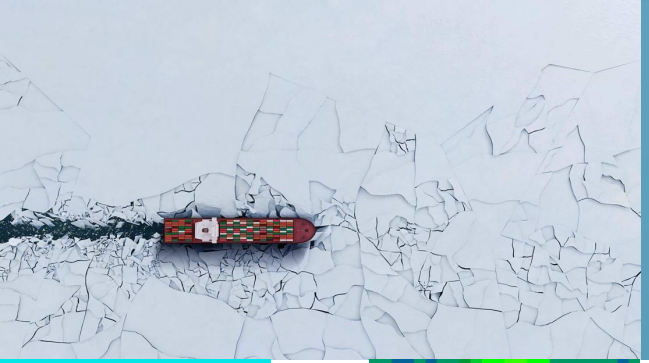
HTAP
SUMMIT 2023

The Never Landing Stream with HTAP and Streaming

Timothy Spann
Principal Developer Advocate

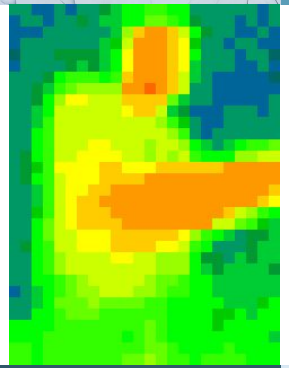
Introduction



IoT

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

CLOUDERA



CLOUDERA

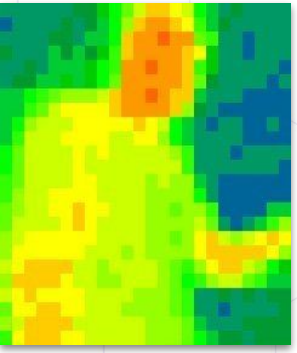
EDGE
2AI

CLOUDERA

ENTERPRISE
DATA
CLOUD

CLOUDERA

FLaNK Stack



Tim Spann

@PaasDev // Blog: **www.datainmotion.dev**

Principal Developer Advocate.

Princeton Future of Data Meetup.

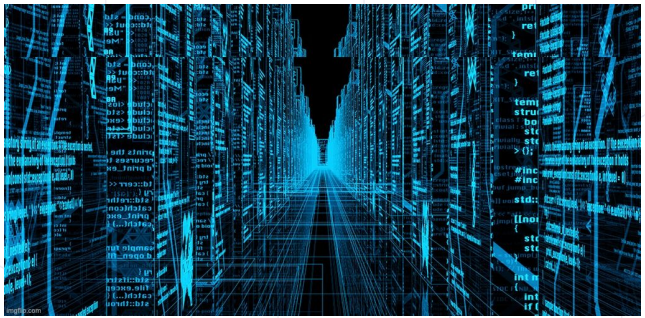
ex-Pivotal, ex-Hortonworks, ex-StreamNative, ex-PwC

https://medium.com/@tspann

https://github.com/tspannhw

Apache NiFi x Apache Kafka x Apache Flink

Future of Data - Princeton + Virtual



https://www.meetup.com/futureofdata-princeton/

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY

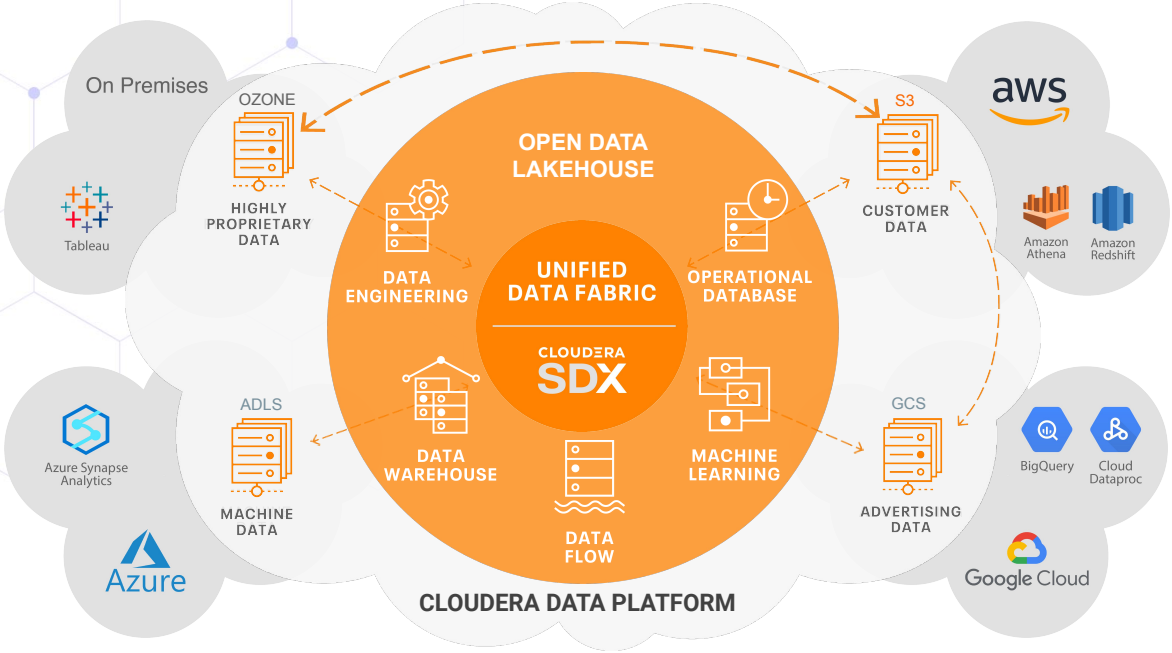
From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...

ICEBERG

@PaasDev

CDP IS THE ONLY HYBRID DATA PLATFORM

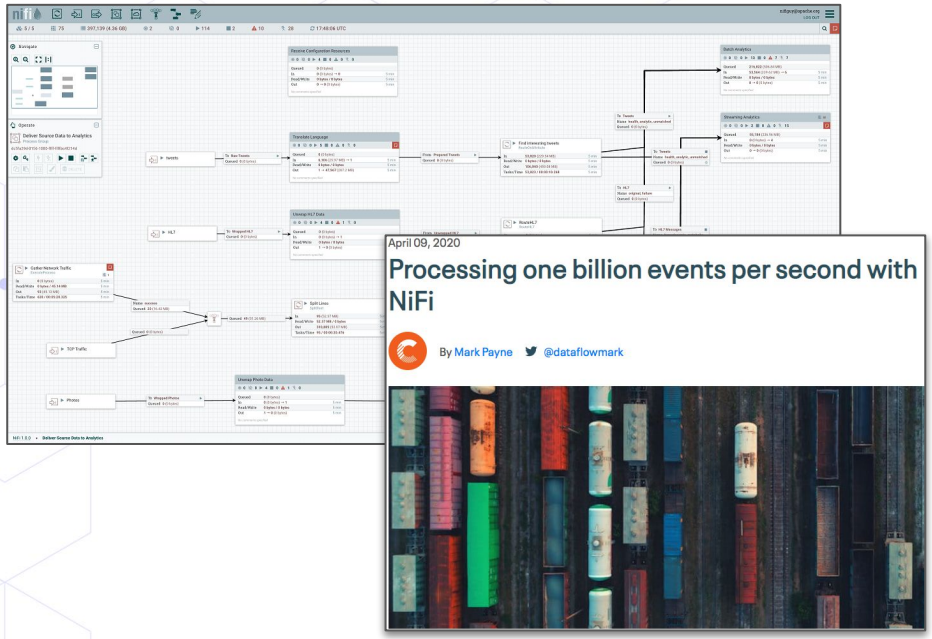
Hybrid. Open. Portable. Secure.



