

# Best Practices for Developing With Data Flow

21-March-2023

---

IMPORTANT

---

## WARNINGS

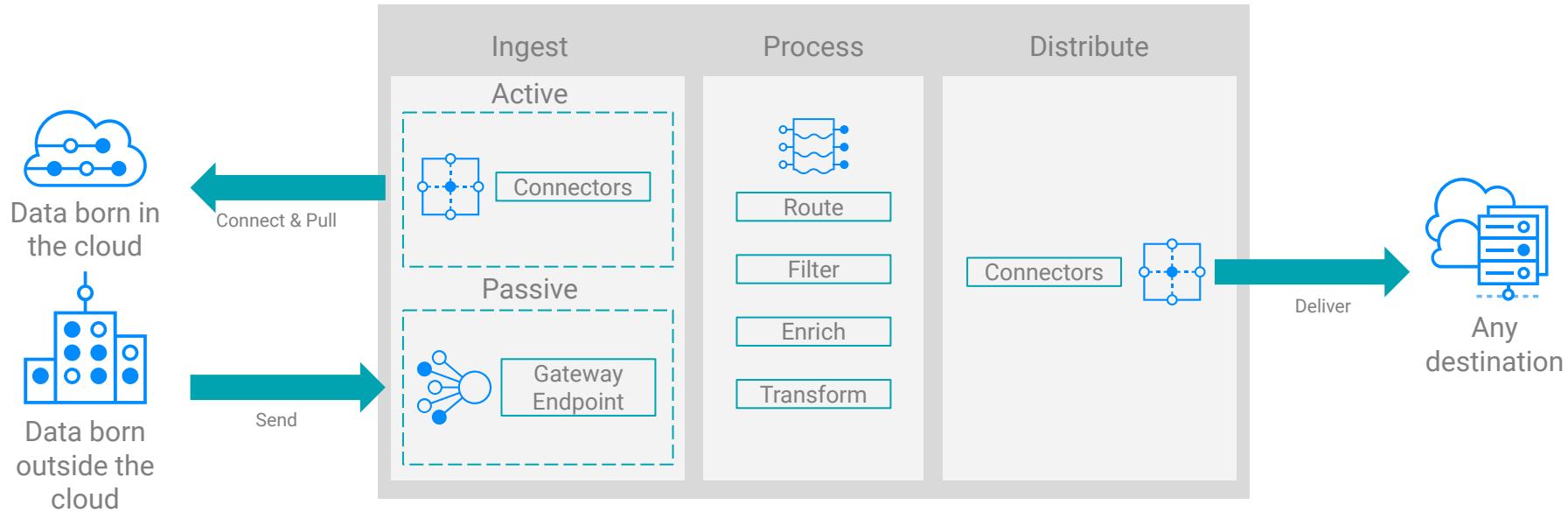
- We will **stop** your design sessions after 4 hours of inactivity
- Starting Sessions takes at least 5 minutes up to 30 minutes be patient
- Ending Sessions takes at least 5 minutes up to 30 minutes be patient
- Starting Sessions requires re-entering Workload User Password in parameters and applying.

---

# DATA IN MOTION - Overview

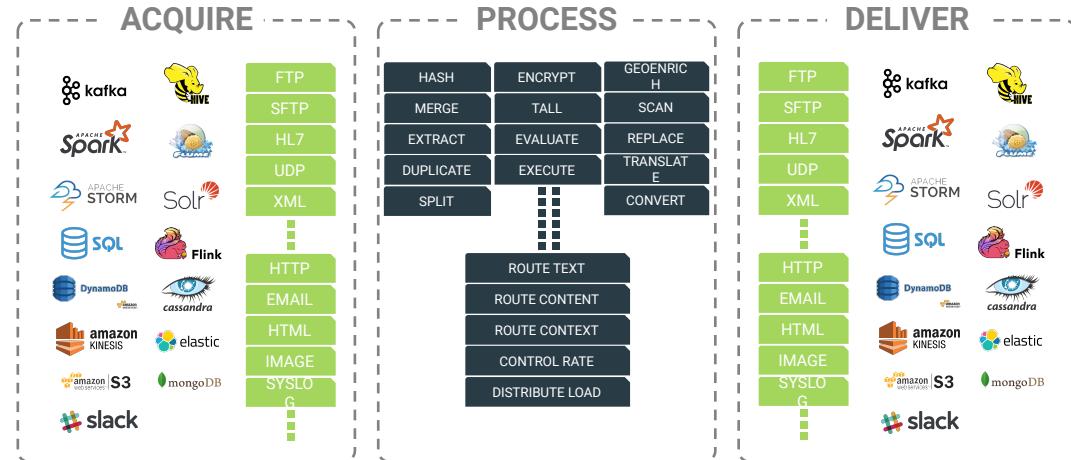
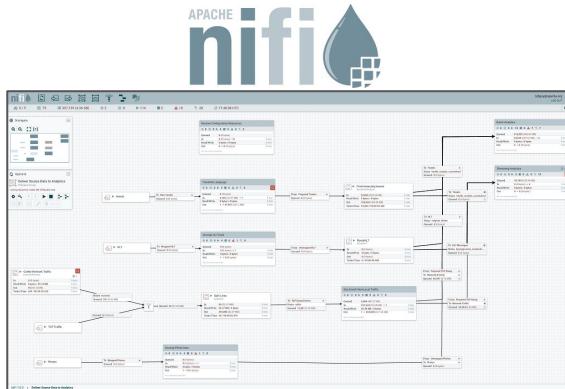
# UNIVERSAL DATA DISTRIBUTION

Connect to any data source anywhere, process and deliver to any destination



# CLOUDERA FLOW MANAGEMENT

Ingest and manage data from edge-to-cloud using a no-code interface



- Over 350 pre-built processors
- Easy to build your own processors
- Parse, enrich & apply schema
- Filter, Split, Merge & Route
- Throttle & Backpressure
- Guaranteed delivery
- Full data provenance
- Eco-system integration

---

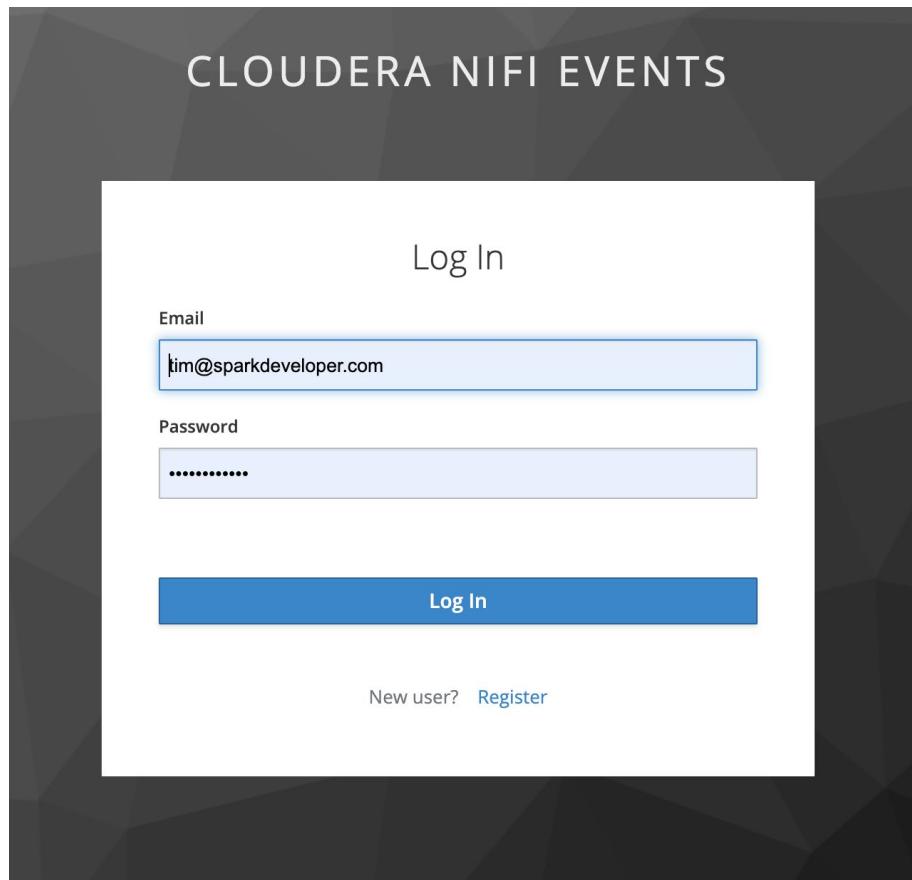
# FAQ

---

# RECOMMENDATIONS

- Kafka Topic Names cannot have blank lines or spaces
- Schema Names cannot have blank lines or spaces
- Make sure everything in parameters is trimmed
- Every action is audited
- Shutdown your flow when you are done for the day or not working
- All flows will be shutdown after 4 hours of running

# RELOGIN FREQUENTLY



# SPACES AND EMPTY LINES MATTER

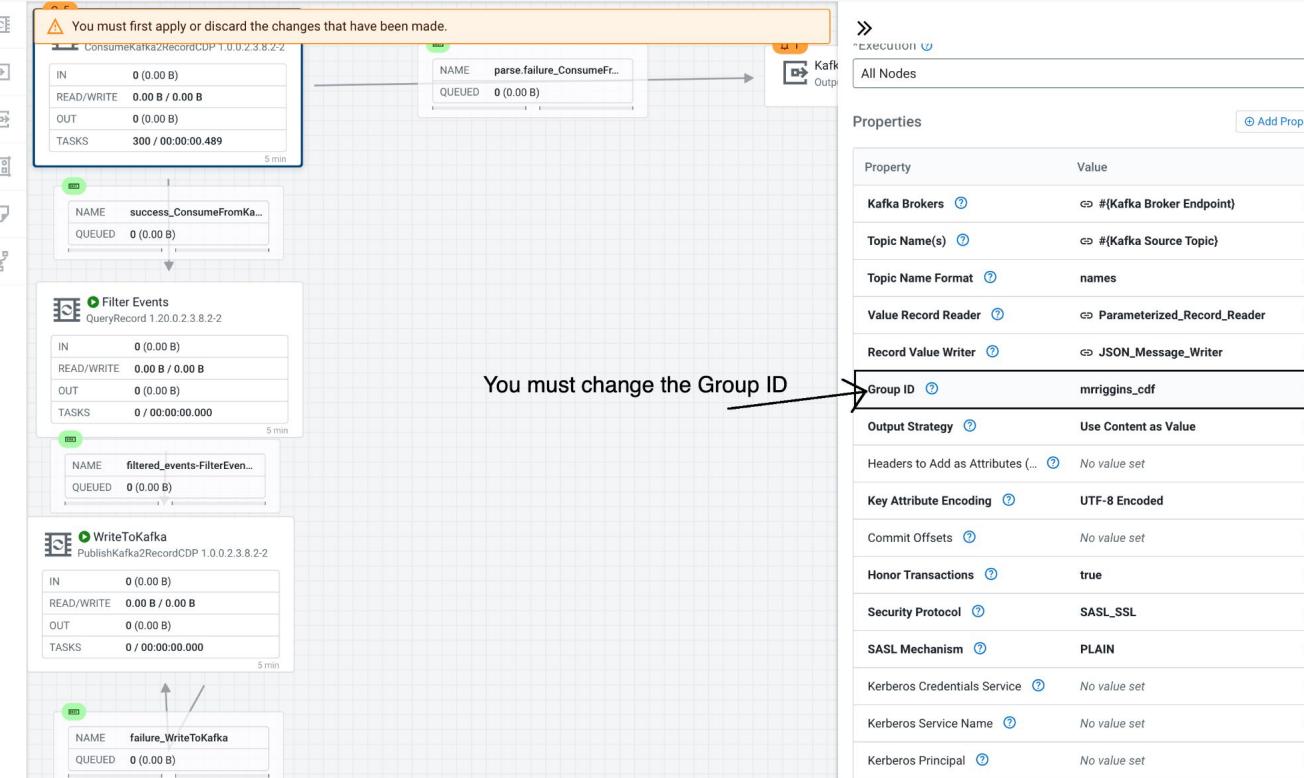
Name ↑	Value	Changed
CDP Workload User	mriggs78729	>
CDP Workload User Password	☒ Sensitive value set	>
CDPEnvironment	☒ hive-site.xml, core-site.xml, ssl-client.xml	>
CSV Delimiter	,	>
Data Input Format	AVRO	ⓘ Modified >

☒ AVRO

Set empty string

Description

Select File  
Drop file or browse



---

# General Tips

---

# RECOMMENDATIONS

## For Developers

- Use the latest Chrome
- Disable virus scans
- Disable VPN
- Use a fast network
- Don't run too many other things
- For security reasons, things time out
- Don't put in extra spaces in parameters, names or anywhere
- Don't use special characters in names
- Keep names unique prefaced with your username\_

---

# IMPORTANT

## For Developers

- Start with a **ReadyFlow** for things like Kafka as they set up a lot of items for you.
- If you have services not working or missing, stop and restart the test session. It will add SSL context.
- Publish your flow to the catalog, this is your backup

---

# USEFUL DOCUMENTATION

## For Developers

- <https://docs.cloudera.com/dataflow/cloud/readyflow-overview-listenhttp-kafka/topics/cdf-ingest-listenhttp-kafka-prerequisites.html>
- 
-

# COMMON ERRORS

## For Developers

- SSL Error
- Authorization
- Doesn't exist
- Timed Out (Login again in another tab:  
<https://login.cdpworkshops.cloudera.com/auth/realms/se-workshop-5/proTOCOL/saml/clients/cdp-sso>)
- If you stopped your session, after you restart you have to reset your sensitive parameters including CD Workload User Password.

