CLOUDERA

Building Real-Time Pipelines with FLaNK: A Case Study with Transit Data

Tim Spann
Principal Developer Advocate

July-2023

Building Real-time Pipelines with FLaNK: A Case Study with Transit Data

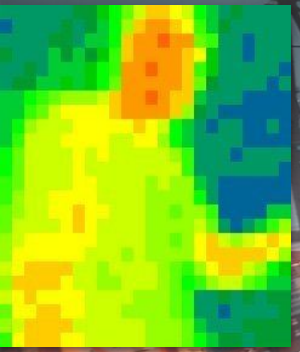
In this session, we will explore the powerful combination of Apache Flink, Apache NiFi, and Apache Kafka for building real-time data processing pipelines. We will present a case study using the FLaNK-MTA project, which leverages these technologies to process and analyze real-time data from the New York City Metropolitan Transportation Authority (MTA). By integrating Flink, NiFi, and Kafka, FLaNK-MTA demonstrates how to efficiently collect, transform, and analyze high-volume data streams, enabling timely insights and decision-making.

Takeaways:

Understanding the integration of Apache Flink, Apache NiFi, and Apache Kafka for real-time data processing
Insights into building scalable and fault-tolerant data processing pipelines
Best practices for data collection, transformation, and analytics with FLaNK-MTA as a reference
Knowledge of use cases and potential business impact of real-time data processing pipelines

https://github.com/tspannhw/FLaNK-MTA/tree/main
https://medium.com/@tspann/finding-the-best-way-around-7491c76ca4cb

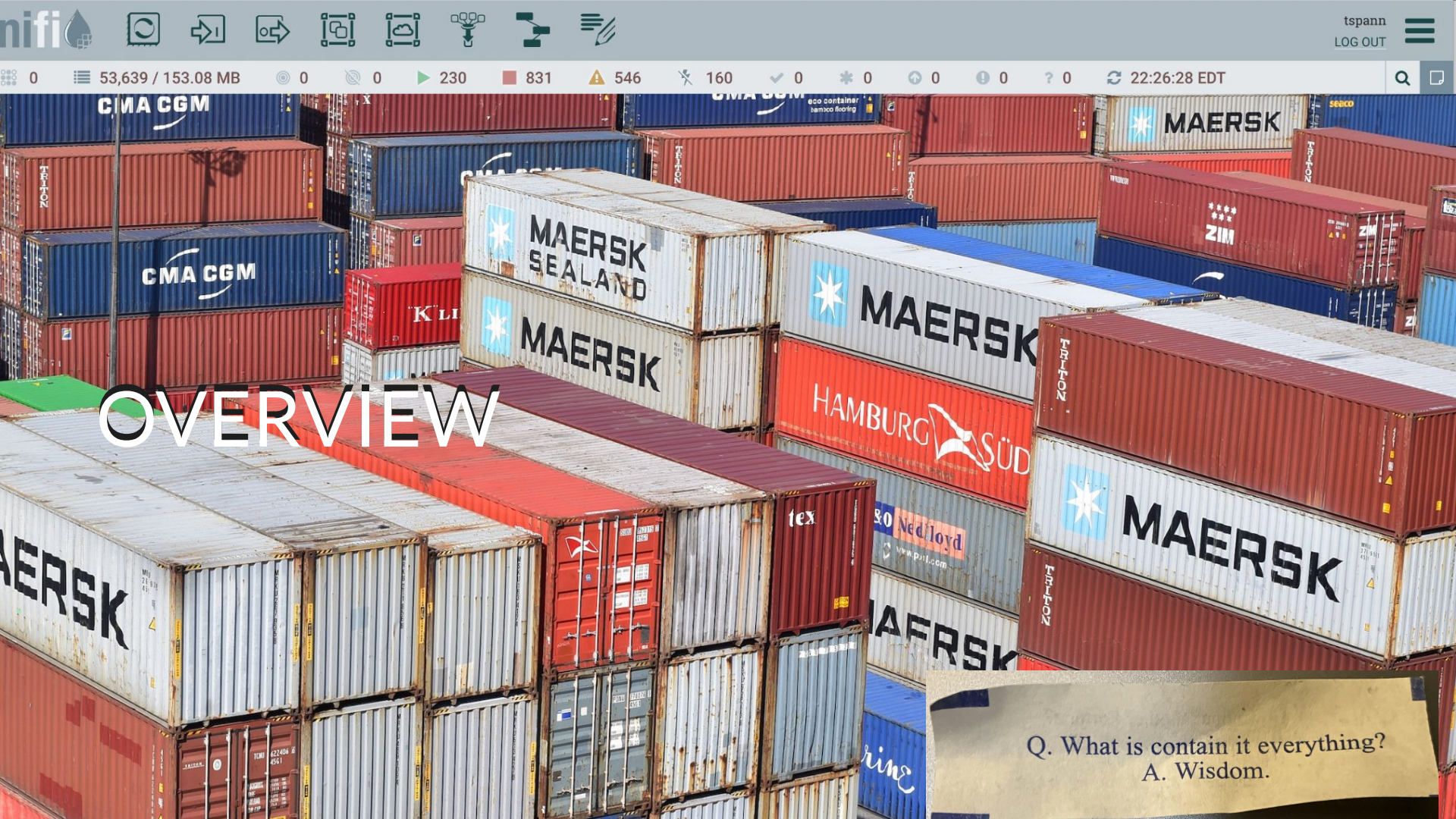
Tim Spann



@PaasDev www.datainmotion.dev
github.com/tspannhw medium.com/@tspann
Principal Developer Advocate

Princeton Future of Data Meetup
ex-Pivotal, ex-Hortonworks, ex-StreamNative,
ex-PwC, ex-EY, ex-HPE.

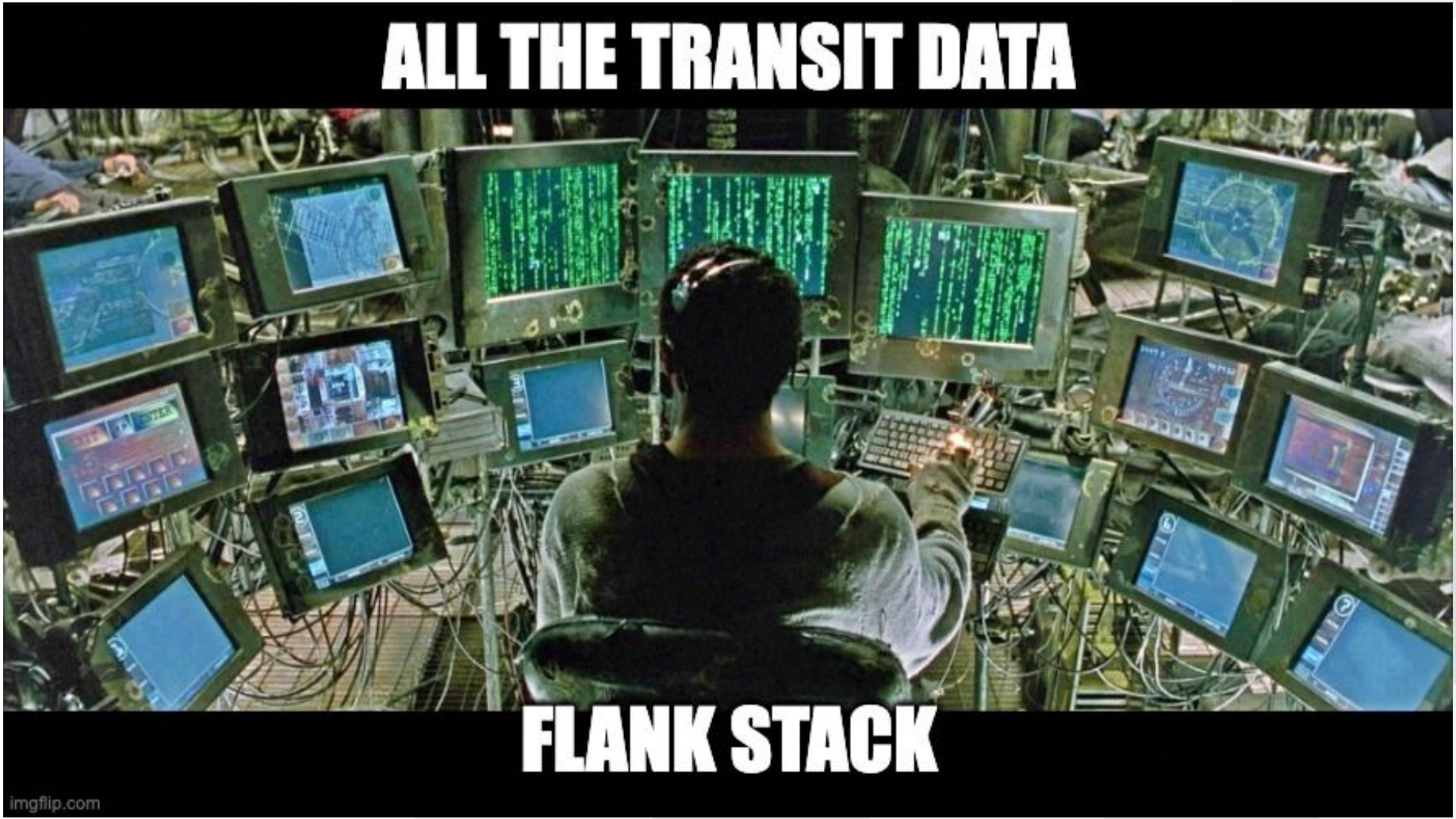
Apache NiFi x Apache Kafka x Apache Flink x Java


nifi
tspann
LOG OUT
0
53,639 / 153.08 MB
0
0
230
831
546
160
0
0
0
0
0
22:26:28 EDT
OVERVIEW
Q. What is contain it everything?
A. Wisdom.

