Best Practices for Developing With Data Flow

21-March-2023

IMPORTANT

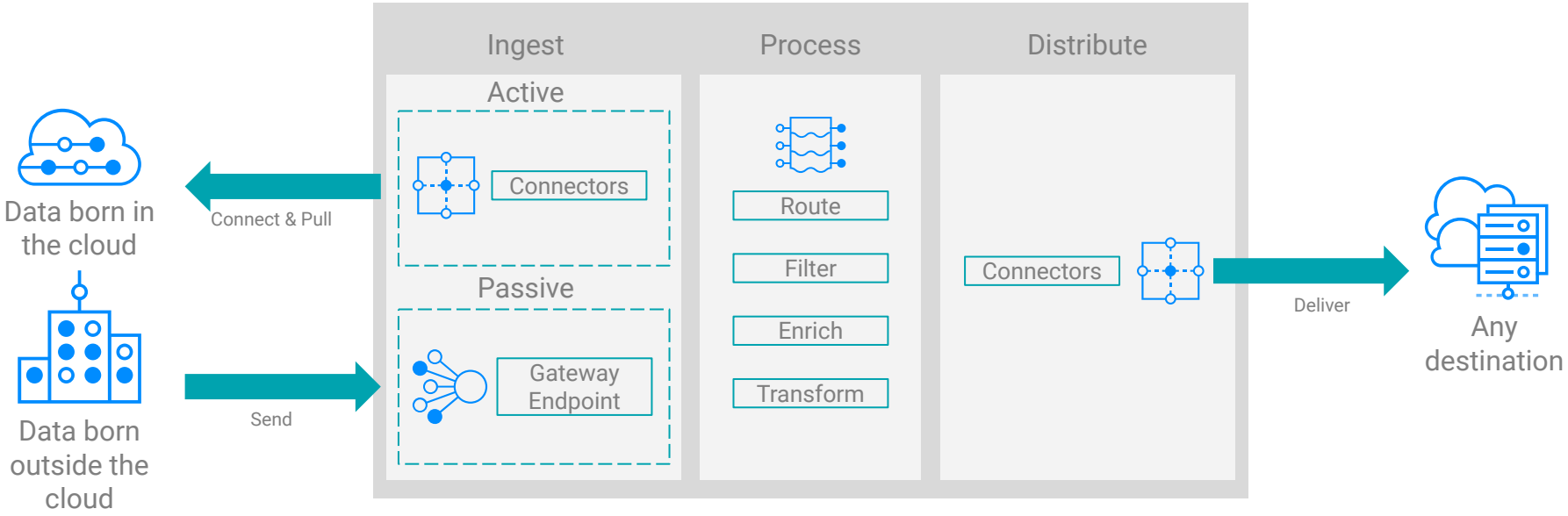
WARNINGS

- 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

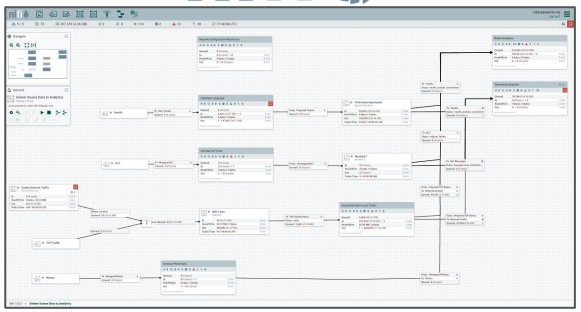
UNIVERSAL DATA DISTRIBUTION

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



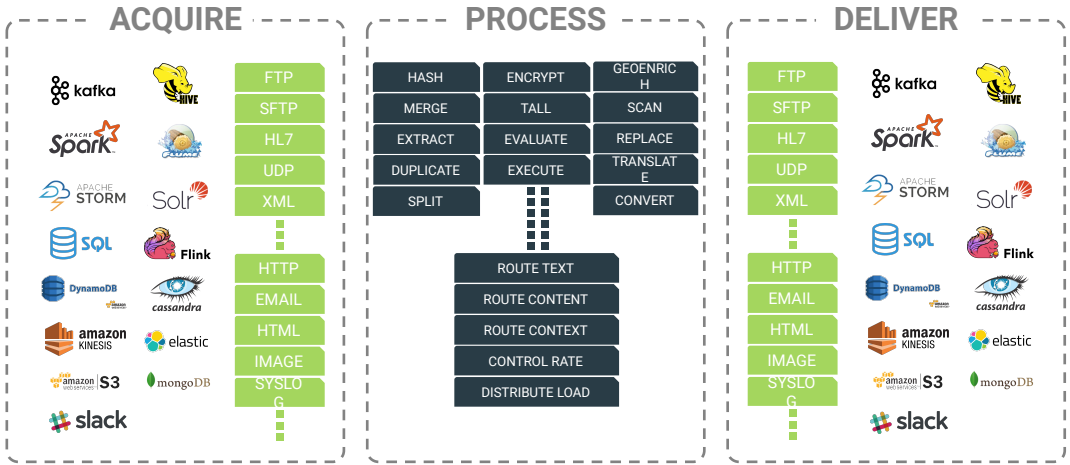
CLOUDERA FLOW MANAGEMENT

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



APACHE NIFI registry

Advanced tooling to industrialize flow development (Flow Development Life Cycle)



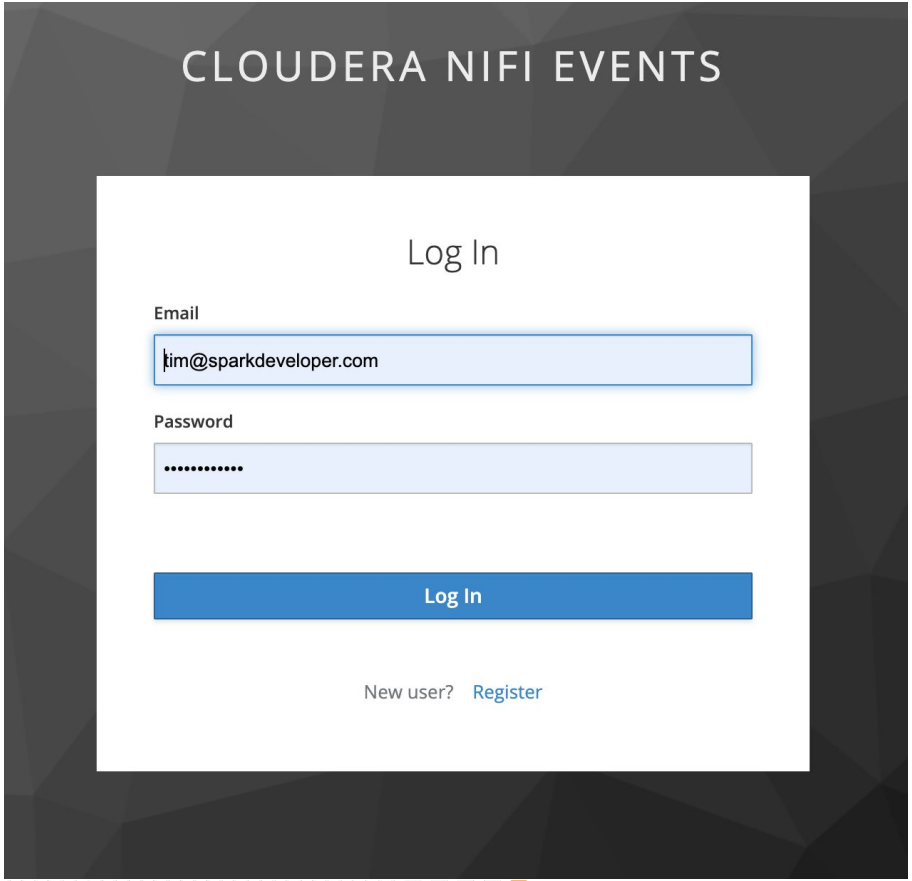
- Over 350 pre-built processors
- Easy to build your own processors
- Parse, enrich & apply schema
- Filter, Split, Merge & Route
- Throttle & Backpressure
- Guaranteed delivery
- Full data provenance
- Eco-system integration

FAQ

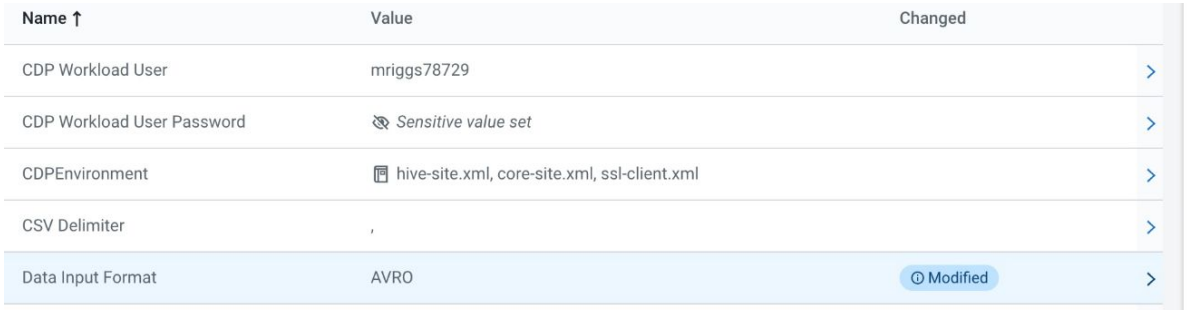
RECOMMENDATIONS

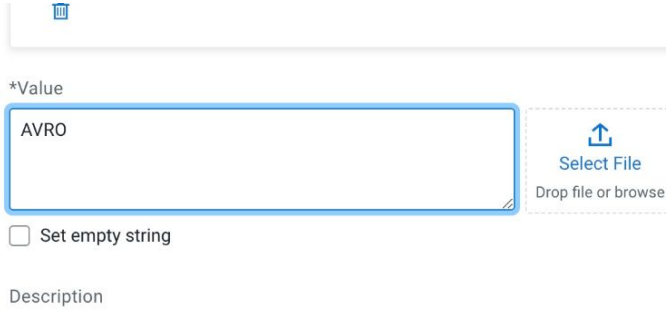
- Kafka Topic Names cannot have blank lines or spaces
- Schema Names cannot have blank lines or spaces
- Make sure everything in parameters is trimmed
- Every action is audited
- Shutdown your flow when you are done for the day or not working
- All flows will be shutdown after 4 hours of running

