



Big Data Conference: Building Real-Time Travel Alerts

Tim Spann
Principal Developer Advocate

22-Nov-2023





CLOUDERA



CLOUDERA



EDGE
2AI

CLOUDERA





Tim Spann

Twitter: @PaasDev // Blog: datainmotion.dev

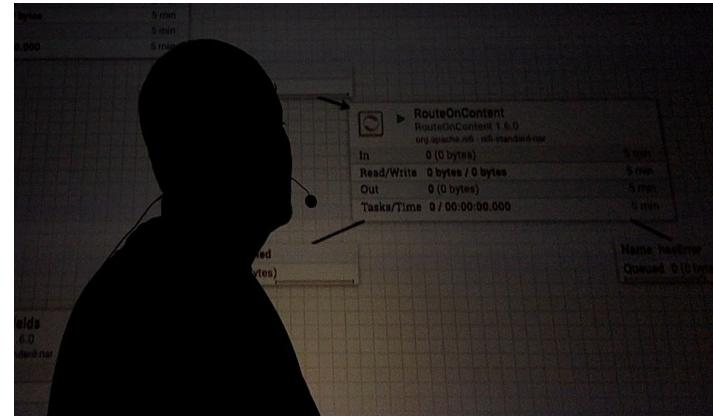
Principal Developer Advocate.

Princeton Future of Data Meetup.

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

<https://medium.com/@tspann>

<https://github.com/tspannhw>



A screenshot of the DZone website showing Tim Spann's profile under the "Top IoT Experts" section. It includes his photo, name, title ("Principal Developer Advocate, Cloudera"), and a link to his GitHub profile.



FLaNK Stack Weekly by Tim Spann



<https://bit.ly/32dAJft>

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



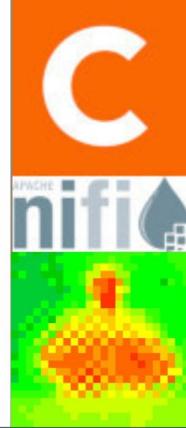
This week in Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java and Open Source friends.

Future of Data - NYC + NJ + Philly + 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 ...



@PaasDev

CODE / DEEP DIVE



DESCRIPTION

In this session, we will walk through how to build a complete streaming application to send alerts based on travel advisories from public data. We will also join in other data sources of relevance and push out alerts.

We will show you how to build this streaming application with Apache NiFi, Apache Kafka, and Apache Flink and show you when/why/how, and what to build to maximize performance, productivity, and ease of development.

Let's get streaming.



