

Meet the Committers Lab Preparation

3-May-2023

IMPORTANT NOTES



Shared Environment

Since it is **shared environment**, each user has access to every other users' flow.

No production data should be used.

We will stop your design sessions after 4 hours of inactivity.



Guidance

Documentation

Daily Zoom Sessions

Examples

Ready Flows

Slack Channel

Flow Proctors

SANDBOX FROM MAY 3, 2023 to MAY 9 MIDNIGHT, 2023

Sandbox will be destroyed at midnight EST May 9, 2023 before May 10, 2023.

You must complete your item, Save and Download Your Flows Before Then.

All data and code will be destroyed on the end of the trial

Submit your flow (CRN), video and text via this [form](#).

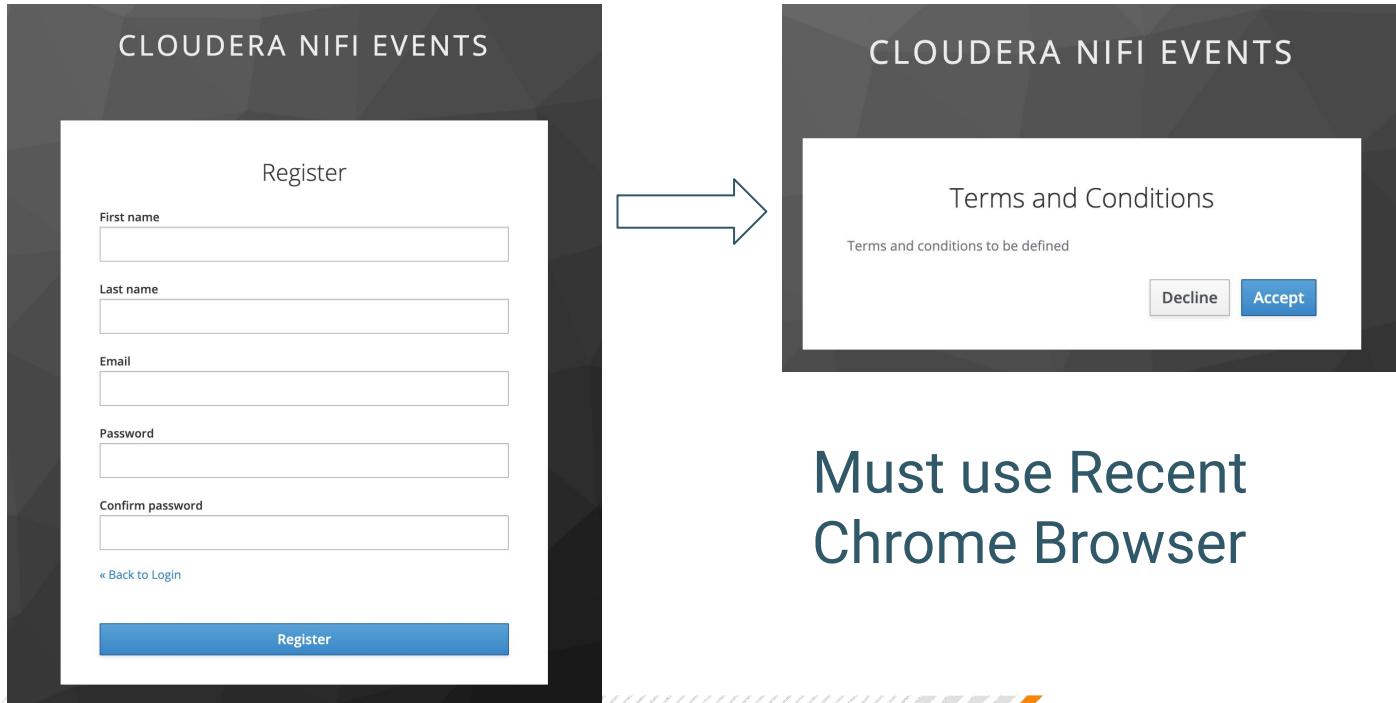
NAVIGATE IN CHROME TO THE SHARED SANDBOX

<https://login.cdpworkshops.cloudera.com/auth/realms/se-workshop-5/protocol/saml/clients/cdp-sso>



REGISTRATION

Click Register



Must use Recent
Chrome Browser

CLOUDERA NIFI EVENTS

Email verification



You need to verify your email address to activate your account.

