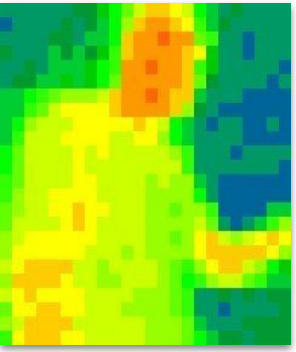
CLOUDERA

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

Tim Spann
Principal Developer Advocate

Feb 8, 2024

Tim Spann

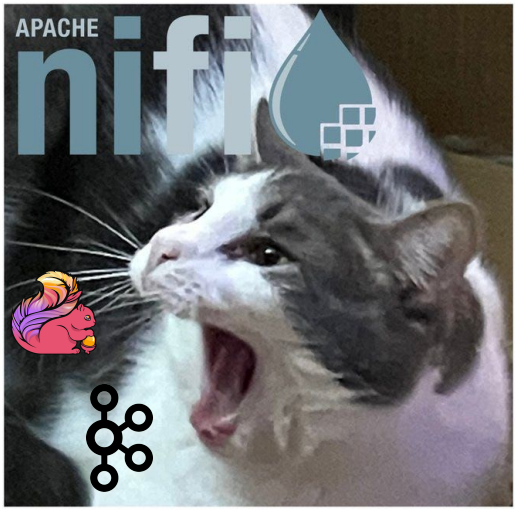


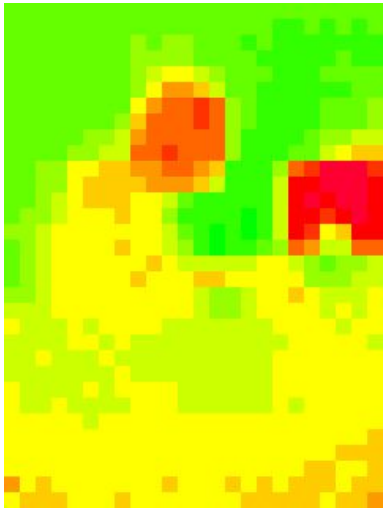
@PaasDev www.datainmotion.dev
github.com/tspannhw medium.com/@tspann
Principal Developer Advocate

Princeton/NYC 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 LLM

FLaNK Stack Weekly by Tim Spann



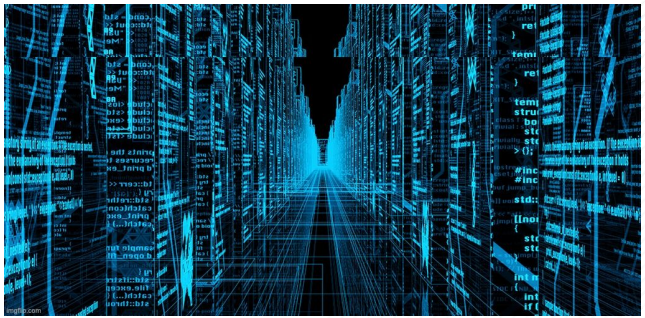


https://bit.ly/32dAJft

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 ...