Introduction

Overview

Examples

Apache Kafka

Apache Flink

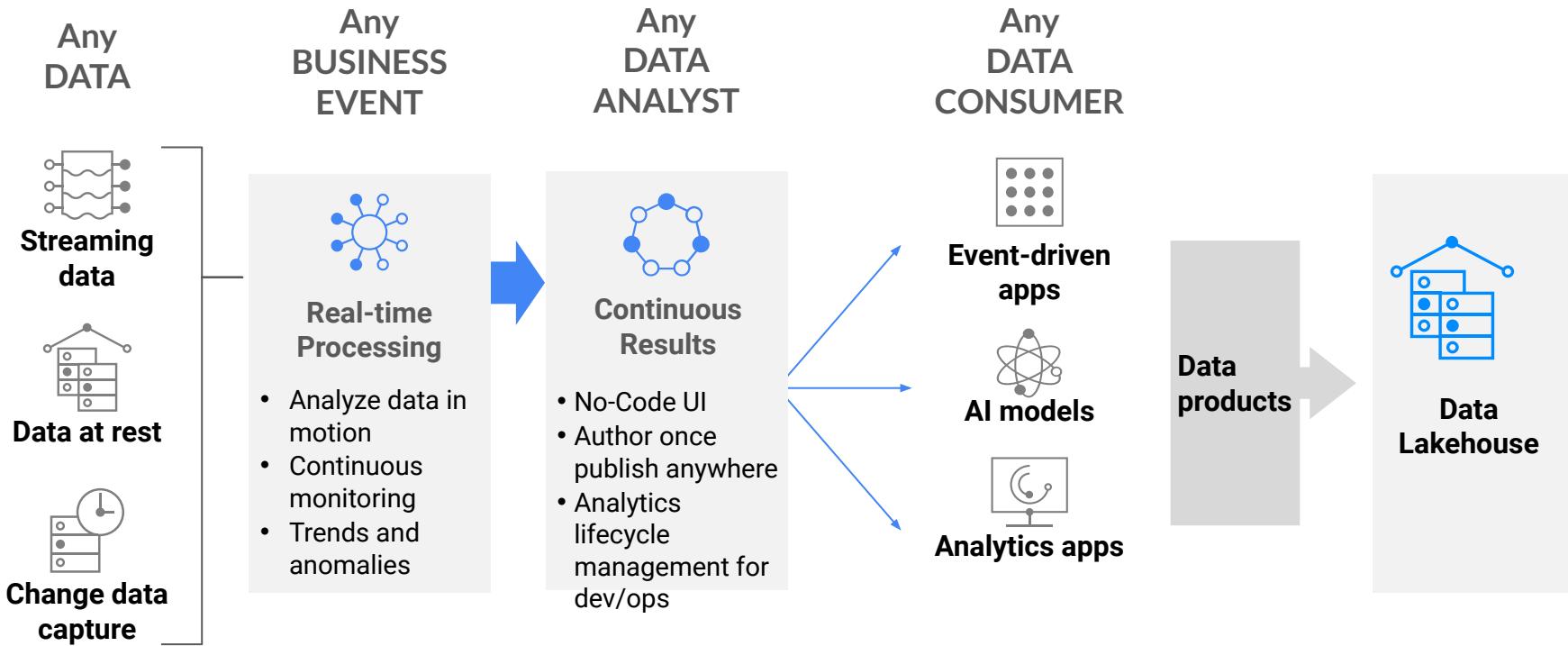
Apache NiFi

Demos

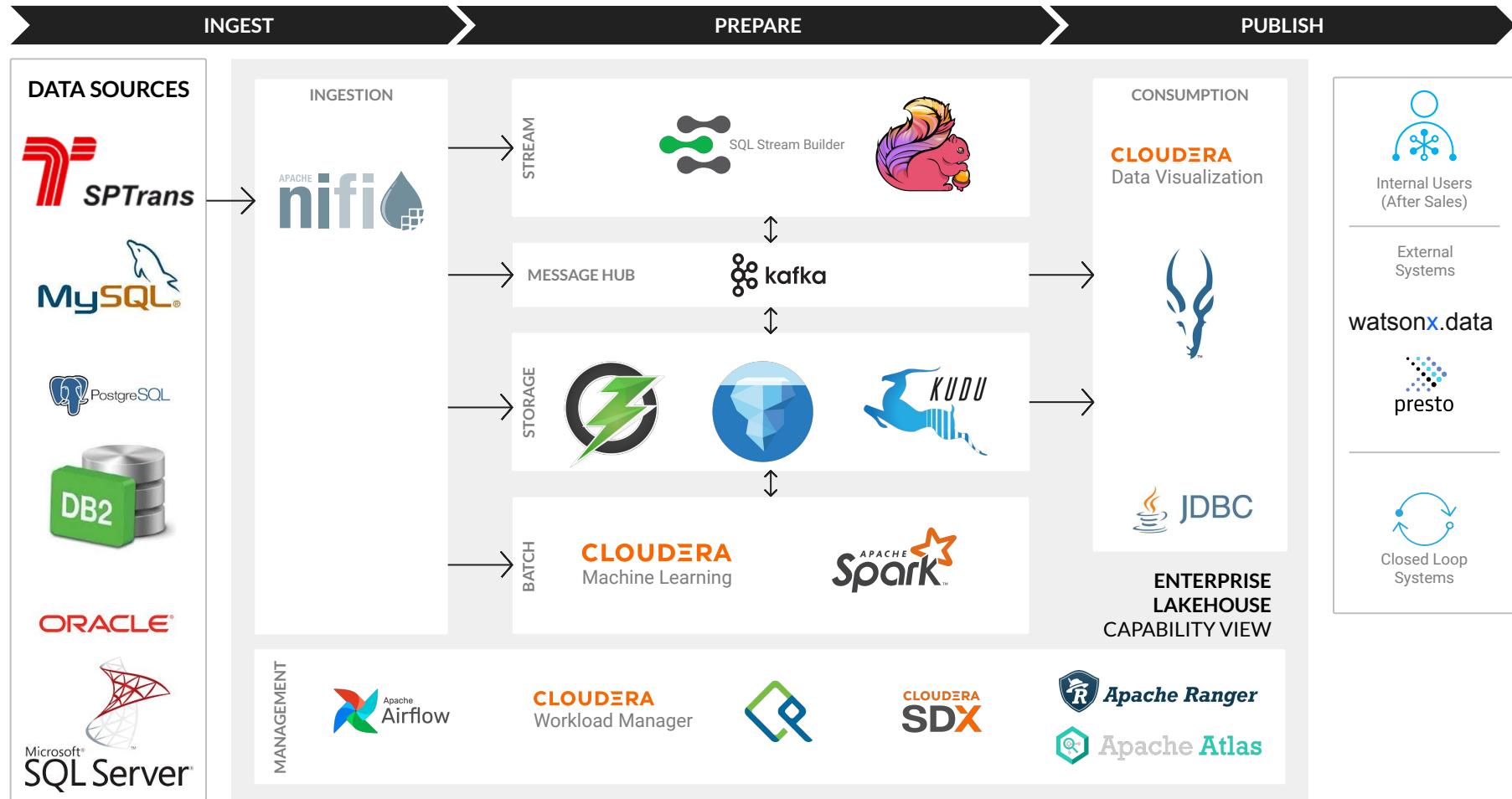
Trains, Planes and Automobiles +++

Information Needed	Data Feed(s)
Local weather conditions	<ul style="list-style-type: none">• XML, JSON, RSS
Mass transit status & alerts	<ul style="list-style-type: none">• XML, JSON, RSS
Regional highways & tunnels	<ul style="list-style-type: none">• GeoRSS, XML, ProtoBuf, JSON
Local social media	<ul style="list-style-type: none">• JSON
ADS-B Plane Data	<ul style="list-style-type: none">• JSON
Local air quality	<ul style="list-style-type: none">• JSON

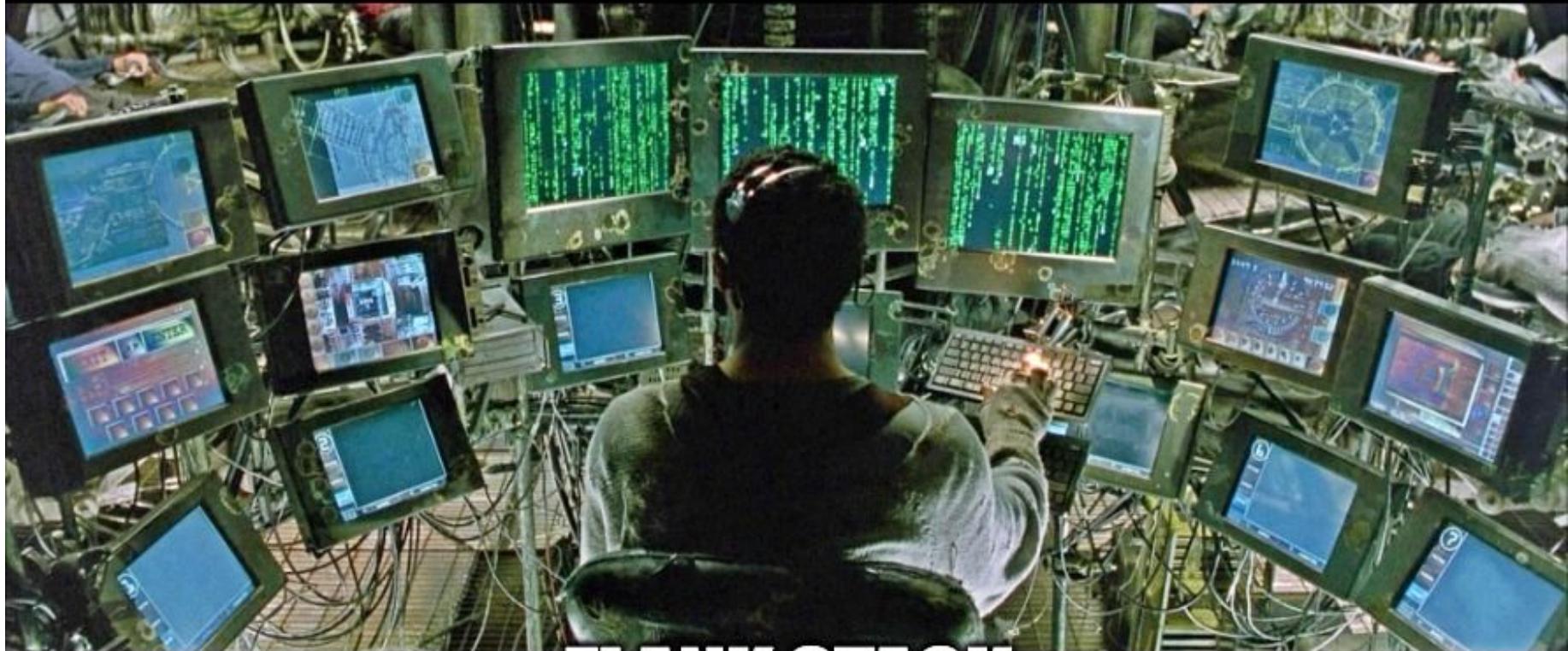
Data Relevance



REST API ARCHITECTURE - Using FLaNK to pull the data out of anything in near-real time



