Cloudera Data Flow

Quickstart

Date published: 2021-04-06
Date modified: 2025-09-30

https://docs.cloudera.com/

Legal Notice

Contents

Cloudera Data Flow Quickstart

Get started with Cloudera Data Flow quickly by walking through a few simple steps. Ensure that administrators have access to Cloudera on cloud, enable Cloudera Data Flow for an environment, give users access to Cloudera Data Flow, and then add to the Catalog and deploy the Hello World ReadyFlow.

Verify cloud infrastructure prerequisites

As the administrator for your environment, ensure that the environment meets the requirements for Cloudera on cloud and Cloudera Data Flow.

Deploy Cloudera using Terraform

Give administrators access

To enable Cloudera Data Flow for an environment, users must have the DFAdmin role. Grant the DFAdmin role to a user or group that is planned to be allowed to enable CDF for an environment.

Before you begin

- You must have the Cloudera PowerUser role.

Procedure

1. On the Cloudera Management Console page, click Environments in the left-hand sidebar.
2. Use the search field to find and select the Cloudera on cloud environment for which you want to grant DFAdmin rights.
3. Go to Actions Manage Access to display the Environments Access page.
4. Find the user or group to whom you want to grant the DFAdmin role, and click the Update Roles button.
5. On the **Update Resource Roles for** page, select the DFAdmin checkbox from the available roles and click the Update Roles button.
6. Click Synchronize Users.

   This propagates the changes you made to the selected environment.

What to do next

When you have finished granting a user or group administrator permissions, they can now proceed by enabling Cloudera Data Flow for an environment.

Related Information

Cloudera Data Flow Security

Enable Cloudera Data Flow for an environment

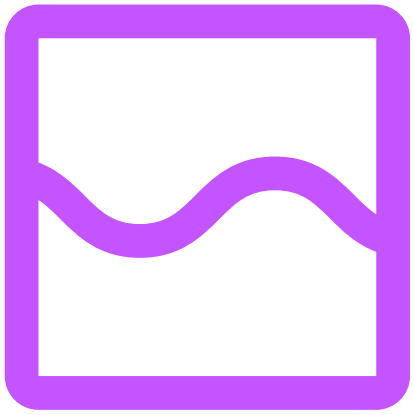
Before you can deploy flow definitions, you must enable Cloudera Data Flow for a Cloudera on cloud environment. Enabling Cloudera Data Flow for an environment means that you are preparing an active and healthy Cloudera on cloud environment for use with Cloudera Data Flow.

Before you begin

- You must have a cloud provider account and meet the infrastructure and network requirements.
- You must have a healthy Cloudera on cloud environment, with FreeIPA and the data lake running and healthy.

- You must have the DFAdmin role for the Cloudera on cloud environment for which you want to enable Cloudera Data Flow.

Procedure

1. 

   Navigate to Cloudera Data Flow, by selecting Data Flow from the Cloudera on cloud Home Page, or from the navigation pane.
2. Go to Environments, and click Enable to launch the **Enable Cloudera Data Flow Service** pane for the environment you want to enable.
3. In the **Enable Cloudera Data Flow Service** form, provide the following information:
   - Instance Type – From the available instance types, select the one appropriate for your use case. For testing and evaluation, the smallest c5.xlarge instance type with 4 vCPUs, 8 GB RAM is sufficient.
   - Cloudera Data Flow Capacity – Specify a minimum and a maximum size for the Kubernetes cluster. You can keep the default settings.
   - Networking
4. Click the Enable button. Enabling Cloudera Data Flow can take up to one hour.

What to do next

When you have finished enabling Cloudera Data Flow for an environment, proceed by giving users permission to import and deploy flow definitions.

Related Information

Enabling Cloudera Data Flow for an environment

Managing Cloudera Data Flow in an environment

Give users access

Cloudera Data Flow restricts who can import flow definitions and deploy them. To get started with Cloudera Data Flow, you must grant users permissions to perform these tasks so that they can import and deploy flow definitions in Cloudera Data Flow.

About this task

To get started with importing and deploying flow definitions, a user must have at least the following Cloudera Data Flow roles:

- DFCatalogPublisher – To import flow definitions to the Cloudera Data Flow Catalog

- DFFlowAdmin – To deploy flow definitions in Cloudera Data Flow

Before you begin

- You must have PowerUser role.
- You must know the user or group name to which you want to grant Cloudera Data Flow user access roles.

