

Pulsar Virtual Summit
Europe 2023

Timothy Spann
Principal Developer Advocate • Cloudera

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



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



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>

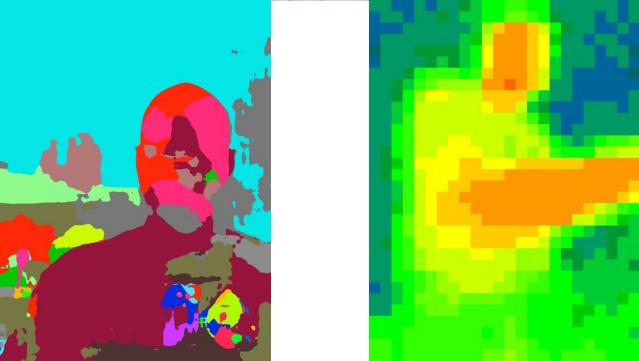
Apache NiFi x Apache Kafka x Apache
Flink x Java x Apache Pulsar

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.

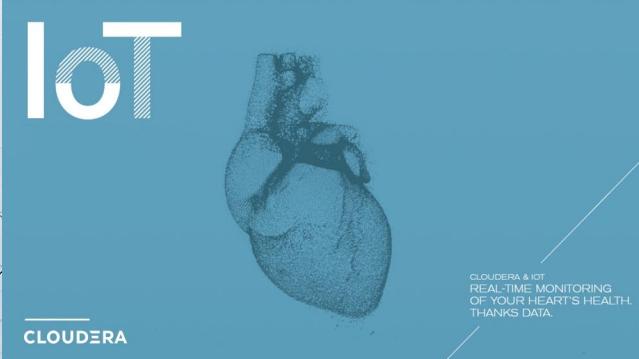
I will walk through how to get started, some use cases and demos and answer questions.