An email with instructions to verify your email address has been sent to you.

Haven't received a verification code in your email? [Click here](#) to re-send the email.



Cloudera NiFi Events <cloudera_nifi_events@cloudera.com>
to me ▾

7:35 AM (0 minutes ago) ⭐

Welcome to Cloudera Nifi Event. Please Verify your email

[Link to e-mail address verification](#)

This link will expire within 15 minutes.

Join Slack with Cloudera employees here: <http://bestinflow.slack.com/>

Submit your flows to be entered for the Amazon Gift Card here: <https://docs.google.com/forms/d/1Ku2KSDFoxJy45jiOWuLRDi9Trpgm-42aaxeAVwy-fpo/edit>

View the developers page here: <https://community.cloudera.com/t5/Community-Articles/Best-in-Flow-Event/ta-p/368947>

Dataflow Example here: <https://github.com/tspannhw/FLaNK-TravelAdvisory>

Reply

Forward

CLOUDERA NIFI EVENTS

Confirm validity of e-mail address
tim@datainmotion.dev.

Confirm validity of e-mail address tim@datainmotion.dev.

» [Click here to proceed](#)



CLOUDERA NIFI EVENTS

Your email address has been verified.

Your email address has been verified.

CLOUDERA NIFI EVENTS

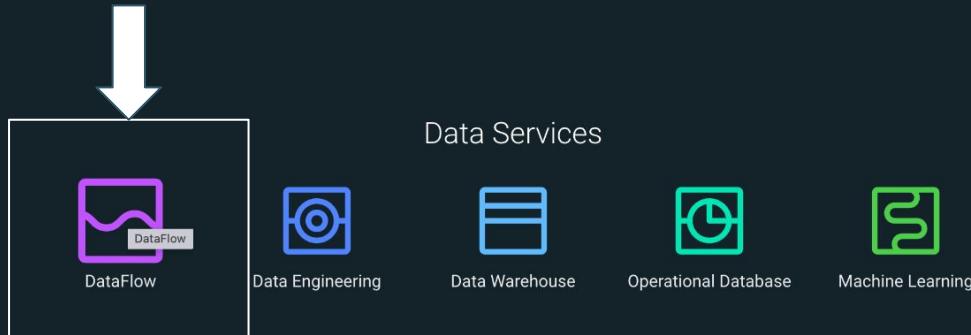
Log In

Email

Password

Log In

New user? [Register](#)

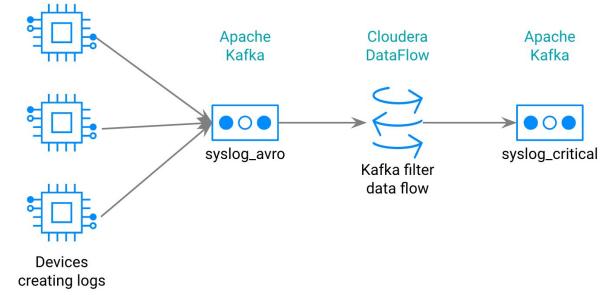


Data Management



GETTING STARTED - GUIDED USE CASES

- Syslog to Kafka topic
- Reading and Filtering a Syslog Stream
- Writing Critical Syslog Events to Apache Iceberg
- Must use Recent Chrome Browser



BEST IN FLOW COMPETITION - BUILD & DOCUMENT A FLOW

A chance to win a \$2,000 Amazon gift card.

A great way to get recognition.

Cloudera public award social media post.

BEST IN FLOW COMPETITION - BUILD & DOCUMENT A FLOW

A chance to win a \$2,000 Amazon gift card.

A great way to get recognition.

Cloudera public award social media post.

WHAT TO BUILD?