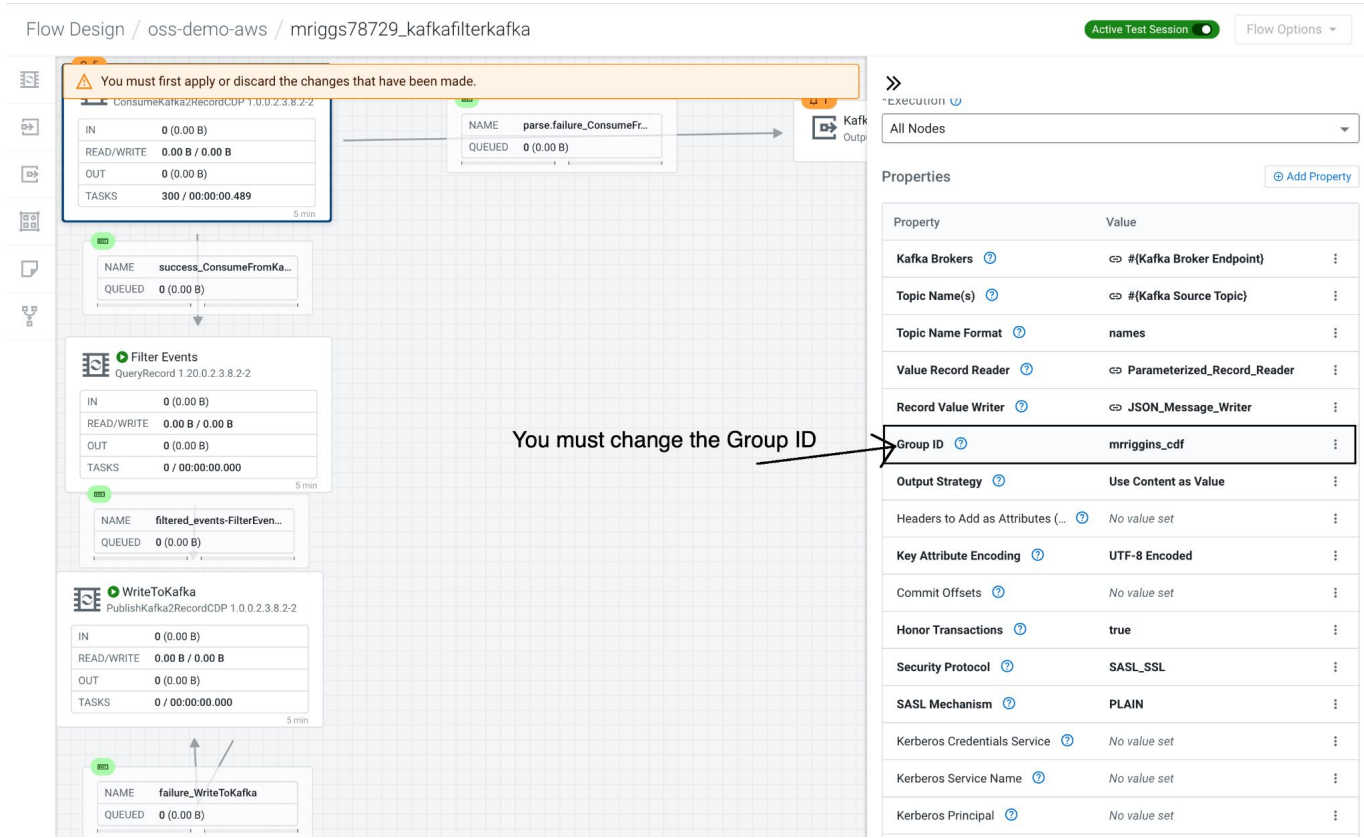
RELOGIN FREQUENTLY



SPACES AND EMPTY LINES MATTER







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
-
-

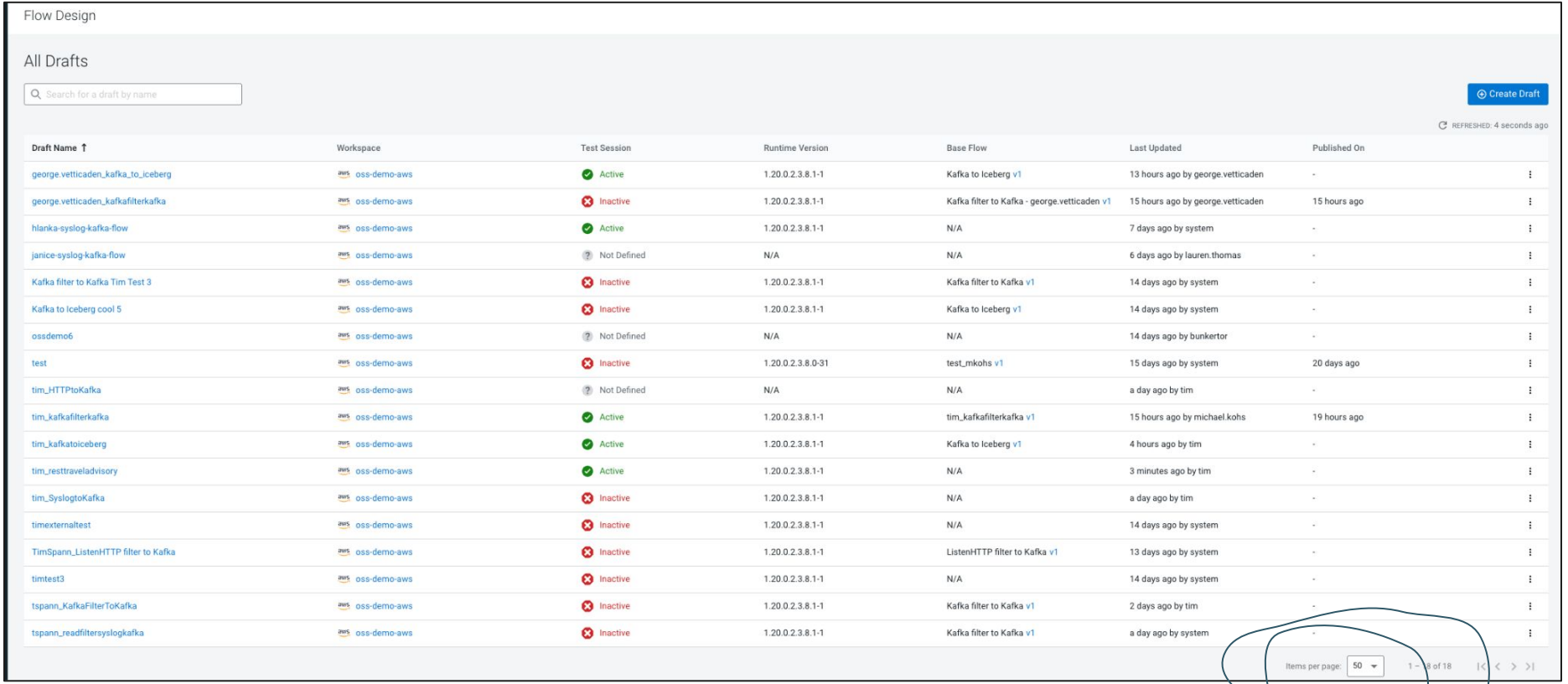
COMMON ERRORS

For Developers

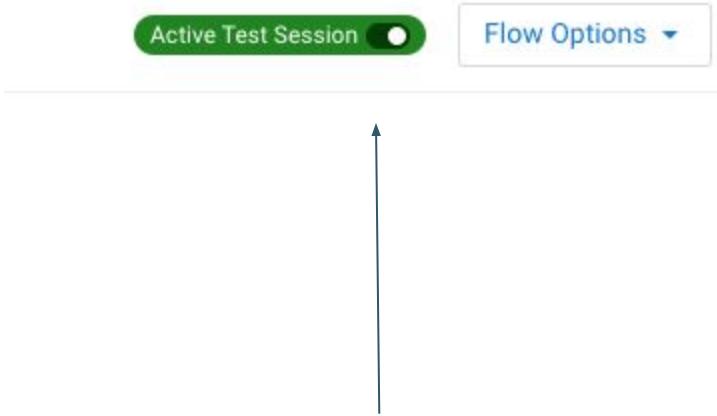
ERROR – 2023-04-19 07:50:53 pm EDT PublishKafka2RecordCDP[id=cb47d8e4-9b13-33ff-a8f8-d4b27bf3da58] Failed to send FlowFile[filename=508557dc-5eb1-4bcf-ade5-029b373a3a8d] to Kafka: org.apache.kafka.common.errors.TopicAuthorizationException: Not authorized to access topics: [tim_traveladvisory]

- 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.

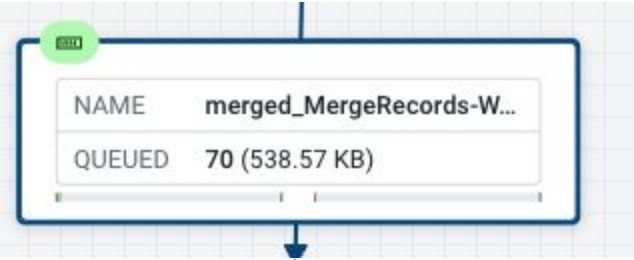
Flow Design Tips



Your items may be on other pages



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

SOURCE
MergeRecords

DESTINATION
WriteTolceberg

More Details

Settings

Connection Name

merged_MergeRecords-WriteTolceberg

Flowfile Expiration

0 sec

Back Pressure Object Threshold

10000

Back Pressure Size Threshold

1 GB

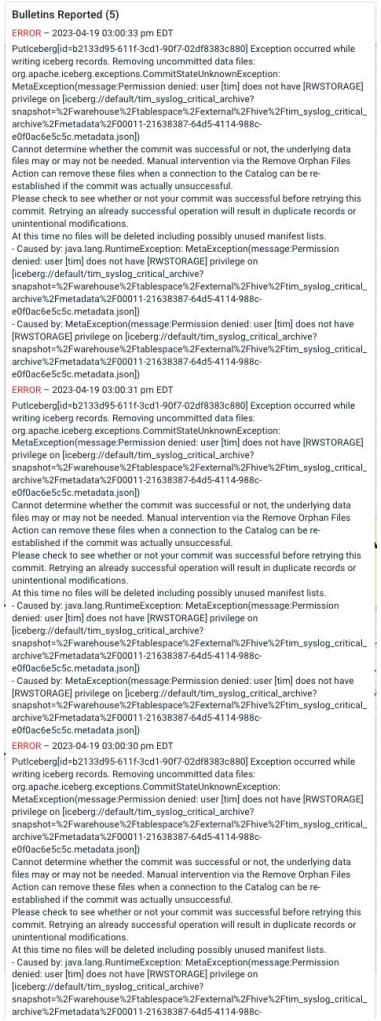
Load Balance Strategy

Do not load balance

Relationships

- [ ] failure
- [ ] original
- [x] merged

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



Check the **ERROR** in Bulletins.

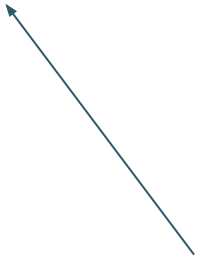
If you see a permission or missing item issue make sure you have your Workload User Name and Workload Password set correctly.

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

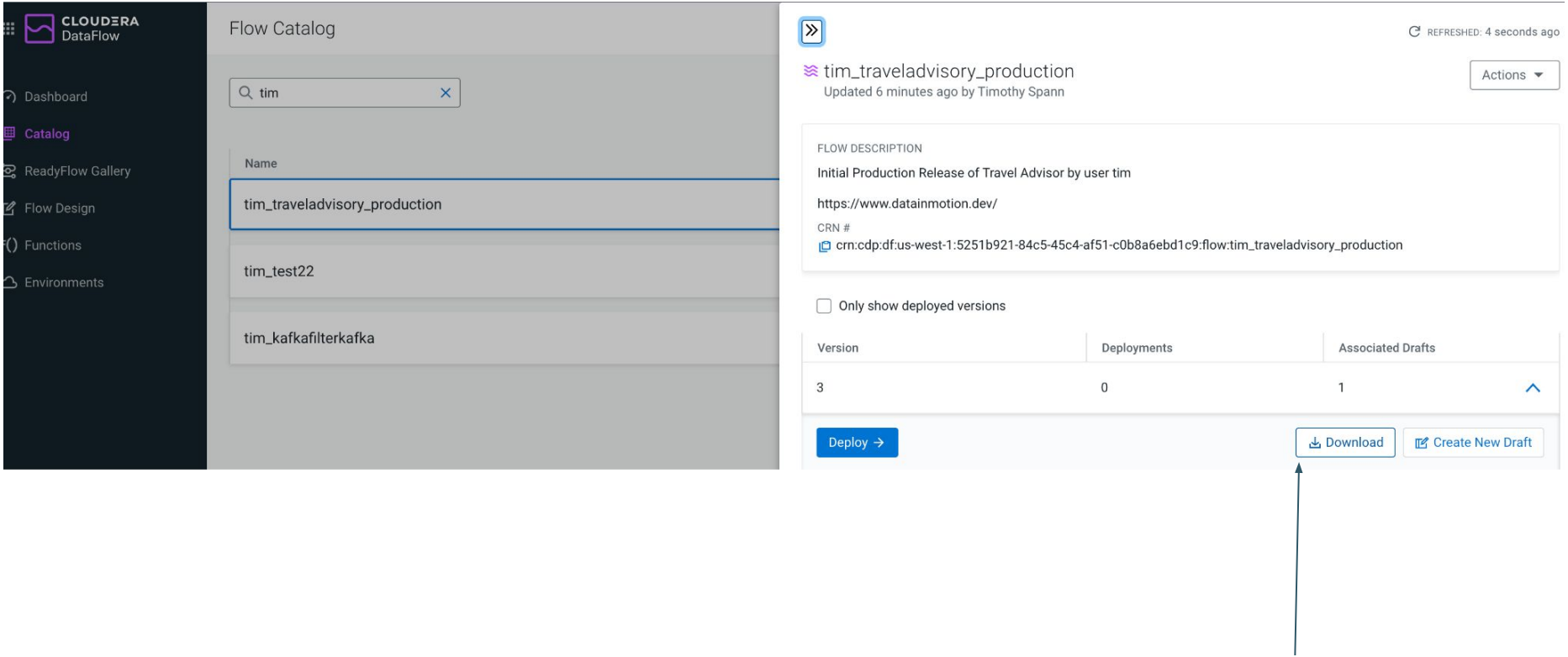
FLOW NAME	VERSION
tim_traveladvisory_produc…	3

Publish

Publish As

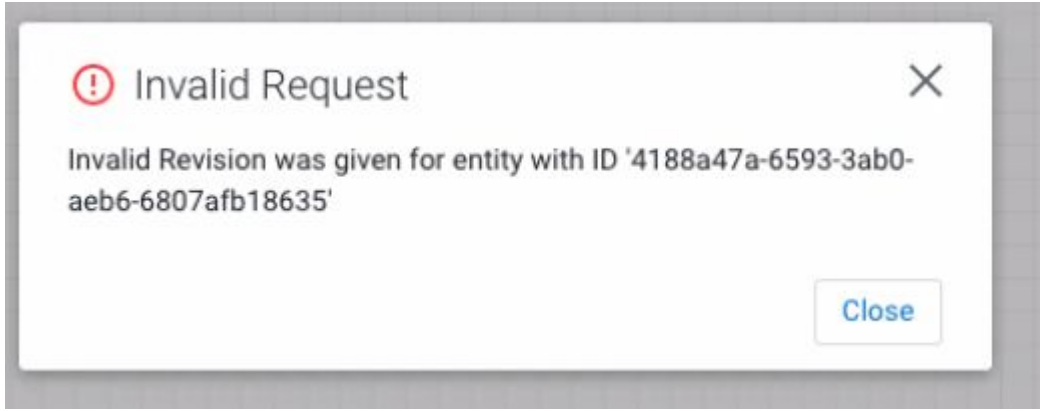
Publish a new version of the existing flow

Publish your flow to back it up



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.



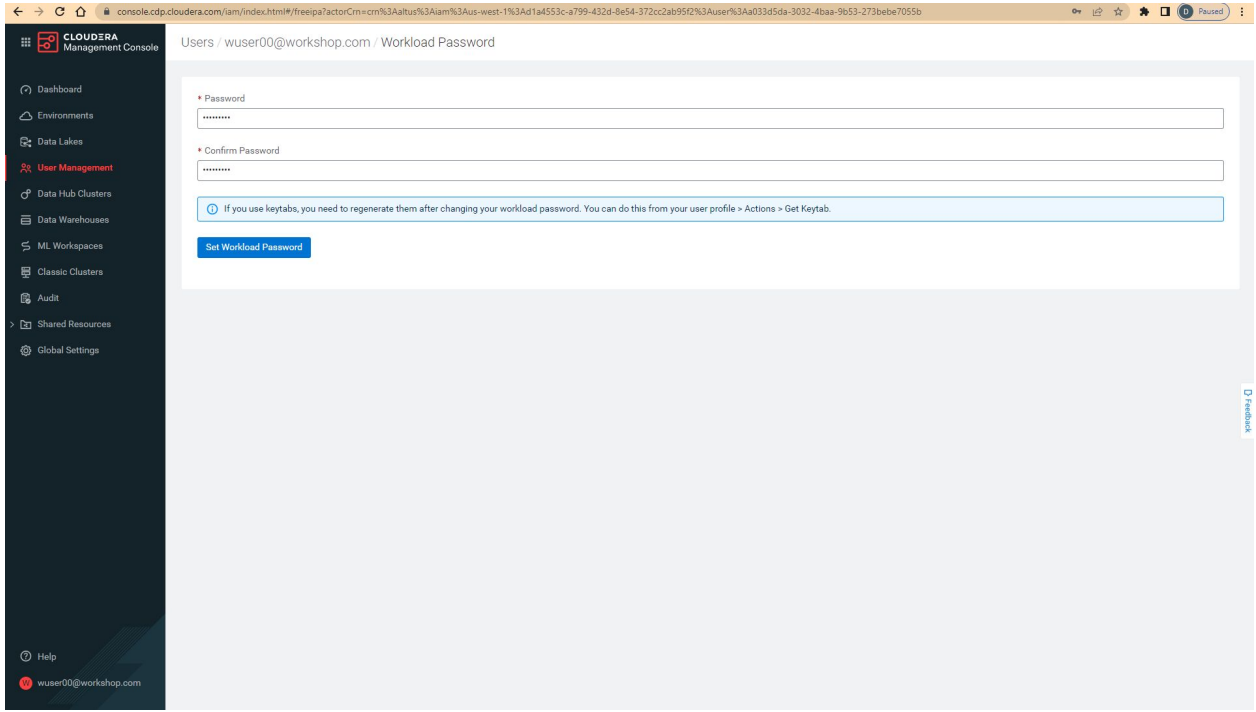
If you see a popup with Invalid Revision or Invalid request.

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

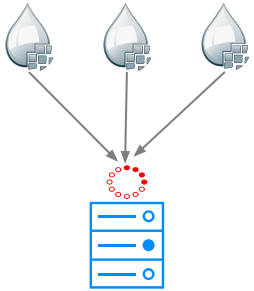
RESOURCES

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

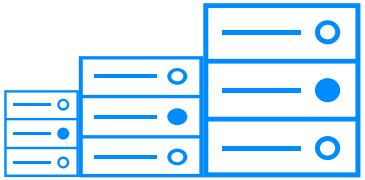
CLOUDERA DataFlow in CDP

OPERATIONAL CHALLENGES WITH NiFi FLOWS



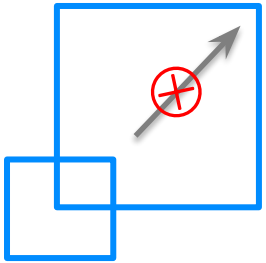
Resource contention

Impacts performance of all flows in the cluster



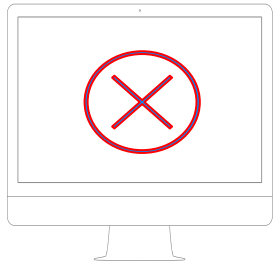
Oversizing clusters

High infrastructure costs



On-demand manual scaling

Operational nightmare



Comprehensive flow visibility

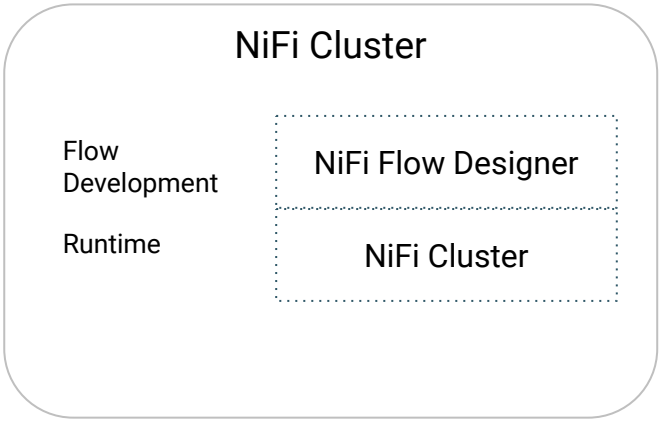
Monitoring NiFi flows across multiple clusters can be challenging

ARCHITECTURAL SHIFT IN DATAFLOW

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

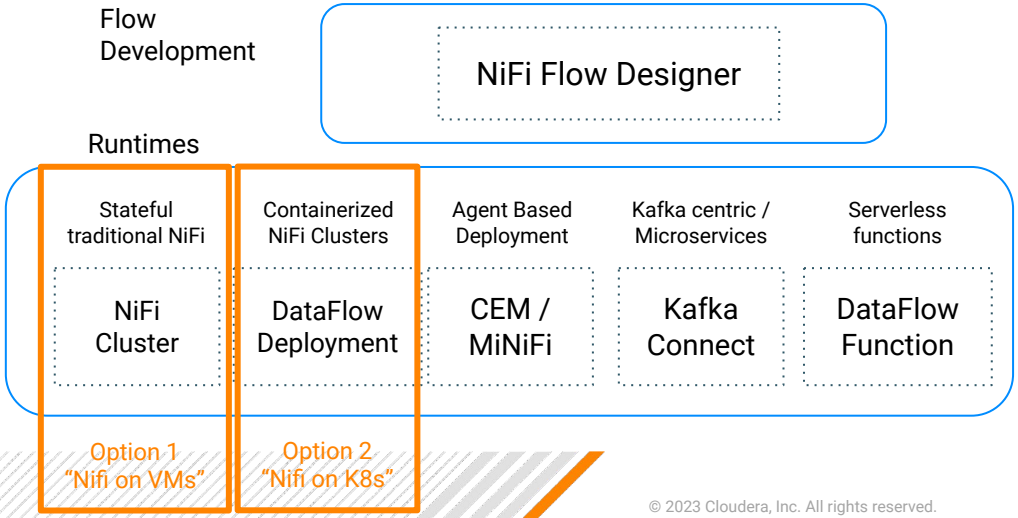
Classic NiFi Architecture

NiFi Flow Designer + NiFi Cluster Runtime are tightly coupled



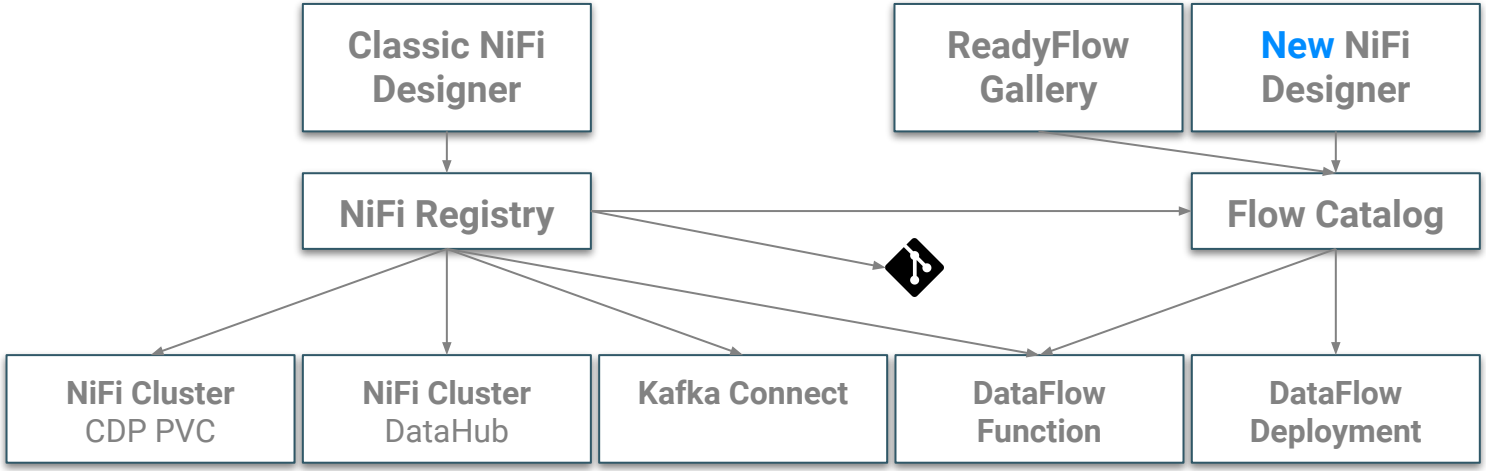
New NiFi Architecture

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



CLOUDERA

DEPLOYMENT OPTIONS FOR DATAFLOW



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

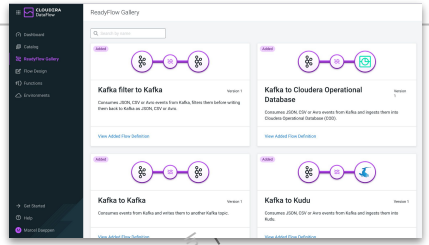
Central monitoring console for all your flows across environments

Monitor flow metrics and infrastructure usage

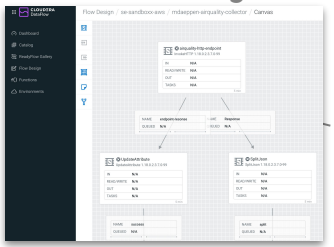
Define alerts for flows breaching assigned KPIs

DEPLOYMENT WITH SDLC SUPPORT

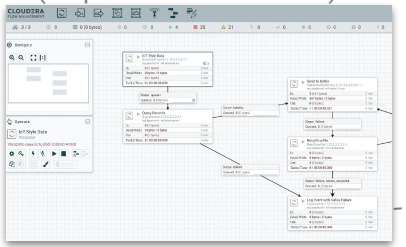
ReadyFlow Gallery



New Flows Designer



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

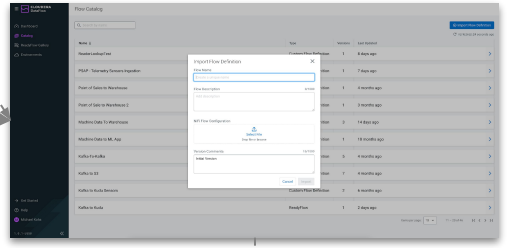


Upload to Catalog

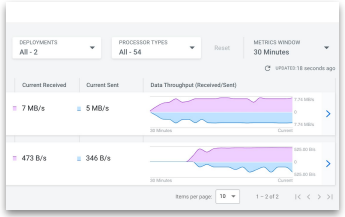
Upload to Catalog

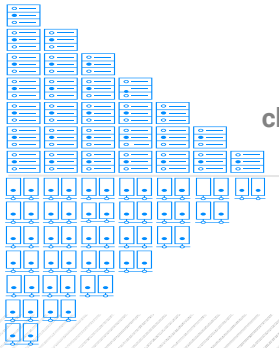
Instantiation into Catalog

Flow Catalog

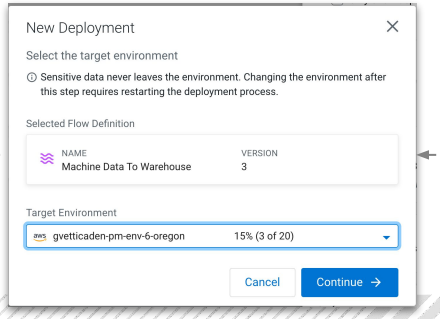


With full monitoring





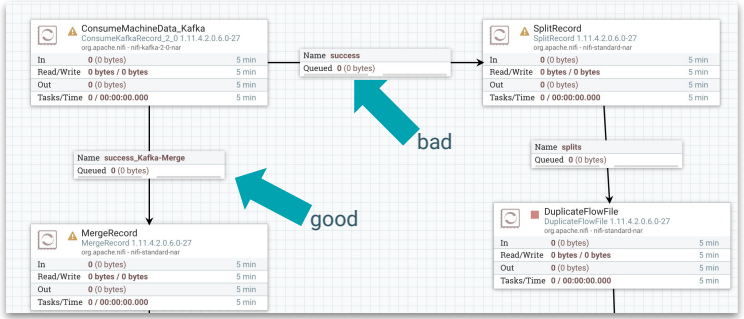
Auto-scale Kubernetes clusters on CDP

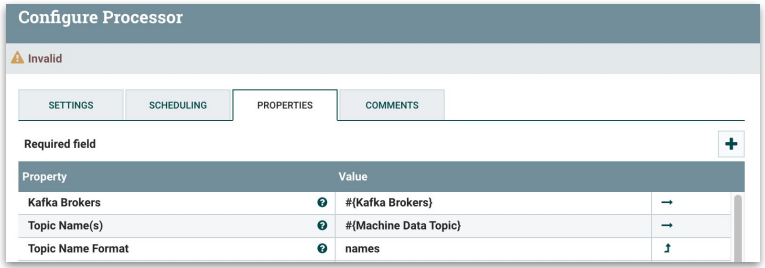


Select a Flow & use the deployment wizard

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*)





FLOW DEVELOPMENT BEST PRACTICES

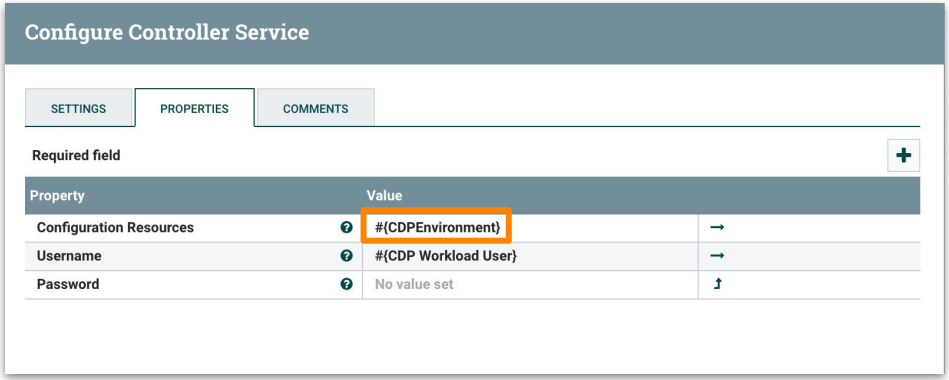
CDPEnvironment Parameter & Default SSLContextService

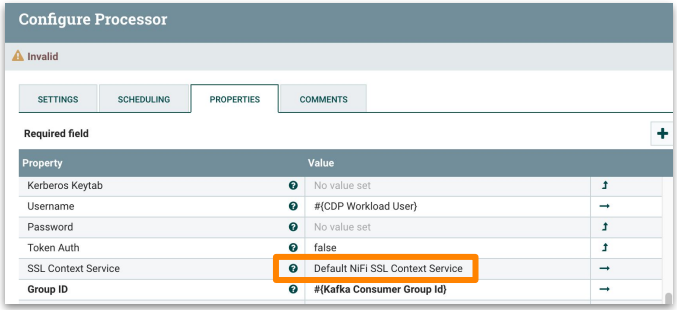
CDP Environment Parameter

- Use whenever Hadoop configuration files are needed
- CDF detects parameter usage, obtains the Hadoop configuration files from SDX, makes them available to NiFi pods and replaces parameter value accordingly
- No copying of config files required anymore

Default NiFi SSL Context Service

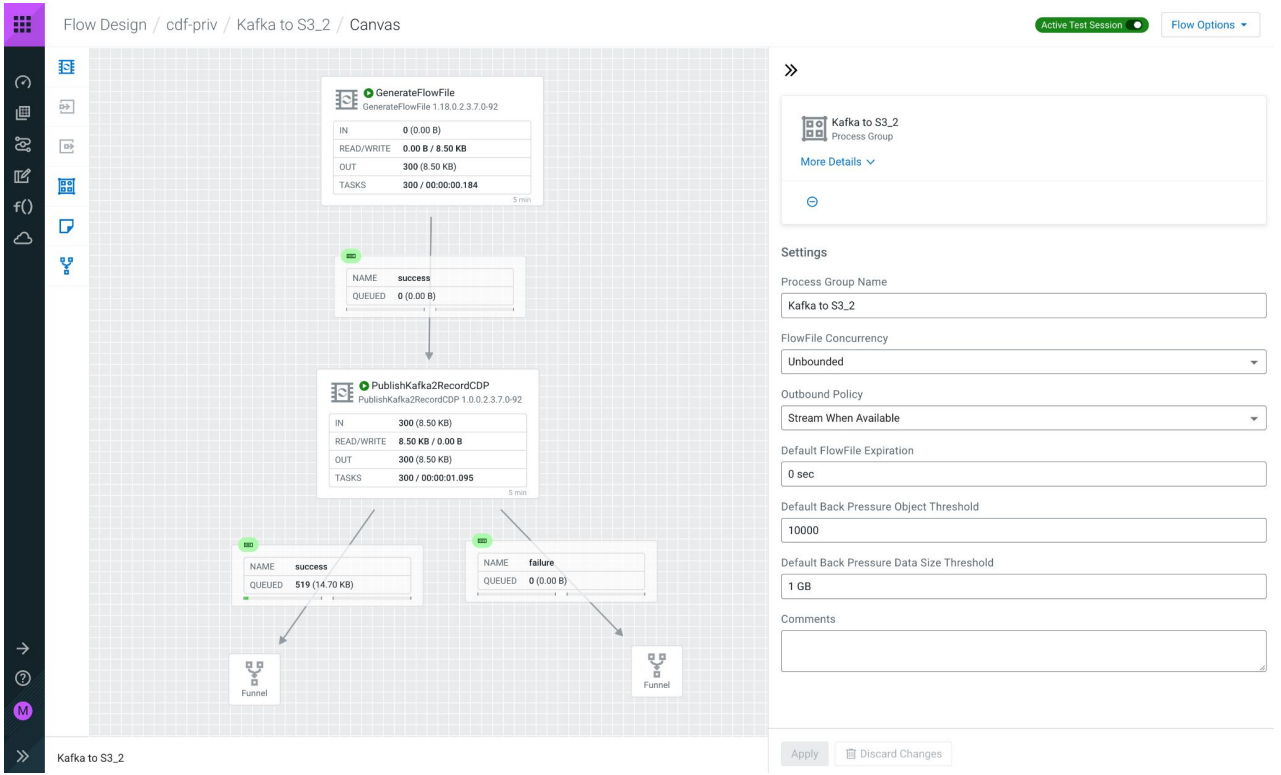
- Use whenever SSL Context Service is required to interact with CDP service in target environment
- CDF detects reference, creates a key and truststore for the target environment and configures a *Default NiFi SSL Context Service* accordingly
- No more manual creation of truststores and moving around of certificates to interact with CDP services
- *Default NiFi SSL Context Service* must be an *external controller service* i.e. defined **outside** of the process group that's exported