**ERROR** – 2023-04-19 07:50:53 pm EDT  
PublishKafka2RecordCDP[id=cb47d8e4-9b13-33ff-a8f8-d4b27bf3da58] Failed to send FlowFile[filename=508557dc-5eb1-4bcf-ade5-029b373a3a8d] to Kafka:  
org.apache.kafka.common.errors.TopicAuthorizationException: Not authorized to access topics: [tim\_traveladvisory]

---

# Flow Design Tips

## All Drafts

>Create Draft
REFRESHED: 4 seconds ago

Draft Name ↑	Workspace	Test Session	Runtime Version	Base Flow	Last Updated	Published On	
george.vetticaden_kafka_to_iceberg	aws oss-demo-aws	Active	1.20.0.2.3.8.1-1	Kafka to Iceberg v1	13 hours ago by george.vetticaden	-	⋮
george.vetticaden_kafkafilterkafka	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	Kafka filter to Kafka - george.vetticaden v1	15 hours ago by george.vetticaden	15 hours ago	⋮
hanka-syslog-kafka-flow	aws oss-demo-aws	Active	1.20.0.2.3.8.1-1	N/A	7 days ago by system	-	⋮
janice-syslog-kafka-flow	aws oss-demo-aws	Not Defined	N/A	N/A	6 days ago by lauren.thomas	-	⋮
Kafka filter to Kafka Tim Test 3	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	Kafka filter to Kafka v1	14 days ago by system	-	⋮
Kafka to Iceberg cool 5	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	Kafka to Iceberg v1	14 days ago by system	-	⋮
ossdemo6	aws oss-demo-aws	Not Defined	N/A	N/A	14 days ago by bunkertor	-	⋮
test	aws oss-demo-aws	Inactive	1.20.0.2.3.8.0-31	test_mkohs v1	15 days ago by system	20 days ago	⋮
tim_HTTPtoKafka	aws oss-demo-aws	Not Defined	N/A	N/A	a day ago by tim	-	⋮
tim_kafkafilterkafka	aws oss-demo-aws	Active	1.20.0.2.3.8.1-1	tim_kafkafilterkafka v1	15 hours ago by michael.kohs	19 hours ago	⋮
tim_kafkatoiceberg	aws oss-demo-aws	Active	1.20.0.2.3.8.1-1	Kafka to Iceberg v1	4 hours ago by tim	-	⋮
tim_restraveladvisory	aws oss-demo-aws	Active	1.20.0.2.3.8.1-1	N/A	3 minutes ago by tim	-	⋮
tim_SyslogtoKafka	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	N/A	a day ago by tim	-	⋮
timexternaltest	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	N/A	14 days ago by system	-	⋮
TimSpann_ListenHTTP filter to Kafka	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	ListenHTTP filter to Kafka v1	13 days ago by system	-	⋮
timtest3	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	N/A	14 days ago by system	-	⋮
tspann_KafkaFilterToKafka	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	Kafka filter to Kafka v1	2 days ago by tim	-	⋮
tspann_readfiltersyslogafka	aws oss-demo-aws	Inactive	1.20.0.2.3.8.1-1	Kafka filter to Kafka v1	a day ago by system	-	⋮

Items per page:
50
1 – 18 of 18
| < > |

Your items may be on other pages

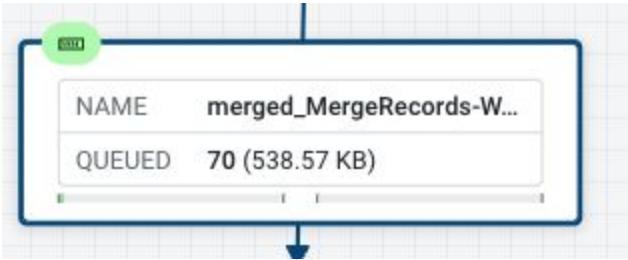
Active Test Session

Flow Options ▾



Make sure your session is active.

If things are slow, weird or not working, click and end your session. Once it is stopped, log off, close your browser. Restart your browser, log back in and restart your session.



Name all your Connections



Click to see your warnings and errors.  
If you click twice or hold you can then  
select and copy your errors.



**Bulletins Reported (5)**

**ERROR** - 2023-04-19 03:00:33 pm EDT  
Puticeberg[id=b2133d95-611f-3cd1-90f7-02df8383c880] Exception occurred while writing iceberg records. Removing uncommitted data files:  
org.apache.iceberg.exceptions.CommitStateUnknownException:  
MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fmetadatas%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)  
Cannot determine whether the commit was successful or not, the underlying data files may or may not be needed. Manual intervention via the Remove Orphan Files Action can remove these files when a connection to the Catalog can be re-established if the commit was actually unsuccessful.  
Please check to see whether or not your commit was successful before retrying this commit. Retrying an already successful operation will result in duplicate records or unintentional modifications.  
At this time no files will be deleted including possibly unused manifest lists.  
- Caused by: java.lang.RuntimeException: MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fwarehouse%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)  
- Caused by: MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fwarehouse%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)  
**ERROR** - 2023-04-19 03:00:31 pm EDT  
Puticeberg[id=b2133d95-611f-3cd1-90f7-02df8383c880] Exception occurred while writing iceberg records. Removing uncommitted data files:  
org.apache.iceberg.exceptions.CommitStateUnknownException:  
MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fwarehouse%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)  
Cannot determine whether the commit was successful or not, the underlying data files may or may not be needed. Manual intervention via the Remove Orphan Files Action can remove these files when a connection to the Catalog can be re-established if the commit was actually unsuccessful.  
Please check to see whether or not your commit was successful before retrying this commit. Retrying an already successful operation will result in duplicate records or unintentional modifications.  
At this time no files will be deleted including possibly unused manifest lists.  
- Caused by: java.lang.RuntimeException: MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fwarehouse%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)  
- Caused by: MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fwarehouse%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)  
**ERROR** - 2023-04-19 03:00:30 pm EDT  
Puticeberg[id=b2133d95-611f-3cd1-90f7-02df8383c880] Exception occurred while writing iceberg records. Removing uncommitted data files:  
org.apache.iceberg.exceptions.CommitStateUnknownException:  
MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fwarehouse%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)  
Cannot determine whether the commit was successful or not, the underlying data files may or may not be needed. Manual intervention via the Remove Orphan Files Action can remove these files when a connection to the Catalog can be re-established if the commit was actually unsuccessful.  
Please check to see whether or not your commit was successful before retrying this commit. Retrying an already successful operation will result in duplicate records or unintentional modifications.  
At this time no files will be deleted including possibly unused manifest lists.  
- Caused by: java.lang.RuntimeException: MetaException(message=Permission denied: user [tim] does not have [RWSTORAGE] privilege on [/ceberg//default/tim\_syslog\_critical\_archive]  
snapshot+;%Fwarehouse%2Ftblspaces%2Fexternal%2Fhive%2Ftim\_syslog\_critical\_archive%2Fmetadatas%2F0001-21638387-44d5-4114-98bc-e0f0ace5e5c-metadata.json)

Check the **ERROR** in Bulletins.

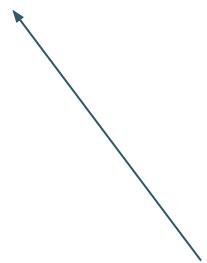
If you see a permission or missing item issue make sure you have your Workload User Name and Workload Password set correctly.

Make sure your table, topic or schema exist - check for typos.

If nothing else works, please go to the Slack channel and post your id, flow and error. We will check it for you.

## Default SSL Context Keystore Password

Sensitive



If you stopped or restarted a Test Session then you may need to re-enter and apply your password.

## Test Session

An active test session will allow you to fully validate your data flow and work with live data.

 Stop

### Inbound Connection Details

## Publish To Catalog

Manage all of your flow definitions from one place.

### Base Flow



FLOW NAME VERSION  
tim\_traveladvisory\_produc... 3

 Publish

 Publish As

Publish a new version of the existing flow

Publish your flow to back it up

[Dashboard](#)[Catalog](#)[ReadyFlow Gallery](#)[Flow Design](#)[Functions](#)[Environments](#)

## Flow Catalog

 [X](#)

Name

[tim\\_traveladvisory\\_production](#)[tim\\_test22](#)[tim\\_kafkafilterkafka](#)[tim\\_traveladvisory\\_production](#)

Updated 6 minutes ago by Timothy Spann

REFRESHED: 4 seconds ago

[Actions ▾](#)

## FLOW DESCRIPTION

Initial Production Release of Travel Advisor by user tim

<https://www.datainmotion.dev/>

CRN #

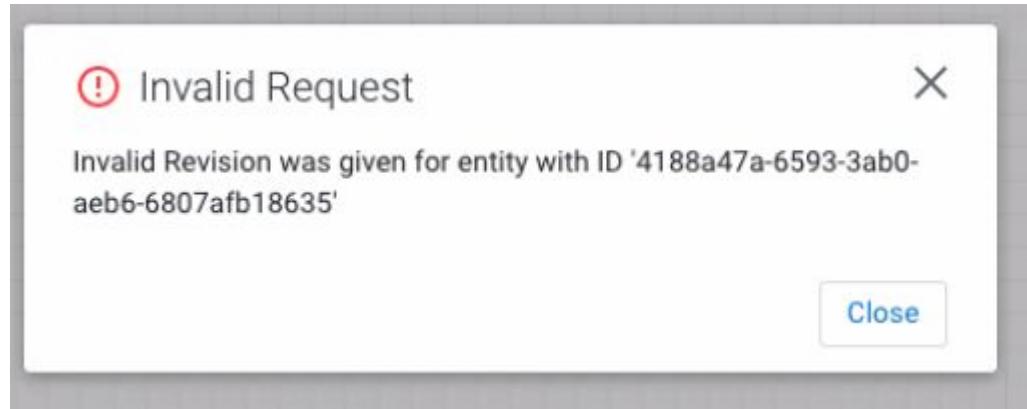
[crn:cdp:df:us-west-1:5251b921-84c5-45c4-af51-c0b8a6ebd1c9:flow:tim\\_traveladvisory\\_production](#) Only show deployed versions

Version	Deployments	Associated Drafts
3	0	1

[Deploy →](#) [Download](#) [Create New Draft](#)

Download your flow as json

If you see error popups, refresh your browser and make sure you have a super fast internet and Chrome browser.



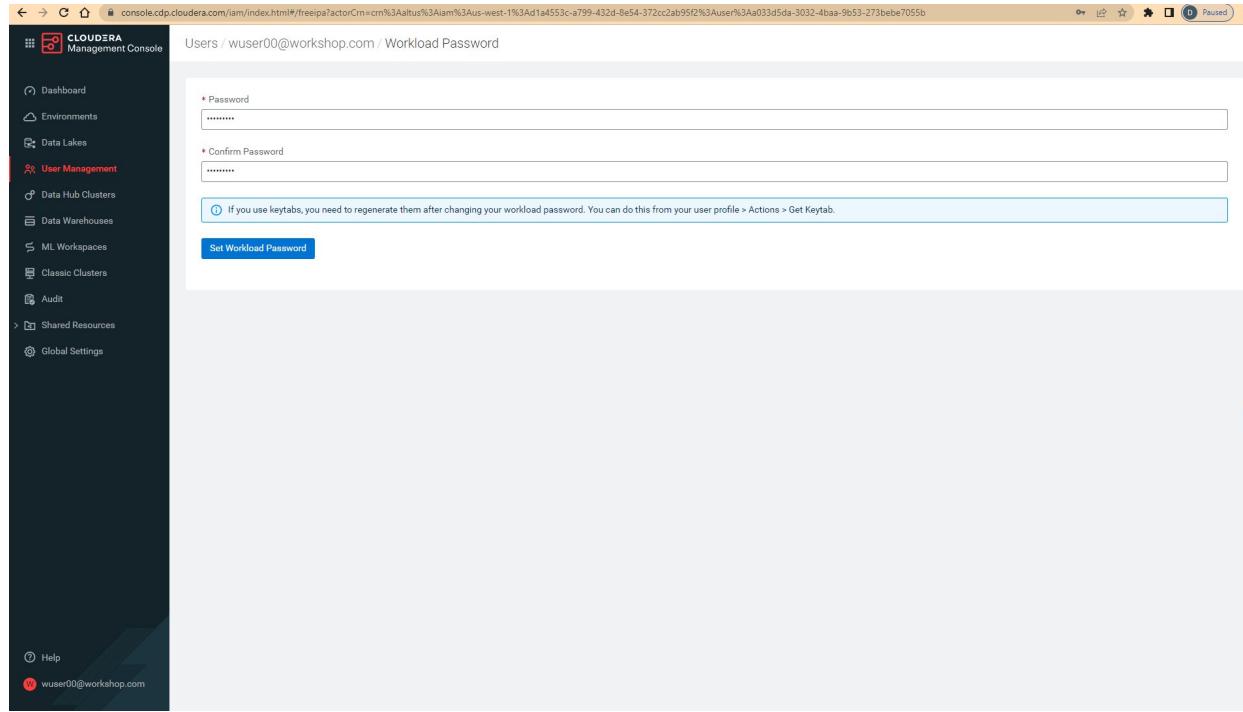
If you see a popup with Invalid Revision or Invalid request.

Refresh your screen, if that doesn't work, restart Chrome or logout and login again.

The cluster is running in the United States, so it may have latency.

Searching for Flow Designs with start, don't add a wild card.

If you want to get earliest messages, set Kafka to earliest



Set your workload passwords.

