Pulsar Virtual Summit
Europe 2023

Timothy Spann
Principal Developer Advocate • Cloudera

Using Apache NiFi with Apache Pulsar for Fast Data On-Ramp

PULSAR SUMMIT

Hosted by StreamNative

Tim Spann
Principal Developer Advocate
Cloudera

Pulsar Virtual Summit
Europe | May 23 2023

Ecosystem

Using Apache NiFi with Apache Pulsar for Fast Data On-Ramp

PULSAR SUMMIT

Hosted by StreamNative

Timothy Spann
Principal Developer Advocate
Cloudera

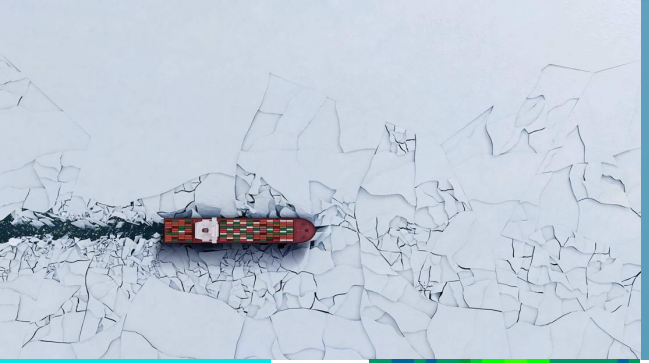
@PaasDev // Blog:
www.datainmotion.dev
Principal Developer Advocate, Cloudera
Princeton Future of Data Meetup.
ex-Pivotal, ex-Hortonworks,
ex-StreamNative, ex-PwC

https://medium.com/@tspann
Apache NiFi x Apache Kafka x Apache Flink x Java x Apache Pulsar

Using Apache Nifi with Apache Pulsar for Fast Data On-Ramp

As the Pulsar communities grows, more and more connectors will be added. To enhance the availability of sources and sinks and to make use of the greater Apache Streaming community, joining forces between Apache NiFi and Apache Pulsar is a perfect fit. Apache NiFi also adds the benefits of ELT, ETL, data crunching, transformation, validation and batch data processing. Once data is ready to be an event, NiFi can launch it into Pulsar at light speed.

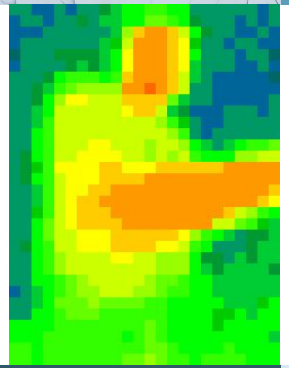
I will walk through how to get started, some use cases and demos and answer questions.



IoT

CLOUDERA & IOT
REAL-TIME MONITORING
OF YOUR HEART'S HEALTH.
THANKS DATA.

CLOUDERA



CLOUDERA

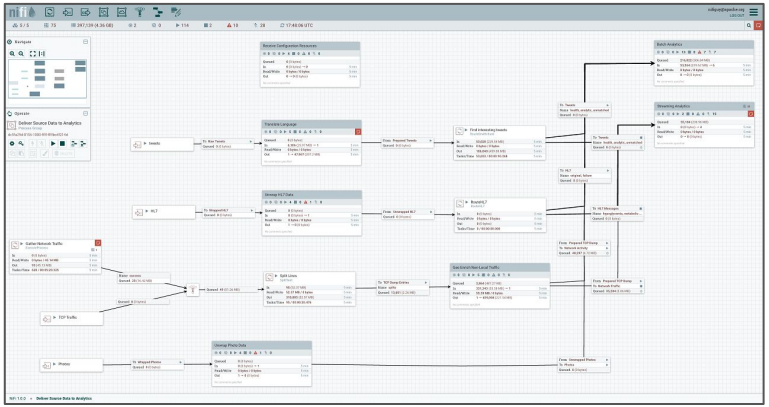
EDGE
2AI

CLOUDERA

ENTERPRISE
DATA
CLOUD

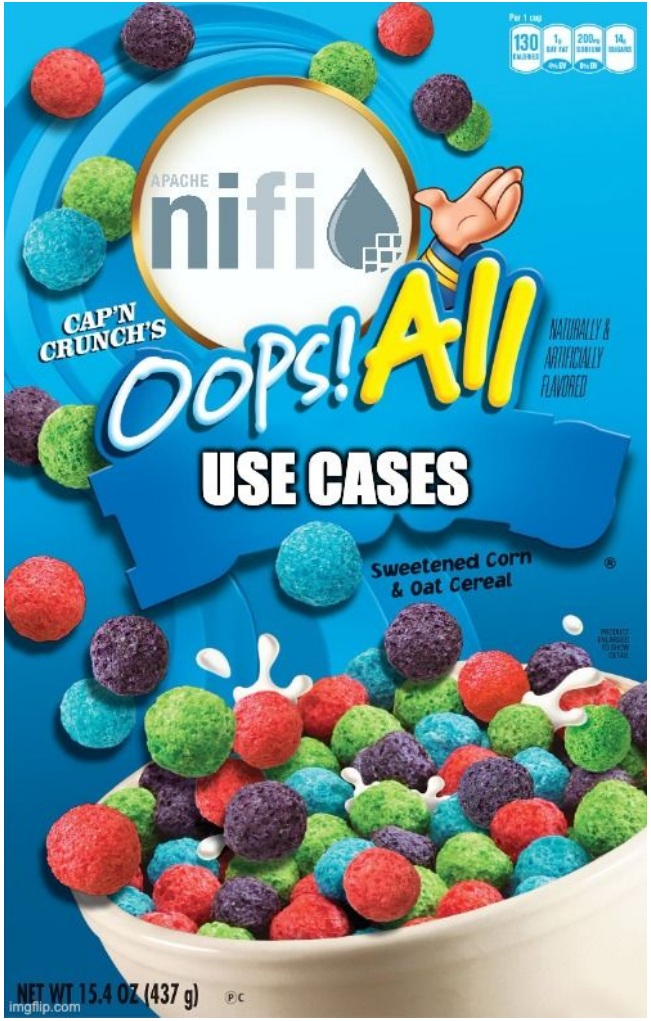
CLOUDERA

APACHE nifi



APACHE NIFI registry

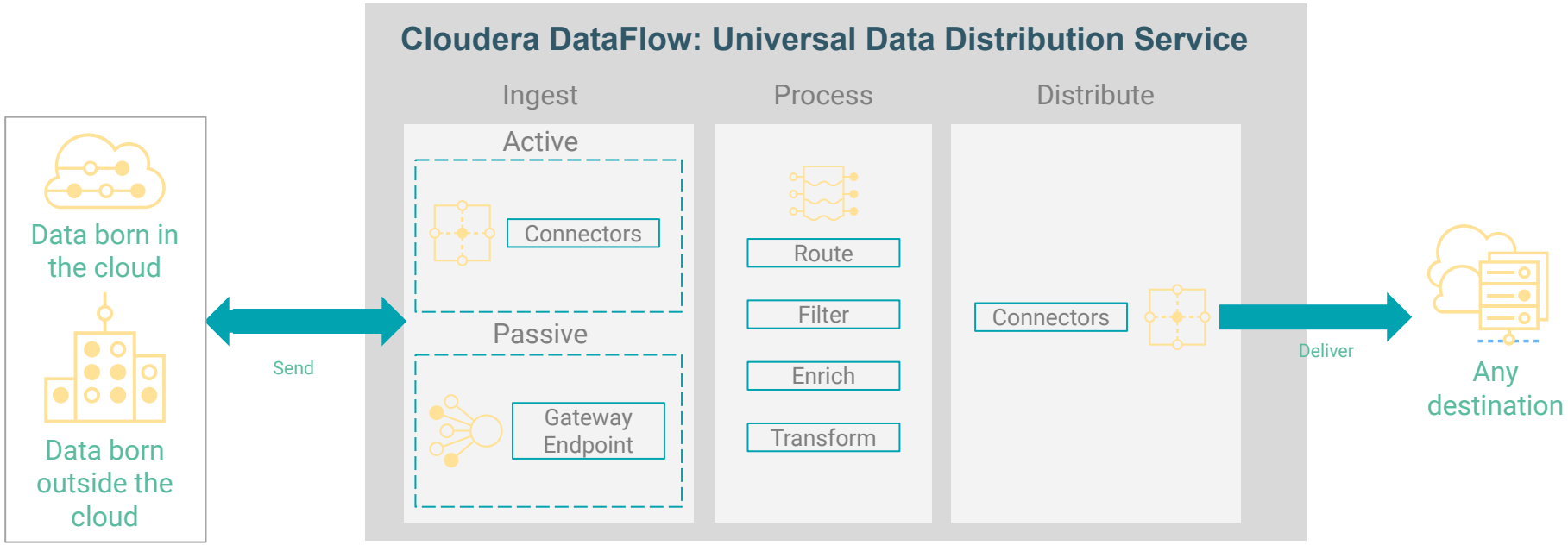
Advanced tooling to industrialize flow development
(Flow Development Life Cycle)



PULSAR SUMMIT

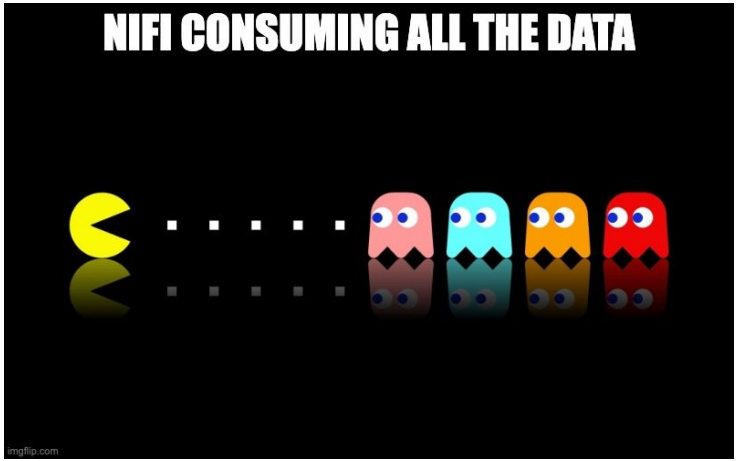
Universal Data Distribution

Connect to Any Data Source Anywhere then Process and Deliver to Any Destination



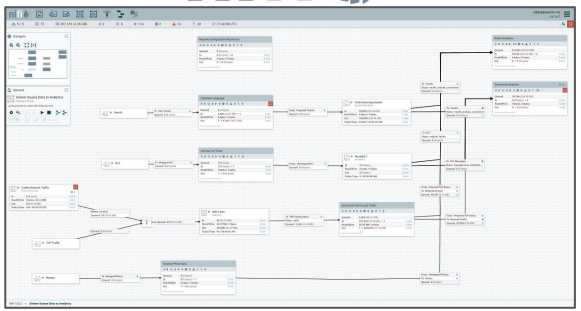
What is Apache NiFi?

Apache NiFi is a scalable, real-time streaming data platform that collects, curates, and analyzes data so customers gain key insights for immediate actionable intelligence.



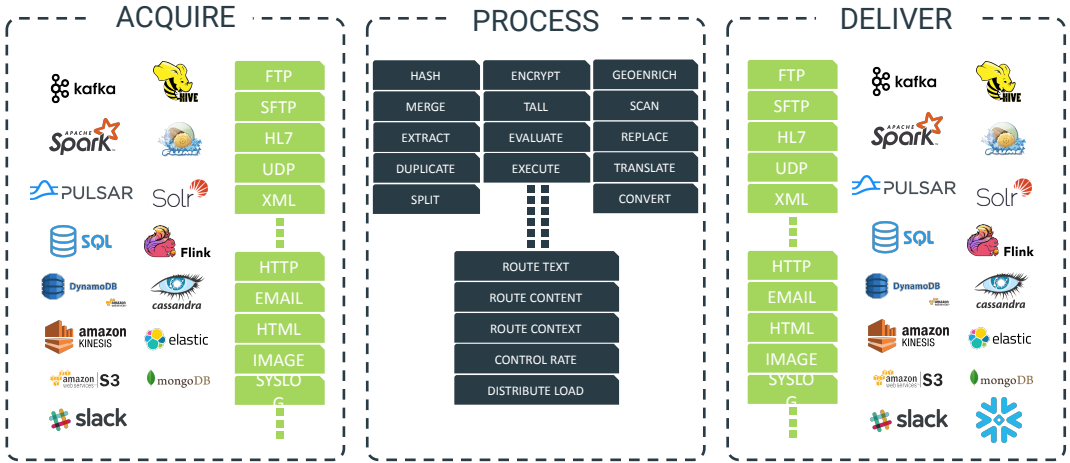
Apache NiFi

Enable easy ingestion, routing, management and delivery of any data anywhere (Edge, cloud, data center) to any downstream system with built in end-to-end security and provenance



APACHE NIFI registry

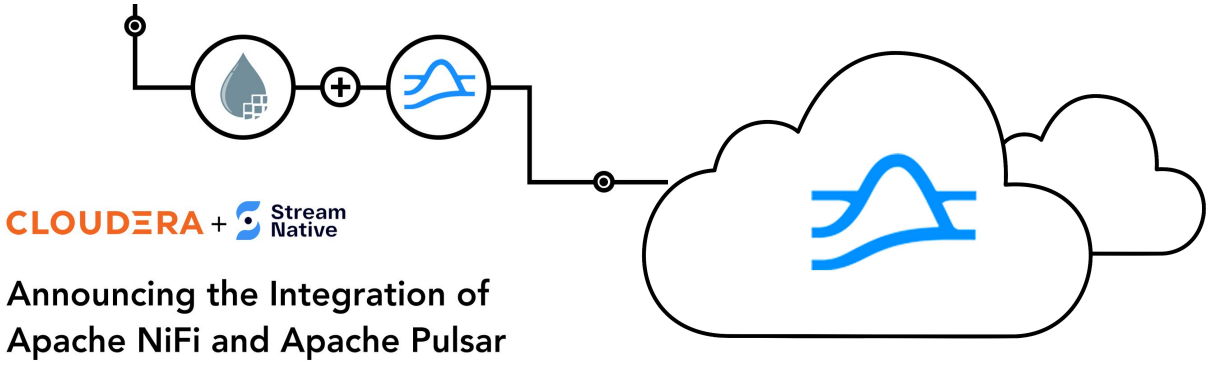
Advanced tooling to industrialize flow development (Flow Development Life Cycle)

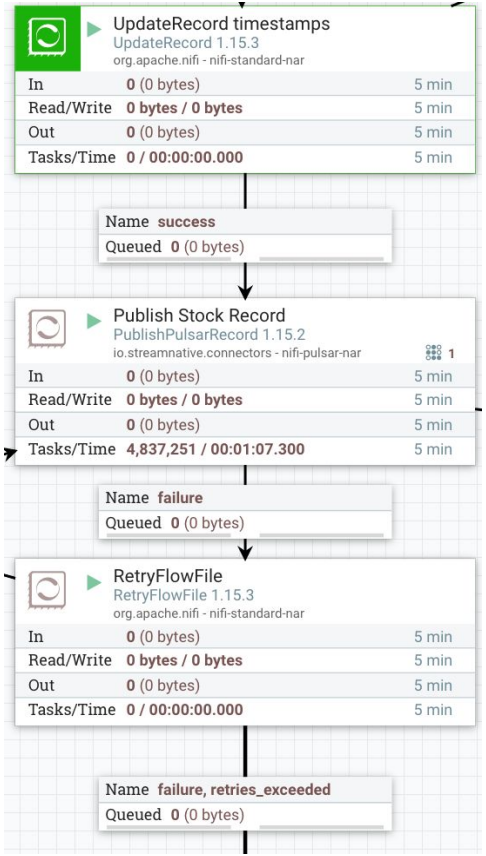


- Over 450 Prebuilt Processors
- Easy to build your own
- Parse, Enrich & Apply Schema
- Filter, Split, Merger & Route
- Throttle & Backpressure

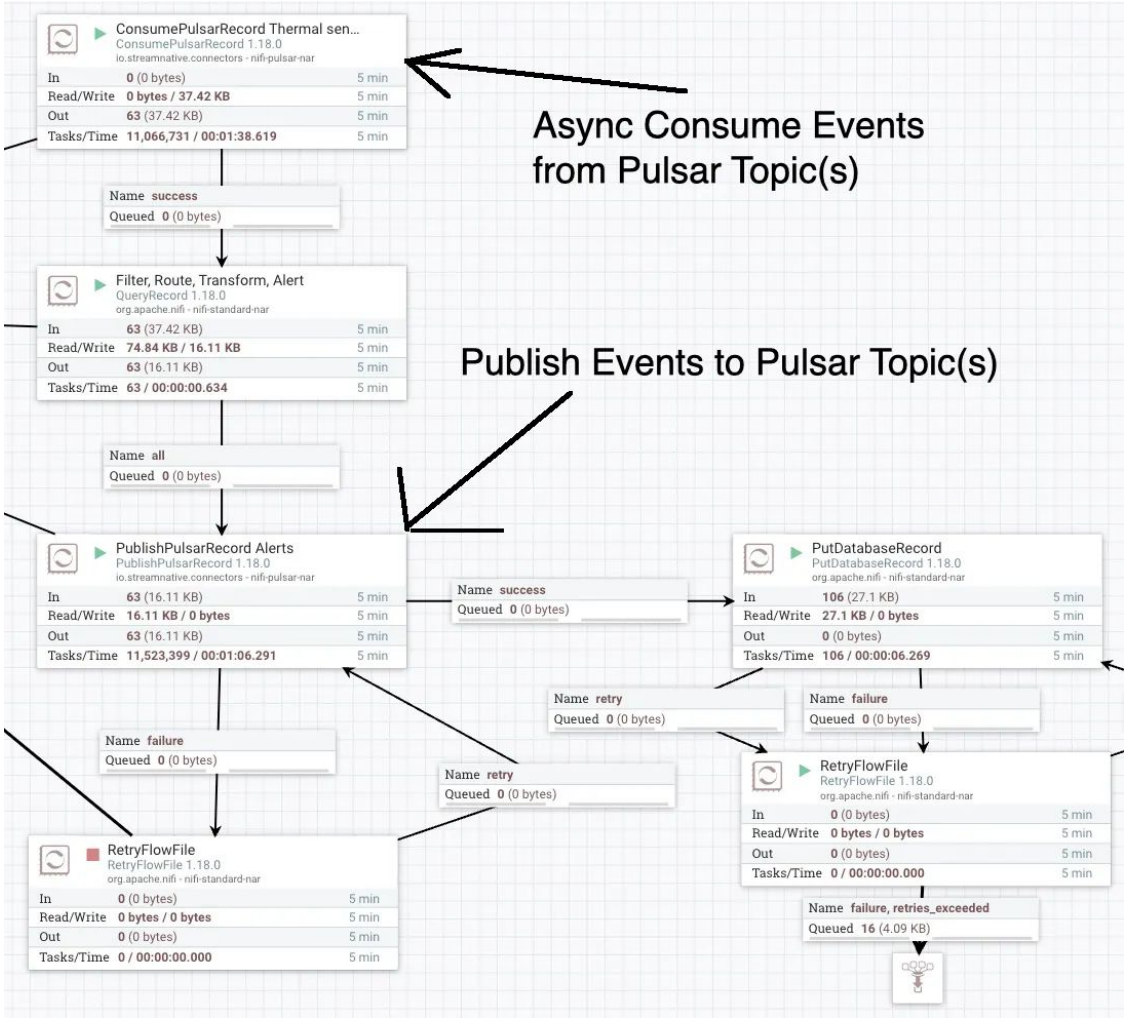
- Guaranteed Delivery
- Full data provenance from acquisition to delivery
- Diverse, Non-Traditional Sources
- Eco-system integration

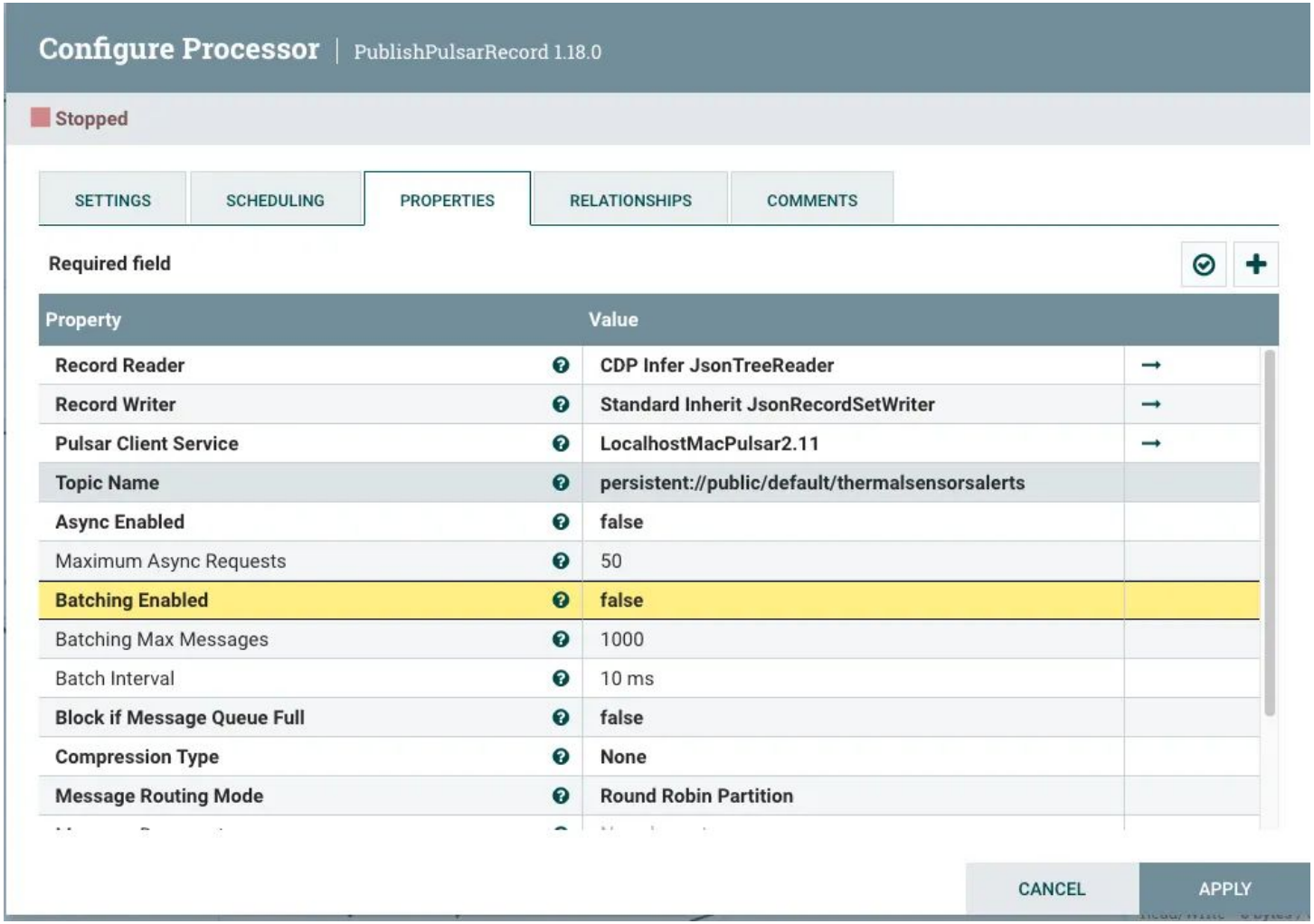
Apache NiFi Pulsar Connector





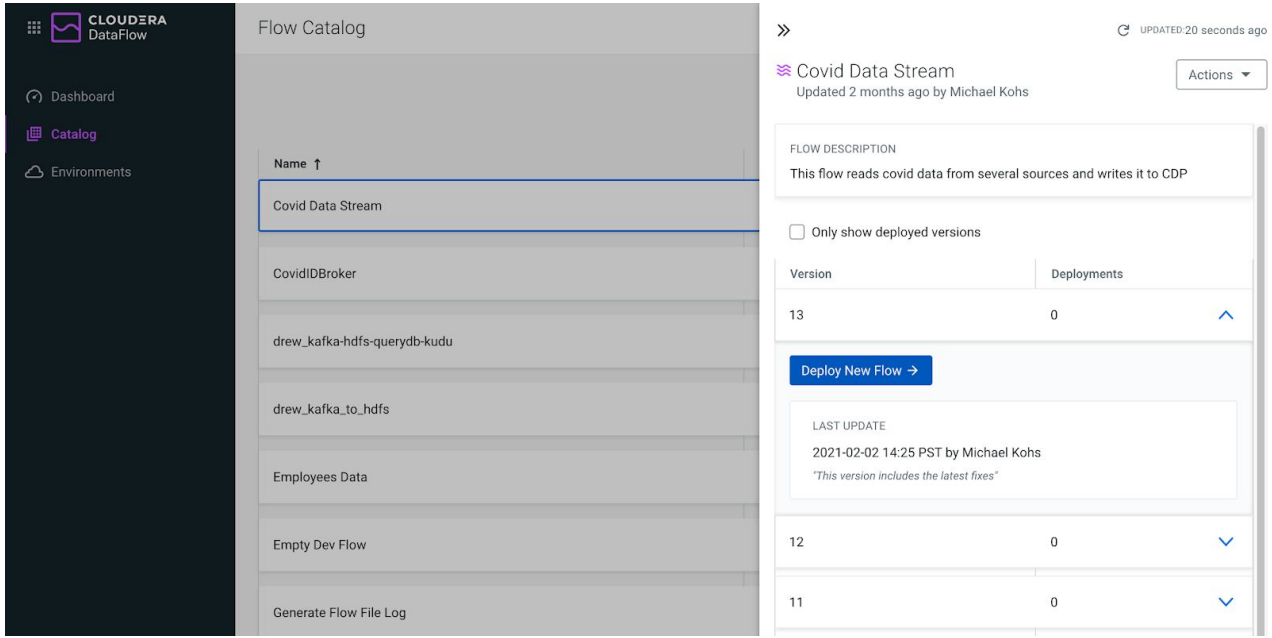
https://streamnative.io/apache-nifi-connector/


ConsumePulsarRecord Thermal sen...
ConsumePulsarRecord 1.18.0
io.streamnative.connectors - nifi-pulsar-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 37.42 KB 5 min
Out 63 (37.42 KB) 5 min
Tasks/Time 11,066,731 / 00:01:38.619 5 min
Async Consume Events
from Pulsar Topic(s)
Name success
Queued 0 (0 bytes)
Filter, Route, Transform, Alert
QueryRecord 1.18.0
org.apache.nifi - nifi-standard-nar
In 63 (37.42 KB) 5 min
Read/Write 74.84 KB / 16.11 KB 5 min
Out 63 (16.11 KB) 5 min
Tasks/Time 63 / 00:00:00.634 5 min
Publish Events to Pulsar Topic(s)
Name all
Queued 0 (0 bytes)
PublishPulsarRecord Alerts
PublishPulsarRecord 1.18.0
io.streamnative.connectors - nifi-pulsar-nar
In 63 (16.11 KB) 5 min
Read/Write 16.11 KB / 0 bytes 5 min
Out 63 (16.11 KB) 5 min
Tasks/Time 11,523,399 / 00:01:06.291 5 min
Name success
Queued 0 (0 bytes)
PutDatabaseRecord
PutDatabaseRecord 1.18.0
org.apache.nifi - nifi-standard-nar
In 106 (27.1 KB) 5 min
Read/Write 27.1 KB / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 106 / 00:00:06.269 5 min
Name retry
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)
Name retry
Queued 0 (0 bytes)
RetryFlowFile
RetryFlowFile 1.18.0
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
RetryFlowFile
RetryFlowFile 1.18.0
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
Name failure, retries_exceeded
Queued 16 (4.09 KB)


Configure Processor | PublishPulsarRecord 1.18.0
Stopped
SETTINGS
SCHEDULING
PROPERTIES
RELATIONSHIPS
COMMENTS
Required field
Property
Value
Record Reader
CDP Infer JsonTreeReader
Record Writer
Standard Inherit JsonRecordSetWriter
Pulsar Client Service
LocalhostMacPulsar2.11
Topic Name
persistent://public/default/thermalsensorsalerts
Async Enabled
false
Maximum Async Requests
50
Batching Enabled
false
Batching Max Messages
1000
Batch Interval
10 ms
Block if Message Queue Full
false
Compression Type
None
Message Routing Mode
Round Robin Partition
CANCEL
APPLY

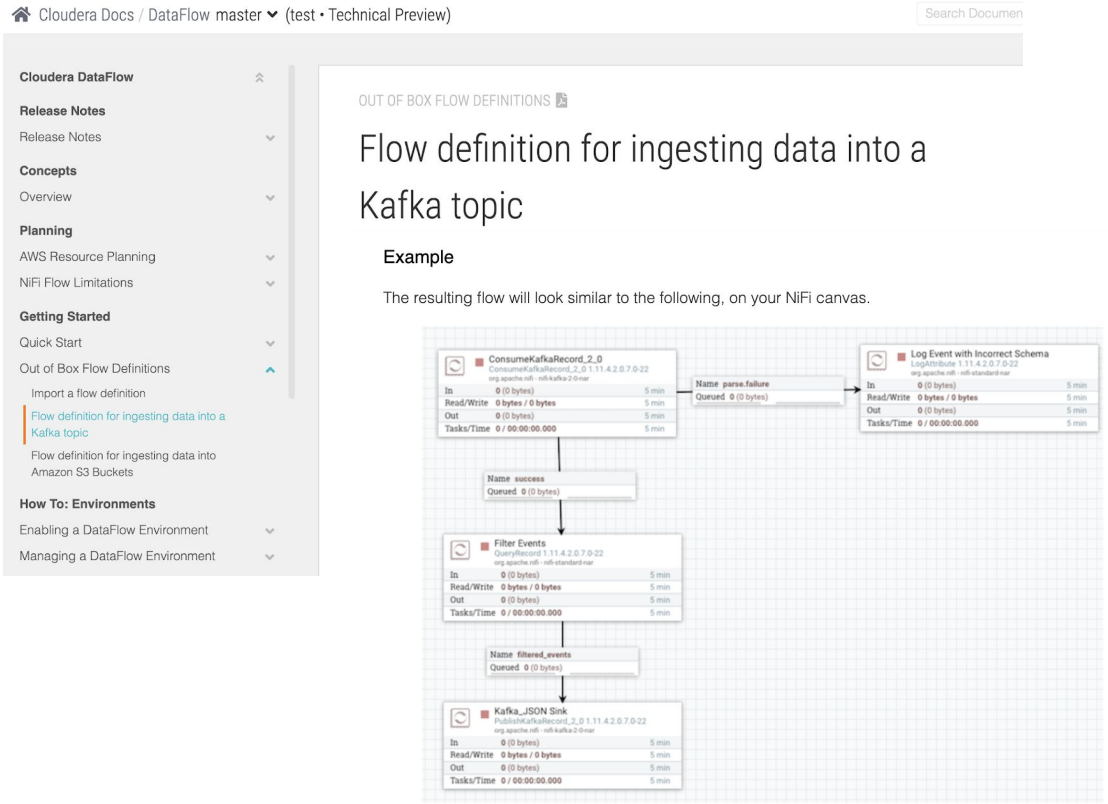
Flow Catalog

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



ReadyFlows

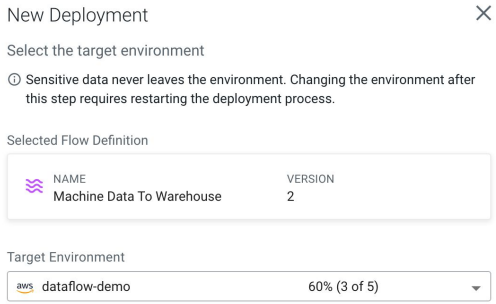
- Cloudera provided flow definitions
- Cover most common data flow use cases
- Can be deployed and adjusted as needed
- Made available through docs during Tech Preview



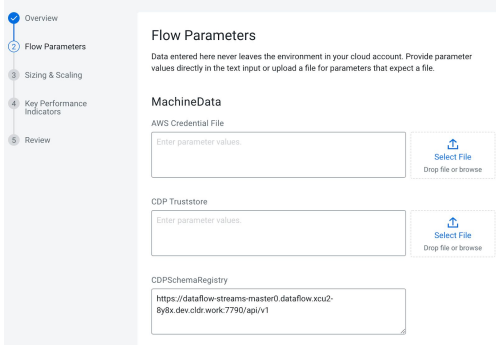
Deployment Wizard

- Turns flow definitions into flow deployments
- Guides users through providing required configuration
- Pick from pre-defined NiFi node sizes
- Define KPIs for the deployment

