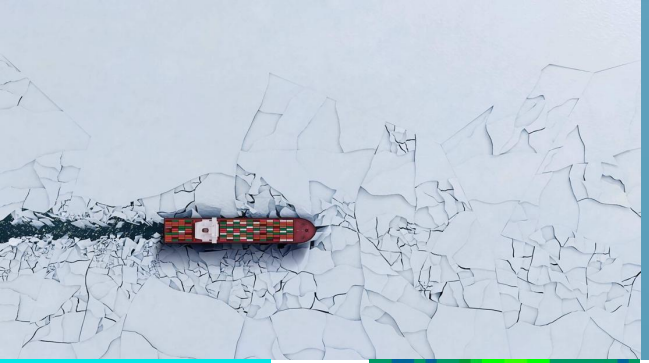
CLOUDERA

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

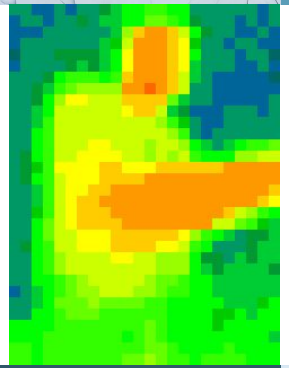
Tim Spann - Principal Developer Advocate
Data In Motion



IoT

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

CLOUDERA



CLOUDERA

EDGE
2AI

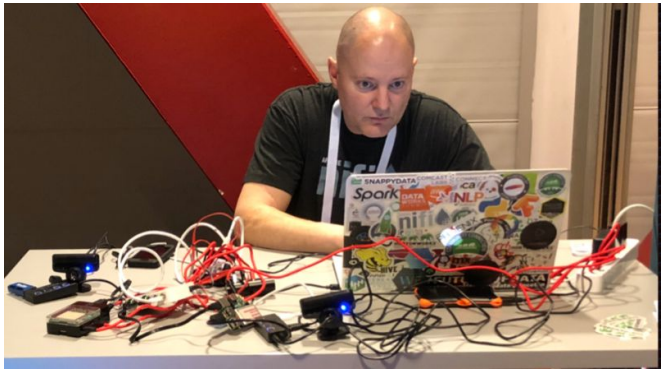
CLOUDERA

ENTERPRISE
DATA
CLOUD

CLOUDERA

TODAY'S LEAD

Who am I?



Principal Data-in-Motion Developer Advocate

DZone Zone Leader and Big Data MVB

Princeton and NYC Future of Data Meetups
ex-Pivotal Field Engineer ex-StreamNative ex-PwC

https://github.com/tspannhw https://twitter.com/PaaSDev
https://www.datainmotion.dev/
https://medium.com/@tspann

