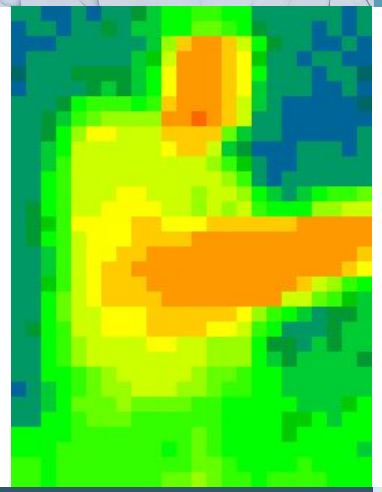
2023 Real-Time Stream Processing unconference organized by Hazelcast

Real-Time Streaming Processing Round Table

Timothy Spann
Principal Developer Advocate

CLOUDERA

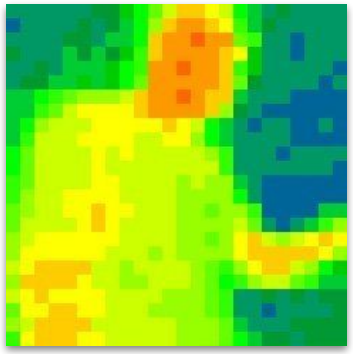


CLOUDERA

EDGE
2AI
CLOUDERA

ENTERPRISE
DATA
CLOUD
CLOUDERA

FLiPN-FLaNK Stack



Tim Spann
@PaasDev // Blog: **www.datainmotion.dev**
Principal Developer Advocate. Princeton Future of Data Meetup.
https://github.com/tspannhw/EverythingApacheNiFi
https://medium.com/@tspann
Apache NiFi x Apache Kafka x Apache Flink

Round Table Discussion

- Current trends of real-time stream processing in 2023
- Current challenges of real-time stream processing and proposed solutions
- Performance / Benchmarking real-time stream processing
- The future of real-time stream processing

Round Table Discussion

Current trends of real-time stream processing in 2023

- Lightweight serverless
- Hazelcast
- SQL
- Flink
- Kafka or Pulsar as Messaging Hub
- Java 17+
- Managed Clusters, Containers and Environments
- Real-Time Analytics
- Fast Storage Options

Round Table Discussion

Current challenges of real-time stream processing and proposed solutions

- Deployment, Automation and Scaling
- Choosing right project/sizing for use case
  - Simple Event Processing vs Complex Event Processing
  - Leveraging existing applications
- Developer Skills
- Self management and monitoring
- Cost issues -> autoscaling, optimizing, performance, hybrid deployment

Round Table Discussion

Performance / Benchmarking real-time stream processing

- Kafka/Pulsar: https://openmessaging.cloud/docs/benchmarks/
- NiFi: https://blog.cloudera.com/benchmarking-nifi-performance-and-scalability/
- Flink: https://github.com/ververica/flink-sql-benchmark
- Hazelcast: https://hazelcast.com/press-release/hazelcast-demonstrates-cloud-efficiency-real-time-stream-processing-of-one-billion-events-per-second/

Round Table Discussion

The future of real-time stream processing

- WASM (Web Assembly)
- Petabyte, 5000 Node Clusters, Smart Hyper Scaling
- Multi-language support (Python, Rust, Kotlin, Golang, Carbon, JVM)
- Machine Learning, Deep Learning, AI and Advanced Math
- Low Code Development like Apache NiFi, DataFlow Designer, SQL
- Dynamic Hybrid Deployment
- Citizen Stream Engineer
- Edge Streaming and Hybrid Edge Streaming
- Java 20, 21; Java Loom Virtual Threading
- Ultra low latency, trillions of events per second, massive RAM/network

Round Table Resources

- https://dzone.com/articles/real-time-stream-processing-with-hazelcast-and-streamnative
- https://flipstackweekly.com/
- https://www.datainmotion.dev/
- https://www.flankstack.dev/
- https://github.com/tspannhw
- https://medium.com/@tspann
- https://medium.com/@tspann/predictions-for-streaming-in-2023-ad4d7395d714
- https://www.apachecon.com/acna2022/slides/04_Spann_Tim_Citizen_Streaming_Engineer.pdf


Tim
SPANN
https://github.com/tspannhw
https://www.datainmotion.dev/

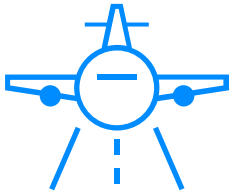
CONNECTED DEVICES ARE EVERYWHERE

EDGE

Miles driven

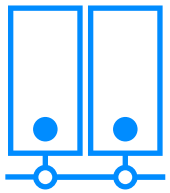
Wear-out tires

Wearing of the doors

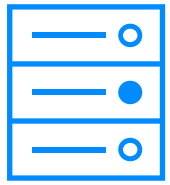


Engine wear

DATA CENTER



Data capacity



Compute speed

Rising temperature

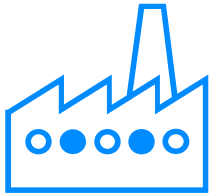
Capture data from all these sources and scale with a data streaming platform in a hybrid architecture

TODAY'S NEEDS FOR DATA STREAMING

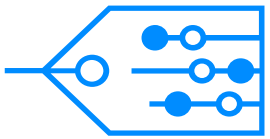
Gain Competitive Advantage

"Many leading enterprises realize that real-time analytics — the analytics of the present — is an incredible competitive advantage because they can act now to ***serve fickle customers, fix operational problems, power internet-of-things (IoT) apps, and respond decisively to competitors.***"

