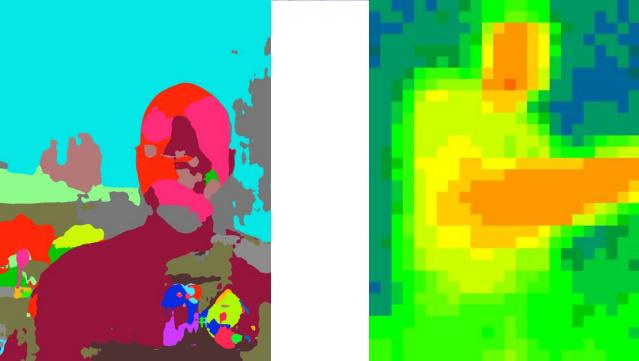




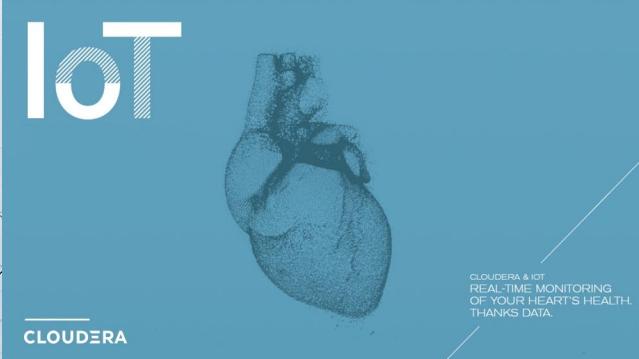
Data in Motion: Overview e Novidades do NiFi, Kafka e Flink