You can extend or try one of our tutorials

You can extend or use one of our Ready Flows

You can connect to external resources (passwords are **visible**,
only use **public data** or **examples**)

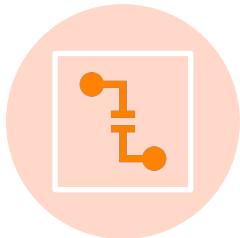
FLOW REQUIREMENTS

The following are the requirements for the flow to be considered eligible for the competition:

1. The flow must be developed using the new **DataFlow Designer in the DataFlow Service sandbox**.
2. The flow must have at least one “**source**” data.
3. The flow must have at least one “**destination**” where the data is delivered
4. The flow must be **functional, tested, and working** using the Test session feature of the DataFlow Service. The Data viewer should be used to inspect the data payload within the different flow steps.
5. The flow must be **checked into the DataFlow Catalog, deployed** using the deployment wizard, and **validated** that it is correctly running.
6. Each submitted Flow must include the following additional details:
 - The CRN of the flow was checked into the flow catalog with a detailed description of the flow and use case.
 - Link to a short blog describing the use case and the flow that was built and deployed using DataFlow Designer
 - Link to a short video showing the flow running in the Flow Designer with the test session and data traversing through flow. The Data viewer should be used to inspect the data payload within the different flow steps.
 - Product feedback on the DataFlow Service.

Criteria	Description
Complete Flow Artifacts	The submitted flow entry contains all the required artifacts, including Flow CRN in the Catalog, a link to the blog describing the use case and the flow, and a short video link showing the flow running with data traversing through the flow.
Adheres to NiFi flow best practices	Follows NiFi flow design best practices like record-oriented processors , controller services, and parameters.
Showcases NiFi processing capabilities	Showcases NiFi processing capabilities including protocol bridging, schema transformation, routing, filtering, enrichment, compression, etc.
Universal Data Distribution	The flow showcases multiple data sources and delivers data to multiple destinations.
Uses the latest NiFi processors and controllers services	Showcases the latest NiFi processors in the latest Apache NiFi release: 1.20 , 1.19 , 1.18 , 1.17 , including PutSnowflakeInternalStage, PutIceberg, UpdateDeltaLakeTable, Amazon ML Processors: Amazon Web Services Polly, Textract, Translate, and Transcribe services, etc.
ReadyFlow	The flow addresses a common data pipeline use case and can be reused by other users hence a good candidate to be added to the ReadyFlow gallery.
Deployable	The flow should be able to be deployed with minimum effort with the appropriate documentation (e.g.: description of parameters in the parameter context, the blog details, etc..)

SANDBOX FLOW DEVELOPMENT BEST PRACTICES



Uniquely Name your
processors/ connections
with yourid_



Parameterize connection
information

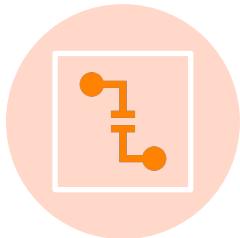


Don't use sensitive data
in sandbox



Don't use or change other
people's assets, only your
own

SANDBOX FLOW DEVELOPMENT BEST PRACTICES



We are here to help reach out via Slack or Zoom.



Don't use or change other people's assets, only your own



Reuse components via Copy and Process Groups

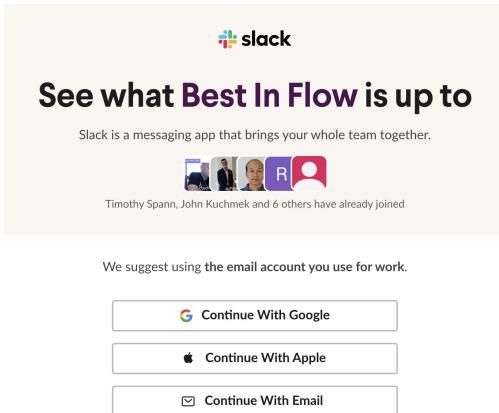
DAILY ZOOM

<https://cloudera.zoom.us/j/96460893376?pwd=eWZEVDhpZmpFSDNRejFzMXkvcHpOdz09>



SLACK CHANNEL

https://bestinflow.slack.com/join/shared_invite/zt-1uj1ti8hc-8mnhmbr_AbOCD7f~A68P0w#/shared-invite/email



SOURCE CODE AND EXAMPLES

<https://github.com/tspannhw/FLaNK-DataFlows>



Submit Your Flow



<https://docs.google.com/forms/d/1Ku2KSDFoxJy45jiOWuLRDi9Trpgm-42aaxeAVwy-fpo>

ADDITIONAL RESOURCES

[Cloudera DataFlow Designer: The Key to Agile Data Pipeline Development](#)

[Streaming Data Ingestion into an Open Data Lakehouse Made Easy with DataFlow Example](#)