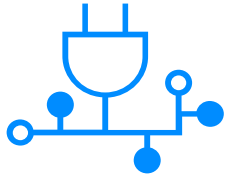
Forrester



Supply chain impacts manufacturing



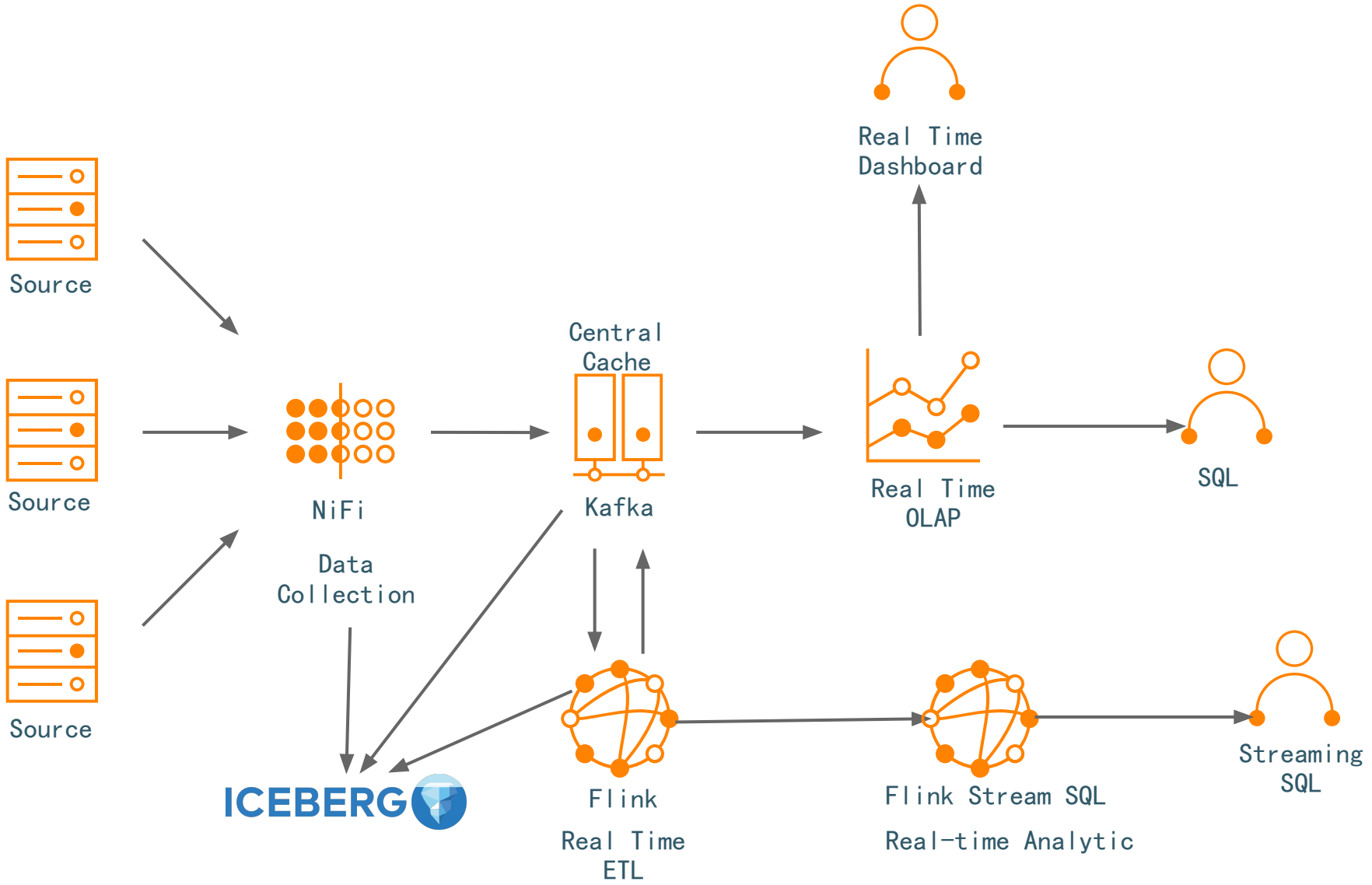
Predict customer buying pattern

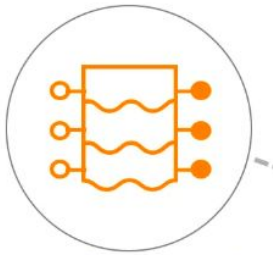


Utilities prevent power outage

Telecoms deliver continuous QoS

Reduce cyber threats





Data-in-motion

High-velocity
streaming data in **Kafka**

Data-at-rest

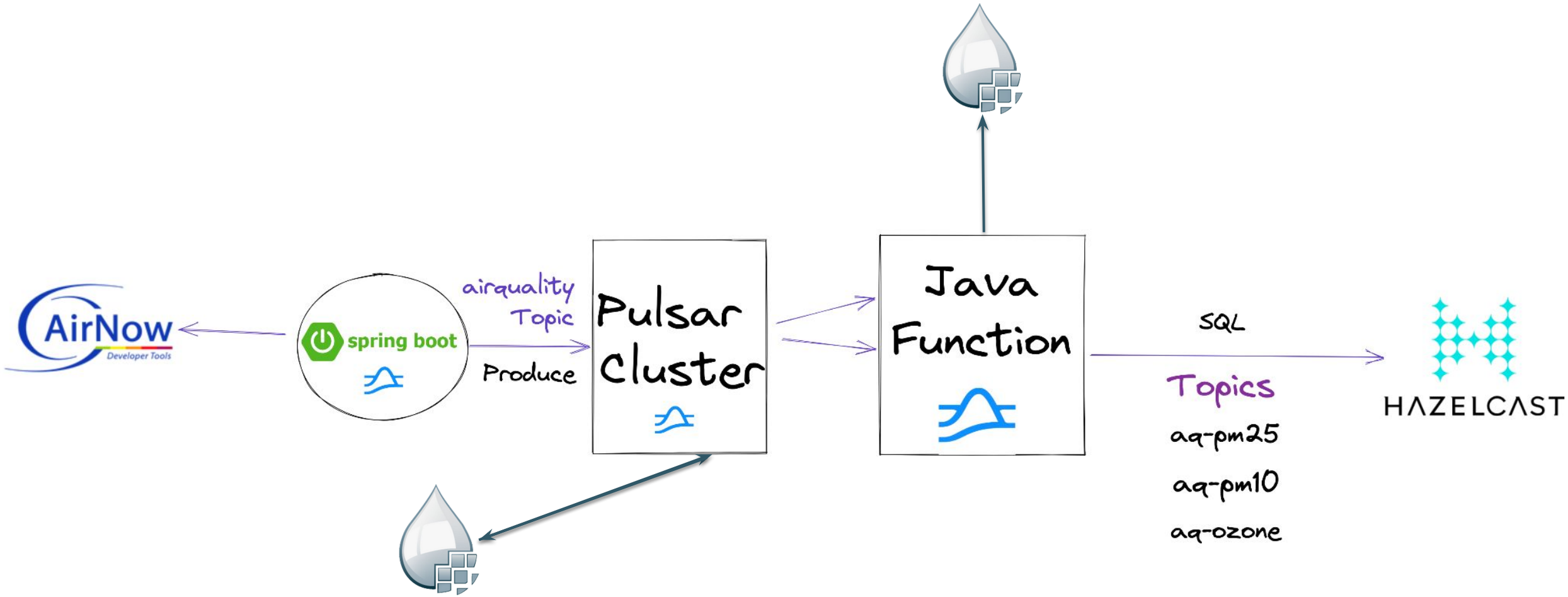
Low-velocity batch data in data warehouses, data lakes, lake houses, operational databases, etc.

Continuous
streaming SQL
(via **Flink**)

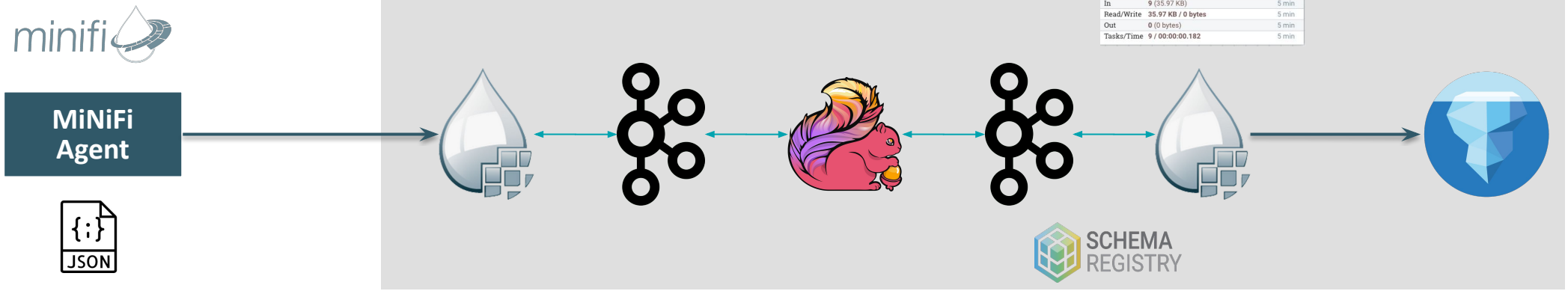
Materialized
views
(via **REST**)

Data products

dashboards, BI apps, microservices, timeseries, single page apps, notebooks, etc.

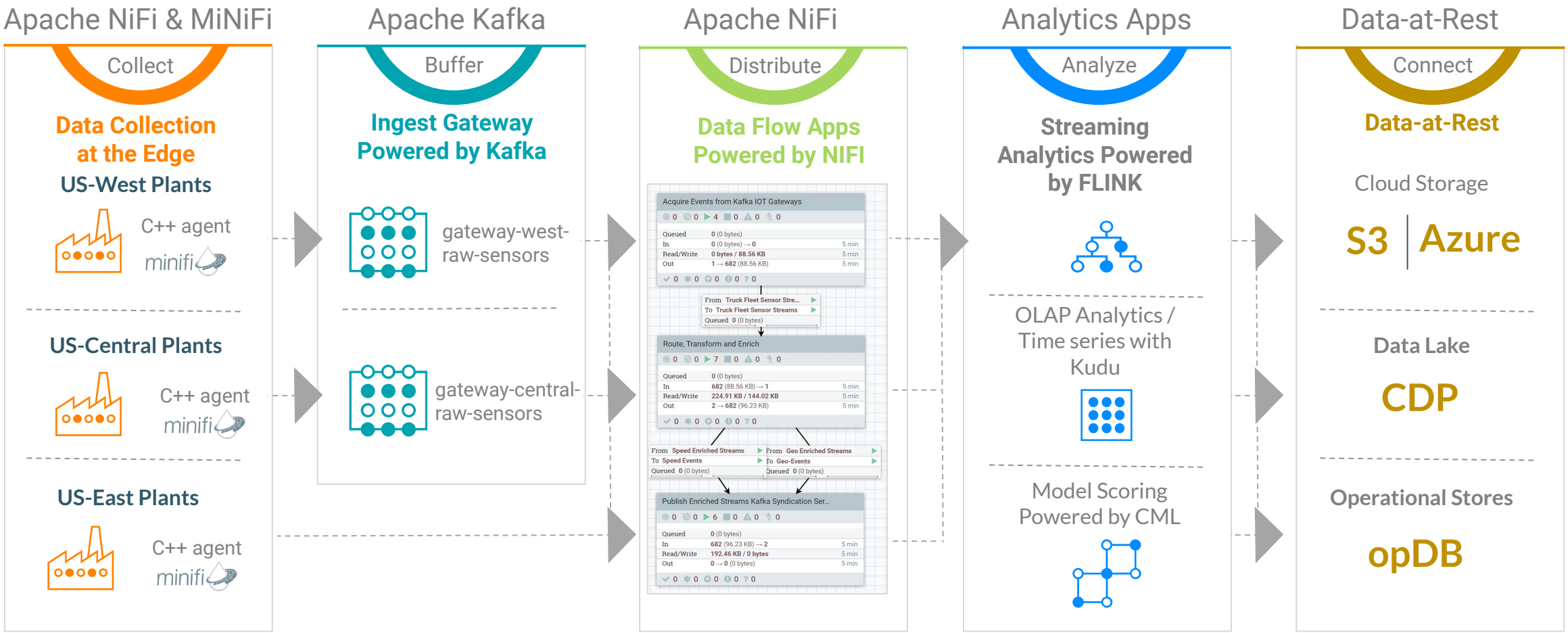

AirNow
Developer Tools
spring boot
airquality
Topic
Produce
Pulsar
Cluster
Java
Function
SQL
Topics
aq-pm25
aq-pm10
aq-ozone
HAZELCAST