Procedure

1. Give a user permission to import flow definitions.
   a) From Cloudera Management Console, click User Management.
   b) Enter the name of the user or group you want to authorize in the Search field.
   c) Select the user or group from the list that displays.
   d) Go to Actions Update Roles to display the **Update Roles** modal window.
   e) Select the DFCatalogPublisher checkbox form the available roles and click the Update button.
2. Give a user or group permission to deploy flow definitions.
   a) From Cloudera Management Console, click Environments to display the **Environment / List** page.
   b) Select the environment where you want the user or group to deploy flow definitions.
   c) Go to Actions Manage Access to display the **Environments Access** page.
   d) Find the user or group o whom you want to grant the DFFlowAdmin role and click the Update Roles button.
   e) On the **Update Resource Roles for** modal window. select the DFFlowAdmin checkbox from the available roles.
   f) Click the Update Roles button.

What to do next

When you have finished giving user or groups permission to import and deploy flow definitions, proceed by importing a flow definition.

Related Information

Cloudera Data Flow Security

Add the Hello World ReadyFlow definition to the Catalog

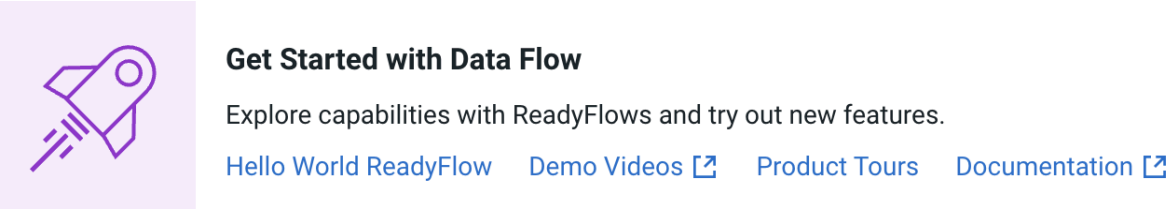
The Hello World ReadyFlow is a flow definition designed to work out of box without any further configuration, to make getting started with Cloudera Data Flow quick and easy. To use it, you must add this ReadyFlow to the Catalog.

Before you begin

- You must have an enabled and healthy Cloudera Data Flow environment.
- You must be assigned the DFCatalogPublisher role granting you access to the Catalog.
- You must be assigned the DFFlowAdmin role for the environment where you want to deploy the flow definition.

About this task

Note: In a new Cloudera Data Flow deployment with an empty Catalog, you can deploy Hello World directly from the **Overview** page.