Apache NiFi

CLOUDERA FLOW MANAGEMENT - POWERED BY APACHE NiFi

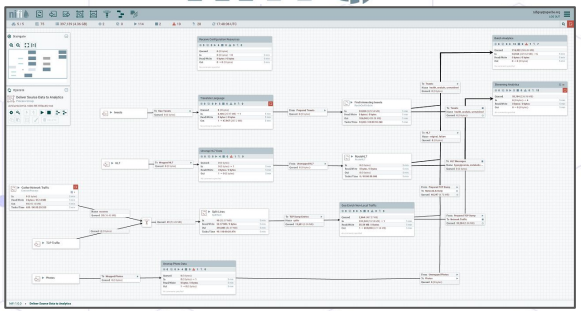
Ingest and manage data from edge-to-cloud using a no-code interface



- #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

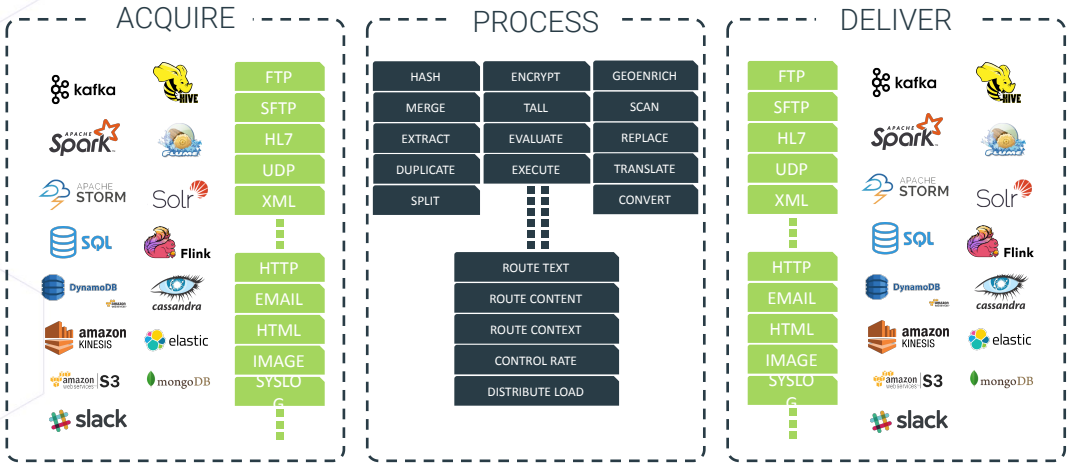
Cloudera Flow Management

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



APACHE NIFI registry

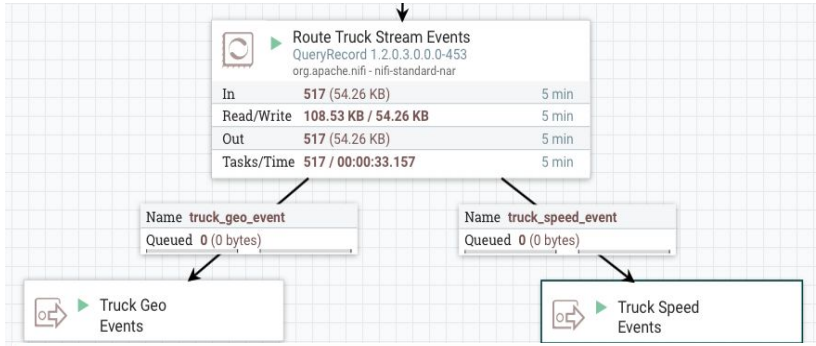
Advanced tooling to industrialize flow development (Flow Development Life Cycle)



- 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

SQL BASED ROUTING WITH NiFi's QueryRecord Processor

- **QueryRecord Processor**– Executes a SQL statement against records and writes the results to the flow file content.
- **CSVReader**: Looking up schema from SR, it will converts CSV Records into ProcessRecords
- **SQL execution via Apache Calcite:** execute configured SQL against the ProcessRecords for routing
- **CSVRecordSetWriter**: Converts the result of the query from Process records into CSV for the for the flow file content



Property	Value
Record Reader	Routing-CSVReader
Record Writer	Routing-CSVWriter
Include Zero Record FlowFiles	false
Cache Schema	false
truck_geo_event	SELECT * FROM FLOWFILE where eventSource = 'truck_geo_event'
truck_speed_event	SELECT * FROM FLOWFILE where eventSource = 'truck_speed_event'

Do routing(routing geo and speed streams) using standard SQL as opposed to complex regular expressions.

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

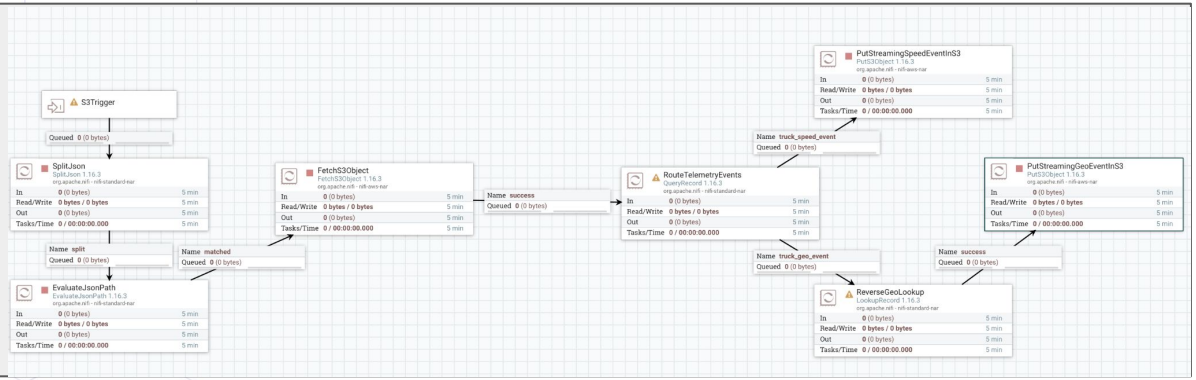
Democratize access to real-time data – Enable data analysts and other personas to quickly build streaming applications with just SQL

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.

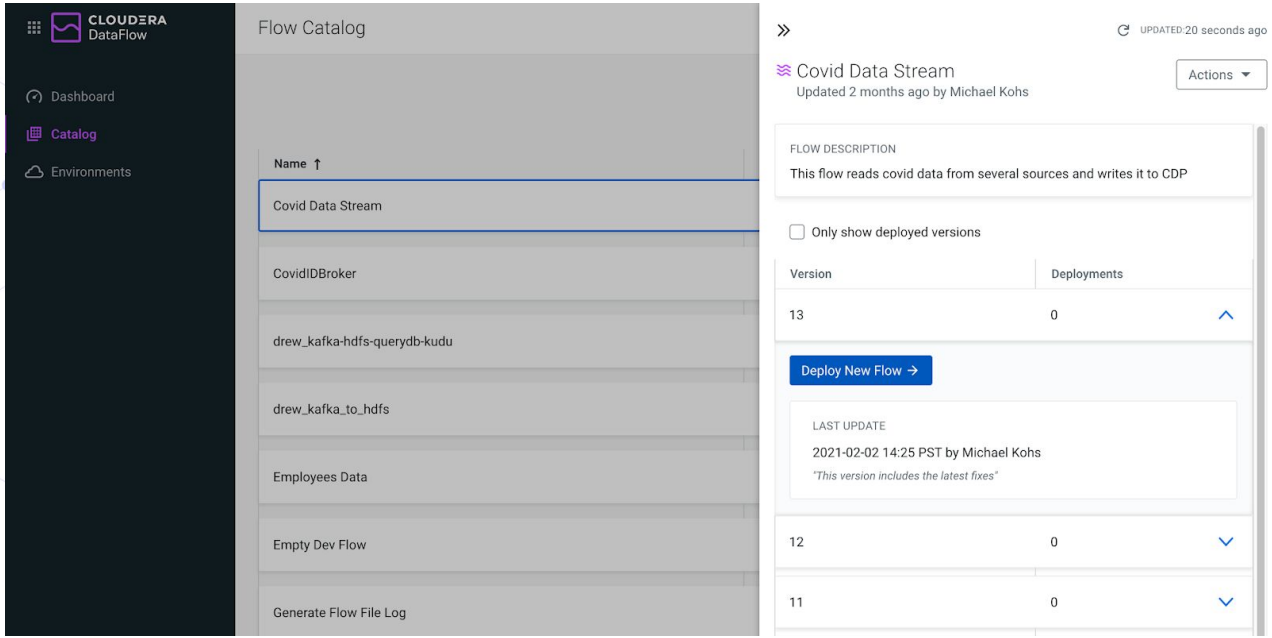
Serverless Customized Triggers

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.