ALL THE TRANSIT DATA

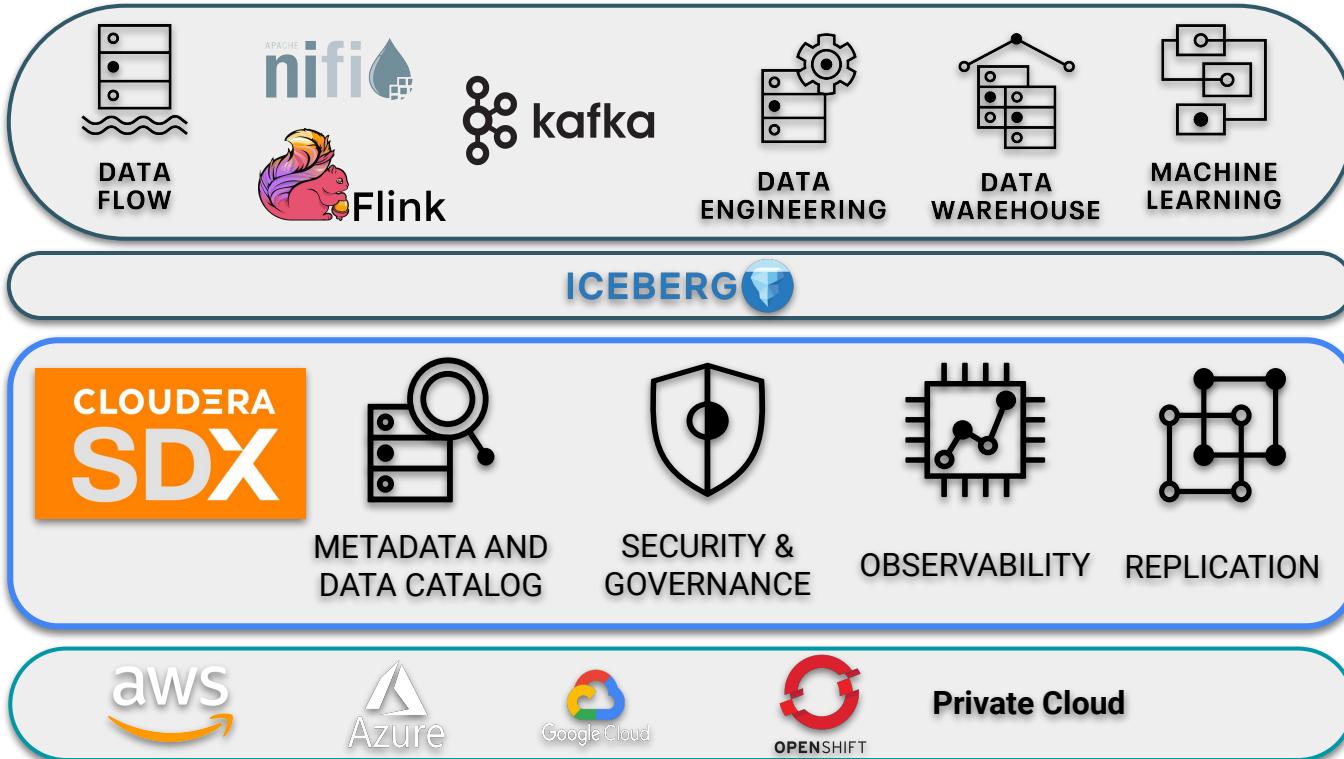


FLANK STACK

REAL-TIME REQUIRES A PLATFORM



CDP: AN OPEN DATA LAKEHOUSE





EXAMPLES



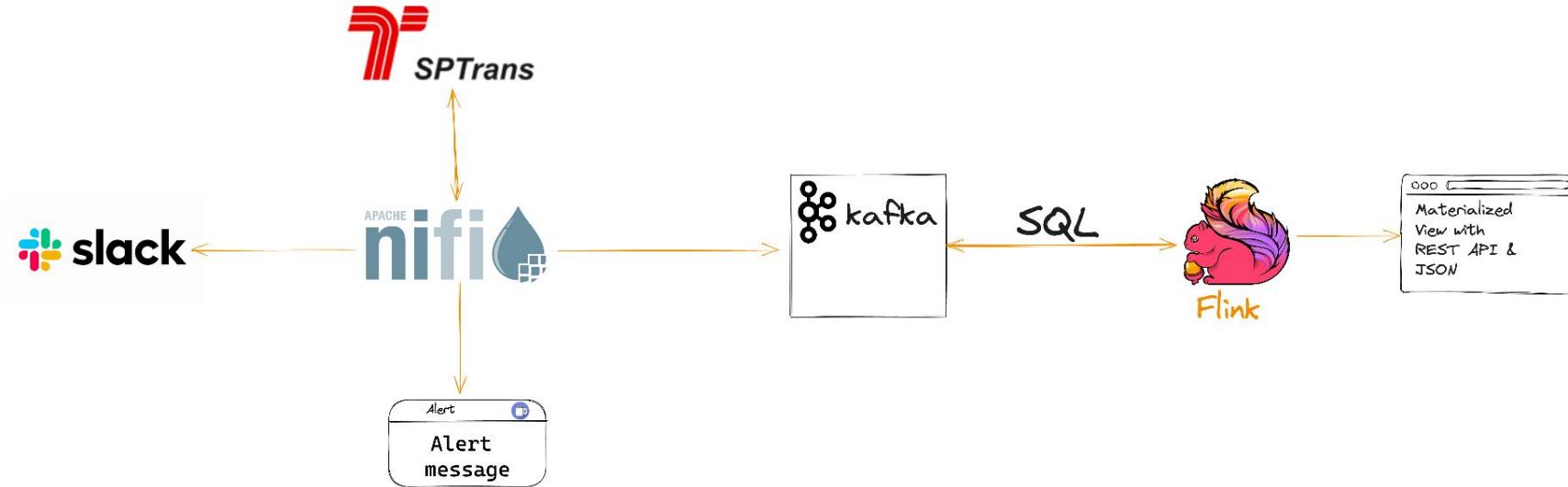
NiFi/Kafka/Flink - Data Tables - Brazil SPTrans



Show 10 entries

HR	Vehicle	Line ID	Line Origin	Line Destination	Lat/Long
17.08	21434	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.537837,-46.63°
17.08	21243	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.529571,-46.59°
17.08	21242	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.4884085,-46.51°
17.08	21399	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.47630525,-46.4°