[Cloudera DataFlow Designer: Kafka to Iceberg in Cloudera Data Warehouse](#)

[Serverless NiFi Flows with DataFlow Functions](#)

[DataFlow Functions Technical Demo](#)

[DataFlow Documentation](#)

Meet the Data-In-Motion Team

Field



Carolyn
Duby
Field CTO



Tim
Spann
Developer Advocate

Marketing



Chris
Joynt
Product Marketing

Engineering



Joe
Witt
Engineering Leader

Product



George
Vetticaden
Product Leader



Richard Walden
DIM SME Lead



John Kuchmek
DIM SME Expert



Michael
Kohs
Product Owner for
DataFlow



Pierre
Villard
Product Owner for
DataFlow



Andre
Araujo
Product Owner for
Stream Processing

WARNING

“Notwithstanding any contrary terms in the Agreement, Customer acknowledges that information shared using the Trial Product is in a shared environment with similarly situated customers. All information in the shared environment is accessible by all other customers participating in the trial and such information will not be deemed Confidential Information.”

<https://www.cloudera.com/legal/commercial-terms-and-conditions/cdp-public-cloud-trial-agreement.html>

CONTAINER BASED DATAFLOW



Flow Catalog

Keep track of your flow definitions and versions in a central catalog

Reuse your existing NiFi flows by uploading them to the catalog

Discover, search and reuse existing flows easily



Flow Deployment

Allows easy flow deployment based on NiFi 1.20 across CDP environments (Dev, QA, Prod)

Define and assign KPIs to your flows

Easy NiFi version upgrades

Update/Add KPIs, Update Parameters, Change sizing configuration

Automatic infrastructure scaling based on CPU utilization



Flow Monitoring

Central monitoring console for all your flows across environments

Monitor flow metrics and infrastructure usage

Define alerts for flows breaching assigned KPIs

FLOW CATALOG

- Central repository for flow definitions
- Import existing NiFi flows
- Manage flow definitions
- Initiate flow deployments

The screenshot shows the Cloudera DataFlow Flow Catalog interface. On the left is a dark sidebar with navigation links: Dashboard, Catalog (which is selected), ReadyFlow Gallery, Environments, Help, and a user icon. At the bottom of the sidebar, it says "1.0.1-b570". The main area is titled "Flow Catalog" and contains a search bar and a refresh button indicating it was last refreshed 25 seconds ago. A blue "Import Flow Definition" button is located in the top right. Below these are two tabs: "Name ↑" and "Last Updated". The main content is a table listing ten flow definitions:

Name ↑	Type	Versions	Last Updated	Action
cc_fraud_template_int101run	Custom Flow Definition	2	a day ago	>
cc_fraud_template_int101run2	Custom Flow Definition	1	9 days ago	>
JSON_Kafka_To_Avro_S3	Custom Flow Definition	2	a day ago	>
Kafka filter to Kafka	ReadyFlow	1	2 days ago	>
Kafka to Cloudera Operational Database	ReadyFlow	1	2 days ago	>
Kafka to S3 Avro	ReadyFlow	1	14 hours ago	>
nifi_flows	Custom Flow Definition	1	2 months ago	>
Weather Data Flow	Custom Flow Definition	1	a day ago	>
Weather_Data	Custom Flow Definition	1	15 days ago	>
Weather_JSON_Kafka_To_Avro_S3	Custom Flow Definition	1	21 days ago	>

At the bottom right, there are buttons for "Items per page:" (set to 10), a page number indicator (1 - 10 of 10), and navigation arrows (< >).

TURNS FLOW DEFINITIONS INTO FLOW DEPLOYMENTS

1.) Start Deployment Wizard

se-sandbox-aws / New Deployment

Overview

Deployment Name: abc_hello_world
Deployment name is valid

Selected Flow Definition: Hello World (Version 1)

Target Environment: se-sandbox-aws

2.) NiFi Config

se-sandbox-aws / New Deployment

NiFi Configuration

NiFi Runtime Version: CURRENT VERSION Latest Version (1.18.0.2.3.7.1-1)

Review the Cloudera DataFlow and CDH Runtime support matrix to ensure the selected NiFi Runtime Version is compatible.