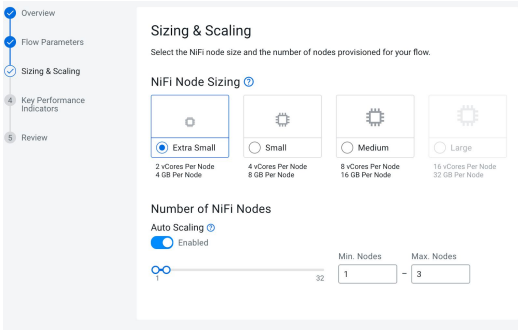
Start Deployment Wizard



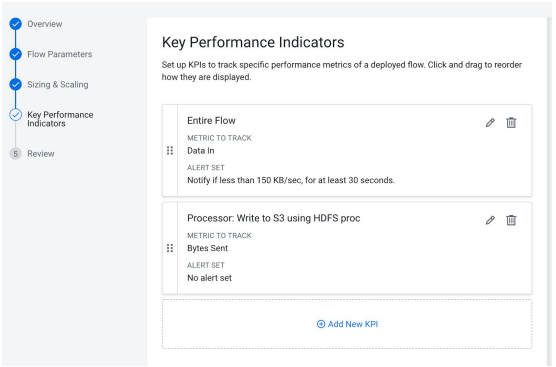
Provide Parameters



Configure Sizing & Scaling



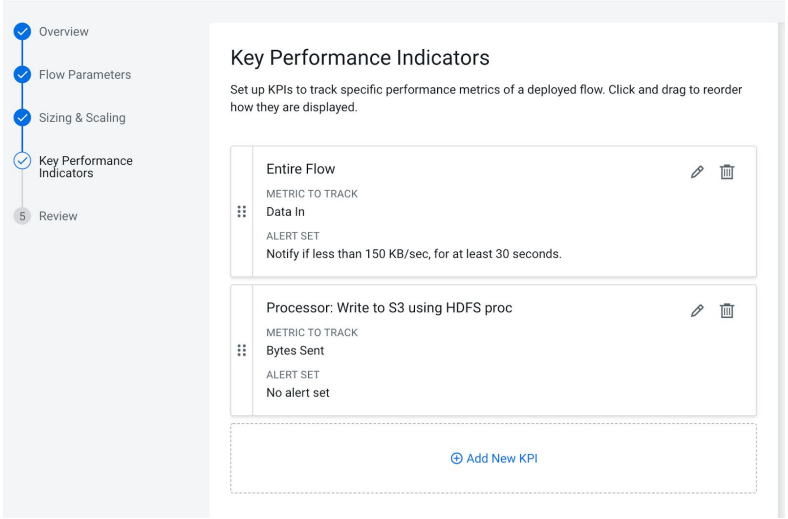
Define KPIs



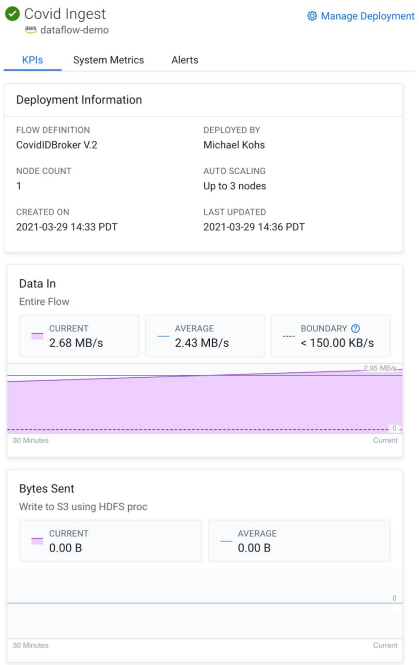
Key Performance Indicators

- 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

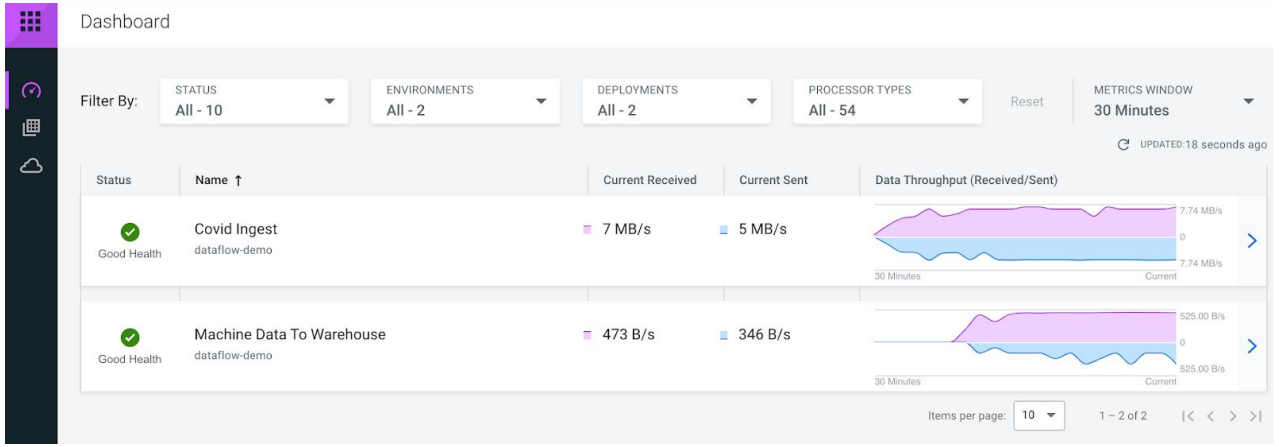


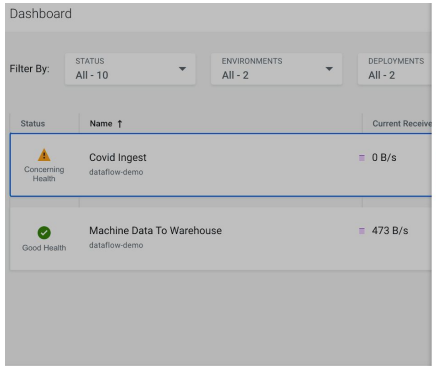
KPI Monitoring

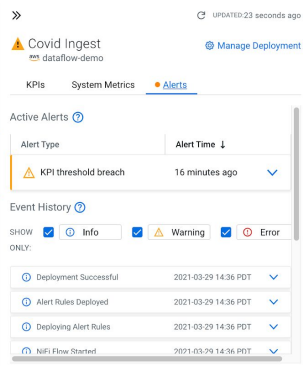