**ENTERPRISE
DATA CLOUD**

CLOUDERA

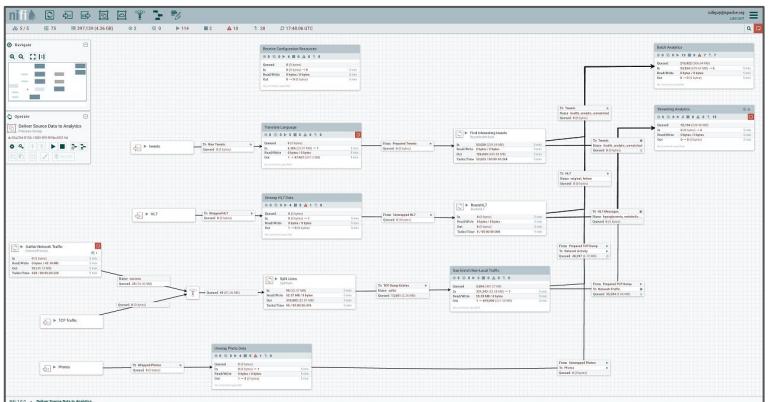


CLOUDERA

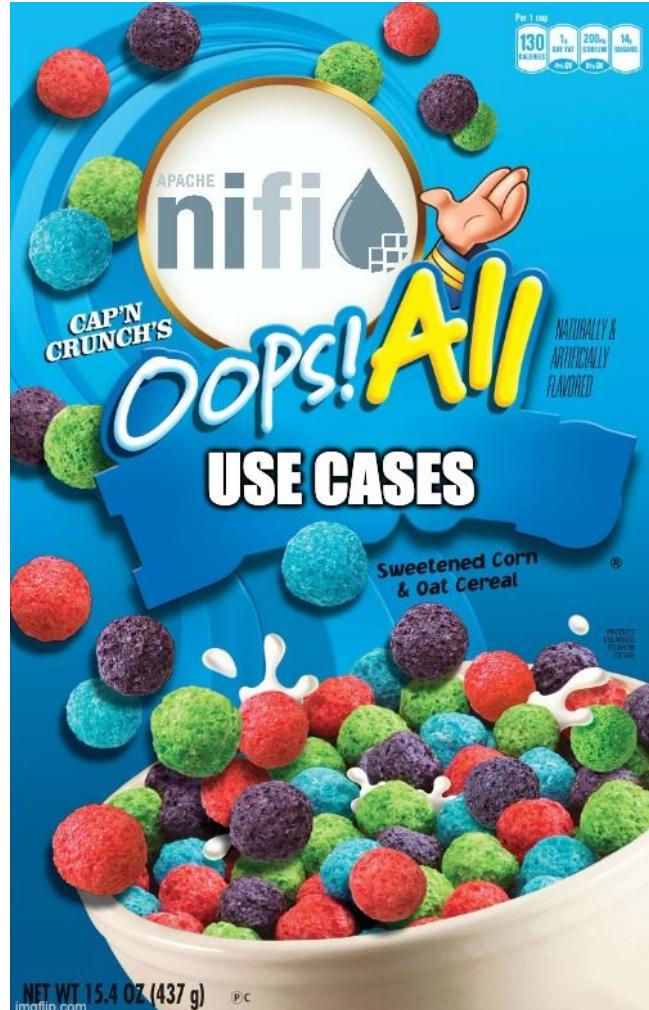


**EDGE
2AI**



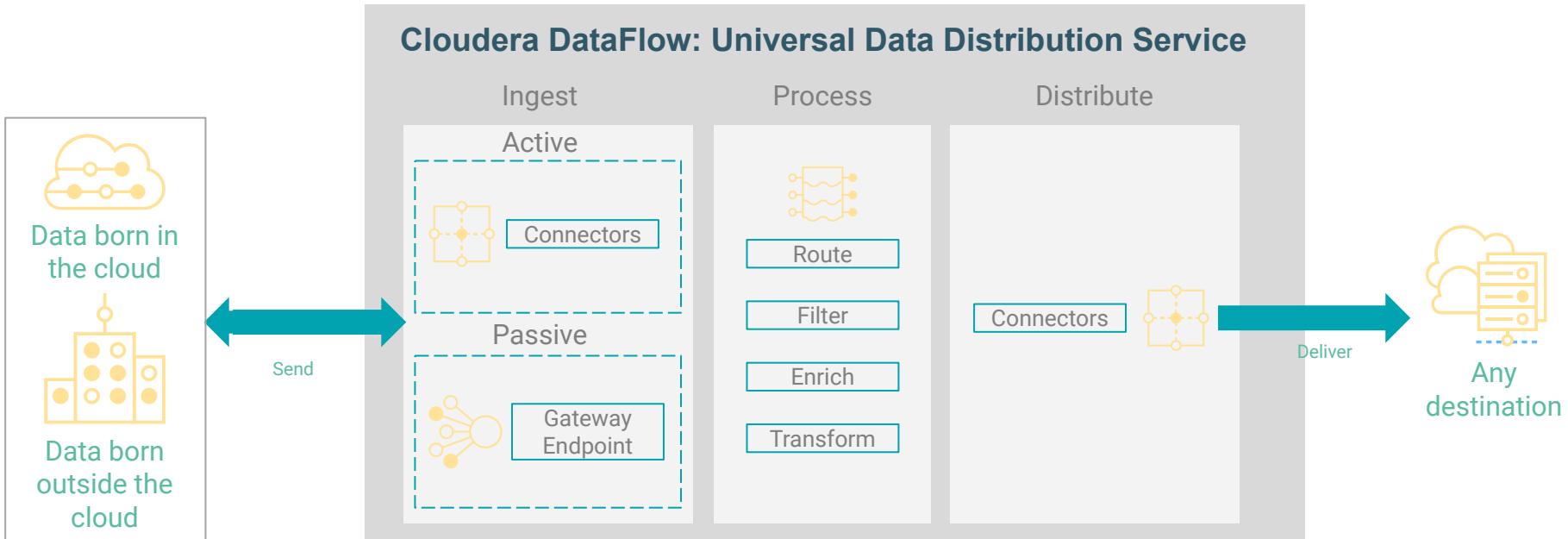


Advanced tooling to industrialize flow development
(Flow Development Life Cycle)



