Cloudera DataFlow for Data Hub 7.2.17

Ingesting Data into Apache Solr

Date published: 2019-12-16
Date modified: 2023-06-27

https://docs.cloudera.com/

Legal Notice

Contents

Ingesting data into Solr

You can use an Apache NiFi data flow to ingest data into Apache Solr in the CDP Public Cloud following these steps.

Understand the use case

Learn how you can use Apache NiFi to move data to Apache Solr in your Data Discovery and Exploration (DDE) cluster in CDP Public Cloud.

Time series use cases analyze data obtained during specified intervals and can improve performance based on available data. These require that you store your documents efficiently in your search engine. You can use Apache NiFi to move data from a range of locations to the Solr service in your Data Discovery and Exploration (DDE) cluster in CDP Public Cloud.

DDE clusters provide an efficient way to have Apache Solr up and running to store and index documents that can be queried by external applications. Beyond search queries, this also allows for building insight applications or exploration services based on text or other unstructured data. For more information on how to use DDE templates in CDP Public Cloud, see the *Discover and Explore Data Faster with the CDP DDE Template* blog post.

This data ingest use case walks you through the steps associated with creating a data flow that moves data from Kafka in a Streams Messaging cluster in CDP Public Cloud into Solr in a DDE cluster, in the same CDP Public Cloud environment. This gets you started with creating an ingest data flow into Solr. If you want to move data from a location other than Kafka, review the *Getting Started with Apache NiFi* for general information about how to build a data flow and about additional data ingest processor options. You can also check the other CDP Public Cloud ingest use cases.

Related Information

Discover and explore data faster with the CDP DDE template

Getting started with Apache NiFi

CDP Public Cloud ingest use cases

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 service account with its username and password set to access Data Hub clusters.

  The predefined resource role of this service account 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 in your CDP Public Cloud environment running Apache NiFi.
- Your CDP user has been added to the appropriate pre-defined Ranger access policies to allow access to the NiFi UI.
- You have a Data Discovery and Exploration cluster running Solr and Hue. This cluster must be in the same CDP environment as the Flow Management cluster.

Related Information

Understanding roles and resource roles

Setting up your Flow Management cluster

Data Discovery and Exploration clusters

Create Solr target collection

Learn how you can create collections in Solr. Data is stored in collections, which are single indexes distributed across multiple servers.

Before you begin

Before you can ingest data into Apache Solr, you need a target Solr collection prepared to receive the data. For more information about Solr, see the *Cloudera Search Overview*.