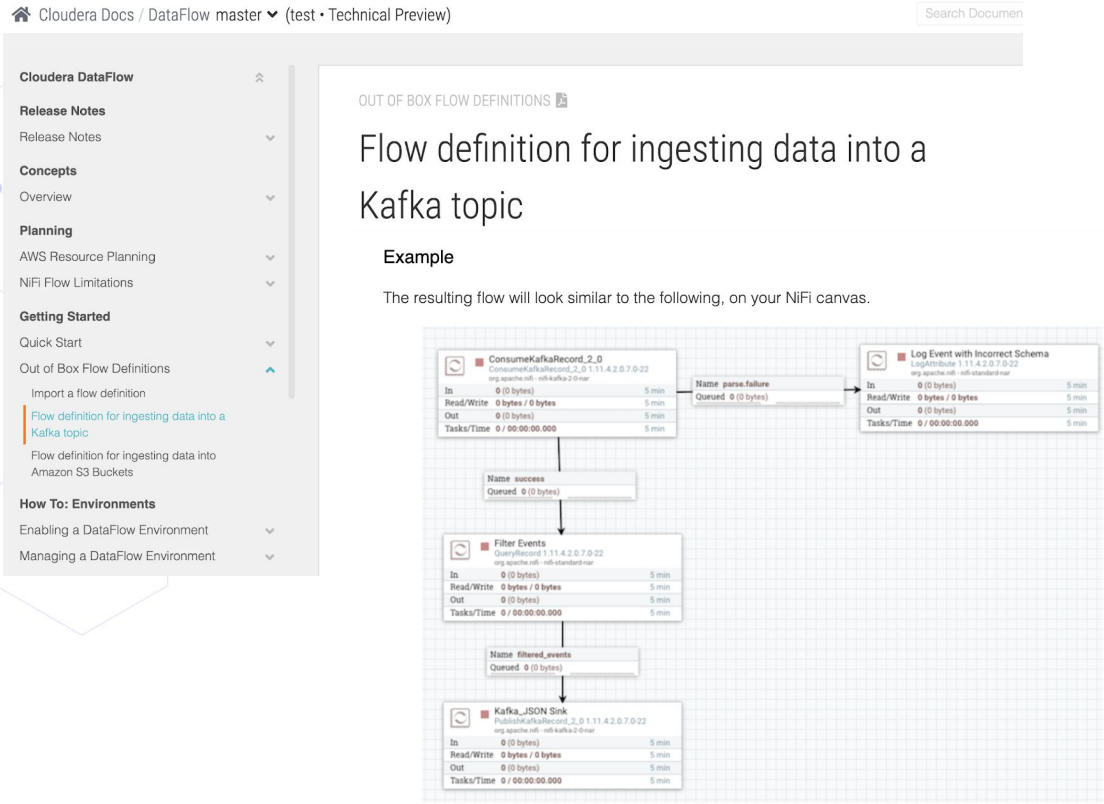
Flow Catalog

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



ReadyFlows

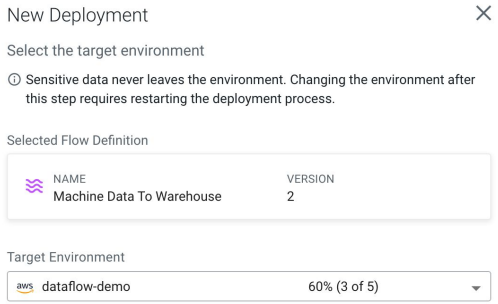
- Cloudera provided flow definitions
- Cover most common data flow use cases
- Can be deployed and adjusted as needed
- Made available through docs during Tech Preview