Trains, Planes and Automobiles +++

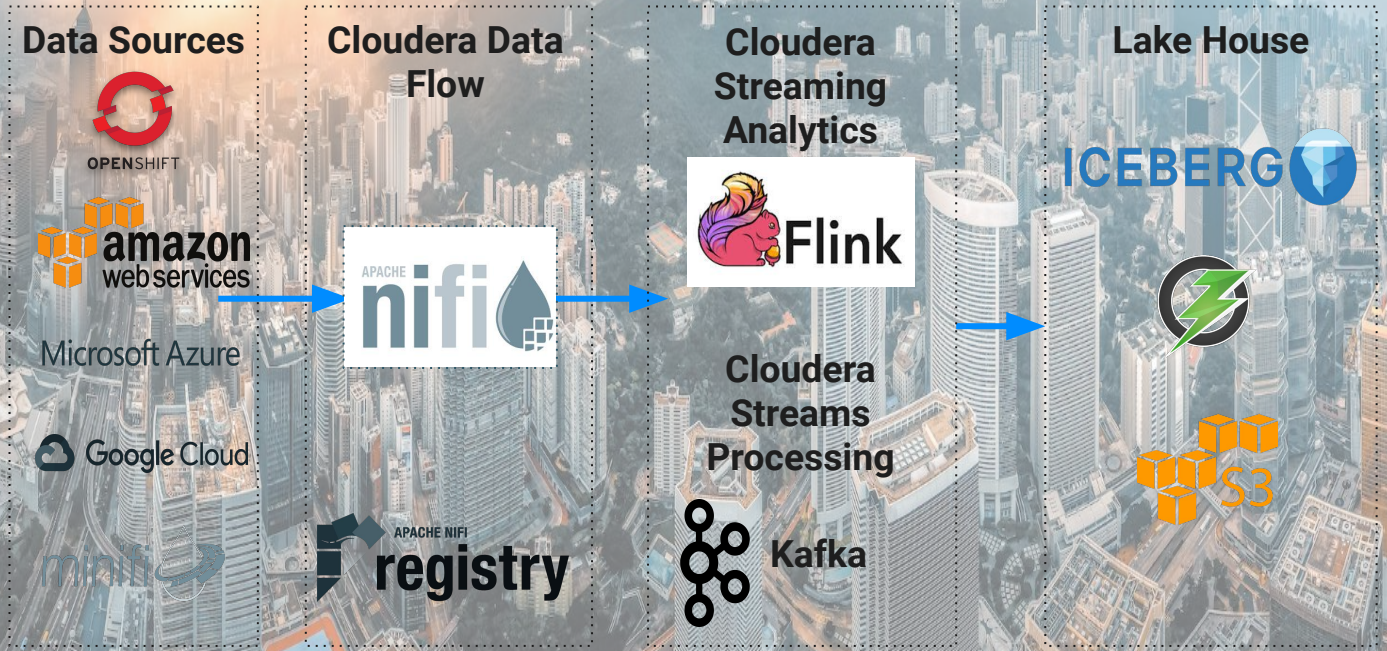
Information Needed	Data Feed(s)
Local weather conditions	• XML, JSON, RSS
Mass transit status & alerts	• XML, JSON, RSS
Regional highways & tunnels	• GeoRSS, XML, ProtoBuf, JSON
Local social media	• JSON
ADS-B Plane Data	• JSON
Local air quality	• JSON