---

# COMMON ERRORS

```
PublishKafka2RecordCDP[id=15c4f0be-11aa-3526-a169-427bbe63dbc1] Failed  
to send FlowFile[filename=923449ec-e460-4cba-a656-28473fd56d06] to  
Kafka: org.apache.kafka.common.errors.TopicAuthorizationException: Not  
authorized to access topics: [ alexvkahan_syslog_critical  
]
```



No blank lines or spaces in topic names

---

# RESOURCES

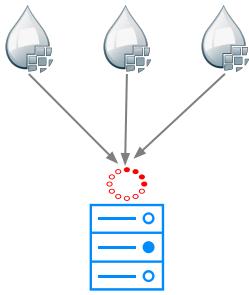
# RESOURCES

- <https://docs.cloudera.com/cdp-public-cloud/cloud/cli/topics/mc-installing-cdp-client.html>
- <https://www.datainmotion.dev/2023/04/dataflow-processors.html>
- <https://github.com/tspannhw/FLaNK-DataFlows>
- <https://github.com/tspannhw/FLaNK-TravelAdvisory>
- <https://github.com/tspannhw/FLaNK-AllTheStreams>
- Bestinflow.slack.com
- <https://docs.google.com/forms/d/1Ku2KSDFoxJy45jiOWuLRDi9Trpgm-42aaxeAVwy-fpo/edit>
- <https://www.cloudera.com/solutions/dim-developer.html>
- <https://community.cloudera.com/t5/Community-Articles/DataFlow-Designer-Event/ta-p/368947>
- [Cloudera DataFlow Designer: The Key to Agile Data Pipeline Development](#)

---

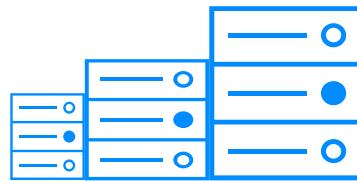
# CLOUDERA DataFlow in CDP

# OPERATIONAL CHALLENGES WITH NiFi FLOWS



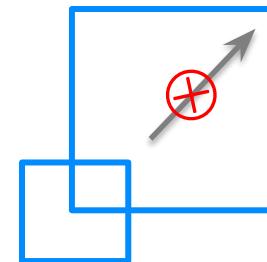
## Resource contention

Impacts performance of all flows in the cluster



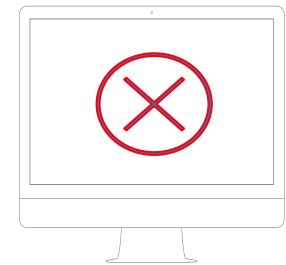
## Oversizing clusters

High infrastructure costs



## On-demand manual scaling

Operational nightmare



## Comprehensive flow visibility

Monitoring NiFi flows across multiple clusters can be challenging

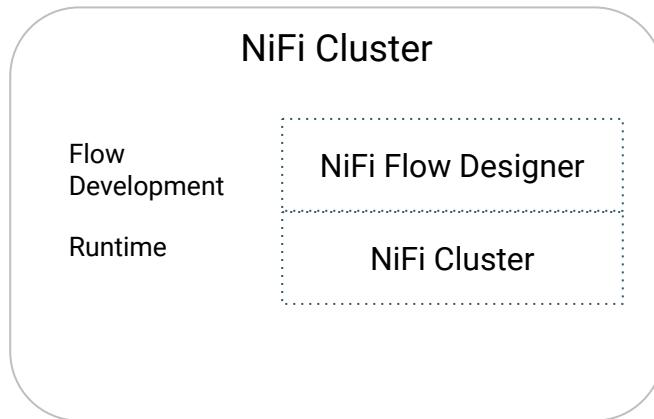
# ARCHITECTURAL SHIFT IN DATAFLOW

Decouple NiFi Flow Designer and the NiFi Cluster Runtime to Support Diverse Runtimes



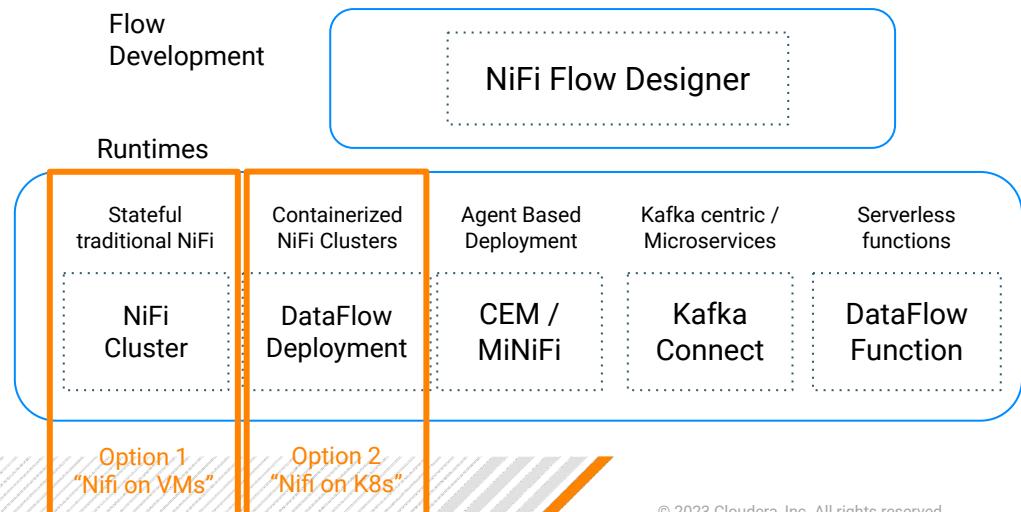
Classic NiFi  
Architecture

NiFi Flow Designer + NiFi Cluster  
Runtime are tightly coupled



New NiFi  
Architecture

Develop flows in designer and deploy in different runtimes  
based on use case



# DEPLOYMENT OPTIONS FOR DATAFLOW

Flow Designer



Version control



Deployment target



Form factor

- CDP PVC BASE

- CDP PC - DataHub

- CDP PC - DataHub  
- CDP PVC BASE

- DataFlow as a Function

- CDP PC - Data Services

Runtime

- VM based/bare metal

- VM based

- VM based/bare metal

- Serverless

- Container based

Workload profile

- Uniform, flat workload profile

- Uniform, flat workload profile

- source/sink for Kafka (stateless)

- not permanent, only used from time to time, event based

- Constantly changing workload profile

# CONTAINER BASED DATAFLOW (AS OF TODAY PUBLIC CLOUD ONLY)



## 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

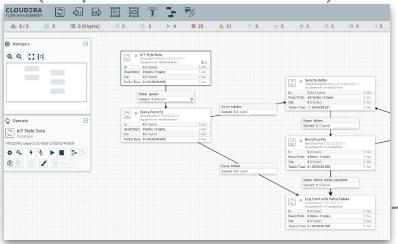
Central monitoring console for all your flows across environments

Monitor flow metrics and infrastructure usage

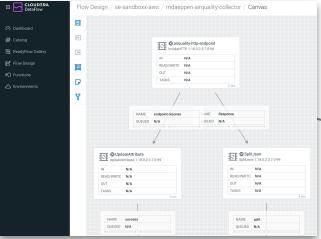
Define alerts for flows breaching assigned KPIs

# DEPLOYMENT WITH SDLC SUPPORT

Classic NiFi Flows  
(on-prem or on Data Hub)

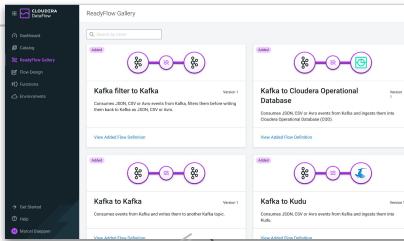


## New Flows Designer



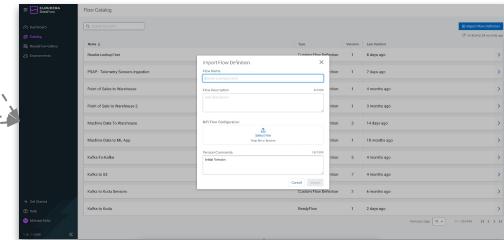
Upload to Catalog

## ReadyFlow Gallery



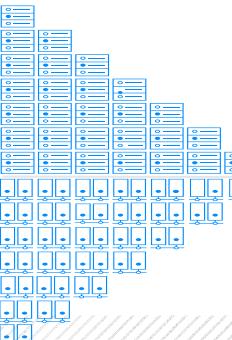
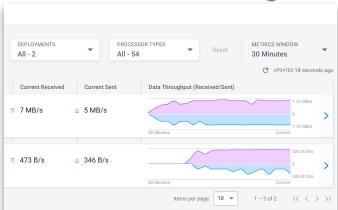
Upload to Catalog

## Flow Catalog



Instantiation into Catalog

With full monitoring



Auto-scale  
Kubernetes  
clusters on CDP

## New Deployment

Select the target environment

- Sensitive data never leaves the environment. Changing the environment after this step requires restarting the deployment process.

Selected Flow Definition

NAME	VERSION
Machine Data To Warehouse	3

Target Environment

aws\_gvtticaden-pm-env-6-oregon 15% (3 of 20)

Cancel

Continue →

Select a Flow &  
use the  
deployment wizard

# FLOW DEVELOPMENT BEST PRACTICES



Name your processors/  
connections



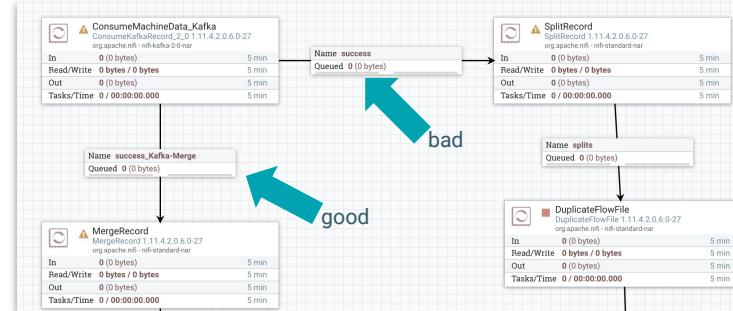
Parameterize  
connection  
information



Tag sensitive  
properties as  
“sensitive”



Define controller  
services on  
process group  
level (except  
*Default NiFi SSL  
Context Service*)



Configure Processor

Invalid

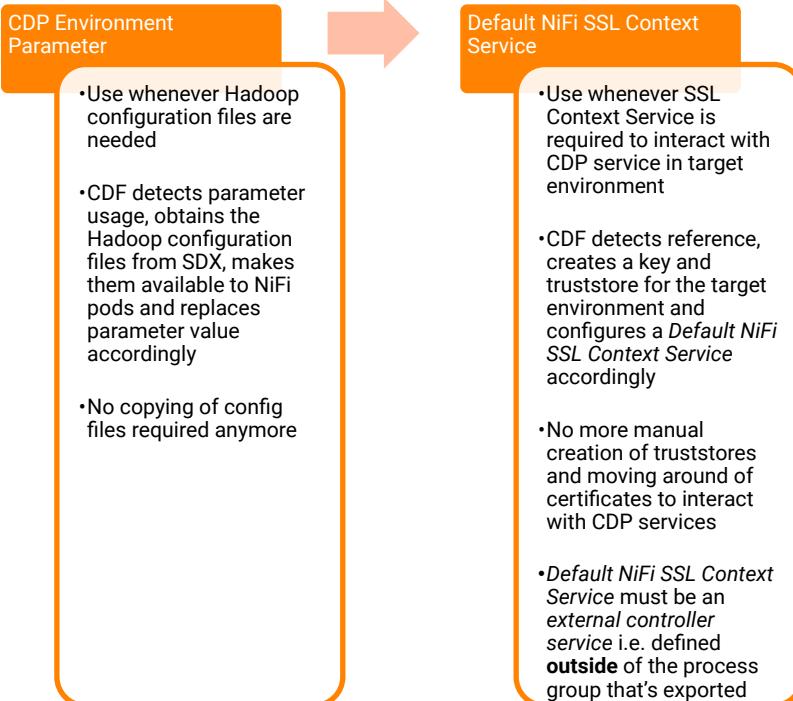
SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Required field			
Property		Value	
Kafka Brokers		#(Kafka Brokers)	
Topic Name(s)		#(Machine Data Topic)	
Topic Name Format		names	

GENERAL CONTROLLER SERVICES

Name	Type	Bundle	State	Scope
AWS Credentials Provider...	AWSCredentialsProviderC...	org.apache.nifi-nifi-aws-n...	Disabled	NiFi Flow
Action Handler Lookup...	ActionHandlerLookup 1.1...	org.apache.nifi-nifi-rules-...	Invalid	MachineDataToWarehouse
CDP SSL Context Service...	StandardRestrictedSSLCo...	org.apache.nifi-nifi-ssl-co...	Invalid	MachineDataToWarehouse
CSV Reader CDP Schema ...	CSVReader 1.11.4.2.0.6.0...	org.apache.nifi-nifi-record...	Invalid	NiFi Flow

# FLOW DEVELOPMENT BEST PRACTICES

## CDPEnvironment Parameter & Default SSLContextService



The screenshot shows two NiFi configuration interfaces side-by-side.