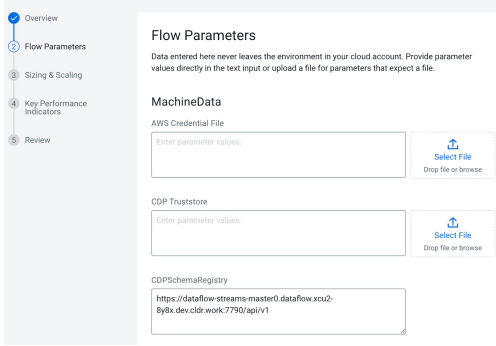
Deployment Wizard

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

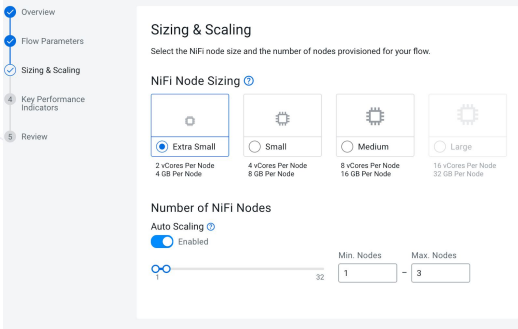
Start Deployment Wizard



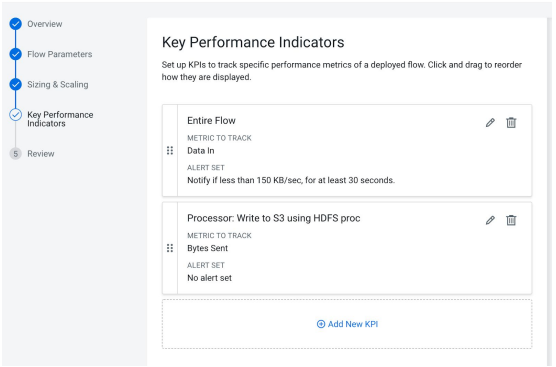
Provide Parameters



Configure Sizing & Scaling



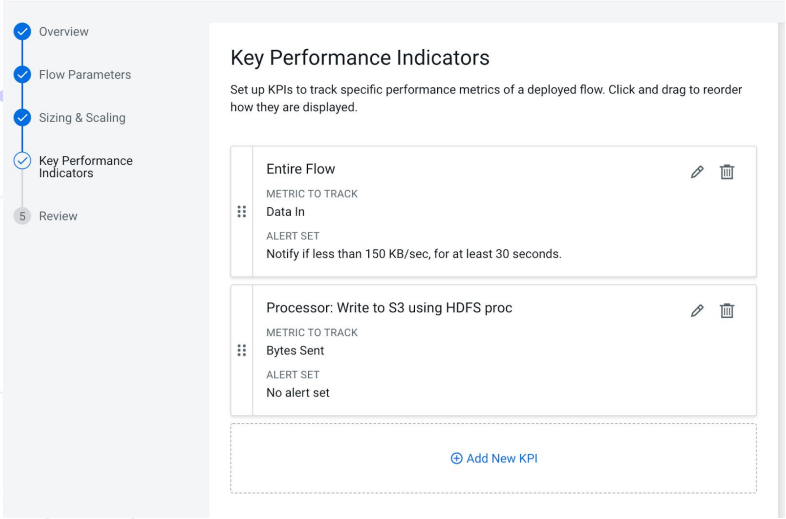
Define KPIs



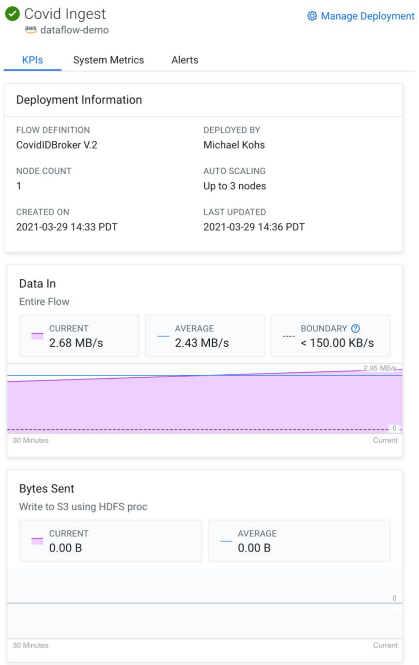
Key Performance Indicators

- 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

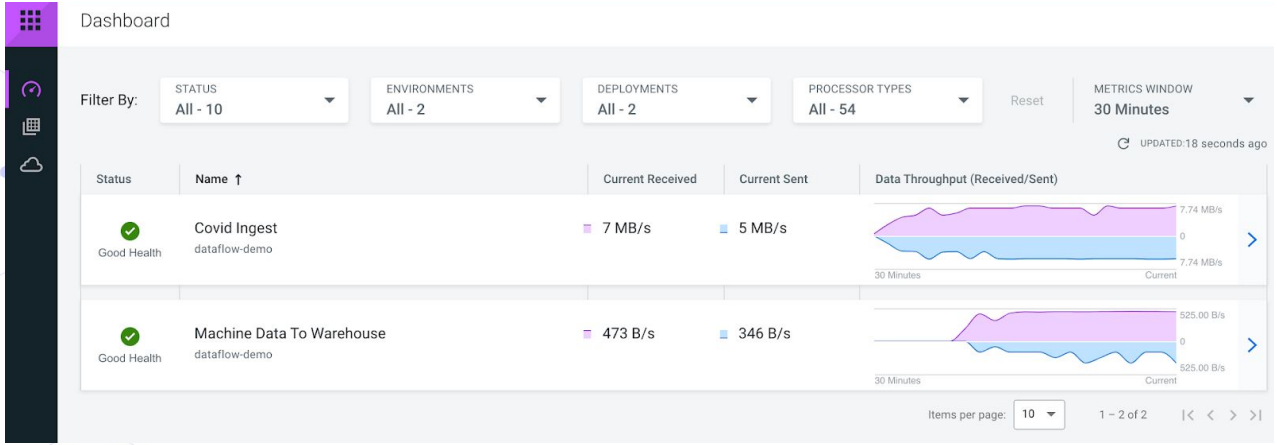


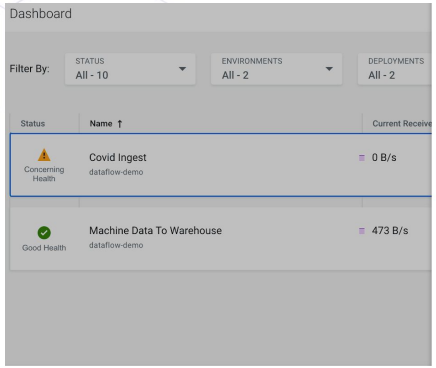
KPI Monitoring