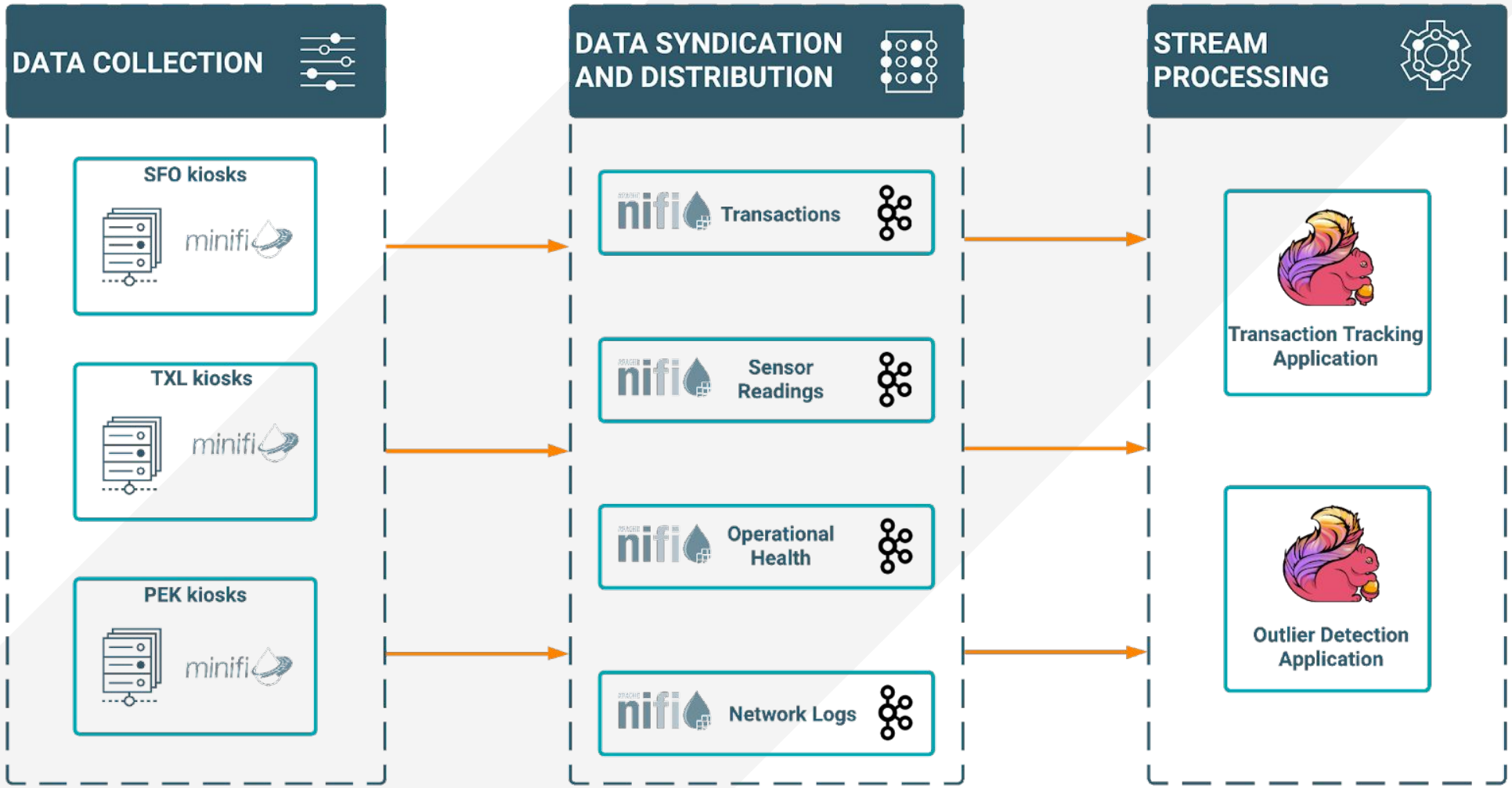
FLaNK Stack



https://flankstack.dev/

A DATA-IN-MOTION REFERENCE ARCHITECTURE

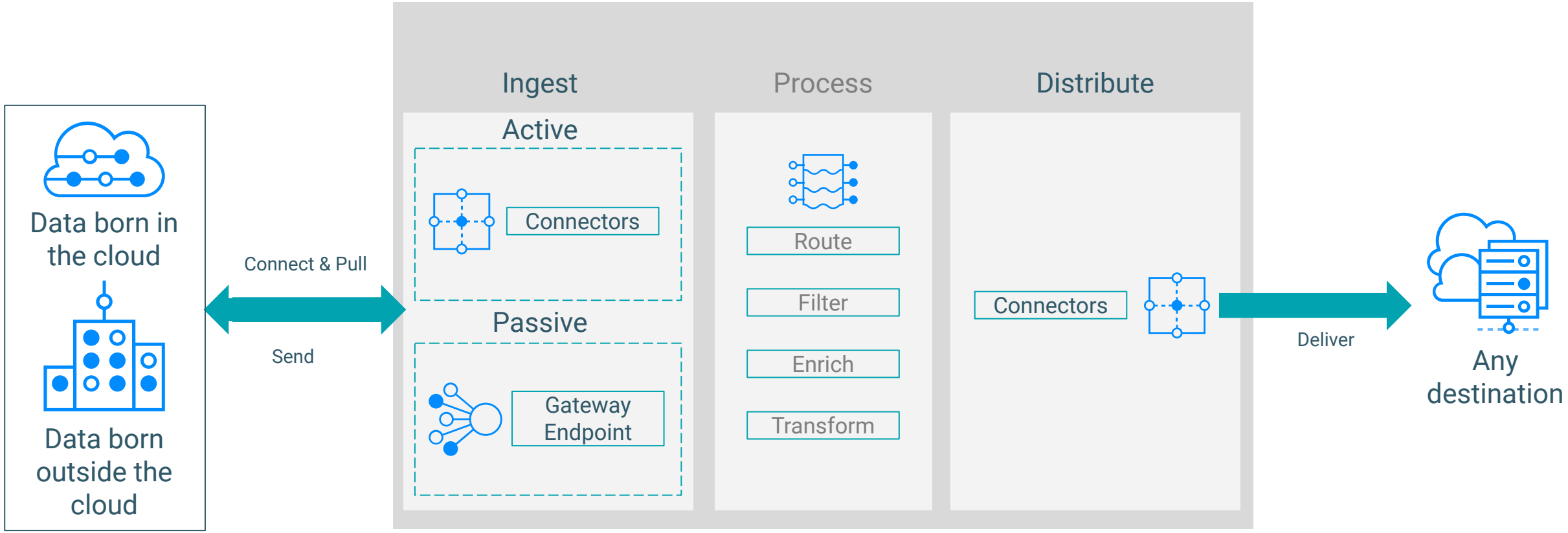



DATA COLLECTION
SFO kiosks
minifi
TXL kiosks
minifi
PEK kiosks
minifi
DATA SYNDICATION AND DISTRIBUTION
nifi
Transactions
nifi
Sensor Readings
nifi
Operational Health
nifi
Network Logs
STREAM PROCESSING
Transaction Tracking Application
Outlier Detection Application

CLOUDERA DATAFLOW DATA-IN-MOTION PLATFORM

CLOUDERA DATAFLOW

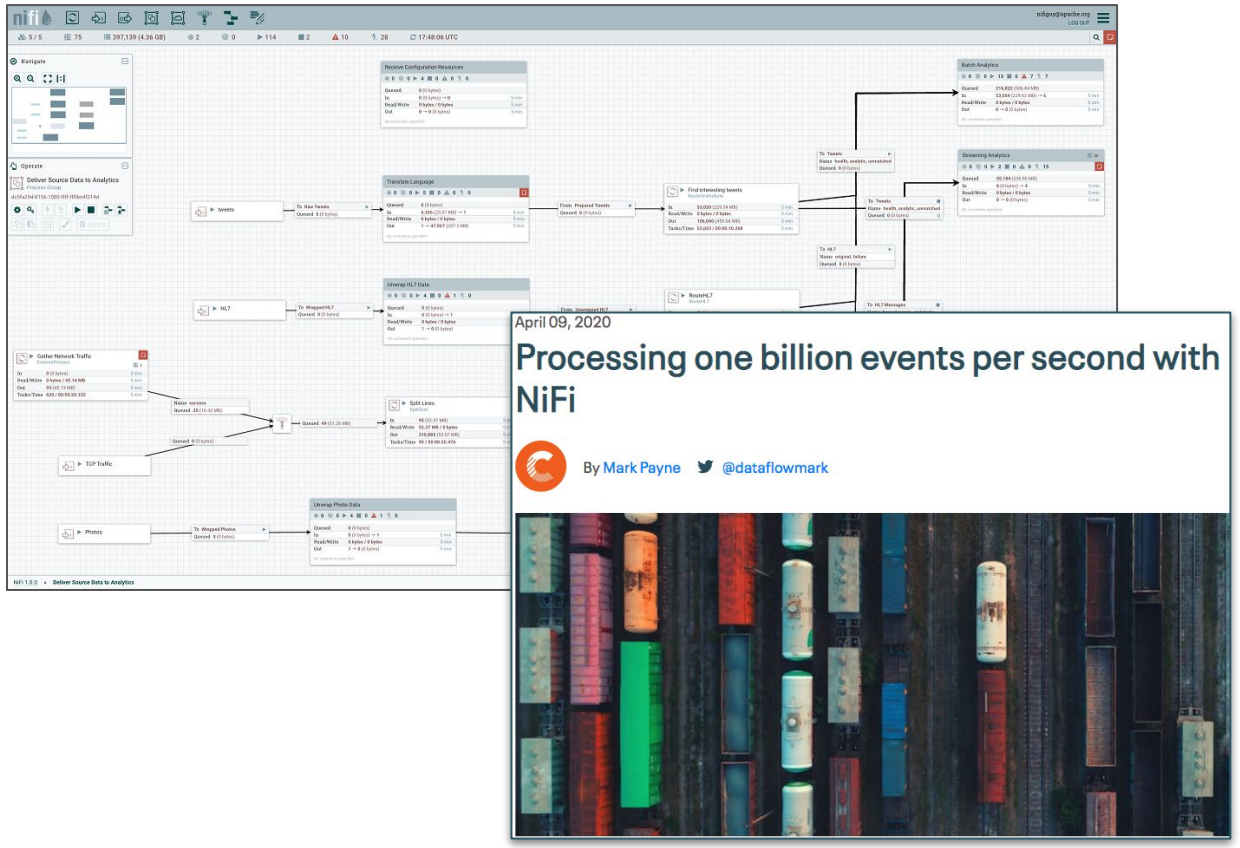
Solve the First Mile Data Collection Problem