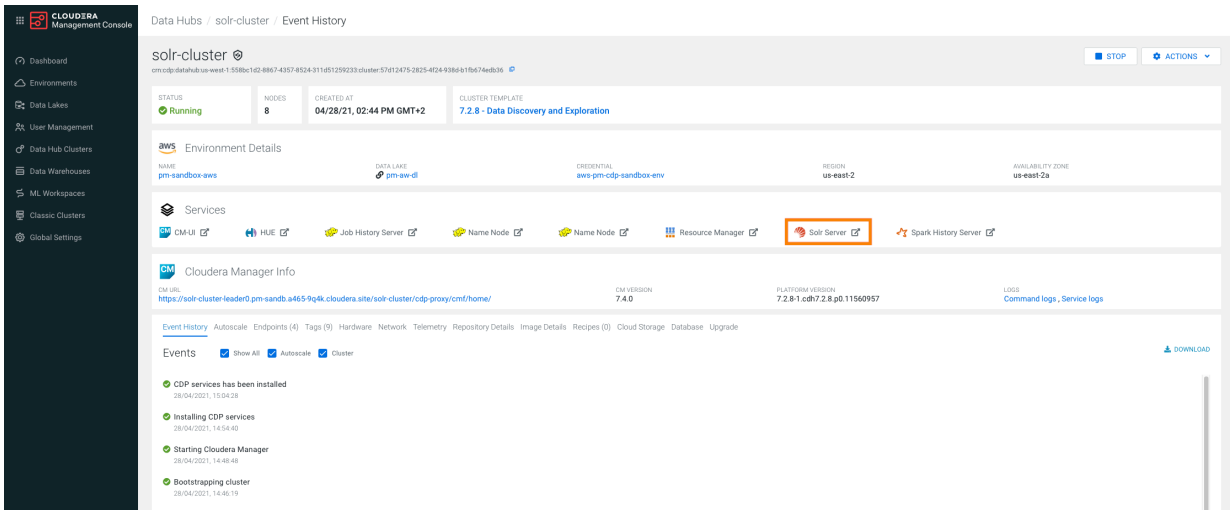
Procedure

1. Navigate to your Data Discovery and Exploration (DDE) cluster and click Hue in the Services section of the DDE cluster overview page.
2. On the Hue web UI, select Indexes+ Create index.
3. From the Type drop-down, select Manually and click Next.
4. Provide a collection name under Destination.

   In this example, we named the collection solr-nifi-demo.
5. Add the following fields, using + Add Field.

Name	Type
name	text_general
initial_release_date	date

6. Click Submit.
7. To check that the collection has been created, go to the Solr web UI by clicking the Solr Server shortcut in the Services section of the DDE cluster overview page.



8. Execute the collection query.
   a) Click the Collections sidebar option or click Select an option.

      In the drop-down you will find the collection you have just created. In our example it is solr-nifi-demo.
   b) Click the name of the collection.
   c) Click QueryExecute Query.

      The query output should look like this:

```
{
  "responseHeader":{
    "zkConnected":true,
    "status":0,
```

```
    "QTime":0,
    "params":{
      "q":"*:*",
      "doAs":"<querying user>",
      "_forwardedCount":"1",
      "_":"1599835760799"}},
  "response":{"numFound":0,"start":0,"docs":[]
}}
```

Results

You have successfully created an empty collection.

Related Information

Cloudera Search overview

Build the data flow

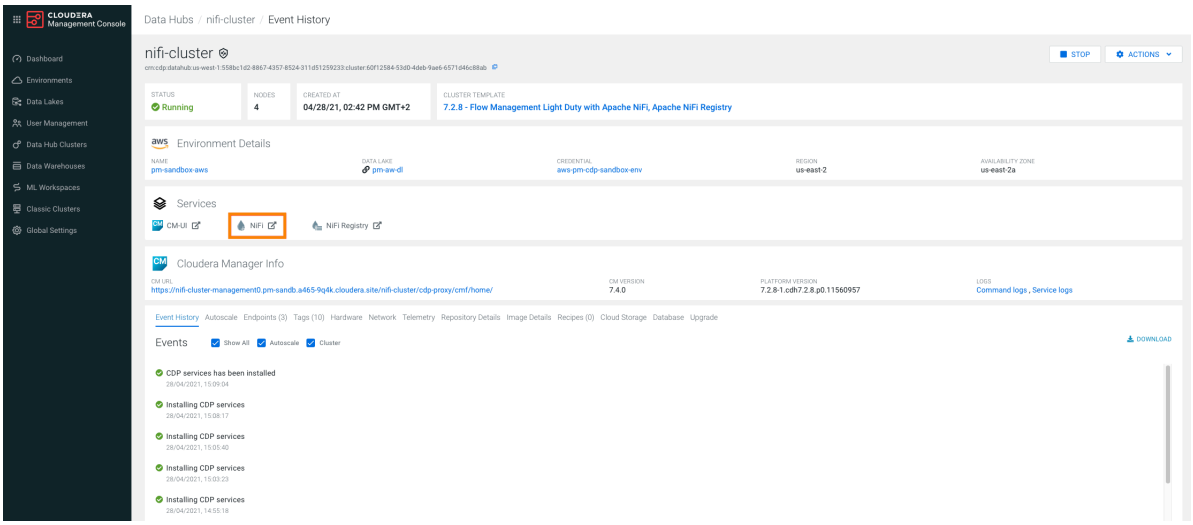
Learn how you can create an ingest data flow to move data from Apache Kafka to Apache Solr. This involves adding processors and other data flow elements to the NiFi canvas, configuring them, and connecting the elements to create the data flow.

About this task

Use the `PutSolrRecord` processor to build your Solr ingest data flow.

Procedure

1. Open NiFi
   in
   CDP Public Cloud.
   a) To access the NiFi service in your Flow Management cluster, navigate to Management ConsoleData Hub Clusters.
   b) Click the tile representing the Flow Management 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.