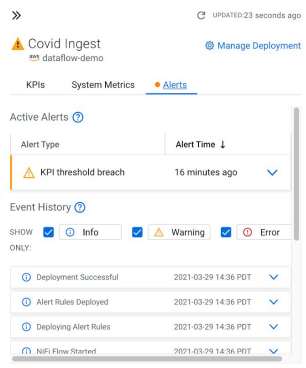


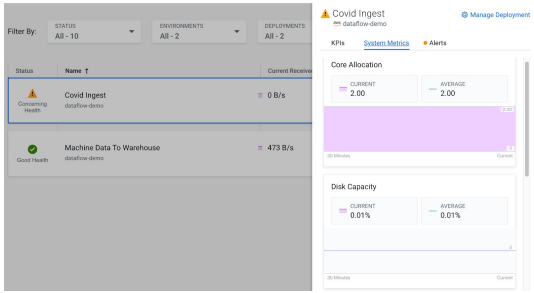
Dashboard

- 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



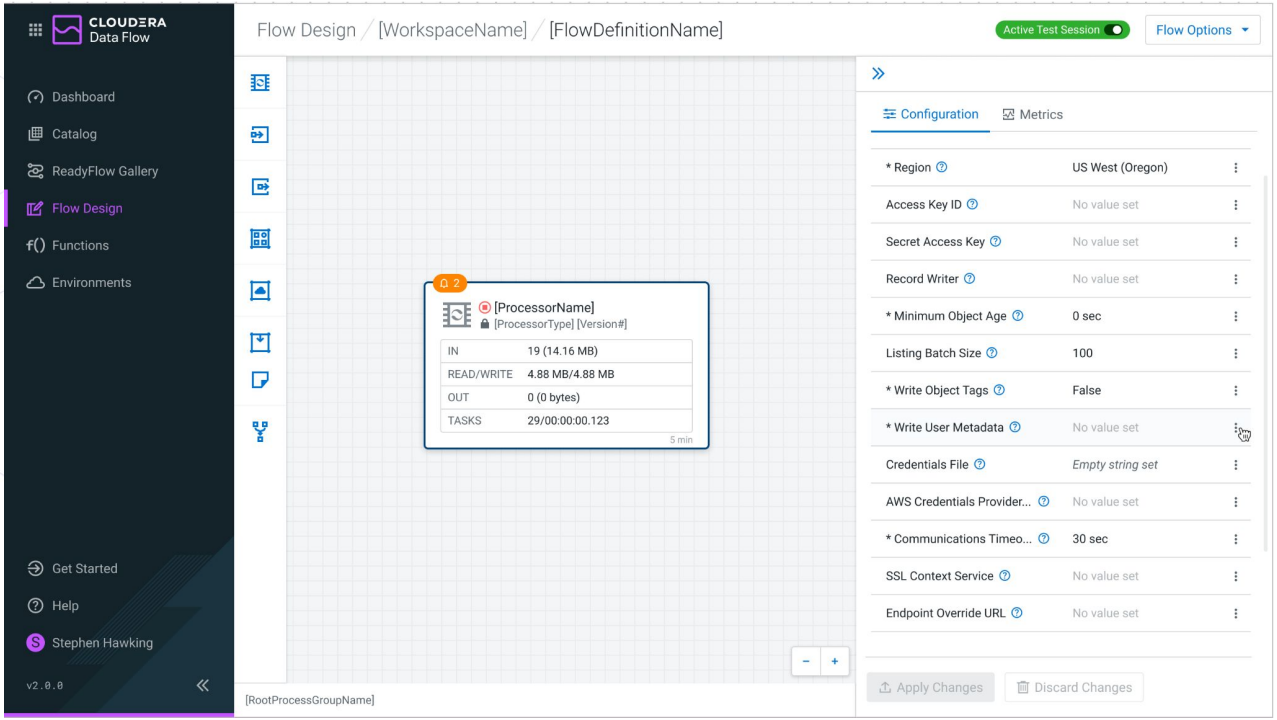






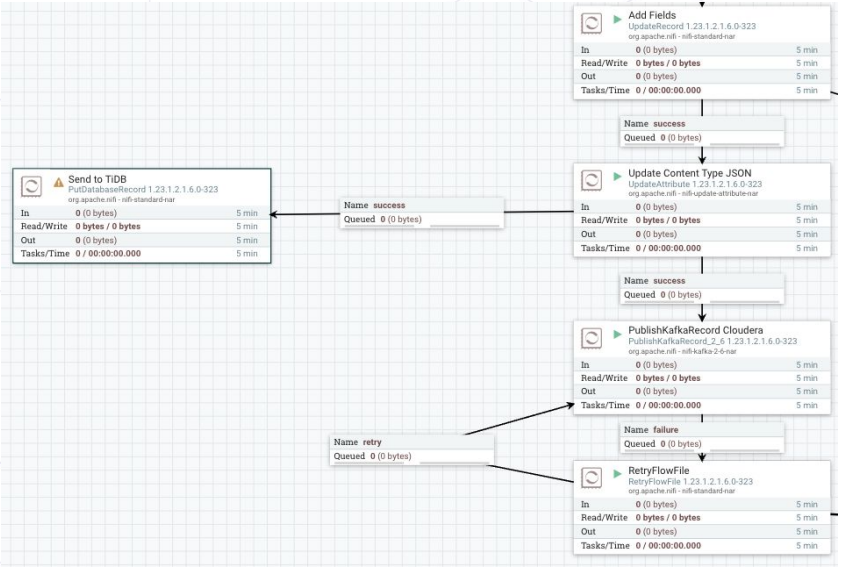
Data Flow Design for Everyone

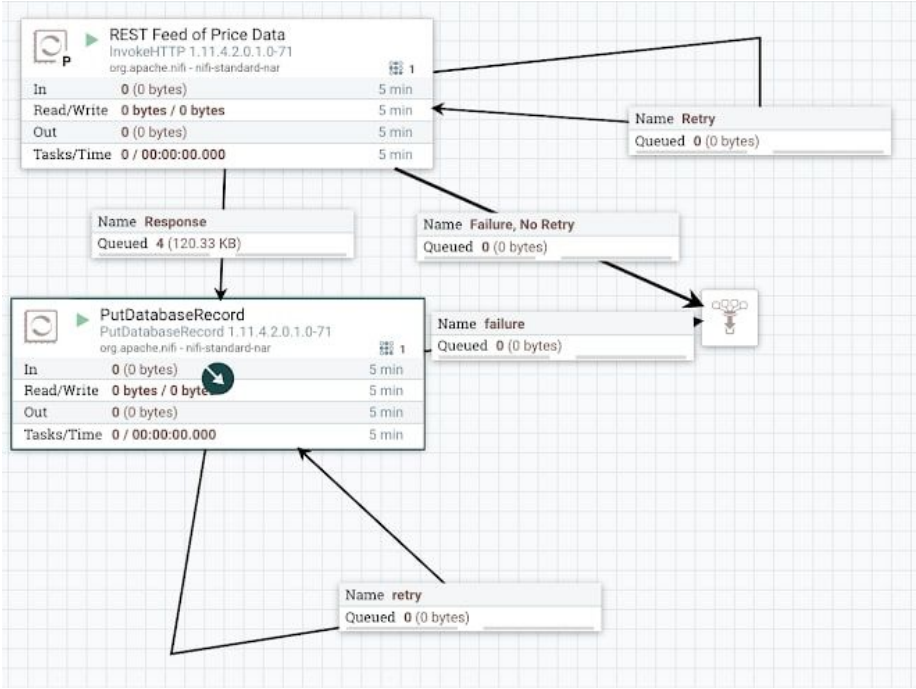
- Cloud-native data flow development
- Developers get their own sandbox
- Start developing flows without installing NiFi
- Redesigned visual canvas
- Optimized interaction patterns
- Integration into CDF-PC Catalog for versioning