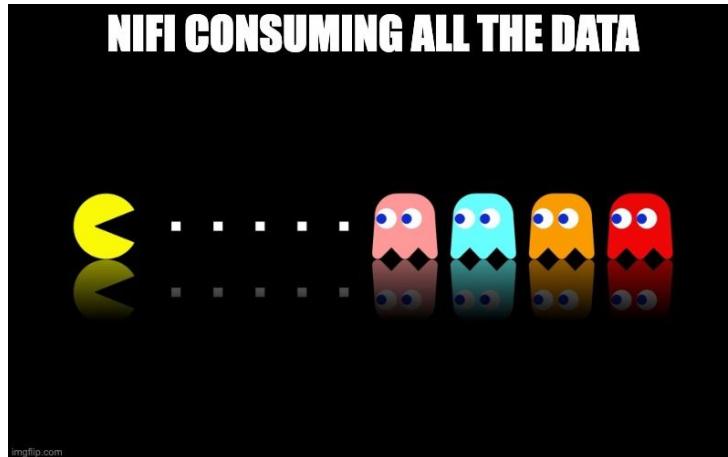
Universal Data Distribution

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



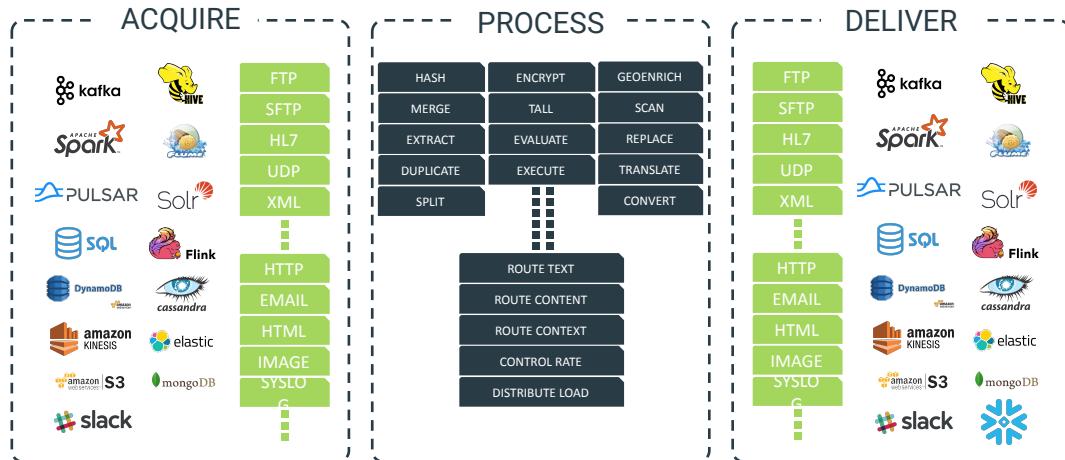
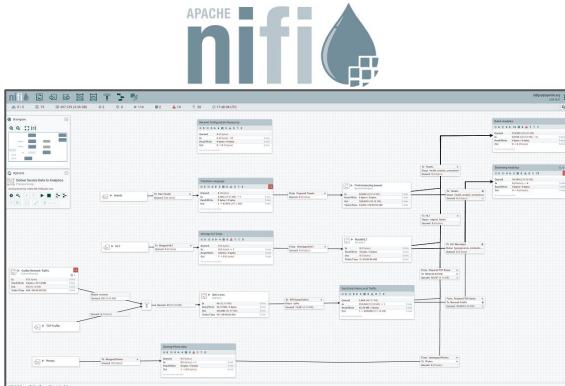
What is Apache NiFi?

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

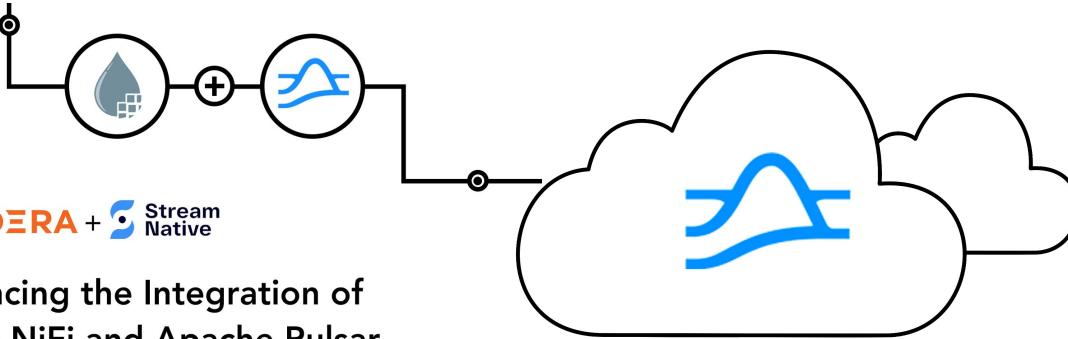
Enable easy ingestion, routing, management and delivery of any data anywhere (Edge, cloud, data center) to any downstream system with built in end-to-end security and provenance



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

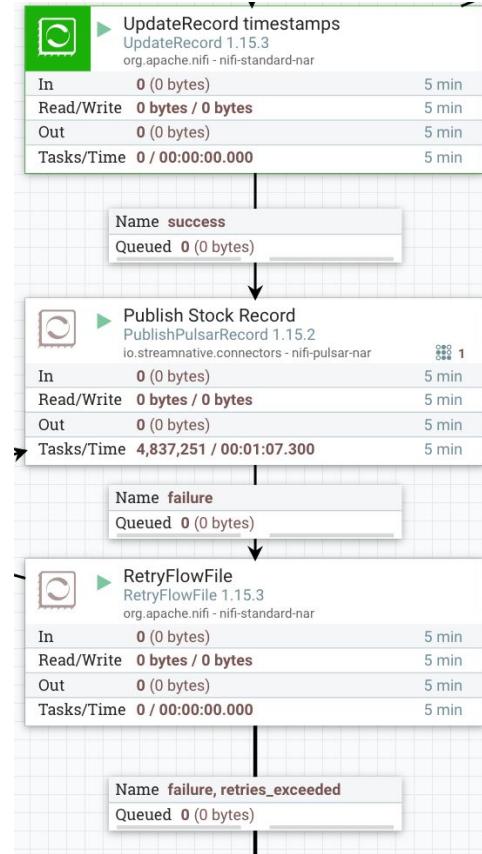
- Guaranteed Delivery
- Full data provenance from acquisition to delivery
- Diverse, Non-Traditional Sources
- Eco-system integration

