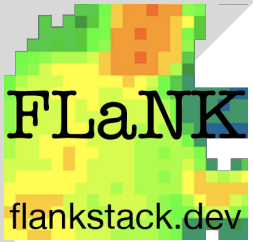
CLOUDERA

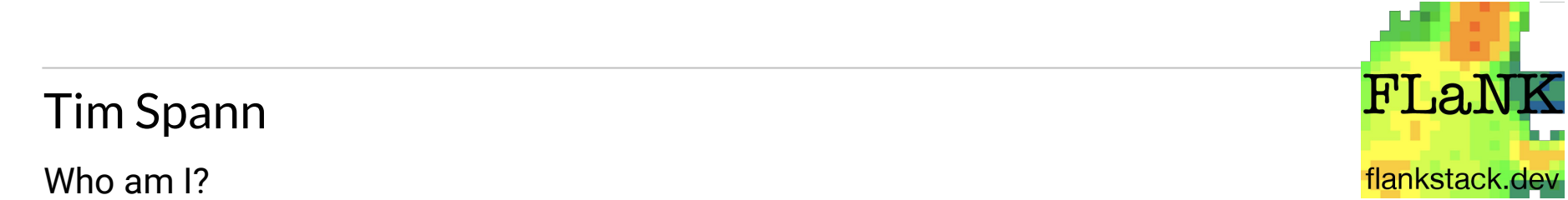


Introduction to the FLaNK Stack

Timothy Spann
Principal DataFlow Field Engineer
Cloudera
@**PaasDev**

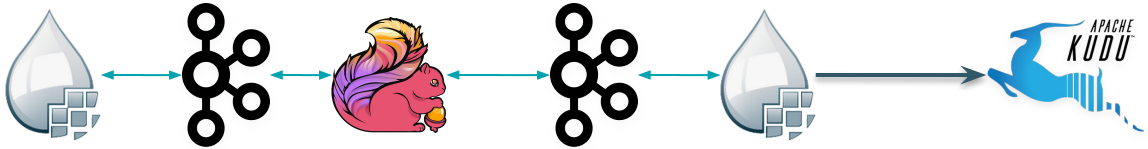
Tim Spann

Who am I?



Cloudera Principal DataFlow Field Engineer

DZone Zone Leader and Big Data MVB
Future of Data Meetup Leader
ex-Pivotal Field Engineer
https://github.com/tspannhw https://www.datainmotion.dev/



Welcome to Future of Data - Princeton - Virtual

FLaNK
flankstack.dev

FUTURE OF DATA

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

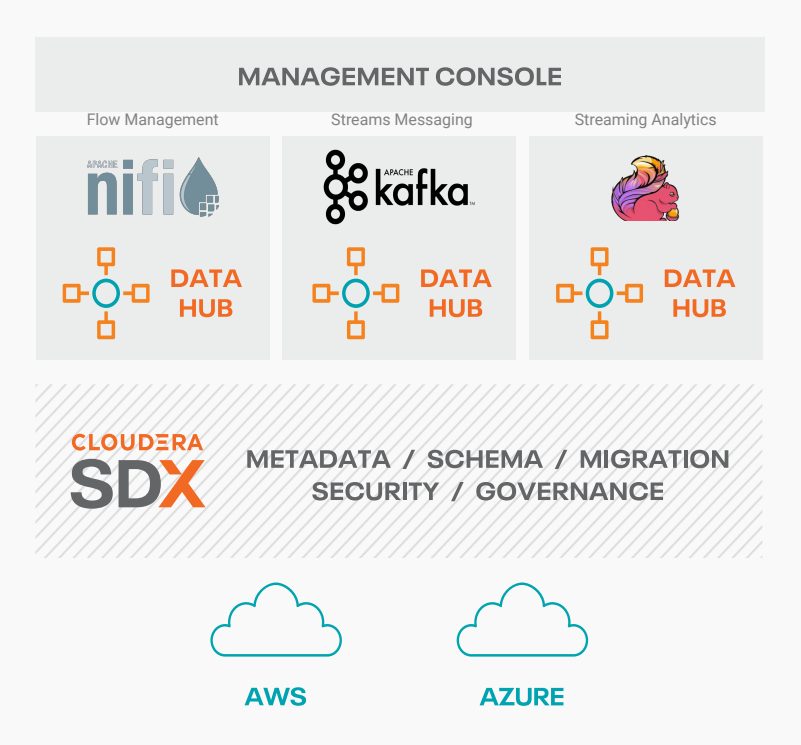
From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...