**Configure Controller Service** (Top):

Property	Value
Configuration Resources	#{CDPEnvironment}
Username	#{CDP Workload User}
Password	No value set

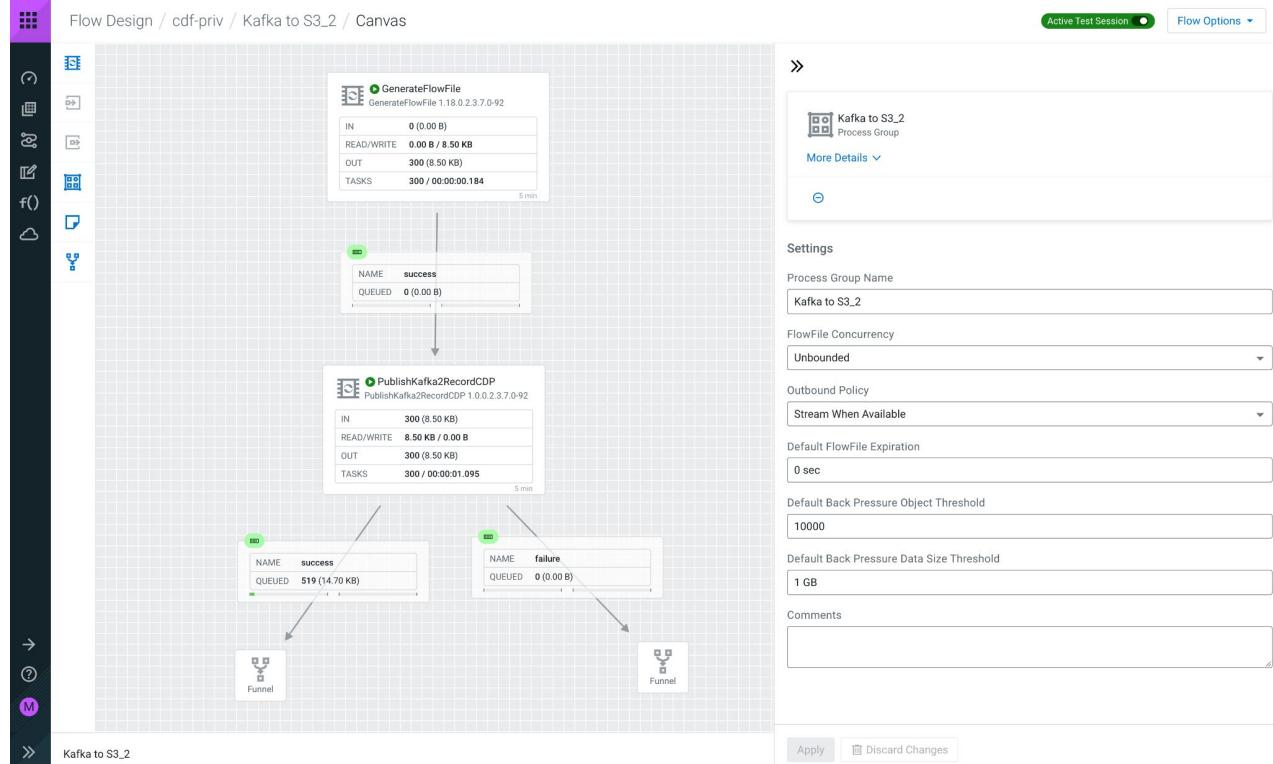
**Configure Processor** (Bottom):

Property	Value
Kerberos Keytab	No value set
Username	#{CDP Workload User}
Password	No value set
Token Auth	false
SSL Context Service	Default NiFi SSL Context Service
Group ID	#{Kafka Consumer Group Id}

In both tables, the "Value" column for the "Configuration Resources" and "SSL Context Service" properties is highlighted with a red box, indicating the use of the CDPEnvironment parameter.

# Data Flow Design for Everyone

- Cloud-native data flow development
- Developers get their own sandbox
- Start developing flows without installing NiFi
- Redesigned visual canvas
- Optimized interaction patterns
- Integration into CDF-PC Catalog for versioning

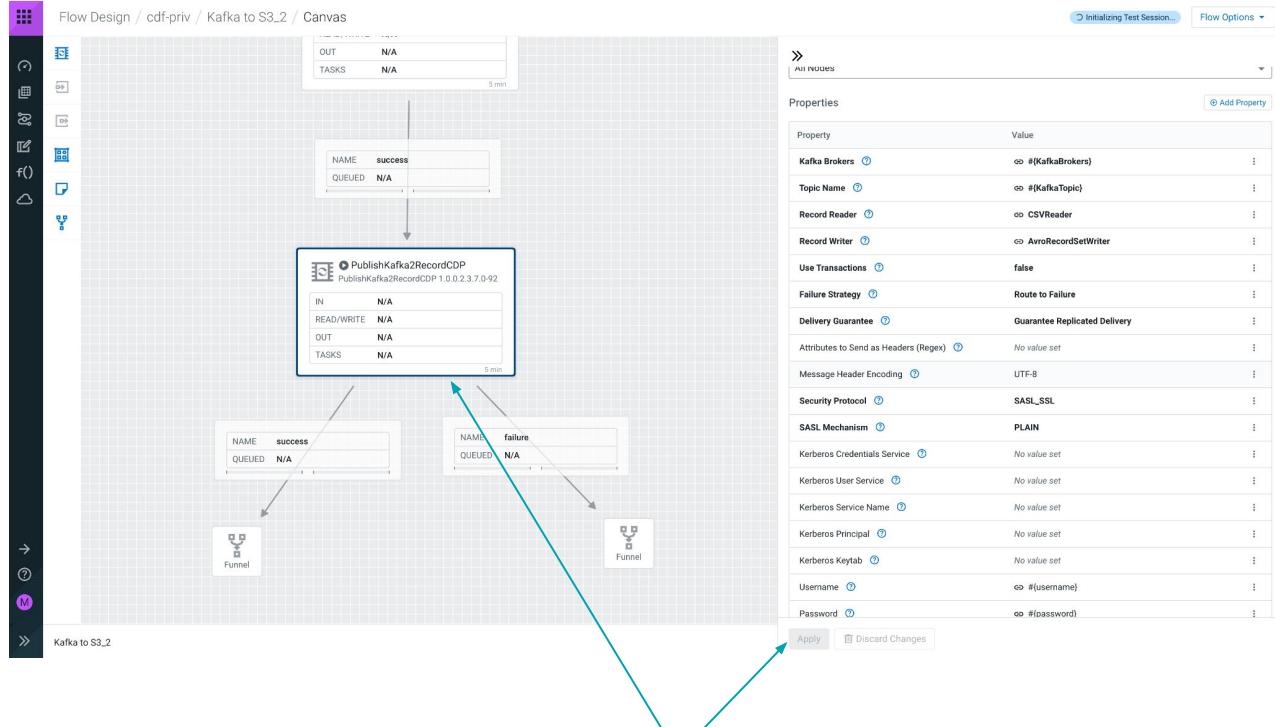


# Context aware configuration panel

Configuration side panel automatically represents selected canvas component

Developers can still navigate on the canvas while having configuration easily accessible

Allows for quick access to configuration while eliminating clicks



Reflects canvas selection

# Simplified Parameters and Controller Services

Manage Services and Parameters centrally for your flow draft

Upload files like JDBC drivers or scripts directly through the Designer UI

Understand impact of changing parameters through *Referenced Components*

DefaultSSLContextService for secure interaction with CDP services is automatically set up

The screenshot shows the Cloudera Data Flow Designer interface. The main area displays a table of parameters for a service named 'AutoDerivedCatWriter'. The table includes columns for Name, Value, and Changed status. Key entries include:

Name	Value	Changed
CDPEnvironment	core-site.xml, ssl-client.xml, hive-site.xml	
Default SSL Context Keystore	/home/nifi/additional/secret/ssl_keystore/ssl-keystore.jks	
Default SSL Context Keystore Password	Sensitive value set	
Default SSL Context Keystore Type	JKS	
Default SSL Context Truststore	/home/nifi/additional/secret/ssl_truststore/ssl-truststore.jks	
Default SSL Context Truststore Password	Sensitive value set	
Default SSL Context Truststore Type	JKS	
jdbcdriver	telemetry-sensor-readings-0.json	
KafkaBrokers	streams-messaging-corebroker1.cdf-priv.xcu2-8y8x.dev.cdr.work:9093,streams-messaging-corebroker1.cdf-priv.xcu2-8y8x.dev.cdr.work:9093,streams-messaging-corebroker2.cdf-priv.xcu2-8y8x.dev.cdr.work:9093	
KafkaTopic	events	
password	Sensitive value set	
username	sr_nifi-kafka-ingest	

Below the table, there are sections for 'Description' (with a placeholder 'Description') and 'Referencing Components'. A single component, 'PublishKafka2RecordCDP', is listed under 'Referencing Processor'.

# Interactive development through test sessions

Start a test session running a specific NiFi version at any point in time

Test sessions provide a NiFi runtime and allow starting/stopping processors and services

Explore data in queues to validate processing logic

Allows you to pin flow file attributes for quick comparison

The screenshot shows the Cloudera DataFlow interface. On the left, a sidebar menu includes options like Dashboard, Catalog, Registry, Flow, Environment, and Metrics. The main area displays two windows: "NiFi Configuration" and "Flow Design / cdf-priv / Kafka to S3\_2 / Test Session". The "NiFi Configuration" window shows a "GenerateFlowFile" processor. The "Flow Design" window shows a "GenerateFlowFile" source connected to a "Kafka to S3\_2" queue. Below these, a "List Queue" table shows 36 flow files, and a detailed view of a single flow file's attributes.

Position	Filename	UUID	File Size	Queue Duration	Lineage Duration	Penalized
1	d02b5fb2-3a1-4ff8-8728-6170e1b...	...	29 B	04:12:34.316	04:12:34.318	No
2	7301a1c83f4-4603-9b43-23adaaa...	...	29 B	04:11:34.183	04:11:34.183	No
3	96a9fd2a-ac23-44ba-a667-1a29031...	...	29 B	04:10:34.019	04:10:34.019	No
4	10f1bd3-07bf-424d-8e08-47386dc...	...	29 B	04:09:34.085	04:09:34.085	No
5	169586bc-405f-47fe-89cc-5c5a9cae...	...	29 B	04:08:33.912	04:08:33.912	No
6	de4f7140-443b-46d1-9100-82a0800...	...	29 B	04:07:34.002	04:07:34.002	No
7	c7089dcd-73e5d-447a-800c-067d6...	...	29 B	04:06:34.108	04:06:34.109	No
8	12f24387-e552-41cf-8c76-8a43352...	...	29 B	04:05:34.220	04:05:34.220	No
9	3e9f9fb3b-d9f1-4467-b5d7-4757cb7d...	...	29 B	04:04:34.279	04:04:34.279	No
10	295536f1-831d-44de-a611-dd8575...	...	29 B	04:03:34.408	04:03:34.408	No
11	9031beff-083d-48fe-x34f-d754380...	...	29 B	04:02:34.419	04:02:34.419	No
12	0f9b6491-7727-4f31-e4b6-8c2ad6f5...	...	29 B	04:01:34.453	04:01:34.453	No
13	b765221-5e09-45db-9990-9919b5f...	...	29 B	04:00:34.278	04:00:34.278	No
14	3a1559e97-8d62-49a1-badd-f523ea3...	...	29 B	03:59:34.361	03:59:34.362	No
15	3a1559e97-8d62-49a1-badd-f523ea3...	...	29 B	03:58:33.420	03:58:33.420	No
16	6919cefb-ab93-42f5-b16d-a98134f5...	...	29 B	03:57:34.582	03:57:34.582	No
17	8cf595e6-e259-4322-ab2d-9a983...	...	29 B	03:56:34.644	03:56:34.645	No
18	03edfd15-8845-49dd-e439-57216691...	...	29 B	03:55:33.717	03:55:34.717	No
19	a510263c-e378-4249-937b-65dc0b1...	...	29 B	03:54:33.802	03:54:33.803	No

# Data Viewer

View your data at every step of the flow

Auto-detects the type and formats accordingly (JSON, Avro, XML, YAML)

Allows you to pin attributes

Download the flowfile content

The screenshot shows the Data Viewer interface with the following details:

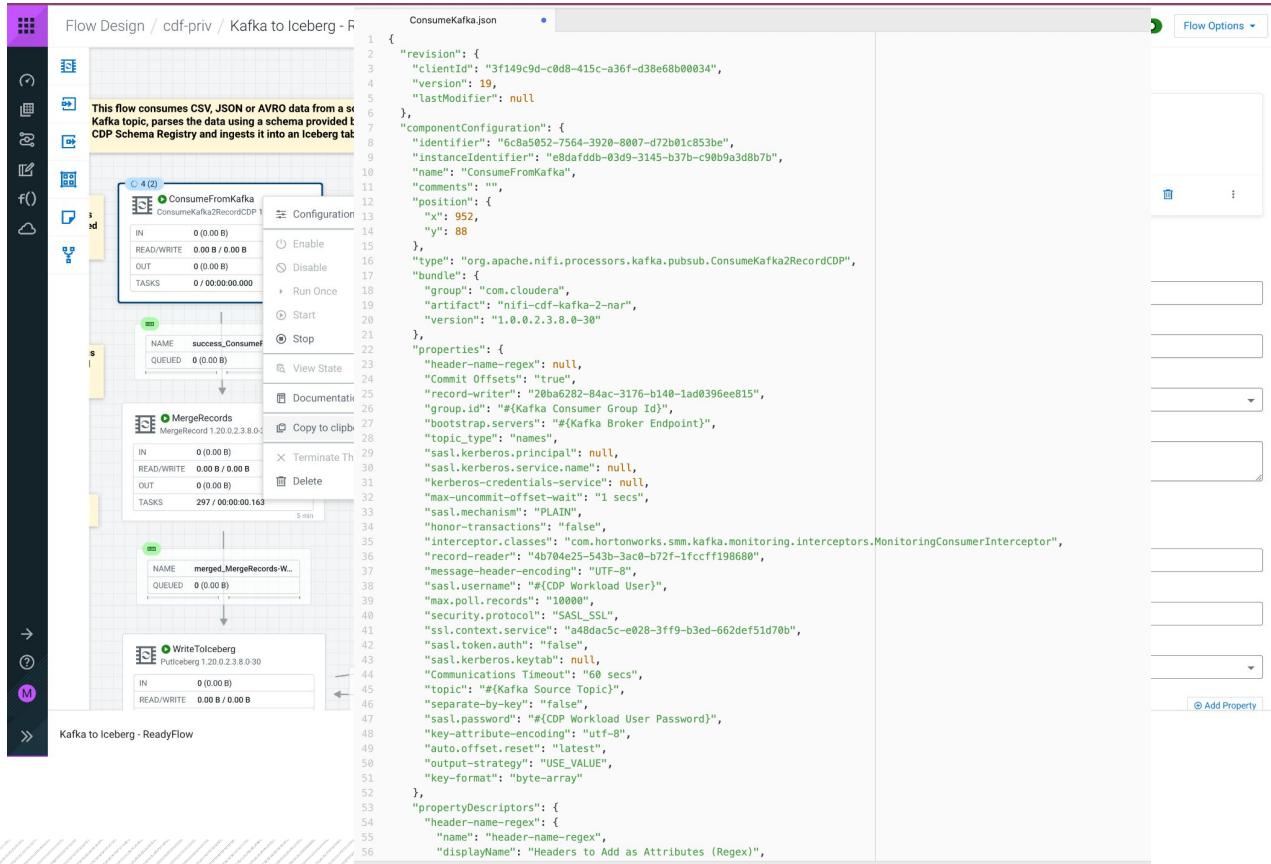
- Content:** A JSON object with five lines of code:

```
1 > [ {  
2   "orderid" : 6453,  
3   "ordertime" : 6453,  
4   "orderaddress" : "299 Thomas E Dunn Memorial Highway, Rutherford, NJ 07070"  
5 } ]
```
- View as:** A dropdown menu set to "Formatted / Avro". Other options include Original, Hex, Formatted, JSON, XML, YAML, and Avro.
- Attributes (16):** A list of pinned attributes:
  - kafka.consumer.id
  - cdf
  - kafka.partition
  - 2
  - mime.type
  - application/avro-binary
  - uuid
  - b8d1d64e-5feb-4036-92cc-d68499d80762
  - merge.bin.age
  - 30004
- Pinned (0) - Copy:** Buttons for managing pinned attributes.