Tim Spann - Principal Developer Advocate
Data In Motion



**ENTERPRISE
DATA CLOUD**

CLOUDERA



CLOUDERA



**EDGE
2AI**

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

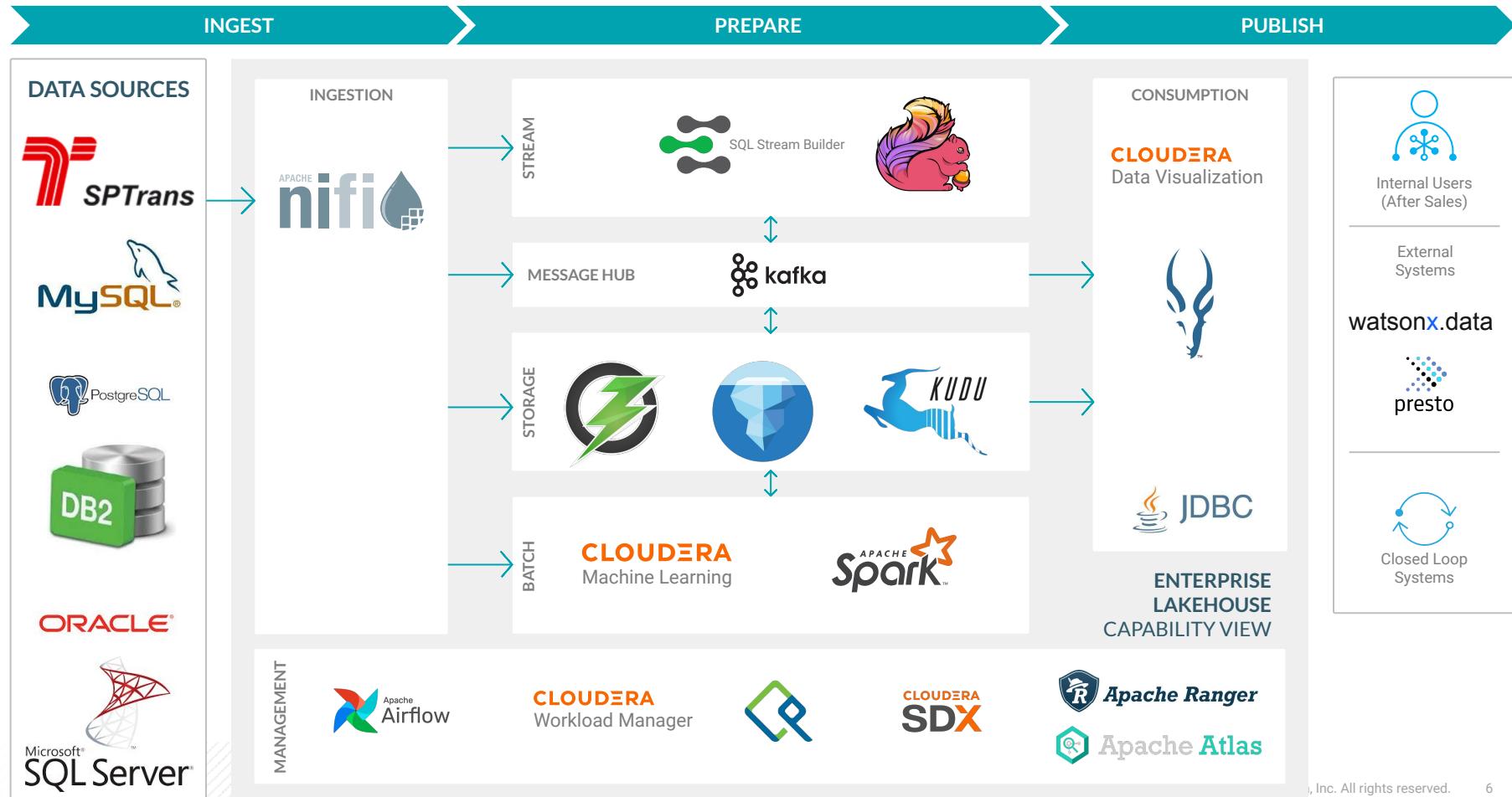
Demos

Q&A

REAL-TIME REQUIRES A PLATFORM



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

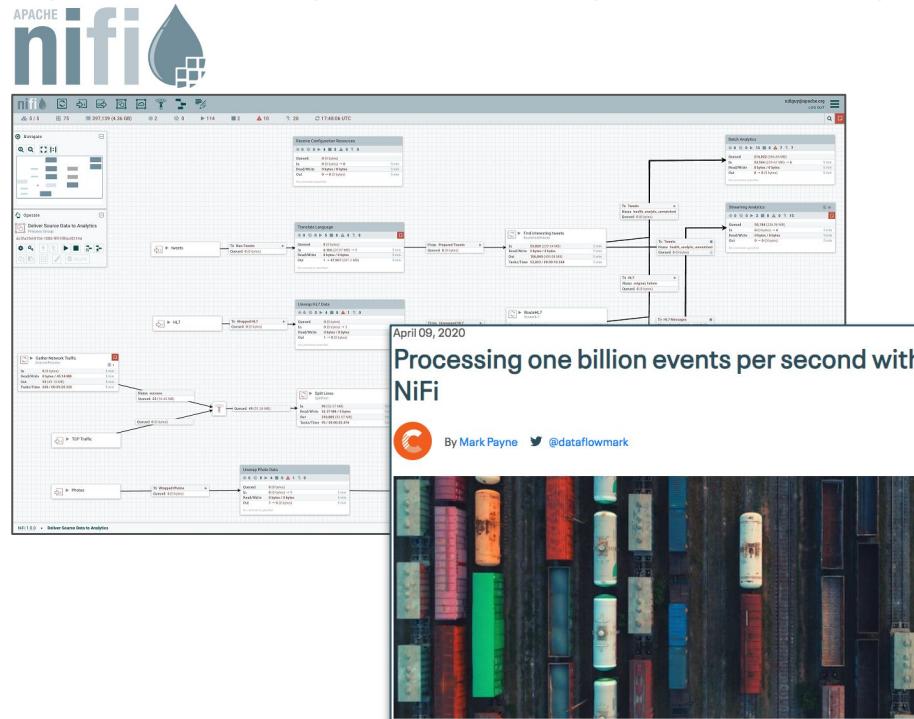


Cloudera DataFlow - Apache NiFi



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

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

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 window titled "Provenance Event".

Provenance Event

DETAILS ATTRIBUTES CONTENT

Attribute Values

filename	328717796819631
kafka.offset	44815
kafka.partition	6
kafka.topic	nifi-testing
path	/
uuid	328717796819631-0000-0000-0000-000000000000

RECORD-ORIENTED DATA WITH NIFI

- **Record Readers** - Avro, CSV, Grok, IPFIX, JSON1, 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.

Filter Events	
QueryRecord 1.13.2.2.2.2.0-127 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

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
Record Reader	CSVReader
Record Writer	JsonRecordSetWriter

+

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.