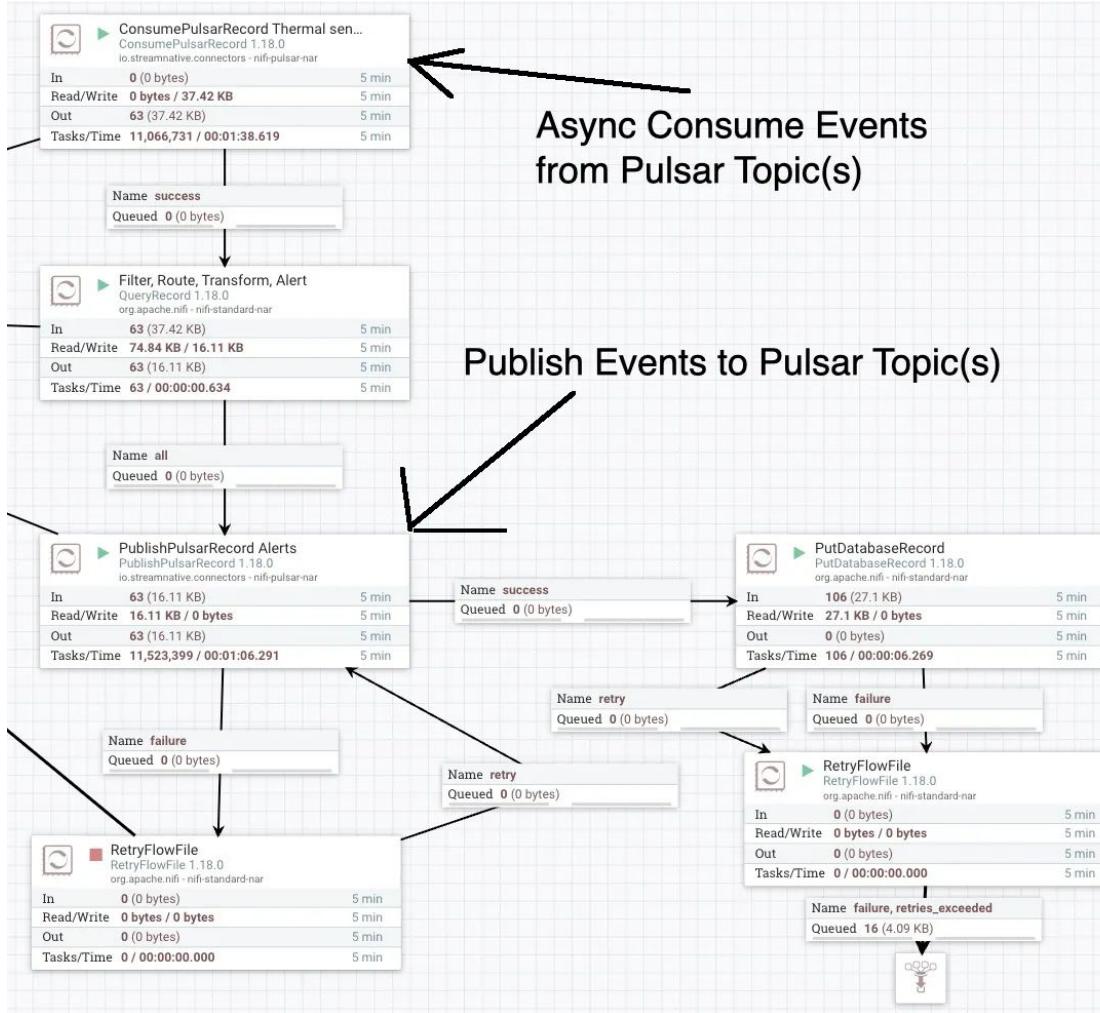
Apache NiFi Pulsar Connector



Announcing the Integration of
Apache NiFi and Apache Pulsar

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





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

The screenshot shows the Cloudera DataFlow interface. On the left is a dark sidebar with navigation links: Dashboard, Catalog (which is selected and highlighted in purple), and Environments. The main area is titled "Flow Catalog". It displays a list of flow definitions:

- Covid Data Stream
- CovidIDBroker
- drew_kafka-hdfs-querydb-kudu
- drew_kafka_to_hdfs
- Employees Data
- Empty Dev Flow
- Generate Flow File Log

To the right of the list, a specific flow is detailed:

