

Tutorial: Customize a ReadyFlow

Date published: 2021-04-06

Date modified: 2024-06-03

CLOUDERA

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Tutorial: customizing a ReadyFlow.....	4
Open the Hello World ReadyFlow as a template for your draft.....	4
Create new services.....	5
Swap the controller services in your draft.....	7
Start a test session.....	9
Publish your flow definition to the Catalog.....	11

Tutorial: customizing a ReadyFlow

Learn how to create a draft using a ReadyFlow as a template.

About this task

This tutorial shows you how you can open a ReadyFlow from the ReadyFlow Gallery and create a customized flow design using the ReadyFlow as a template. This tutorial uses the 'Hello World' ReadyFlow, a simple flow design that retrieves the latest changes from Wikipedia through invoking the Wikipedia API. It converts JSON events to Avro, before filtering and routing them to two different processors which merge events together before a file is written to local disk.

You will learn about:

- Opening a ReadyFlow as a draft on the Flow Design Canvas.
- Creating a Controller Service
- Changing the configuration of processors
- Running a Test Session
- Publishing a draft to the Catalog as a flow definition.

Before you begin

The 'Hello World' ReadyFlow that you are about to customize can be deployed without any external dependencies and does not require any parameter values during deployment. Still, there are prerequisites you have to meet before you can start building your first draft.

- You have an enabled and healthy Cloudera DataFlow environment.
- You have been assigned the DFDeveloper role granting you access to the Flow Designer.
- You have been assigned the DFCatalogAdmin or DFCatalogViewer role granting you access to the Catalog. You will need this authorization to publish your draft as a flow definition to the Catalog.
- You have been assigned the DFFlowAdmin role for the environment to which you want to deploy the flow definition.

Open the Hello World ReadyFlow as a template for your draft

About this task


ReadyFlows are read-only, therefore you cannot publish a new version of them. When you are done with creating your customized version of the given ReadyFlow, you can only publish your work to the Catalog as a new flow definition.

Before you begin

- You have an enabled and healthy Cloudera DataFlow environment.
- You have been assigned the DFDeveloper role granting you access to the Flow Designer.
- You have been assigned the DFCatalogAdmin or DFCatalogViewer role granting you access to the Catalog. You will need this authorization to publish your draft flow as a flow definition to the Catalog.
- You have been assigned the DFFlowAdmin role for the environment to which you want to deploy the flow definition.

Procedure

1. Open Cloudera DataFlow by clicking the DataFlow tile in the Cloudera sidebar.

2. Select  ReadyFlow Gallery in the left navigation pane.
3. Select the Hello World ReadyFlow.
4. Click Create New Draft.
5. Select the target Workspace where you want to create the draft flow.
6. Provide a valid Flow Name for the draft flow.
Flow names must be unique within a workspace. If a draft with the provided name already exists, you need to provide a different name.
7. Click Create.
Hello World opens as a draft in the designated Flow Designer workspace with the Flow Name you provided on the **Flow Design** canvas.

What to do next

Proceed with creating the necessary Controller Services.

Related Tasks

[Create new services](#)

Create new services

Learn about creating Controller Services in Cloudera DataFlow Flow Designer.

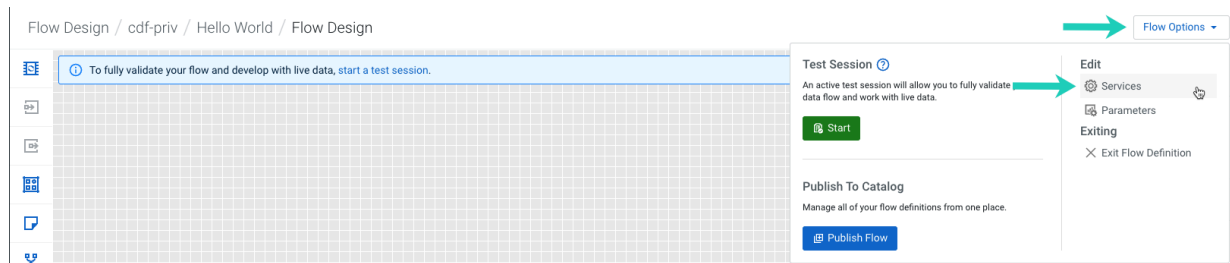
About this task

Services are extension points that provide information for use by other components (such as processors or other controller services). The idea is that, rather than configure this information in every processor that might need it, the service provides it for any processor to use as needed.

You will use the services you create now to configure the behavior of several processors you will add to your flow as you are building it.