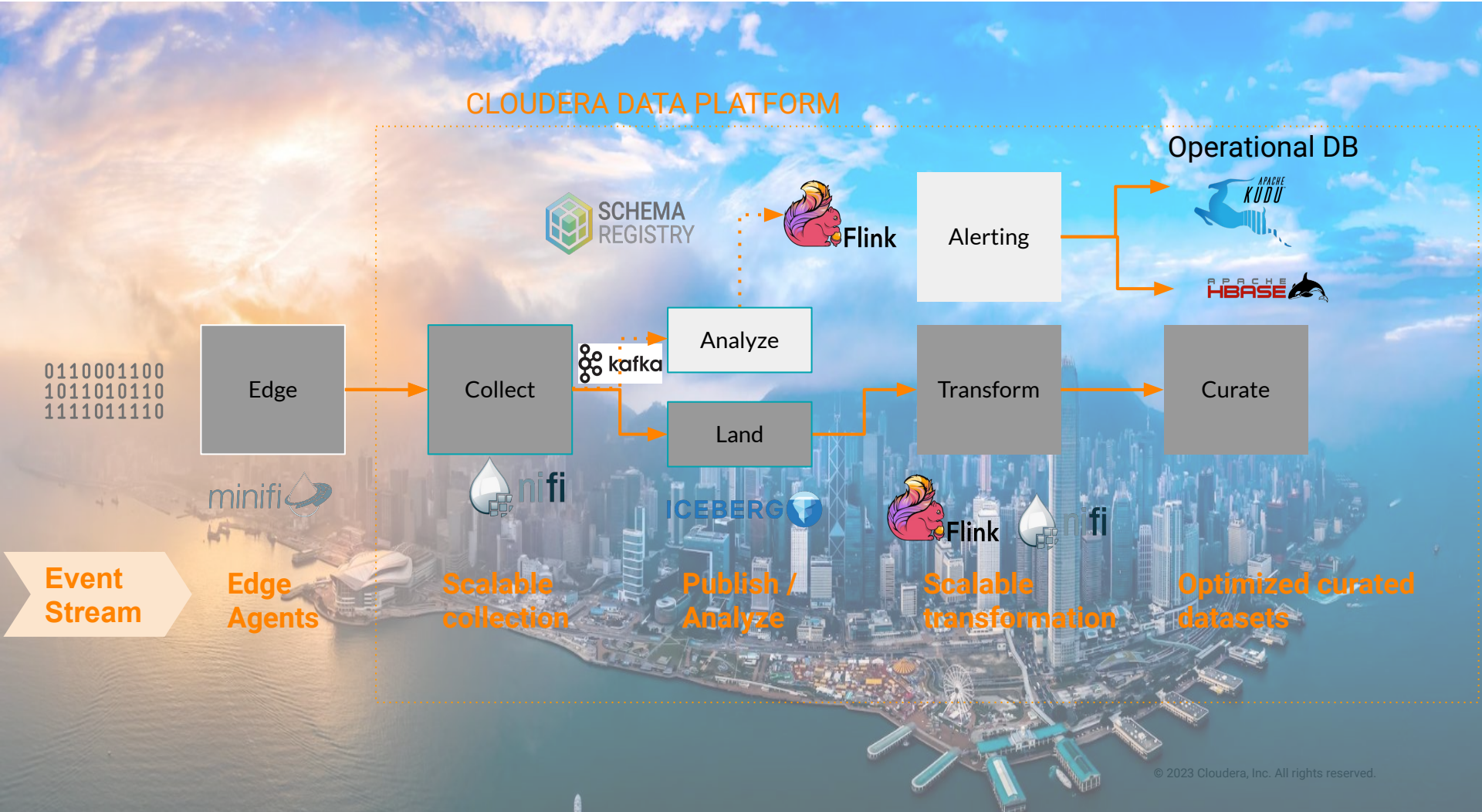
REAL-TIME REQUIRES A PLATFORM

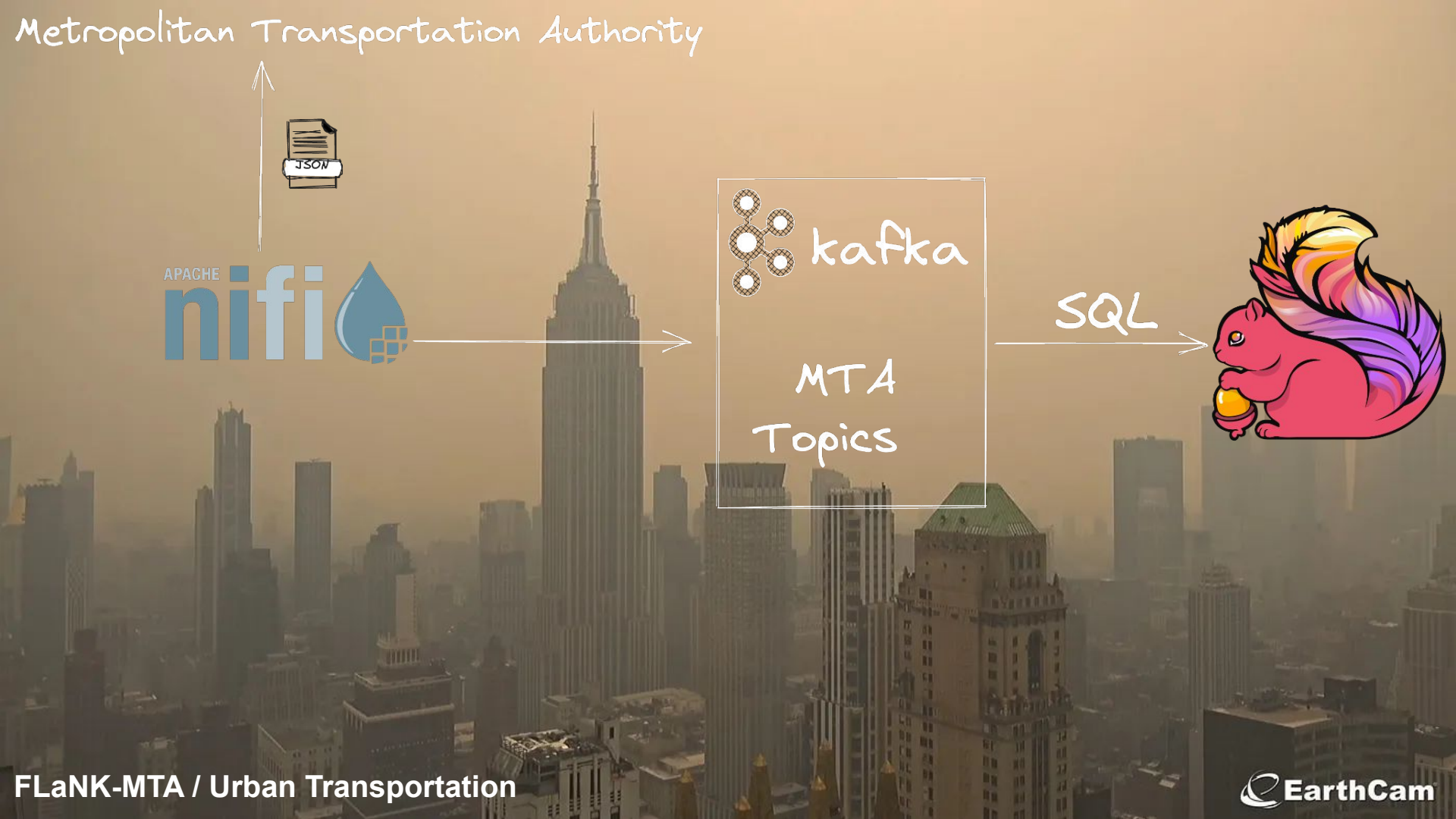

ALL THE TRANSIT DATA
FLANK STACK
imgflip.com

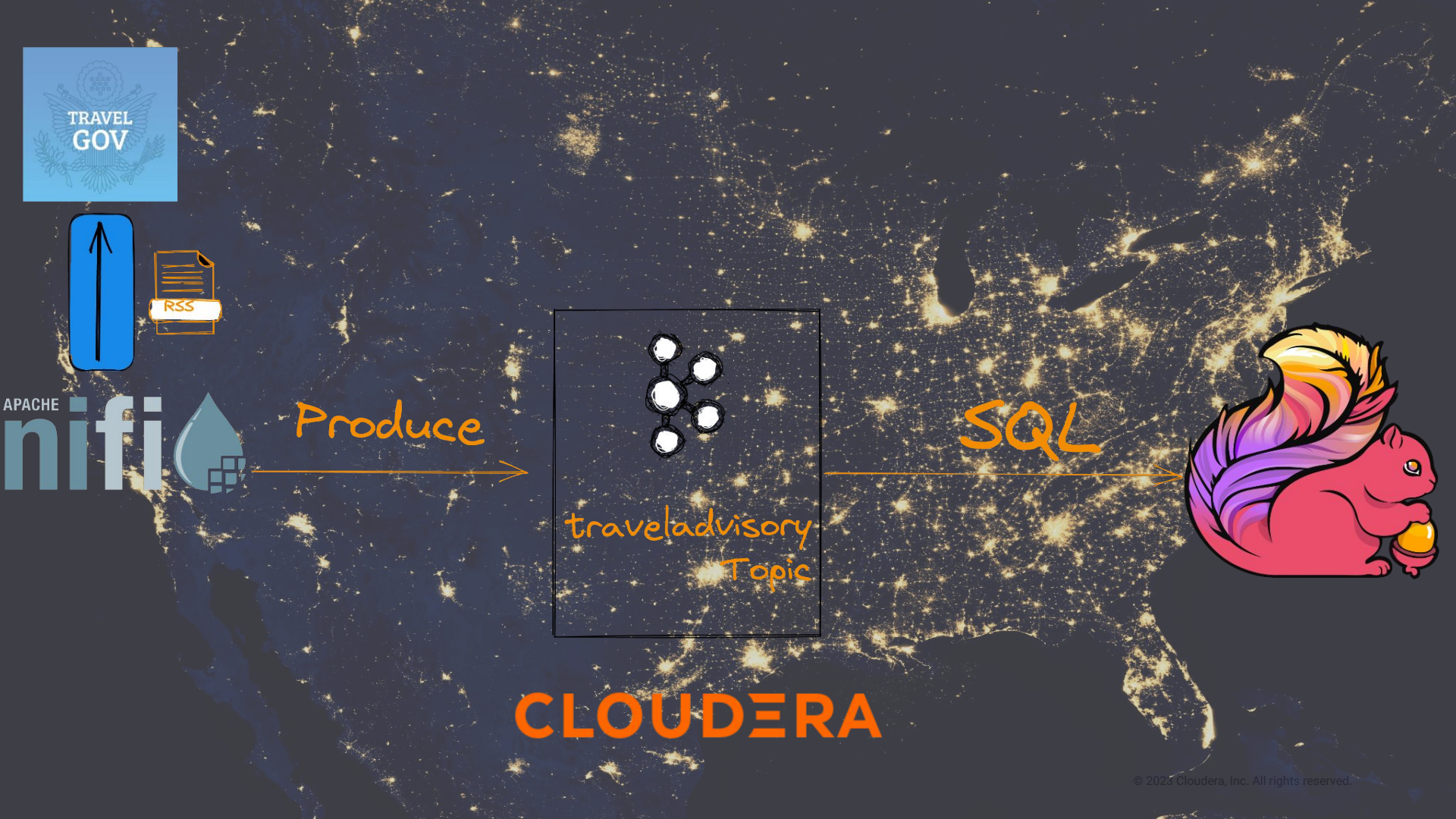
INGEST OF ALL TRANSIT DATA

Run collection and streaming on any cloud, server, container or VM





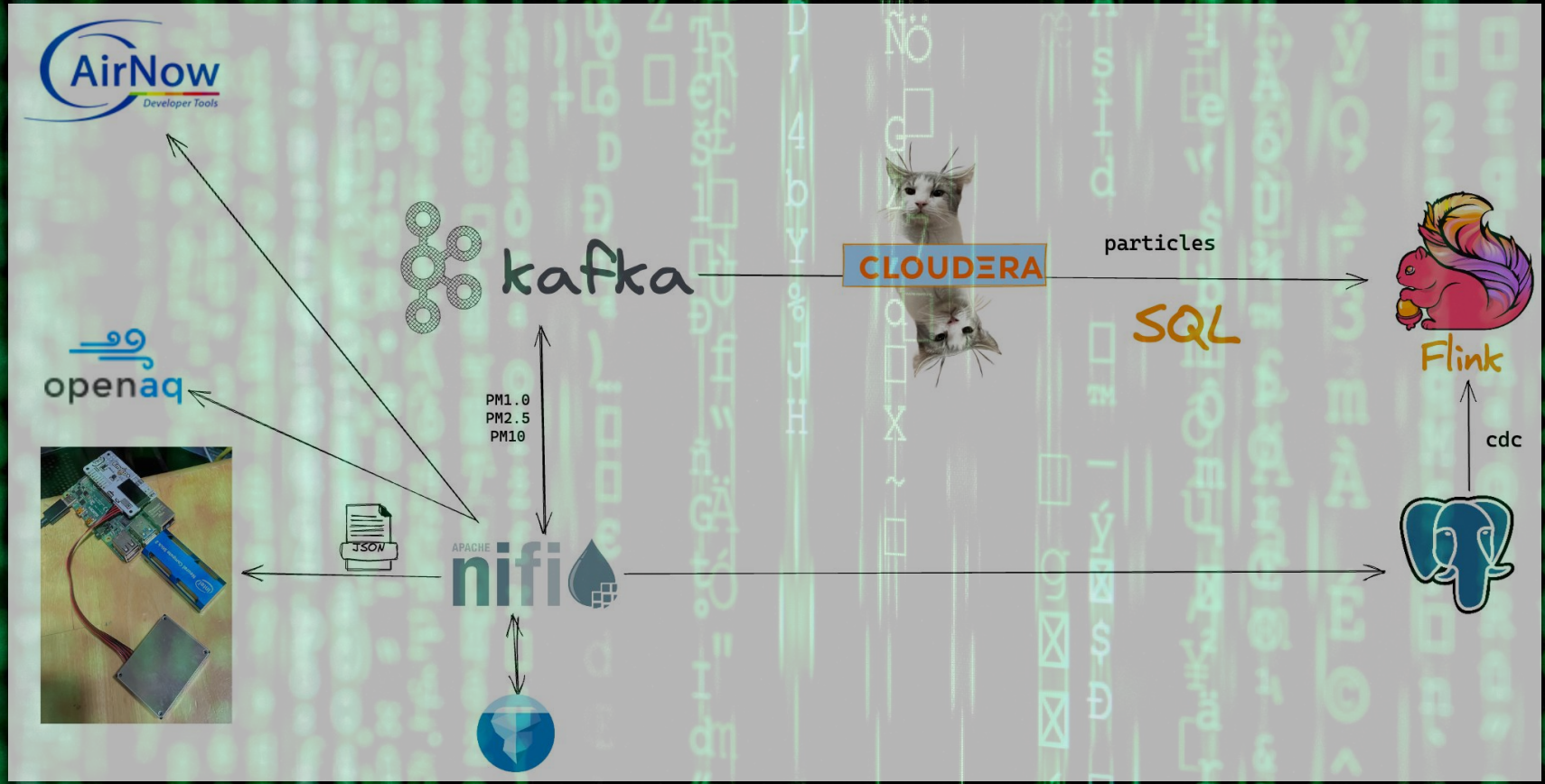

Metropolitan Transportation Authority
JSON
APACHE nifi
kafka
MTA Topics
SQL
FLaNK-MTA / Urban Transportation
EarthCam



CLOUDERA


AirNow
Developer Tools
APACHE
nifi
openaq

Particulate Matter Sensor Pipeline

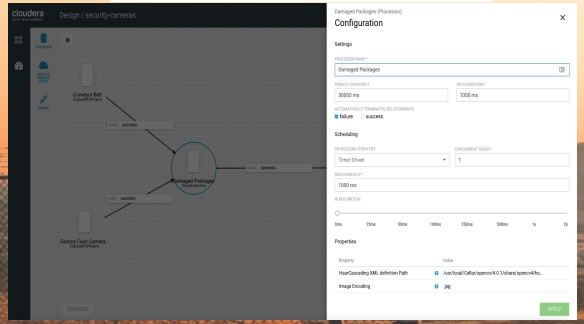


APACHE NiFi - MiNiFi Agents

Cloudera Edge Management

Edge device data collection and processing with easy to use central command and control

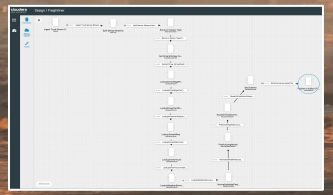
Edge Flow Manager



APACHE NIFI minifi

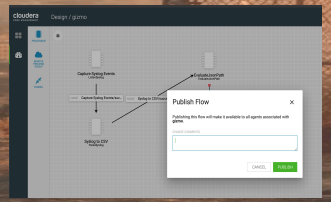
A lightweight edge agent that implements the core features of Apache NiFi, focusing on data collection and processing at the edge

Flow Authorship

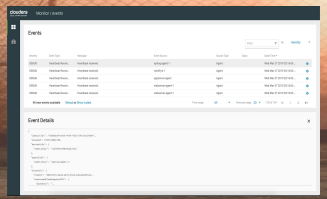


- Small footprint agent with MiNiFi
- Java and C++ agents
- Rich edge processors (edge collection & processing)
- End to end lineage and security
- Kubernetes support

Flow Deployment



Flow Monitoring



- Central Command and Control (C2)
- Design and deploy to millions of agents
- Edge Applications lifecycle management
- Multitenancy with Agent classes
- Native integration with other CDF services

CLOUDERA

APACHE KAFKA

Apache Kafka

MiNiFi

DATA COLLECTION AT THE EDGE

US-West Fleet

US-Central Fleet

US-East Fleet

Apache Kafka

INGEST GATEWAY POWERED BY KAFKA

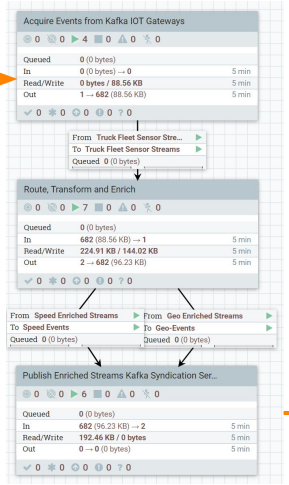
gateway-west-raw-sensors

gateway-central-raw-sensors

gateway-east-raw-sensors

Apache NiFi

DATA FLOW APPS POWERED BY NIFI



Apache Kafka

KAFKA CLUSTER GEO 1 DATA SYNDICATE SERVICES

Kafka Topic syndicate-transmission	Kafka Topic syndicate-speed
Kafka Topic syndicate-temp	Kafka Topic syndicate-geo

Replication / Data Deployment

KAFKA CLUSTER GEO 2 DATA SYNDICATE SERVICES

Kafka Topic syndicate-transmission	Kafka Topic syndicate-speed
Kafka Topic syndicate-temp	Kafka Topic syndicate-geo

Apache Flink

STREAMING ANALYTICS APPS

Micro Batch Analytics Stream Analytics App

Apache Spark Structured Streaming

Micro Services Stream Analytics App

kafka

Complex Low Latent Stream Analytics App

Apache Flink

APACHE FLINK

SQL STREAM BUILDER (SSB)

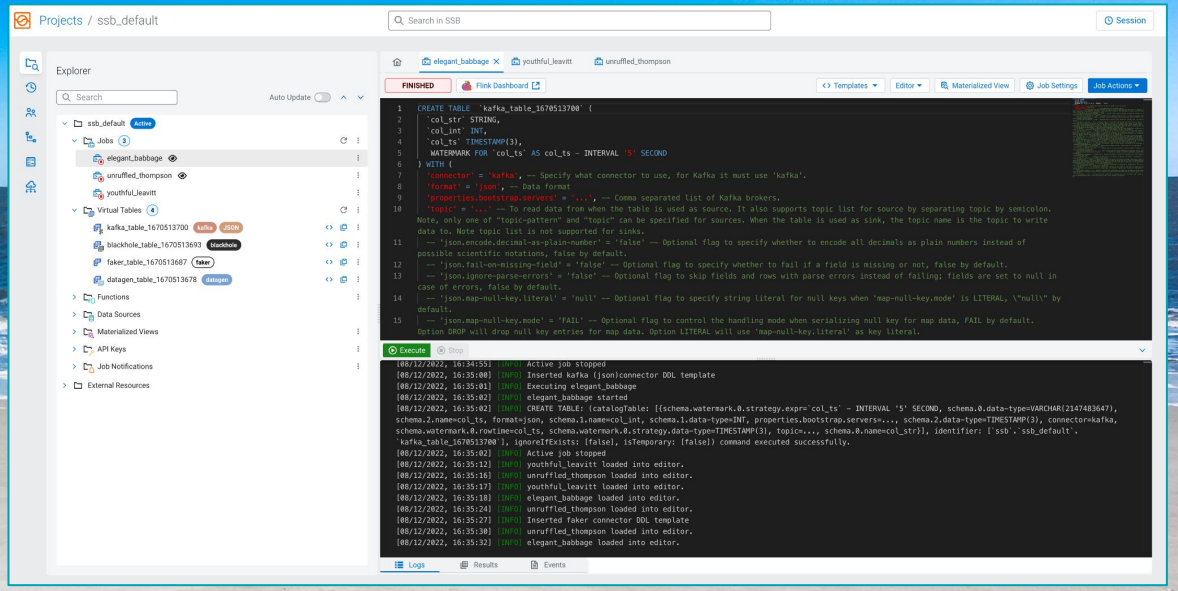
Democratize access to real-time data with just SQL