Filter Events		
QueryRecord 1.13.2.2.2.2.0-127 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

Configure Processor | QueryRecord 1.13.2.2.2.2.0-127

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
Record Reader	Syslog_5424_Reader
Record Writer	JSON_Syslog_5424_Writer
Include Zero Record FlowFiles	false
Cache Schema	false
Default Decimal Precision	10
Default Decimal Scale	0
filtered_events	#(Filter Rule)

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

The screenshot shows the Cloudera DataFlow ReadyFlow Gallery interface. On the left is a dark sidebar with the Cloudera DataFlow logo, navigation links for Dashboard, Catalog, ReadyFlow Gallery, and Environments, and footer links for Help and alpha_intcookieuser cookie. The main area is titled "ReadyFlow Gallery" and contains six flow definitions:

- Kafka filter to Kafka**: Version 1. Consumes JSON, CSV or Avro events from Kafka, filters them before writing them back to Kafka as JSON, CSV or Avro. [View Added Flow Definition](#)
- Kafka to Cloudera Operational Database**: Version 1. Consumes JSON, CSV or Avro events from Kafka and ingests them into Cloudera Operational Database (COD). [View Added Flow Definition](#)
- Kafka to Kafka**: Version 1. Consumes events from Kafka and writes them to another Kafka topic. [Add To Catalog](#)
- Kafka to Kudu**: Version 1. Consumes JSON, CSV or Avro events from Kafka and ingests them into Kudu. [Add To Catalog](#)
- Kafka to S3 Avro**: Version 1. Consumes JSON, CSV or Avro events from Kafka and writes Avro files to S3. [View Added Flow Definition](#)
- S3 to S3 Avro**: Version 1. Consumes JSON, CSV or Avro files from source S3 location and writes Avro files to a destination S3 location. [Add To Catalog](#)

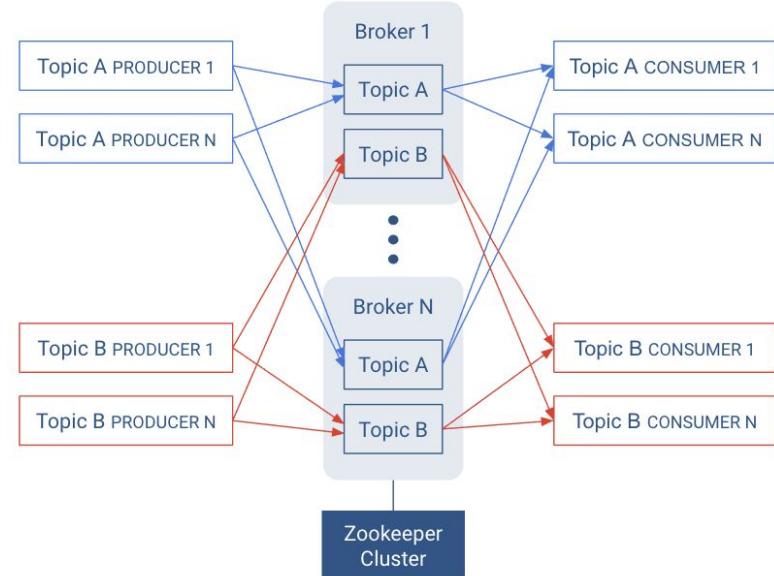
Cloudera Streams Messaging Manager - Apache Kafka

STREAMS MESSAGING WITH KAFKA



WriteToKafka	PublishKafka2RecordCDP 1.0.0.2.2.2.0-127 com.cloudera - nifi-cdf-kafka-2-nar	5 min
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

- 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.