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

CLOUDERA FLOW MANAGEMENT - POWERED BY APACHE NiFi

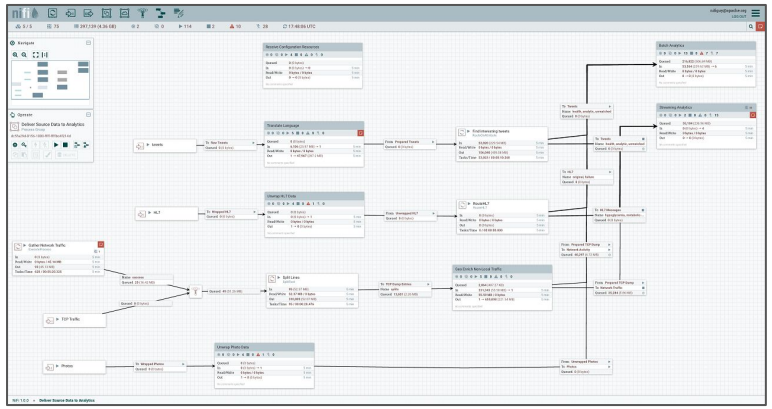
Ingest and manage data from edge-to-cloud using a no-code interface



- #1 data ingestion/movement engine
- Strong community
- Product maturity over 11 years
- Deploy on-premises or in the cloud
- Over 400+ pre-built processors
- Built-in data provenance
- Guaranteed delivery
- Throttling and Back pressure

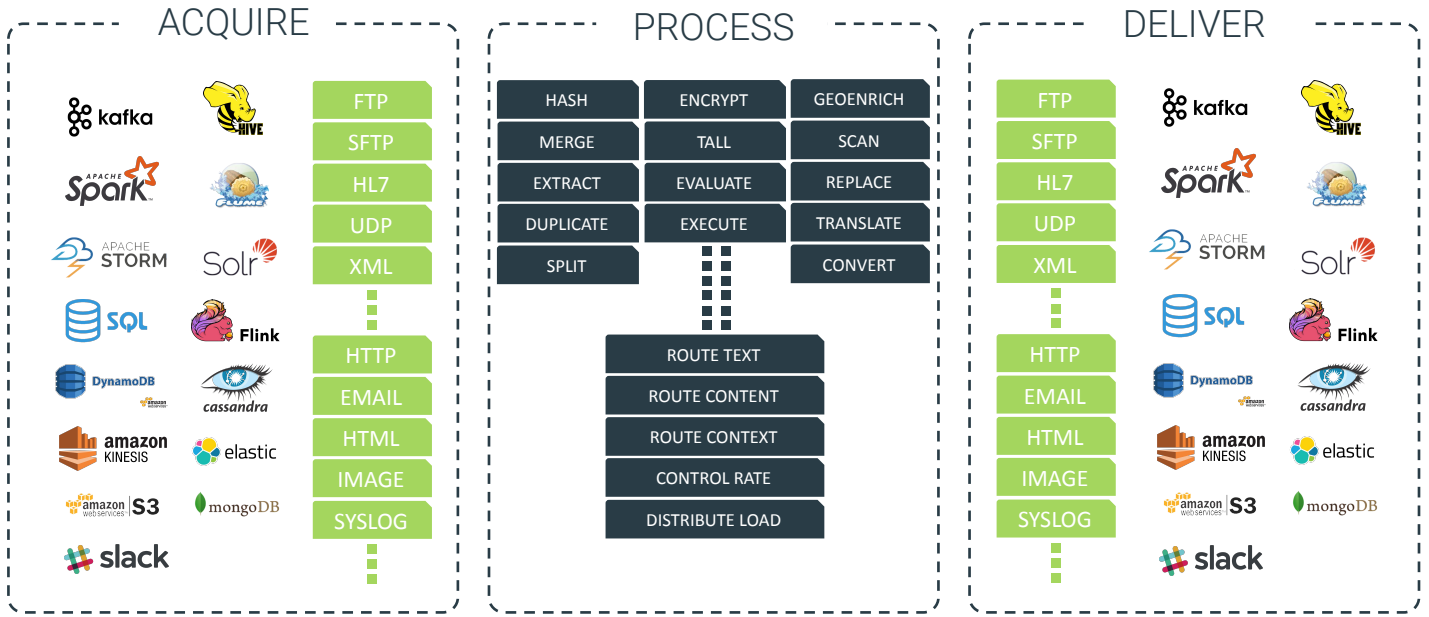
Cloudera Flow Management

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)



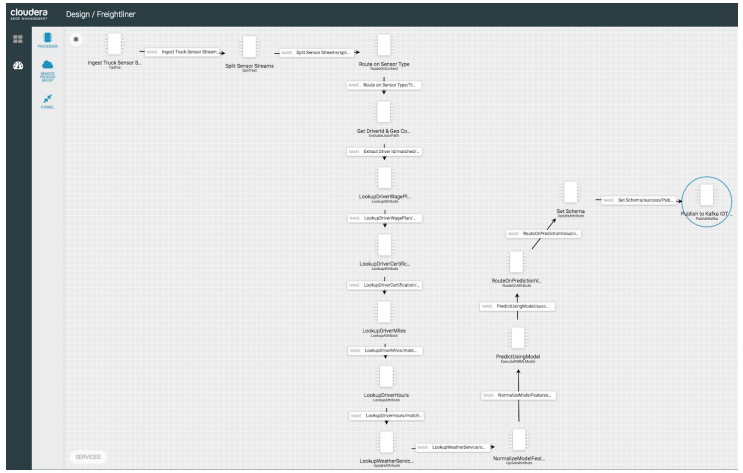
- Over 300 pre-built processors
- Easy to build your own processors
- Parse, enrich & apply schema
- Filter, Split, Merge & Route

- Throttle & Backpressure
- Guaranteed delivery
- Full data provenance
- Eco-system integration

Cloudera Edge Management

Manage, control, and monitor the edge for all your streaming and IoT initiatives

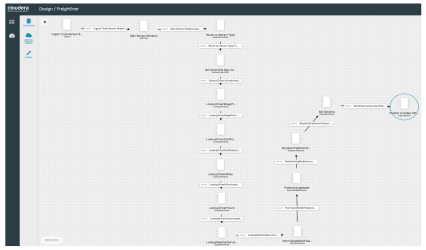
Edge Flow Manager

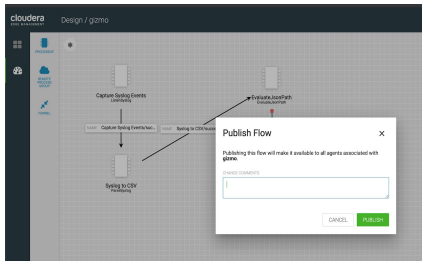


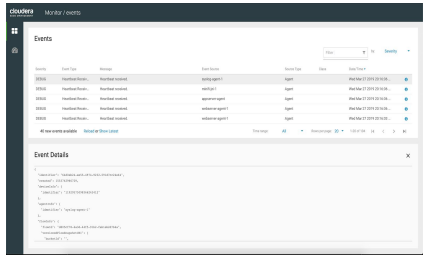
APACHE NIFI minifi

A lightweight edge agent that implements the core features of Apache NiFi, focusing on data collection and processing at the edge

Flow Authorship	Flow Deployment	Flow Monitoring







- Small footprint agent with MiNiFi
- Java and C++ agents
- Rich edge processors (edge collection & processing)
- End to end lineage and security

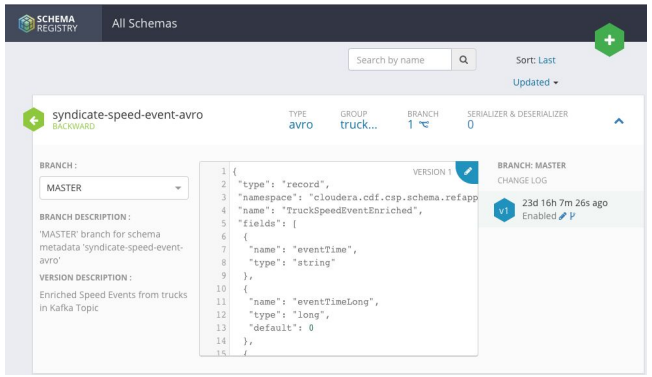
- Central Command and Control
- Design and deploy to thousands of agents
- Edge Applications lifecycle management
- Multitenancy with Agent classes
- Native integration with other CDF services