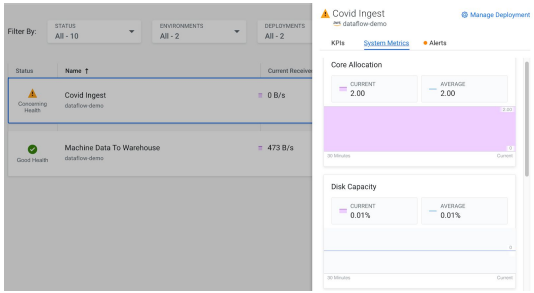
Dashboard

- 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



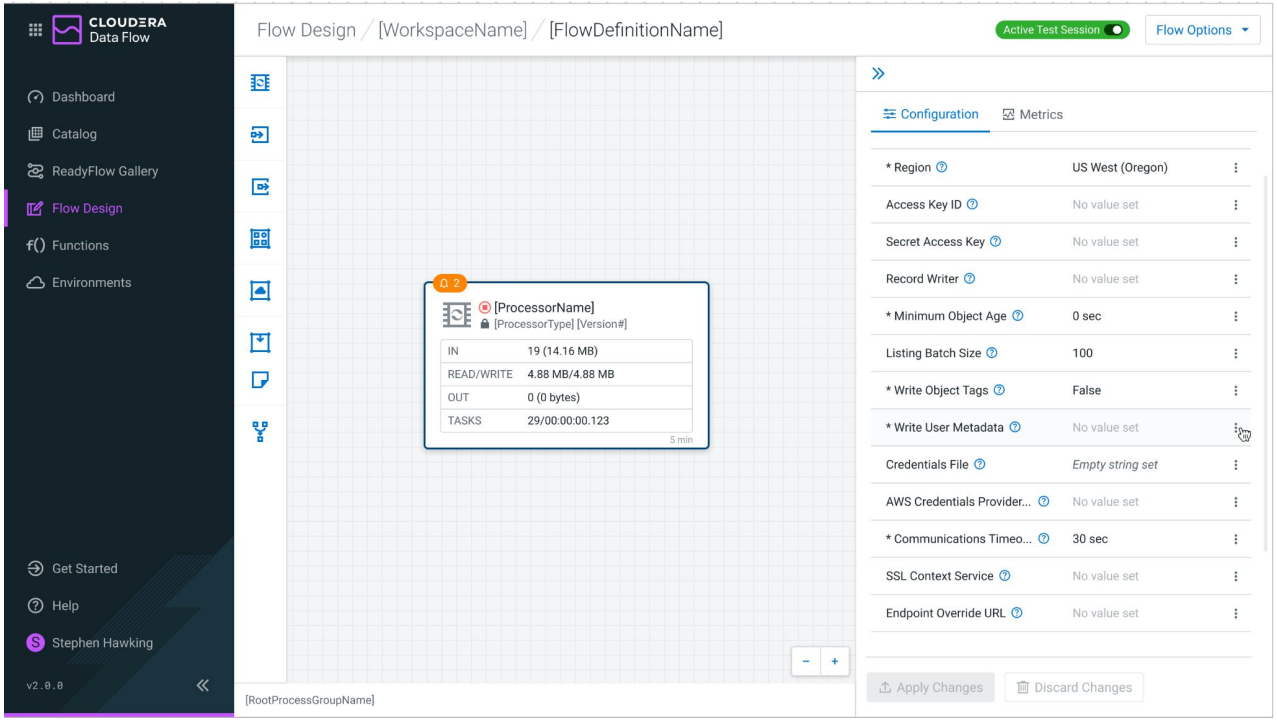






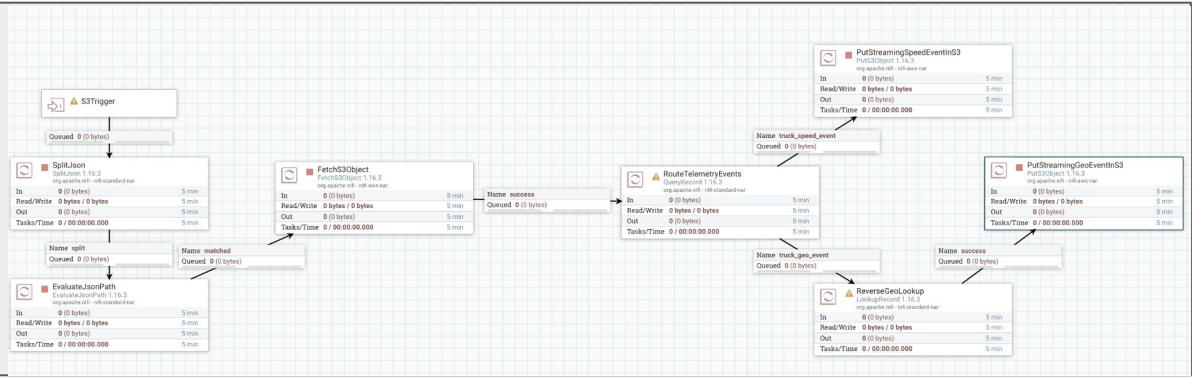
DATA FLOW DESIGN FOR EVERYONE

- 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



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

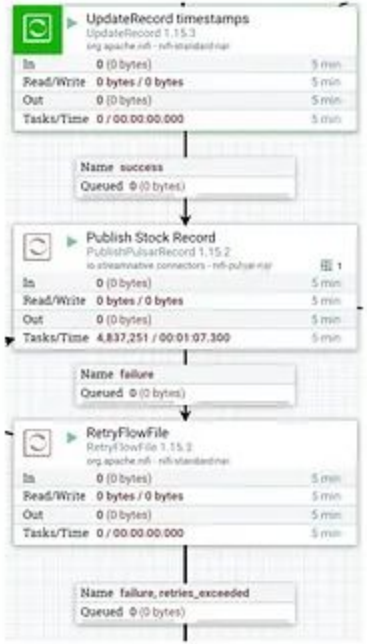
With the DFF State feature, build flows to create customized triggers allowing access to on-premises or external services