Procedure

1. Go to Flow Options Services .



- Click Add Service.
The Add Service page opens.

Add Service ✕

- AccumuloService >
- ActionHandlerLookup
- ADLSCredentialsControllerService
- ADLSIDBrokerCloudCredentialsProviderCont...
- AlertHandler
- AvroReader
- AvroRecordSetWriter
- AvroSchemaRegistry
- AWSCredentialsProviderControllerService
- AWSIDBrokerCloudCredentialsProviderContr...
- AzureBlobIDBrokerCloudCredentialsProvider...
- AzureCosmosDBClientService
- AzureEventHubRecordSink
- AzureStorageCredentialsControllerService
- AzureStorageCredentialsControllerService_v...

Service Name

Type

AccumuloService

IMPLEMENTS SERVICE

BaseAccumuloService 1.18.0.2.3.7.0-64 from nifi-accumulo-services-api-nar

VERSION	GROUP
1.18.0.2.3.7.0-64	org.apache.nifi

BUNDLE

nifi-accumulo-services-nar

DESCRIPTION

A controller service for accessing an Accumulo Client.

TAGS

accumulo, service, client

- In the text box, filter for CSVReader.
- Provide Service Name: CSVReader_Recent_Changes.
- Click Add.
- Click Add Service to create another service.
- In the text box, filter for CSVRecordSetWriter.
- Provide Service Name: CSVRecordSetWriter_Recent_Changes.
- Click Add.

You do not need to configure the CSVReader_Recent_Changes service. You can leave all properties with their default values.

You do not need to configure the CSVRecordSetWriter_Recent_Changes service. You can leave all properties with their default values.

- Select the AvroReader_Recent_Changes service and check the list of Referencing Components in the Service Details pane on the right.

Make a note of the listed components, because that is where you will need to replace AvroReader_Recent_Changes with the CSVReader_Recent_Changes service.

11. Select the AvroWriter_Recent_Changes service and check the list of Referencing Components in the Service Details pane on the right.

Make a note of the listed components, because that is where you will need to replace AvroWriter_Recent_Changes with the CSVRecordSetWriter_Recent_Changes service.

State	Name ↑	Type	Notifications
🔍	AvroReader_Recent_Changes	AvroReader V.1.20.0.2.3.8.0-21	>
🔍	AvroWriter_Recent_Changes	AvroRecordSetWriter V.1.20.0.2.3.8.0-21	>
🔍	CSVReader_Recent_Changes	CSVReader V.1.20.0.2.3.8.0-21	>
🔍	CSVRecordSetWriter_Recent_Changes	CSVRecordSetWriter V.1.20.0.2.3.8.0-21	>
🔍	JSON_Reader_Recent_Changes	JsonTreeReader V.1.20.0.2.3.8.0-21	>

Properties

Property	Value
Schema Access Strategy ⓘ	Use Embedded Avro Schema ⋮
Cache Size ⓘ	1000 ⋮

Referencing Components

State	Referencing Processor	Path
⊗	Filter Edits QueryRecord V.1.20.0.2.3.8.0-21	🔗
⊗	Merge Edit Events MergeRecord V.1.20.0.2.3.8.0-21	🔗
⊗	Merge Edit Events MergeRecord V.1.20.0.2.3.8.0-21	🔗
⊗	Route On Content Size QueryRecord V.1.20.0.2.3.8.0-21	🔗

12. Click Back To Flow Designer to return to the flow design Canvas.

What to do next

After creating the necessary services, you can start customizing your flow.

Related Tasks

[Swap the controller services in your draft](#)

Swap the controller services in your draft

Customize your draft by replacing services to change the destination file format.

Procedure

- 1. Select the Convert JSON TO AVRO processor by clicking on it.

2.

In the Configuration pane scroll down to Properties, click the



drop-down next to Record Writer and select

CSVRecordSetWriter_Recent_Changes.

3. Click OK.

4. Rename the processor from Convert JSON to AVRO to Convert JSON to CSV.

5. Click Apply.

6. Select the Filter Edits processor by clicking on it.

7.

In the Configuration pane scroll down to Properties, click the



drop-down next to Record Reader and select

CSVReader_Recent_Changes.

8. Click OK.

9.

Click the



drop-down next to Record Writer and select the CSVRecordSetWriter_Recent_Changes service.

10. Click OK.

11. Click Apply.

12. Proceed with updating the remaining Record Reader and Record Writer references.

You need to modify the configurations of the Route on Content Size and the two Merge Edit Events processors.