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

Apresentador: Tim Spann - Principal DIM Specialist and Developer Advocate

Intro to NiFi

Intro to Kafka

Intro to Flink

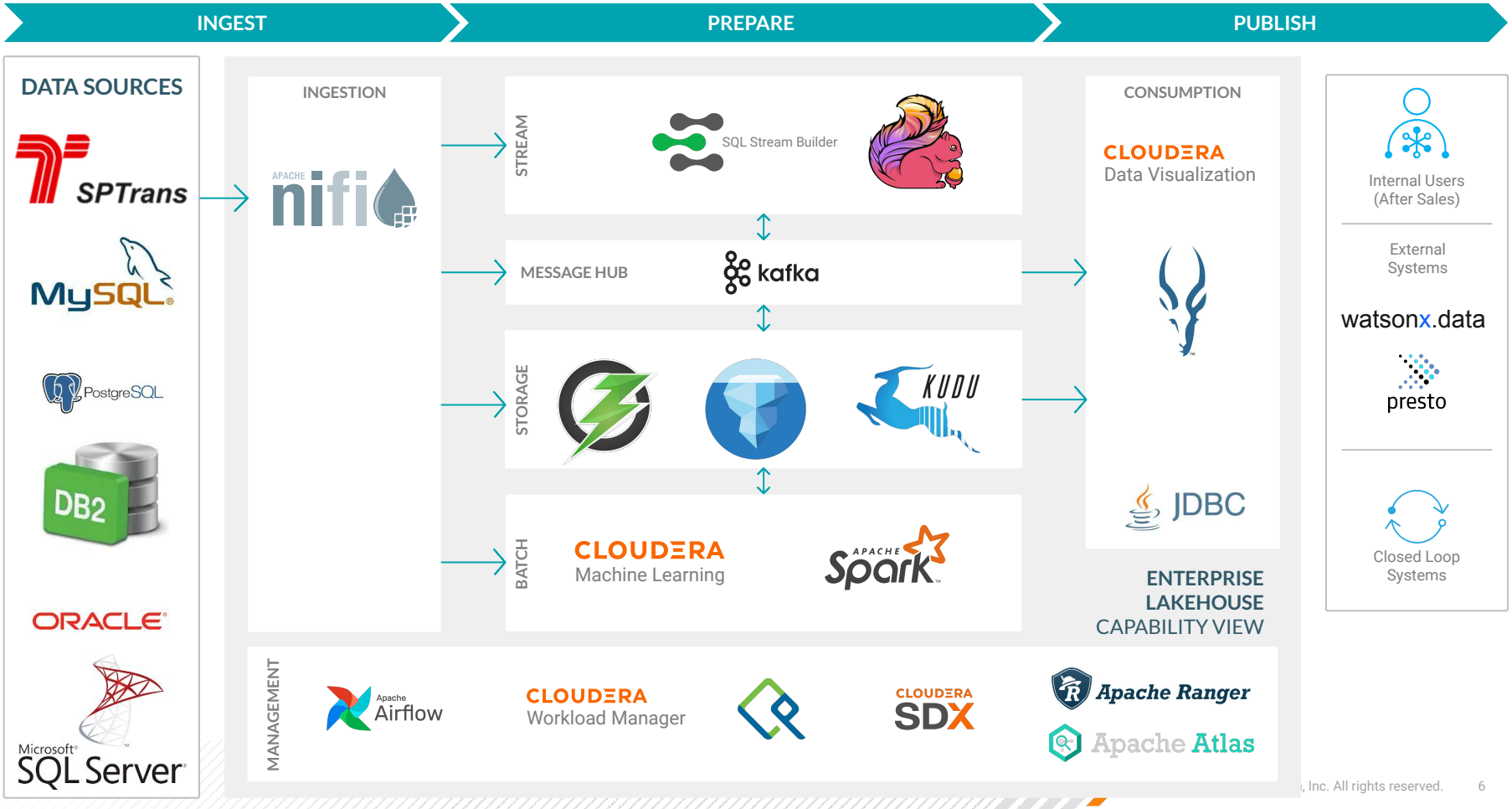
Together as FLaNK

Demos

Q&A

REAL-TIME REQUIRES A PLATFORM

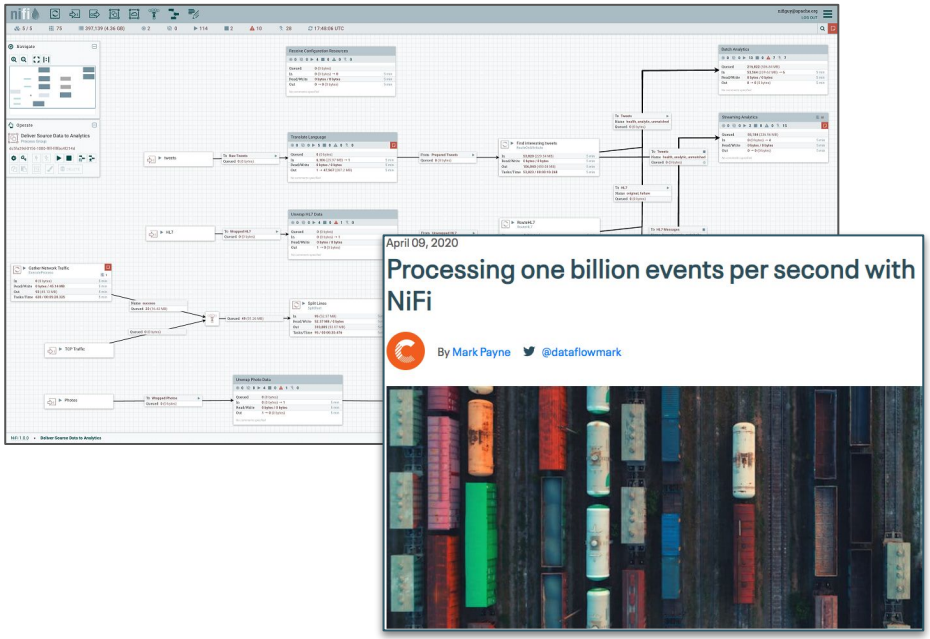
REST API ARCHITECTURE - Using FLaNK to pull the data out of anything in near-real time