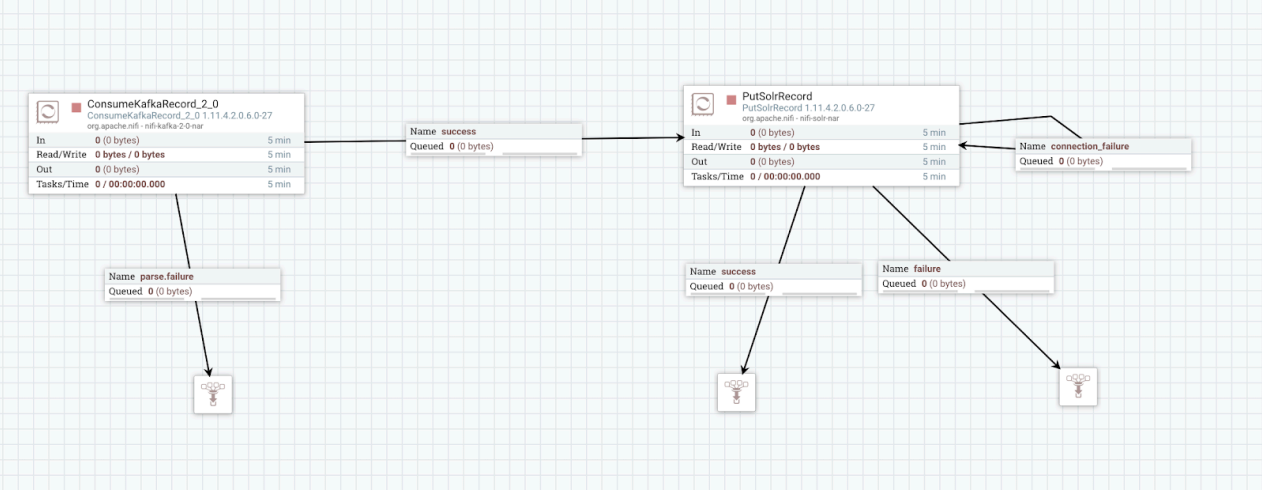
2. Add the NiFi processors to your canvas.
   a) Select the processor icon from the Cloudera Flow Management actions pane, and drag it to the canvas.
   b) Use the Add Processor filter box to search for the processors you want to add.
   c) Select the following processor and click Add.
      - `ConsumeKafkaRecord_2_0`
      - `PutSolrRecord`
3. Connect the two processors to create the data flow.
   a) Click the connection icon in the first processor, and drag it to the second processor.

      You can see the Create Connection dialog 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.
   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 data flow, which helps you to see where flow files are routed when your data flow is running.

      For more information about working with funnels, see the *Apache NiFi Documentation*.

Results

Your data flow may look similar to the following:



What to do next

Create a controller service for your data flow. You will need this service later as you configure your data flow target processor.

Related Information

Apache NiFi documentation

Create controller services for your data flow

Learn how you can create and configure controller services to provide shared services for the processors in your Solr ingest data flow. You will use these controller services later when you configure your processors.