13. Check that you have updated all the processors.

- a) Select Flow Options Services .
- b) Select the CSVReader_Recent_Changes service and check Referencing Components.

You should see:


- Filer Edits
- Merge Edit Events
- Merge Edit Events
- Route On Content Size

- c) Select the CSVRecordSetWriter_Recent_Changes service and check Referencing Components.

You should see:

- Convert JSON to AVRO
- Filer Edits
- Merge Edit Events
- Merge Edit Events
- Route On Content Size

- d) If you check the AvroReader_Recent_Changes and AvroWriter_Recent_Changes services, you should see No referencing Processors to display.

14. Delete the AvroReader_Recent_Changes and AvroWriter_Recent_Changes services by selecting the service and clicking  Delete.

In the confirmation pop-up select Delete.

15. Click Back To Flow Designer to return to the flow design Canvas.

What to do next

Start a Test Session to validate your draft.

Start a test session

To validate your draft, start a test session. This provisions an Apache NiFi cluster where you can test your draft.

About this task

Starting a Test Session provisions NiFi resources, acting like a development sandbox for a particular draft. It allows you to work with live data to validate your data flow logic while updating your draft. You can suspend a test session any time and change the configuration of the NiFi cluster then resume testing with the updated configuration.

Procedure

1. Click start a test session in the banner on top of the Canvas.

 To fully validate your flow and develop with live data, start a test session.

2. Click Start Test Session.

Test Session status  Initializing Test Session... Initializing Test Session... appears on top of the page.

3. Wait for the status to change to


Active Test Session.



This may take several minutes.

4. Click Flow Options Services to enable Controller Services.


5.

Select a service you want to enable, then click  Enable Service and Referencing Components.

This option does not only enable the controller service, but also any component that references it. This way, you do not need to enable the component separately to run the test session. In the context of this tutorial, enabling the 'AvroReader_Recent_Changes' controller service will also enable 'Filter Edits', 'Route on Content Size', and 'Merge Edit Events' processors as well.

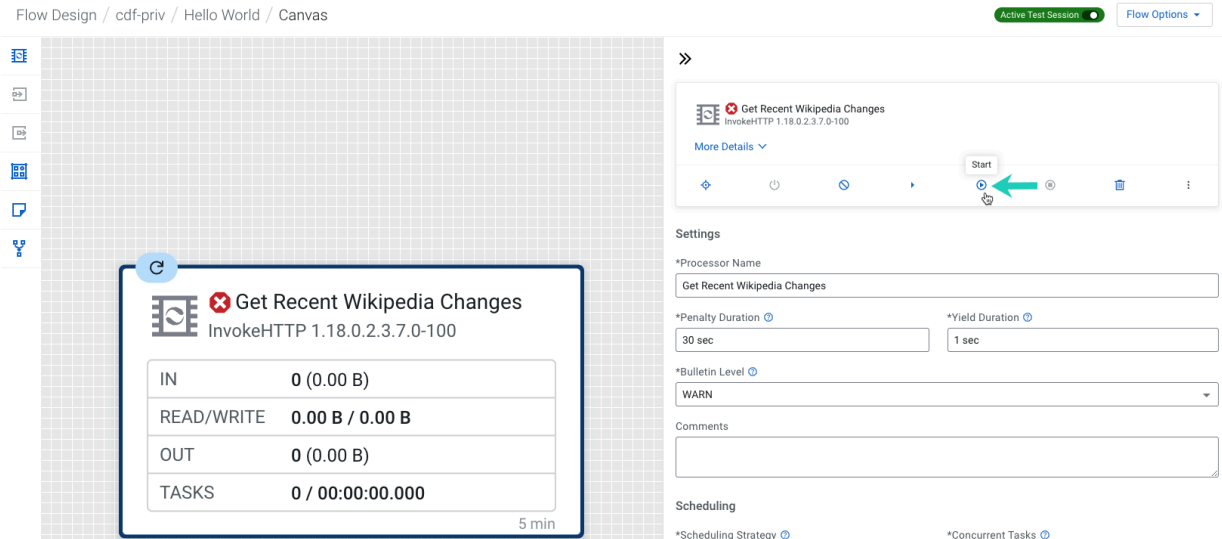
Repeat this step for all Controller Services.

6. Click Back To Flow Designer to return to the flow design Canvas.

7. Start the Get Recent Wikipedia Changes, Write "Added Content" Events To File, and Write "Removed Content" Events To File components by selecting them on the Canvas then clicking  Start.