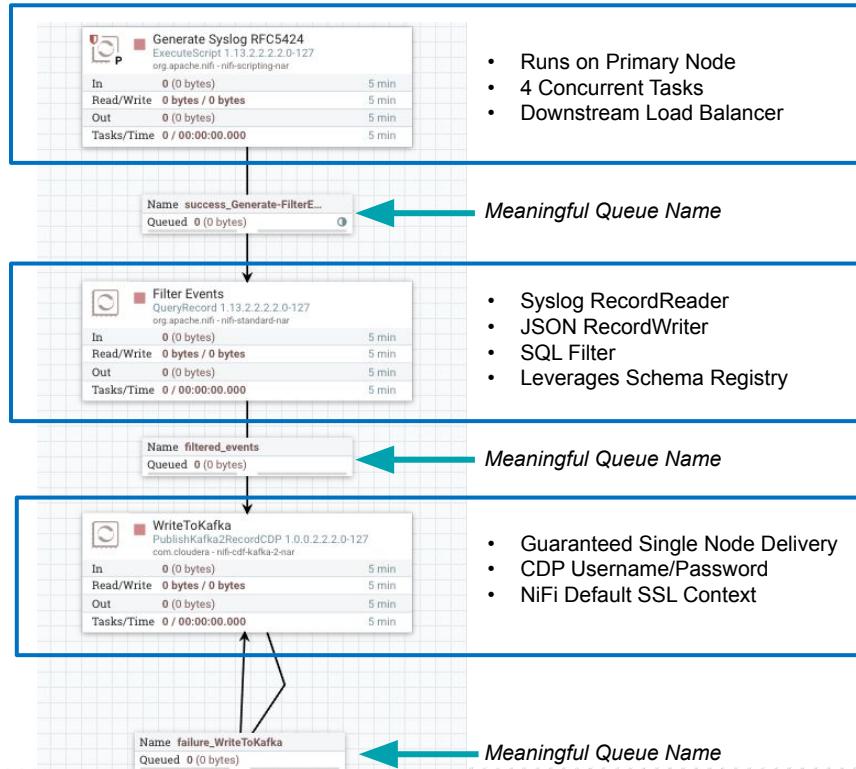
# Copy & Paste

Copy & Paste components  
between different drafts

Paste clipboard content in text  
editor to get JSON representation



# LOG ANALYTICS IMPLEMENTATION



- Runs on Primary Node
- 4 Concurrent Tasks
- Downstream Load Balancer

- Syslog RecordReader
- JSON RecordWriter
- SQL Filter
- Leverages Schema Registry

- Guaranteed Single Node Delivery
- CDP Username/Password
- NiFi Default SSL Context

GENERAL		CONTROLLER SERVICES	
<b>Name</b> ▾			
CDP_Schema_Registry			
Default NiFi SSL Context Service			
JSON_Syslog_5424_Reader			
JSON_Syslog_5424_Writer			
Syslog_5424_Reader			

SETTINGS		PARAMETERS	
<b>Name</b> ▾		<b>Value</b>	
CDP Workload User	?	nismaily	
CDP Workload User Password	?	Sensitive value set	
Filter Rule	?	SELECT * FROM FLOWFILE	
Kafka Broker Endpoint	?	nismaily-kafka-broker1.se-sandb.a46...	
Kafka Destination Topic	?	syslog	
Kafka Producer ID	?	nifi_dh_1	
Schema Name	?	syslog	
Schema Registry Hostname	?	nismaily-kafka-master0.se-sandb.a4...	

# SYSLOG RFC 5424

Generate Syslog RFC5424		
ExecuteScript 1.13.2.2.2.0-127 org.apache.nifi - nifi-scripting-nar		
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

VERSION	PRI	TIMESTAMP	HOSTNAME	APP-NAME	PROCID	MSGID	STRUCTURED-DATA	MSG
<165>1	2003-10-11T22:14:15.003Z	mymachine.example.com	evntslog	-	ID47		[exampleSDID@32473 iut="3" eventSource="Application" eventID="1011"]	BOMAn application event log entry...

- **PRI** – or "priority", Facility (what kind of message) \* 8 + **Severity** (how urgent is the message)
- **VERSION** – version is always "1" for RFC 5424
- **TIMESTAMP** – valid timestamp examples (must follow ISO 8601 format with uppercase "T" and "Z")
- **HOSTNAME** – using FQDN (fully qualified domain name) is recommended
- **APP-NAME** – usually the name of the device or application that provided the message
- **PROCID** – often used to provide the process name or process ID (is - "nil" in the example)
- **MSGID** – should identify the type of message
- **STRUCTURED-DATA** – named lists of key-value pairs for easy parsing and searching
- **MSG** – details about the event

Numerical Code	Severity
0	Emergency: system is unusable
1	Alert: action must be taken immediately
2	Critical: critical conditions
3	Error: error conditions
4	Warning: warning conditions
5	Notice: normal but significant condition
6	Informational: informational messages
7	Debug: debug-level messages

# QUEUE CONFIGURATION

- **FlowFile Expiration** - Data that cannot be processed in a timely fashion can be automatically removed from the flow
- **Back Pressure Thresholds** - Thresholds indicate how much data should be allowed to exist in the queue before the component that is the source of the Connection is no longer scheduled to run. This allows the system to avoid being overrun with data
- **Load Balance Strategy** – Strategy to distribute the data in a flow across the nodes in the cluster. When enabled, compression can be configured on FlowFile contents and attributes
- **Prioritization** – Determines the order in which flow files are processed

	Generate Syslog RFC5424 ExecuteScript 1.13.2.2.2.0-127 org.apache.nifi - nifi-scripting-nar
In	0 (0 bytes) 5 min
Read/Write	0 bytes / 0 bytes 5 min
Out	0 (0 bytes) 5 min
Tasks/Time	0 / 00:00:00.000 5 min

### Configure Connection

DETAILS SETTINGS

Name: success\_Generate-FilterEvents

Available Prioritizers:

- FirstInFirstOutPrioritizer
- NewestFlowFileFirstPrioritizer
- OldestFlowFileFirstPrioritizer
- PriorityAttributePrioritizer

Id: 64146cca-d197-3c27-9c47-015dd7b7a6c6

FlowFile Expiration: 0 sec

Back Pressure Object Threshold: 10000

Size Threshold: 1 GB

Selected Prioritizers: (empty)

Load Balance Strategy: Round robin

Load Balance Compression: Do not compress

# RECORD-ORIENTED DATA WITH NIFI

- **Record Readers** - Avro, CSV, Grok, IPFIX, JSAN1, JSON, Parquet, Scripted, Syslog5424, Syslog, WindowsEvent, XML
- **Record Writers** - Avro, CSV, FreeFromText, Json, Parquet, Scripted, XML
- Record Reader and Writer support referencing a schema registry for retrieving schemas when necessary.
- Enable processors that accept any data format without having to worry about the parsing and serialization logic.
- Allows us to keep FlowFiles larger, each consisting of multiple records, which results in far better performance.

Filter Events		
QueryRecord 1.13.2.2.2.2.0-127 org.apache.nifi - nifi-standard-nar		
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Required field			
Property	Value		
Record Reader	CSVReader	→	
Record Writer	JsonRecordSetWriter	→	

# RUNNING SQL ON FLOWFILES

- Evaluates one or more SQL queries against the contents of a FlowFile.
- This can be used, for example, for field-specific filtering, transformation, and row-level filtering.
- Columns can be renamed, simple calculations and aggregations performed.
- The SQL statement must be valid ANSI SQL and is powered by Apache Calcite.

Filter Events		
QueryRecord 1.13.2.2.2.2.0-127 org.apache.nifi - nifi-standard-nar		
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Configure Processor | QueryRecord 1.13.2.2.2.2.0-127

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
Record Reader	Syslog_5424_Reader
Record Writer	JSON_Syslog_5424_Writer
Include Zero Record FlowFiles	false
Cache Schema	false
Default Decimal Precision	10
Default Decimal Scale	0
filtered_events	#{\$Filter Rule}

# CLOUDERA DATA FLOW – PUBLIC CLOUD

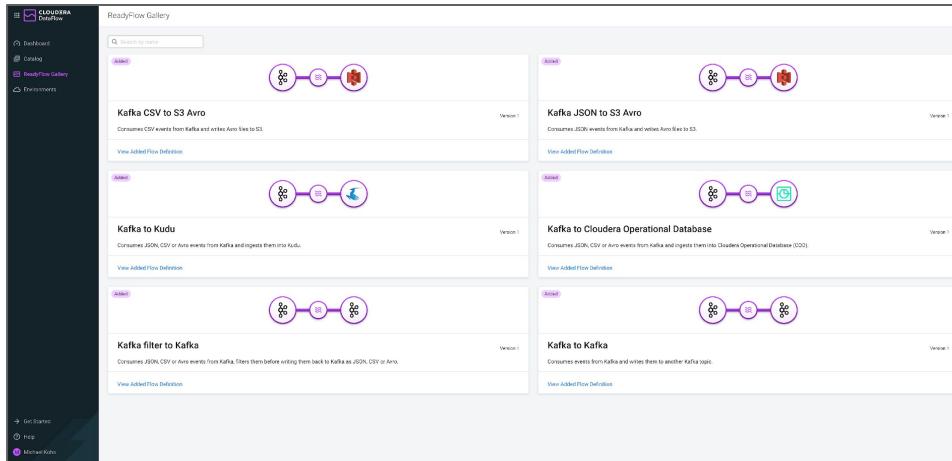


# READYFLOWS - FOR THE MOST COMMON USE CASES

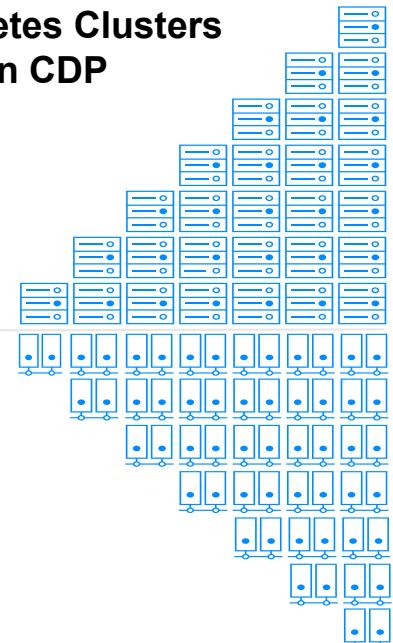
For new NiFi users

## ReadyFlows Gallery

A list of pre-defined flows called **ReadyFlows** to accelerate flow authorship and deployment



## Auto-Scale Kubernetes Clusters on CDP



Select a ReadyFlow  
and use the  
deployment wizard

# READYFLOW GALLERY

- Cloudera provided flow definitions
- Cover most common data flow use cases
- Optimized to work with CDP sources/destinations
- Can be deployed and adjusted as needed

ReadyFlow Gallery

Search by name

Added

**Kafka filter to Kafka** Version 1

Consumes JSON, CSV or Avro events from Kafka, filters them before writing them back to Kafka as JSON, CSV or Avro.

[View Added Flow Definition](#)

Added

**Kafka to Cloudera Operational Database** Version 1

Consumes JSON, CSV or Avro events from Kafka and ingests them into Cloudera Operational Database (COD).

[View Added Flow Definition](#)

**Kafka to Kafka** Version 1

Consumes events from Kafka and writes them to another Kafka topic.

[Add To Catalog](#)

**Kafka to Kudu** Version 1

Consumes JSON, CSV or Avro events from Kafka and ingests them into Kudu.

[Add To Catalog](#)

**Kafka to S3 Avro** Version 1

Consumes JSON, CSV or Avro events from Kafka and writes Avro files to S3.