ICEBERG

@PaasDev

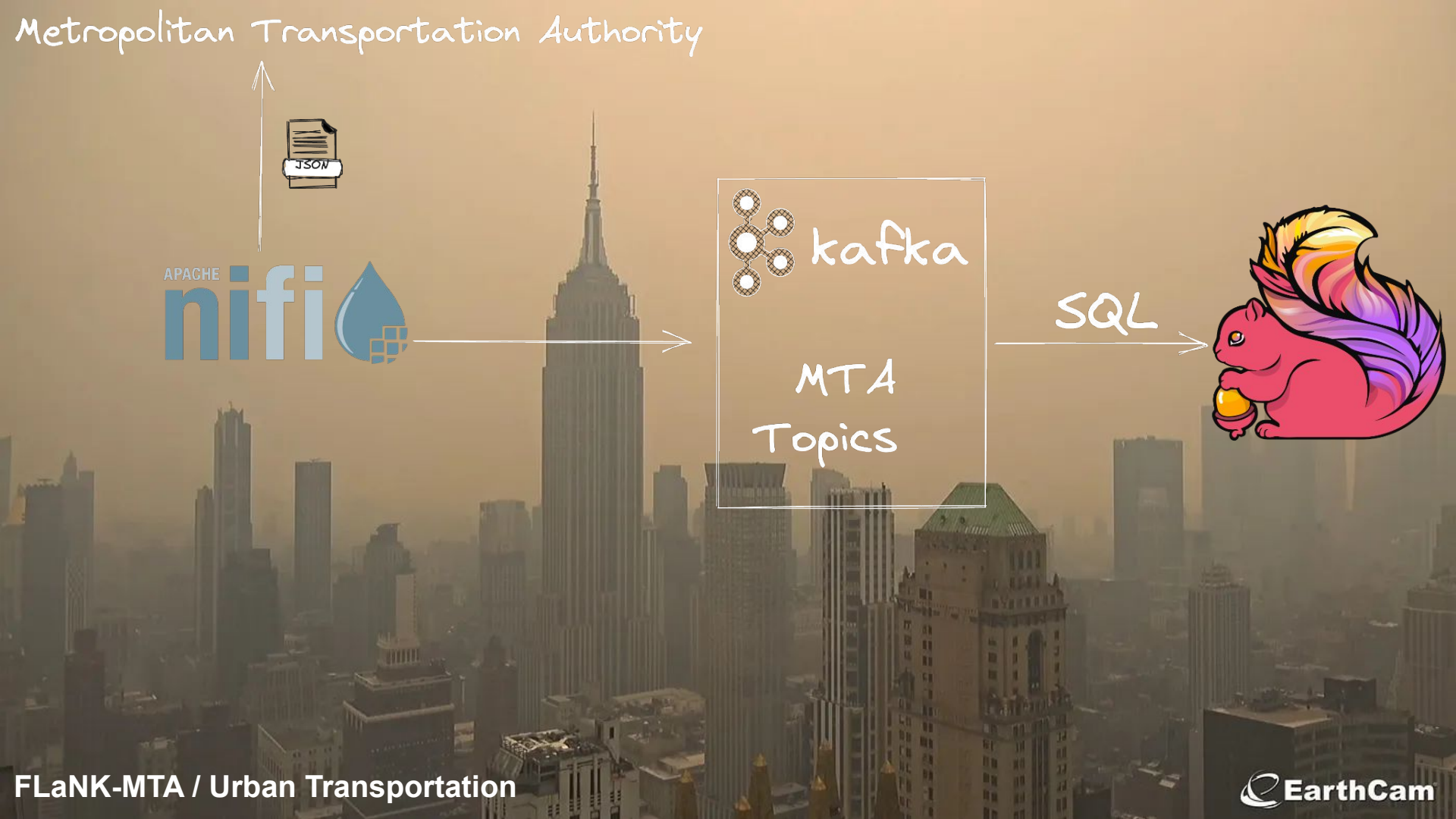
Introduction

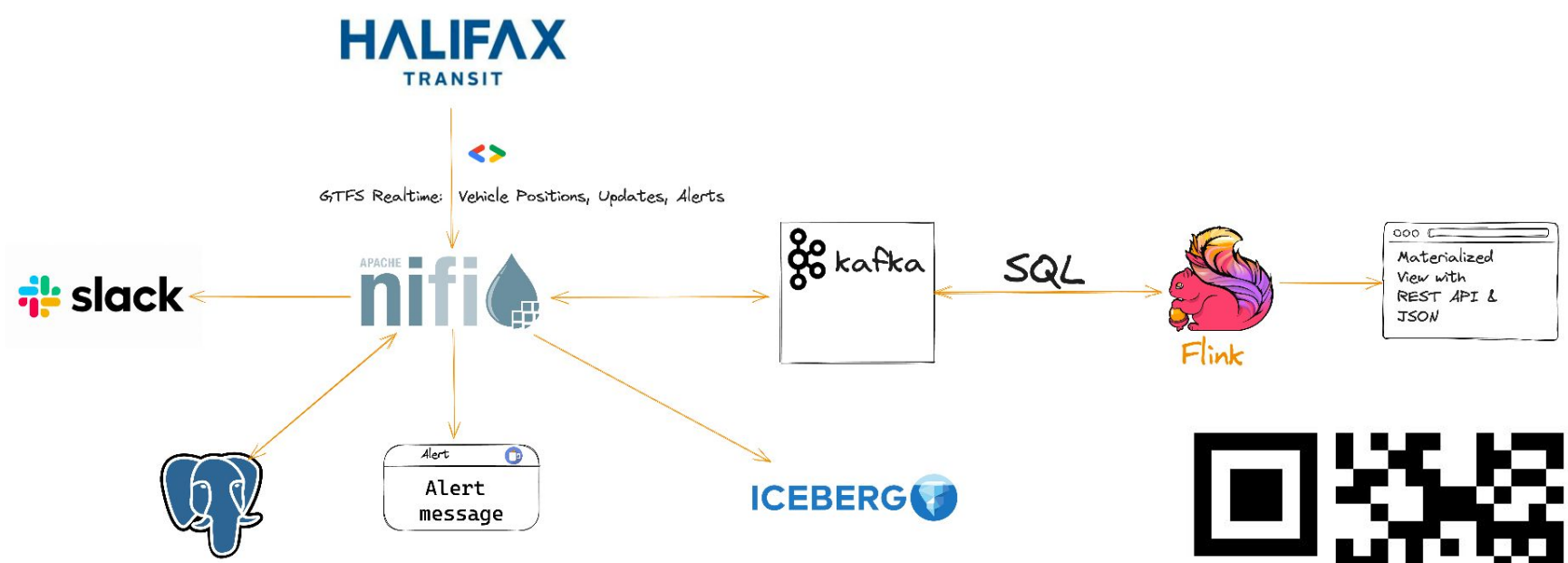
Use Cases

Overview

Apache NiFi, Kafka, Flink

Demos


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


HALIFAX
TRANSIT
GTFS Realtime: Vehicle Positions, Updates, Alerts
slack
APACHE
nifi
kafka
SQL
Flink
Materialized
View with
REST API &
JSON
Alert
Alert
message
ICEBERG