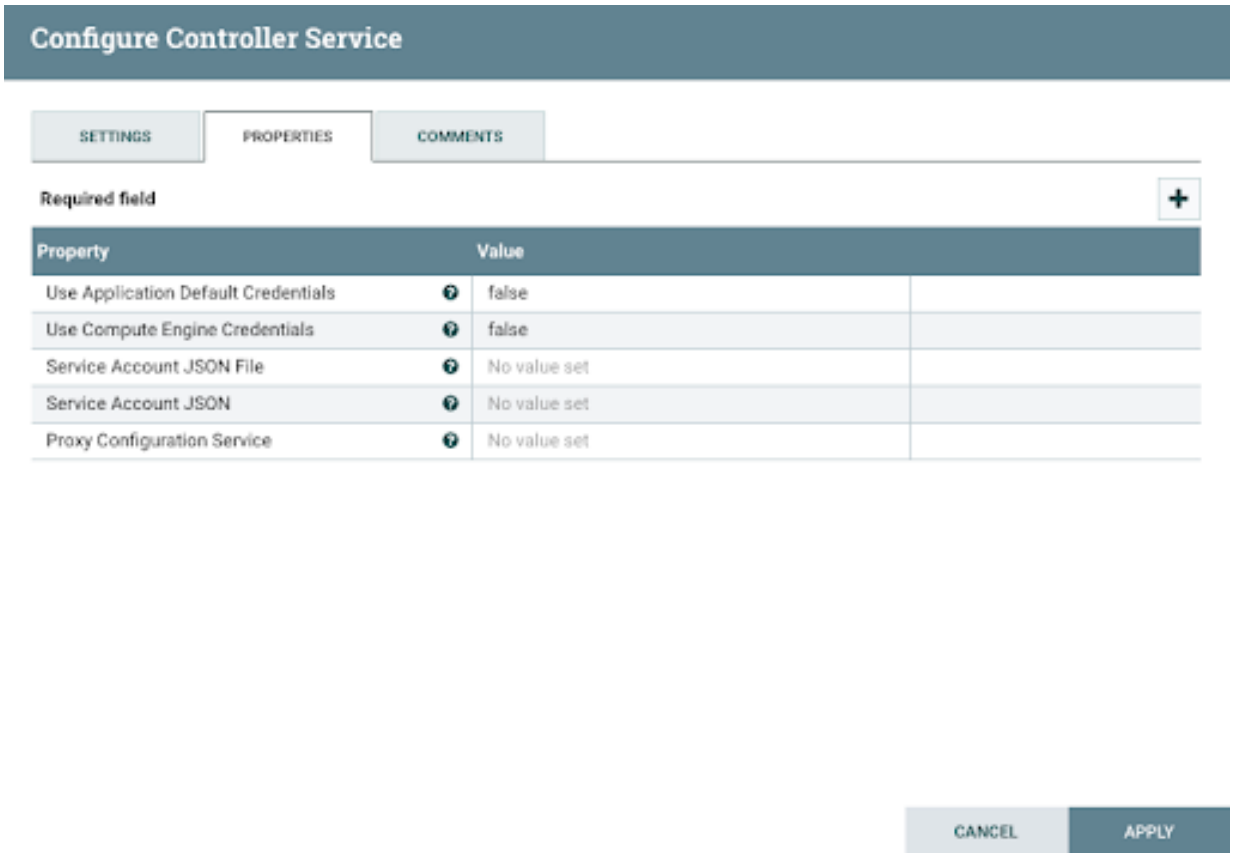
About this task

Note: You must define the controller services for the processors in your data flow in the configuration of the root process group where they will be used.

For more information on controller services in NiFi, see *Adding Controller Services for Data Flows*.

Procedure

1. To add a controller service to your data flow, right-click the canvas and select Configure from the pop-up menu.
   You can see the Controller Services Configuration window for the process group you are in.
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.
   For this use case, add the RecordReader and RecordWriter controller services to support record-based processing in your data flow.
5. Perform any necessary controller service configuration tasks by clicking the Configure icon in the right-hand column.



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

Configure the processor for your data source

You can set up a data flow to move data to Apache Solr from many different locations. This example assumes that you are transferring data from Kafka using the `ConsumeKafkaRecord_2_0` processor.

Before you begin

- Ensure that you have the Ranger policies required to access the Kafka consumer group.
- This use case demonstrates a data flow using data in Kafka and moving it to Solr. To review how to ingest data to Kafka, see *Ingesting Data to Apache Kafka*.

Procedure

1. Launch the Configure Processor window, by right clicking the `ConsumeKafkaRecord_2_0` processor and selecting Configure.

   You can see 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.
   a) Click the Properties tab.
   b) Configure `ConsumeKafkaRecord_2_0` with the required values.

   The following table includes a description and example values for the properties required for this Kafka to Solr ingest data flow. For a complete list of `ConsumeKafkaRecord_2_0` options, see the processor documentation in *Getting Started with Apache NiFi*.