Data Distribution and Sharing with TiDB

HTAP SUMMIT 2023



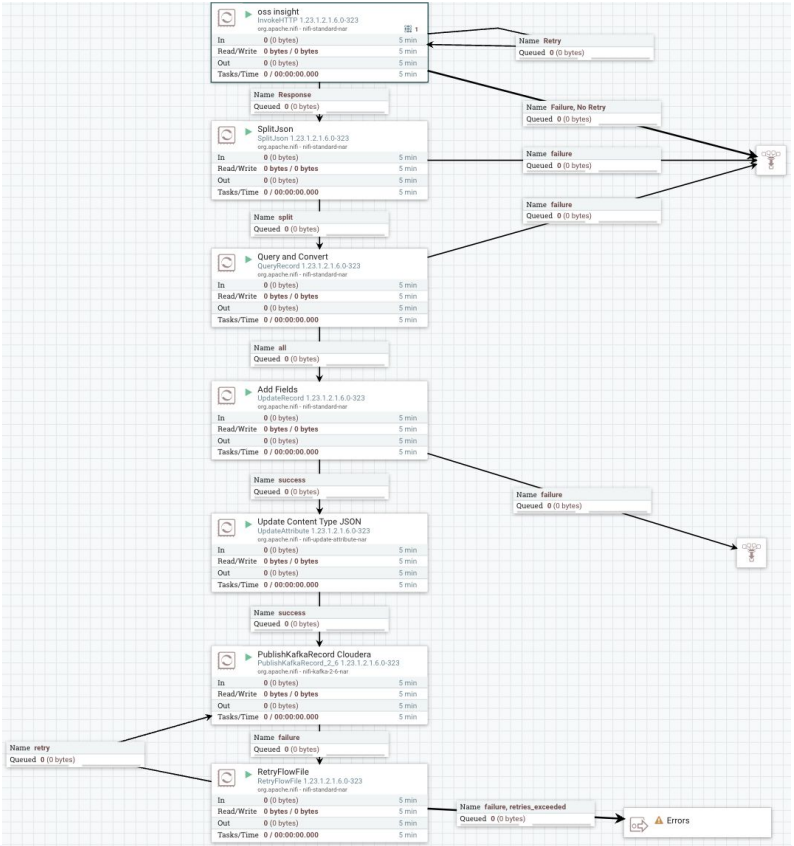


https://docs.pingcap.com/tidb/dev/mysql-compatibility

NiFi Ingesting REST API

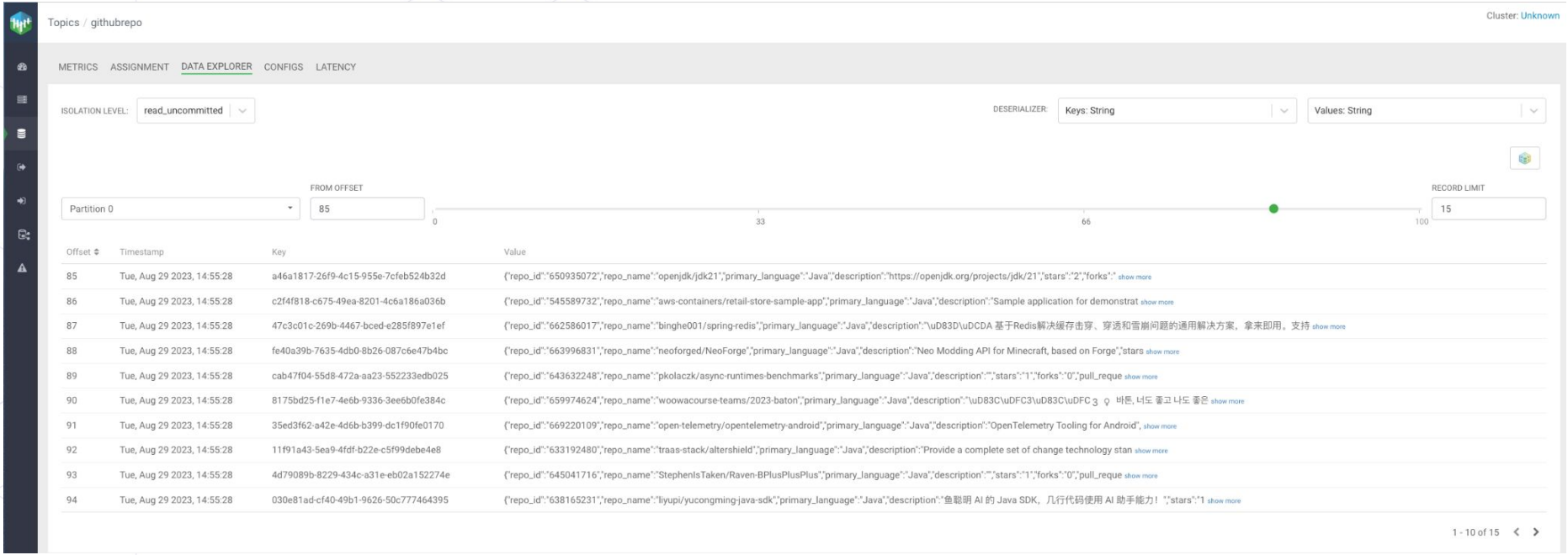
- NiFi consumes stream (cdc, rest, sensors)
- Distributes real-time to
- Kafka and MySQL at same time
- Flink SQL consumes from Kafka
- TiDB CDC -> Kafka

https://ossinsight.io/docs/api



Apache Kafka

Data Distribution with Apache Kafka



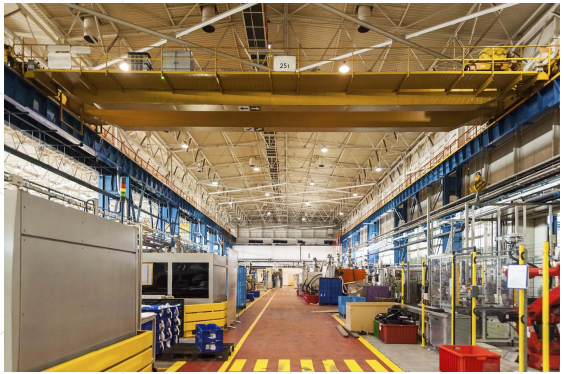
Apache Kudu

Why Kudu?

A simultaneous combination of sequential and random reads and writes

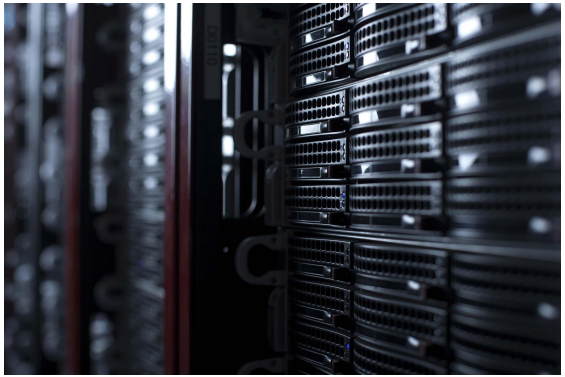
Time Series Data

Can you insert time series data in real time? How long does it take to prepare it for analysis? Can you get results and act fast enough to change outcomes?



Machine Data Analytics

Can you handle large volumes of machine-generated data? Do you have the tools to identify problems or threats? Can your system do machine learning?



Online Reporting

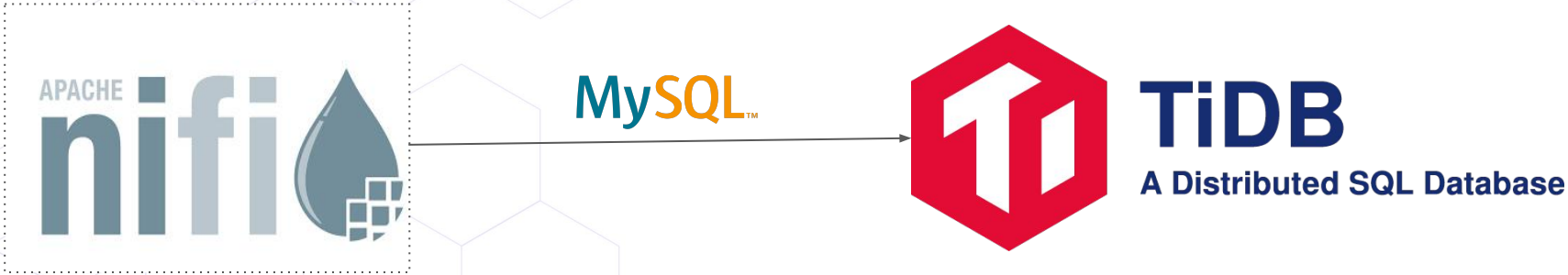
How fast can you add data to your data store? Are you trading off the ability to do broad analytics for the ability to make updates? Are you retaining only part of your data?

HTAP Options - Apache Kudu

Online Transactional Processing (OLTP) and Online Analytical Processing (OLAP)

https://hpi.de/fileadmin/user_upload/hpi/navigation/10_forschung/20_future_soc_lab/Poster/2019-1/Tozun_FSOC-Poster_20191_150443.pdf

HTAP Options - TiDB



Apache Flink SQL

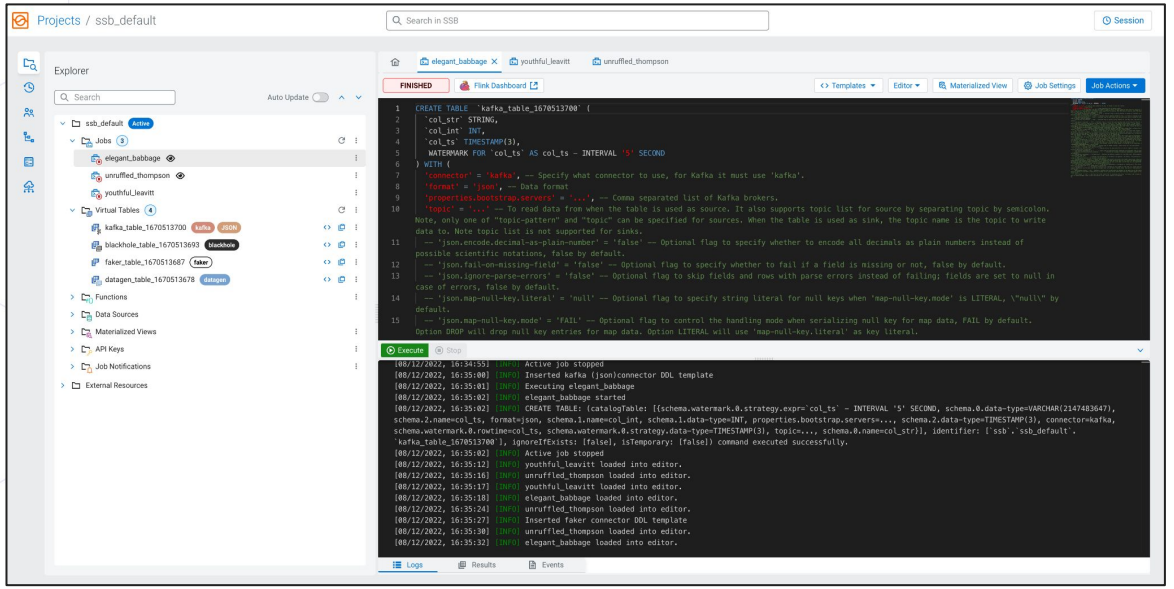
SQL STREAM BUILDER (CLOUDERA SSB)

SQL STREAM BUILDER allows developers, analysts, and data scientists to write streaming applications with industry standard SQL.