SPTrans

slack

APACHE nifi

kafka

SQL

Flink

Materialized View with REST API & JSON

Alert

Alert message

FLaNK

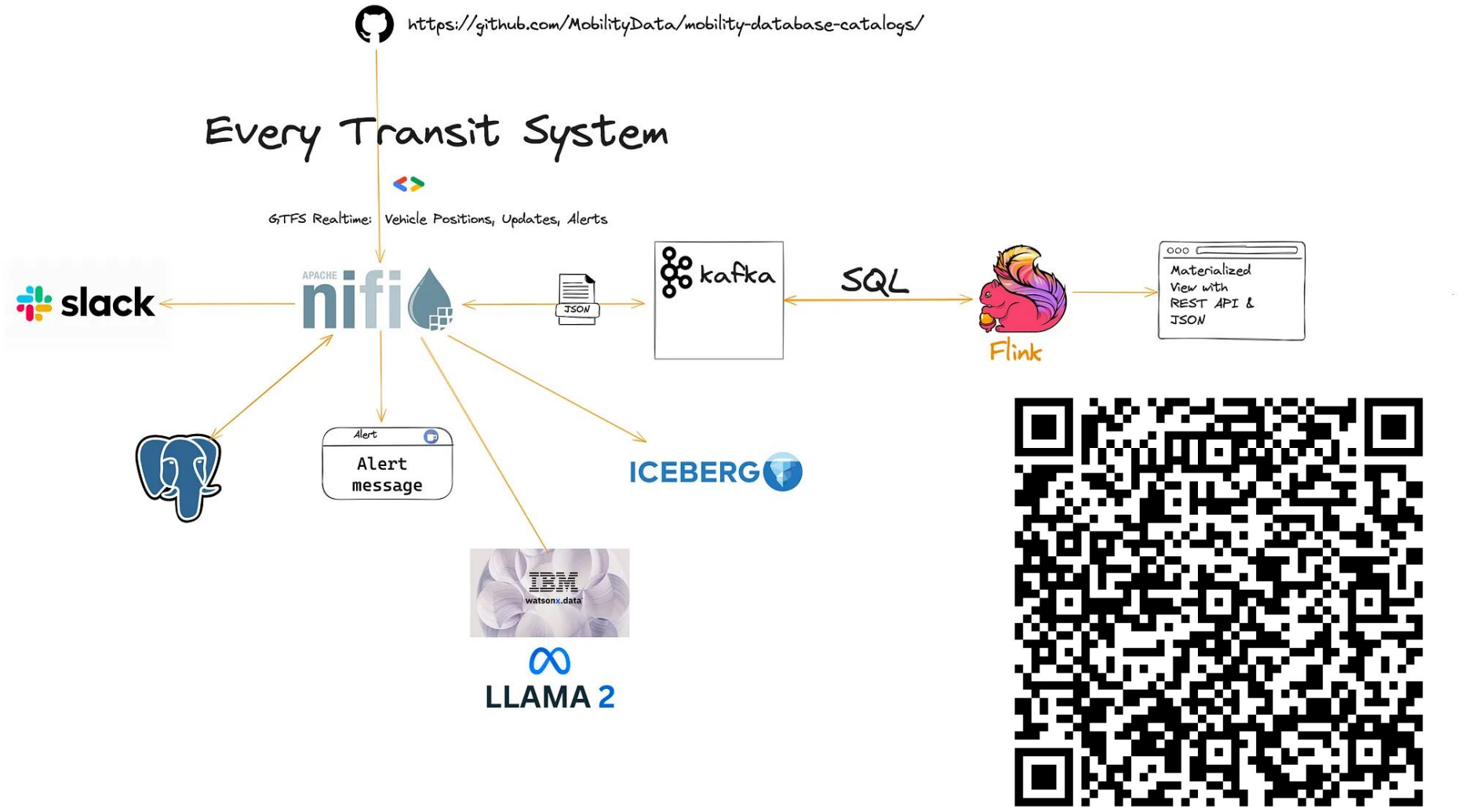
flankstack.dev

NiFi/Kafka/Flink - Data Tables - Brazil SPTrans

Show 10 entries

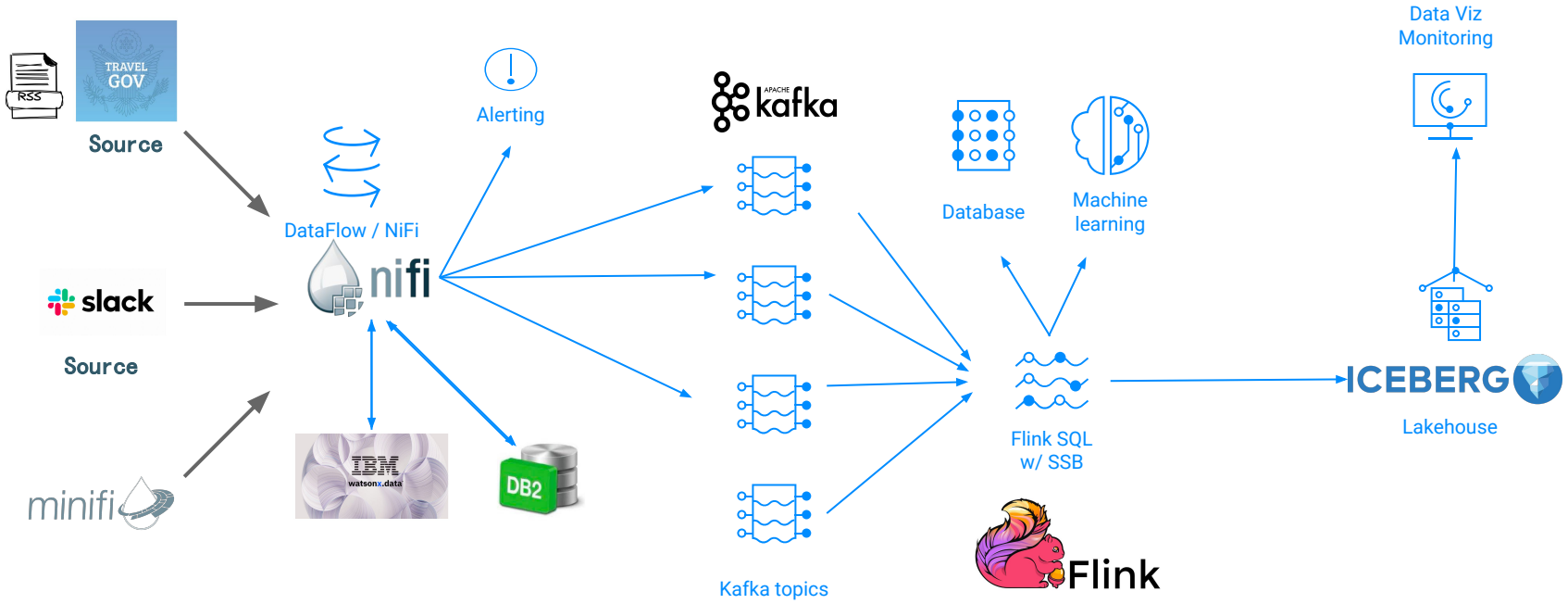
Search:

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


https://github.com/MobilityData/mobility-database-catalogs/
Every Transit System
GTFS Realtime: Vehicle Positions, Updates, Alerts
slack
APACHE nifi
JSON
kafka
SQL
Flink
Materialized View with REST API & JSON
Alert
Alert message
ICEBERG
IBM watsonx.data
LLAMA 2

Architecture in the context of Travel Advisories

WatsonX.AI LLM, NiFi, Kafka & Flink

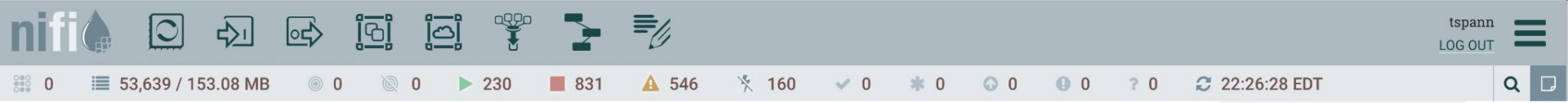


[FLaNK for Halifax Canada Transit — NiFi, Kafka, Flink, SQL, GTFS-RT | by Tim Spann | Cloudera | Dec, 2023 | Medium]

[Never Get Lost in the Stream. NiFi-Kafka-Flink for getting to work… | by Tim Spann | Cloudera | Dec, 2023 | Medium]

[Iteration 1: Building a System to Consume All the Real-Time Transit Data in the World At Once | by Tim Spann | Cloudera | Medium]

[Watching Airport Traffic in Real-Time | by Tim Spann | Cloudera | Medium]



DATAFLOW APACHE NIFI

Apache NiFi - developed 17 years ago by the NSA

2006
NiagaraFiles (NiFi) was first incepted
at the National Security Agency (NSA)

APACHE INCUBATOR

November 2014
NiFi is donated to the Apache Software Foundation (ASF) through NSA's Technology Transfer Program and enters ASF's incubator.