APACHE HBASE

APACHE KUDU

@PaasDev

Where Can I Run Edge AI Easily?



Web service hosted and managed by Cloudera

Hosted in the your cloud environment, but managed by the CDP Management Console

Shared Data Experience (SDX) technologies form a secure and governed data lake backed by object storage (S3, ADLS, GCS)

CDP services are optimized for the elastic compute & 'always-on' storage services provided by any cloud provider

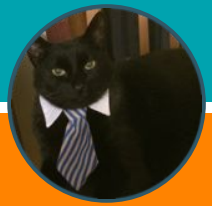
Edge AI for Big Data Engineers

Multiple users, frameworks, languages, devices, data sources & clusters

FLaNK
flankstack.dev

CLOUD DATA ENGINEER

- Experience in ETL/ELT
- Coding skills in Python or Java
- Knowledge of database query languages such as SQL
- Experience with Streaming
- Knowledge of Cloud Tools



CAT

- Expert in ETL (Eating, Ties and Laziness)
- Edge Camera Interaction
- Typical User
- No Coding Skills
- Can use NiFi
- Questions your cloud spend

AI / Deep Learning / ML / DS

- Can run in Apache NiFi
- Can run in Kafka Streams
- Can run in Apache Flink
- Can run in MiNiFi
- Can run in Cloudera Machine Learning
- Use Native ML/DL in Jetson, Movidius, Coral, TPU/GPU on Edge Devices

Streaming Data Pipelines with Apache NiFi + Kafka + Flink

Apache Tools and Frameworks Used

FLaNK
flankstack.dev

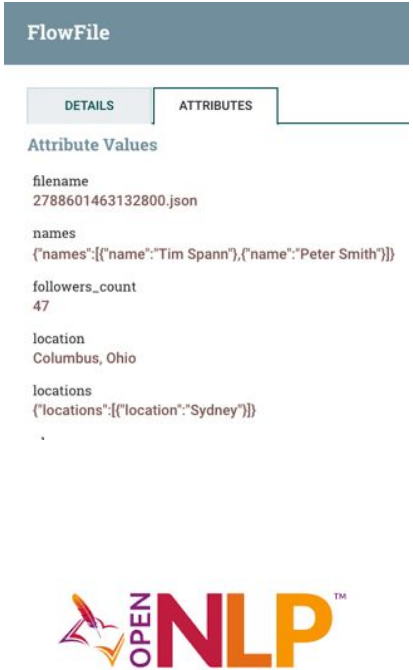
Apache OpenNLP with Apache NiFi

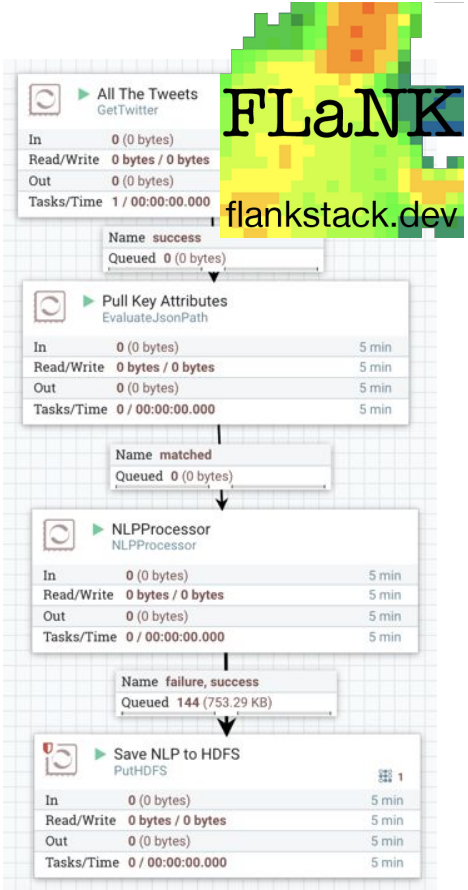
Apache OpenNLP for Entity Resolution Processor
https://github.com/tspannhw/nifi-nlp-processor