Covid Data Stream
Updated 2 months ago by Michael Kohs

FLOW DESCRIPTION
This flow reads covid data from several sources and writes it to CDP

Only show deployed versions

Version	Deployments
13	0
12	0
11	0

Deploy New Flow →

LAST UPDATE
2021-02-02 14:25 PST by Michael Kohs
"This version includes the latest fixes"

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

Cloudera Docs / DataFlow master ▾ (test • Technical Preview) Search Document

Cloudera DataFlow

Release Notes

Release Notes

Concepts

Overview

Planning

AWS Resource Planning

NiFi Flow Limitations

Getting Started

Quick Start

Out of Box Flow Definitions

Import a flow definition

Flow definition for ingesting data into a Kafka topic

Flow definition for ingesting data into Amazon S3 Buckets

How To: Environments

Enabling a DataFlow Environment

Managing a DataFlow Environment

OUT OF BOX FLOW DEFINITIONS

Flow definition for ingesting data into a Kafka topic

Example

The resulting flow will look similar to the following, on your NiFi canvas.

```
graph TD; A[ConsumeKafkaRecord_2.0] -- In: 0 bytes --> B[Name parse.failure  
Queued: 0 (0 bytes)]; A -- Out: 0 bytes --> C[Log Event with Incorrect Schema]; B --> D[Name success  
Queued: 0 (0 bytes)]; D --> E[Filter Events]; E --> F[Name filtered.events  
Queued: 0 (0 bytes)]; F --> G[Kafka_JSON_Sink]
```

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

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	2

Target Environment

aws dataflow-demo	60% (3 of 5)
-------------------	--------------

Configure Sizing & Scaling

- ✓ Overview
- ✓ Flow Parameters
- ✓ Sizing & Scaling
- ⓘ Key Performance Indicators
- ⓘ Review

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	4 vCores Per Node 8 GB Per Node	8 vCores Per Node 16 GB Per Node	16 vCores Per Node 32 GB Per Node

Number of NiFi Nodes

Auto Scaling ⓘ
 Enabled

Min. Nodes: Max. Nodes:

Provide Parameters

Flow Parameters

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.

MachineData

AWS Credential File

Enter parameter values.



Select File

Drop file or browse

CDP Truststore

Enter parameter values.



Select File

Drop file or browse

CDPSchemaRegistry

https://dataflow-streams-master0.dataflow.xcu2-8y8x.dev.cdr.work:7790/api/v1

Define KPIs

Key Performance Indicators

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

Entire Flow

METRIC TO TRACK

Data In
ALERT SET
Notify if less than 150 KB/sec, for at least 30 seconds.



Processor: Write to S3 using HDFS proc

METRIC TO TRACK

Bytes Sent
ALERT SET
No alert set



ⓘ Add New KPI

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

Key Performance Indicators

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

Entire Flow

METRIC TO TRACK
Data In

ALERT SET
Notify if less than 150 KB/sec, for at least 30 seconds.