No Java or Scala code development required.

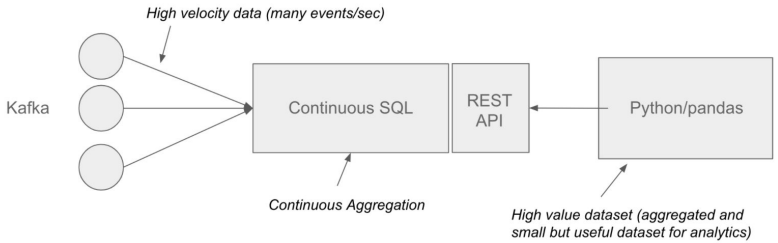
Simplifies access to data in Kafka & Flink. Connectors to batch data in HDFS, Kudu, Hive, S3, JDBC, CDC and more

Enrich streaming data with batch data in a single tool



SSB MATERIALIZED VIEWS

Key Takeaway; MV's allow data scientist, analyst and developers consume data from the firehose




```sql
SELECT userid,
max(amount) as max_amount,
sum(amount) as sum_amount,
count(*) as thecount,
tumble_end(eventTimestamp, interval '5' second) as ts
FROM authorizations
GROUP BY userid, tumble(eventTimestamp, interval '5' second)
HAVING count(*) > 1
```

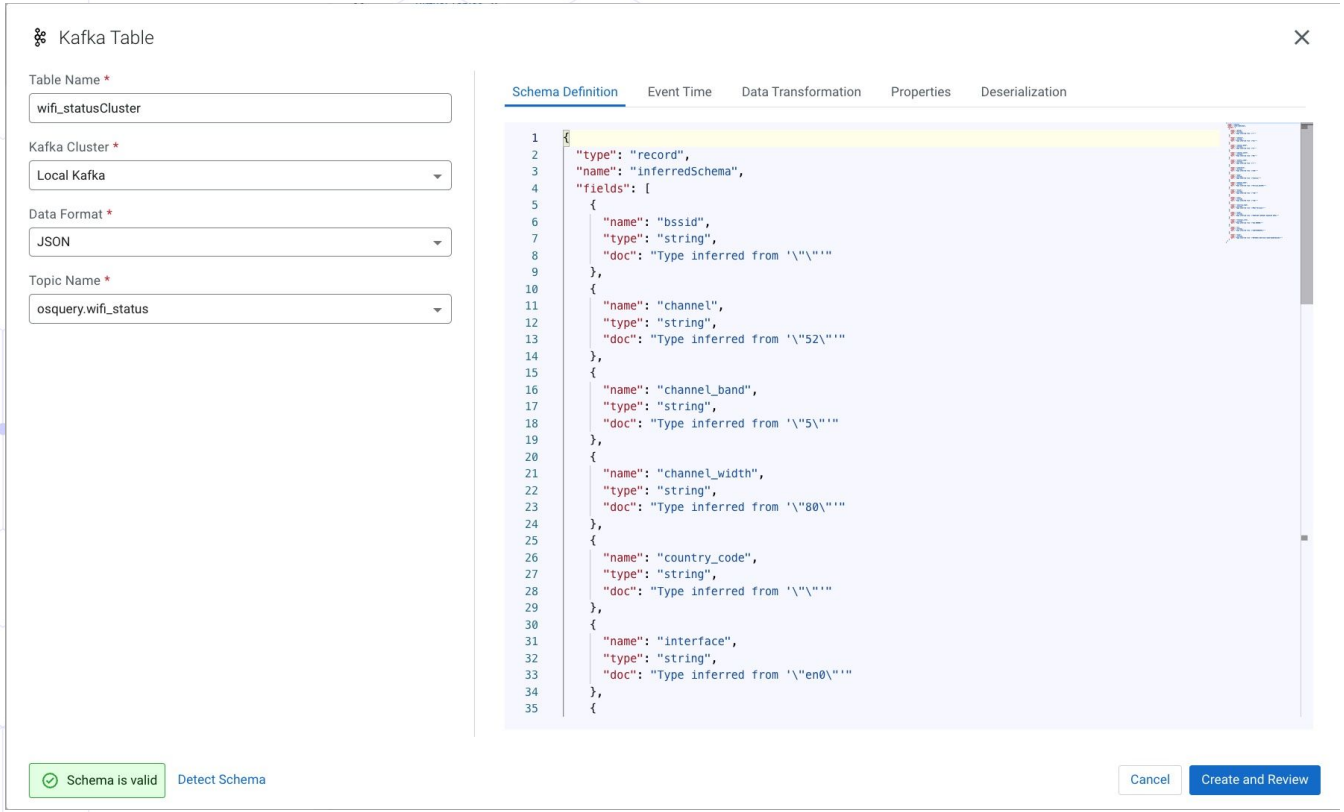
```
[90]: import pandas as pd
[91]: mv = "https://xxxyyyyzzz"
[92]: df = pd.read_json(mv)
[93]: len(df.keys())
[93]: 5
[95]: df['ts'] = pd.to_datetime(df['ts'])
[97]: df.dtypes
[97]: max_amount              int64
      sum_amount              int64
      thecount                int64
      ts             datetime64[ns]
      userid                  int64
      dtype: object
[98]: df.set_index('userid').sort_values(by=['thecount'], ascending=False).head()
```

[98]:

userid	max_amount	sum_amount	thecount	ts
787	34911	57304	10	2020-06-16 19:52:15
744	77407	95407	9	2020-06-16 19:52:15
78	86761	330397	9	2020-06-16 19:52:15
541	78762	282682	8	2020-06-16 19:52:15
926	85636	129728	8	2020-06-16 19:52:15