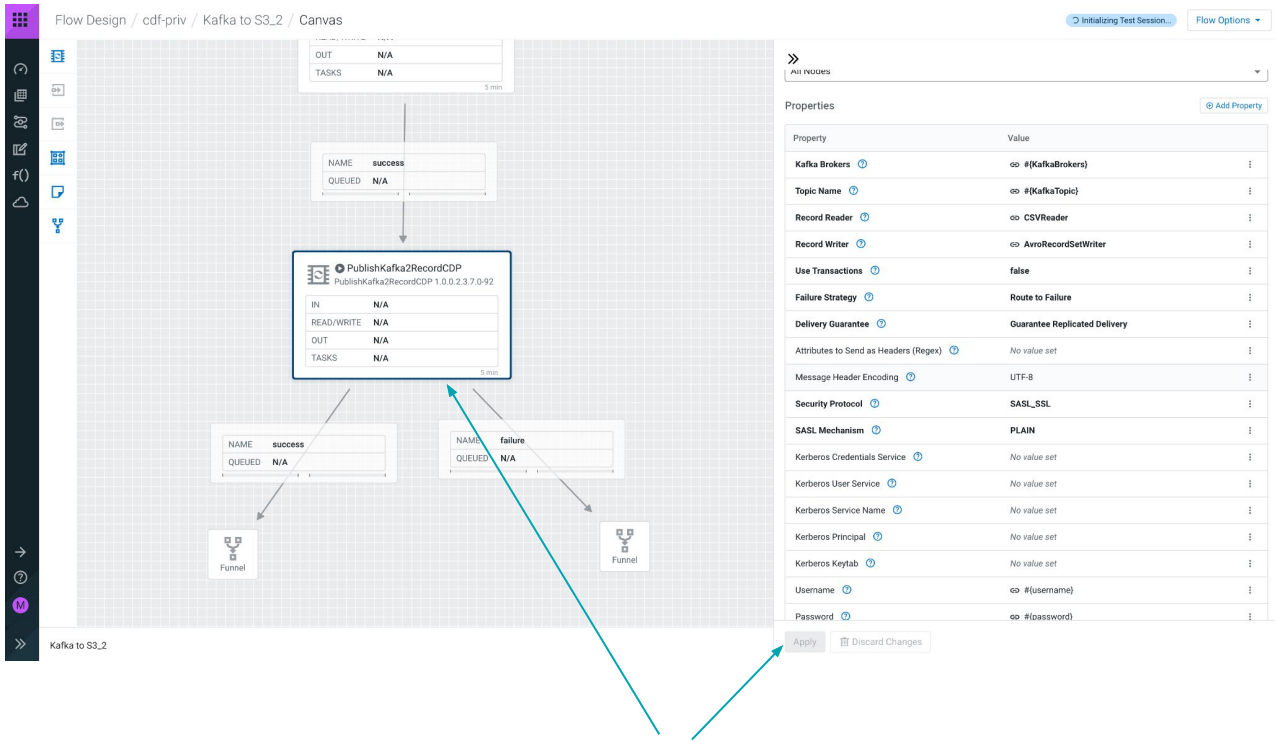
Data Flow Design for Everyone

- 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

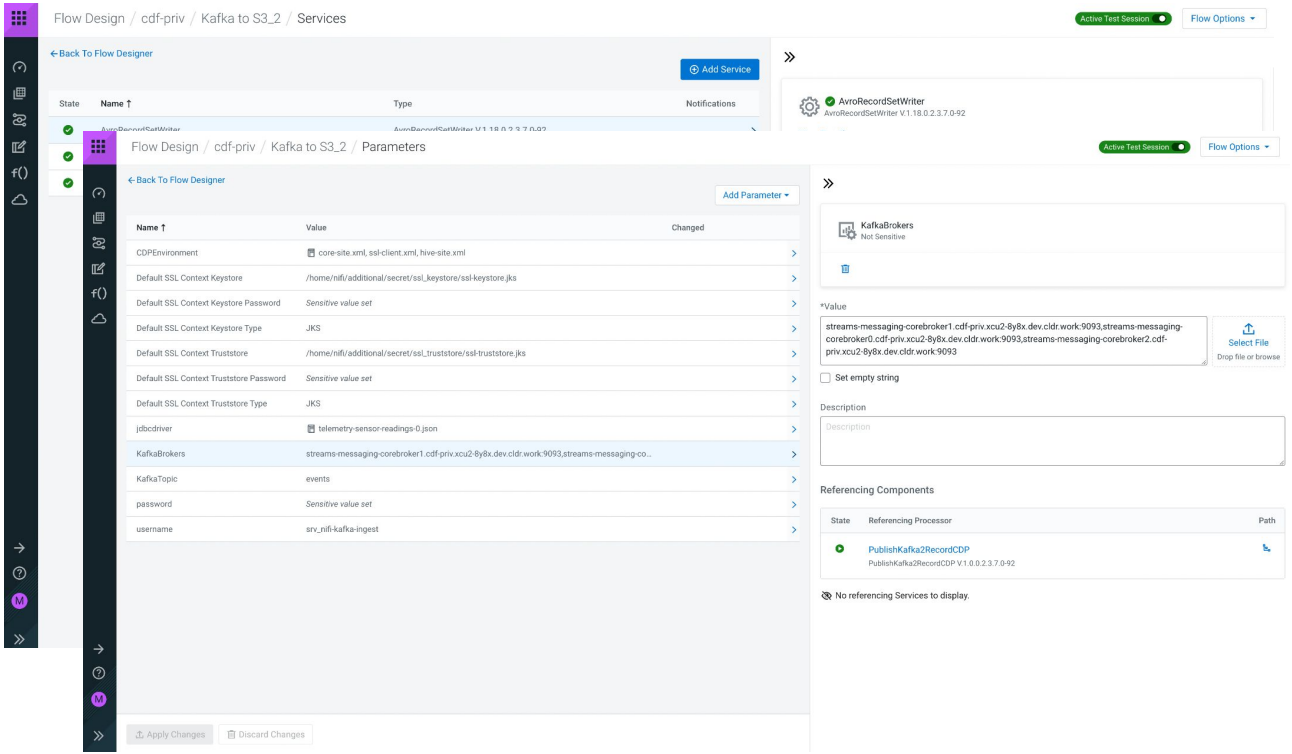
- Configuration side panel automatically represents selected canvas component
- Developers can still navigate on the canvas while having configuration easily accessible
- Allows for quick access to configuration while eliminating clicks



Reflects canvas selection

Simplified Parameters and Controller Services

- Manage Services and Parameters centrally for you 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



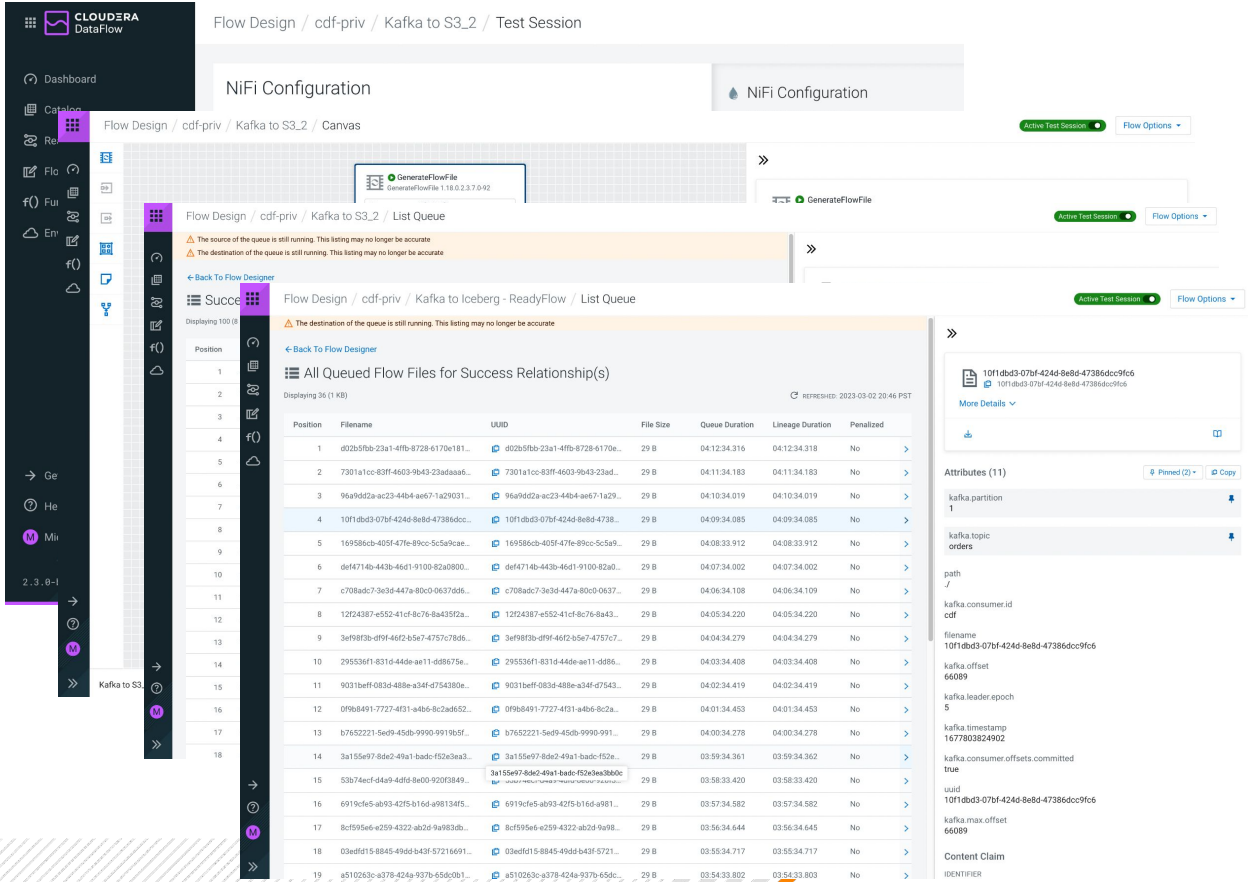
Interactive development through test sessions

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



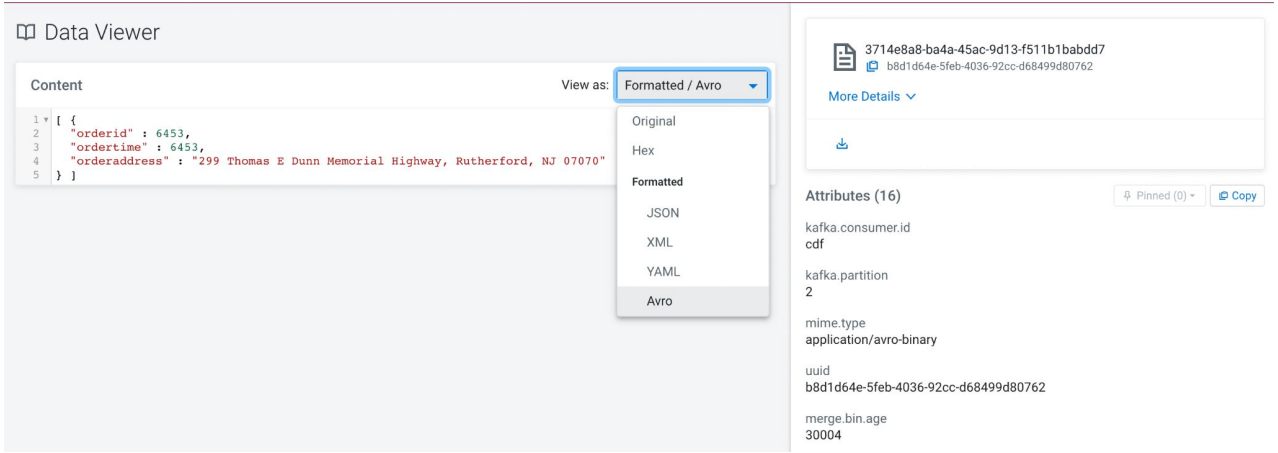
Data Viewer

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

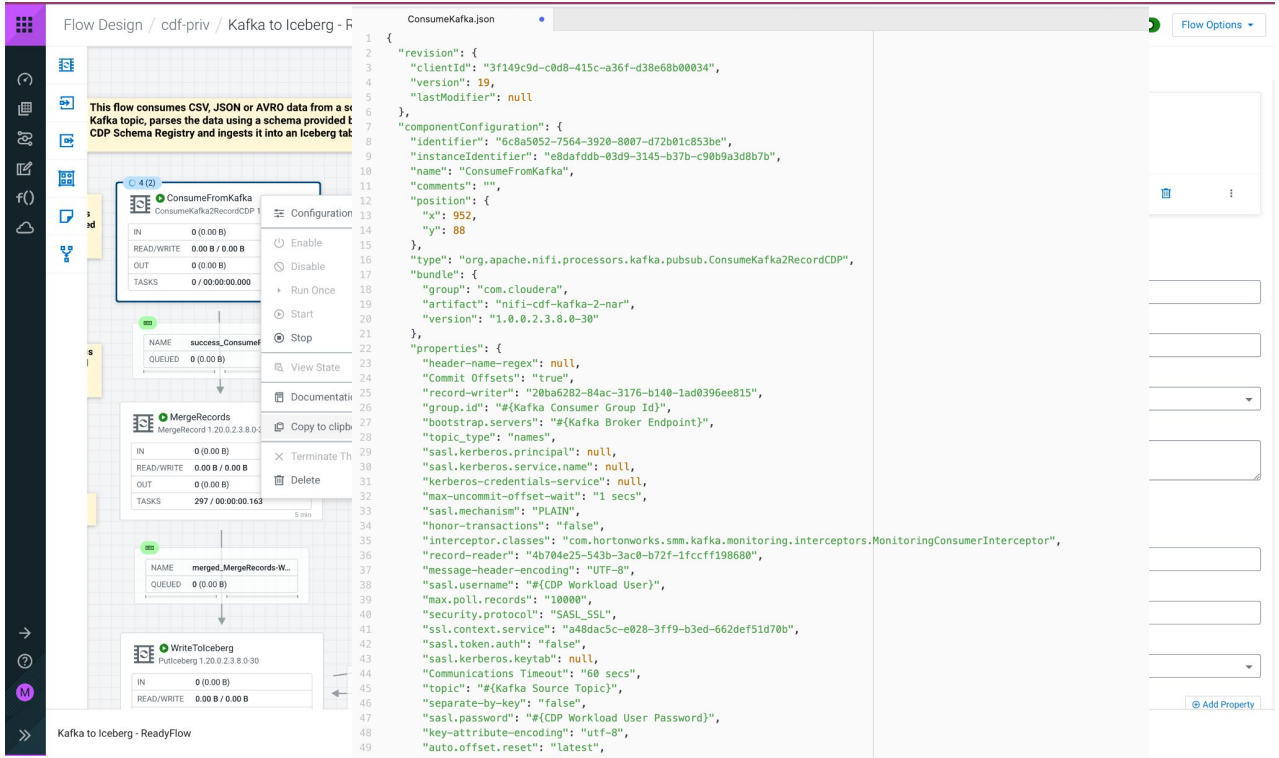
Download the flowfile content



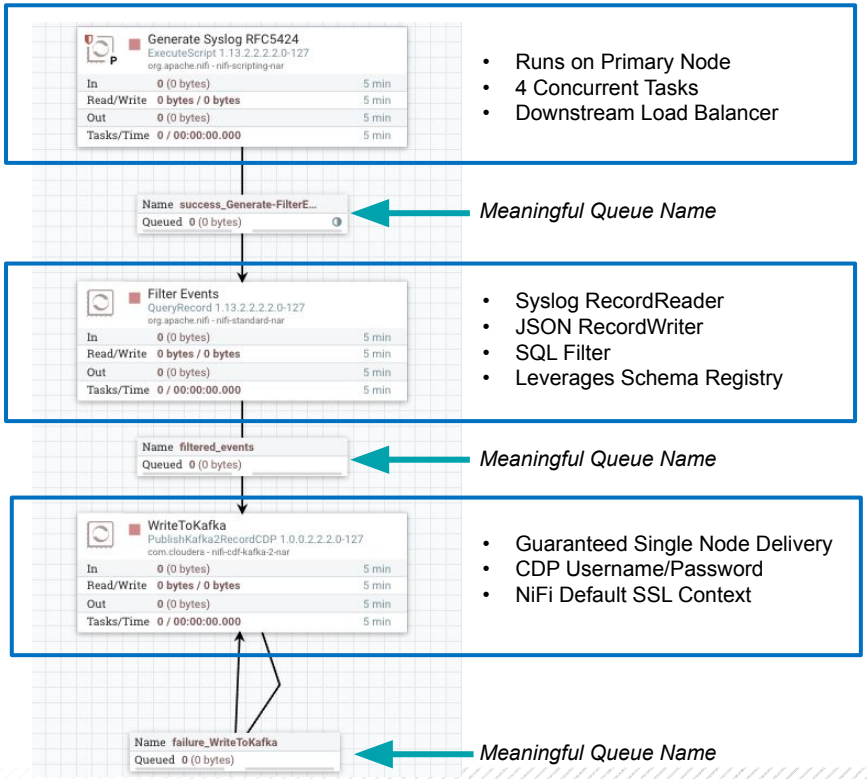
Copy & Paste

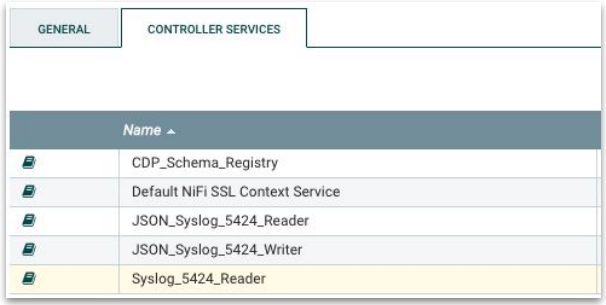
Copy & Paste components between different drafts

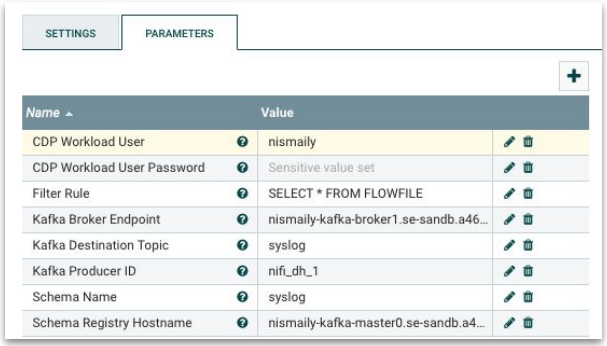
Paste clipboard content in text editor to get JSON representation