[View Added Flow Definition](#)

**S3 to S3 Avro** Version 1

Consumes JSON, CSV or Avro files from source S3 location and writes Avro files to a destination S3 location.

[Add To Catalog](#)

# FLOW CATALOG

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

The screenshot shows the Cloudera DataFlow interface with the 'Catalog' tab selected. The main area is titled 'Flow Catalog' and contains a table of flow definitions. A search bar at the top allows users to 'Search by name'. A blue button labeled 'Import Flow Definition' is located in the top right corner, along with a timestamp indicating the catalog was 'REFRESHED 25 seconds ago'.

Name ↑	Type	Versions	Last Updated	
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 of the catalog page, there are navigation controls for 'Items per page' (set to 10), a page number indicator '1 - 10 of 10', and standard left and right arrow navigation buttons.

# TURNS FLOW DEFINITIONS INTO FLOW DEPLOYMENTS

## 1.) Start Deployment Wizard

se-sandboxx-aws / New Deployment

Overview

Deployment Name: abc\_hello\_world  
Deployment name is valid

Selected Flow Definition: Hello World

Target Environment: se-sandboxx-aws

Review

## 2.) NiFi Config

se-sandboxx-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-sandboxx-aws

DEPLOYMENT NAME abc\_hello\_world

## 3.) Provide Parameters for NiFi

se-sandboxx-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 ↗

Review

## 4.) Configure Sizing & Scaling

se-sandboxx-aws / New Deployment

Sizing & Scaling

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

NIFI Node Sizing

Small

Medium

Large

Number of NiFi Nodes

Auto Scaling: Enabled

Min. Nodes: 1

Max. Nodes: 32

Review

## 5.) Define KPIs

se-sandboxx-aws / New Deployment

Key Performance Indicators

Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.

Learn more ↗

Entire Flow

METRIC TO TRACK: Data Out

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-sandboxx-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

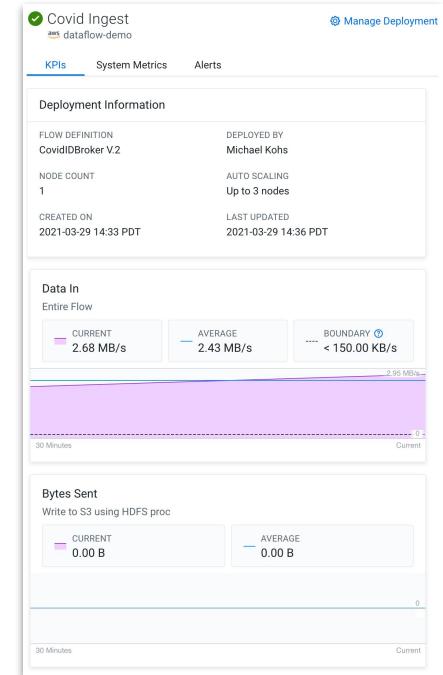
# KEY PERFORMANCE INDICATORS

- 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

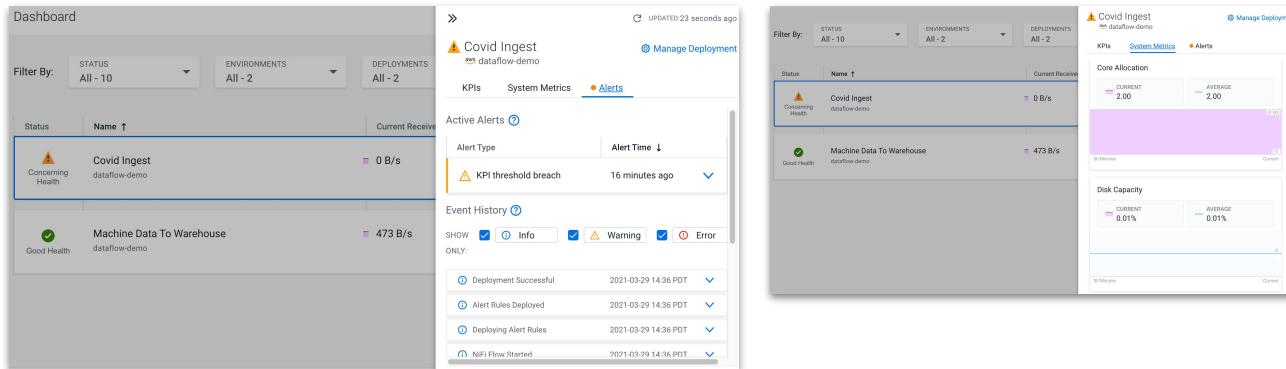
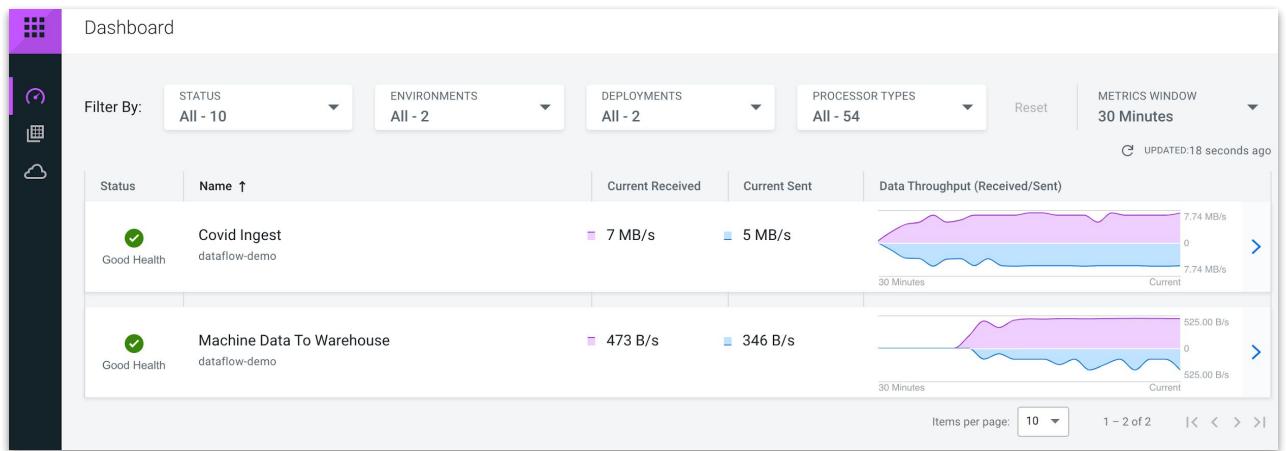
The screenshot shows the 'New Deployment' step of the deployment wizard. On the left, a sidebar lists steps: Overview, NiFi Configuration, Parameters, Sizing & Scaling, Key Performance Indicators (selected), and Review. The main area is titled 'Key Performance Indicators' with the sub-section 'Entire Flow'. It shows 'METRIC TO TRACK' set to 'Flow Files Queued', 'ALERT SET' as 'No alert set', and a note: 'Notify if outside the range of 999 MB/sec - 1 MB/sec, for at least 5 minutes.' A button at the bottom right says '(+) Add New KPI'. To the right, there are two panels: 'Overview' (Flow Definition: Hello World v1, Environment: se-sandbox-aws, Deployment Name: abc\_hello\_world) and 'NiFi Configuration' (NIIF Runtime Version: Latest Version (1.18.0.2.3.7.1-1), Auto-start Flow: Yes, Inbound Connections: No, Custom NAR Configuration: No). Below these is a 'Parameters' panel stating 'No parameters are available for this flow'.

## 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 V1

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:

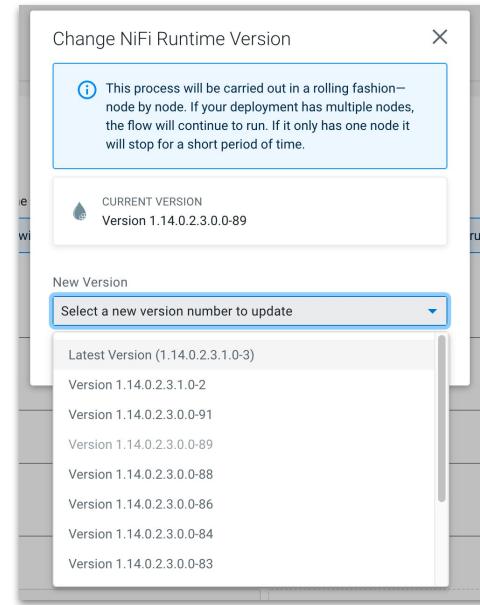
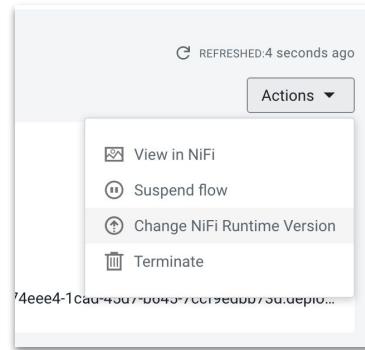
CDP Workload User Password:

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

Discard Changes Apply Changes

# NIFI VERSION UPGRADES

- Pick up NiFi hotfixes easily
- Upgrade (or downgrade) the hotfix version of existing deployments
- Rolling upgrade (if the deployment has >1 NiFi nodes)



---

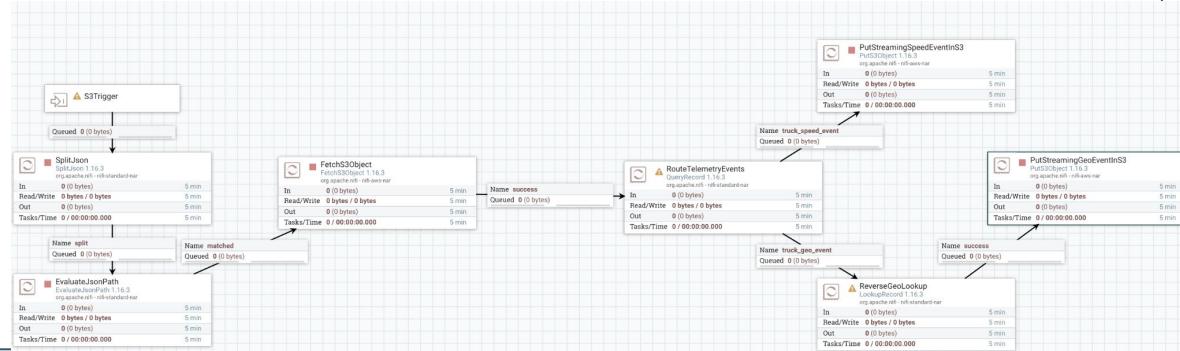
# Deploy a DataFlow Function

---

DataFlow Functions provides an  
**efficient, cost optimized, scalable** way to  
run NiFi flows in a completely **serverless**  
fashion for **event-driven** use cases.

# Development & Runtime of DataFlow Functions

**Step1. Develop** functions on local workstation or in CDP Public Cloud using **no-code**, UI designer



**Step 2. Run** functions on serverless compute services in AWS, Azure & GCP



AWS Lambda



Azure Functions



Google Cloud Functions

# DataFlow Functions Use Cases

## Trigger Based, Batch, Scheduled and Microservice Use Cases

### Serverless Trigger-Based File Processing Pipeline

Develop & run data processing pipelines when files are created or updated in any of the cloud object stores

**Example:** When a photo is uploaded to object storage, a data flow is triggered which runs image resizing code and delivers resized image to different locations.

### Serverless Workflows / Orchestration

Chain different low-code functions to build complex workflows

**Example:** Automate the handling of support tickets in a call center or orchestrate data movement across different cloud services.

### Serverless Scheduled Tasks

Develop and run scheduled tasks without any code on pre-defined timed intervals

**Example:** Offload an external database running on-premises into the cloud once a day every morning at 4:00 a.m.

### Serverless Microservices

Build and deploy serverless independent modules that power your applications microservices architecture

**Example:** Event-driven functions for easy communication between thousands of decoupled services that power a ride-sharing application.

### Serverless Web APIs

Easily build endpoints for your web applications with HTTP APIs without any code using DFF and any of the cloud providers' function triggers

**Example:** Build high performant, scalable web applications across multiple data centers.

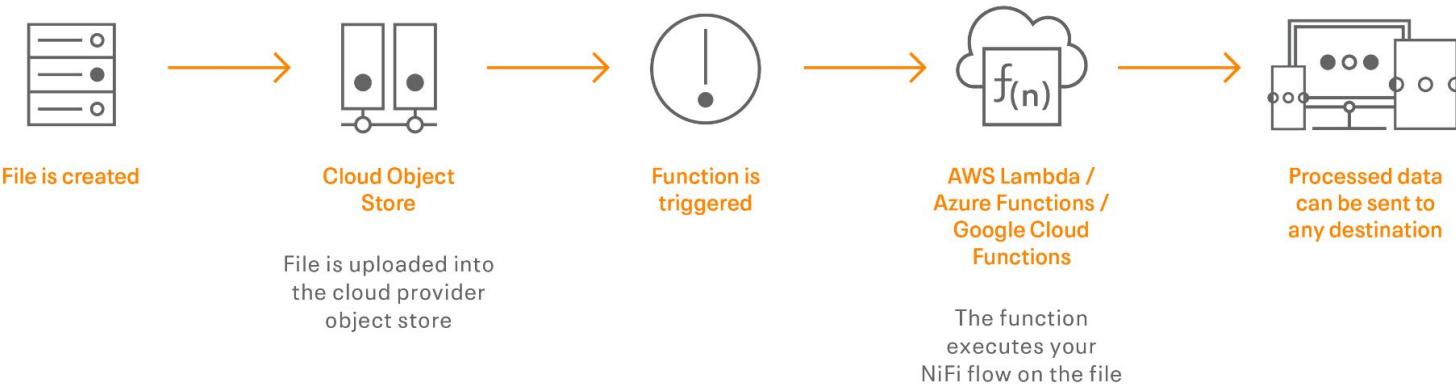
### Serverless Customized Triggers

With the DFF State feature, build flows to create customized triggers allowing access to on-premises or external services

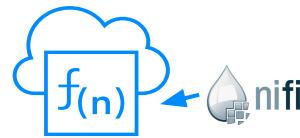
**Example:** Near real time offloading of files from a remote SFTP server.