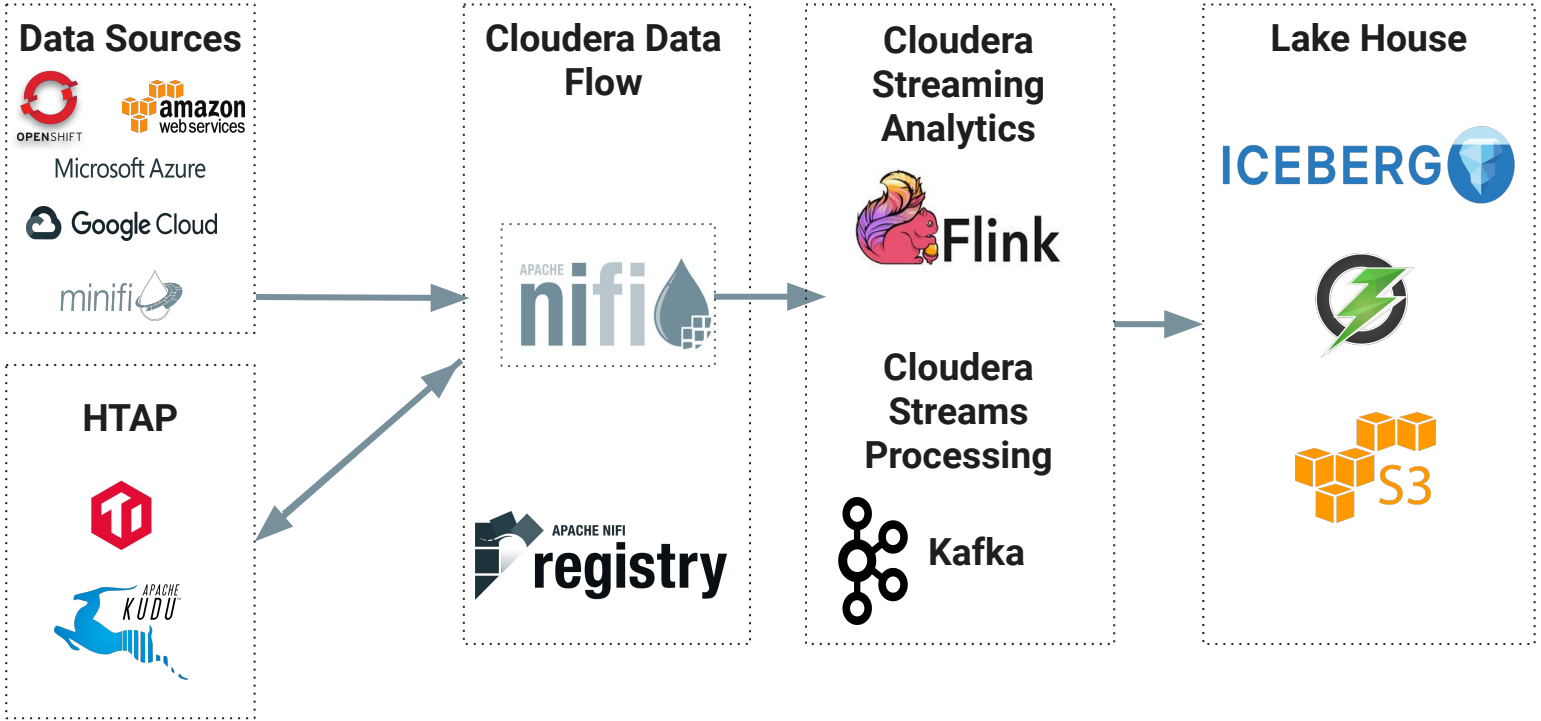
```
[ ]:
```

Infer Tables from Kafka Topics with JSON or Avro



Demos

INGEST OF ALL DATA



LLM USE CASE

Unstructured file types

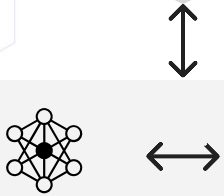
Structured Sources

GURU

Other enterprise data

Data in Motion on Cloudera Data Platform (CDP)

Capture, process & distribute any data, anywhere



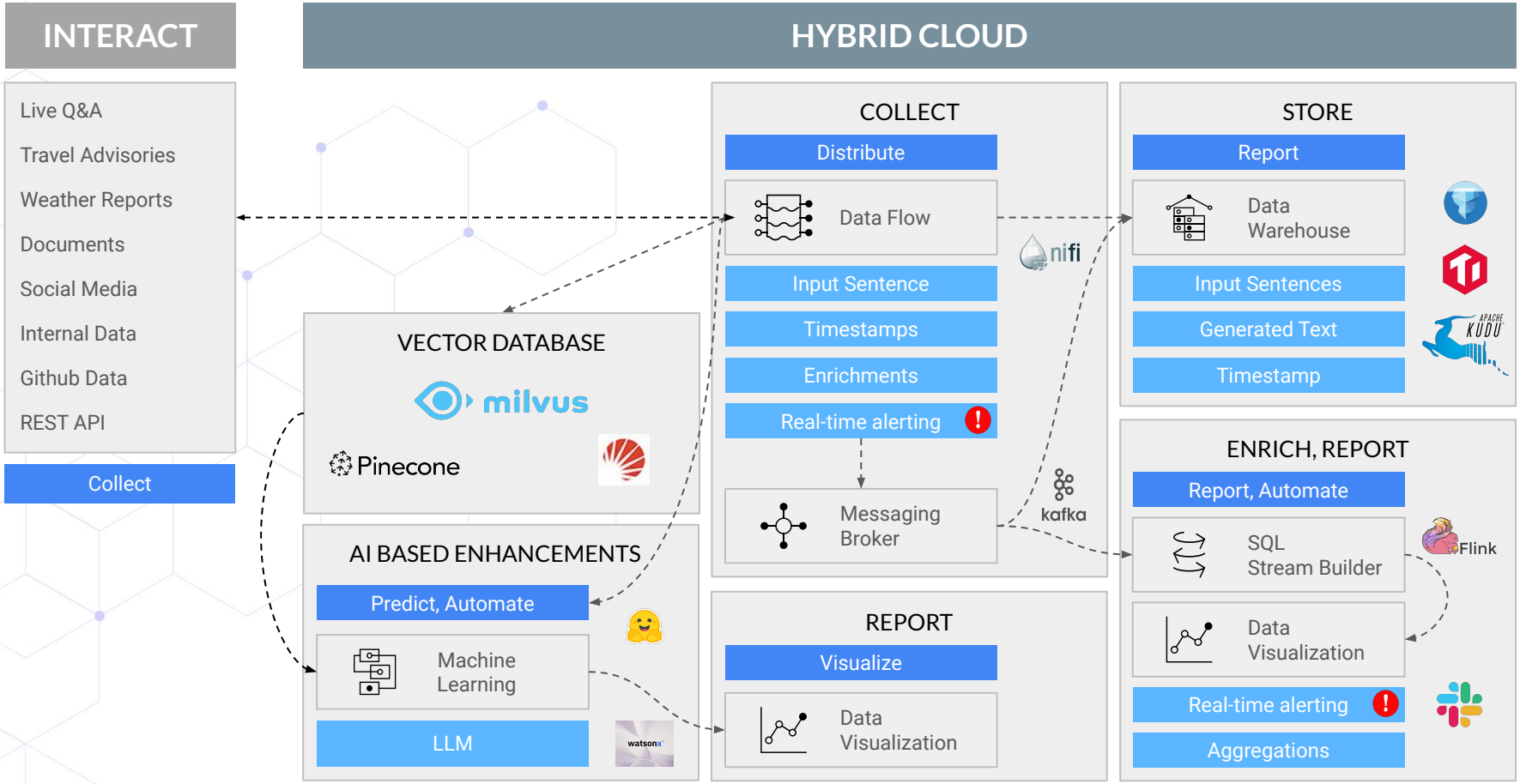
AI Model ⟷ Vector DB

Applications/API's

Streams

Materialized Views

Open Data Lakehouse


INTERACT
HYBRID CLOUD
Live Q&A
Travel Advisories
Weather Reports
Documents
Social Media
Internal Data
Github Data
REST API
Collect
VECTOR DATABASE
milvus
Pinecone
AI BASED ENHANCEMENTS
Predict, Automate
Machine Learning
LLM
watsonx
COLLECT
Distribute
Data Flow
nifi
Input Sentence
Timestamps
Enrichments
Real-time alerting
Messaging Broker
kafka
REPORT
Visualize
Data Visualization
STORE
Report
Data Warehouse
Input Sentences
Generated Text
Timestamp
APACHE KUDU
ENRICH, REPORT
Report, Automate
SQL Stream Builder
Flink
Data Visualization
Real-time alerting
Aggregations

RUN AT HOME

CSP Community Edition

- Kafka, KConnect, SMM, SR, Flink, and SSB in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

- Docker compose file of CSP to run from command line w/o any dependencies, including Flink, SQL Stream Builder, Kafka, Kafka Connect, Streams Messaging Manager and Schema Registry
  - `$> docker compose up`
- Licensed under the Cloudera Community License
- **Unsupported**
- Community Group Hub for CSP
- Find it on docs.cloudera.com under Applications

CSP Community Edition

A readily available, dockerized deployment of Apache Kafka and Apache Flink that allows you to test the features and capabilities of Cloudera Stream Processing.

Learn More

Open Source Edition

- Apache NiFi in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

- Licensed under the ASF License
- **Unsupported**

https://hub.docker.com/r/apache/nifi

Thank You