APACHE nifi

July 2015
NiFi reaches ASF top-level project status

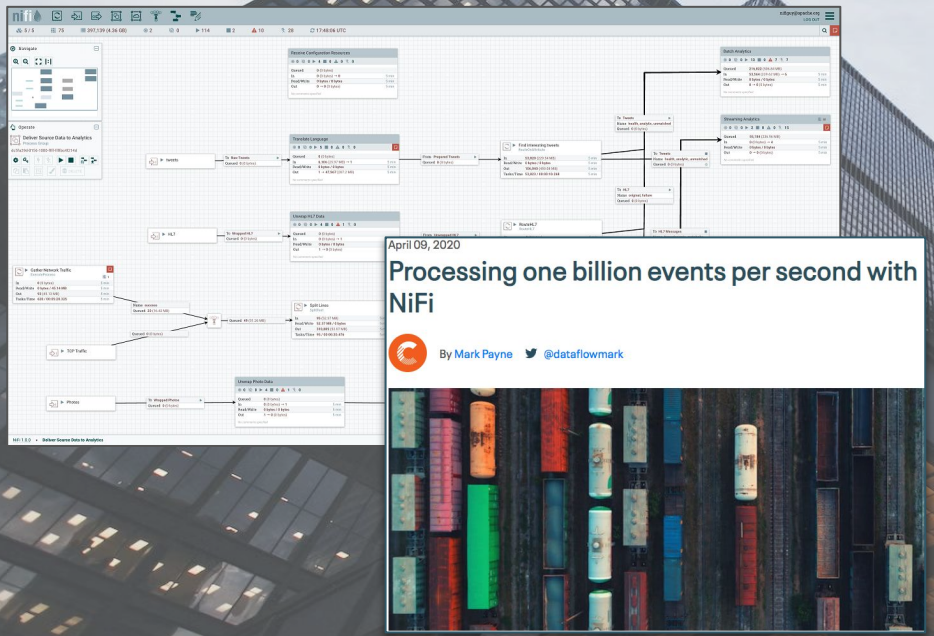
Apache NiFi in a few numbers

A very active project with a dynamic community & comparison with ACEU 2019

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 1.25.0 is the latest release, NiFi 2.0.0-M2 is in alpha.
14M+ docker pulls of the Apache NiFi image (1M+ - 4 years ago)

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

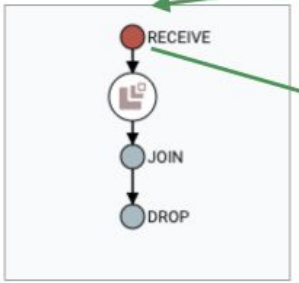
Displaying 13 of 104
Oldest event available: 11/15/2016 13:34:50 EST

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

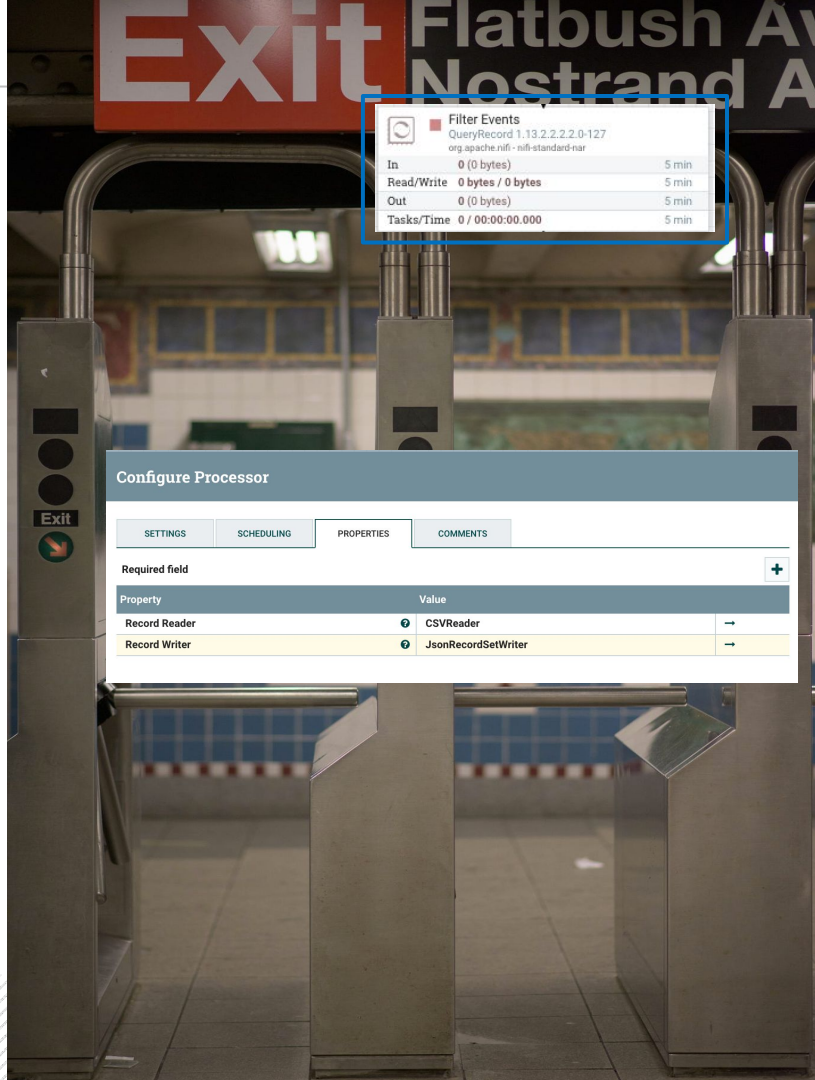
kafka.topic
nifi-testing
No value previously set