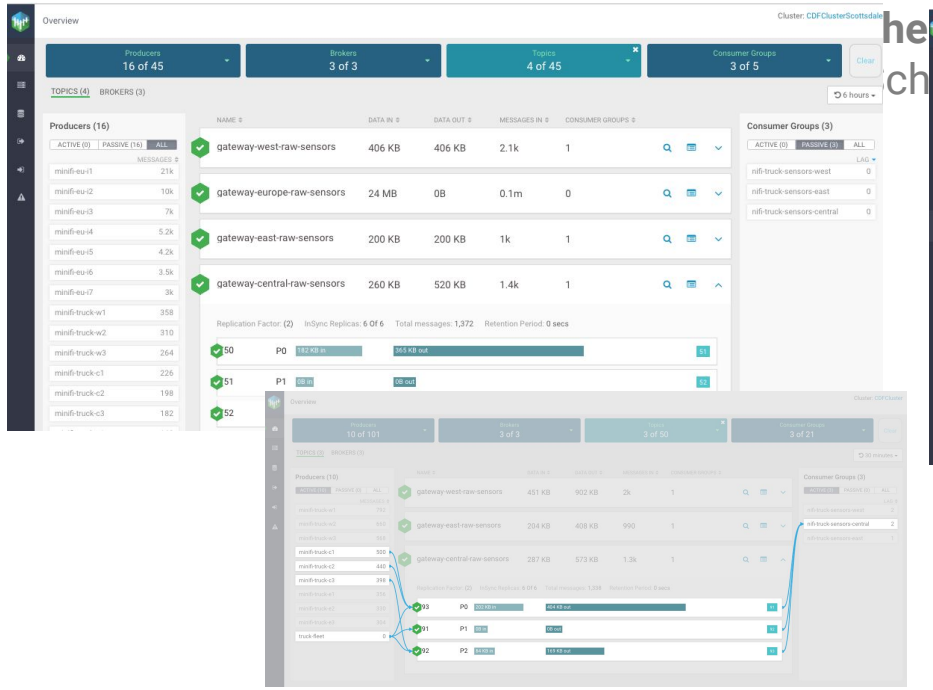
Comprehensive Streams Messaging

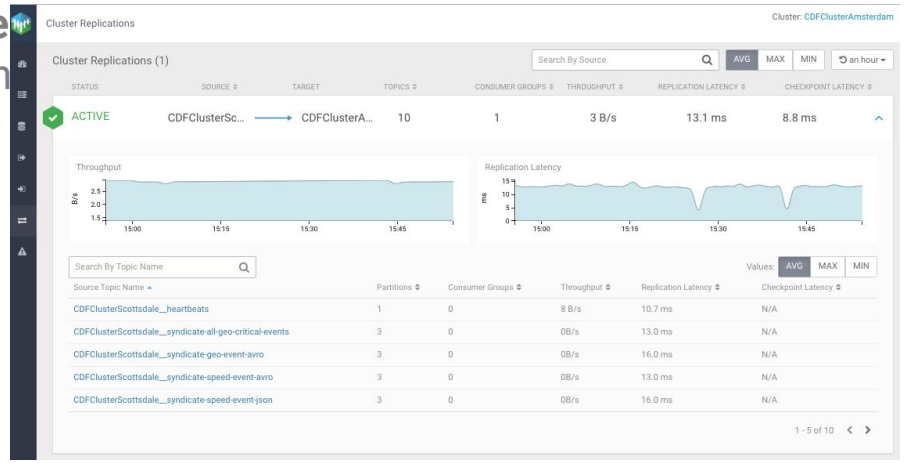
Extend streams messaging services for Schema Mgmt, Replication & Monitoring

Streams Messaging Manager
Management & Monitoring Service
for all of your Kafka clusters

Streams Replication Manager
Kafka Replication Service powered by
MirrorMaker2

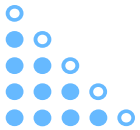






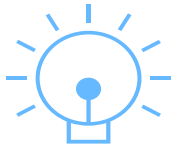
Next Generation Stream Processing & Analytics

Low latency stateful stream processing



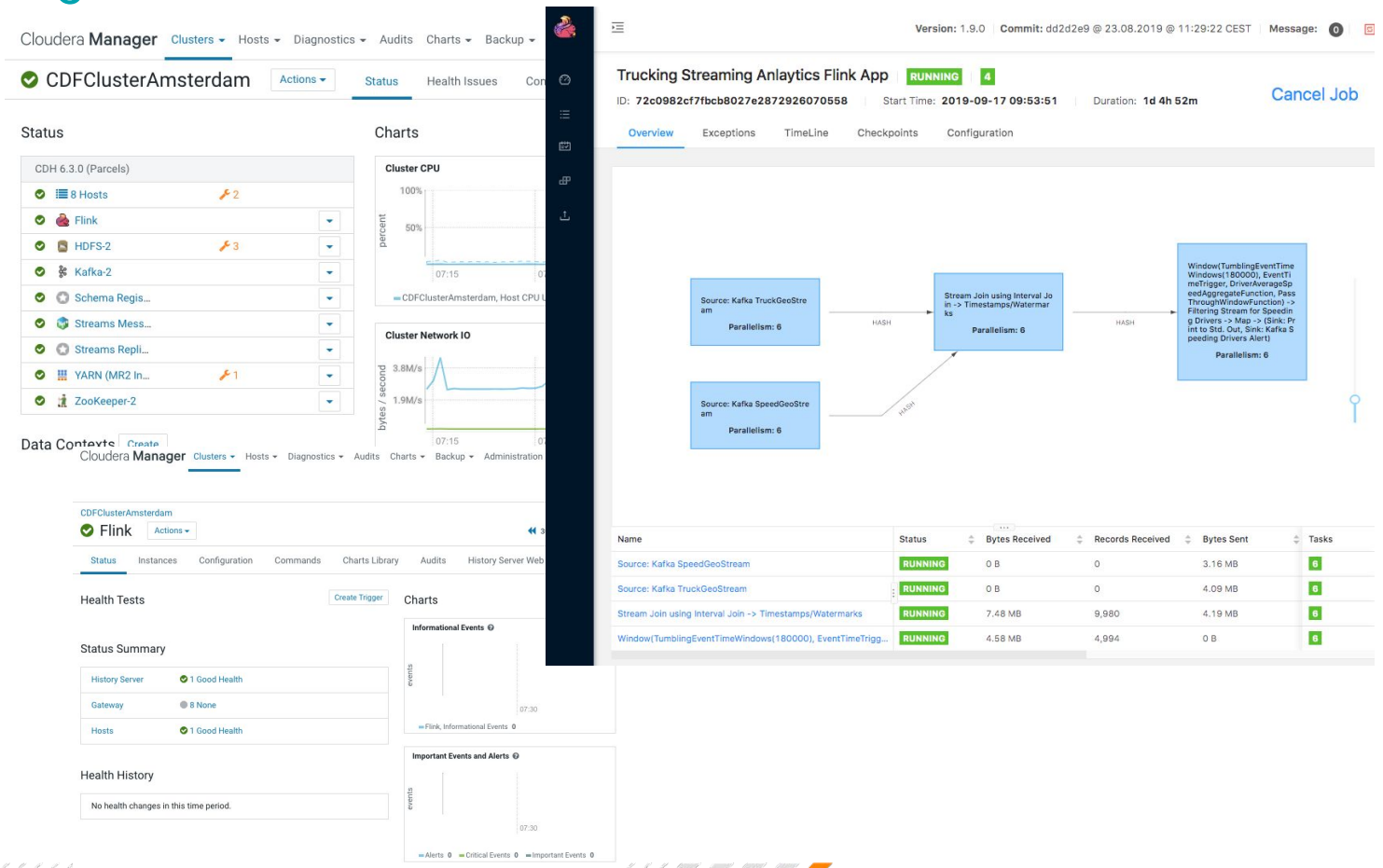
Low Latency

Event Processing



Real-Time Insights

- Advanced features - late arriving data, checkpointing, event time processing, Exactly Once Processing
- We support Apache Flink along with other stream processing engines like Kafka Streams and Spark Structured Streaming.