Cloudera DataFlow - Apache NiFi

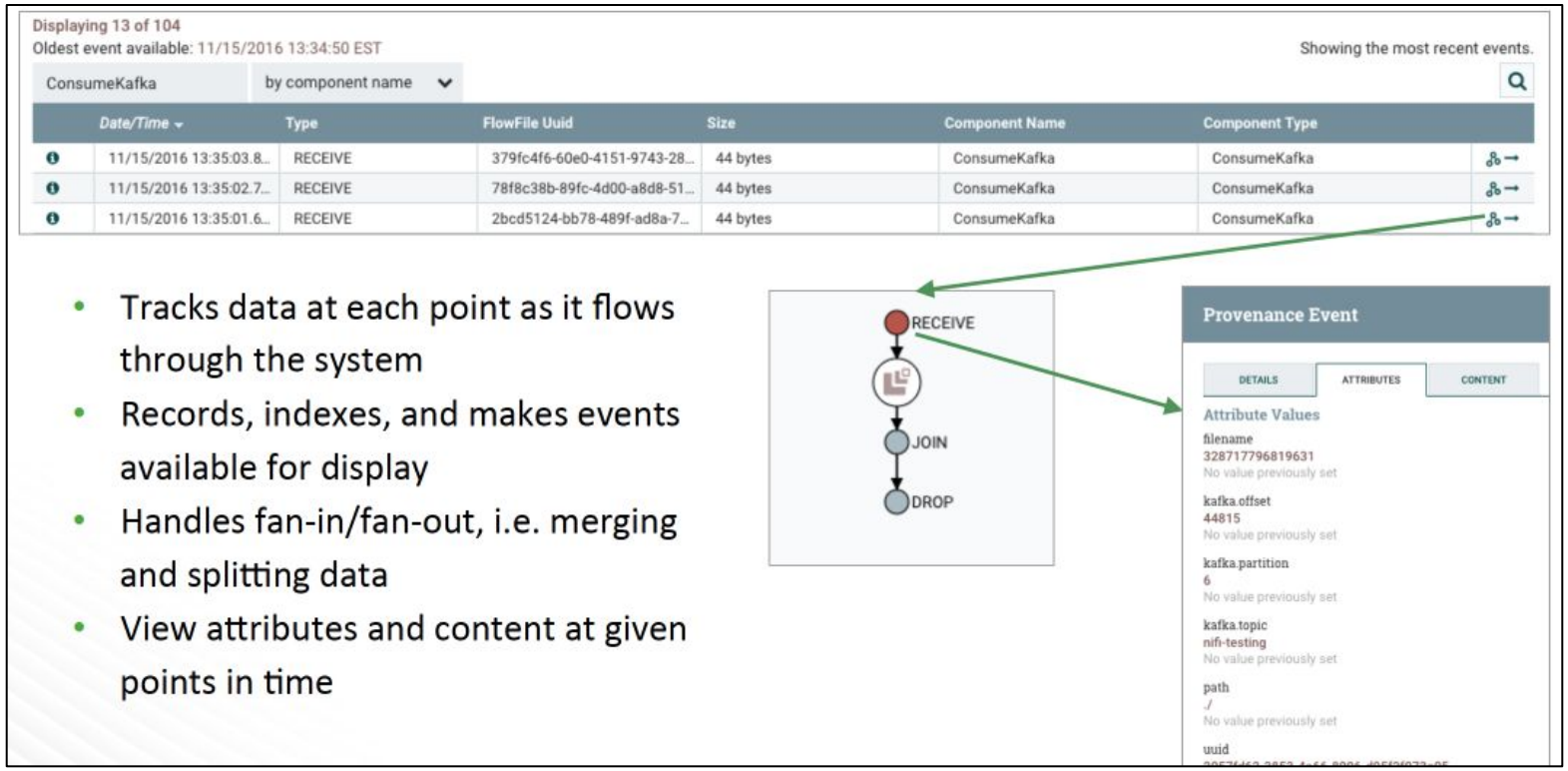
CLOUDERA DATAFLOW - POWERED BY APACHE NiFi

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

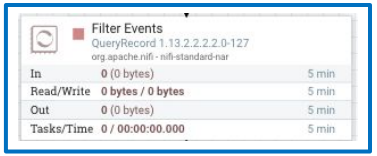


- #1 data ingestion/movement engine
- Strong community
- Product maturity over 11 years
- Deploy on-premises or in the cloud
- Over 400+ pre-built processors
- Built-in data provenance
- Guaranteed delivery
- Throttling and Back pressure

PROVENANCE

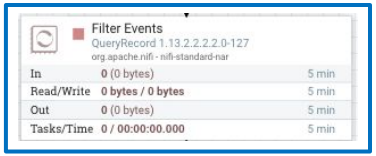


RECORD-ORIENTED DATA WITH NIFI

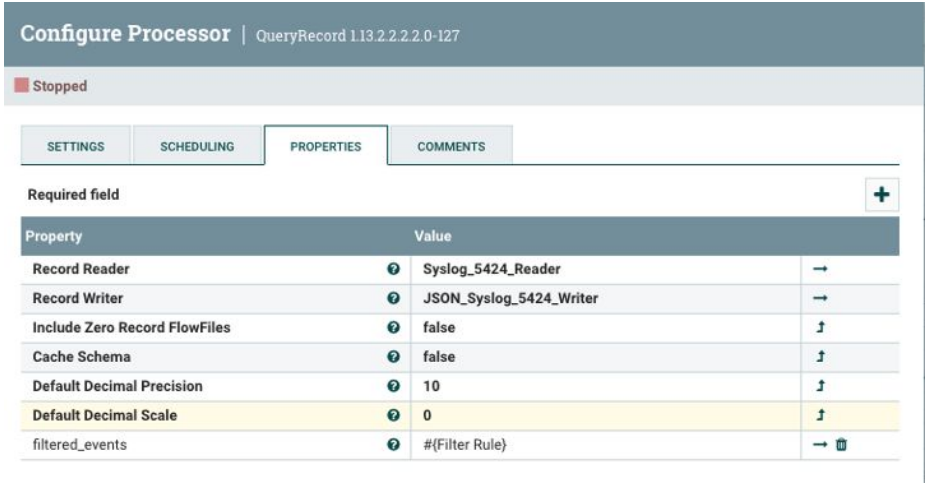


- **Record Readers** - Avro, CSV, Grok, IPFIX, JSAN1, JSON, Parquet, Scripted, Syslog5424, Syslog, WindowsEvent, XML
- **Record Writers** - Avro, CSV, FreeFromText, Json, Parquet, Scripted, XML
- Record Reader and Writer support referencing a schema registry for retrieving schemas when necessary.
- Enable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

RUNNING SQL ON FLOWFILES

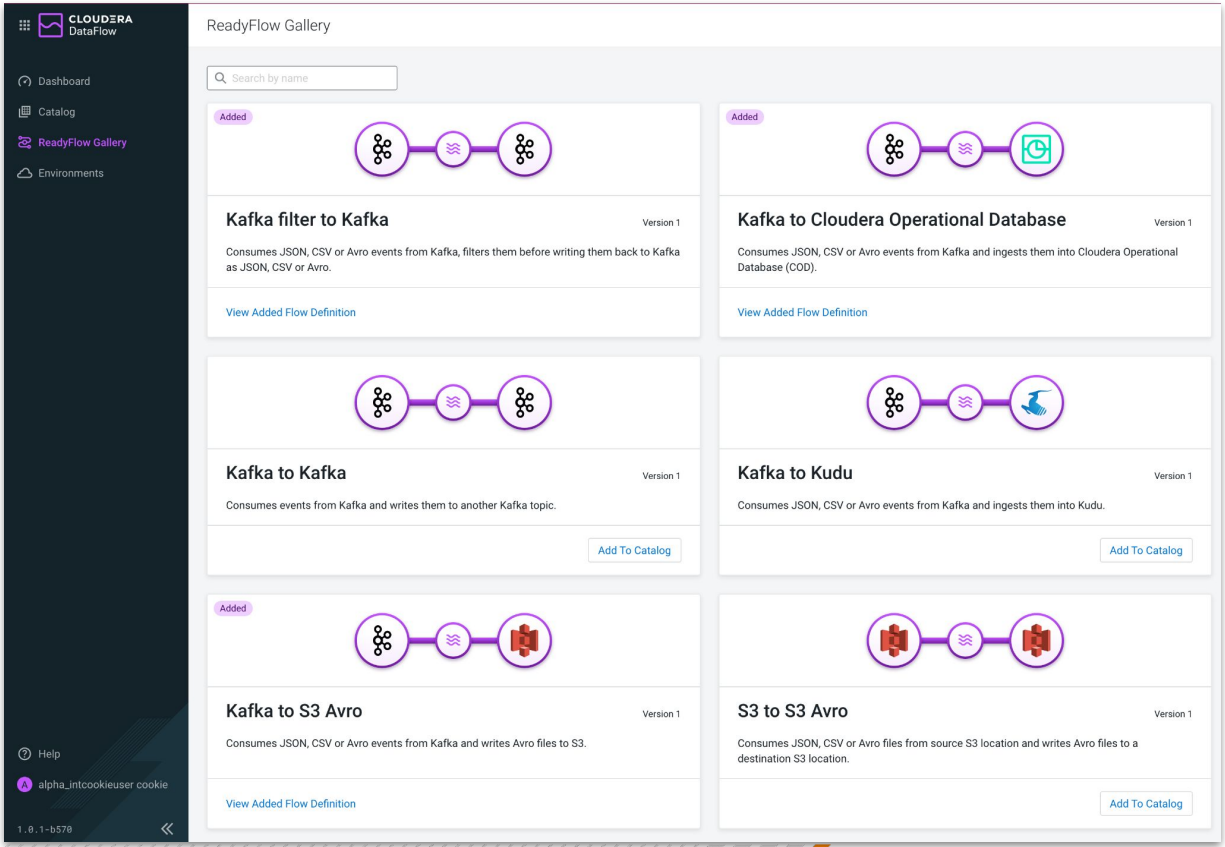


- Evaluates one or more SQL queries against the contents of a FlowFile.
- This can be used, for example, for field-specific filtering, transformation, and row-level filtering.
- Columns can be renamed, simple calculations and aggregations performed.
- The SQL statement must be valid ANSI SQL and is powered by Apache Calcite.



