run schedule to a reasonable value is important so that the flow does not overwhelm the system.
3. When you have finished configuring the options you need, save the changes by clicking the Apply button.

   Make sure that you set all required properties, because you cannot start the processor until all mandatory properties have been configured.

Example

The following settings and properties are used in this example:

Table 3: 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.	500 ms

Table 4: GenerateFlowFile processor properties

	Description	Example value for ingest data flow
Custom text	If Data Format is text and if Unique FlowFiles is 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.	

What to do next

Configure the processor for your data target.

Configure the processor for your data target

Learn how to configure a data target processor for the CDW Iceberg ingest data flow.

About this task

You can set up a data flow to move data to many locations. This example assumes that you are moving data to Cloudera Data Warehouse (CDW). Once your data is collected and available in your flow, you can use the PutIceberg processor to send the data in the CDW table.

Procedure

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

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

   With the PutIceberg processor, you can ingest data using Iceberg table format.

3. When you have finished configuring the options you need, click Apply to save the changes.

   Make sure that you set all required properties, because you cannot start the processor until all mandatory properties are configured.

Example

The following settings and properties are used in this example:

Table 5: PublishKafka2RecordCDP processor properties

Property	Description	Example value for ingest data flow
Record Reader	Select the Record Reader of your choice based on the format of your incoming data that should be pushed in the table.	JSON Reader
Catalog Service	Add the Hive Catalog Controller Service you have created.	Hive Catalog Service
Catalog Namespace	Set the Catalog Namespace property to the name of the database where you created the destination table.	default
Table Name	Set the Table Name property to the name of the table previously created.	customer
Kerberos User Service	Add the Kerberos User Service of your choice (based on username/password or based on username/keytab) that you have created.	Kerberos Password User Service

What to do next

Your data flow is ready to ingest data into an Iceberg table in CDW. You can now start the data flow.

Start the data flow

Learn how run the flow and confirm that data is successfully sent to the Iceberg table.

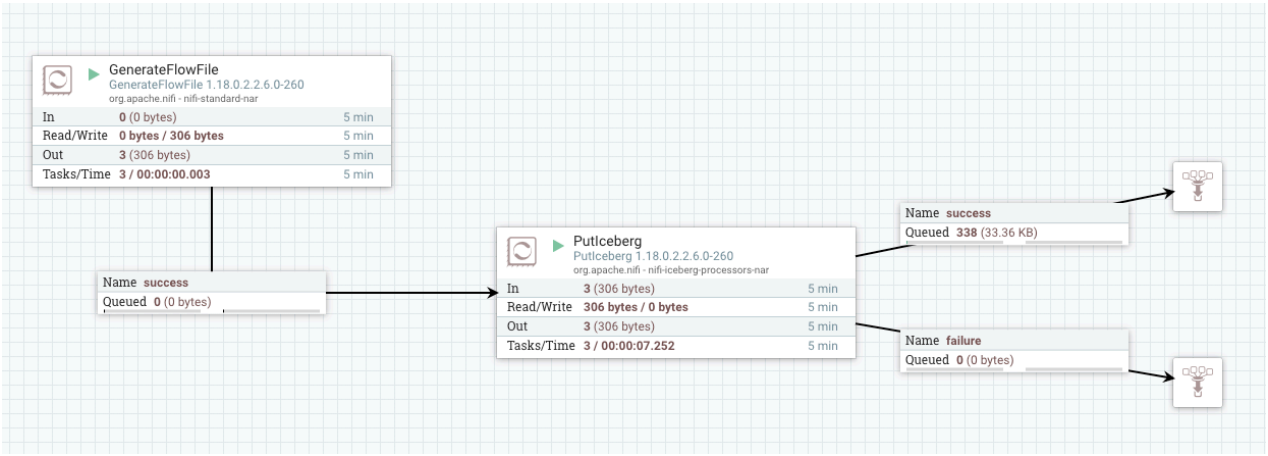
Procedure

1. Select all the data flow components you want to start.
2. Click the Start icon in the Actions toolbar.

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

Results

Your flow should be running without any errors. Data should be generated in the form of FlowFiles, and the files should be written to your designated Kafka topic.



You can verify the operation of your Iceberg ingest data flow by checking that data is running through the flow you have built and it actually appears in the target table.

- You can verify that NiFi processors are not producing errors.
- You can look at the processors in the UI to see the amount of data that has gone through them. You can also right-click on the processors, or on connections to view status history.
- You can query the data in Hue: