Cloudera DataFlow for Data Hub 7.2.16

Ingesting Data into Google Cloud Storage

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

CLOUDERA

Legal Notice

Contents

Ingesting data into Google Cloud Storage

You can use an Apache NiFi data flow to ingest data into Google Cloud Storage (GCS) object stores in CDP Public Cloud by following these steps.

Understand the use case

Learn how you can use a Flow Management cluster in CDP Public Cloud to build a flow that ingests data to Google Cloud Storage (GCS). This example use case shows you how to use Apache NiFi to move data into GCS buckets.

Why move data to object stores?

Cloud environments offer numerous deployment options and services. There are many ways to store data in the cloud, but the easiest option is to use object stores. Object stores are extremely robust and cost-effective storage solutions with multiple levels of durability and availability. You can include them in your data pipeline, both as an intermediate step and as an end state. Object stores are accessible to many tools and connecting systems, and you have a variety of options to control access.

Apache NiFi for cloud ingestion

As a part of Cloudera Data Flow in CDP Public Cloud, Flow Management clusters run Apache NiFi. You can use NiFi to move data from a range of locations into object stores. NiFi supports data transfer from nearly any type of data source to cloud storage solutions in a secure and governed manner. It also offers a scalable solution for managing data flows with guaranteed delivery, data buffering / pressure release, and prioritized queuing.

NiFi helps you to build a cloud ingestion data flow in a quick and user-friendly way, by simply dragging and dropping a series of processors on the NiFi user interface. When the data flow is ready, you can visually monitor and control the pipeline you have built.

This use case walks you through the steps associated with creating an ingest-focused data flow by tailing the NiFi log files and sending the data into a GCS bucket. If you are moving data from a different source, see the *Getting Started with Apache NiFi* for information about how to build a data flow, and about other data get and consume processor options.

Related Information

Getting Started with Apache NiFi

Ingesting Data into Apache Kafka

Ingesting Data into Apache HBase

Ingesting data into Apache Hive

Ingesting Data into Apache Kudu

Ingesting Data into Azure Data Lake Storage

Ingesting Data into Amazon S3

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 CDP 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 Data Hub cluster running in your CDP Public Cloud environment.

- Your CDP user has been added to the appropriate pre-defined Ranger access policies to allow access to the NiFi UI.
- You have created a target folder in your GCS bucket through the Google Cloud console for the data to be moved to the cloud.
- You have created a GCP service account with the right role and policies to write into the GCS bucket you want to use in your data flow.

Related Information

[Understanding roles and resource roles](#)

[Setting up your Flow Management cluster](#)

[IDBroker - Cloud identity federation](#)

[Google Cloud Console User Guide](#)

Build the data flow

Learn how you can create an ingest data flow to move data to Google Cloud Storage (GCS) buckets. This involves opening Apache NiFi in your Flow Management cluster, adding processors and other data flow objects to your canvas, and connecting your data flow elements.

About this task

You can use the `PutHDFS` or `PutGCSObject` processors to build your Google Cloud ingest data flows. Regardless of the type of flow you are building, the first steps in building your data flow are generally the same. Open NiFi, add your processors to the canvas, and connect the processors to create the flow.

Procedure

1. Open NiFi in Data Hub.

   a) To access the NiFi service in your Flow Management Data Hub cluster, navigate to Management Console serviceData Hub Clusters.

   b) Click the tile representing the Flow Management Data Hub cluster you want to work with.

   c) Click Nifi in the Services section of the cluster overview page to access the NiFi UI.
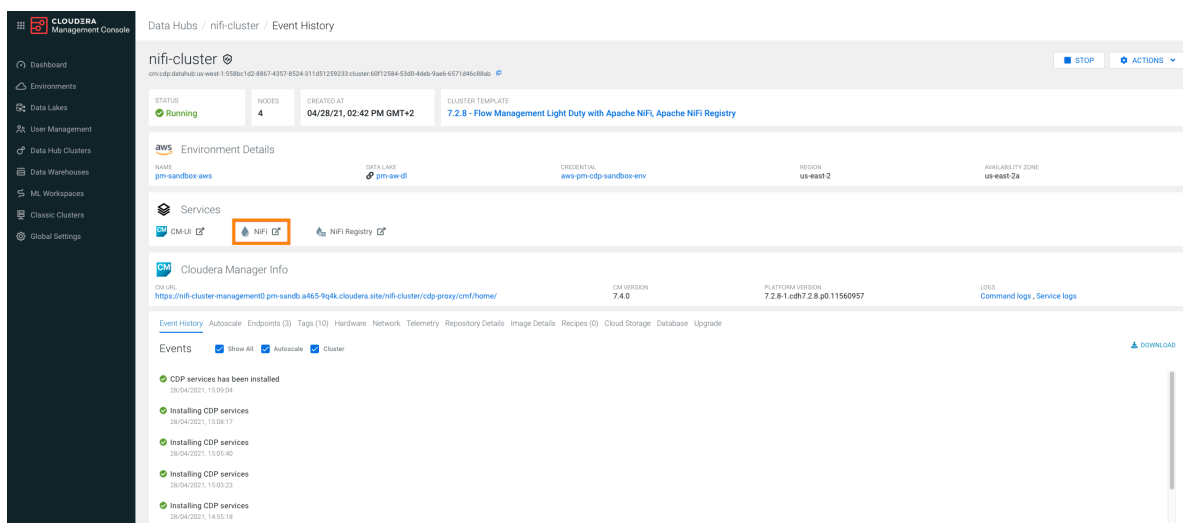
   You will be logged into NiFi automatically with your CDP credentials.