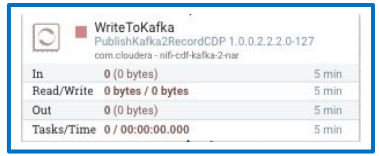
READYFLOW GALLERY

- 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

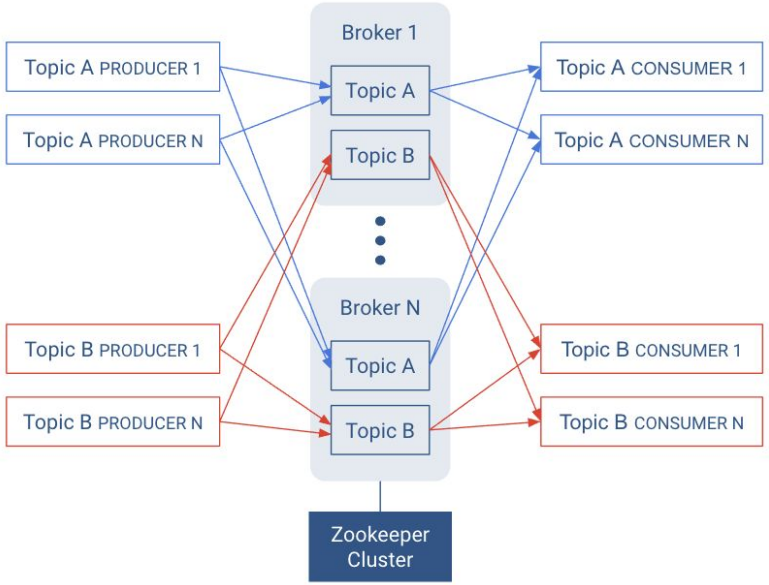


Cloudera Streams Messaging Manager - Apache Kafka

STREAMS MESSAGING WITH KAFKA



- Highly reliable distributed messaging system.
- Decouple applications, enables many-to-many patterns.
- Publish-Subscribe semantics.
- Horizontal scalability.
- Efficient implementation to operate at speed with big data volumes.
- Organized by topic to support several use cases.

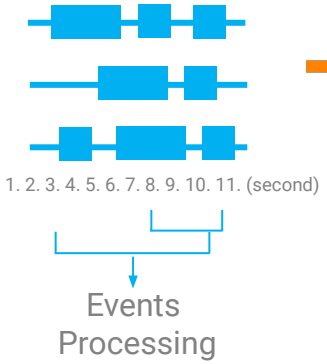


Cloudera SQL Stream Builder - Flink SQL

DELIVERING STREAMING ANALYTICS

Capture Events that Matter

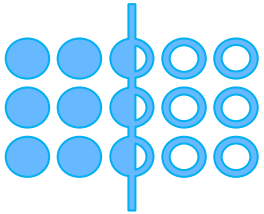
Low-latency analytics use cases



+

Parsing and Blending Data

Both offline and streaming data



+

Data Analysts Can Write Queries

Across the Lines of Business

SQL

=

Streaming Analytics

SQL STREAM BUILDER (SSB)

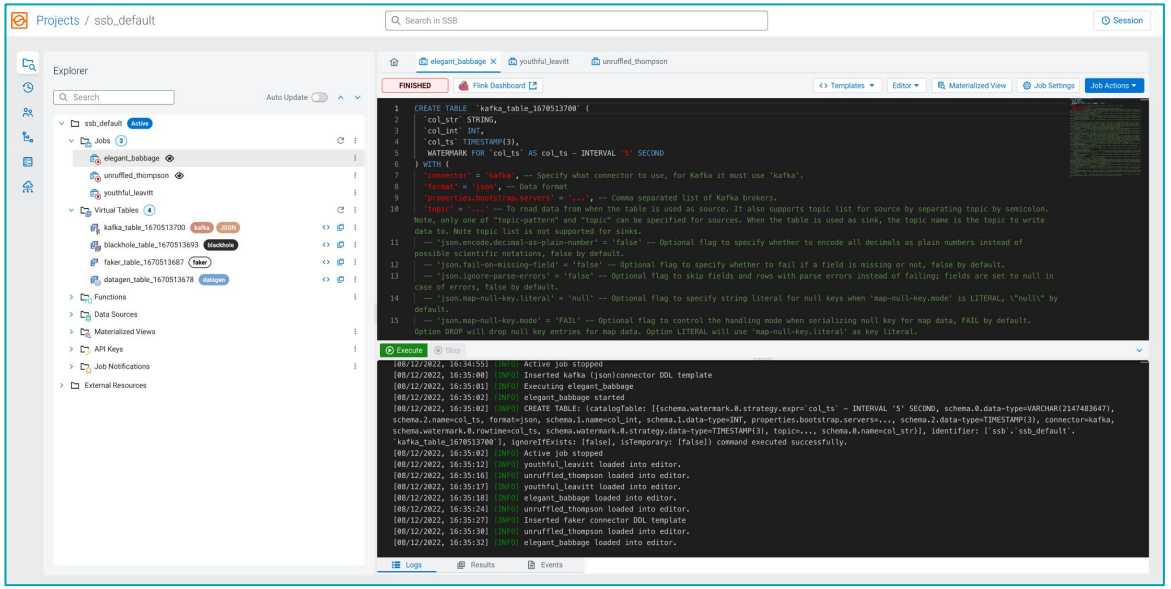
Democratize access to real-time data with just SQL