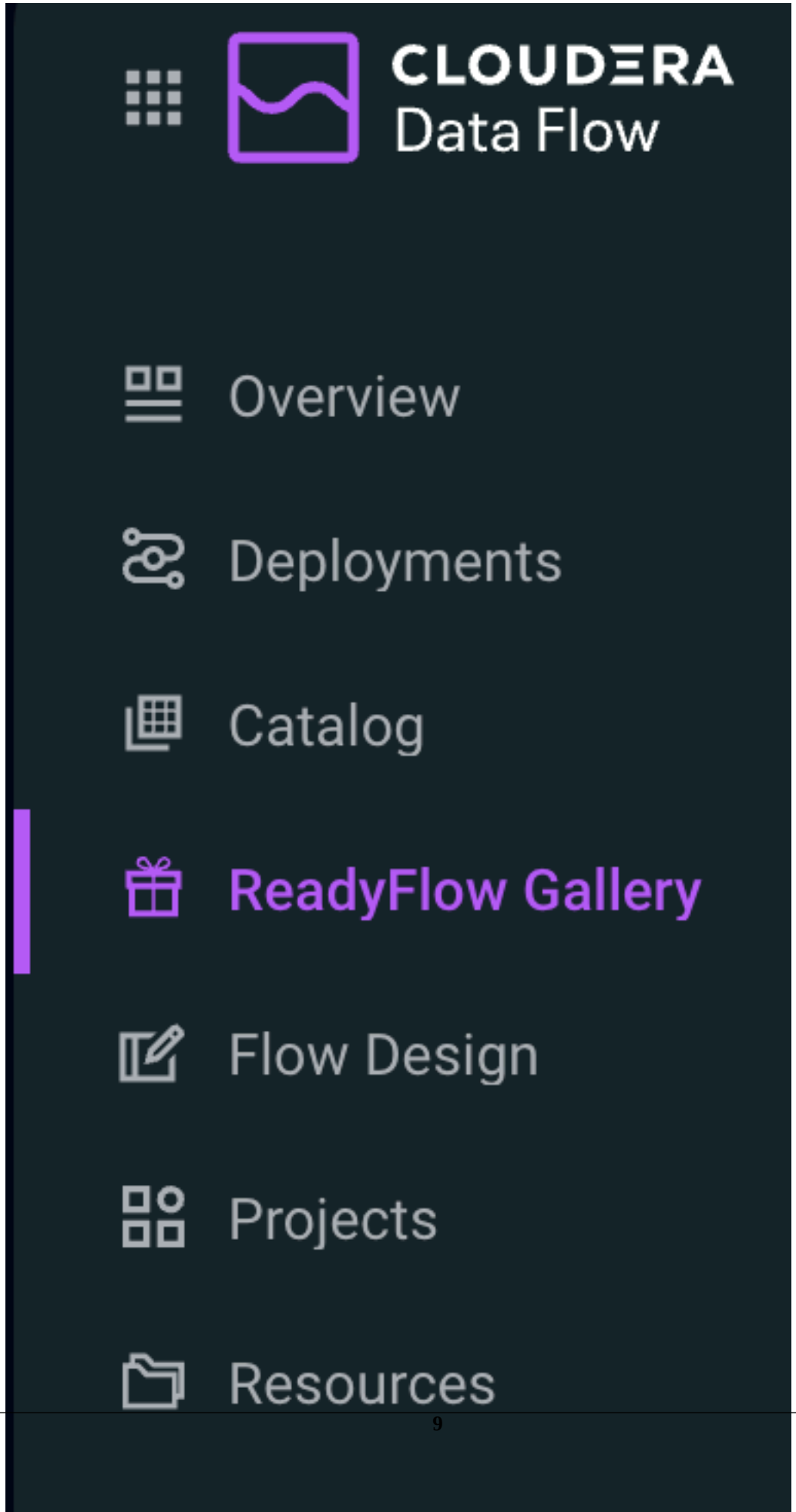


Just click the Hello World ReadyFlow option on the **Get Started with Data Flow** card on the **Overview** page. You are forwarded to the **Catalog**, from where you can deploy the ReadyFlow.

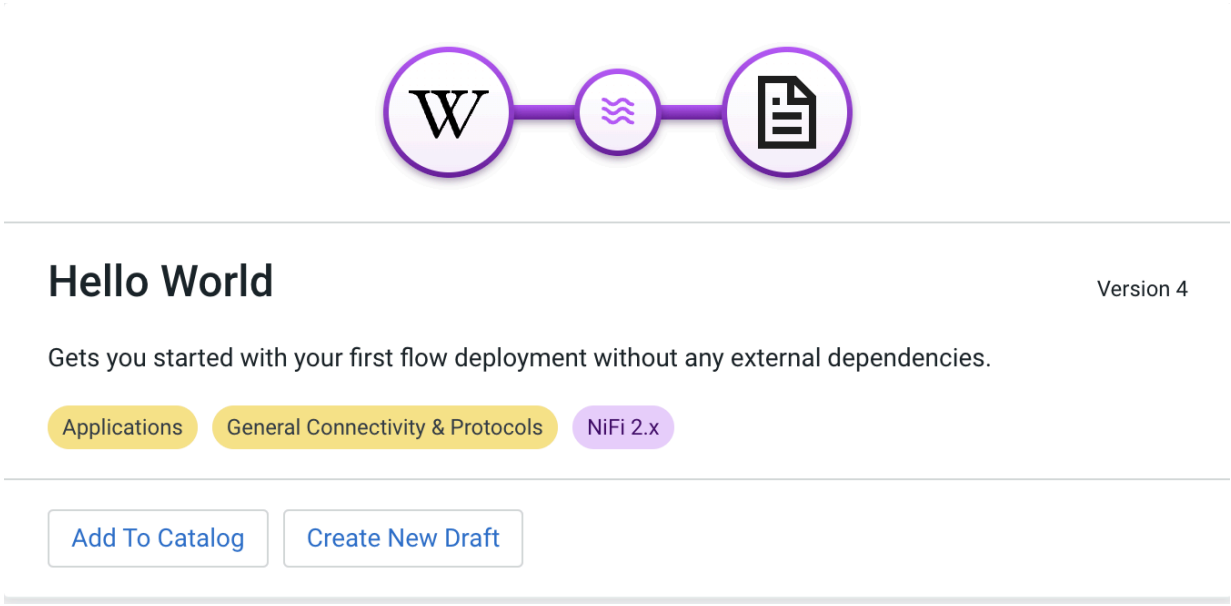
If the catalog already contains at least one flow definition, you can add Hello World from the ReadyFlow gallery.