KEY DIFFERENTIATORS

Stream to Cloud – Extend the same on-premises streaming capabilities to the cloud with full support for multi-cloud and hybrid cloud models

400+ pre-built processors – Only product to offer such comprehensive connectivity to a wide range of data sources from edge to cloud

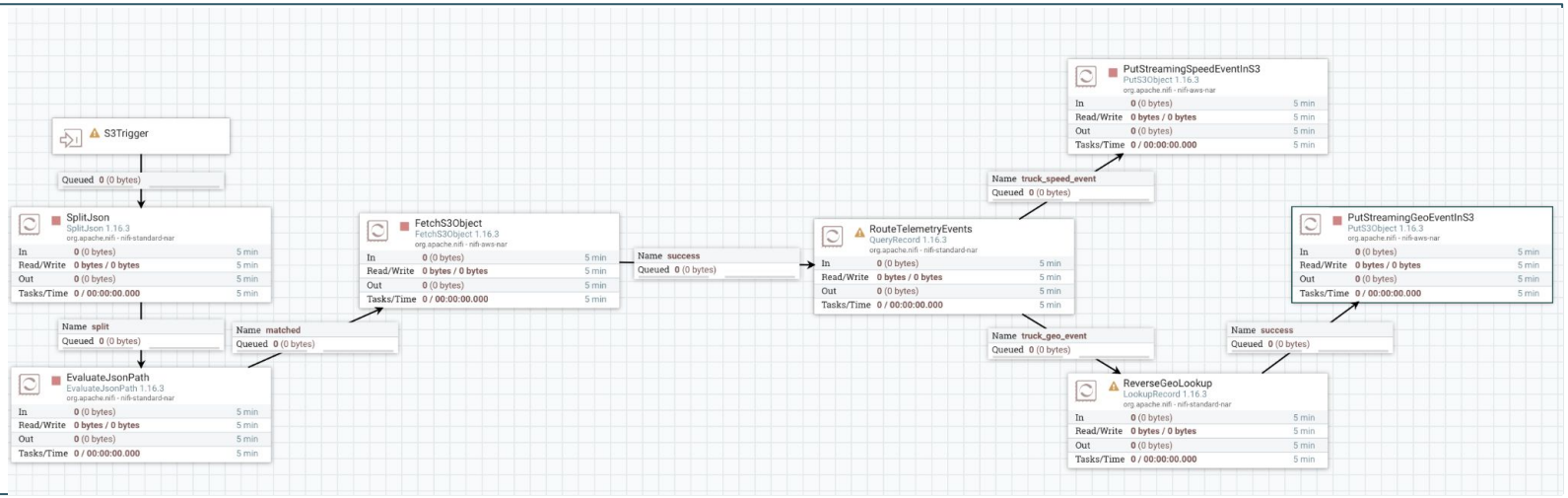
Democratize access to real-time data – Enable data analysts and other personas to quickly build streaming applications with just SQL

Enterprise-Grade Security & Governance – Deploy your streaming applications with confidence and trust with Cloudera SDX offering unified security and governance across the entire platform

Comprehensive streaming platform – Only vendor to offer a open and comprehensive streaming platform for real-time data ingestion and processing to produce prescriptive and predictive analytics

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.

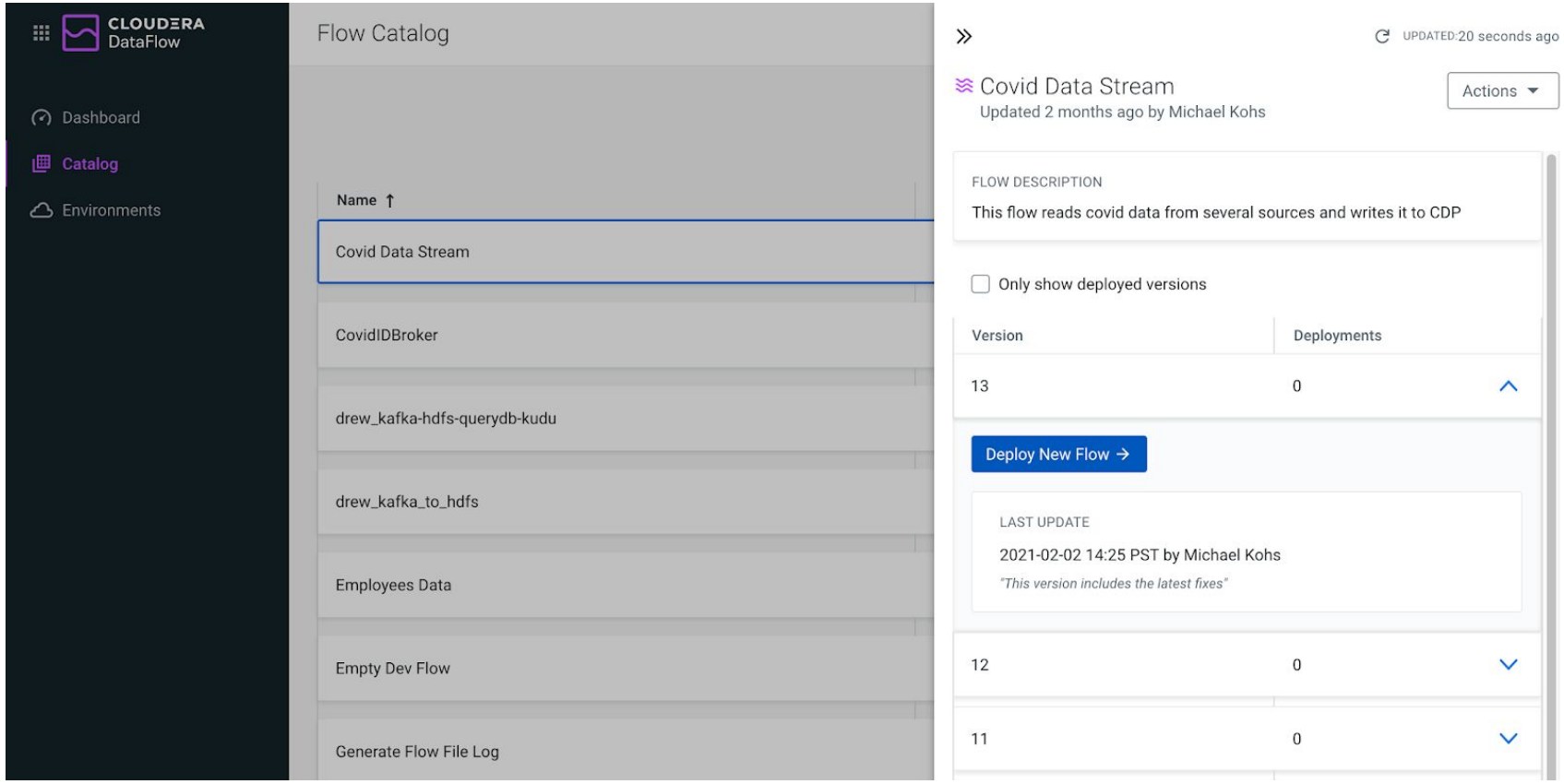
Serverless Customized Triggers

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.

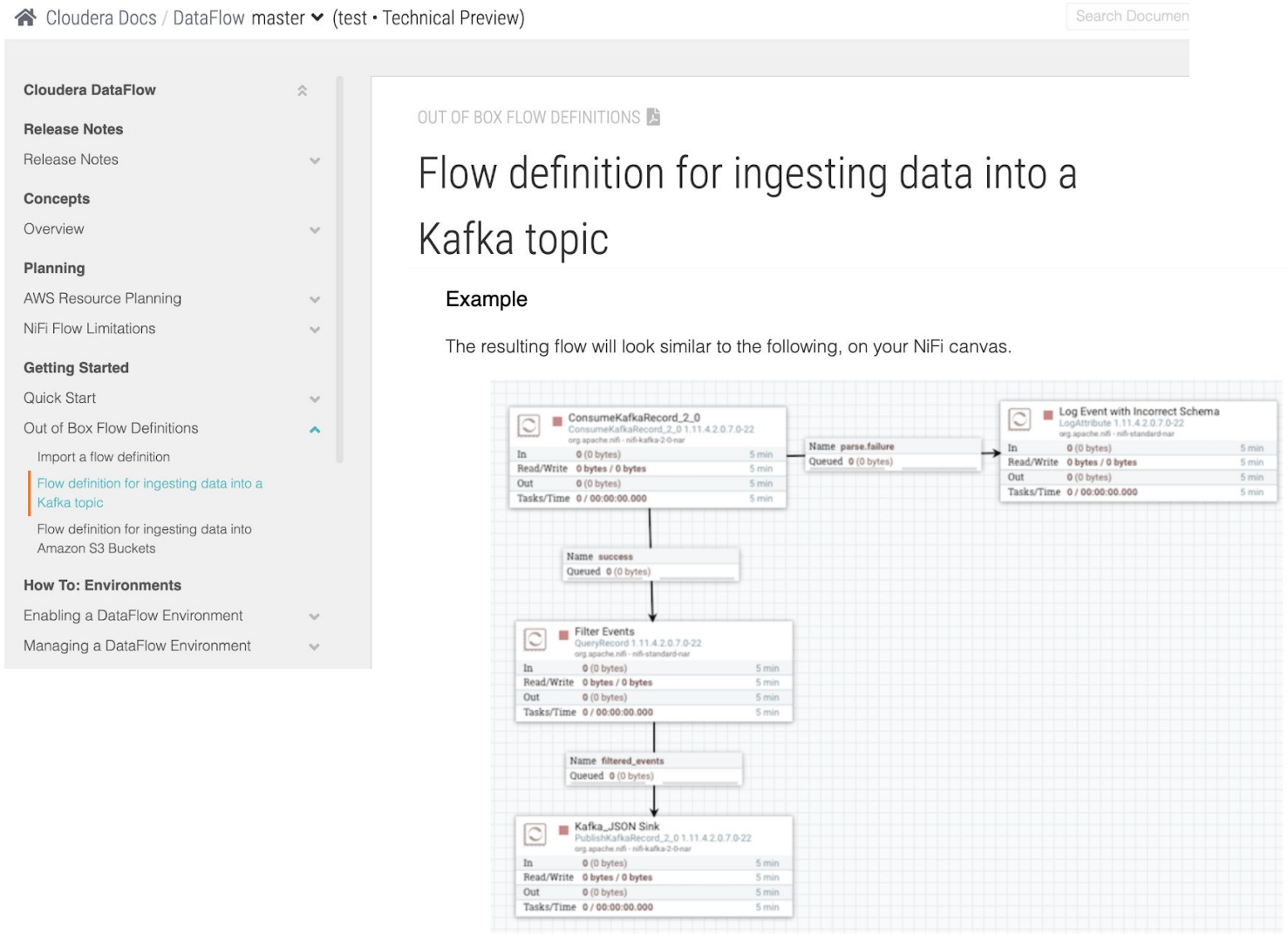
Flow Catalog

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



ReadyFlows

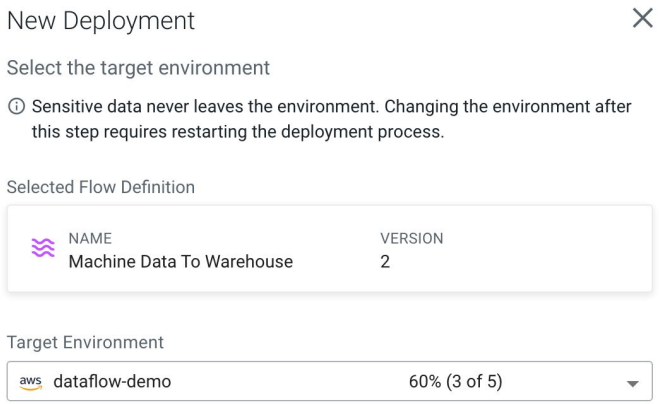
- Cloudera provided flow definitions
- Cover most common data flow use cases
- Can be deployed and adjusted as needed
- Made available through docs during Tech Preview