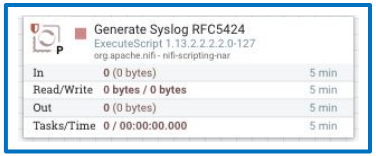
LOG ANALYTICS IMPLEMENTATION







SYSLOG RFC 5424

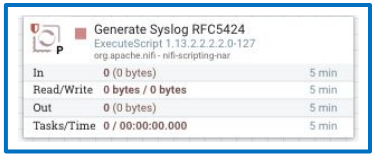



```
        VERSION                                                    PROCID
PRI     | TIMESTAMP                  HOSTNAME                APP-NAME | MSGID
<165>1 2003-10-11T22:14:15.003Z mymachine.example.com evntslog - ID47
[exampleSDID@32473 iut="3" eventSource="Application" eventID="1011"] BOMAn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

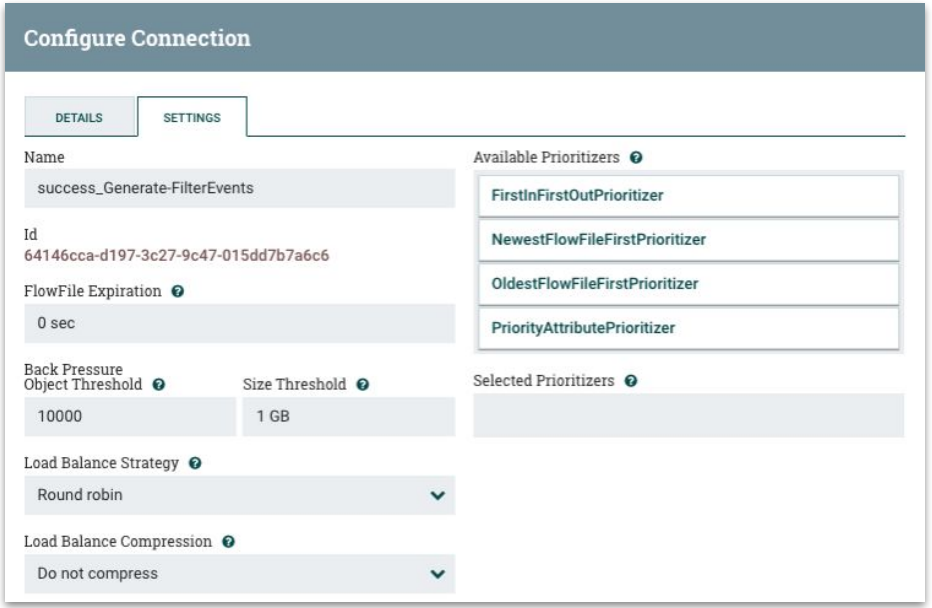
- **PRI** — or "priority", Facility (what kind of message) * 8 + **Severity** (how urgent is the message)
- **VERSION** — version is always "1" for RFC 5424
- **TIMESTAMP** — valid timestamp examples (must follow ISO 8601 format with uppercase "T" and "Z")
- **HOSTNAME** — using FQDN (fully qualified 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

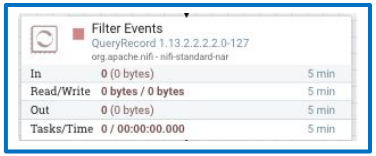
QUEUE CONFIGURATION



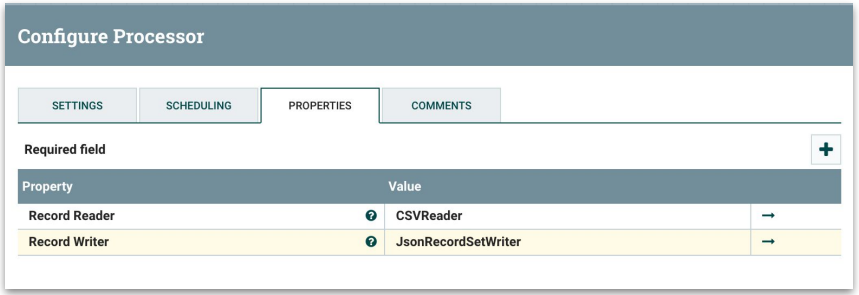
- **FlowFile Expiration -** Data that cannot be processed in a timely fashion can be automatically removed from the flow
- **Back Pressure Thresholds -** Thresholds indicate how much data should be allowed to exist in the queue before the component that is the source of the Connection is no longer scheduled 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



RECORD-ORIENTED DATA WITH NIFI

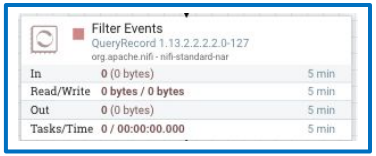


- **Record Readers** - Avro, CSV, Grok, IPFIX, JSAN1, JSON, Parquet, Scripted, Syslog5424, Syslog, WindowsEvent, XML
- **Record Writers** - Avro, CSV, FreeFromText, Json, Parquet, Scripted, XML
- Record Reader and Writer support referencing a schema registry for retrieving schemas when necessary.
- Enable 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.

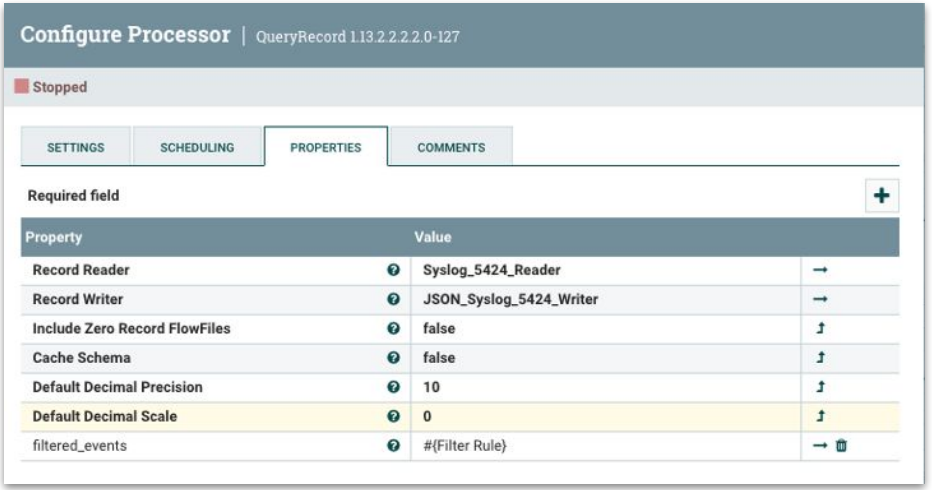


CLOUDERA

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.



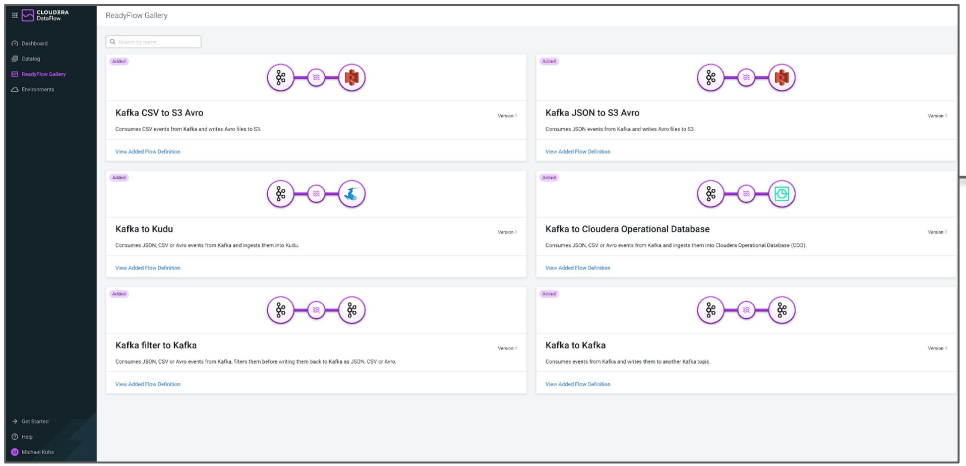
CLOUDERA DATA FLOW – PUBLIC CLOUD

READYFLOWS - FOR THE MOST COMMON USE CASES

For new NiFi users

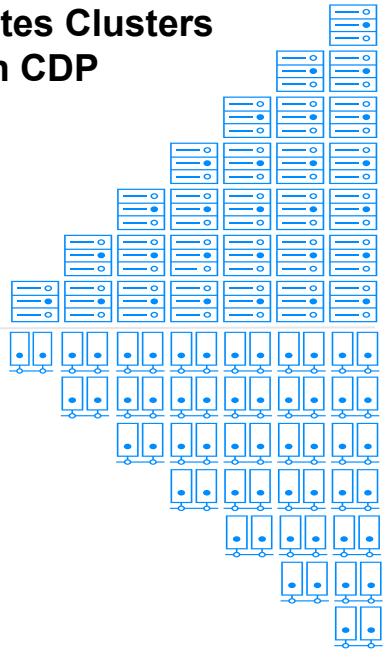
ReadyFlows Gallery

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



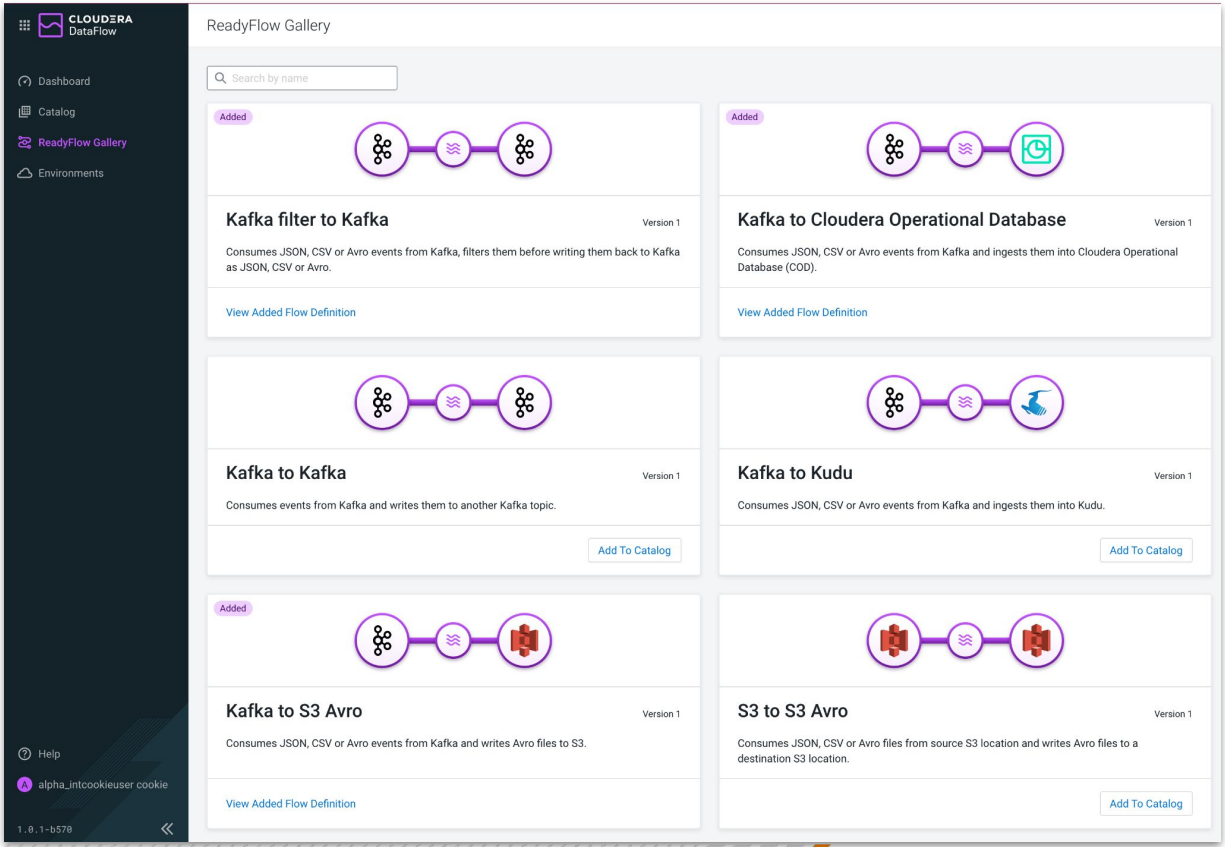
Select a ReadyFlow and use the deployment wizard

Auto-Scale Kubernetes Clusters on CDP



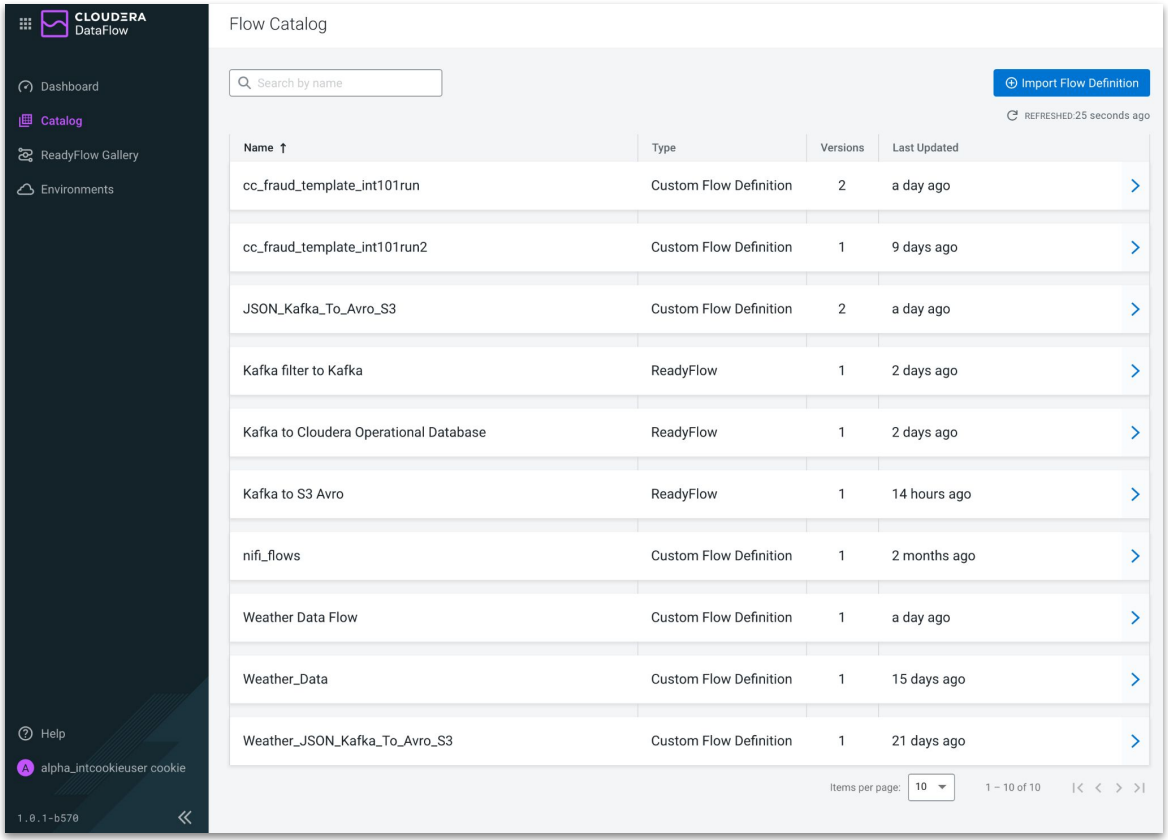
READYFLOW GALLERY

- 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



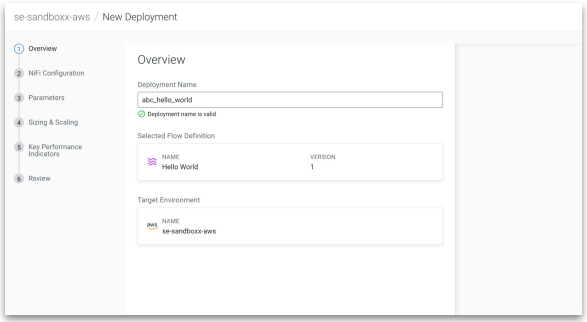