Procedure

1. In the navigation pane, click the ReadyFlow Gallery item.


CLOUDERA
Data Flow
Overview
Deployments
Catalog
ReadyFlow Gallery
Flow Design
Projects
Resources

2. On the **ReadyFlow Gallery** pane, filter for and select the Hello World ReadyFlow.



Tip: Click anywhere on the card to review the ReadyFlow details.

3. Click the Add to Catalog button to add the ReadyFlow to the catalog and make it ready for deployment.
4. Click the Add button to confirm that you want to add Hello World to the catalog.

Results
The Hello World flow definition is added to the catalog and is ready for deployment.

What to do next
You can now proceed with deploying the ReadyFlow.

Deploy the Hello World ReadyFlow using the deployment wizard

Learn about the steps to deploy the Hello World ReadyFlow to get started with Cloudera Data Flow.

About this task

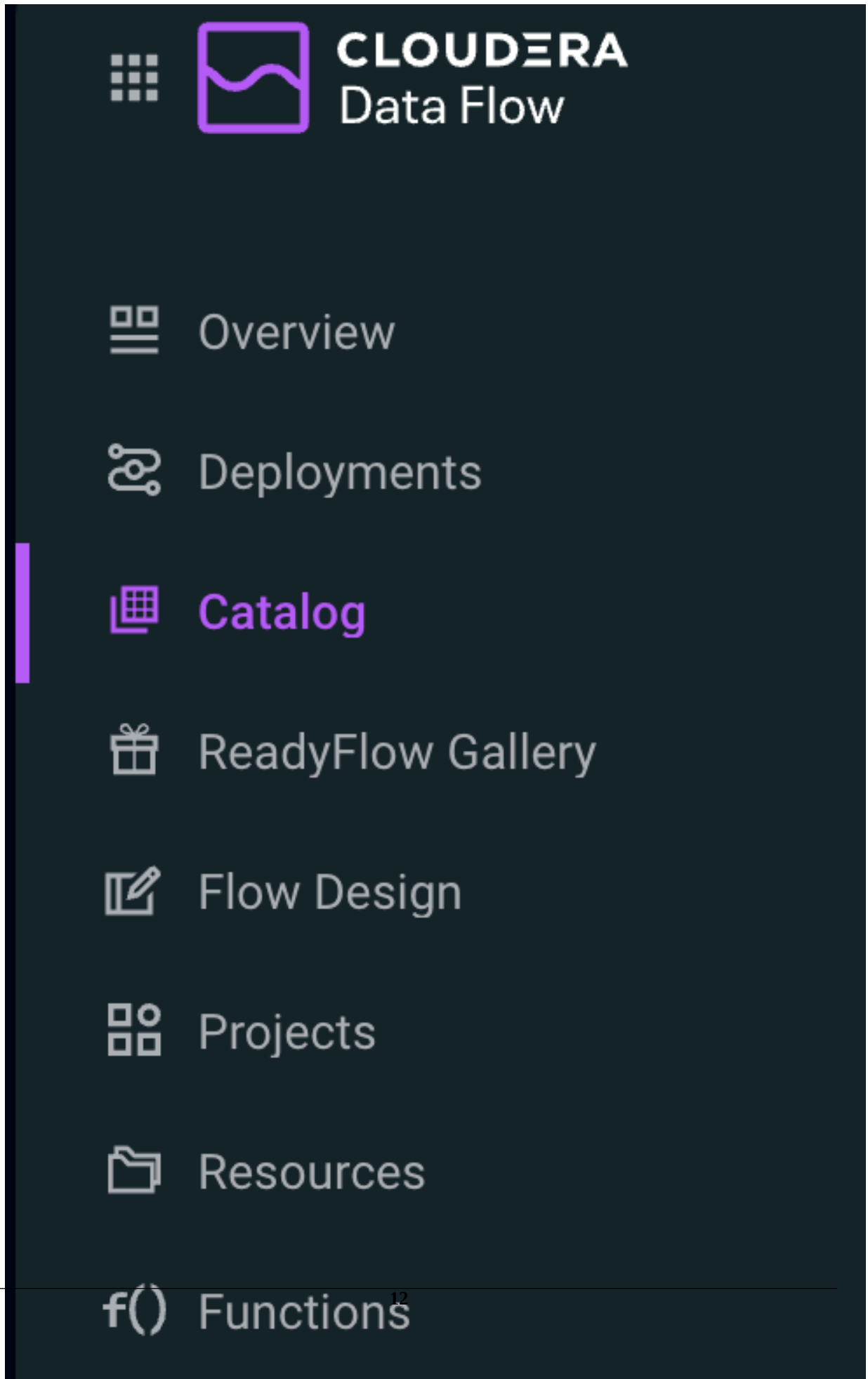
Once you have added the Hello World ReadyFlow into the Catalog, stay in the Catalog and use the Deployment wizard to deploy that flow definition.