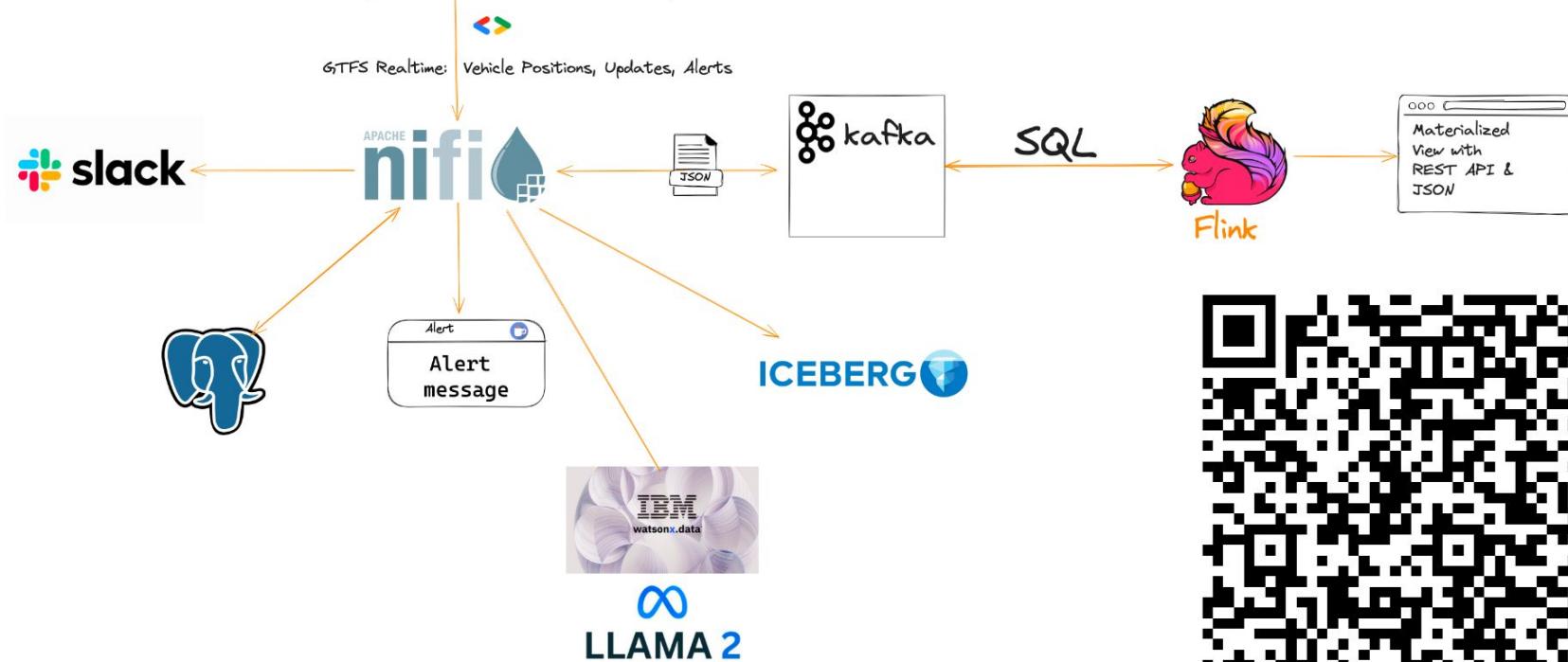


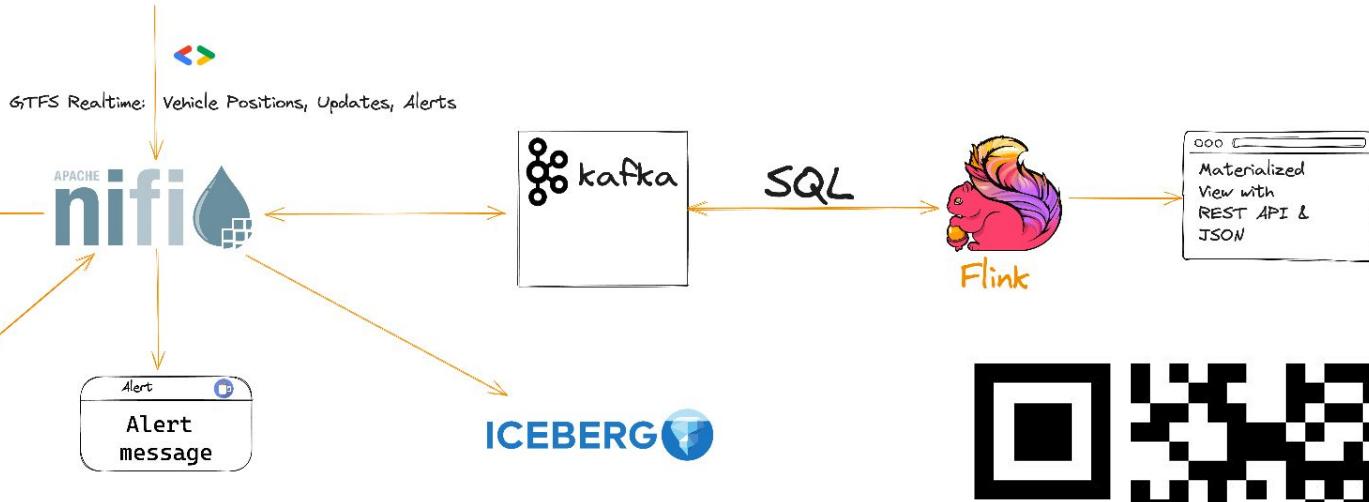




<https://github.com/MobilityData/mobility-database-catalogs/>

Every Transit System







Metropolitan Transportation Authority



Produce

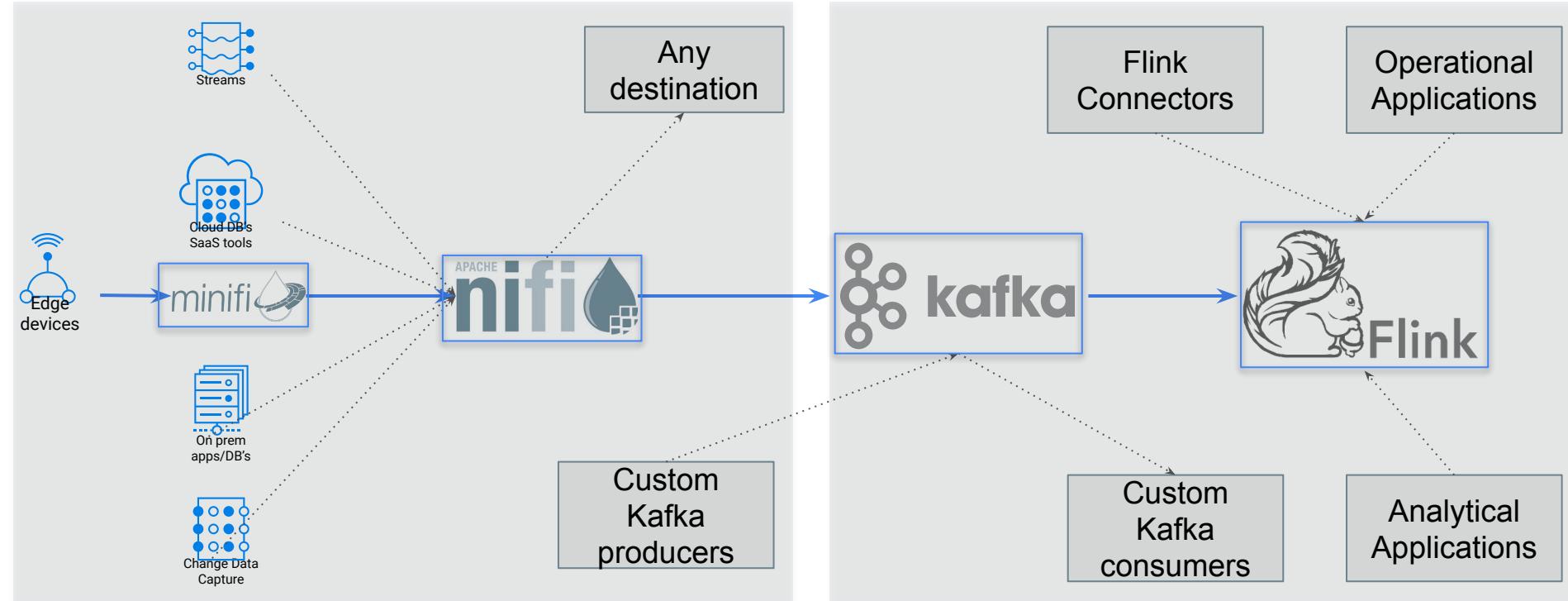


SQL



CLOUDERA

Streaming Data Pipelines with Cloudera Data Platform



STREAMING DATA MOVEMENT & PROCESSING

STREAMING DATA PROCESSING & ANALYTICS

APACHE KAFKA

What is Apache Kafka?

Distributed: horizontally scalable

Partitioned: the data is split-up and distributed across the brokers

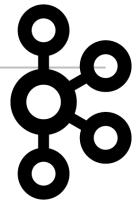
Replicated: allows for automatic failover

Unique: Kafka does not track the consumption of messages (the consumers do)

Fast: designed from the ground up with a focus on performance and throughput

Kafka was built at LinkedIn in 2011

Open sourced as an Apache project



Yes, Franz, It's Kafka

Let's do a metamorphosis on your data. Don't fear changing data.

You don't need to be a brilliant writer to stream data.

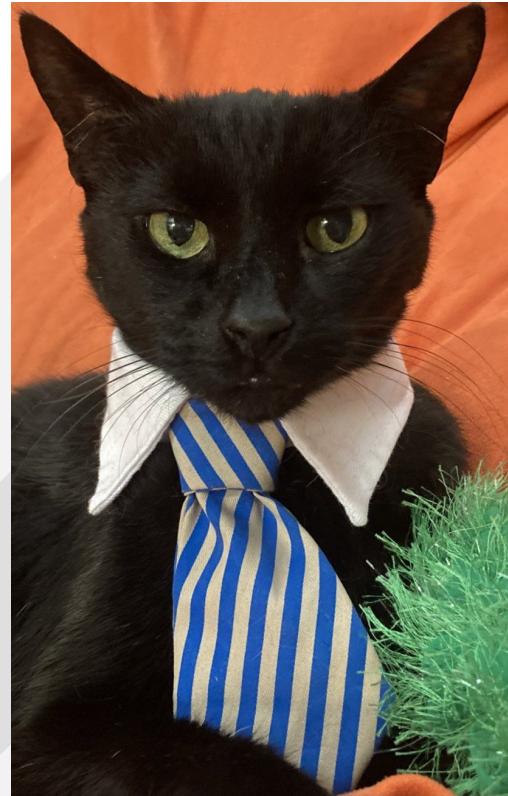


Franz Kafka was a German-speaking Bohemian novelist and short-story writer, widely regarded as one of the major figures of 20th-century literature. His work fuses elements of realism and the fantastic.

[Wikipedia](#)



APACHE FLINK



Flink SQL



- Streaming Analytics
- Continuous SQL
- Continuous ETL
- Complex Event Processing
- Standard SQL Powered by Apache Calcite

The screenshot shows the Apache Flink Dashboard interface. At the top, it displays the job ID (aa4436584d28078f7c758f8903667b84), status (RUNNING), start time (2021-04-07 10:08:37), and duration (3h 56m 21s). Below this, the 'Overview' tab is selected, showing the job's configuration. The configuration details include a Kafka source reading from topic 'weather2' and writing to a Flink TS sink. The sink is configured with a SourceConversionTablet, a Map function, and a SinkConversionTablet. The map function includes logic for handling missing values, setting default values, and filtering. The sink function is a 'MapToTabletSink'. The dashboard also shows task details, including task managers, sub-tasks, and metrics like吞吐量 (Throughput) and延迟 (Latency).