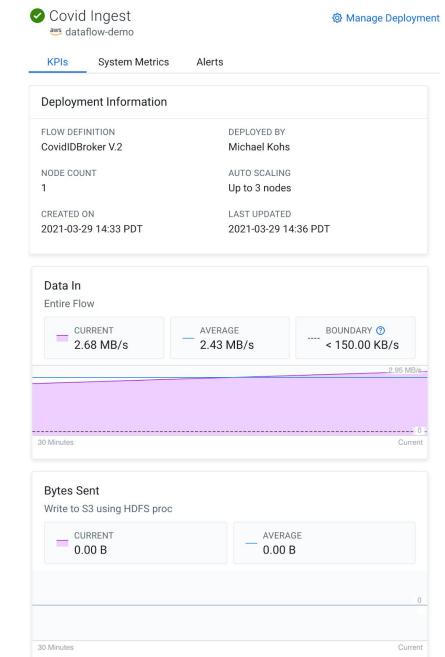
Processor: Write to S3 using HDFS proc

METRIC TO TRACK
Bytes Sent

ALERT SET
No alert set

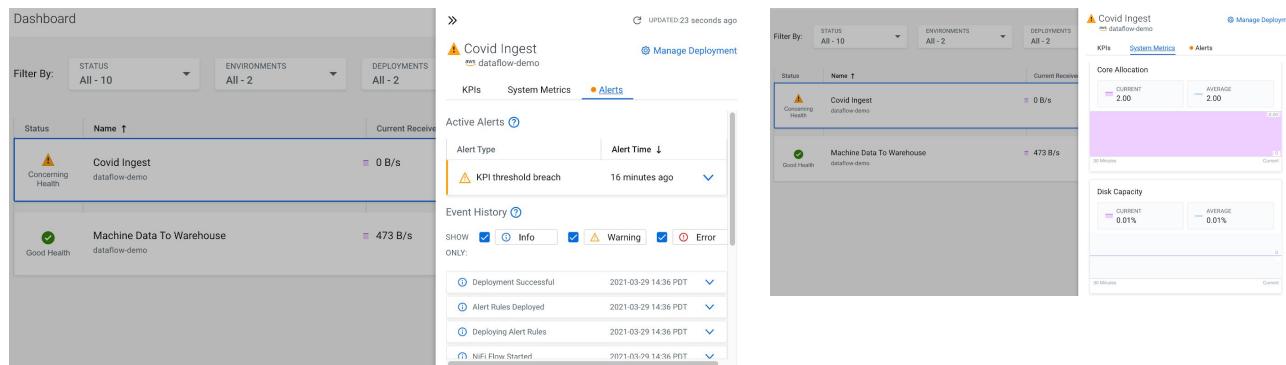
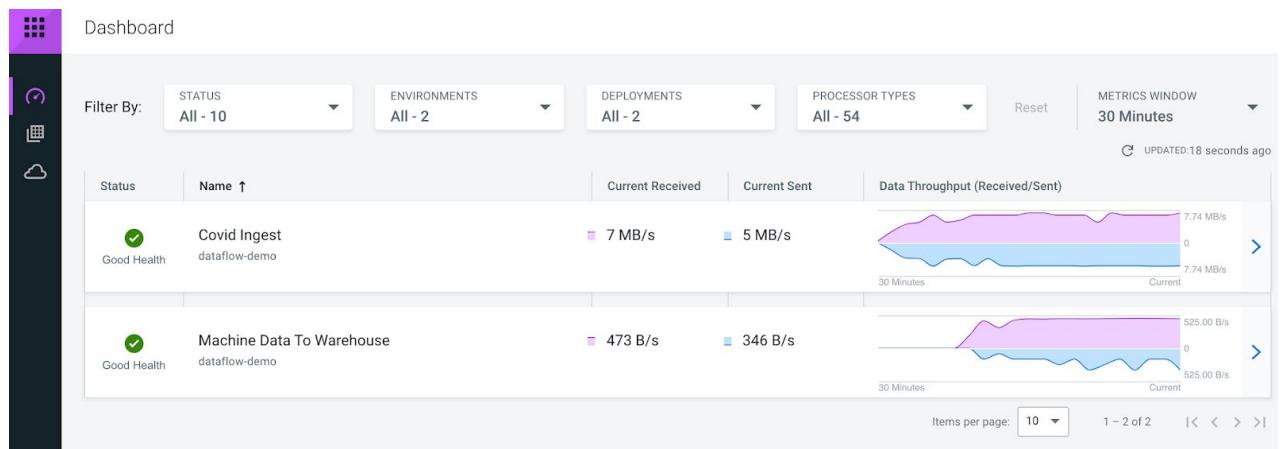
+ Add New KPI

KPI Monitoring



Dashboard

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



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

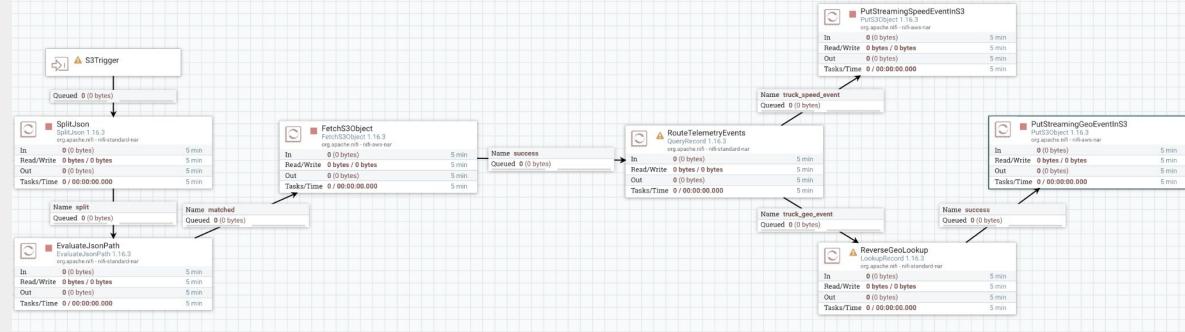
The screenshot shows the Cloudera Data Flow interface. On the left is a dark sidebar with the Cloudera logo and navigation links: Dashboard, Catalog, ReadyFlow Gallery, Flow Design (which is selected and highlighted in purple), Functions, Environments, Get Started, Help, and Stephen Hawking. At the bottom of the sidebar, it says v2.0.0. To the right is the main workspace titled "Flow Design / [WorkspaceName] / [FlowDefinitionName]". The workspace contains a visual canvas with a single component icon (a square with a triangle) and a configuration panel. The configuration panel is titled "[ProcessorName] [ProcessorType] [Version#]" and displays the following metrics:

IN	19 (14.16 MB)
READ/WRITE	4.88 MB/4.88 MB
OUT	0 (0 bytes)
TASKS	29/00:00:00.123

A timestamp at the bottom right of the panel says "5 min". Below the configuration panel is a text input field with the placeholder "[RootProcessGroupName]". On the far right, there is a sidebar titled "Configuration" which lists various configuration options with their current values and dropdown menus for modification. Buttons at the bottom right of the sidebar allow users to "Apply Changes" or "Discard Changes".