Before you begin

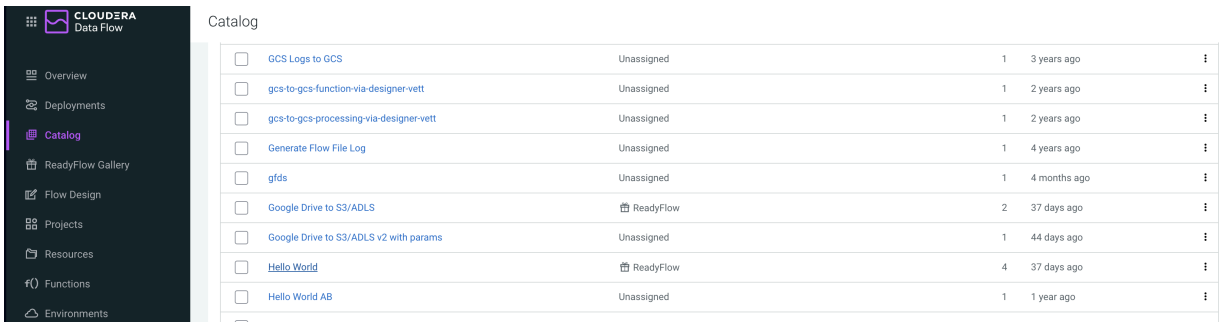
- You must have an enabled and healthy Cloudera Data Flow environment.
- You must be assigned the DFCatalogAdmin role granting you access to the Catalog.
- You must have added the Hello World flow definition to the Catalog.
- You must be assigned the DFFlowAdmin role for the environment where you want to deploy the flow definition.

Procedure

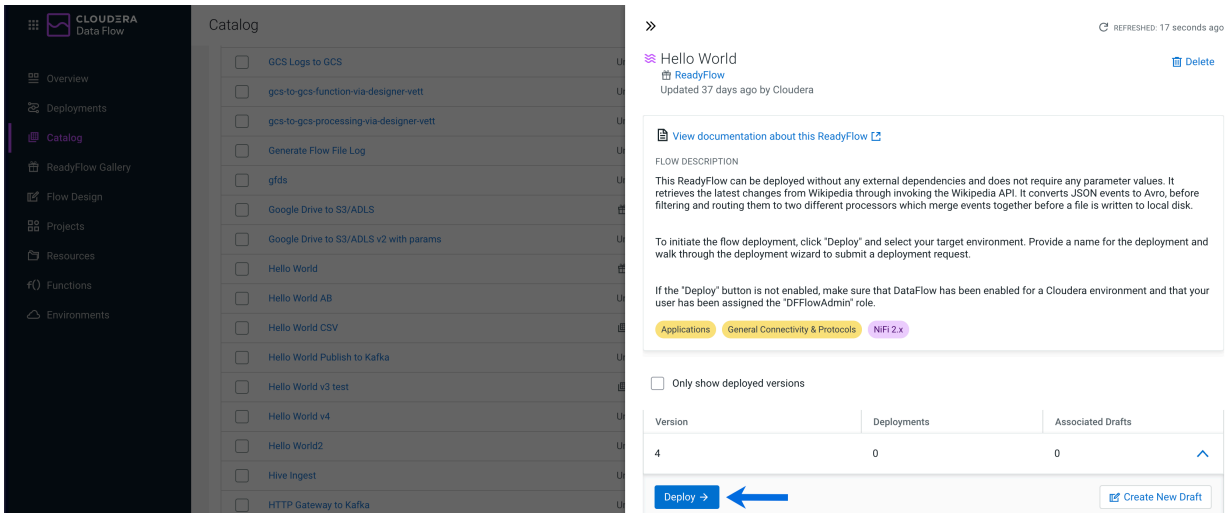
1. In the navigation pane, click the Catalog item.