Property	Description	Value to set for ingest to Solr data flow
Kafka Brokers	Provide a comma-separated list of known Kafka Brokers in the format <host>:<port>.	Docs-messaging-broker1.cdf-docs.a465-9q4k.cloudera.site:9093, docs-messaging-broker2.cdf-docs.a465-9q4k.cloudera.site:9093, docs-messaging-broker3.cdf-docs.a465-9q4k.cloudera.site:9093
Topic Name(s)	Enter the name of the Kafka topic from which you want to get data.	customer
Topic Name Format	Specify whether the topics are a comma separated list of topic names or a single regular expression.	names
Record Reader	Specify the record reader to use for incoming flowfiles. This can be a range of data formats.	AvroReader
Record Writer	Specify the format you want to use to serialize data before for sending it to the data target. **Note:** Ensure that the source record writer and the target record reader are using the same schema.	AvroRecordSetWriter
Honor Transactions	Specify whether or not NiFi should honor transactional guarantees when communicating with Kafka.	false
Security Protocol	Specify the protocol used to communicate with brokers. This value corresponds to Kafka's 'security.protocol' property.	SASL_SSL
SASL Mechanism	Specify the SASL mechanism to be used for authentication. This value corresponds to Kafka's 'sasl.mechanism' property.	plain
Username	Specify the username when the SASL Mechanism is PLAIN or SCRAM-SHA-256.	srv_nifi-solr-ingest
Password	Specify the password for the given username when the SASL Mechanism is PLAIN or SCRAM-SHA-256.	Password1!
SSL Context Service	Specify the SSL Context Service to be used for communicating with Kafka.	default

Property	Description	Value to set for ingest to Solr data flow
Group ID	Specify the Group ID used to identify consumers that are within the same consumer group. This value corresponds to Kafka's grou p.id property.	nifi-solr-ingest

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

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

   If you want to move data to Solr from a location other than Kafka, see Getting Started with Apache NiFi for general information about how to build a data flow and about other data consumption processor options. You can also check the other *CDP Public Cloud ingest use cases* for more information.

What to do next
Configure the processor for your data target.

Related Information
Ingesting data into Apache Kafka
Getting started with Apache NiFi
CDP Public Cloud ingest use cases

Configure the processor for your data target

You can set up a data flow to move data to many locations. This example assumes that you are moving data to Apache Solr in a Data Discovery and Exploration cluster in CDP Public Cloud, using the `PutSolrRecord` processor.

Before you begin
Create a machine user in CDP User Management and synchronize this user to your CDP environment.

Procedure

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

   You can see 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.
   a) Click the Properties tab.
   b) Configure `PutSolrRecord` with the required values.

   The following table includes a description and example values for the properties required for this Kafka to Solr ingest data flow. For a complete list of `PutSolrRecord` options, see the processor documentation in *Getting Started with Apache NiFi*.

Property	Description	Value for example Solr ingest data flow
Solr Type	Provide the type of Solr instance. It can be Cloud or Standard.	Cloud
Solr Location	Specify the Solr URL for a StandardSolr type. For example:http://localhost:8984/solr/gettingstarted Specify the ZooKeeper hosts for a Cloud Solr type. For example: localhost:9983). You can find this value on the dashboard of the Solr web UI as the zkHost parameter value.	zookeeper1.cloudera.site:2181,zookeeper2 .cloudera.site:2181,zookeeper3.cloudera.site :2181/solr-dde
Collection	Specify the name of the Solr collection.	solr-nifi-demo

Property	Description	Value for example Solr ingest data flow
Solr Update Path	Provide the path in Solr where flowfile records are posted.	/update
Kerberos Principal	Specify the CDP user name you are using to perform this workflow.	Provide the CDP user name you created and synced with your CDP environment in *Meet the prerequisites*.
Kerberos Password	Specify the password for the CDP user you are using to perform this workflow.	Password1!
RecordReader	Specify the service for reading records from incoming flowfiles.	AvroReader
SSL Context Service	Specify the controller service to use to obtain an SSL context. This property must be set when communicating with Solr over https. Reference the default SSL context service that has been created for you.	In this example the value is: Default NiFi SSL Context Service

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.

   If you want to move data between other locations, see Getting Started with Apache NiFi for general information about how to build a data flow and about other data consumption processor options. You can also check the other *CDP Public Cloud ingest use cases* for more information.

What to do next
Your data flow is ready to ingest data into Solr. Start the data flow.

Related Information
Getting started with Apache NiFi

Meet the prerequisites

Start the data flow

When you have built and configured your data flow, you can begin ingesting data into Solr. Learn how to start your Solr ingest data flow.