Flow Design / cdf-priv / Hello World / Canvas

Active Test Session • Flow Options ▾



Get Recent Wikipedia Changes
InvokeHTTP 1.18.0.2.3.7.0-100

IN	0 (0.00 B)
READ/WRITE	0.00 B / 0.00 B
OUT	0 (0.00 B)
TASKS	0 / 00:00:00.000

5 min

Settings

*Processor Name
Get Recent Wikipedia Changes

*Penalty Duration 30 sec *Yield Duration 1 sec

*Bulletin Level WARN

Comments

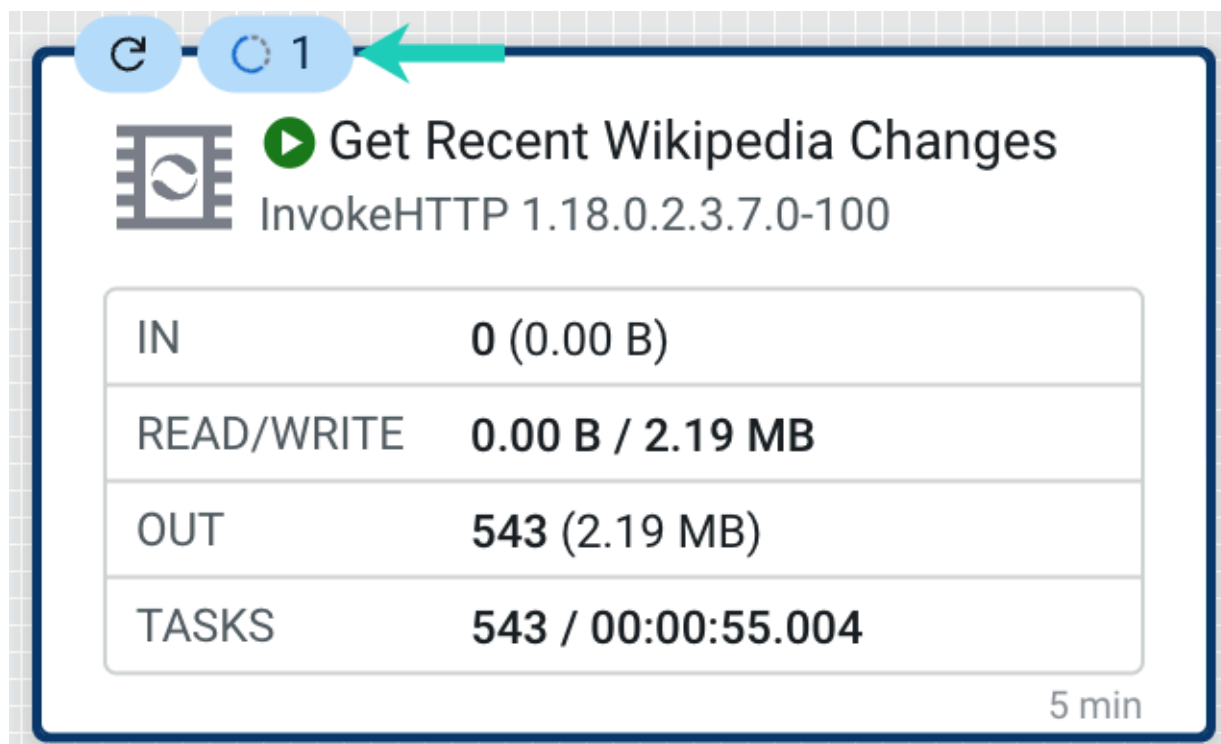
Scheduling

*Scheduling Strategy *Concurrent Tasks

All other components were auto-started when you selected the Enable Service and Referencing Components option.

8. Observe your first draft flow processing data.

On the Flow Design Canvas you can observe statistics on your processors change as they consume and process data from Wikipedia. You can also observe one or more blue Notification Pills, providing information about the current task.



Publish your flow definition to the Catalog

Now that you have tested your draft and it works fine, you can go on and publish it to the Catalog as a flow definition so that you can create a Cloudera DataFlow deployment.

Procedure

1. On the **Flow Designer** canvas, click **Flow Options Publish To Catalog Publish**.
2. Fill in the fields in the **Publish A New Flow** box.
 - Provide a Flow Name for your flow definition.
You can only provide a name when you publish your flow for the first time.
 - Optionally provide a Flow Description.
You can only provide a description when you publish your flow for the first time.
 - Optionally provide Custom Tags.
You can filter flow definition versions by tags in the Catalog.
 - Optionally provide Version Comments.
3. Click **Publish**.

Results

Your draft is published to the Catalog as a flow definition.

Related Information

[Deploying a flow definition](#)