CLOUDERA
Data Flow
Overview
Deployments
Catalog
ReadyFlow Gallery
Flow Design
Projects
Resources
Functions

2. To initiate deployment, click the Hello World flow definition on the **Catalog** pane to display the flow definition details.



3. Click the Deploy → button to launch the New Deployment wizard.



4. Select the Target Workspace where you want to deploy the ReadyFlow.

   The New Deployment wizard only displays workspaces that meet the following conditions:

   - Cloudera Data Flow is enabled for the environment
   - The environment is in healthy state
   - You are allowed to access the environment

5. Click the Continue → button.
6. In the Overview, give your flow deployment a unique name.

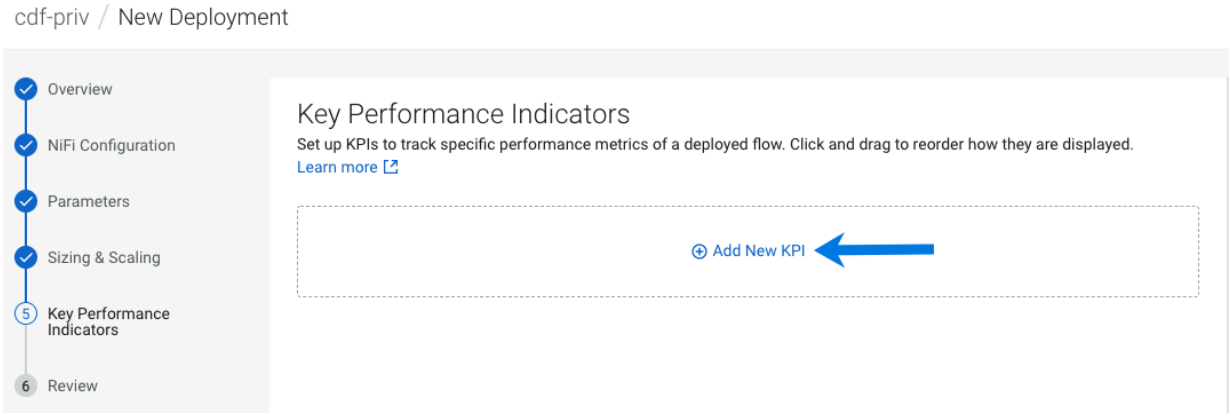
   Note: Flow Deployment names must be unique. The Deployment Wizard indicates whether a name is valid by displaying a green check below the Deployment Name text box.

   Note: As you have created no projects for this quickstart, you can leave your flow deployment Unassigned.

7. Click the Next → button.
8. In the NiFi Configuration step, click the Next → button.
9. In the Parameters step, click the Next → button.
10. In the Sizing & Scaling step, click the Next → button.

11. To learn about the use of Key Performance Indicators (KPIs), add a KPI that will track how much data the processor is writing to the destination files.

The KPI will raise an alert whenever the value goes below 1MB.



a) In the Key Performance Indicators step, click the ⊕ Add New KPI button.

b) Set the following properties:

KPI Scope

Select Processor.

Processor Name

Select Write "Added Content" Events to File.