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
Data is being read from Kafka and it is written to the Solr target collection you have created.

What to do next
It is useful to check that data is running through the flow without any issues.

Verify data flow operation

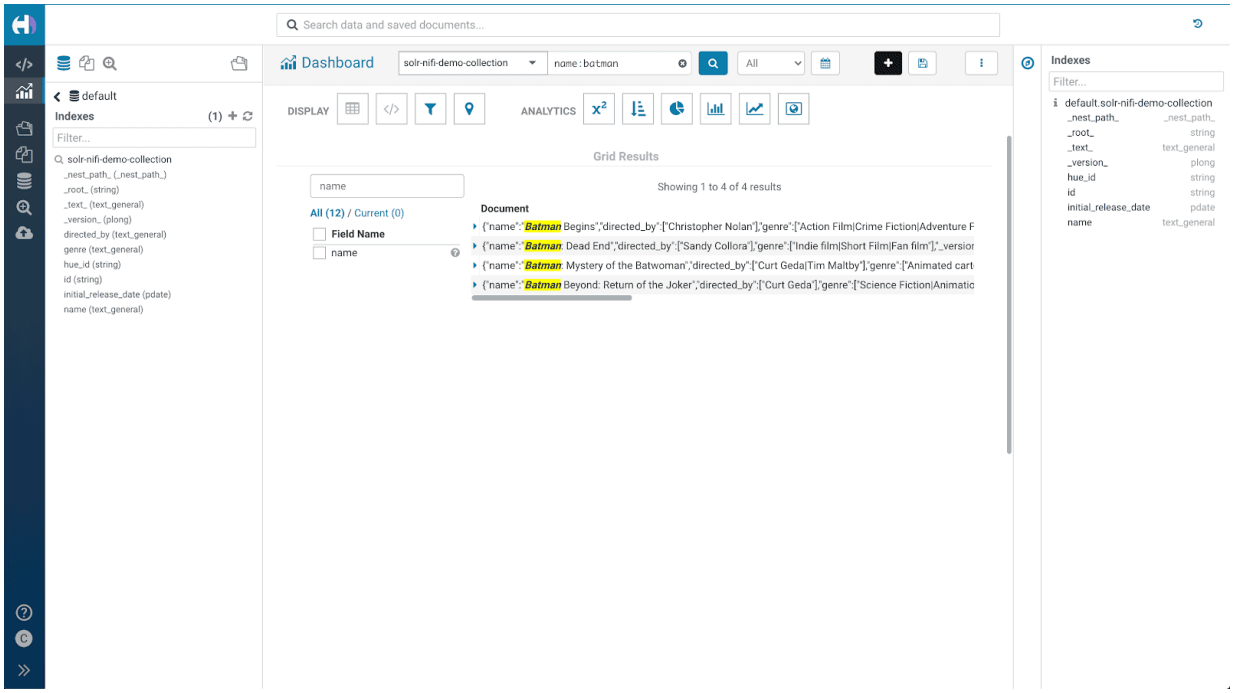
Once you have configured and started your data flow, you can verify that you can write data to Solr. There are a number of ways you can check that data is running through the flow and it actually appears in Solr.

Procedure

- Review the Solr ingest data flow on the NiFi canvas and confirm that the NiFi processors are not producing errors.

- From your Data Discovery and Exploration (DDE) cluster, go to the Hue web UI and search for your data ingest target collection.

  You should see data being ingested into your collection:



Monitoring your data flow

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. Learn about the different monitoring options for your Solr ingest data flow in CDP Public Cloud.

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.

Next steps

Learn about the different options on what you can do once you have moved data into Solr in CDP Public Cloud.

This example data flow shows a basic setup, but it also offers endless opportunities to implement more complex solutions for moving and processing data in Solr. You can explore different data ingest and indexing options where Cloudera Data Platform (CDP) components collaborate with each other, while still being resource isolated and managed separately.

For more information on data flows and Solr in CDP Public Cloud:

- Review the *NiFi documentation* to learn more about building and managing data flows.

- Review the *NiFi Registry documentation* for more information about versioning data flows.
- Review the *Cloudera Search Overview*.

Related Information

Apache NiFi documentation
Apache NiFi Registry documentation
Cloudera Search overview