# Sample DataFlow Functions Use Case

DataFlow Functions easily enables near real-time file processing in a serverless architecture



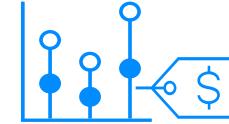
# Summary



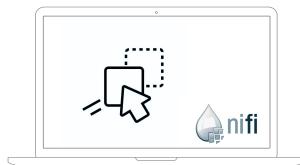
Serverless NiFi  
New Use Cases



Faster ROI



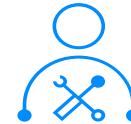
Pay for Usage  
True Pay for Value



No code UI  
Rapid dev & test



Lower TCO  
Cost optimization



Reduce  
Operational Overhead

# CDF-PC: Run serverless NiFi with DataFlow Functions

## f(x) True serverless compute

Leverages AWS Lambda, Azure Functions or Google Cloud Functions for compute. No servers to manage.

The screenshot shows the AWS Lambda console interface for the 'CustomerProcessing' function. It includes sections for Overview, Trigger configuration, Code source, and Code properties. A note at the bottom indicates that the deployment package is too large for inline editing.



## Execute on events

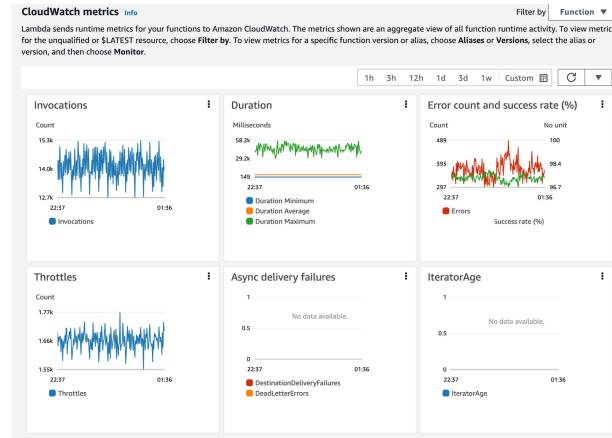
Supports cloud provider native triggers to launch a function.

The screenshot shows the 'Add trigger' configuration screen in the AWS Lambda console. It lists various AWS services as potential triggers, such as SNS, Kinesis, and CloudWatch Metrics.



## Guaranteed delivery

Acknowledges receipt of event to source system only once it has been delivered to the destination .



# CDF-PC Deployments: Resource Isolation & Monitoring



## Resource Isolation

Turn process groups into separate flow deployments and assign minimum and maximum resources

The screenshot shows two separate flow deployment panels in the NiFi interface:

- Azure\_Event\_Hub\_to\_ADLS\_v1**: Shows metrics for Queued, In, Read/Write, and Out operations. A context menu is open over the Out operation, showing options like "Configure", "Parameters", "Variables", "Enter group", "Start", "Stop", "Enable", "Disable", "Enable all controller services", "Disable all controller services", "View status history", "View connections", "Center in view", "Group", "Download flow definition", "Create template", "Copy", "Empty all queues", and "Delete".
- Log Processing**: Shows metrics for Queued, In, Read/Write, and Out operations. A context menu is open over the Out operation, showing options like "Configure", "Parameters", "Variables", "Enter group", "Start", "Stop", "Enable", "Disable", "Enable all controller services", "Disable all controller services", "View status history", "View connections", "Center in view", "Group", "Download flow definition", "Create template", "Copy", "Empty all queues", and "Delete".



## Central Monitoring

Monitor health and performance of all flow deployments across environments or clouds in a single dashboard

The screenshot shows a NiFi Dashboard with the following interface elements:

- Filter By:** STATUS All - 13
- ENVIRONMENTS:** 2 of 12
- DEPLOYMENTS:** All - 18
- Search Bar:** Search
- Status Legend:** Good Health (Green), Degraded Health (Yellow), Critical Health (Red)
- Table Headers:** Status, Name ↓, Received, Current
- Table Data:**
  - Zero-To-Dashboard (se-sandbox-aws): Good Health, Received: 0 B/, Current: 0 B/
  - Stock market Iceberg (se-sandbox-aws): Good Health, Received: 0 B/, Current: 0 B/



## Inbound Connections

Send data from clients to flow deployments for further distribution and leave the load balancer, DNS and security configuration to CDF-PC

### Inbound Connections

The screenshot shows the "Inbound Connections" configuration section in the NiFi interface:

- Allow NiFi to receive data:**
- Endpoint Hostname:** syslogtokafka.inbound.dfx.pu5t6mxt.a465-9q4k.cloudera.site
- This endpoint is valid:**
- Listening Ports:**
  - Protocol: TCP, Port: 7800
  - Protocol: TCP, Port: 8100
- Edit Ports**

# CDF-PC Deployments: Auto-Scaling & Custom NARs



## Configure Auto Scaling with Cost Controls

For each flow select container node size and min/max node count for cost control

### Sizing & Scaling

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

#### NiFi Node Sizing

<input checked="" type="radio"/> Extra Small	<input type="radio"/> Small	<input type="radio"/> Medium	<input type="radio"/> Large
2 vCores Per Node 4 GB Per Node	3 vCores Per Node 6 GB Per Node	6 vCores Per Node 12 GB Per Node	12 vCores Per Node 24 GB Per Node

#### Number of NiFi Nodes

##### Auto Scaling

Enabled

Min. Nodes      Max. Nodes

 - 


## Zero Data Loss Guarantee when scaling down

Support scaling down which requires complex coordination ensuring existing data sheds to other avail nodes.



## Support for Custom Nars

Run your existing NiFi data flows that rely on custom NARs/components

The screenshot shows the NiFi Configuration interface for a new deployment named 'rnd-sysIntegration-usWest-ca-xx / New Deployment'. The process is at step 6: Review. On the left, a vertical list of steps is shown: Overview, NiFi Configuration (selected), Parameters, Sizing and Scaling, KPIs, and Review. In the main area, the 'Custom Processor Configuration' section is expanded, containing fields for 'CDP Workload Username', 'Password', 'Confirm Password', and 'Storage Location' (set to 's3a://bucket\_name/folder1/folder2/file1.json').

# CDF-PC: Upgrades & Automation



## High velocity NiFi releases

New NiFi releases & Hotfixes can be shipped at any time and are immediately available for flow deployments

The screenshot shows the Cloudera Manager interface for a 'New Flow Deployment'. On the left, a sidebar lists steps: 1 Overview, 2 NiFi Configuration (selected), 3 Parameters, 4 Sizing & Scaling, 5 Key Performance Indicators, and 6 Review. The main area is titled 'NiFi Configuration' and shows the 'NiFi Runtime Version' as 'CURRENT VERSION Latest Version (1.16.0.2.3.4.0-33)'. A modal window titled 'Change NiFi Runtime Version' is open, showing a dropdown menu with the same option: 'Latest Version (1.16.0.2.3.4.0-33)'.



## Powerful CLI for automation

Automate the entire flow lifecycle with the CLI including single command flow deployment

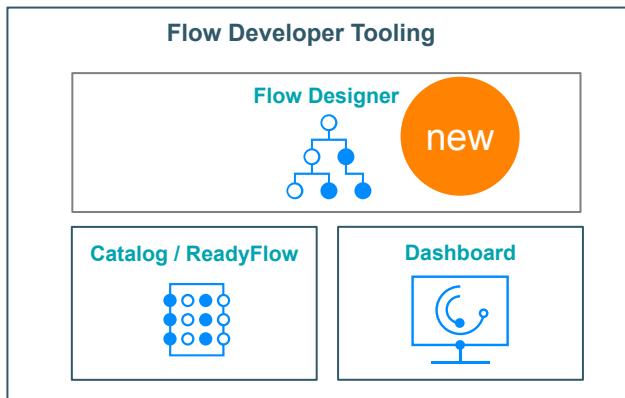
The screenshot shows a terminal window with a copy-paste box containing the following command:

```
1 cdp df create-deployment \
2   --service-crn crn:cdp:df:us-west-1:558bc1d2-8867-4357-8524-311d51259233:service:fb5a721e-e893-40b7-8b53-b837f95adfca
3   --flow-version-crn "crn:cdp:df:us-west-1:altus:readyFlow:Azure-Event-Hub-to-ADLS/v.1"
4   --deployment-name "Syslog to Kafka"
5   --fm-nifi-version 1.16.0.2.3.4.0-33 \
6   --auto-start-flow \
7   --cluster-size-name EXTRA_SMALL \
8   --static-node-count 1 \
9   --no-auto-scaling-enabled \
10  --parameter-groups "file://<>PATH_TO_UPDATE>>/Syslog to Kafka-parameter-groups.json"
```

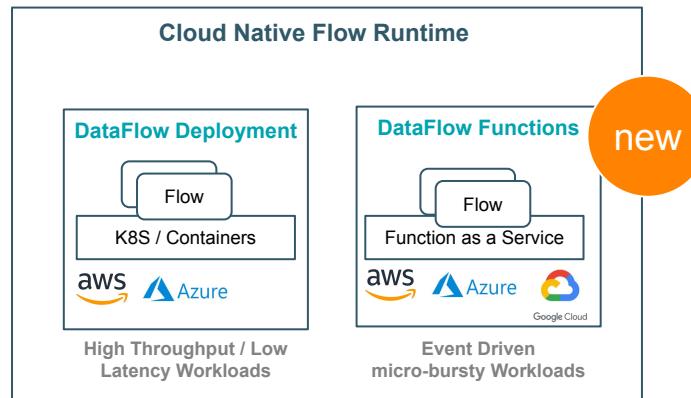
# CLOUDERA DATAFLOW FOR THE PUBLIC CLOUD (CDF-PC)

## Cloud Native Data Distribution Powered by Apache NiFi

**Solves the First & Last Mile Problem** -- Easily connect to any data born on the edge, on-prem or in the cloud and deliver it to any destination.



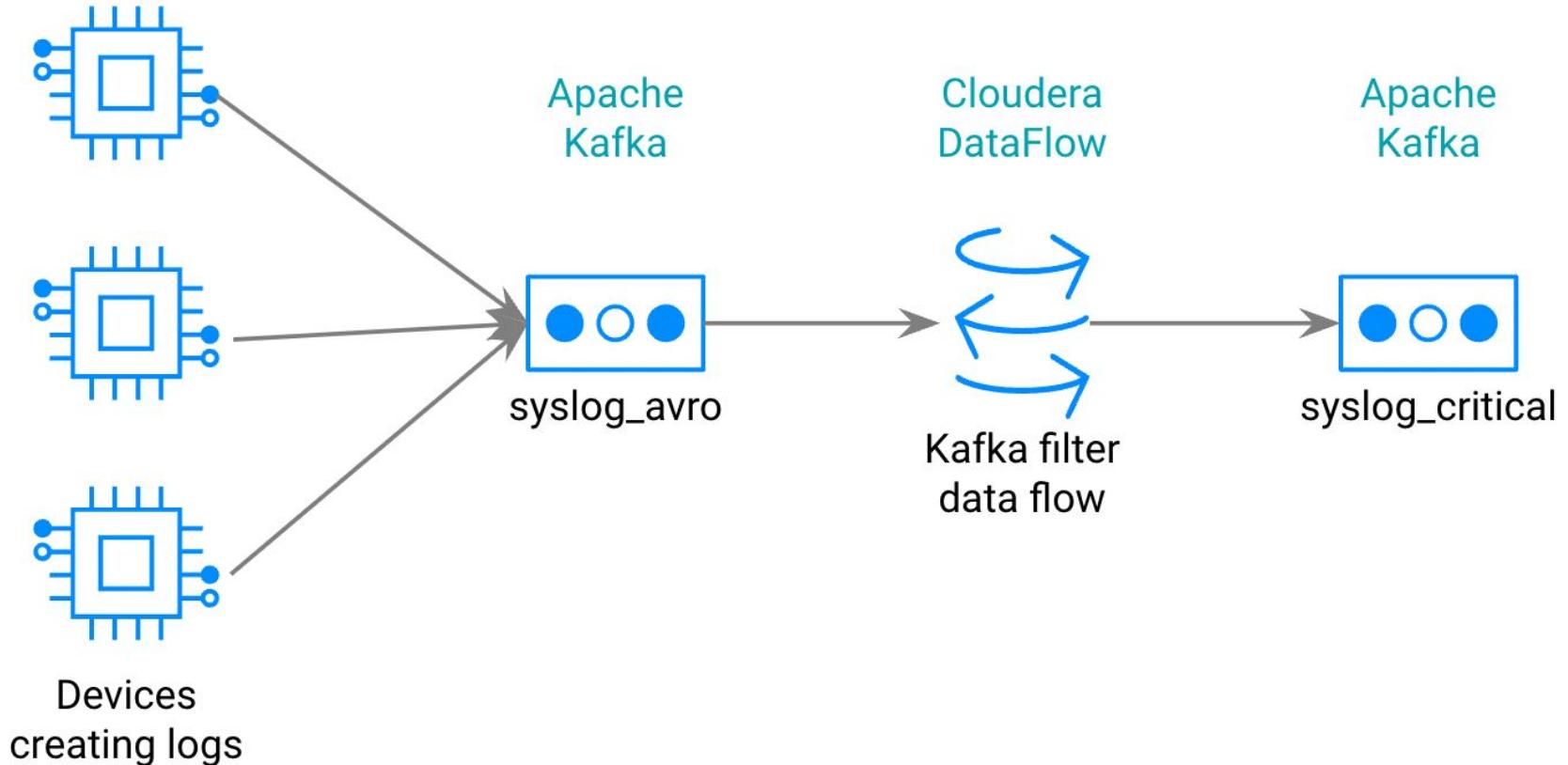
**Productivity Tooling for Developers** – Flow designer combined w/ catalog of flows provides developers the agility & extensibility to build data movement flows in minutes