<https://www.datainmotion.dev/2021/04/cloudera-sql-stream-builder-ssb-updated.html>

DATAFLOW APACHE NIFI



Apache NiFi in a few numbers

A very active project with a dynamic community

2800+ members on the Slack channel (535+ - 4 years ago)

475+ contributors on Github across the repositories (260+ - 4 years ago)

65 committers in the Apache NiFi community (45 - 4 years ago)

Apache NiFi 1.23.2 is the latest release, NiFi 2.0 coming soon (NiFi 1.10 - 4 years ago)

14M+ docker pulls of the Apache NiFi image (1M+ - 4 years ago)

Provenance

Displaying 13 of 104
Oldest event available: 11/15/2016 13:34:50 EST
Showing the most recent events.

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

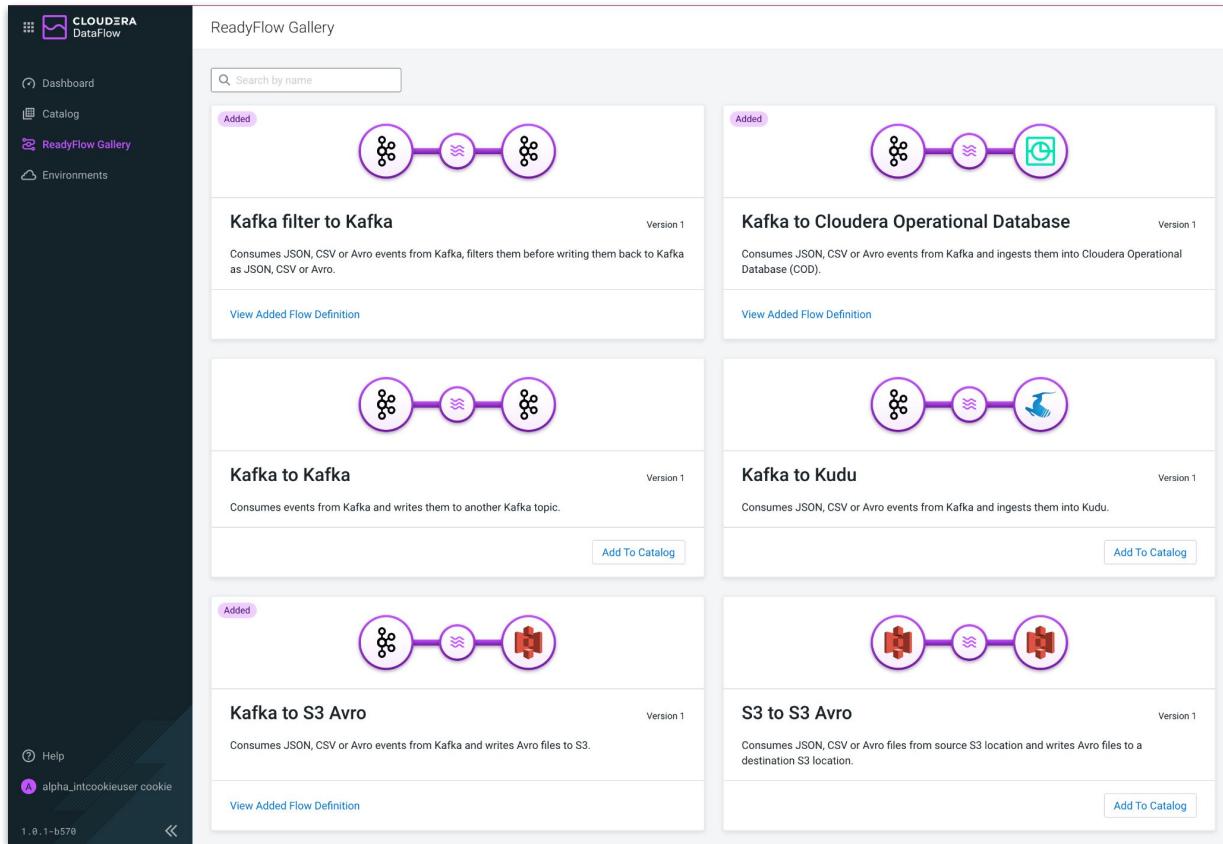
The diagram illustrates a data flow process. It starts with a red circle labeled "RECEIVE", which has an arrow pointing down to a grey circle labeled "JOIN". From the "JOIN" circle, an arrow points down to a grey circle labeled "DROP". Two green arrows originate from the "RECEIVE" and "JOIN" circles and point to a separate "Provenance Event" panel on the right.

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	328717796819631-44800-10519073-0E	

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



Flow Catalog

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

The screenshot shows the Cloudera DataFlow interface with the 'Catalog' tab selected. The main area is titled 'Flow Catalog' and displays a list of available flow definitions. A search bar at the top allows users to search by name. A blue button labeled 'Import Flow Definition' is located in the top right corner. The catalog table has columns for Name, Type, Versions, and Last Updated. Each row in the table represents a different flow definition, with a 'View' icon (a blue arrow) to the right of each entry. The table shows ten entries, with the last two being 'Weather' related flows.

Name ↑	Type	Versions	Last Updated
cc_fraud_template_int101run	Custom Flow Definition	2	a day ago
cc_fraud_template_int101run2	Custom Flow Definition	1	9 days ago
JSON_Kafka_To_Avro_S3	Custom Flow Definition	2	a day ago
Kafka filter to Kafka	ReadyFlow	1	2 days ago
Kafka to Cloudera Operational Database	ReadyFlow	1	2 days ago
Kafka to S3 Avro	ReadyFlow	1	14 hours ago
nifi_flows	Custom Flow Definition	1	2 months ago
Weather Data Flow	Custom Flow Definition	1	a day ago
Weather_Data	Custom Flow Definition	1	15 days ago
Weather_JSON_Kafka_To_Avro_S3	Custom Flow Definition	1	21 days ago

Apache NiFi with Python Custom Processors

Python as a 1st 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

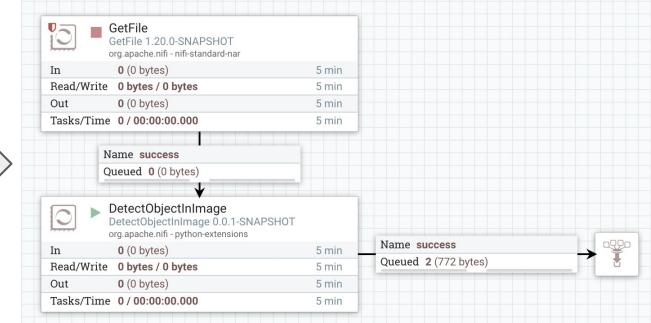
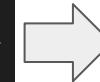
class DetectObjectInImage:
    class Java:
        implements = ['org.apache.nifi.python.processor.FlowFileTransform']
        class ProcessorDetails:
            version = '0.0.1-SNAPSHOT'
            dependencies = ['numpy >= 1.23.5', 'opencv-python >= 4.6']