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

No Java or Scala code development required

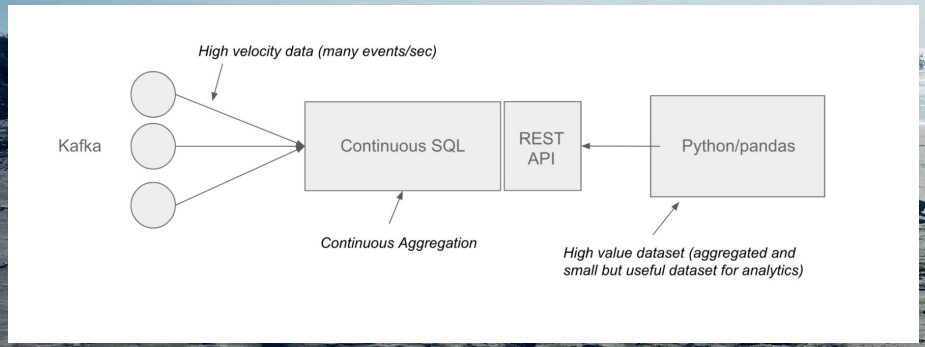
Simplifies access to data in Kafka & Flink. Connectors to batch data in HDFS, Kudu, Hive, S3, JDBC, CDC and more

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




```
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
```

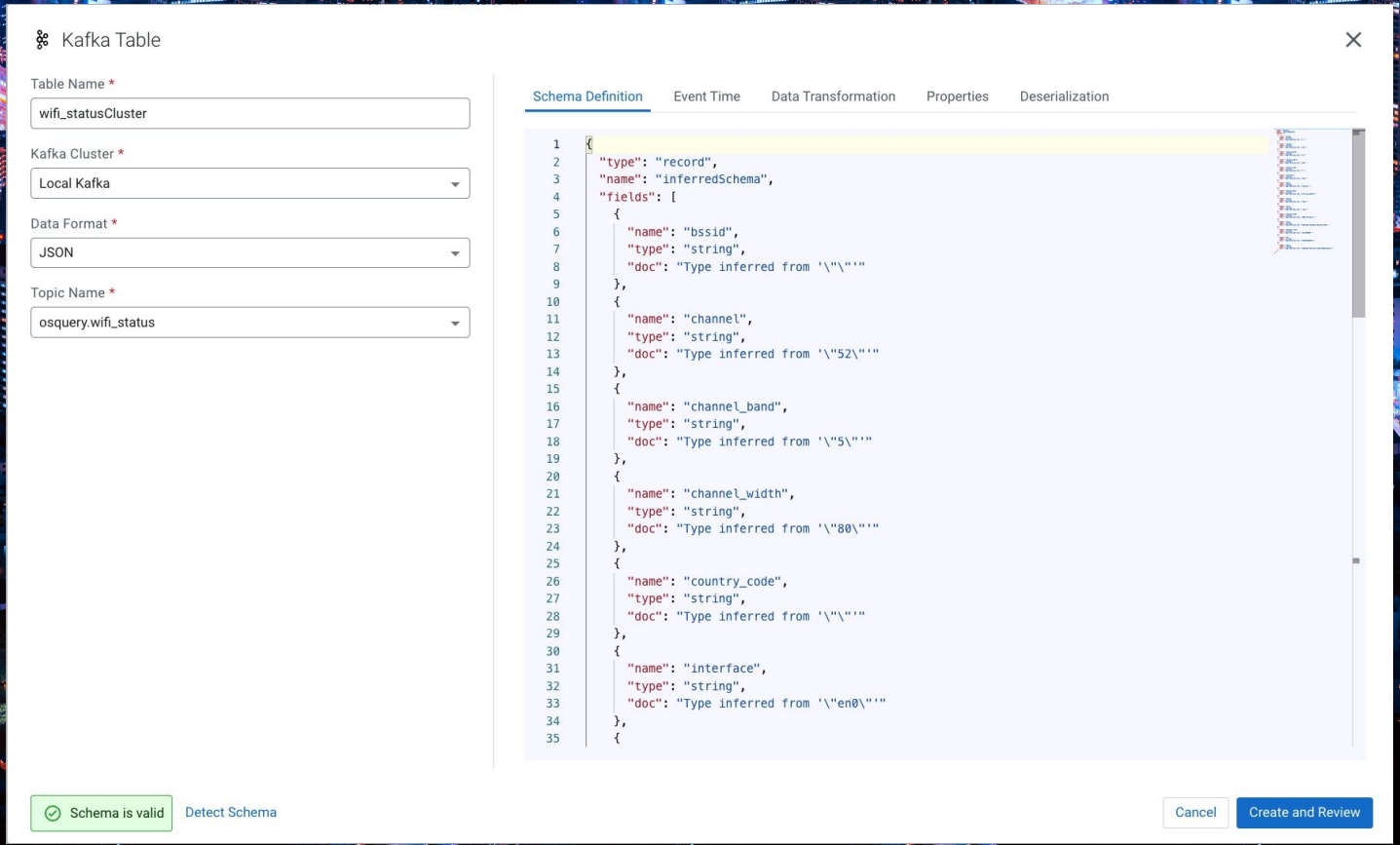
```
[90]: import pandas as pd
[91]: mv = "https://xxxyyyyzzz"
[92]: df = pd.read_json(mv)
[93]: len(df.keys())
[93]: 5
[95]: df['ts'] = pd.to_datetime(df['ts'])
[97]: df.dtypes
[97]: max_amount              int64
      sum_amount              int64
      thecount                int64
      ts       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

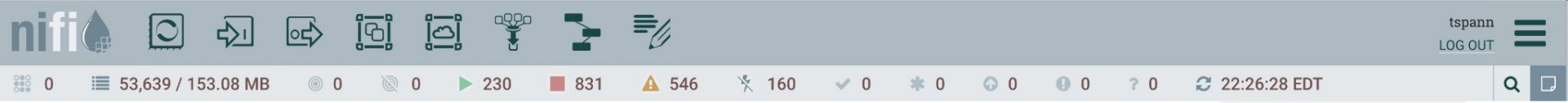
[98]:

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

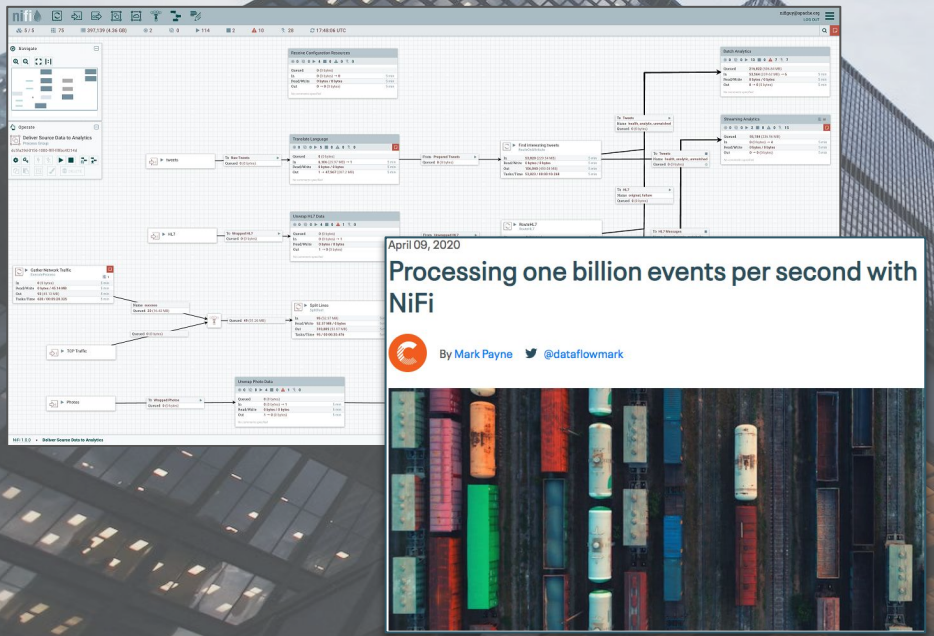




DATAFLOW APACHE NIFI

CLOUDERA DATAFLOW - 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

PROVENANCE

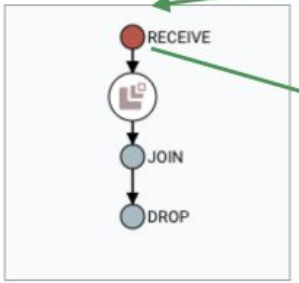
Displaying 13 of 104
Oldest event available: 11/15/2016 13:34:50 EST

Showing the most recent events.

ConsumeKafka | by component name

Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type
11/15/2016 13:35:03.8...	RECEIVE	379fc4f6-60e0-4151-9743-28...	44 bytes	ConsumeKafka	ConsumeKafka
11/15/2016 13:35:02.7...	RECEIVE	78f8c38b-89fc-4d00-a8d8-51...	44 bytes	ConsumeKafka	ConsumeKafka
11/15/2016 13:35:01.6...	RECEIVE	2bcd5124-bb78-489f-ad8a-7...	44 bytes	ConsumeKafka	ConsumeKafka

- Tracks data at each point as it flows through the system
- Records, indexes, and makes events available for display
- Handles fan-in/fan-out, i.e. merging and splitting data
- View attributes and content at given points in time



Provenance Event

DETAILS | ATTRIBUTES | CONTENT

Attribute Values

filename
328717796819631
No value previously set

kafka.offset
44815
No value previously set

kafka.partition
6
No value previously set

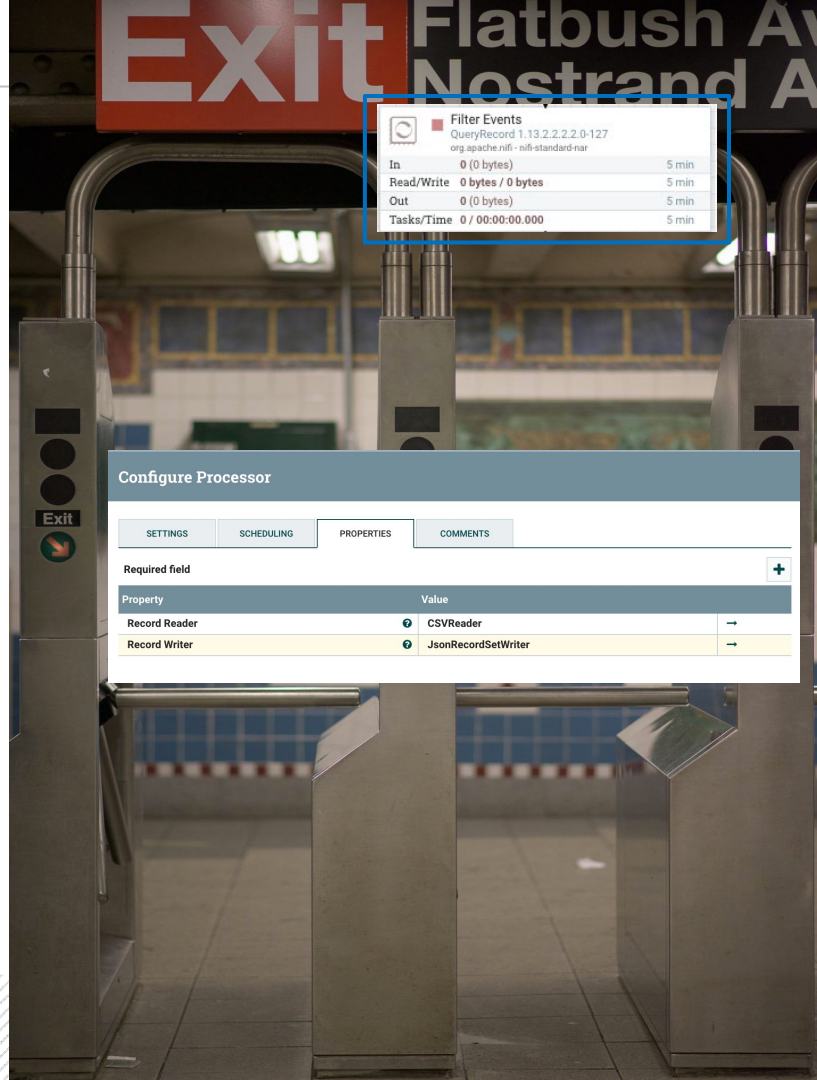
kafka.topic
nifi-testing
No value previously set

path
./
No value previously set

uuid

RECORD-ORIENTED DATA WITH NIFI

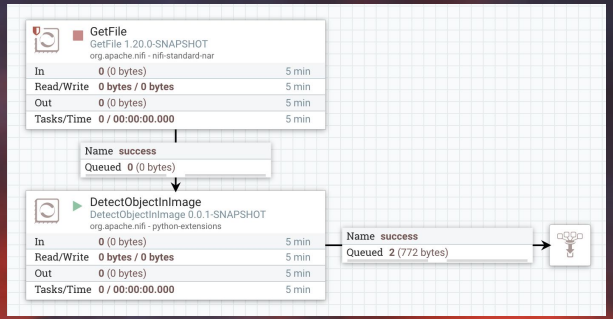
- **Record Readers** - Avro, CSV, Grok, IPFIX, JSAN1, JSON, Parquet, Scripted, Syslog5424, Syslog, WindowsEvent, XML
- **Record Writers** - Avro, CSV, FreeFromText, Json, Parquet, Scripted, XML
- Record Reader and Writer support referencing a schema registry for retrieving schemas when necessary.
- Enable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



APACHE NIFI WITH PYTHON CUSTOM PROCESSORS

Python as a First Class Citizen

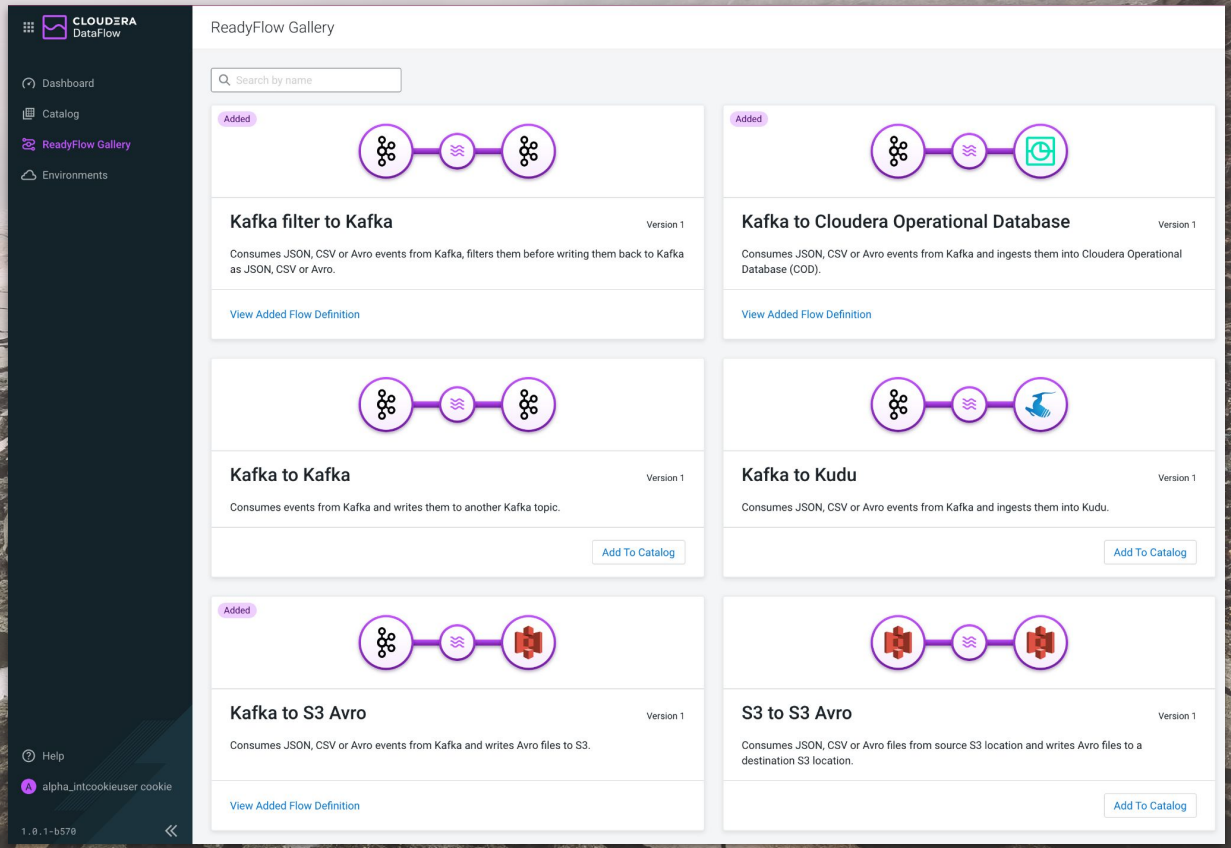
```
import cv2
import numpy as np
import json
from nifiapi.properties import PropertyDescriptor
from nifiapi.properties import ResourceDefinition
from nifiapi.flowfiletransform import FlowFileTransformResult