Autostart Behavior: Automatically start flow upon successful deployment

Inbound Connections: Allow NiFi to receive data

Custom NAR Configuration: This flow deployment uses custom NARs

Overview

Hello World v1

ENVIRONMENT DEPLOYING TO se-sandbox-aws

DEPLOYMENT NAME abc_hello_world

3.) Provide Parameters for NiFi

se-sandbox-aws / New Deployment

Parameters

No parameters to configure. Flow parameters allow you to use your data flow in different contexts. Check out our documentation to learn more about parameters in Cloudera DataFlow. [Learn more](#)

Overview

FLOW DEFINITION Hello World v1

ENVIRONMENT DEPLOYING TO se-sandbox-aws

DEPLOYMENT NAME abc_hello_world

NiFi Configuration

NiFi Runtime Version: Latest Version (1.18.0.2.3.7.1-1)

AUTO-START FLOW Yes

INBOUND CONNECTIONS No

CUSTOM NAR CONFIGURATION No

4.) Configure Sizing & Scaling

se-sandbox-aws / New Deployment

Sizing & Scaling

Select the NiFi node size and the number of nodes provisioned for your flow.

NiFi Node Sizing

Extra Small, Small (selected), Medium, Large

2 Cores Per Node, 4 GB Per Node

3 Cores Per Node, 6 GB Per Node

4 Cores Per Node, 12 GB Per Node

12 Cores Per Node, 24 GB Per Node

Number of NiFi Nodes

Auto Scaling Enabled

Min. Nodes: 1 Max. Nodes: 32

5.) Define KPIs

se-sandbox-aws / New Deployment

Key Performance Indicators

FLOW DEFINITION Hello World v1

ENVIRONMENT DEPLOYING TO se-sandbox-aws

DEPLOYMENT NAME abc_hello_world

Entire Flow

METRIC TO TRACK File Queued

ALERT SET No alert set

Entire Flow

METRIC TO TRACK Data Out

ALERT SET Notify if outside the range of 999 MB/sec - 1 MB/sec, for at least 5 minutes.

Add New KPI

Parameters

No parameters are available for this flow.

Overview

Hello World v1

ENVIRONMENT DEPLOYING TO se-sandbox-aws

DEPLOYMENT NAME abc_hello_world

NiFi Configuration

NiFi Runtime Version: Latest Version (1.18.0.2.3.7.1-1)

AUTO-START FLOW Yes

INBOUND CONNECTIONS No

CUSTOM NAR CONFIGURATION No

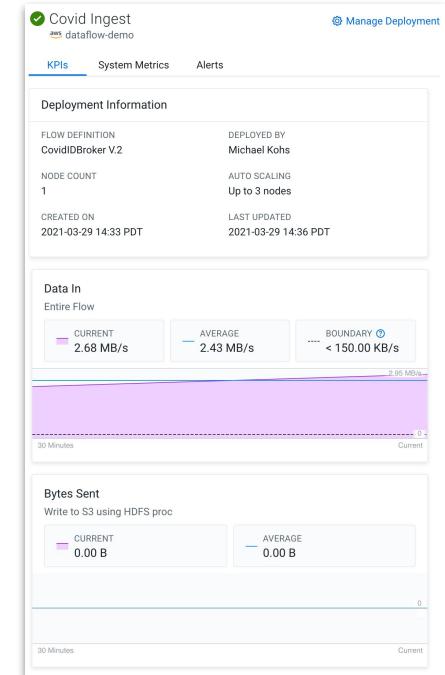
KEY PERFORMANCE INDICATORS

- Visibility into flow deployments
- Track high level flow performance
- Track in-depth NiFi component metrics
- Defined in Deployment Wizard
- Monitoring & Alerts in Deployment Details