    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:
```



FREE LEARNING ENVIRONMENT

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



<https://www.cloudera.com/downloads/cdf/csp-community-edition.html>



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

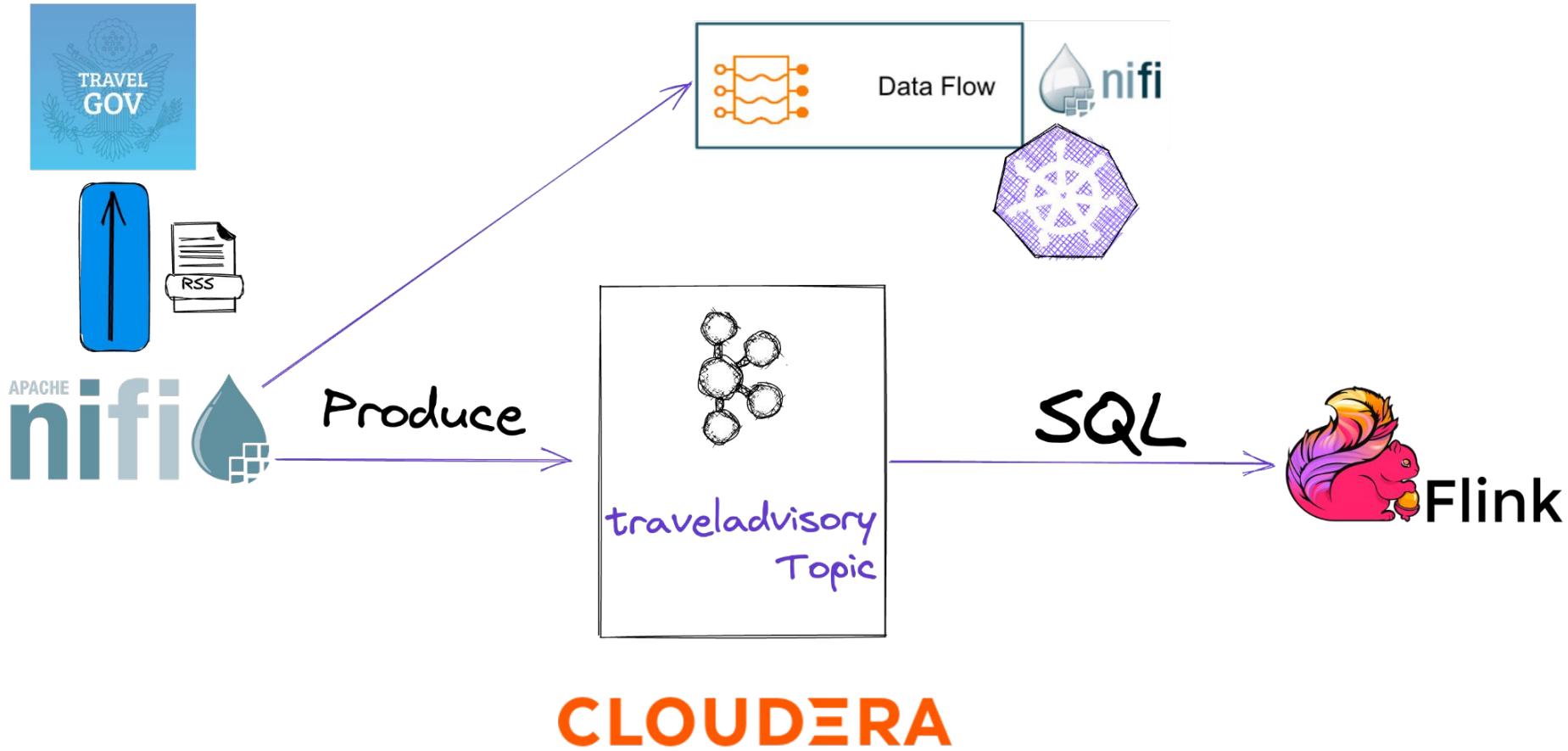


- Apache NiFi in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally
- Docker NiFi
 - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`
 - Licensed under the ASF License
 - **Unsupported**

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

DEMO / CODE







traveladvisories X

searchplanes

RUNNING



Flink Dashboard

Templates

Editor

Materialized View

Job Settings

Job Actions

```
1 select title, domain, category, link, pubdate, ts, uuid, advisoryId
2 FROM
3 `sr1`.`default_database`.traveladvisory
4
```

 Restart Stop Stop Polling Polling samples...

<input type="checkbox"/> title	domain	category	link	pubdate	ts	uuid
<input type="checkbox"/> Bhutan - Level 1: Exercise Normal Precautions	BT,advisory	Level 1: Exercise Normal ...	http://travel.state.gov/co...	Wed, 05 Oct 2022	1680277517680	0412509-8e00-4000-93...
<input type="checkbox"/> China - Level 3: Reconsider Travel	CH,advisory,MC,HK	CH	http://travel.state.gov/co...	Fri, 10 Mar 2023	1680277517682	79e7912a-5d40-4afb-96...
<input type="checkbox"/> China - Level 3: Reconsider Travel	CH,advisory,MC,HK	HK	http://travel.state.gov/co...	Fri, 10 Mar 2023	1680277517682	528c584a-e2cc-4119-ac...
<input type="checkbox"/> Tajikistan - Level 2: Exercise Increased Caution	TI,advisory	Level 2: Exercise Increas...	http://travel.state.gov/co...	Wed, 05 Oct 2022	1680277517683	24fef95e-42a9-4011-9f3...
<input type="checkbox"/> Zambia - Level 1: Exercise Normal Precautions	ZA,advisory	advisory	http://travel.state.gov/co...	Tue, 28 Mar 2023	1680277517684	a4e8106e-5f55-4ef9-a5e...
<input type="checkbox"/> Taiwan - Level 1: Exercise Normal Precautions	TW,advisory	advisory	http://travel.state.gov/co...	Mon, 24 Oct 2022	1680277517688	ed3bad9e-96a0-42ca-a6...
<input type="checkbox"/> Chad - Level 3: Reconsider Travel	CD,advisory	Level 3: Reconsider Travel	http://travel.state.gov/co...	Tue, 04 Oct 2022	1680277517690	1ac6673c-dd29-4186-b8...

Logs

Results

Events

1 to 7 of 7

<

>

Page 1 of 1

»

🔍 Materialized View

Configuration

Primary Key ⓘ

uuid

 Enable MV ⓘ

Retention (Seconds) ⓘ

 Recreate on Job Start ⓘ

Min Row Retention Count ⓘ

10000

 Ignore NULLs ⓘ

API Key ⓘ

traveladvisory1



Queries

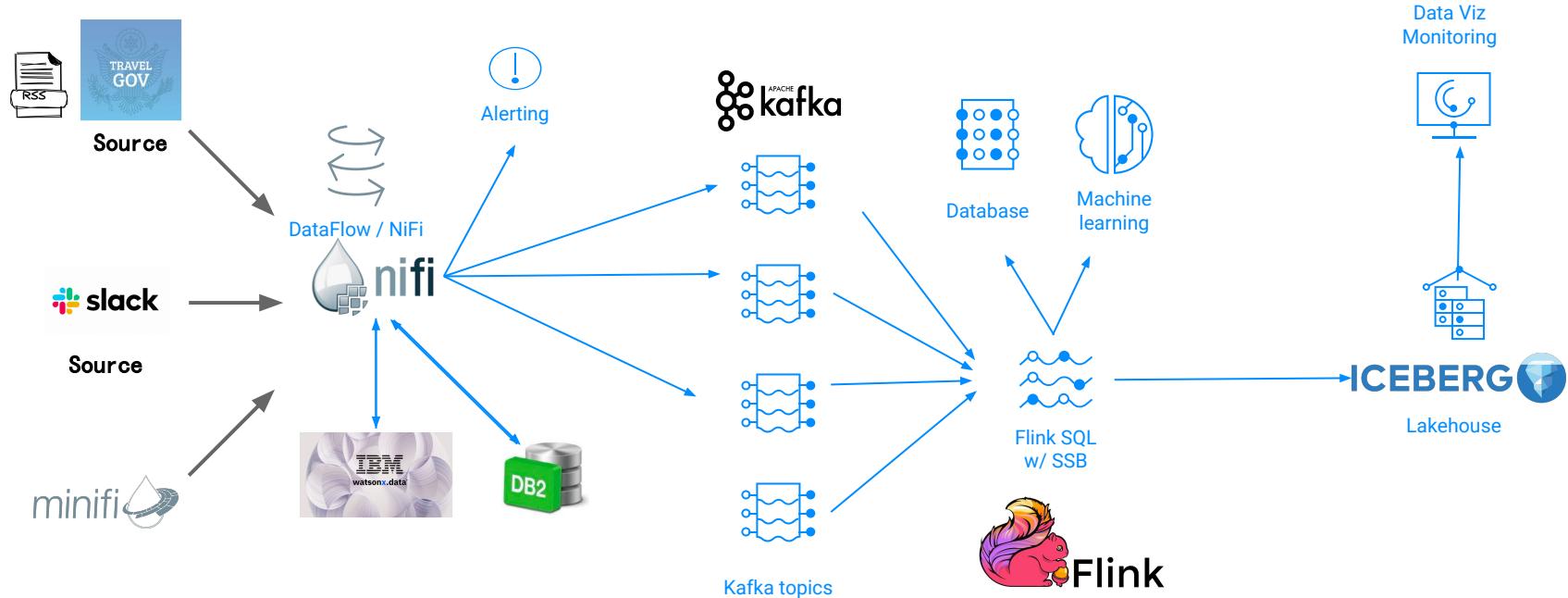
[⊕ Add New Query](#)