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 KAFKA

Apache Kafka

MiNiFi

DATA COLLECTION AT THE EDGE

US-West Fleet

C++ agent

minifi

US-Central Fleet

C++ agent

minifi

US-East Fleet

C++ agent

minifi

Apache Kafka

INGEST GATEWAY POWERED BY KAFKA

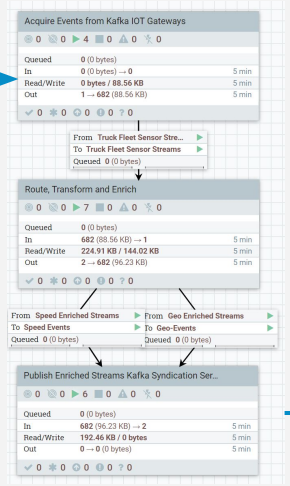
gateway-west-raw-sensors

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

Structured Streaming

Micro Services Stream Analytics App

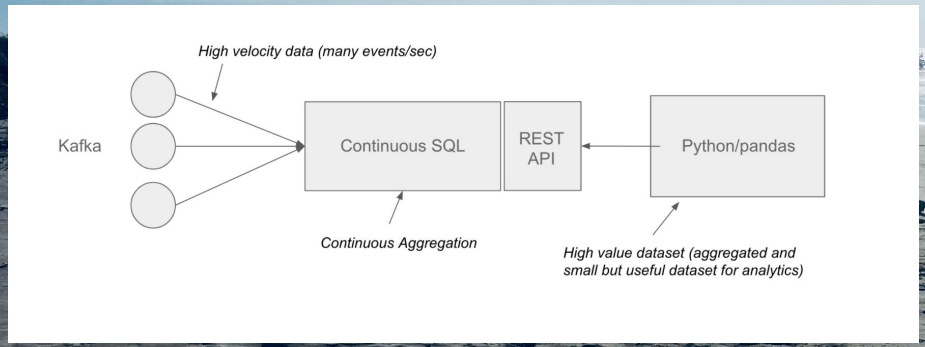
Complex Low Latent Stream Analytics App

Apache Flink

APACHE FLINK

SSB MATERIALIZED VIEWS

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



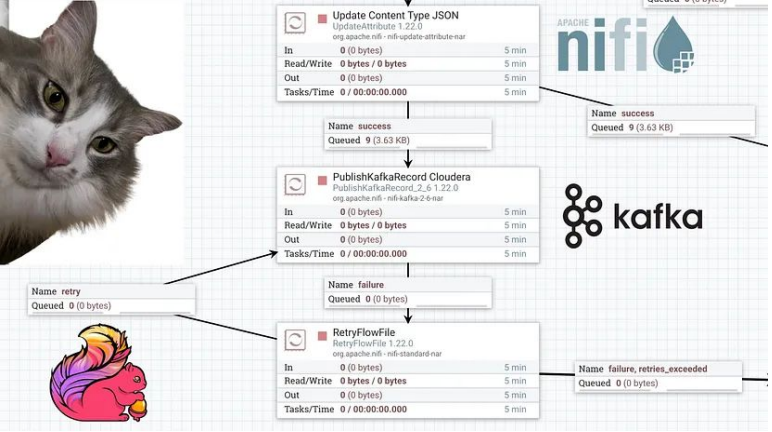

```
SELECT userid,
max(amount) as max_amount,
sum(amount) as sum_amount,
count(*) as thecount,
tumble_end(eventTimestamp, interval '5' second) as ts
FROM authorizations
GROUP BY userid, tumble(eventTimestamp, interval '5' second)
HAVING count(*) > 1
```

```
[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

```
[ ]:
```



DEMO

https://github.com/tspannhw/FLaNK-EveryTransitSystem

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


THANK YOU