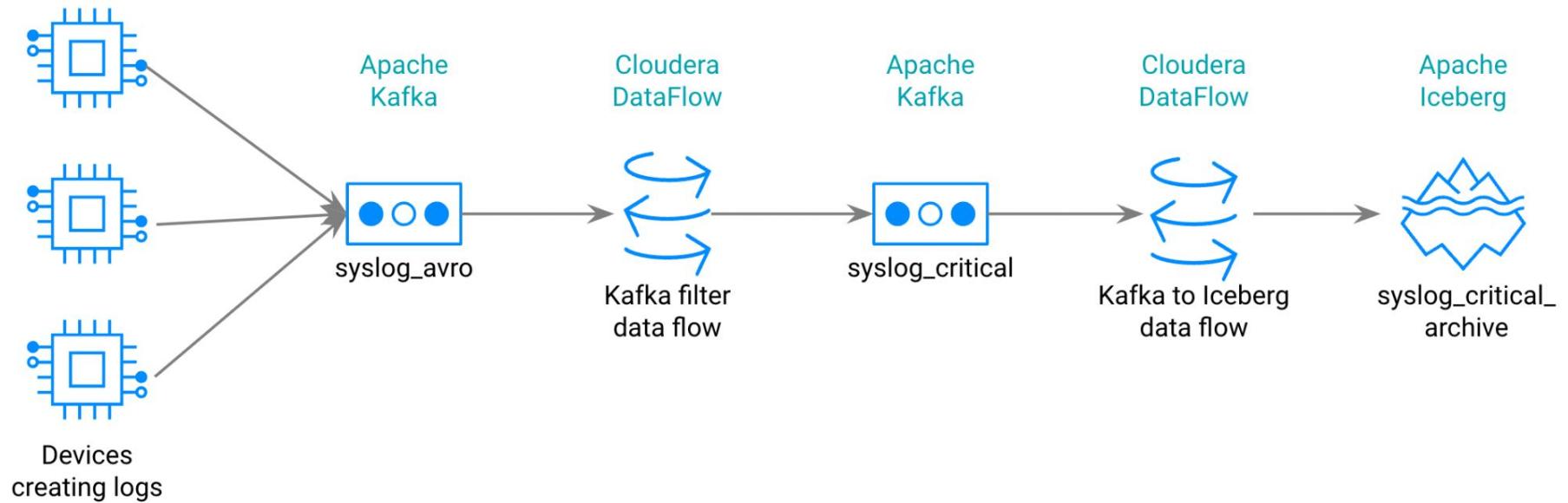


**Cloud Native Flow Runtime** – Multi-Cloud support for deploying flows on auto-scaling K8S NiFi clusters or as serverless functions in any cloud providers' Function as a Service runtime

# Resources

- [New - GA Announcement Blog Post](#)
- [New - Technical Blog: Self-service data pipeline development](#)
- [New - DataFlow Designer Product Tour](#)
- [New - Kafka to Iceberg Demo Video](#)
- [New - Kafka to Snowflake Demo Video](#)
- [New - What's New Post](#)
- [Deploying Functions](#)
- [Updated - Product Page](#)
- [Updated - Product Documentation](#)
- [Universal Data Distribution Blog series](#)





## DAILY ZOOM

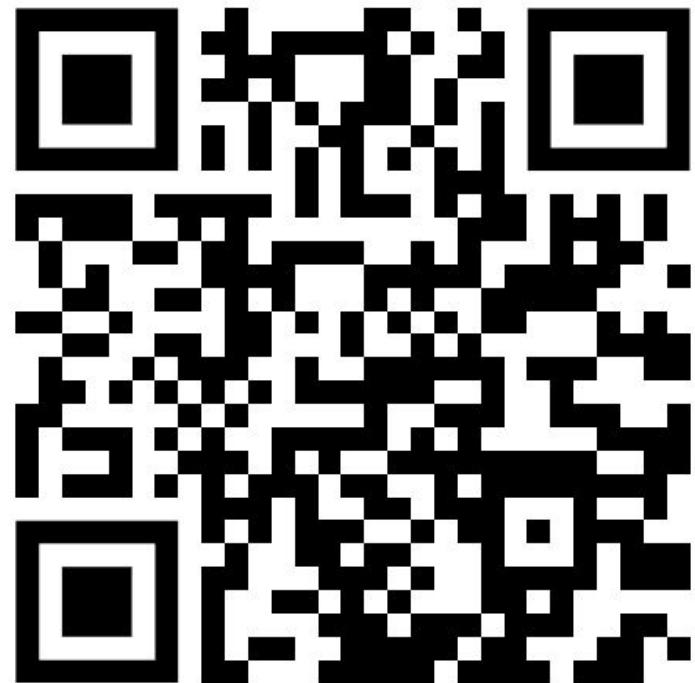
<https://cloudera.zoom.us/j/96460893376?pwd=eWZEVDhpZmpFSDNRejFzMXkvcHpOdz09>



## SLACK CHANNEL

---

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



# SOURCE CODE AND EXAMPLES

---

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



## Submit Your Flow



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



## ReadyFlow Gallery



Iceberg X

Added



### Kafka to Iceberg

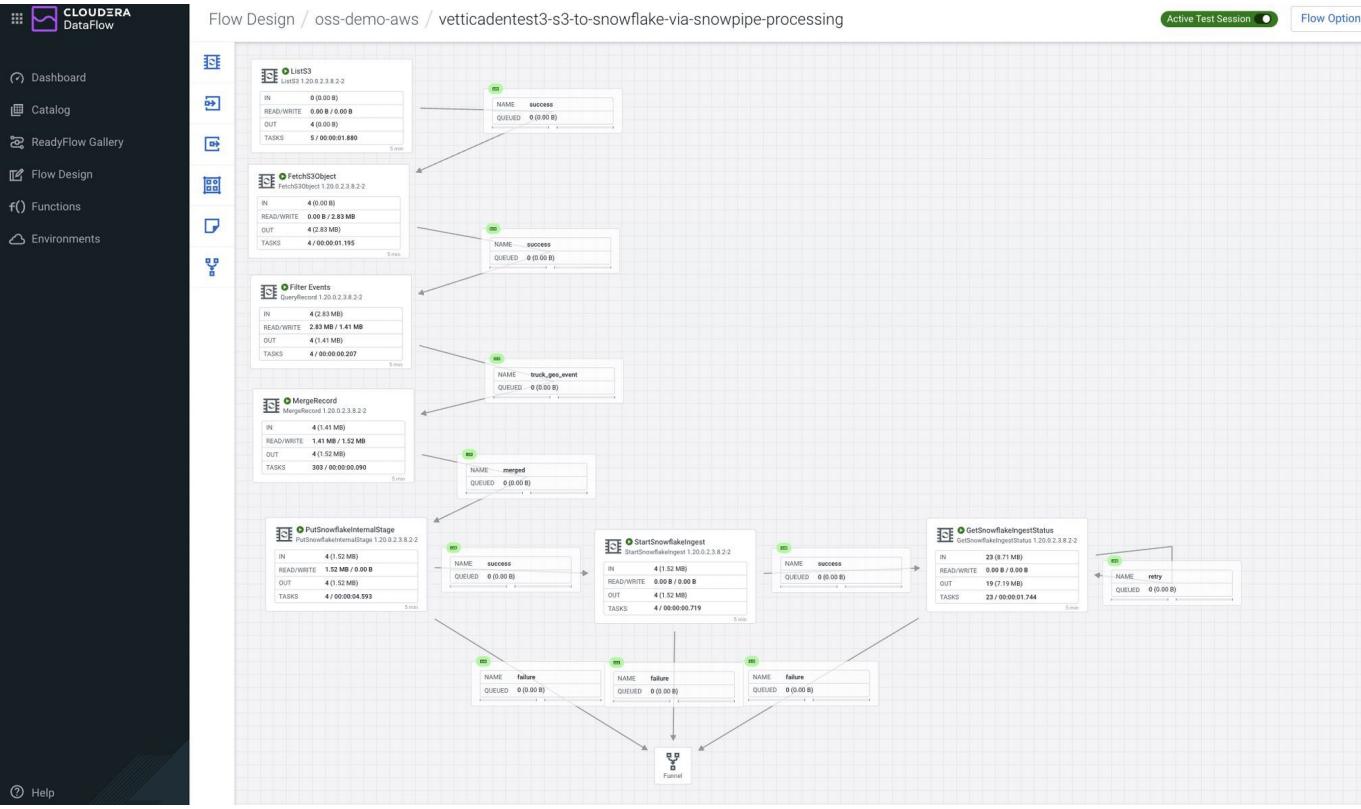
Version 1

Consumes JSON, CSV or Avro events from Kafka and writes them as Parquet files to a destination Iceberg table.

[View Added Flow Definition](#)

[Create New Draft](#)

# EXAMPLE



[https://www.linkedin.com/posts/georgevetticaden\\_just-finished-up-a-trial-run-of-the-new-dat-a-activity-7058557234556907520-W6O2/](https://www.linkedin.com/posts/georgevetticaden_just-finished-up-a-trial-run-of-the-new-dat-a-activity-7058557234556907520-W6O2/)

---

# HELP

---

## Tim Spann

@PaasDev // Blog: [www.datainmotion.dev](http://www.datainmotion.dev)

Principal Developer Advocate

Princeton Future of Data Meetup

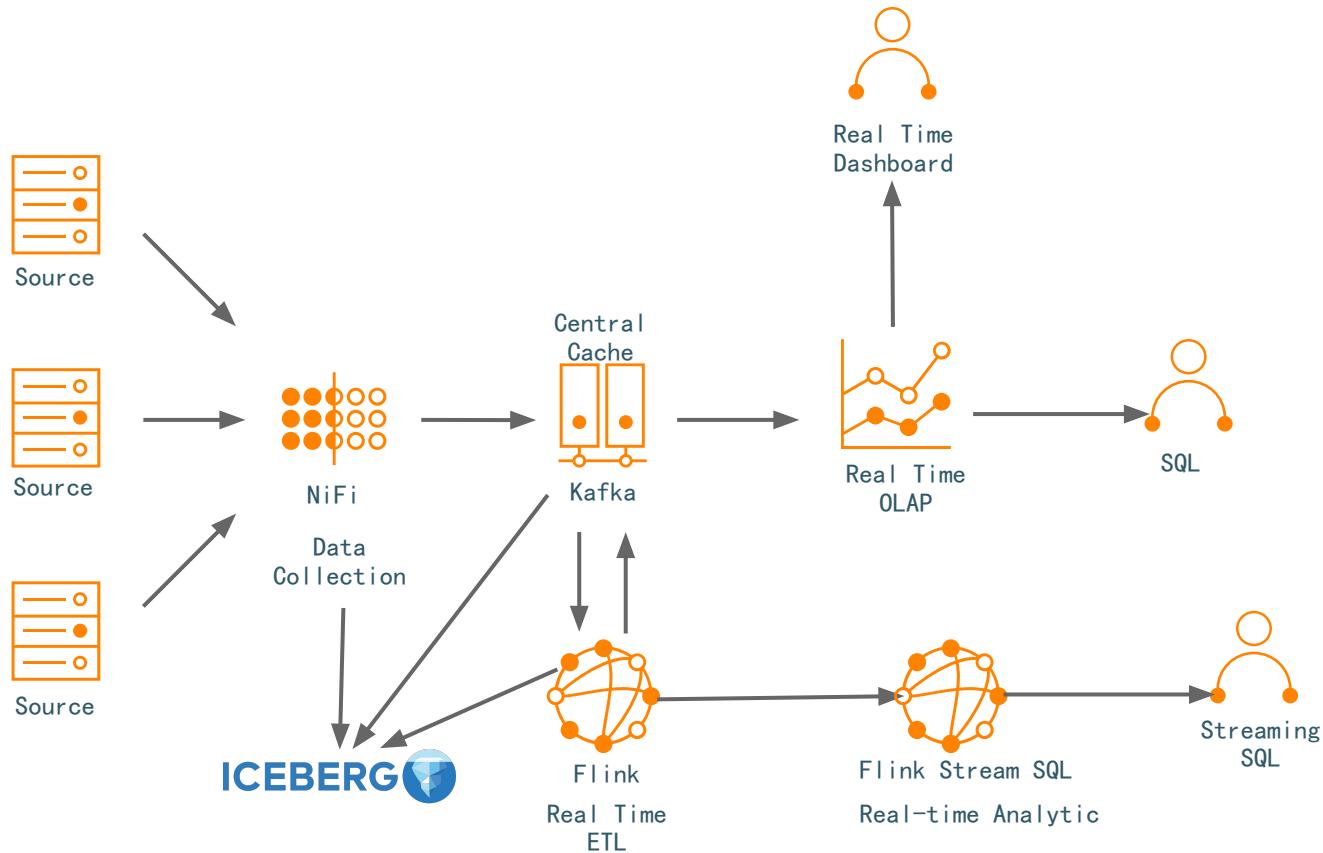
<https://medium.com/@tspann>

<https://github.com/tspannhw>

<https://bestinflow.slack.com/>

<https://docs.cloudera.com/dataflow/cloud/release-notes/topics/cdf-whats-new-latest.html>

Contact Me!



# THANK YOU

CLOUDERA