```
/api/v1/query/5201/travel?key=66ba91a9-507f-422c-bbb4-86250a9f7bb1&limit=100
```



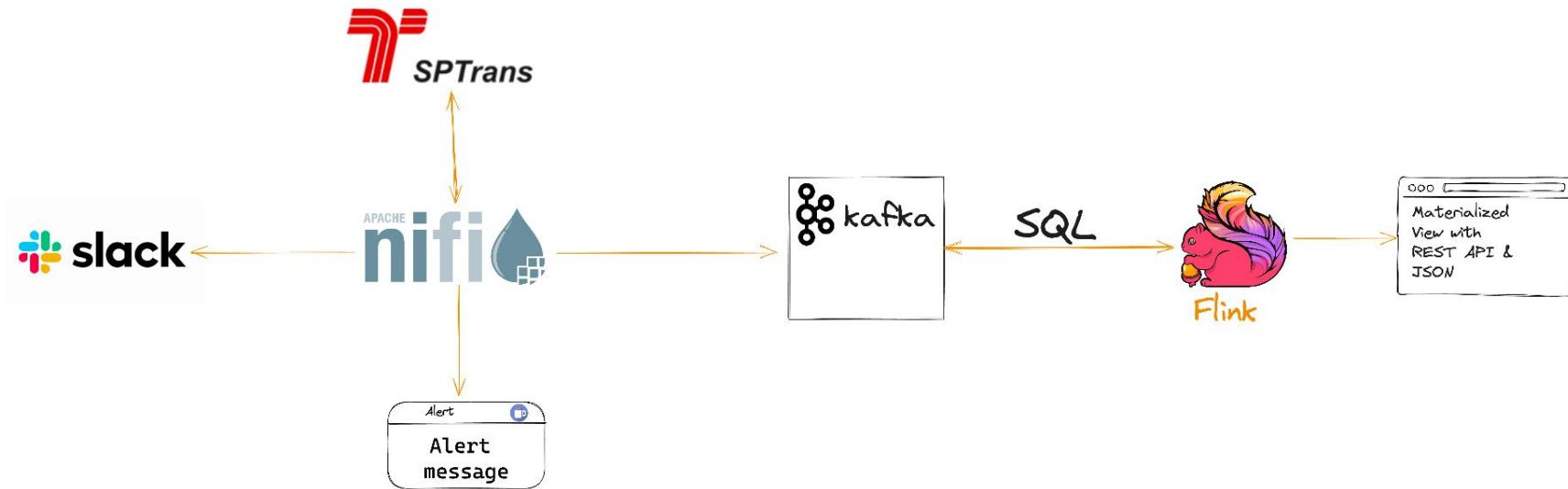
Architecture in the context of Travel Advisories

WatsonX.AI Granite LLM, NiFi, Kafka & Flink



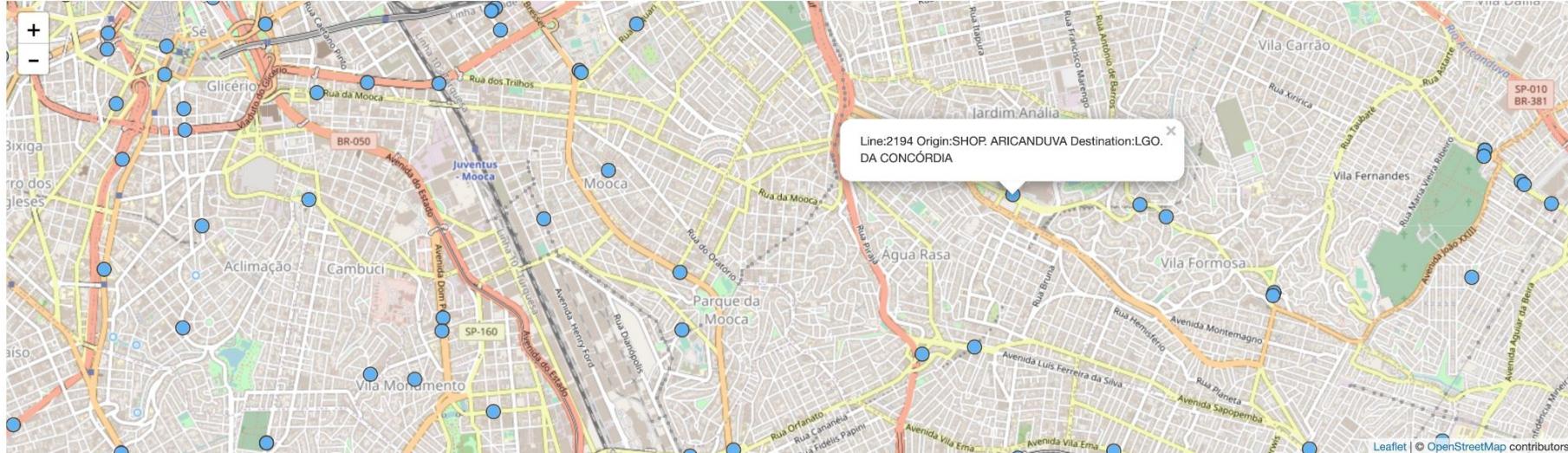
Data in Motion: Overview e Novidades do NiFi, Kafka e Flink

Apresentador: Tim Sparr - Principal DIM Specialist and Developer Advocate



<https://medium.com/cloudera-inc/transit-in-sao-paulo-brasil-flank-style-eaec6753cc63>

NiFi/Kafka/Flink - Data Tables - Brazil SPTrans



Show 10 entries

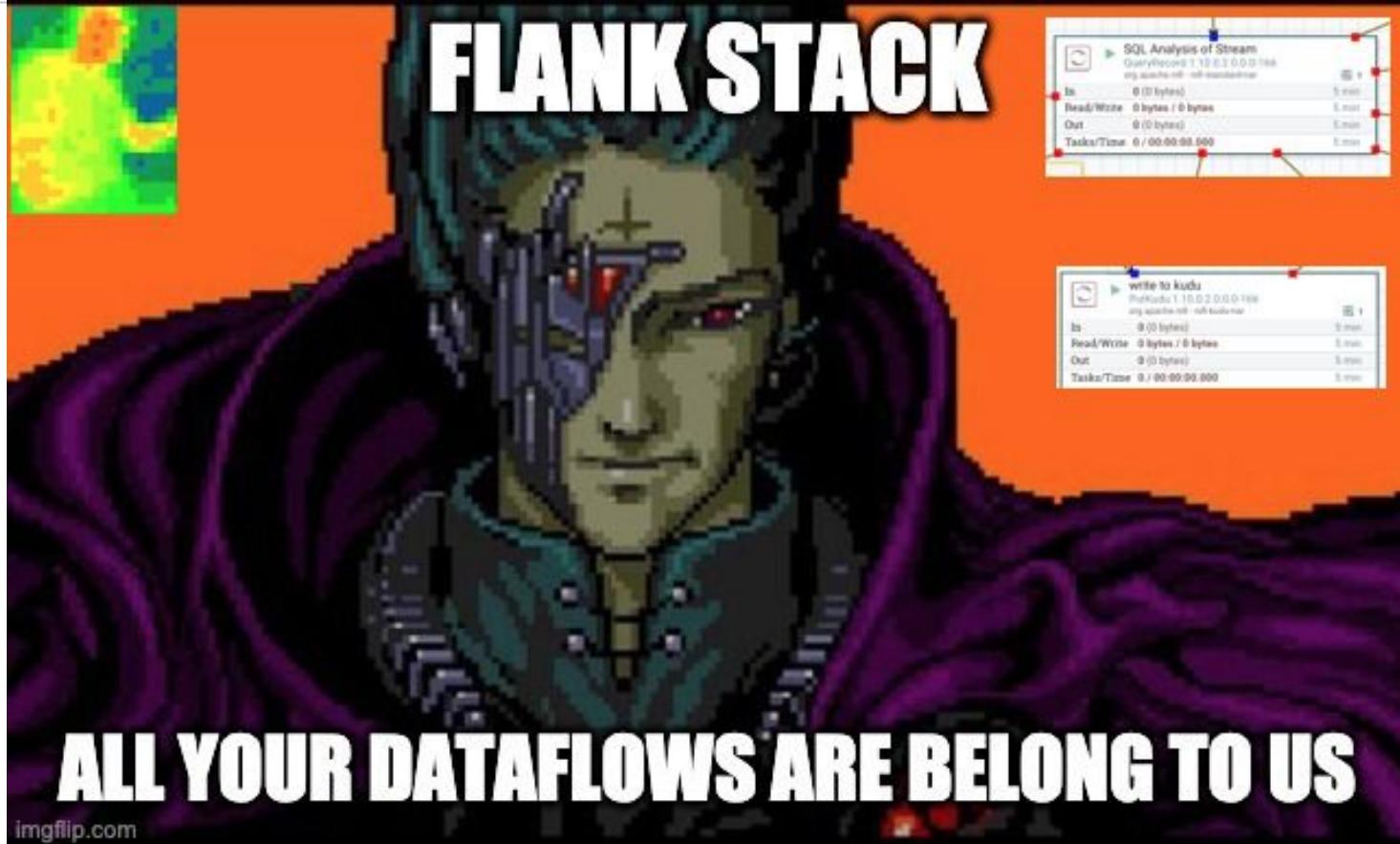
Search:

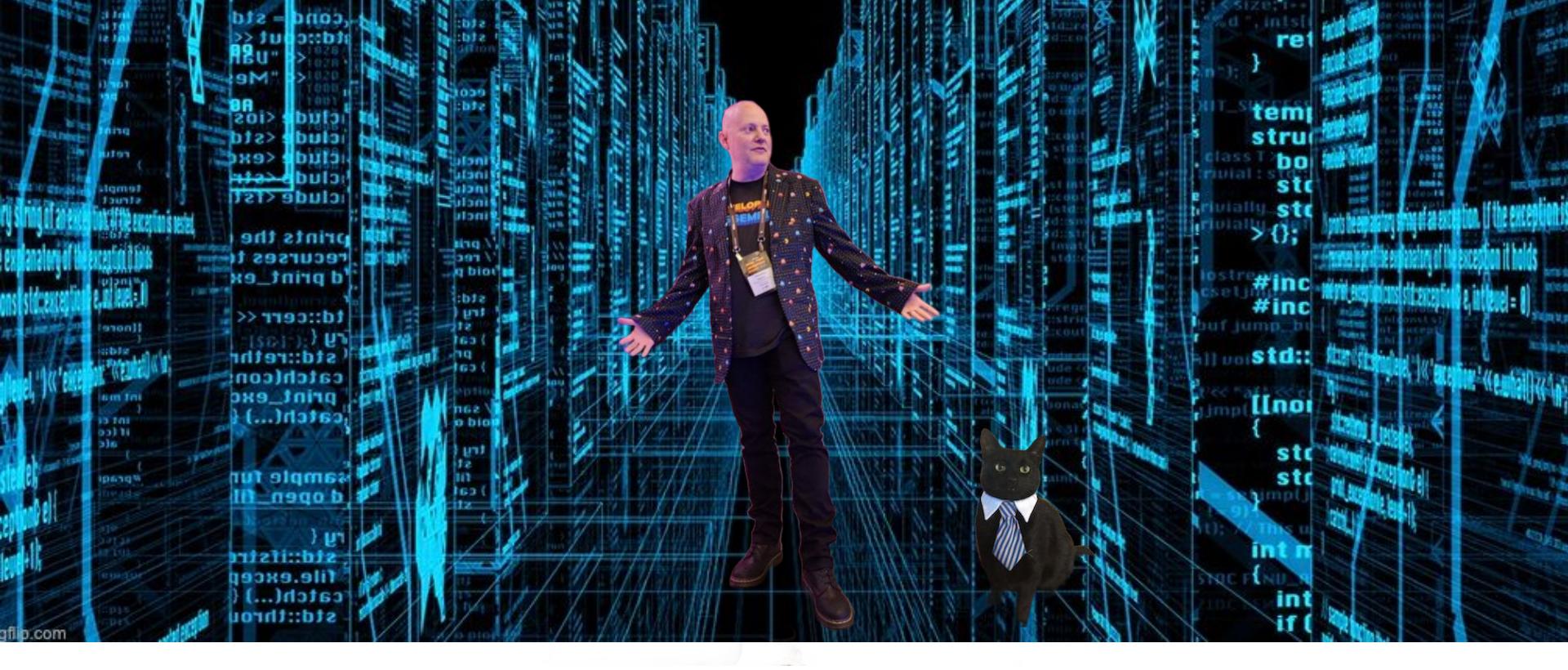
HR	Vehicle	Line ID	Line Origin	Line Destination	Lat/Long	Date/Time
17:08	21434	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.537837,-46.6328475	2023-09-08T20:07:30Z
17:08	21243	33462	PQ. EDU CHAVES	PÇA. DO CORREIO	-23.529571,-46.5984615	2023-09-08T20:07:31Z
17:08	61677	32840	PQ. RES. COCAIA	PQ. IBIRAPUERA	-23.6532785,-46.7017075	2023-09-08T20:07:35Z