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

NiFi → Pulsar

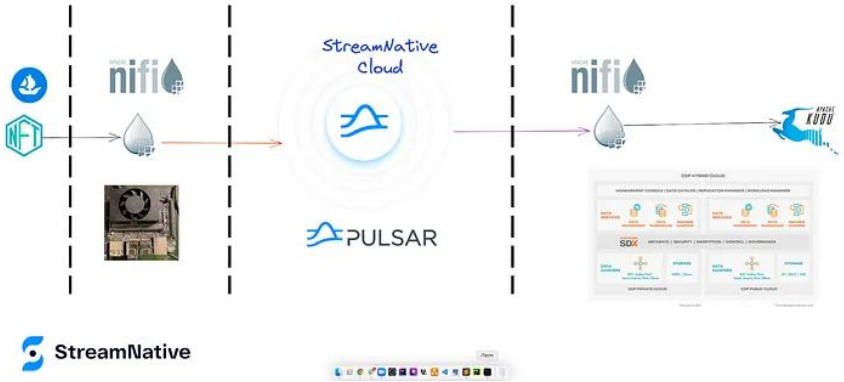
Apache NiFi Pulsar Connector

https://github.com/streamnative/pulsar-nifi-bundle

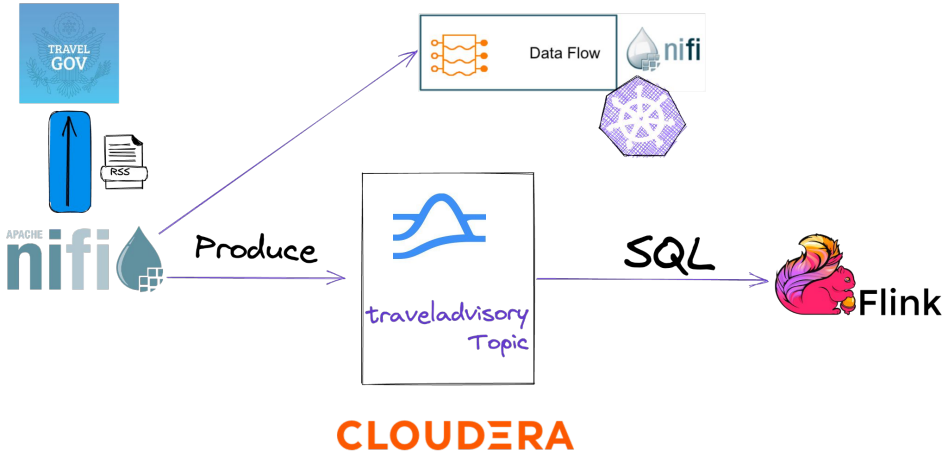


StreamNative

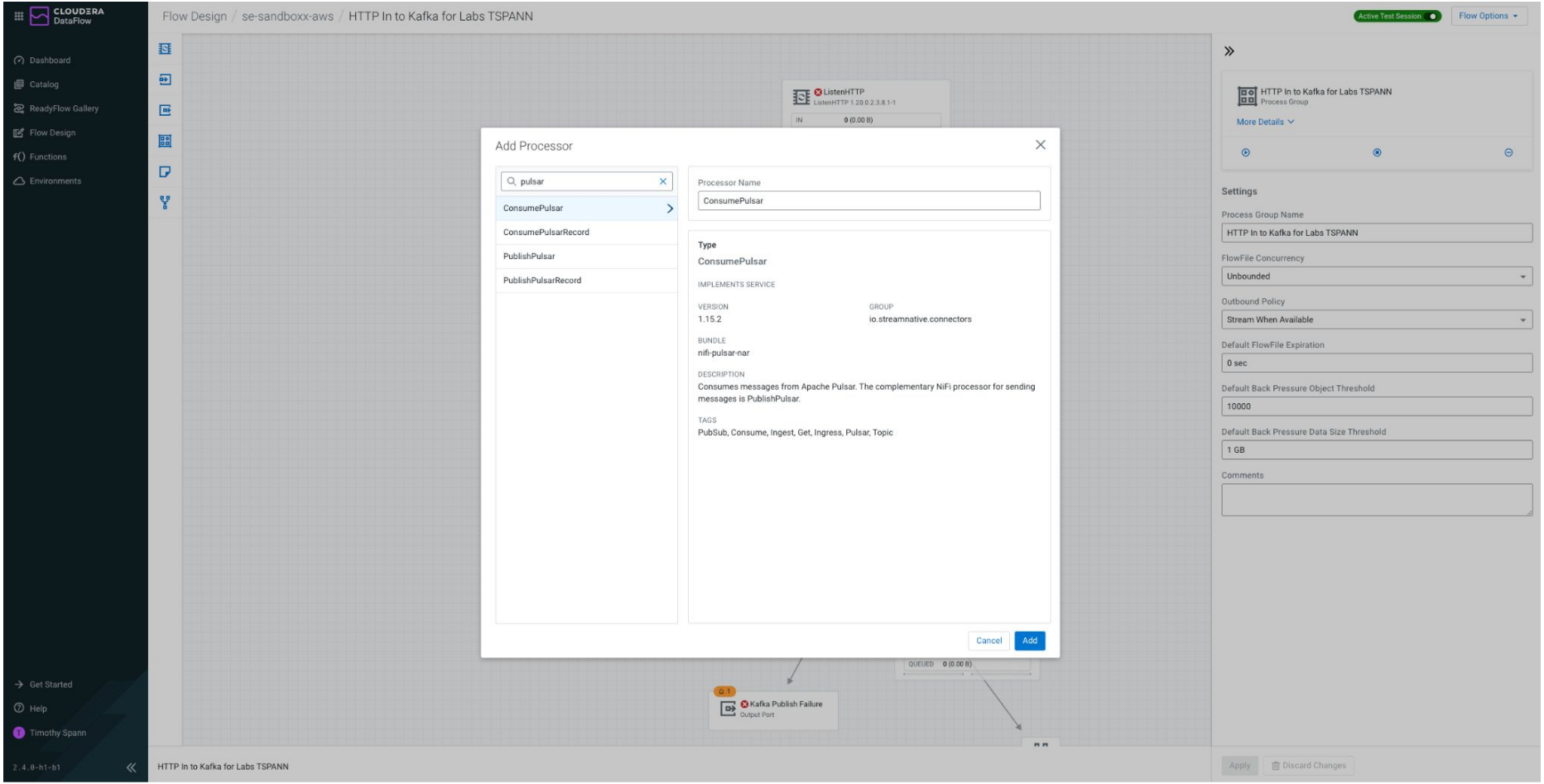
Streaming NFTs

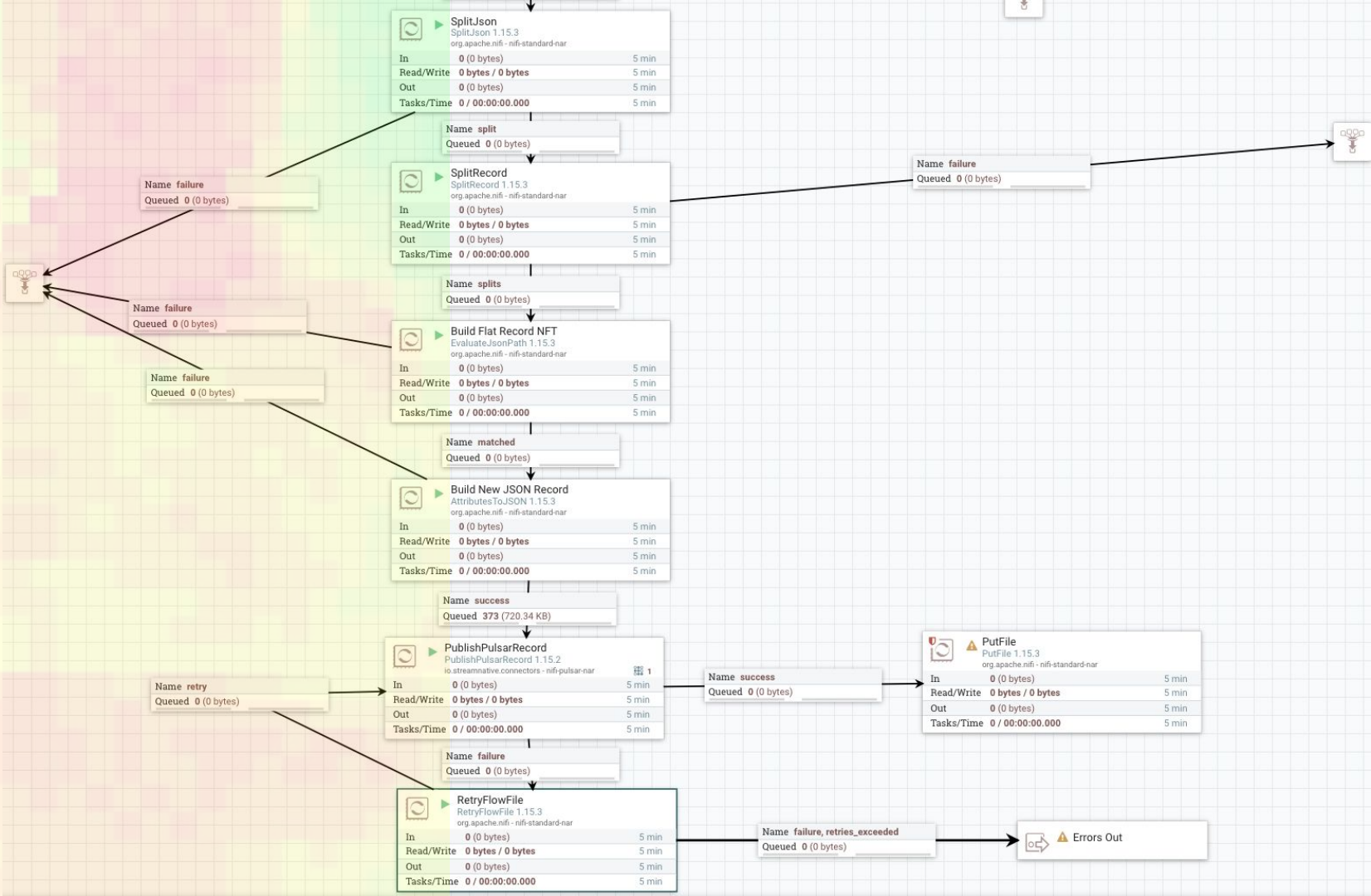


https://github.com/tspannhw/FLiPN-NFT



https://github.com/tspannhw/FLaNK-TravelAdvisory


CLOUDERA DataFlow
Flow Design / se-sandboxx-aws / HTTP In to Kafka for Labs TSPANN
Active Test Session
Flow Options
Dashboard
Catalog
ReadyFlow Gallery
Flow Design
Functions
Environments
Get Started
Help
Timothy Spann
2.4.0-h1-b1
ListenHTTP
ListenHTTP 1.20.0.2.3.8.1-1
IN 0 (0.00 B)
Add Processor
pulsar
ConsumePulsar
ConsumePulsarRecord
PublishPulsar
PublishPulsarRecord
Processor Name
ConsumePulsar
Type
ConsumePulsar
IMPLEMENTS SERVICE
VERSION
1.15.2
GROUP
io.streamnative.connectors
BUNDLE
nifi-pulsar-nar
DESCRIPTION
Consumes messages from Apache Pulsar. The complementary NiFi processor for sending messages is PublishPulsar.
TAGS
PubSub, Consume, Ingest, Get, Ingress, Pulsar, Topic
Cancel
Add
QUEUED 0 (0.00 B)
Q 1
Kafka Publish Failure
Output Port
HTTP In to Kafka for Labs TSPANN
Process Group
More Details
Settings
Process Group Name
HTTP In to Kafka for Labs TSPANN
FlowFile Concurrency
Unbounded
Outbound Policy
Stream When Available
Default FlowFile Expiration
0 sec
Default Back Pressure Object Threshold
10000
Default Back Pressure Data Size Threshold
1 GB
Comments
Apply
Discard Changes
HTTP In to Kafka for Labs TSPANN



Questions

Pulsar Virtual Summit
Europe 2023

Your name
Contact information

Thank you!

PULSAR SUMMIT

Hosted by StreamNative