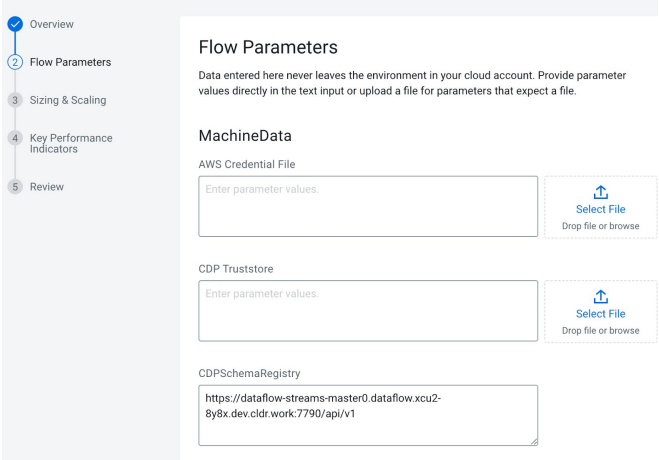
Deployment Wizard

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

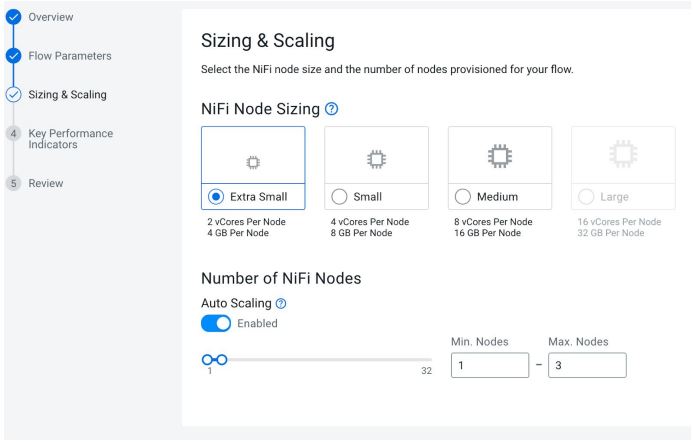
Start Deployment Wizard



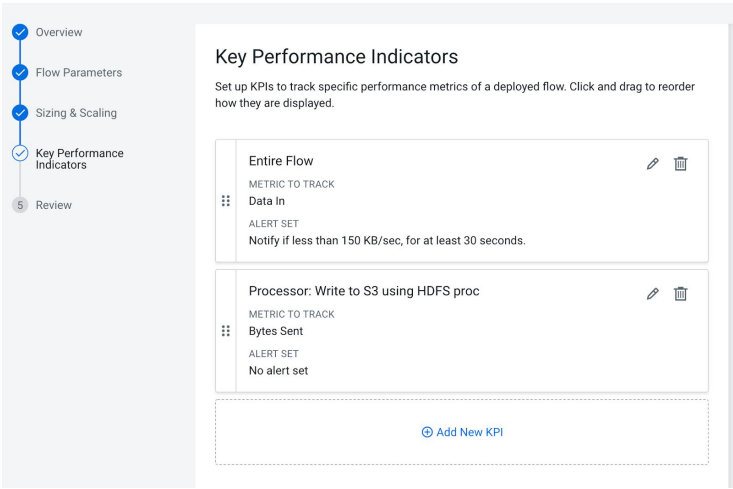
Provide Parameters



Configure Sizing & Scaling



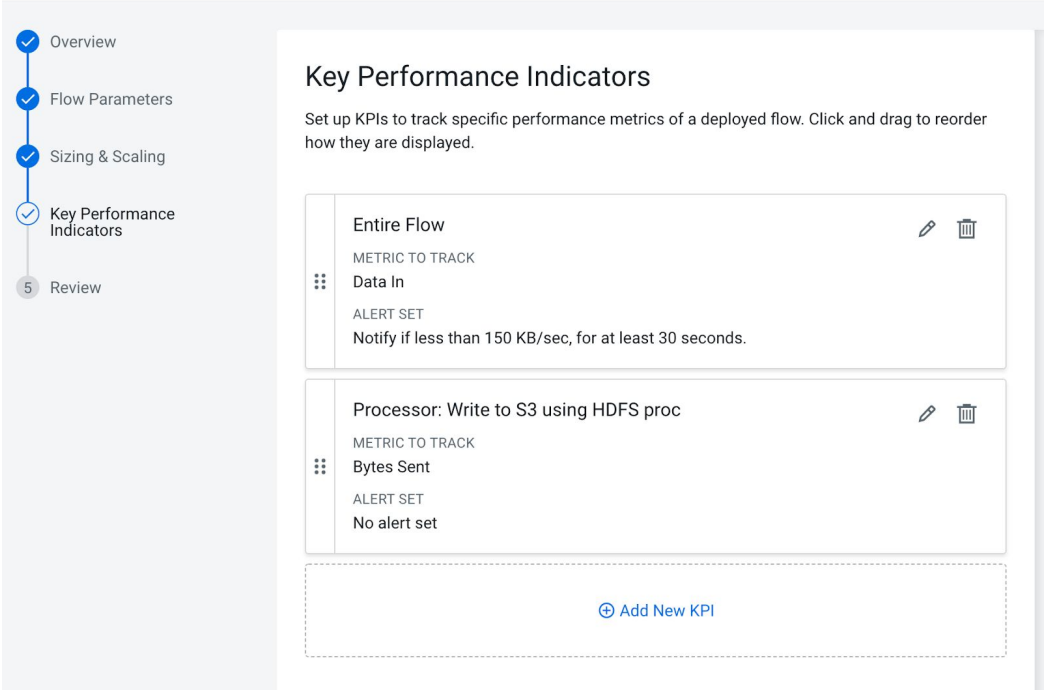
Define KPIs



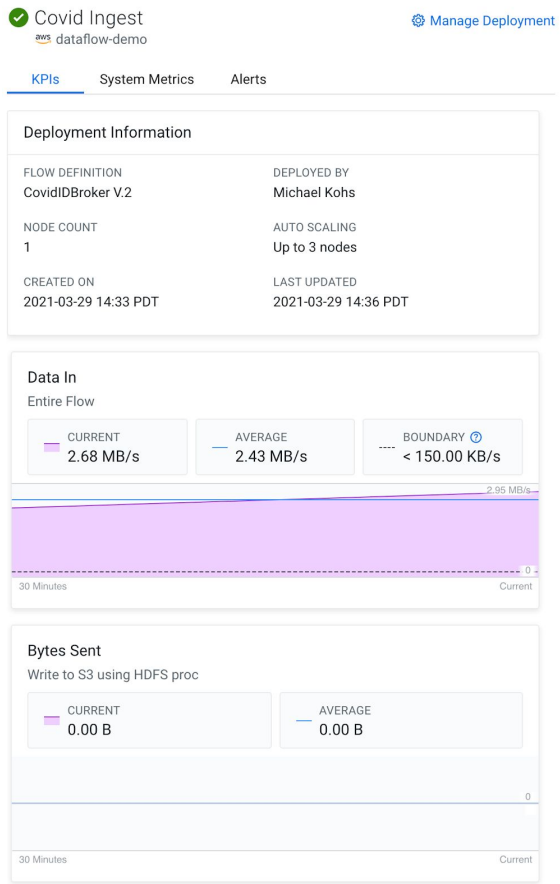
Key Performance Indicators

- 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