KPI Definition in Deployment Wizard

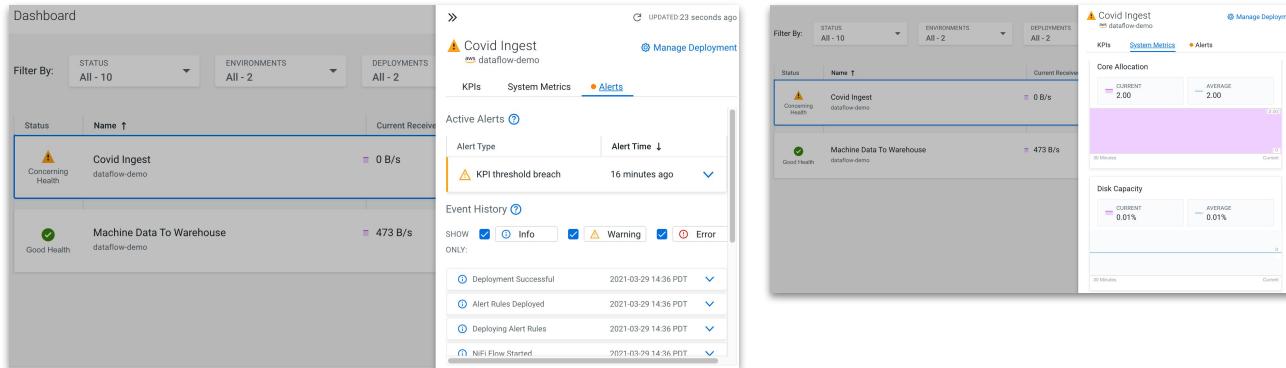
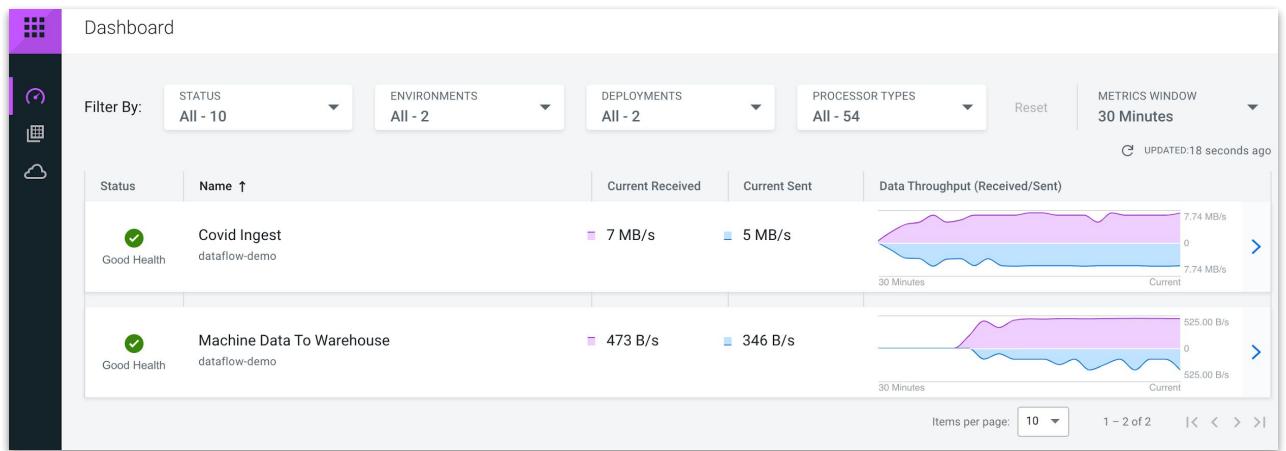
The screenshot shows the 'New Deployment' step of the deployment wizard. On the left, a sidebar lists steps: Overview, NiFi Configuration, Parameters, Sizing & Scaling, Key Performance Indicators (selected), and Review. The main area is titled 'Key Performance Indicators' with the sub-instruction 'Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.' Below this is a 'Learn more' link. There are two sections for defining KPIs: 'Entire Flow' and 'Data Out'. Under 'Entire Flow', it says 'METRIC TO TRACK: Flow Files Queued' and 'ALERT SET: No alert set'. Under 'Data Out', it says 'METRIC TO TRACK: Data Out' and 'ALERT SET: Notify if outside the range of 999 MB/sec - 1 MB/sec, for at least 5 minutes.' At the bottom is a button labeled 'Add New KPI'.