FLOW CATALOG

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

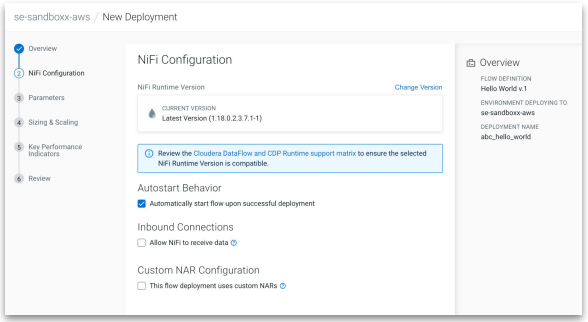


TURNS FLOW DEFINITIONS INTO FLOW DEPLOYMENTS

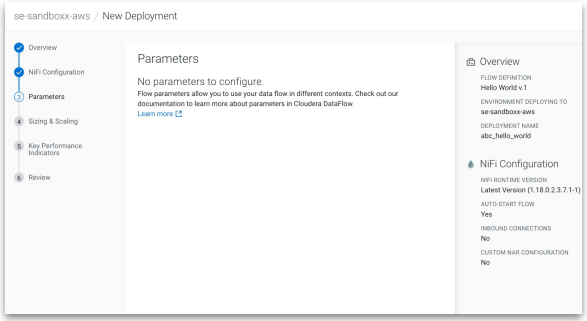
1.) Start Deployment Wizard



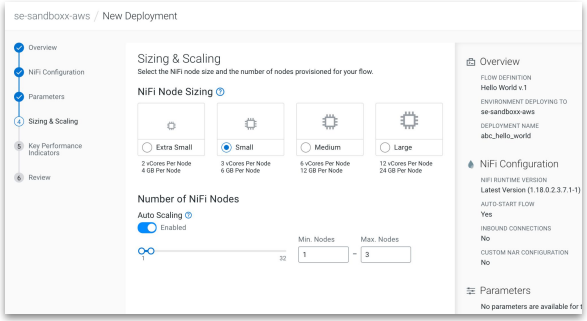
2.) NiFi Config



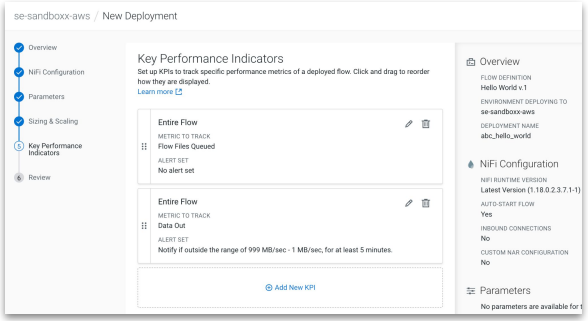
3.) Provide Parameters for NiFi



4.) Configure Sizing & Scaling



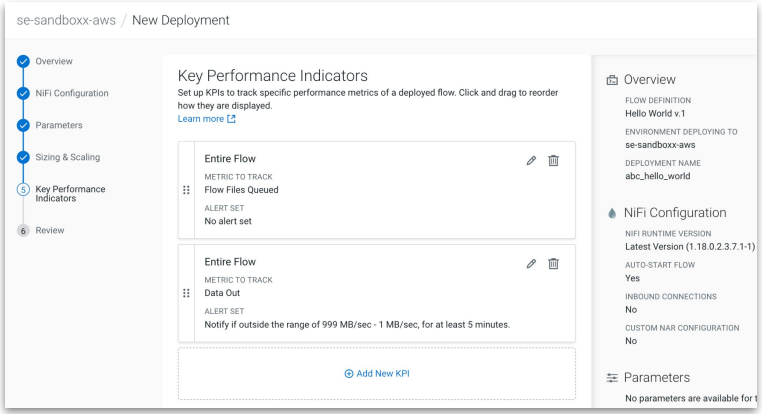
5.) Define KPIs



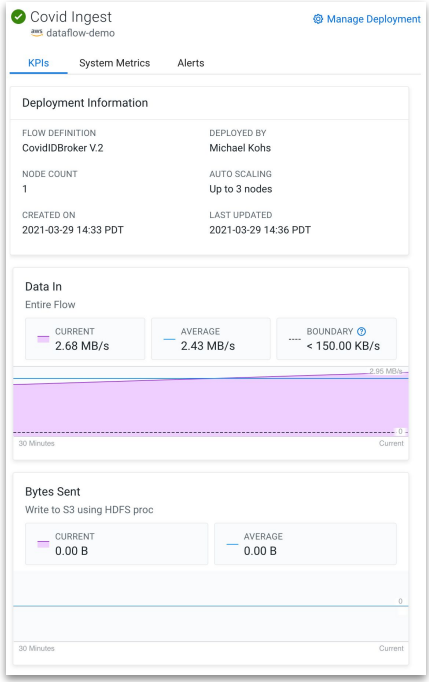
KEY PERFORMANCE INDICATORS

- 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

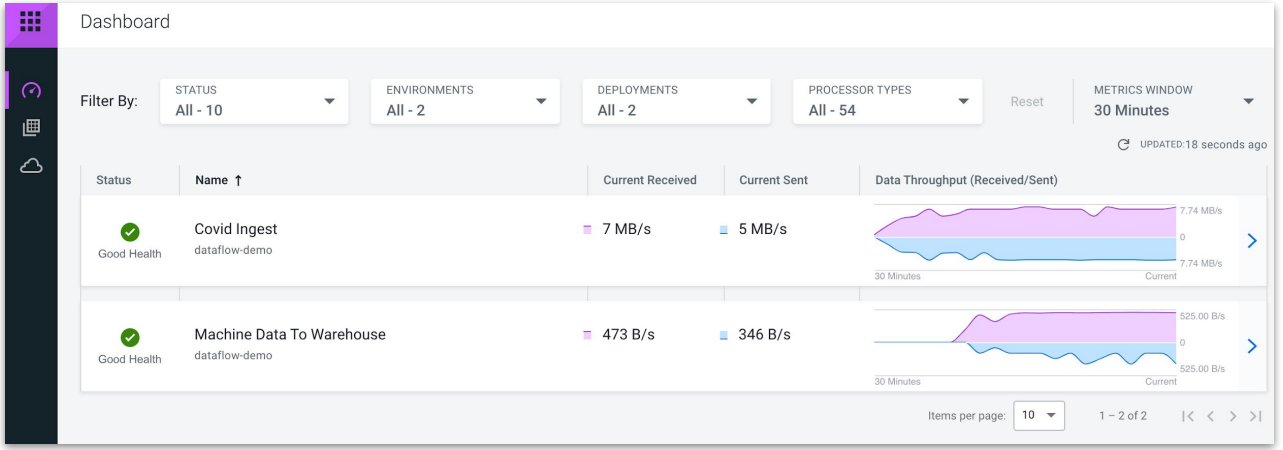


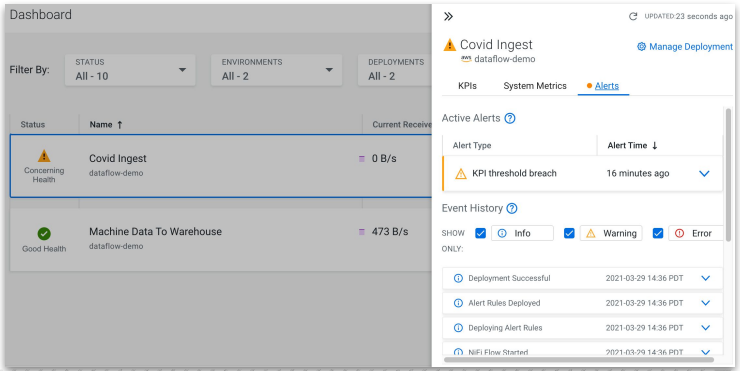
KPI Monitoring

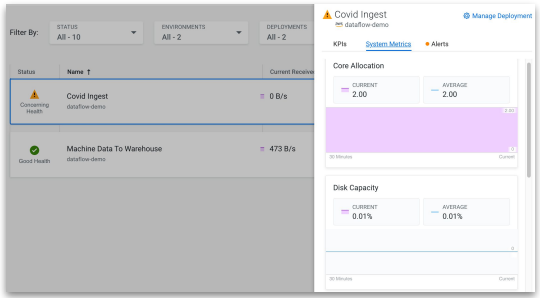


DASHBOARD

- 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

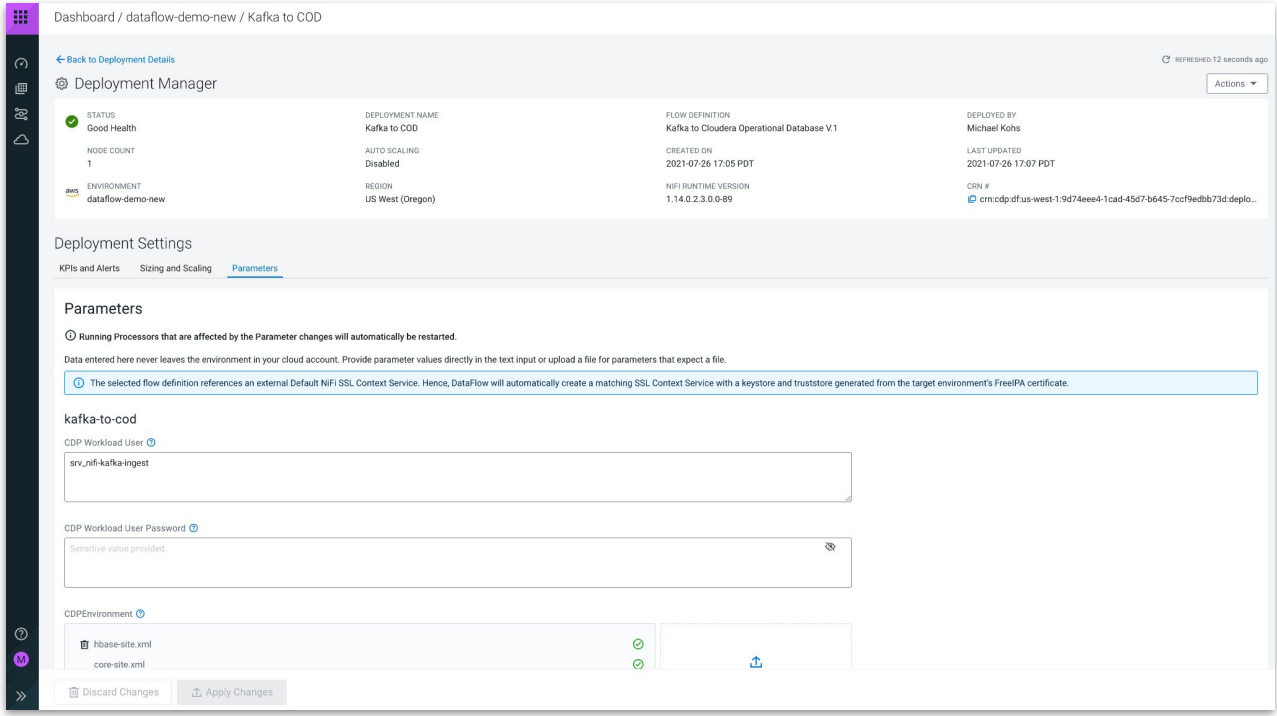






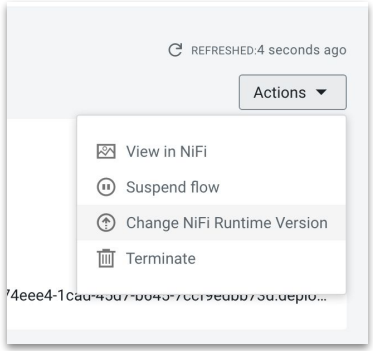
DEPLOYMENT MANAGER

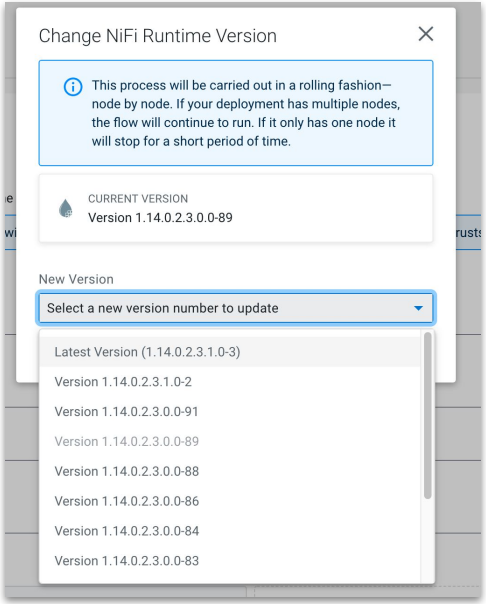
- Manage flow deployment lifecycle (Suspend/Start/Terminate)
- Add/Edit KPIs
- Change sizing configuration
- Update parameters
- Change NiFi version of the deployment
- Gateway to NiFi canvas



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)



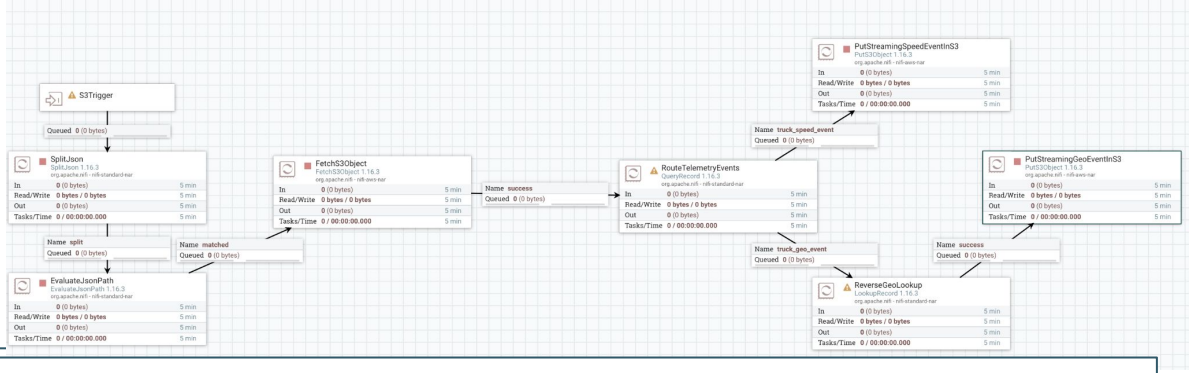


Deploy a DataFlow Function

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

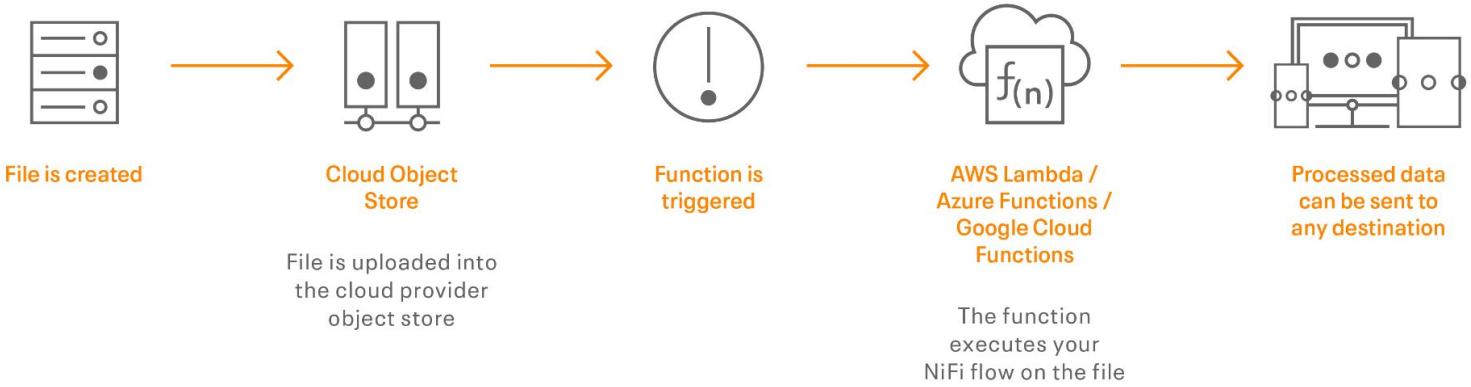
With the DFF State feature, build flows to create customized triggers allowing access to on-premises or external services

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

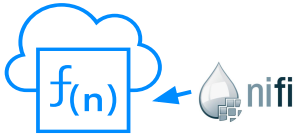
CLOUDERA

Sample DataFlow Functions Use Case

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



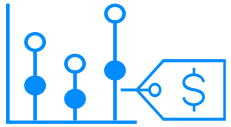
Summary



Serverless NiFi

New Use Cases

Faster ROI



Pay for Usage

True Pay for Value

No code UI

Rapid dev & test

Lower TCO

Cost optimization

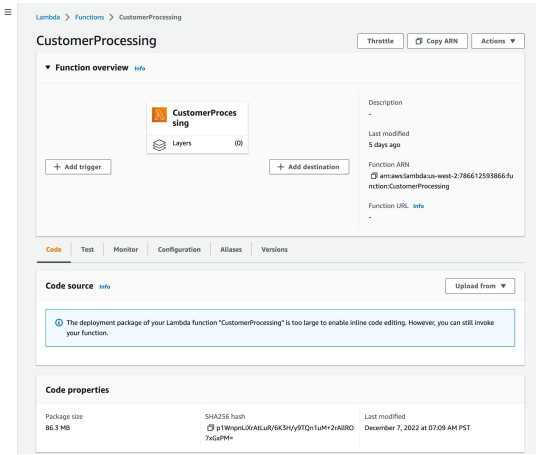
Reduce

Operational Overhead

CDF-PC: Run serverless NiFi with DataFlow Functions

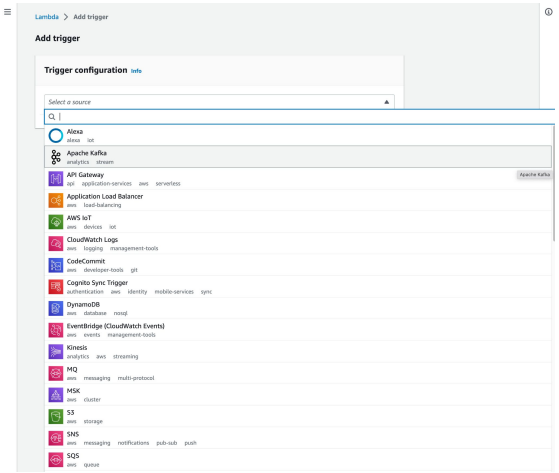
f(x) True serverless compute

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



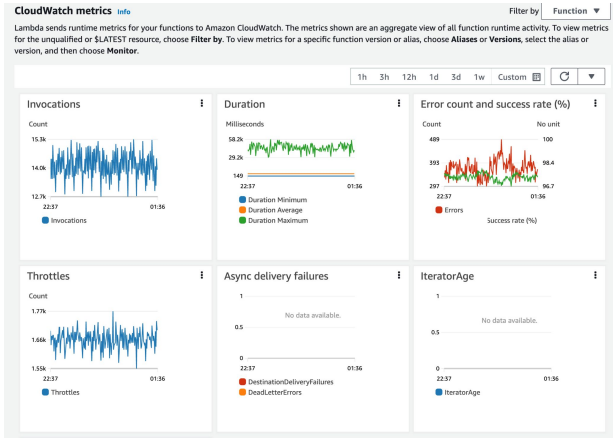
Execute on events

Supports cloud provider native triggers to launch a function.



Guaranteed delivery

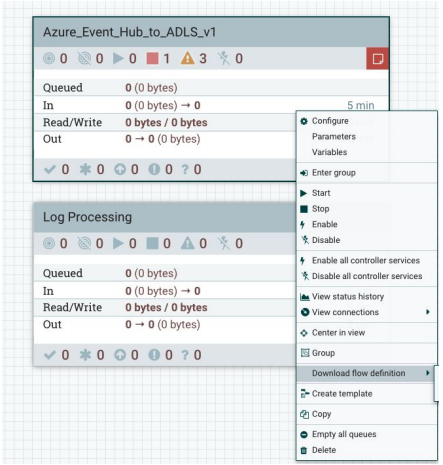
Acknowledges receipt of event to source system only once it has been delivered to the destination .