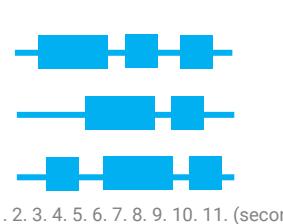


Cloudera SQL Stream Builder - Flink SQL

DELIVERING STREAMING ANALYTICS

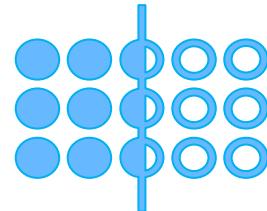
Capture Events that Matter

Low-latency analytics use cases



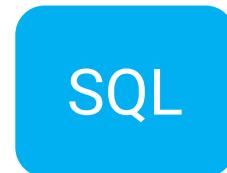
Parsing and Blending Data

Both offline and streaming data

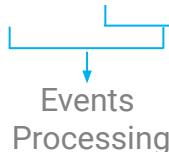


Data Analysts Can Write Queries

Across the Lines of Business



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

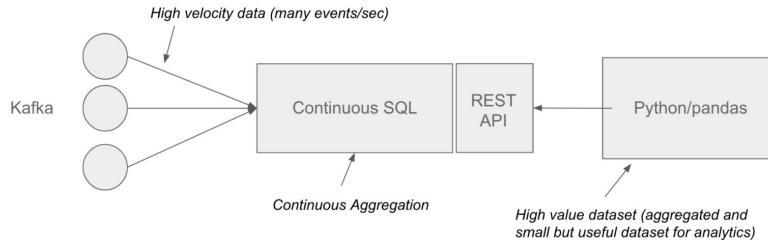
Enrich streaming data with batch data in a single tool

```
CREATE TABLE `kafka_table_1670513700` (
  `col_str` STRING,
  `col_int` INT,
  `col_ts` TIMESTAMP(3),
  WATERMARK FOR `col_ts` AS col_ts - INTERVAL '5' SECOND
) WITH (
  'connector' = 'kafka', -- Specify what connector to use, for Kafka it must use 'kafka'.
  'format' = 'json', -- Topic name to read from.
  'topic' = '...', -- Comma separated list of Kafka brokers.
  'bootstrap.servers' = '...', -- Optional flag to specify whether to encode all decimals as plain numbers instead of
  Note, only one of 'topic-pattern' and 'topic' can be specified for sources. When the table is used as sink, the topic name is the topic to write
  data to. Note topic list is not supported for sinks.
  'json.decode.decimals-as-plain-number' = 'false' -- Optional flag to parse integers as plain numbers by default.
  'parse-as-records' = 'true' -- Optional flag to parse records by default.
  'topic' = '...', -- To read from a topic when the table is used as source. It also supports topic list for source by separating topic by semicolon.
  'ignore-key-errors' = 'false' -- Optional flag to skip fields and rows with parse errors instead of failing; fields are set to null in
  case of errors, false by default.
  'ignore-parse-failures' = 'false' -- Optional flag to skip fields and rows with fail if a field is missing or not, false by default.
  'map-null-key.literal' = 'null' -- Optional flag to specify string literal for null keys when 'map-null-key.mode' is LITERAL, '\\"null\\"' by
  default.
  'map-null-key.mode' = 'FAIL' -- Optional flag to control the handling mode when serializing null key for map data, FAIL by default.
  Option DROP will drop null key entries for map data. Option LITERAL will use 'map-null-key.literal' as key literal.
)
```

Logs Results Events

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
```



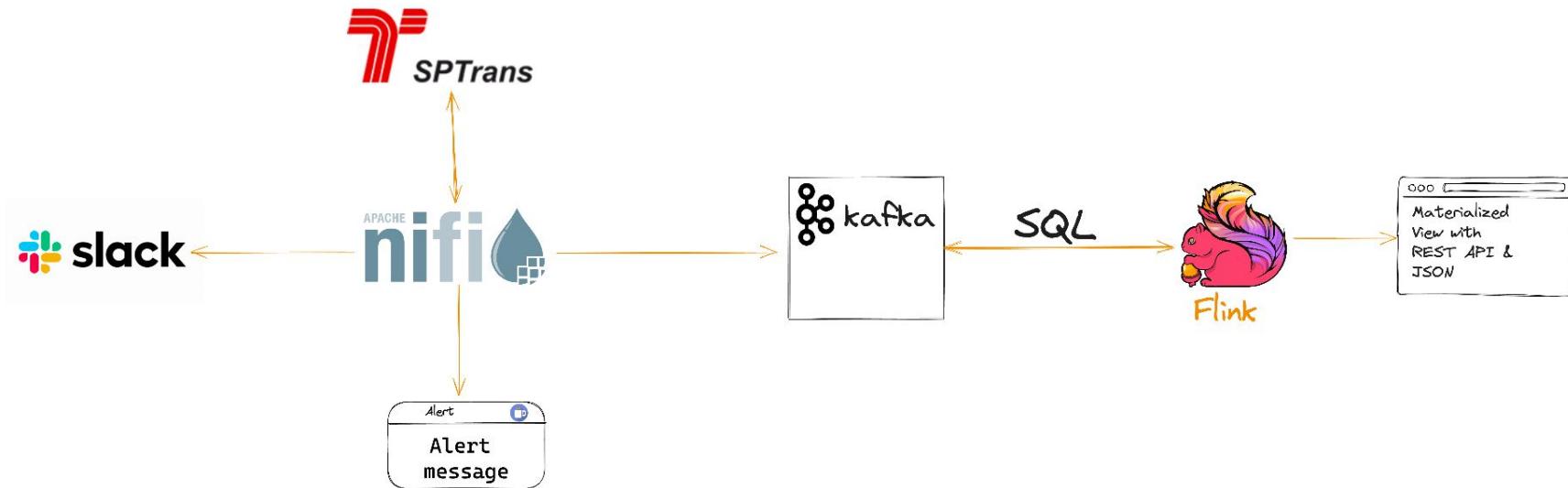
```
[90]: import pandas as pd
[91]: mv = "https://xxxxxxxxxx"
[92]: df = pd.read_json(mv)
[93]: len(df.keys())
[93]: 5
[95]: df['ts'] = pd.to_datetime(df['ts'])
[97]: df.dtypes
[97]: max_amount      int64
       sum_amount      int64
       thecount        int64
       ts              datetime64[ns]
       userid          int64
       dtype: object
[98]: df.set_index('userid').sort_values(by=['thecount'], ascending=False).head()
[98]:
       max_amount  sum_amount  thecount      ts
userid
    787      34911     57304     10 2020-06-16 19:52:15
    744      77407     95407      9 2020-06-16 19:52:15
     78      88761     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