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

No Java or Scala code development required.

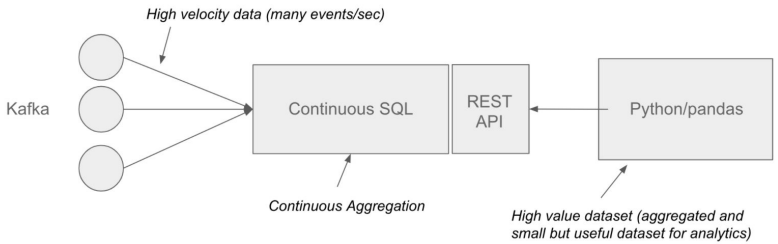
Simplifies access to data in Kafka & Flink. Connectors to batch data in HDFS, Kudu, Hive, S3, JDBC, CDC and more

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

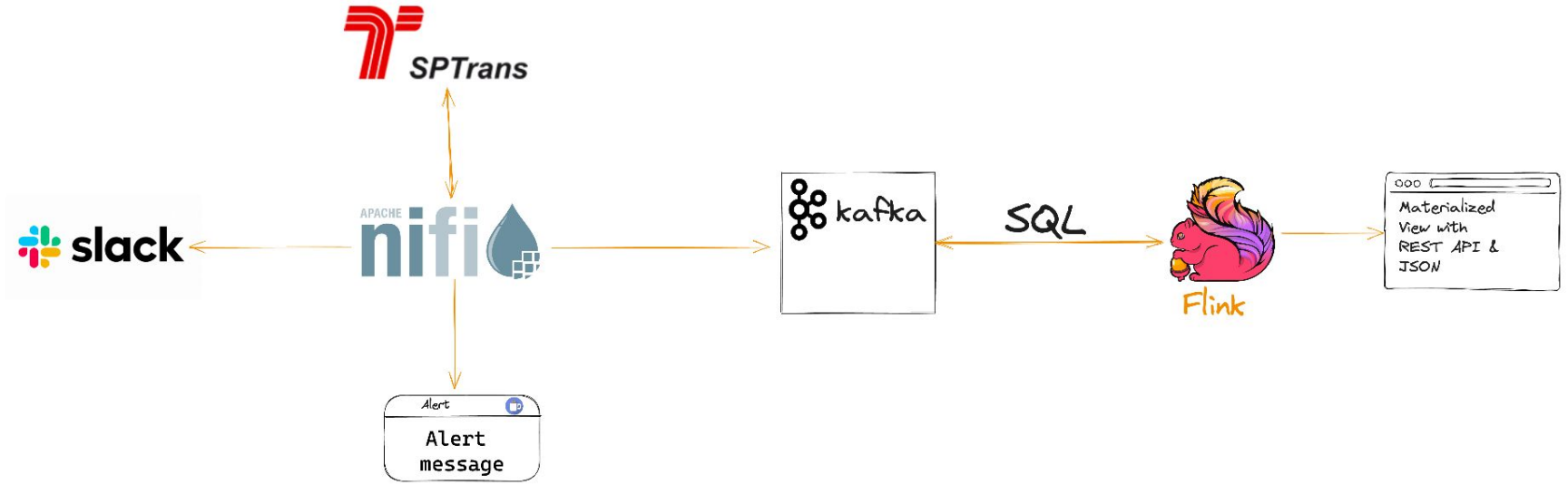
```
[ ]:
```

Demo


Denúncia 181
Vamos combater a violência
Sigilo absoluto - Atendimento 24 horas
TRANSPPASS
SPTrans
8 0255
SP 156
CIDADE DE SÃO PAULO
50 Km/h
FAM-9980
Mantenha Distância

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

Apresentador: Tim Spann - Principal DIM Specialist and Developer Advocate



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

FREE LEARNING ENVIRONMENT

Cloudera Streams Processing - 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

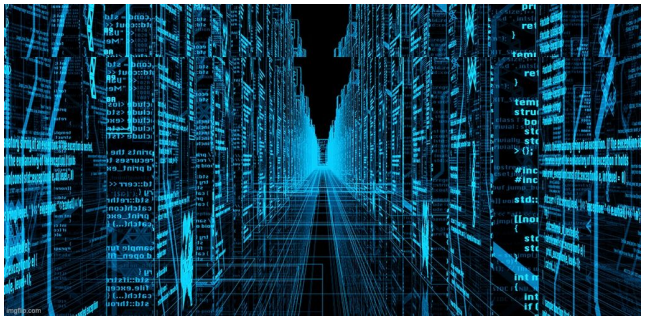
- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

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

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