2. Add the `TailFile` processor for data input in your data flow.

   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 `TailFile` processor from the list.

   c) Click Add or double-click the required processor type to add it to the canvas.

   > **Note:** This processor continuously tails a file to get its data as flow files.

3. Add the `MergeContent` processor.

   > **Note:** This processor merges a group of flow files together based on a user-defined strategy and packages them into a single flow file. It updates the mime.type attribute as appropriate.

   When using the `TailFile` processor, you are pulling small-sized flow files, so it is practical to merge them into larger files before writing them to GCS.

4. Add a processor for writing data to GCS.

   You have two options:

   - `PutHDFS` processor: The HDFS client writes to GCS through the GCS API. This solution leverages centralized CDP security. You can use the usernames and passwords you set up in CDP for central authentication, and all requests go through IDBroker.

   - `PutGCSObject` processor: This is a GCS-specific processor that directly interacts with the Google Cloud Storage API. It is not integrated into CDP's authentication and authorization frameworks. If you use it for data ingestion, the authorization process is handled in GCP, and you need to provide the Google Cloud access credentials file / information when configuring the processor. You can do this by using the GCPCredentialsCo ntrollerService controller service in your data flow.

   a) Drag and drop the processor icon into the canvas.

   In the dialog box you can choose which processor you want to add.

   b) Select the processor of your choice from the list.

   c) Click Add or double-click the required processor type to add it to the canvas.

5. Connect the processors to create the data flow by clicking the connection icon in the first processor, and dragging and dropping it on the second processor.

   A Create Connection dialog appears with two tabs: Details and Settings. You can configure the connection's name, flowfile expiration time period, thresholds for back pressure, load balance strategy and prioritization.

   a) Connect `TailFile` with `MergeContent`.

   b) Add the success flow of the `TailFile` processor to the `MergeContent` processor.

   c) Click Add to close the dialog box and add the connection to your data flow.

   d) Connect `MergeContent` with your target data processor (`PutHDFS` / `PutGCSObject`).

   e) Add the merged flow of the `MergeContent` processor to the target data processor.

   f) Click Add to close the dialog box and add the connection to your data flow.

6. Optionally, you can add funnels to your flow.

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

Results

This example data flow has been created using the `PutHDFS` processor.

What to do next

If you are using the `PutHDFS` processor, configure IDBroker mapping authorization.

If you are using the `PutGCSObject`processor, you do not need IDBroker mapping, so proceed to configuring controller services for the processors in your data flow.

Related Information

Apache NiFi Documentation

Create IDBroker mapping

You must create IDBroker mapping for a user or group to access cloud storage. As a part of Knox, the IDBroker allows a user to exchange cluster authentication for temporary cloud credentials. To enable your CDP user to utilize the central authentication features CDP provides and to exchange credentials for Google Cloud access tokens, you have to map your CDP user to the correct IAM role. Learn how you can create the IDBroker mapping for the `PutHDFS` processor for your data flow.

About this task

This task is only needed if you use the `PutHDFS` processor for accessing your GCS bucket.

You have to set the mapping between the CDP user that will be used in your NiFi flow and the GCP service account you created. In the Technical Preview, you can manage ID Broker mappings only using the CDP CLI.

Use the cdp environments   set-id-broker-mappings command to set the mappings.

What to do next
Configure controller services for your data flow.
Related Information
IDBroker - Cloud identity federation
Onboarding CDP users and groups for cloud storage
Using CDP CLI

Create controller services for your data flow

Controller services provide shared services that can be used by the processors in your data flow. You will use these Controller Services later when you configure your processors. Learn how you can create and configure the controller services for a Google Cloud ingest data flow in CDP Public Cloud.

Before you begin

This task is only needed if you use the `PutGCSObject` processor for accessing your GCS bucket.

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.

In this case, add the GCPCredentialsControllerService controller service and configure it by providing the credentials file of the GCP service account you created.

Note: The credentials file is usually a JSON file.