SCALE_FACTOR = 0.00392
NMS_THRESHOLD = 0.4     # non-maximum suppression threshold
CONFIDENCE_THRESHOLD = 0.5

class DetectObjectInImage:
    class Java:
        implements = ['org.apache.nifi.python.processor.FlowFileTransform']
    class ProcessorDetails:
        version = '0.0.1-SNAPSHOT'
        dependencies = ['numpy >= 1.23.5', 'opencv-python >= 4.6']

    def __init__(self, jvm=None, **kwargs):
        self.jvm = jvm

        # Build Property Descriptors
        self.model_file = PropertyDescriptor(
            name = 'Model File',
            description = 'The binary file containing the trained Deep Neural Network weights. Supports Caffe (*.caffemodel), TensorFlow (*.pb), Torch (*.t7, *.net), Darknet (*.weights), ' +
                          'DLDT (*.bin), and ONNX (*.onnx)',
            required = True,
            resource_definition = ResourceDefinition(allow_file = True)
        )
        self.config_file = PropertyDescriptor(
            name = 'Network Config File',
            description = 'The text file containing the Network configuration. Supports Caffe (*.prototxt), TensorFlow (*.pbtxt), Darknet (*.cfg), and DLDT (*.xml)',
            required = False,
            resource_definition = ResourceDefinition(allow_file = True)
        )
        self.class_name_file = PropertyDescriptor(
            name = 'Class Names File',
            description = 'A text file containing the names of the classes that may be detected by the model. Expected format is one class name per line, new-line terminated."',
            required = True,
            resource_definition = ResourceDefinition(allow_file = True)
        )
        self.descriptors = [self.model_file, self.config_file, self.class_name_file]