RESOURCES, WRAP-UP, Q&A

Future of Data - NYC / Princeton + Virtual



FUTURE OF DATA

AN OPEN SOURCE COMMUNITY

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

From Big Data to AI to Streaming to LLM to Cloud to Analytics to NLP to Fast Data to Machine Learning to Microservices to ...

ICEBERG

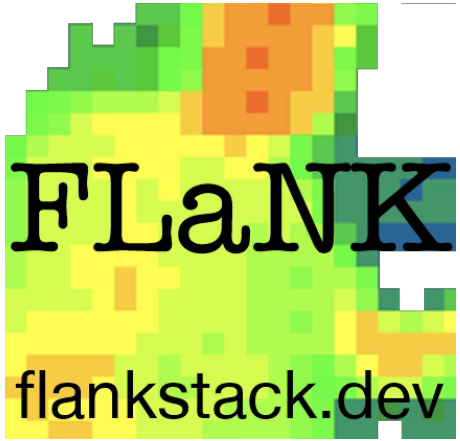
@PaasDev

https://medium.com/cloudera-inc/streaming-llm-with-apache-nifi-huggingface-ad2f0d367468

Streaming 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

FLaNK Stack Weekly

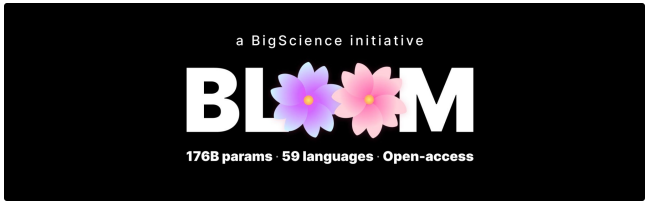


https://bit.ly/32dAJft

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

Generative AI

https://github.com/tspannhw/FLaNK-HuggingFace-DistilBert-SentimentAnalysis



watsonx.ai

CLOUDERA
Machine Learning

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

LLM USE CASE

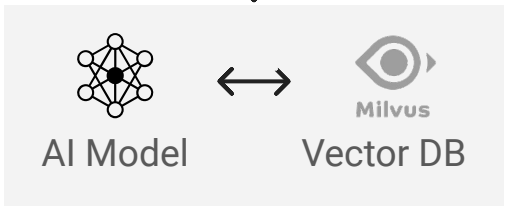
Unstructured file types

Structured Sources

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


COLLECTING DATA WITH NIFI
DATA INC
imgflip.com