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



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.

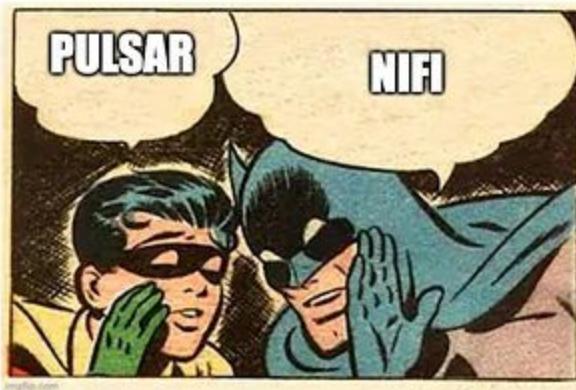
Serverless Customized Triggers

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

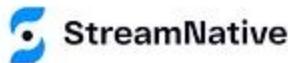
Example: Near real time offloading of files from a remote SFTP server.

NiFi → Pulsar

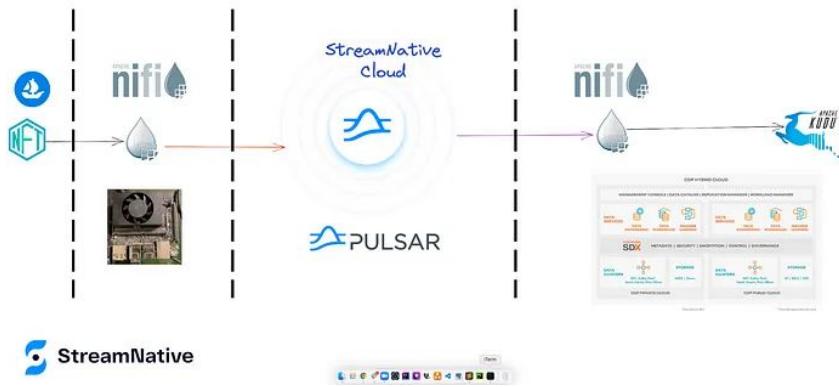
Apache NiFi Pulsar Connector



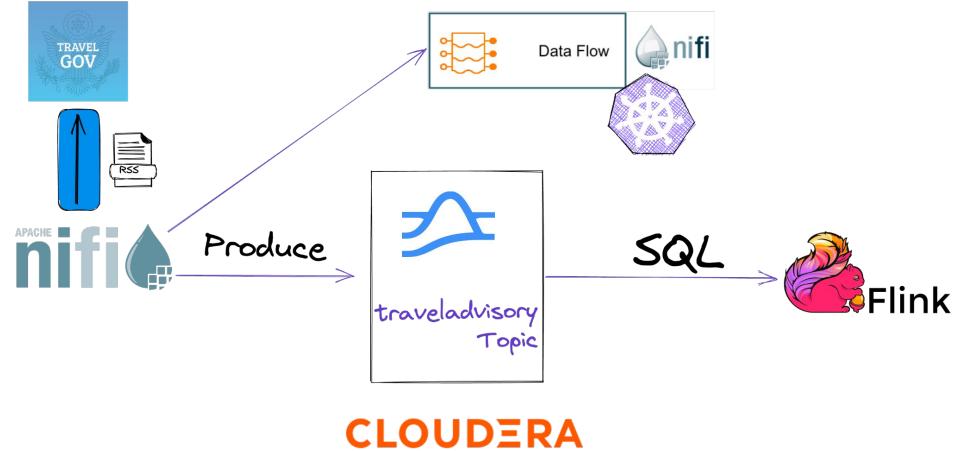
<https://github.com/streamnative/pulsar-nifi-bundle>



Streaming NFTs



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



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

- [!\[\]\(0cf1de0a013c12b724e3acde3b2f780a_img.jpg\) Dashboard](#)
- [!\[\]\(1c108ad8fe9e61d7178ebe04ab721b96_img.jpg\) Catalog](#)
- [!\[\]\(7dcf48fddd351e3ddacf30e5bc93fcd2_img.jpg\) ReadyFlow Gallery](#)
- [!\[\]\(27098a06ea2a054d76bd80b8907db857_img.jpg\) Functions](#)
- [!\[\]\(9e2a99a7e5be25d92c34927621c21f02_img.jpg\) Environments](#)