    def getPropertyDescriptors(self):
        return self.descriptors

    def onScheduled(self, context):
        # read class names from text file
        class_name_file = context.getProperty(self.class_name_file.name).getValue()
        if class_name_file is None:
```



https://github.com/apache/nifi/blob/614947e4ac6798ad80817e82514c39349d5faacb/nifi-docs/src/main/asciidoc/python-developer-guide.adoc

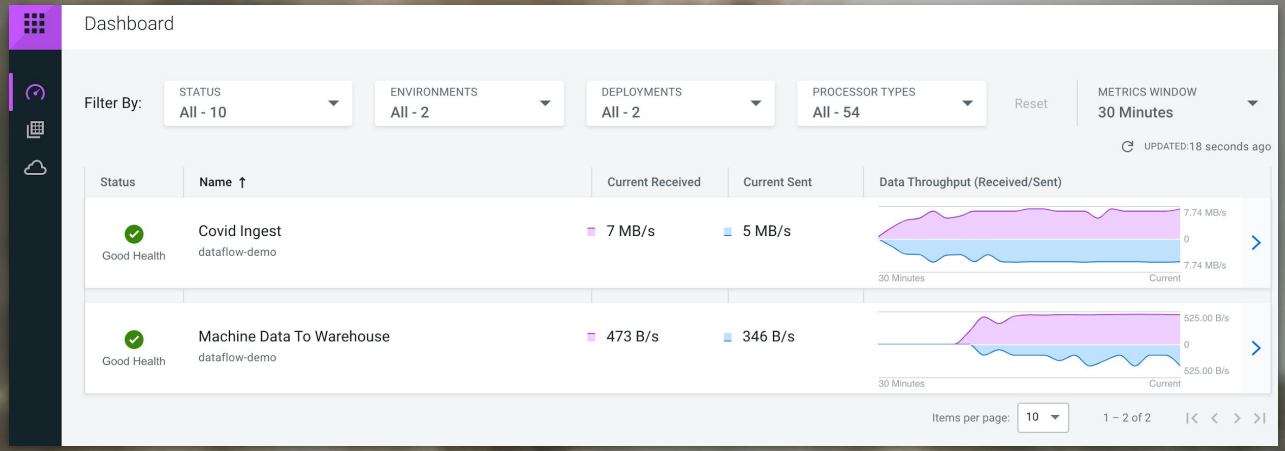
READYFLOW GALLERY

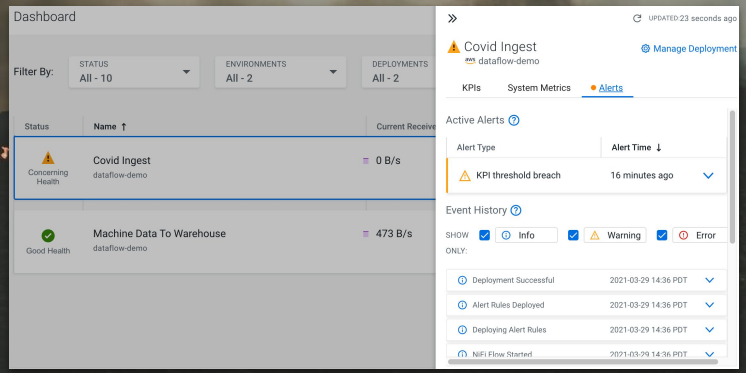
- Cloudera provided flow definitions
- Cover most common data flow use cases
- Optimized to work with CDP sources/destinations
- Can be deployed and adjusted as needed

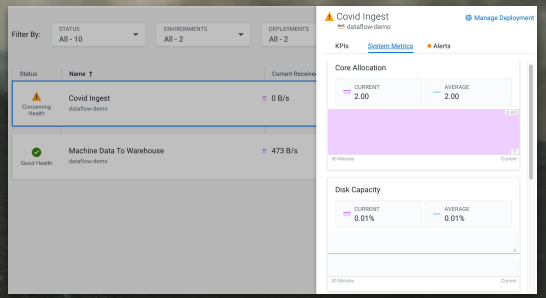


DASHBOARD

- 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







RESOURCES AND WRAP-UP

Future of Data - Princeton + Virtual

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

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 ...

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY

ICEBERG

@PaasDev

FLaNK Stack Weekly

https://bit.ly/32dAJft

This week in Apache NiFi, Apache Flink, Apache Kafka, Apache Spark, Apache Iceberg, Python, Java, AI, ML, LLM and Open Source friends.

CSP Community Edition

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

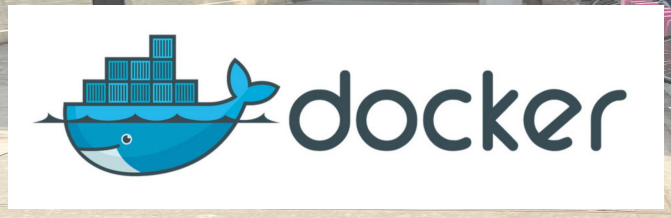
- Docker compose file of CSP to run from command line w/o any dependencies, including Flink, SQL Stream Builder, Kafka, Kafka Connect, Streams Messaging Manager and Schema Registry
  - `$> docker compose up`
- Licensed under the Cloudera Community License
- Unsupported
- Community Group Hub for CSP
-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

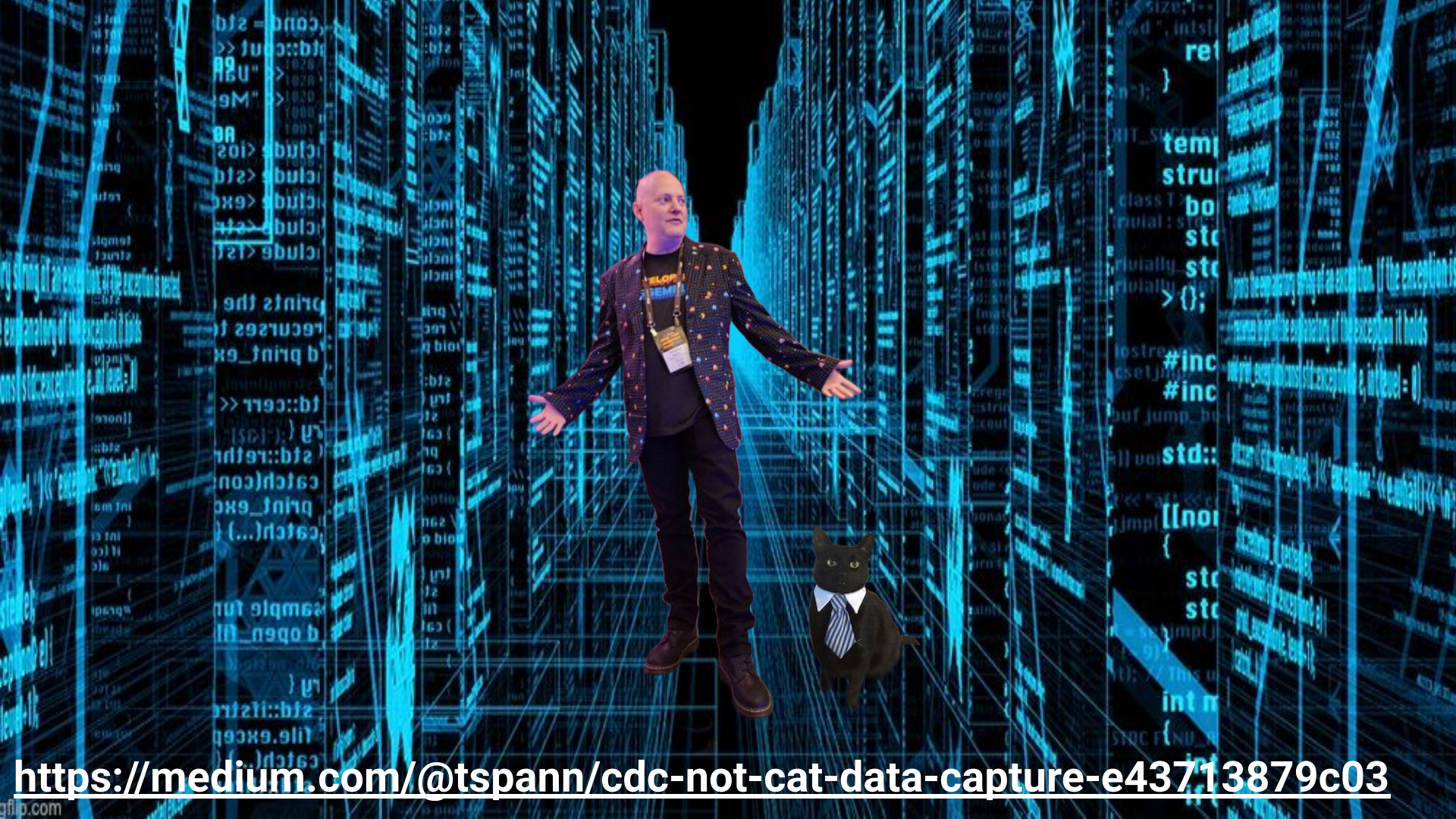
- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

- Licensed under the ASF License
- Unsupported

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

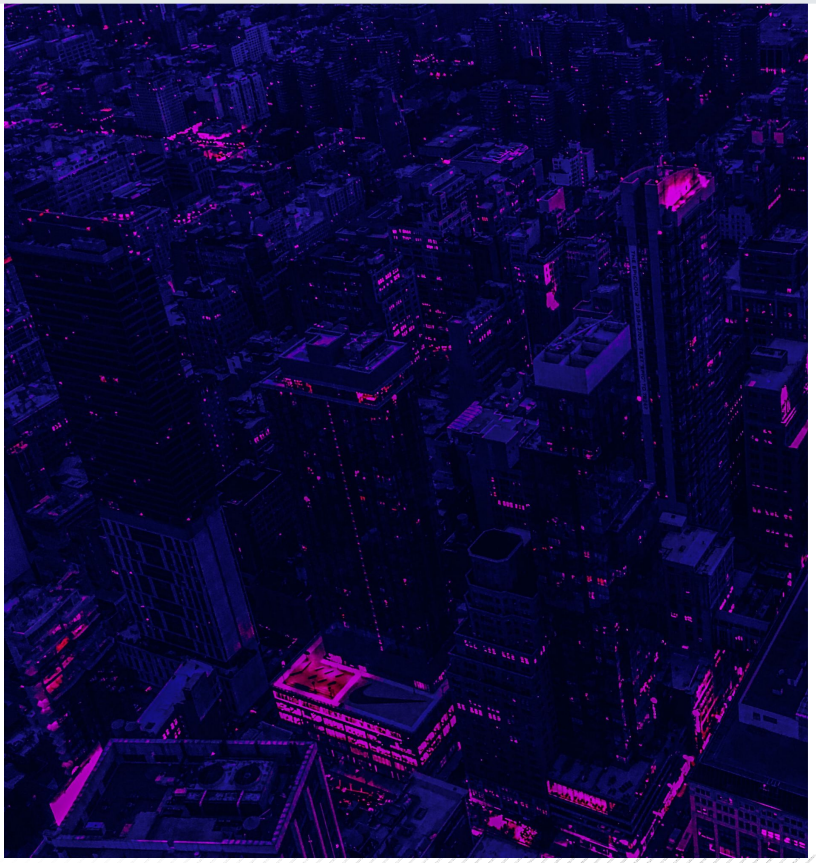
Resources

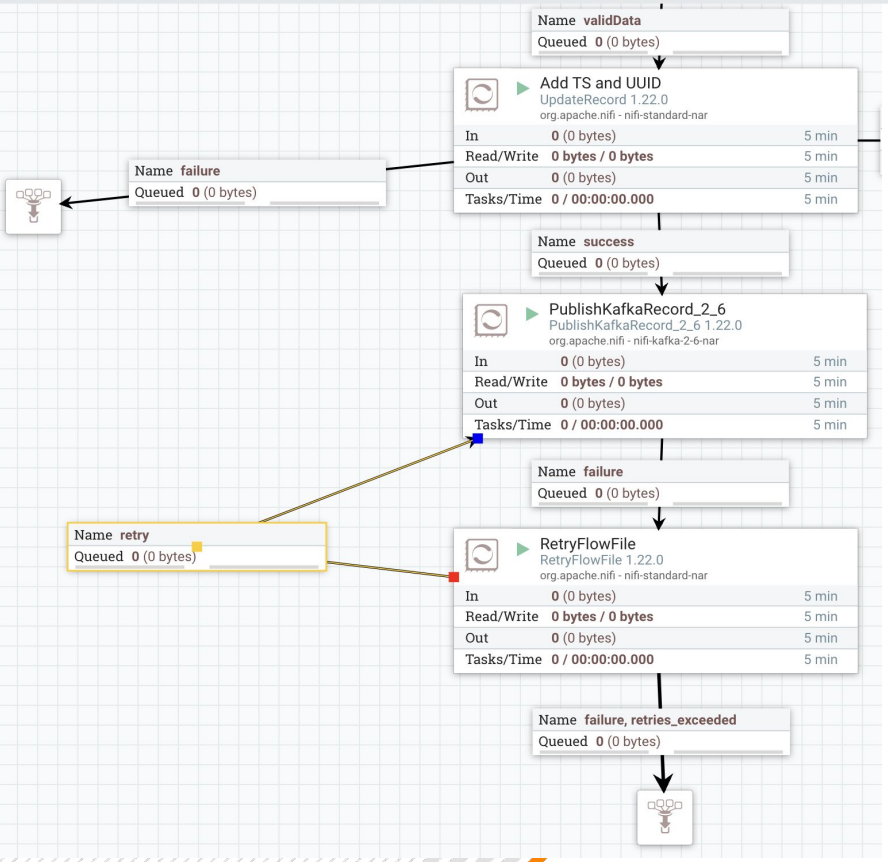
threads.net/@tspannhw

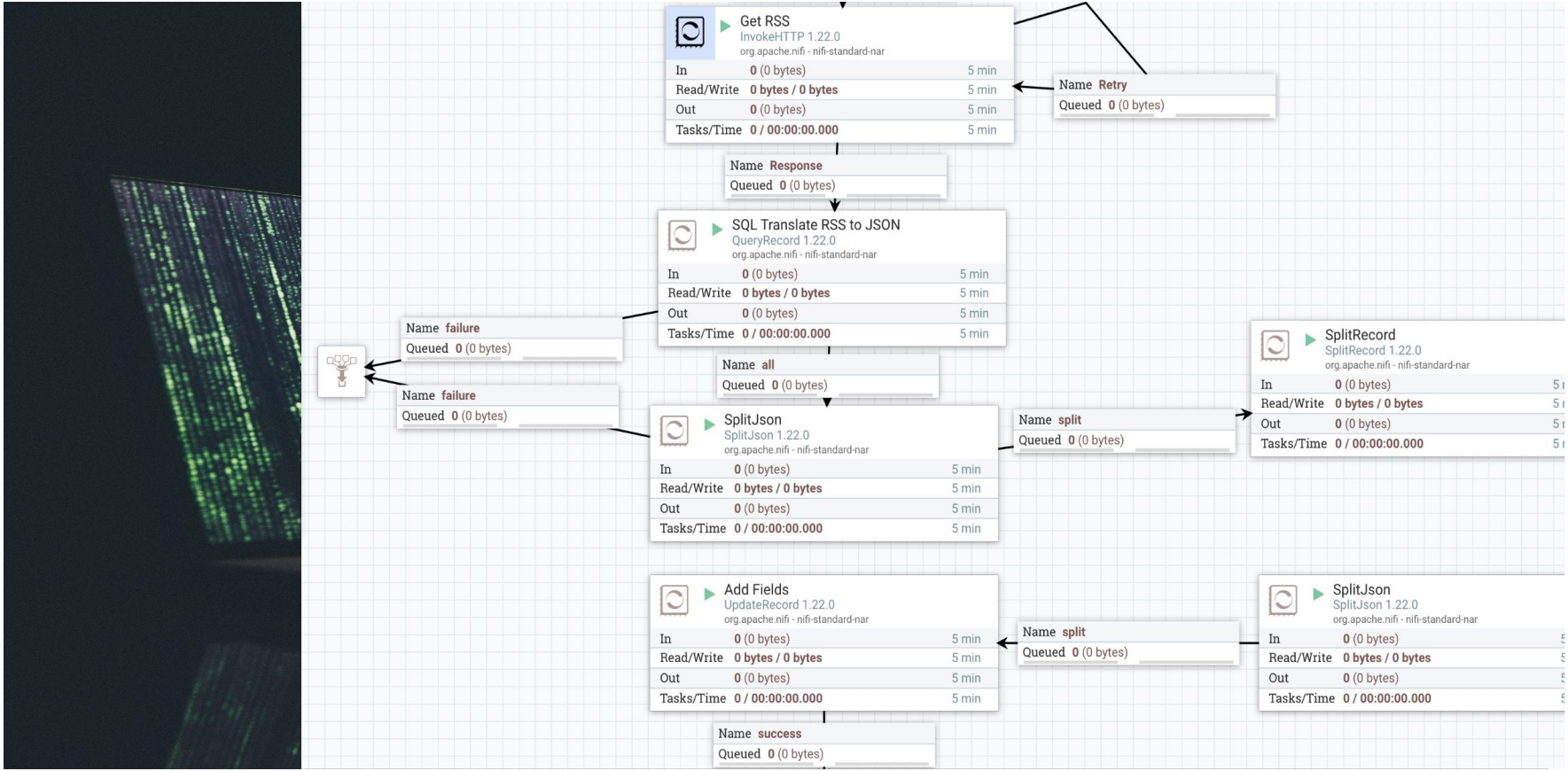

https://medium.com/@tspann/cdc-not-cat-data-capture-e43713879c03

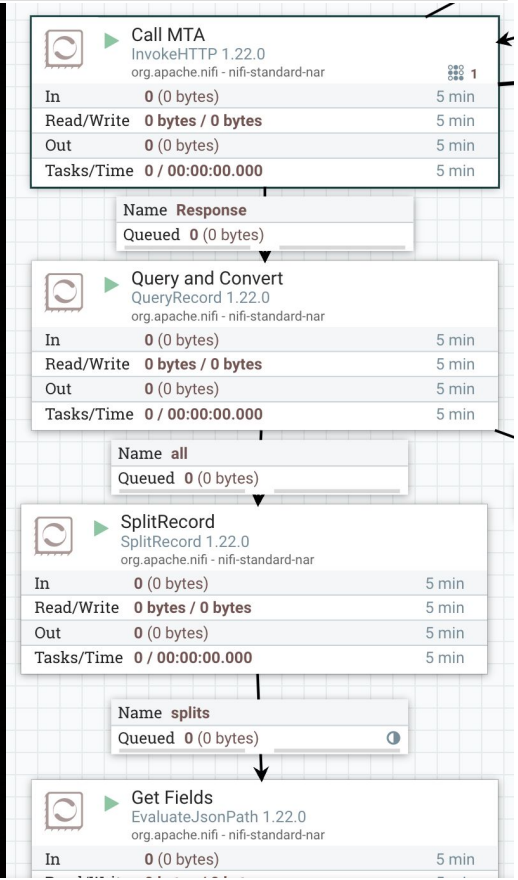

COLLECTING DATA WITH NIFI
DATA INC