5. Perform any necessary Controller Service configuration tasks by clicking the Configure icon in the right-hand column.

Configure Controller Service

SETTINGS	PROPERTIES	COMMENTS

Required field   +

Property		Value
Use Application Default Credentials	❷	false
Use Compute Engine Credentials	❷	false
Service Account JSON File	❷	No value set
Service Account JSON	❷	No value set
Proxy Configuration Service	❷	No value set

CANCEL   APPLY

6. When you have finished configuring the options you need, save the changes by clicking Apply.

7. Enable the controller service by clicking the Enable button (flash) in the far-right column of the Controller Services tab.

What to do next
Configure the processors in your data flow.
Related Information
Adding Controller Services for data flows
Apache NiFi Documentation

Configure the processor for your data source

You can set up a data flow to move data to Google Cloud Storage (GCS) from many different locations. This example assumes that you are streaming the data that you are tailing from a local file on the NiFi host. Learn how you can configure the `TailFile` data source processor for your GCS ingest data flow.

Procedure

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

   This gives you a configuration dialog with the following tabs: Settings, Scheduling, Properties, Comments.

2. Configure the processor according to the behavior you expect in your data flow.

   You can provide the path to the nifi-app.log file that contains the log of the NiFi service running on the host.

3. When you have finished configuring the options you need, save the changes by clicking Apply.

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

What to do next
Configure the processor for merging your records.
Related Information
Configuring a processor
Apache NiFi Documentation
Getting Started with Apache NiFi

Configure the processor for merging records

You can use merge together multiple record-oriented flow files into a large flow file that contains all records of your data input. Learn how you can configure the `MergeContent` processor for your Google Cloud Storage (GCS) ingest data flow.

Procedure

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

   This gives you a configuration dialog with the following tabs: Settings, Scheduling, Properties, Comments.

2. Configure the processor according to the behavior you expect in your data flow.

3. When you have finished configuring the options you need, save the changes by clicking Apply.

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

   For a complete list of `MergeContent` properties, see the *processor documentation*.

What to do next
Configure the processor for your data target.
Related Information
Configuring a processor
Apache NiFi Documentation
Getting Started with Apache NiFi

Configure the processor for your data target

This example use case shows you two options for ingesting data to Google Cloud Storage (GCS). Learn how you can configure these data NiFi processors for your GCS ingest flow.

Procedure

1. Launch the Configure Processor window by right clicking the processor you added for writing data to GCS (`PutHDFS`or `PutGCSObject`) and selecting Configure.

   This gives you a configuration dialog with the following tabs: Settings, Scheduling, Properties, Comments.

2. Configure the processor according to the behavior you expect in your data flow.

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

3. When you have finished configuring the options you need, save the changes by clicking Apply.

   The following properties are used for `PutHDFS`:

Table 1: PutHDFS processor properties

Property	Description	Example value for ingest data flow
Hadoop Configuration Resources	Specify the path to the core-site.xml configuration file. Make sure that the default file system (fs, default.FS) points to the GCS bucket you are writing to. **Note:** The core-site.xml file used for the Hadoop Configuration Resources property is present on every Flow Management cluster. Cloudera Manager stores the right core-site.xml file in the same /etc directory for every cluster.	/etc/hadoop/conf.cloudera.core_settings/ core-site.xml
Kerberos Principal	Specify the Kerberos principal (your username) to authenticate against CDP.	srv_nifi-gcs-ingest
Kerberos Password	Provide the password that should be used for authenticating with Kerberos.	password

Property	Description	Example value for ingest data flow
Directory	Provide the path to your target directory in Google Cloud expressed in a GCS compatible path. It can also be relative to the path specified in the core-site.xml file.	gs://your path/customer

You can leave all other properties as default configurations.

For a complete list of `PutHDFS` properties, see the *processor documentation*.

If you want to use the `PutGCSObject` processor to store the data in Google Cloud Storage, you have to configure your processor to reference the controller service you configured before:

If you want to move data to a different location, review the other use cases in the *Cloudera Data Flow for Data Hub library*.

What to do next
Your data flow is ready to ingest data into Google Cloud Storage. Start the flow.
Related Information
Configuring a processor

Apache NiFi Documentation

Data ingest use cases in Cloudera Data Flow for Data Hub

Start the data flow

When your flow is ready, you can begin ingesting data into Google Cloud Storage buckets. Learn how to start your GCS ingest data flow.

Procedure

1. To initiate your data flow, 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. Data should be read from Kafka and it should be written to the defined folder of your GCS bucket.

Results

What to do next

It is useful to check that data is running through the flow you have created.

Verify data flow operation

Learn how you can verify the operation of your Google Cloud Storage (GCS) ingest data flow.

About this task

There are a number of ways you can check that data is running through the flow you have built.

Procedure

- You can verify that NiFi processors are not producing errors.
- You can look at the processors in the UI, where you can 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 check that the data generated appears in your Google Cloud Storage bucket. To do this, return to the dedicated folder of your GCS bucket in the Google Cloud console, where you should see your files listed. You may have to refresh the page depending on your browser/settings.