KPI Monitoring



DASHBOARD

- Central Monitoring View
- Monitors flow deployments across CDP environments
- Monitors flow deployment health & performance
- Drill into flow deployment to monitor system metrics and deployment events



DEPLOYMENT MANAGER

- Manage flow deployment lifecycle
(Suspend/Start/Terminate)
- Add/Edit KPIs
- Change sizing configuration
- Update parameters
- Change NiFi version of the deployment
- Gateway to NiFi canvas

Dashboard / dataflow-demo-new / Kafka to COD

REFRESHED 12 seconds ago

Actions ▾

Deployment Manager

Status: Good Health

Deployment Name: Kafka to COD

Flow Definition: Kafka to Cloudera Operational Database V.1

Deployed By: Michael Kohs

Node Count: 1

Auto Scaling: Disabled

Created On: 2021-07-26 17:05 PDT

Last Updated: 2021-07-26 17:07 PDT

Environment: dataflow-demo-new

Region: US West (Oregon)

NIFI Runtime Version: 1.14.0-2.3.0-0-89

Deployment Settings

KPIs and Alerts Sizing and Scaling Parameters

Parameters

Running Processors that are affected by the Parameter changes will automatically be restarted.

Data entered here never leaves the environment in your cloud account. Provide parameter values directly in the text input or upload a file for parameters that expect a file.

The selected flow definition references an external Default NiFi SSL Context Service. Hence, DataFlow will automatically create a matching SSL Context Service with a keystore and truststore generated from the target environment's FreeIPA certificate.

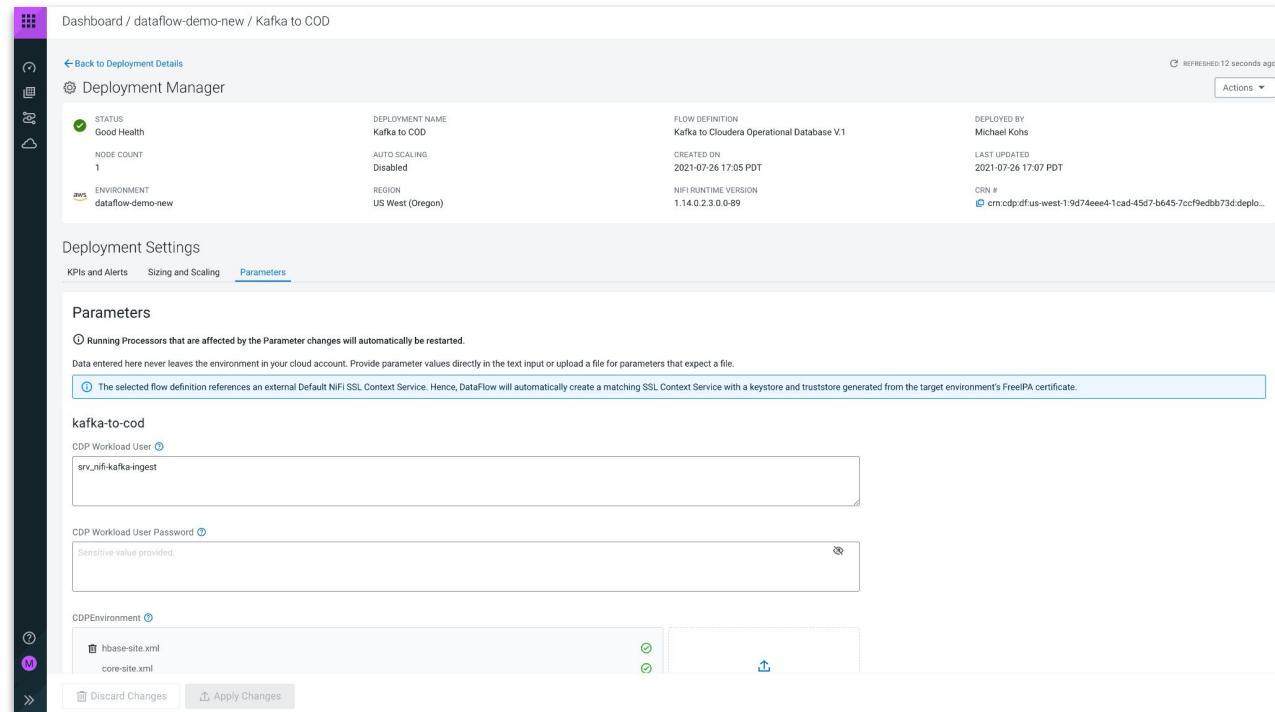
kafka-to-cod

CDP Workload User:

CDP Workload User Password:

CDP Environment: hbase-site.xml core-site.xml

Discard Changes Apply Changes



TH^ON^G Y^OU[★]