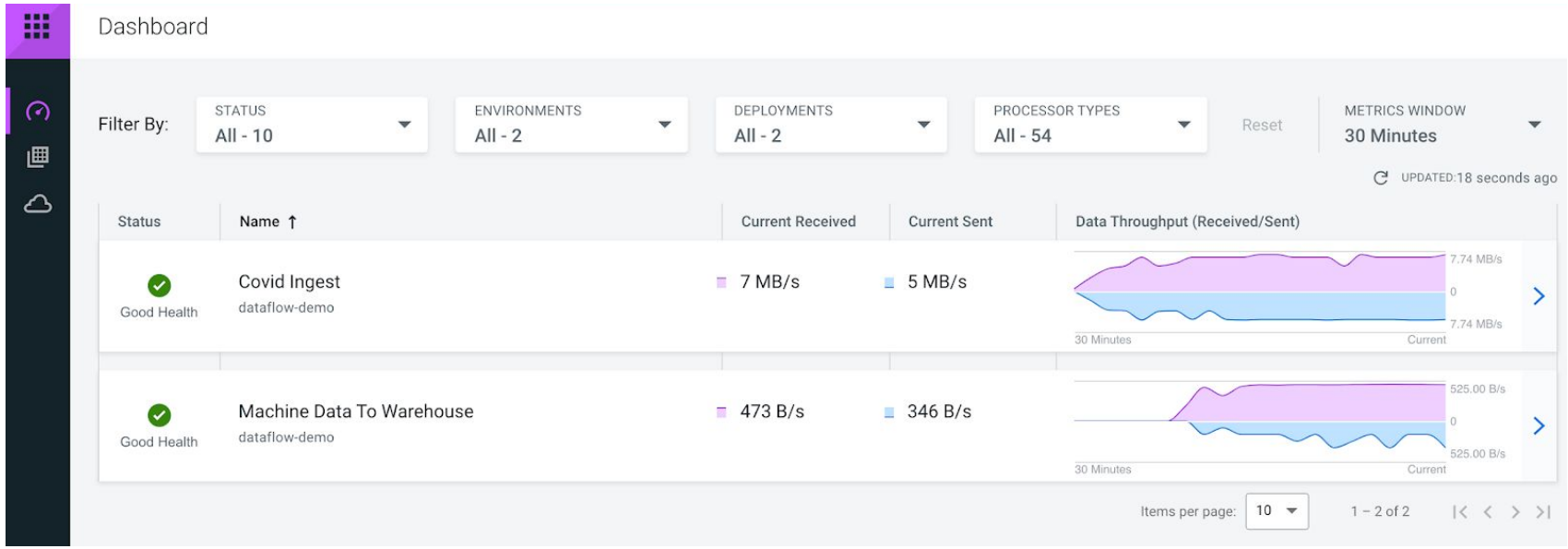


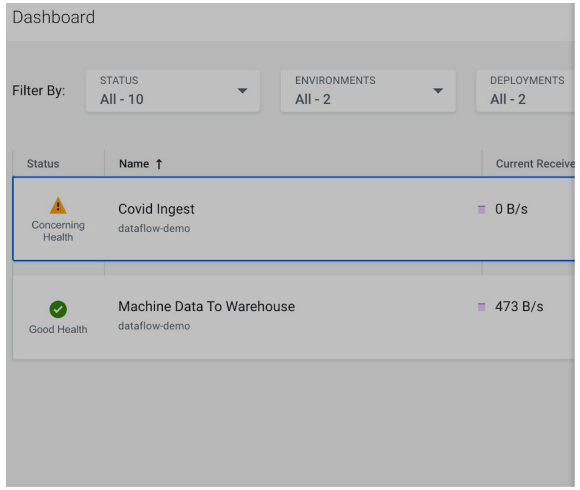
KPI Monitoring

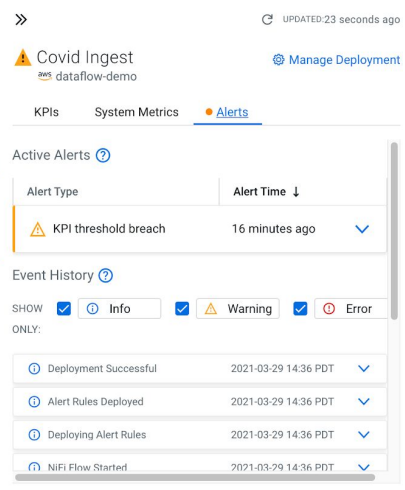


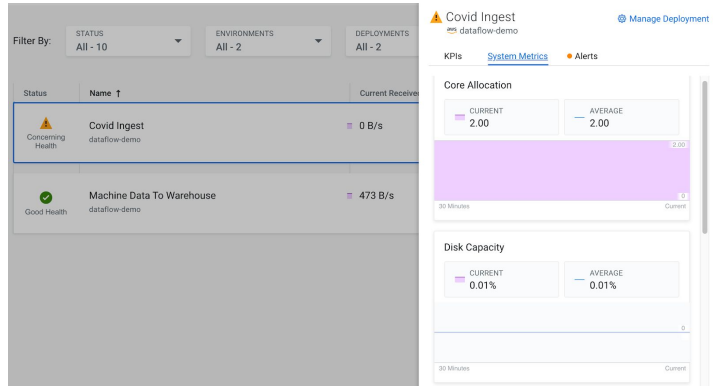
Dashboard

- 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



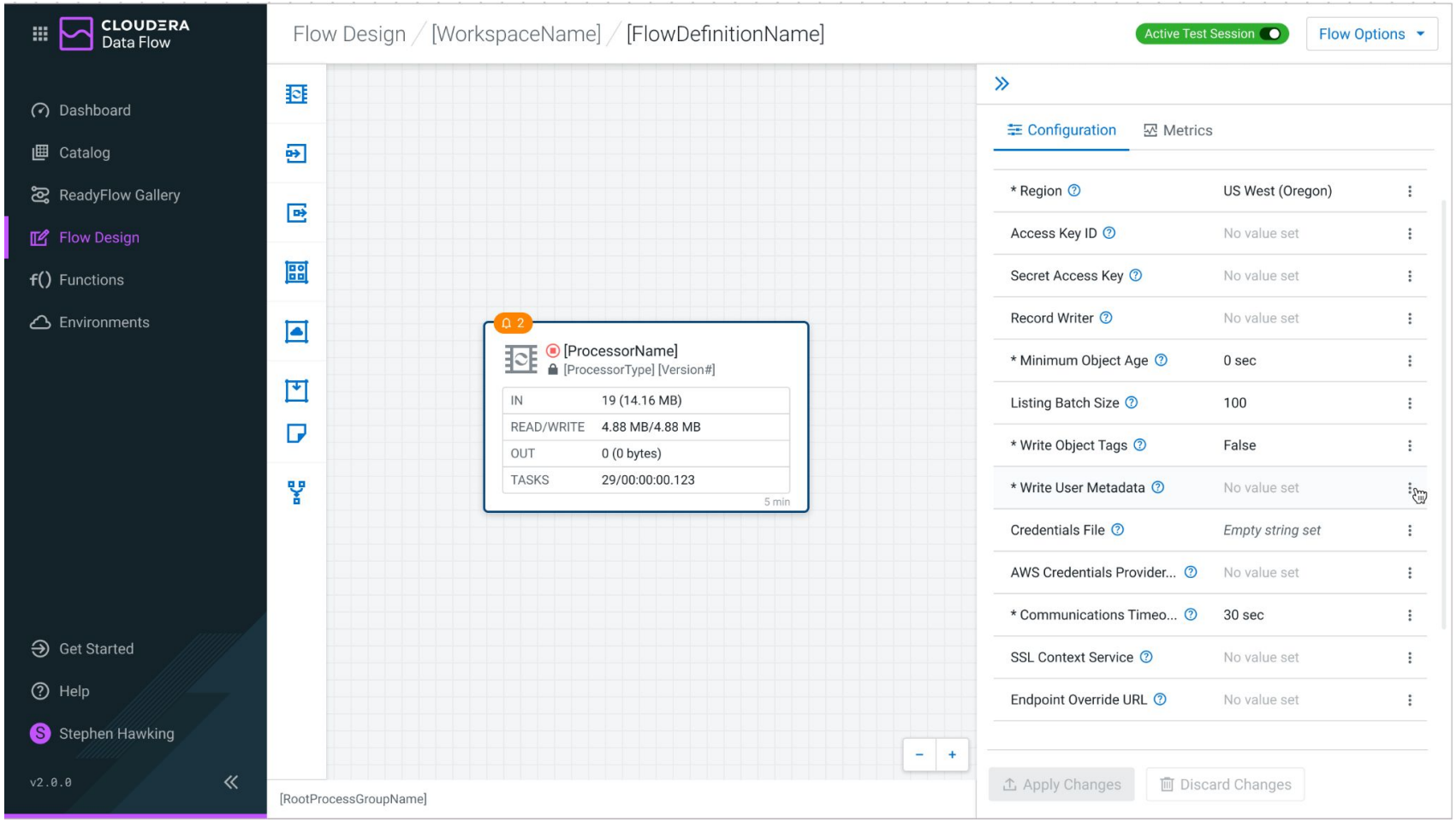


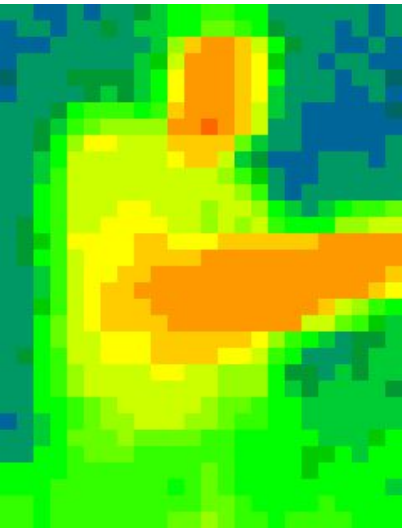




DATA FLOW DESIGN FOR EVERYONE

- 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





THANK YOU