Monitoring your data flow

Learn about the different monitoring options for your Google Cloud Storage (GCS) ingest data flow in CDP Public Cloud.

You can monitor your data flow for information about health, status, and details about the operation of processors and connections. NiFi records and indexes data provenance information, so you can conduct troubleshooting in real time.

Data statistics are rolled up on a summary screen (the little table icon on the top right toolbar which lists all the processors). You can use the `MonitorActivity` processor to alert you, if for example you have not received any data in your flow for a specified amount of time.

If you are worried about data being queued up, you can check how much data is currently queued. Process groups also conveniently show the totals for any queues within them. This can often indicate if there is a bottleneck in your flow somewhere, and how far the data has got through that pipeline.

Another option to check that data has fully passed through your flow is to check out data provenance to see the full history of your data.

Viewing data lineage in Apache Atlas

Learn how you can follow your data lifecycle conveying the origin of data and where it moves over time. You may view dataset level lineage graphs in the Atlas UI.

When deploying a NiFi cluster in CDP Public Cloud, a reporting task is automatically configured that sends lineage information about your flows into Apache Atlas. In the case of the example described in this simple use case, you can confirm in the Atlas UI that the data is correctly collected.
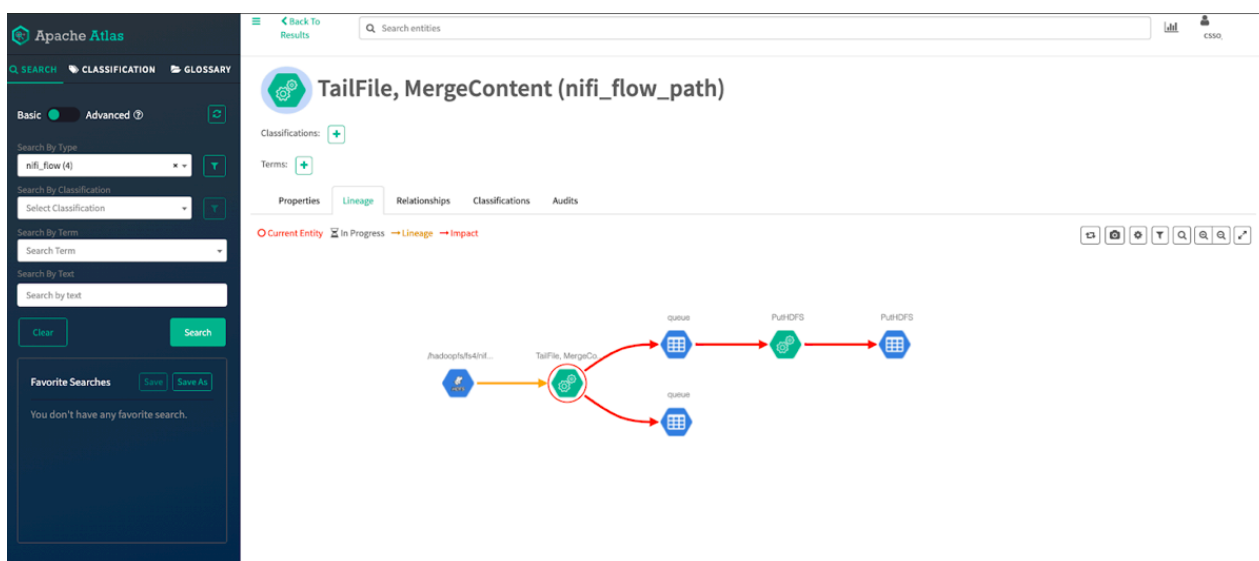
Related Information
[Viewing data lineage in Apache Atlas](#)

Next steps

Learn about the different options that you have after building a simple Google Cloud Storage ingest data flow in CDP Public Cloud.

Moving data to the cloud is one of the cornerstones of any cloud migration. Cloud environments offer numerous deployment options and services. This example data flow provides you with a model to design more complex data flows for moving and processing data as part of cloud migration efforts.

You can build a combination of on-premise and public cloud data storage. You can use this solution as a path to migrate your entire data to the cloud over time—eventually transitioning to a fully cloud-native solution or to extend your existing on-premise storage infrastructure, for example for a disaster recovery scenario. Cloud storage can provide secure, durable, and extremely low-cost options for data archiving and long-term backup for on-premise datasets.

You can also use cloud services without storing your data in the cloud. In this case you would continue to use your legacy on-premise data storage infrastructure, and work with on-demand, cloud-based services for data transformation, processing and analytics with the best performance, reliability and cost efficiency for your needs. This way you can manage demand peaks, provide higher processing power, and sophisticated tools without the need to permanently invest in computer hardware.