17:08	61683	32840	PQ. RES. COCAIA	PQ. IBIRAPUERA	-23.718092,-46.699059	2023-09-08T20:07:20Z
17:08	61517	32840	PQ. RES. COCAIA	PQ. IBIRAPUERA	-23.58114725,-46.6574995	2023-09-08T20:07:28Z
17:08	41014	33514	VL. DALILA	TERM. PQ. D. PEDRO II	-23.5383225,-46.563772	2023-09-08T20:08:04Z
17:08	41019	33514	VL. DALILA	TERM. PQ. D. PEDRO II	-23.5443805,-46.5217695	2023-09-08T20:07:45Z

RESOURCES AND WRAP-UP

Resources







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



DZone. Events

Data Pipelines Virtual Roundtable

REGISTER NOW



Friday, October 27, 2023 | 12 PM ET



Timothy Spann
Principal Developer Advocate,
Cloudera



Eric Sammer
CEO,
Decodeable



Jesse Davis
Moderator,
DZone Chief Technologist

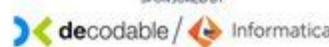


Amol Dongre
Sr Director of Product Management,
Informatica

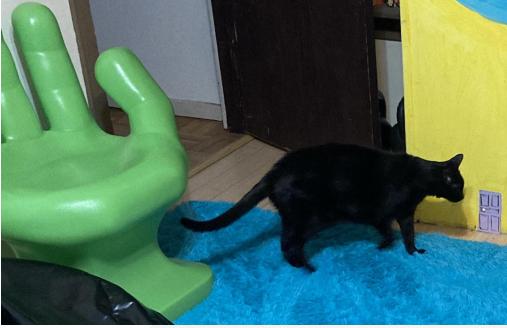
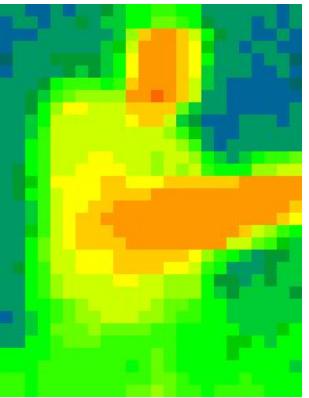


Miguel Lorenzo
VP of Engineering,
Nextall

SPONSORED BY



<https://events.dzone.com/dzone/Data-Pipelines-Investigating-the-Modern-Day-Stack>



TH_ON_G Y_OU