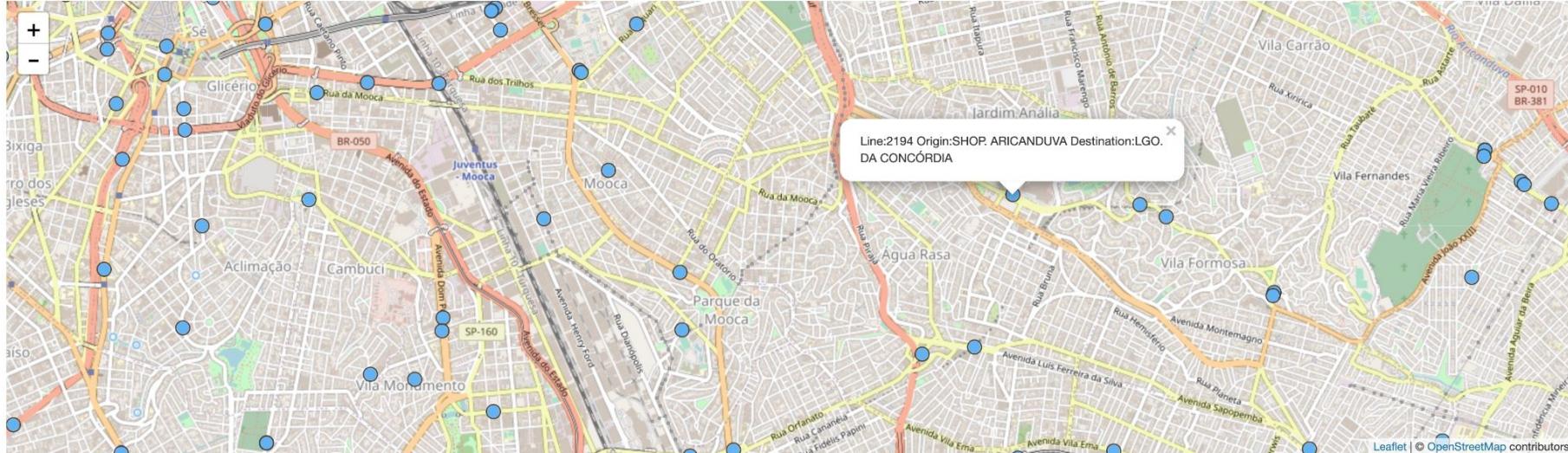


Data in Motion: Overview e Novidades do NiFi, Kafka e Flink

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



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

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

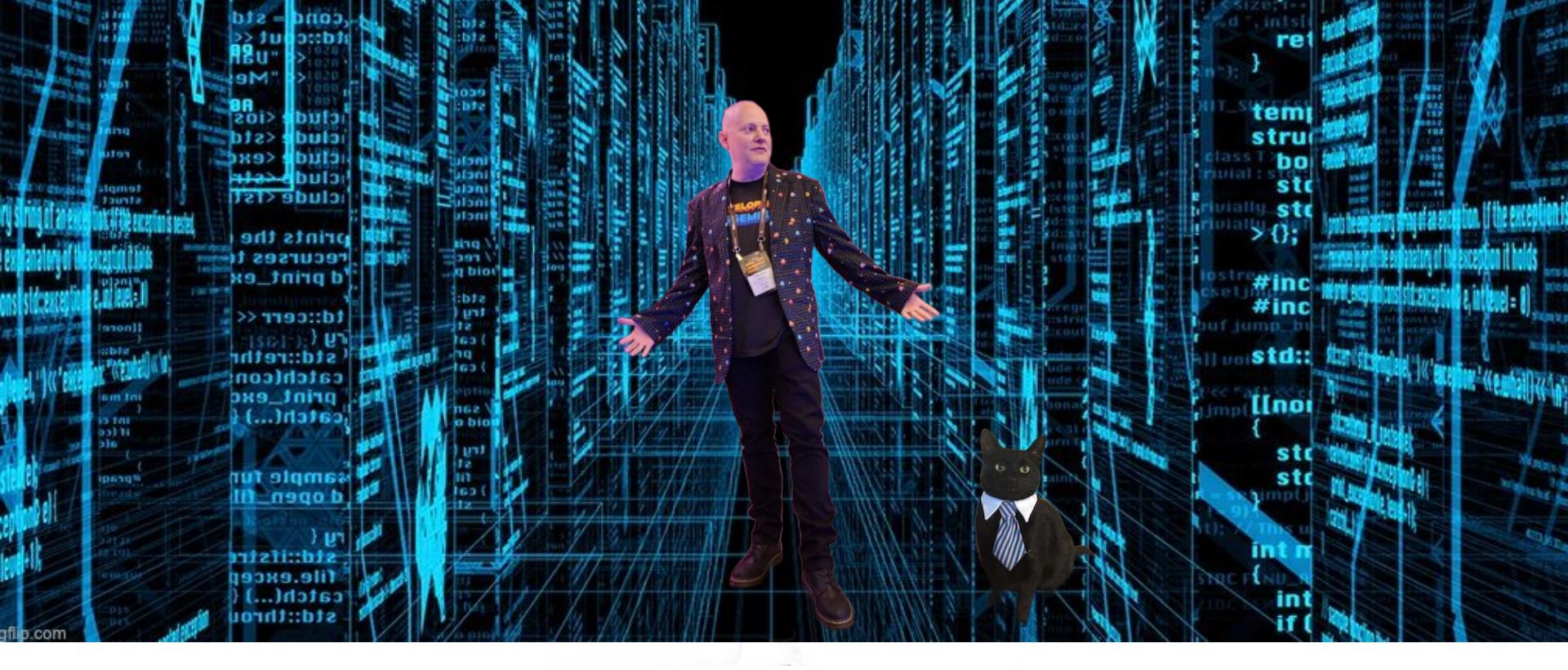
From Big Data to AI to Streaming to LLM to Cloud to
Analytics to NLP to Fast Data to Machine Learning to
Microservices to ...



CLOUDERA



@PaasDev

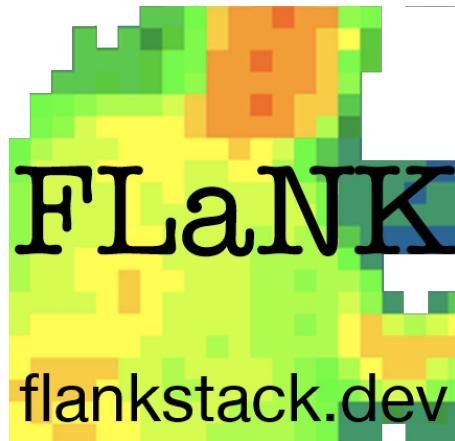


<https://medium.com/cloudera-inc/streaming-llm-with-apache-nifi-huggin-qface-ad2f0d367468>