imgflip.com

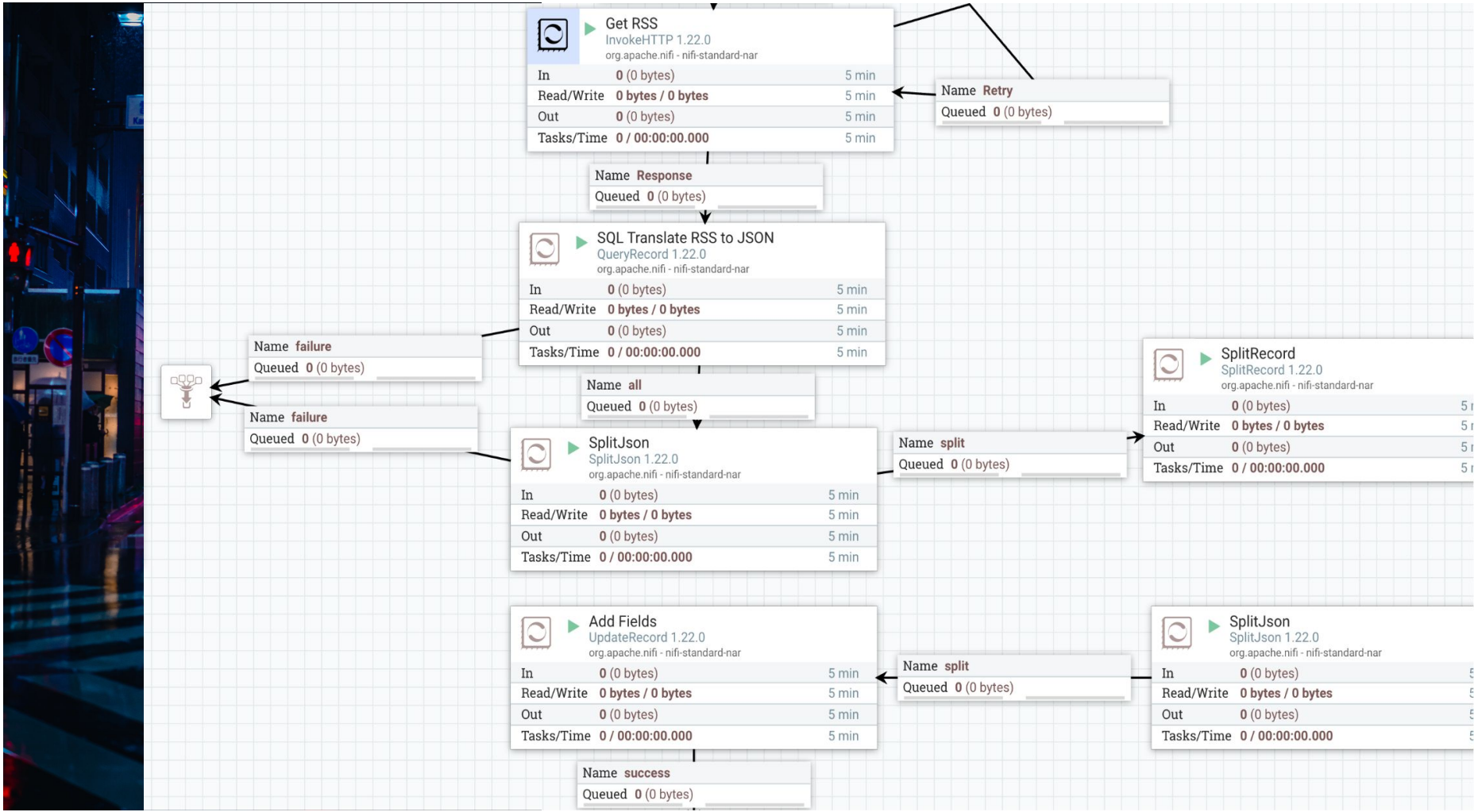
WALK THROUGH

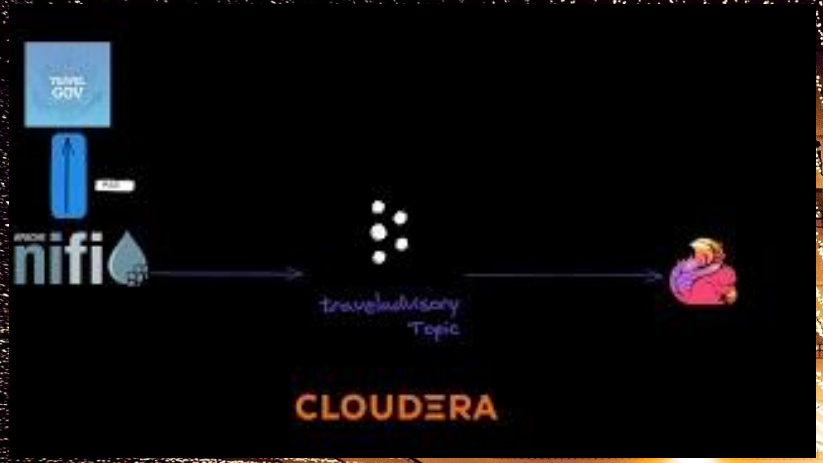









Get RSS
InvokeHTTP 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name Retry
Queued 0 (0 bytes)
Name Response
Queued 0 (0 bytes)
SQL Translate RSS to JSON
QueryRecord 1.22.0
org.apache.nifi - nifi-standard-nar
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
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)
Name all
Queued 0 (0 bytes)
SplitRecord
SplitRecord 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes)
Read/Write 0 bytes / 0 bytes
Out 0 (0 bytes)
Tasks/Time 0 / 00:00:00.000
SplitJson
SplitJson 1.22.0
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 split
Queued 0 (0 bytes)
Add Fields
UpdateRecord 1.22.0
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 split
Queued 0 (0 bytes)
SplitJson
SplitJson 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes)
Read/Write 0 bytes / 0 bytes
Out 0 (0 bytes)
Tasks/Time 0 / 00:00:00.000
Name success
Queued 0 (0 bytes)



GetTravel Advisories
InvokeHTTP 1.20.0.2.3.8.1-1

IN	0 (0.00 B)
READ/WRITE	0.00 B / 0.00 B
OUT	0 (0.00 B)
TASKS	0 / 00:00:00.000

5 min

NAME	Response
QUEUED	0 (0.00 B)

Convert RSS XML To JSON
QueryRecord 1.20.0.2.3.8.1-1

IN	0 (0.00 B)
READ/WRITE	0.00 B / 0.00 B
OUT	0 (0.00 B)
TASKS	0 / 00:00:00.000

5 min

NAME	all
QUEUED	0 (0.00 B)

NAME	failure
QUEUED	0 (0.00 B)

SplitJson
SplitJson 1.20.0.2.3.8.1-1

IN	0 (0.00 B)
READ/WRITE	0.00 B / 0.00 B
OUT	0 (0.00 B)
TASKS	0 / 00:00:00.000

5 min

NAME	failure
QUEUED	0 (0.00 B)

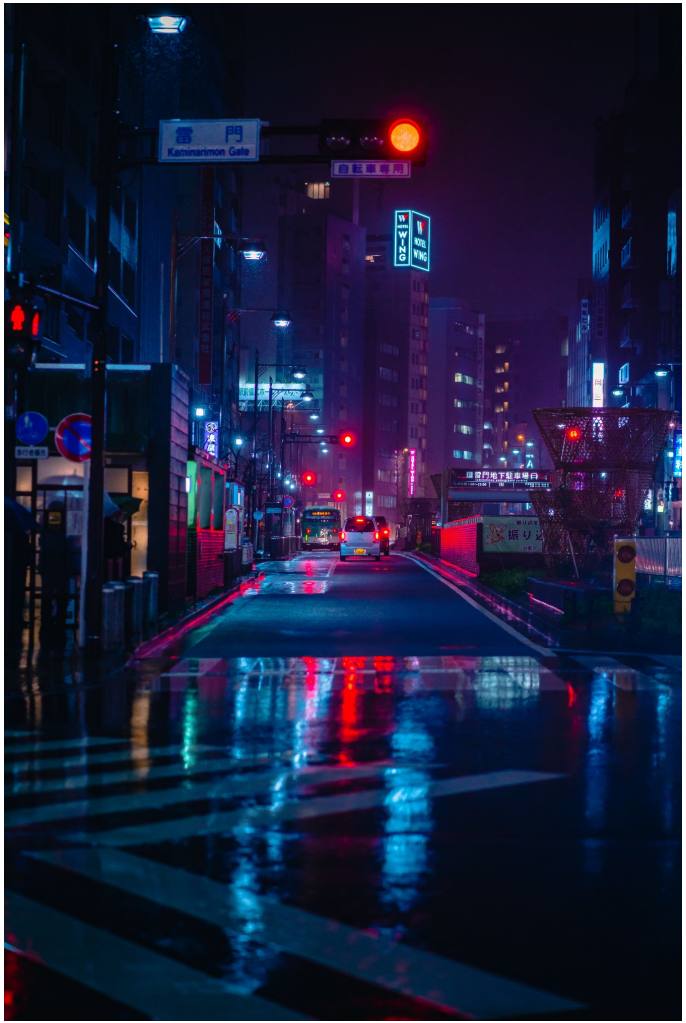
NAME	split
QUEUED	0 (0.00 B)

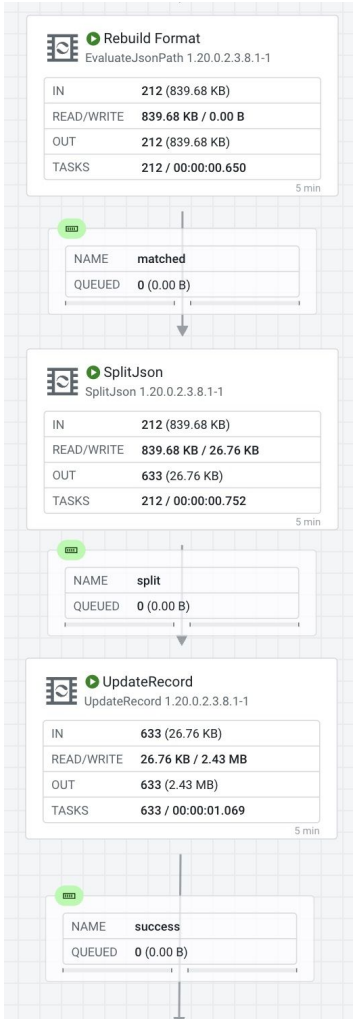
Funnel

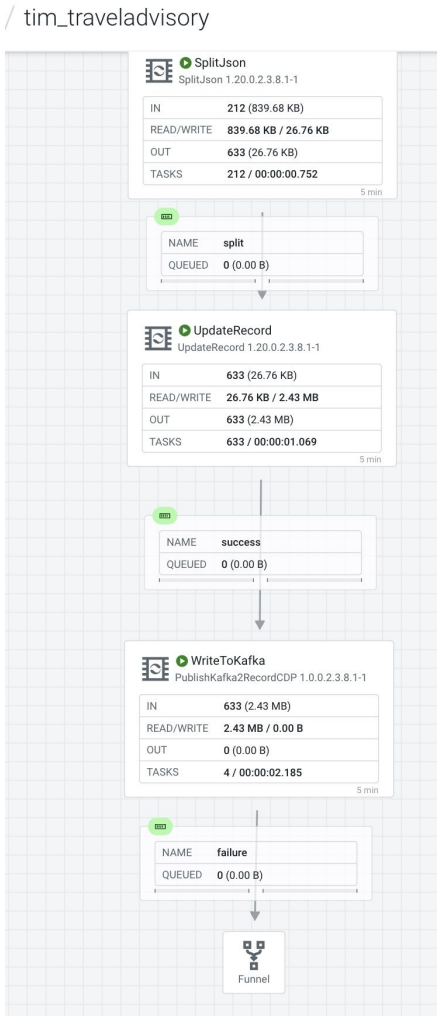
QUEUED	0 (0.00 B)

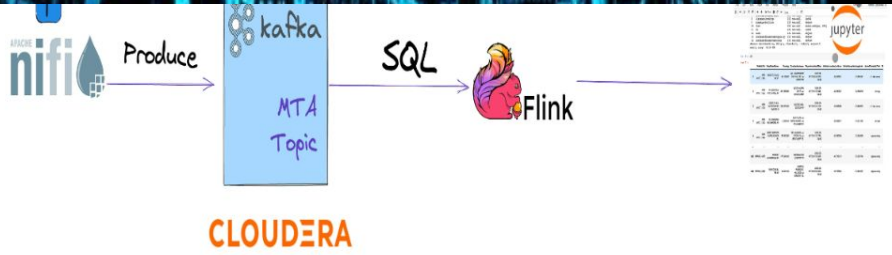
SplitRecord
SplitRecord 1.20.0.2.3.8.1-1

IN	0 (0.00 B)
READ/WRITE	0.00 B / 0.00 B
OUT	0 (0.00 B)


雷門
Kaminarimon Gate
WING
ROYAL WING


Rebuild Format
EvaluateJsonPath 1.20.0.2.3.8.1-1
IN 212 (839.68 KB)
READ/WRITE 839.68 KB / 0.00 B
OUT 212 (839.68 KB)
TASKS 212 / 00:00:00.650
5 min
NAME matched
QUEUED 0 (0.00 B)
SplitJson
SplitJson 1.20.0.2.3.8.1-1
IN 212 (839.68 KB)
READ/WRITE 839.68 KB / 26.76 KB
OUT 633 (26.76 KB)