Add Processor

ConsumePulsar >

ConsumePulsarRecord

PublishPulsar

PublishPulsarRecord

Processor Name

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

X

Cancel

Add



>

 HTTP In to Kafka for Labs TSPANN
Process Group

[More Details ▾](#)

Settings

Process Group Name

Flowfile Concurrency

Outbound Policy

Default Flowfile Expiration

Default Back Pressure Object Threshold

Default Back Pressure Data Size Threshold

Comments

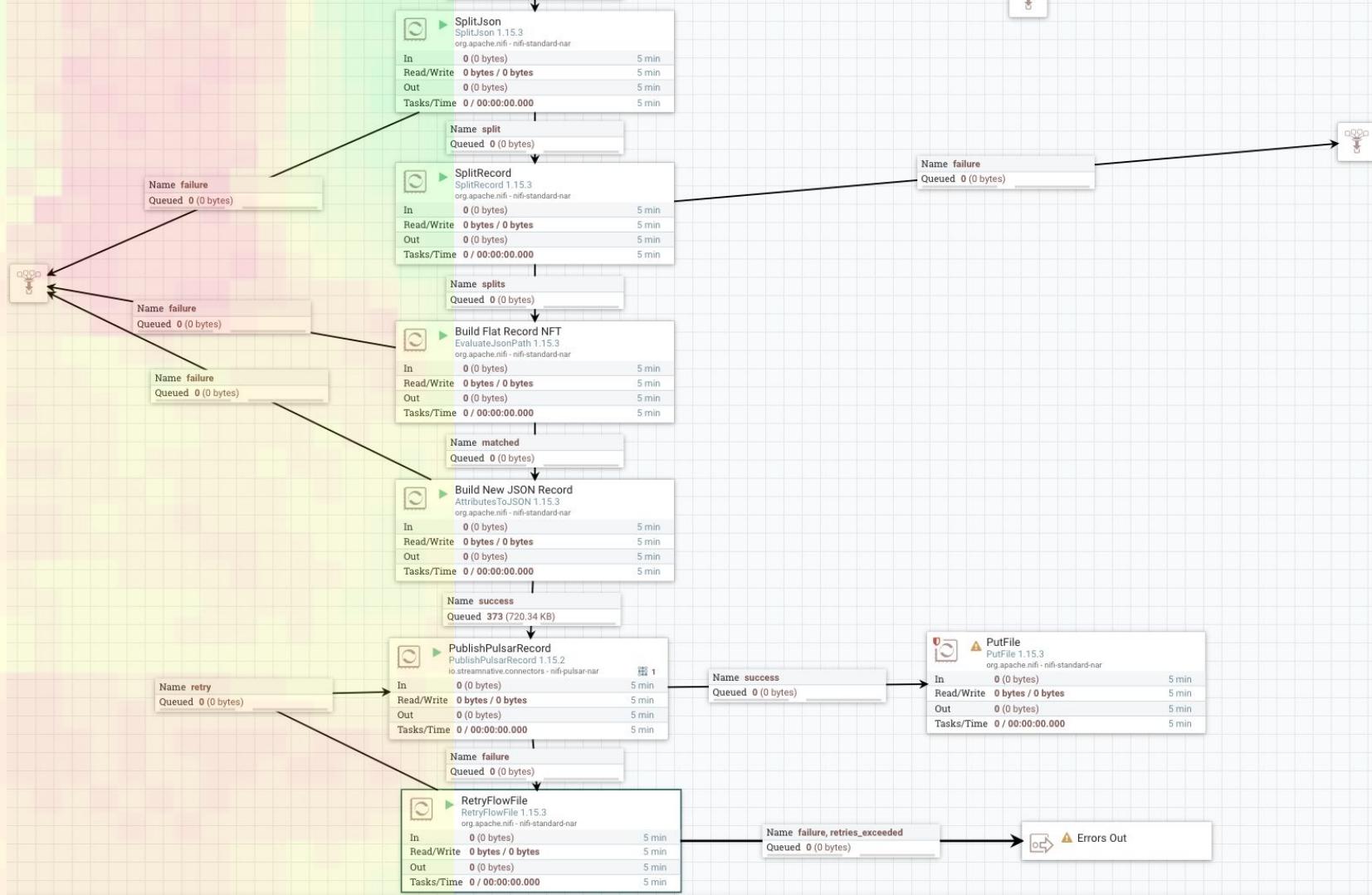
→ Get Started

① Help

Timothy Spann

2.4.0-h1-b1

HTTP In to Kafka for Labs TSPANN



Questions

Pulsar Virtual Summit
Europe 2023

Your name
Contact information

Thank you!



Hosted by  StreamNative