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



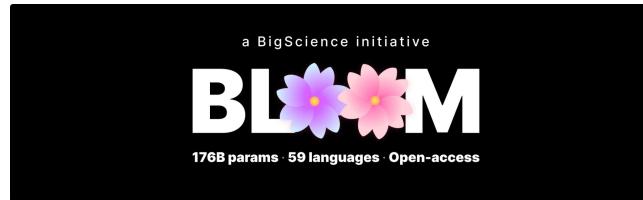
<https://bit.ly/32dAJft>



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>



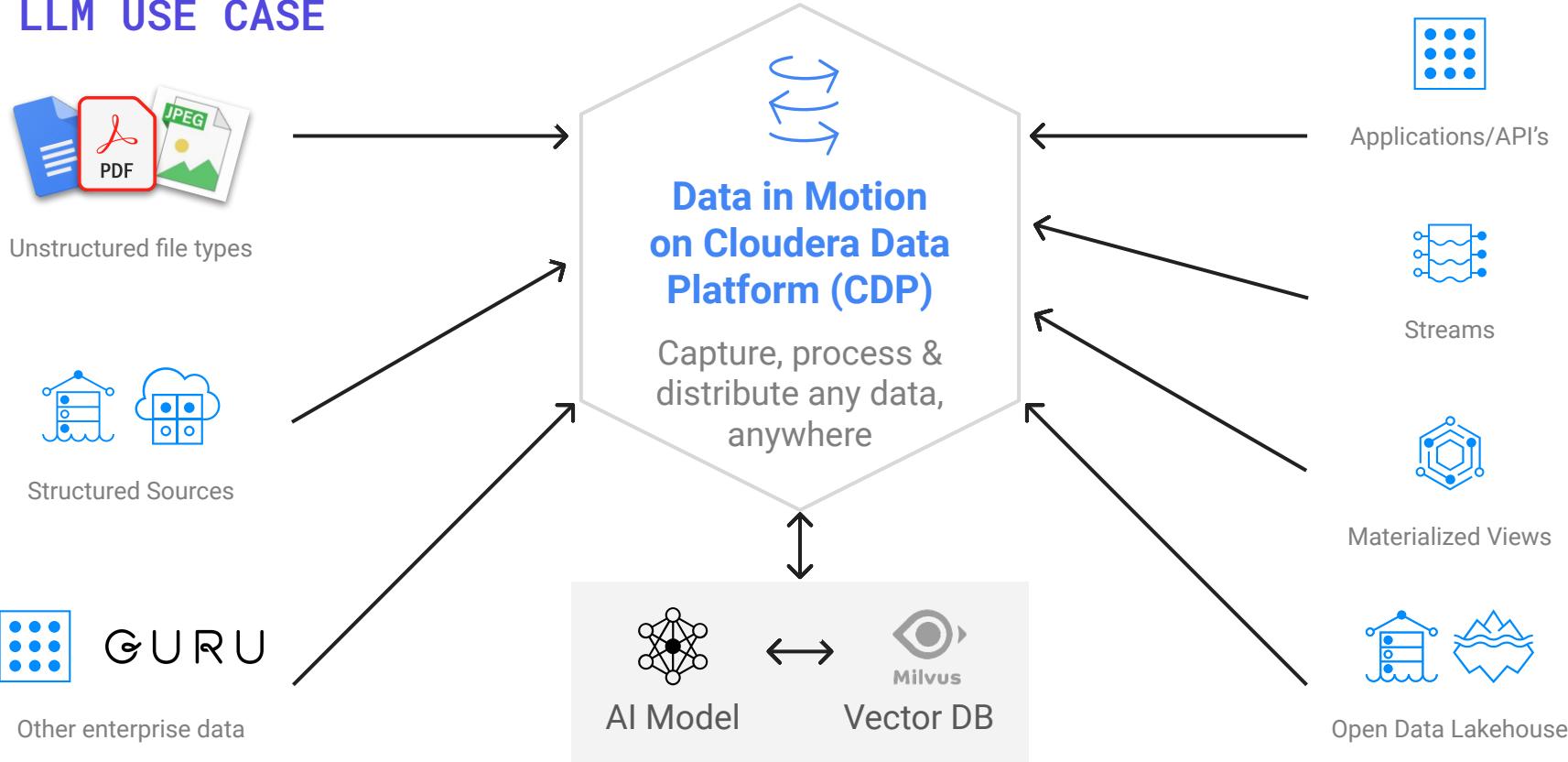
watsonx.ai

CLOUDERA
Machine Learning

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



LLM USE CASE



COLLECTING DATA WITH NIFI

DATA INC

softsys

TH^ON^G Y^OU