Metric to Track

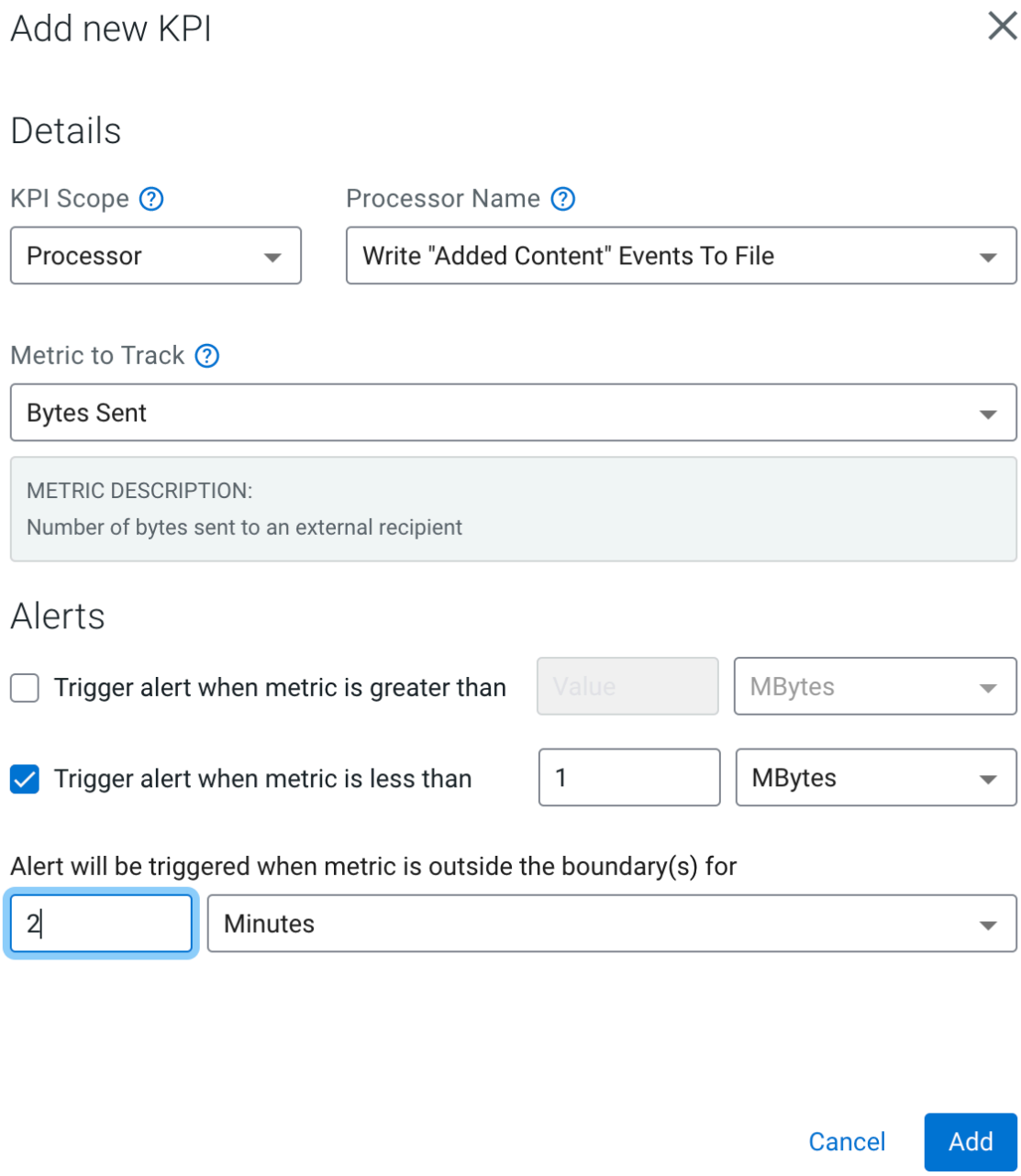
Select Bytes Sent

Alerts

Select Trigger alert when metric is less than and set the value to 1 MBytes.

Alert will be triggered when metric is outside the boundary(s) for

Set to 2 Minutes.



c) Click the Add button to create the KPI.

12. Click the Next → button.
13. Review a summary of the information provided. When you are finished, complete flow deployment by clicking the ☁ Deploy button.

Results

Once you click the Deploy button, you are redirected to the **Active Alerts** tab in the detail view for the deployment where you can track the deployment progress.

Related Information

Deploying Flow Definitions