TASKS 212 / 00:00:00.752
5 min
NAME split
QUEUED 0 (0.00 B)
UpdateRecord
UpdateRecord 1.20.0.2.3.8.1-1
IN 633 (26.76 KB)
READ/WRITE 26.76 KB / 2.43 MB
OUT 633 (2.43 MB)
TASKS 633 / 00:00:01.069
5 min
NAME success
QUEUED 0 (0.00 B)


/ tim_traveladvisory
SplitJson
SplitJson 1.20.0.2.3.8.1-1
IN 212 (839.68 KB)
READ/WRITE 839.68 KB / 26.76 KB
OUT 633 (26.76 KB)
TASKS 212 / 00:00:00.752
5 min
NAME split
QUEUED 0 (0.00 B)
UpdateRecord
UpdateRecord 1.20.0.2.3.8.1-1
IN 633 (26.76 KB)
READ/WRITE 26.76 KB / 2.43 MB
OUT 633 (2.43 MB)
TASKS 633 / 00:00:01.069
5 min
NAME success
QUEUED 0 (0.00 B)
WriteToKafka
PublishKafka2RecordCDP 1.0.0.2.3.8.1-1
IN 633 (2.43 MB)
READ/WRITE 2.43 MB / 0.00 B
OUT 0 (0.00 B)
TASKS 4 / 00:00:02.185
5 min
NAME failure
QUEUED 0 (0.00 B)
Funnel



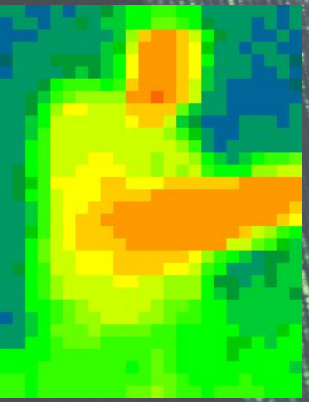
Show 10 entries

Search:

Vehicle	Stop	Destination	Expected Arrival	How Close?	Distance Away	# of Passengers
MTA NYCT_9686	4 AV/E 10 ST	EAST VILLAGE 8 ST via 5 AV	2023-07-11T14:53:00.000-04:00	approaching	118	
MTA NYCT_9651	MURRAY ST/CHURCH ST	BATTERY PK CITY via AVENUE C via PK ROW	2023-07-11T14:51:10.862-04:00	< 1 stop away	351	5
MTA NYCT_9639	FT WASHINGTON AV/W 168 ST	MIDTOWN 32 ST & 5 AV via BWAY 5 AV	2023-07-11T14:51:06.076-04:00	approaching	148	33
MTA NYCT_8387	W 231 ST/BROADWAY	NORWOOD 205 ST STA	2023-07-11T14:50:22.856-04:00	< 1 stop away	287	20
MTA NYCT_9649	CENTRE ST/CHAMBERS ST	BATTERY PK CTY VESEY ST	2023-07-11T14:50:09.630-04:00	< 1 stop away	317	14
MTA NYCT_7890	PENNSYLVANIA AV/PITKIN AV	SPRING CREEK GATEWAY MALL	2023-07-11T14:48:55.081-04:00	< 1 stop away	349	19
MTA NYCT_8382	OGDEN AV/W 162 ST	PLIMPTON E.L. GRANT HWY	2023-07-11T14:48:53.459-04:00	< 1 stop away	346	47
MTA NYCT_8350	KISSENA BL/45 AV	FLUSHING MAIN ST STATION	2023-07-11T14:48:51.609-04:00	< 1 stop away	213	17
MTA NYCT_8331	ARCHER AV/SUTPHIN PL	JAMAICA LIRR	2023-07-11T14:48:47.233-04:00	approaching	150	
MTA NYCT_8388	BROADWAY/W 178 ST	MANHTTNVILLE W 125 via BWAY via AMSTRDM	2023-07-11T14:48:44.718-04:00	< 1 stop away	161	21
Vehicle	**Stop**	**Destination**	**Expected Arrival**	**How Close?**	**Distance Away**	**# of Passengers**

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 … 10 Next



THANK YOU