CDF-PC Deployments: Resource Isolation & Monitoring

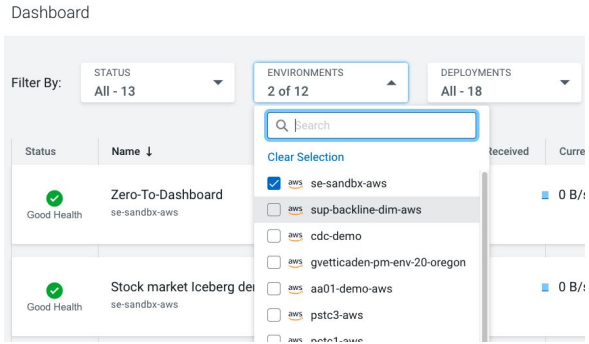
Resource Isolation

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



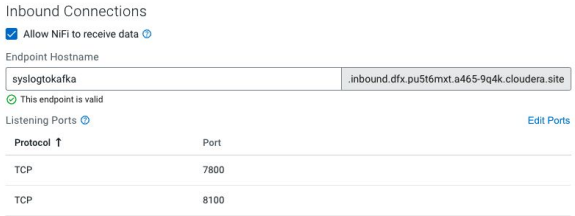
Central Monitoring

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



Inbound Connections

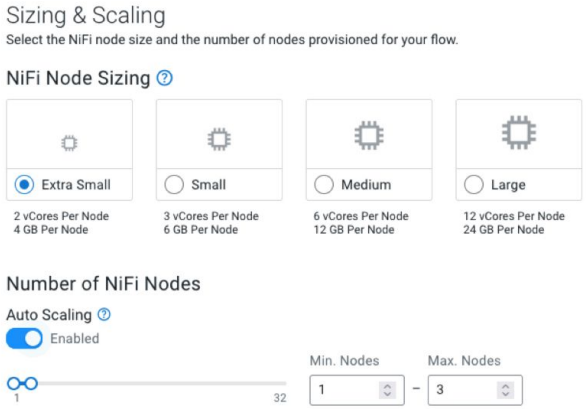
Send data from clients to flow deployments for further distribution and leave the load balancer, DNS and security configuration to CDF-PC



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

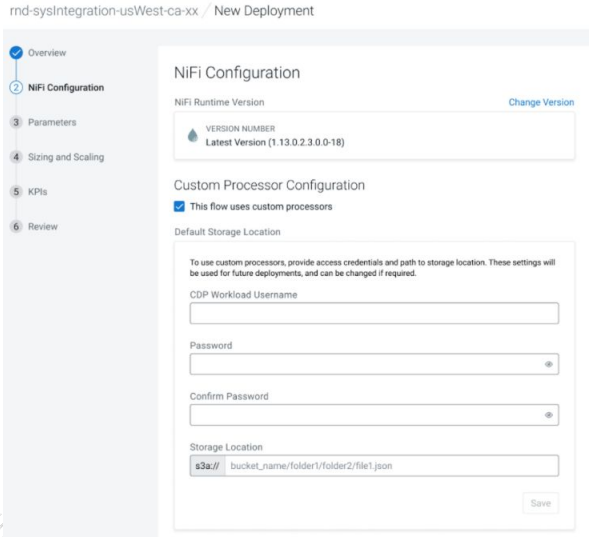


Zero Data Loss Guarantee when scaling down

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

Support for Custom Nars

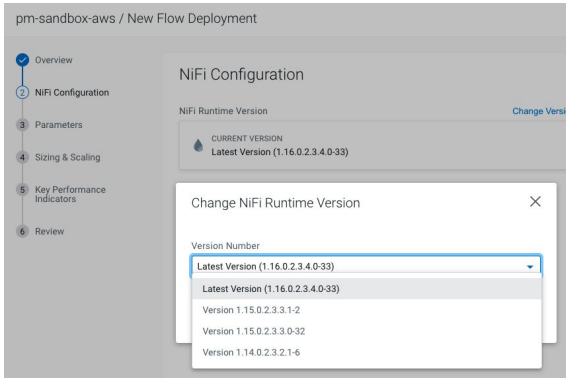
Run your existing NiFi data flows that rely on custom NARs/components



CDF-PC: Upgrades & Automation

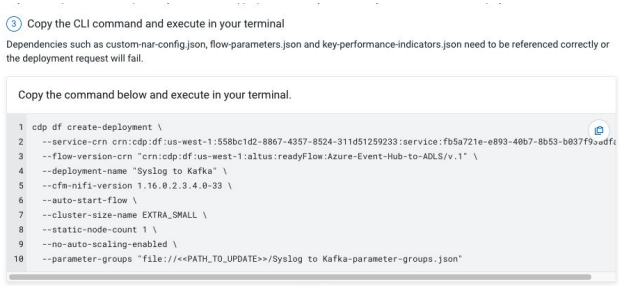
High velocity NiFi releases

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



Powerful CLI for automation

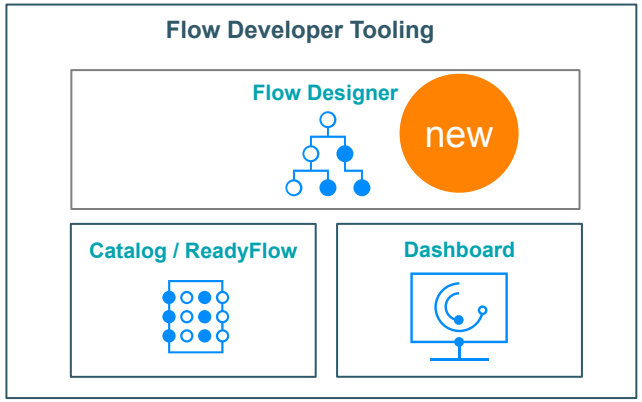
Automate the entire flow lifecycle with the CLI including single command flow deployment

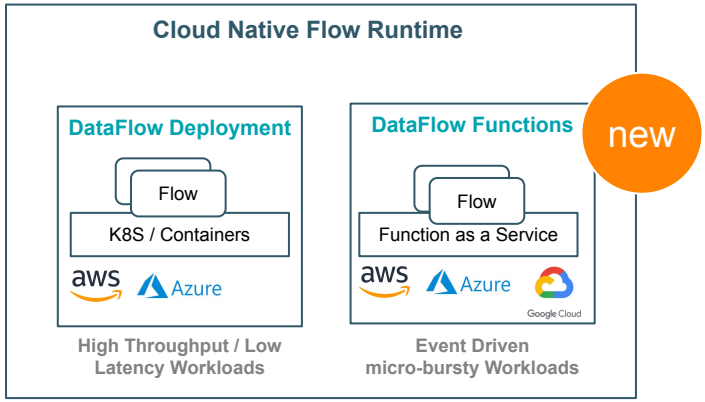


CLOUDERA DATAFLOW FOR THE PUBLIC CLOUD (CDF-PC)

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.



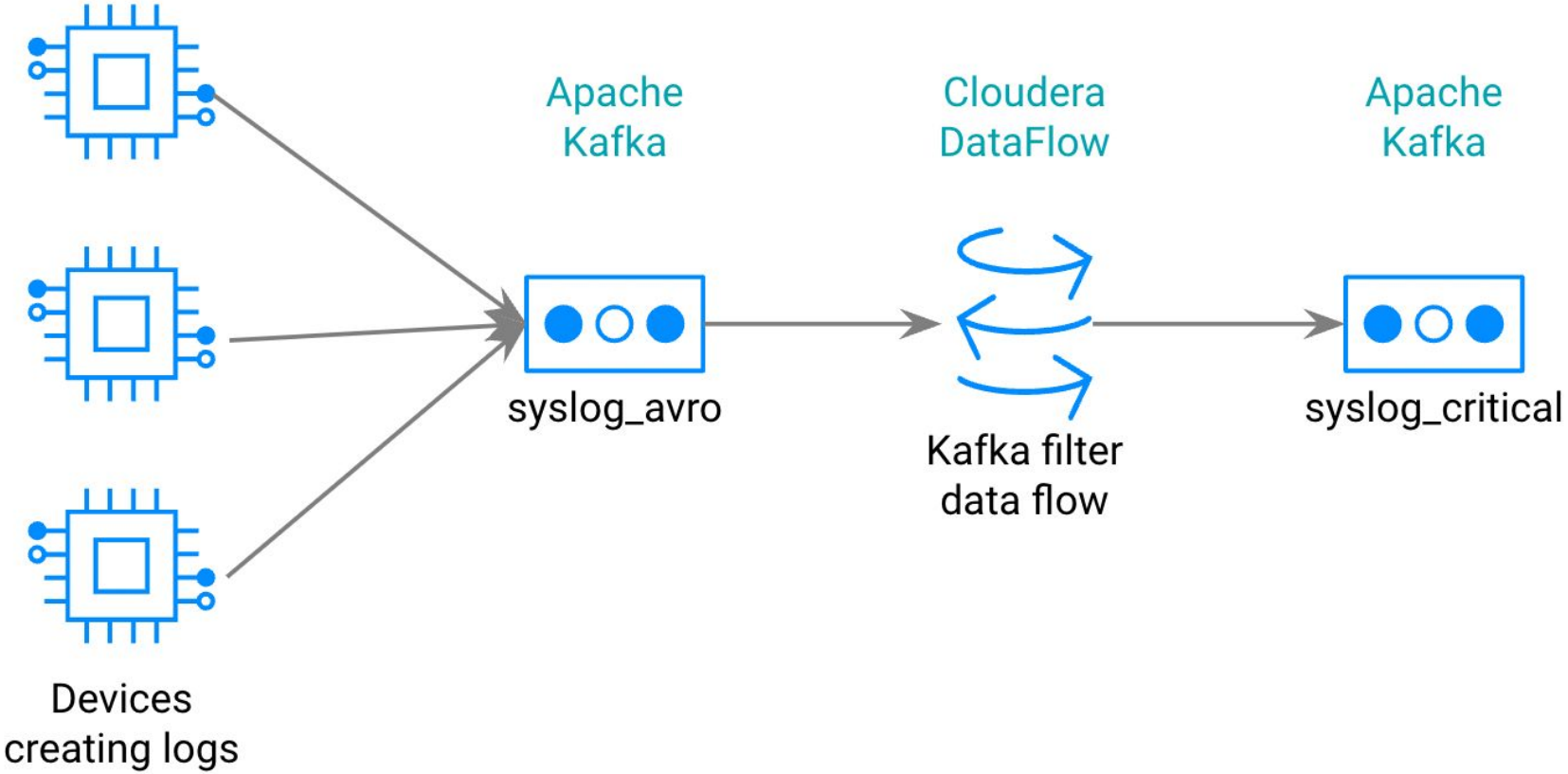


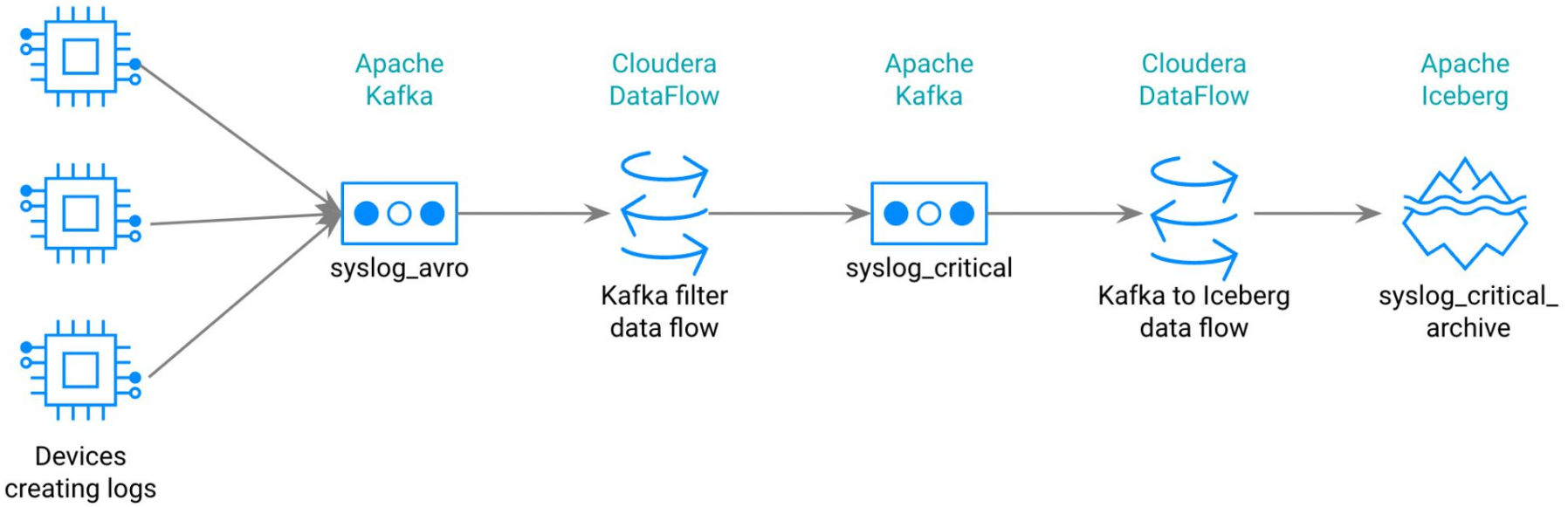
Productivity Tooling for Developers – Flow designer combined w/ catalog of flows provides developers the agility & extensibility to build data movement flows in minutes

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


Apache
Kafka
Cloudera
DataFlow
Apache
Kafka
syslog_avro
Kafka filter
data flow
syslog_critical
Devices
creating logs


Apache
Kafka
Cloudera
DataFlow
Apache
Kafka
Cloudera
DataFlow
Apache
Iceberg
syslog_avro
Kafka filter
data flow
syslog_critical
Kafka to Iceberg
data flow
syslog_critical_
archive
Devices
creating logs

DAILY ZOOM

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

CLOUDERA

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

Added

Kafka to Iceberg

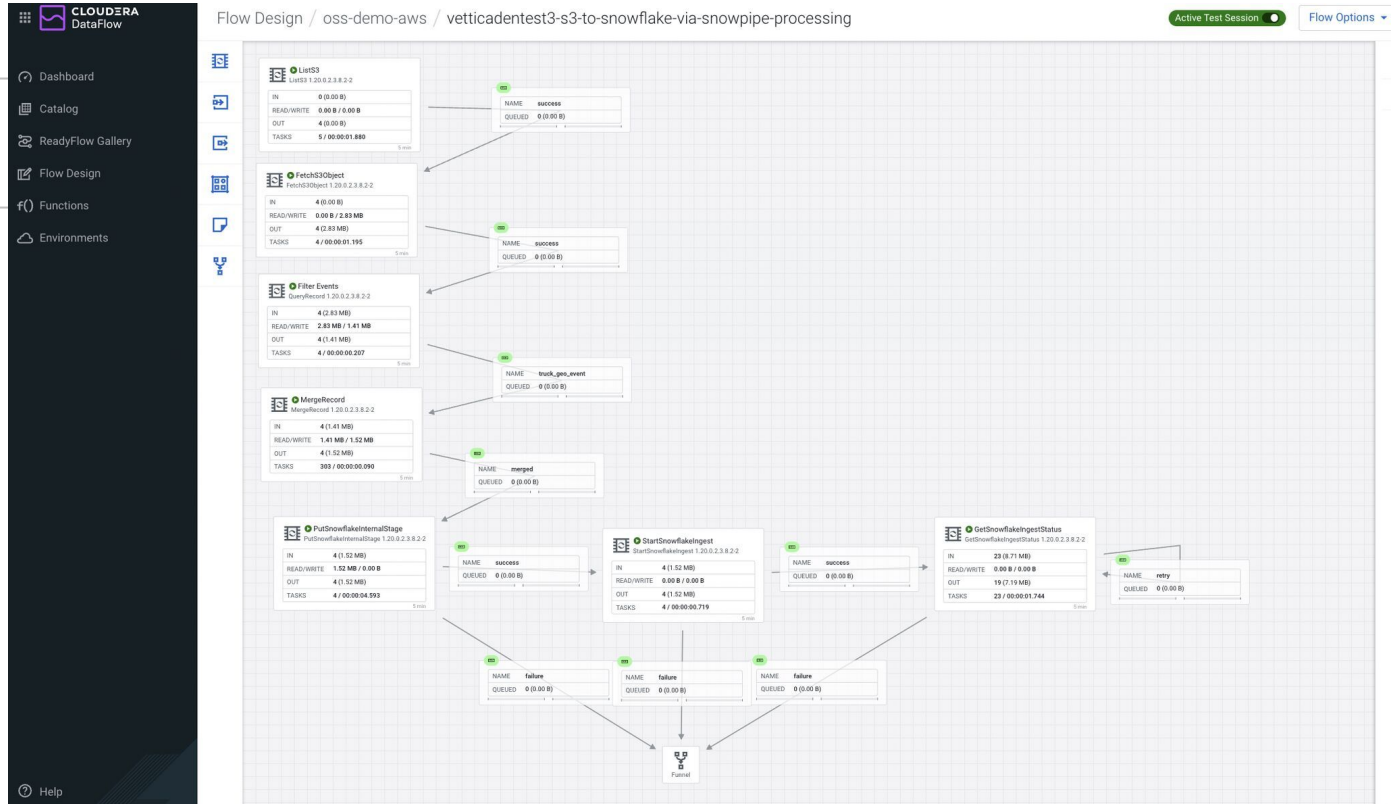
Version 1

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

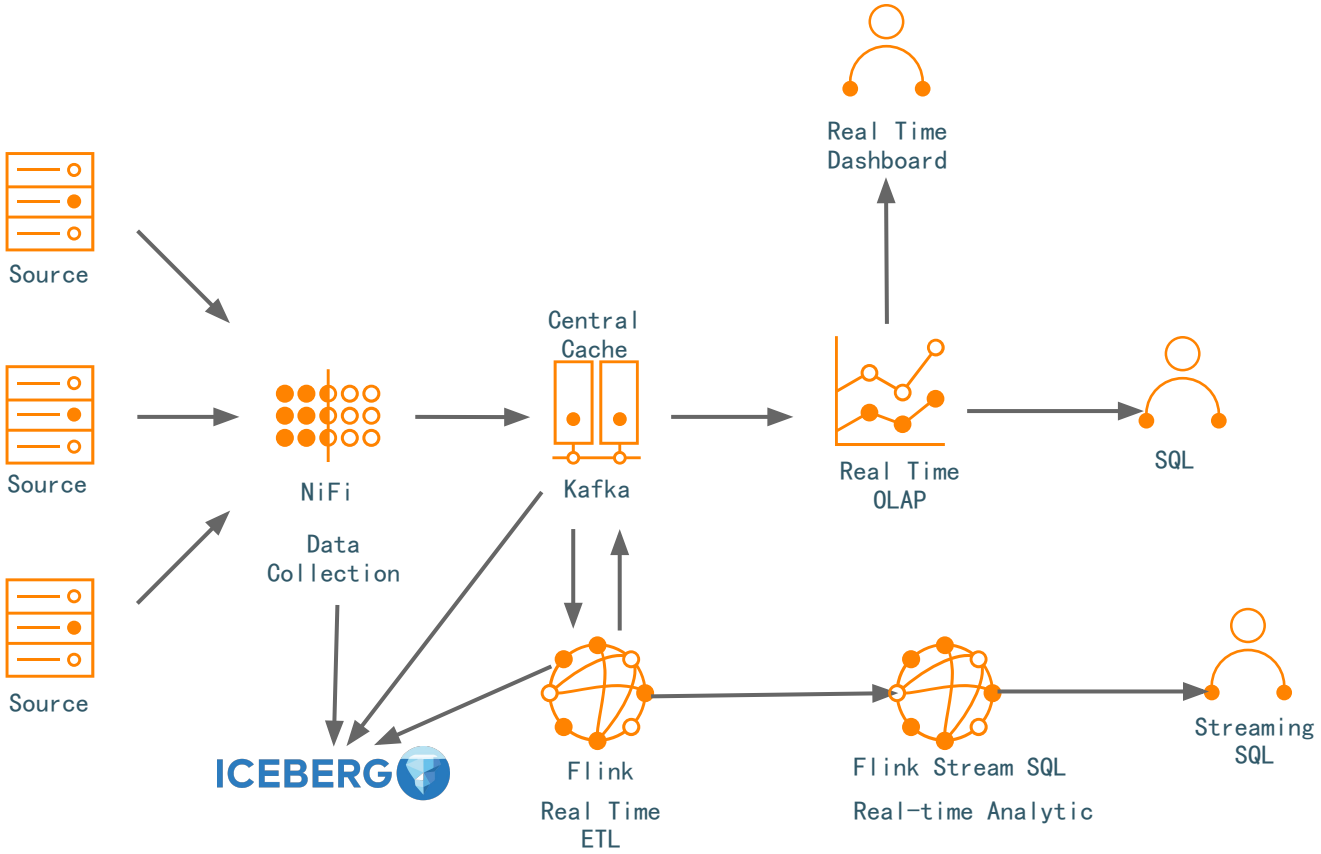


https://www.linkedin.com/posts/georgevetticaden_just-finished-up-a-trial-run-of-the-new-data-activity-7058557234556907520-W6O2/

HELP

Tim Spann
@PaasDev // Blog: **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