Requires installation of NAR and Apache OpenNLP Models
http://opennlp.sourceforge.net/models-1.5/

This is a non-supported processor that I wrote and put into the community. You can write one too!

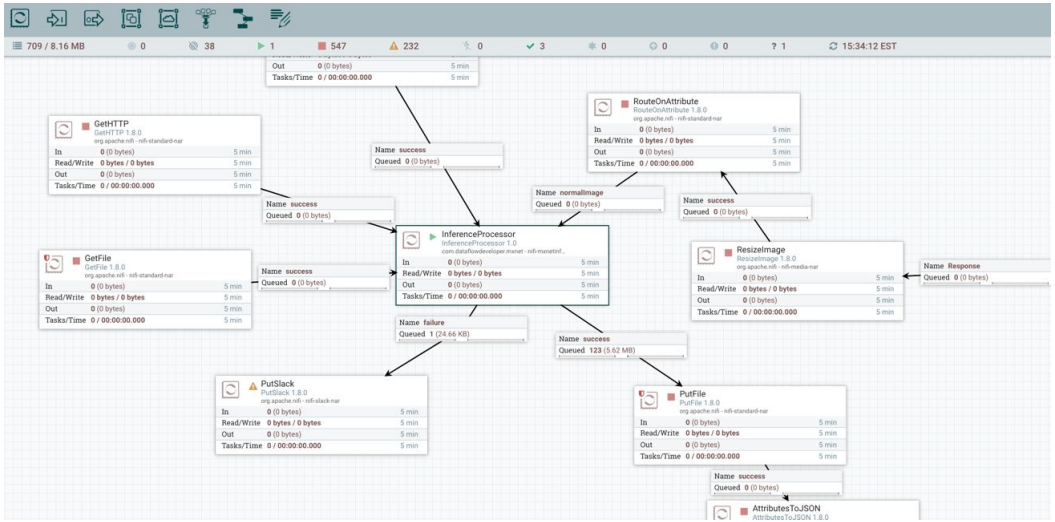
https://opennlp.apache.org/

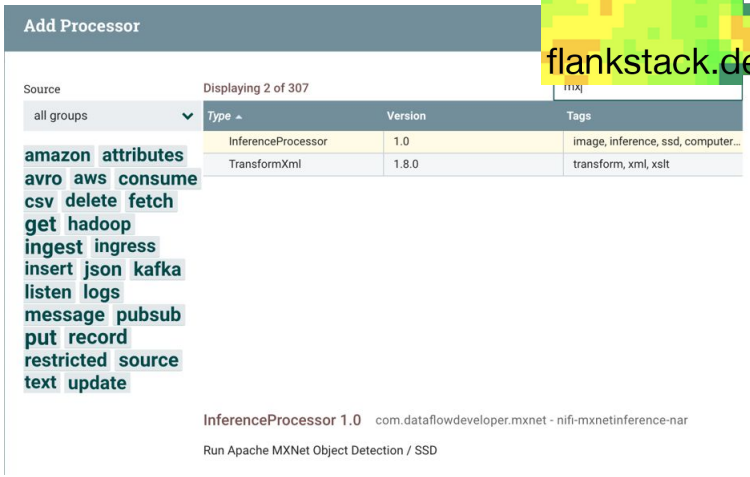




https://community.cloudera.com/t5/Community-Articles/Open-NLP-Example-Apache-NiFi-Processor/ta-p/249293

Apache MXNet Native Processor for Apache NiFi





This is a beta, community release by me using the new beta Java API for Apache MXNet.

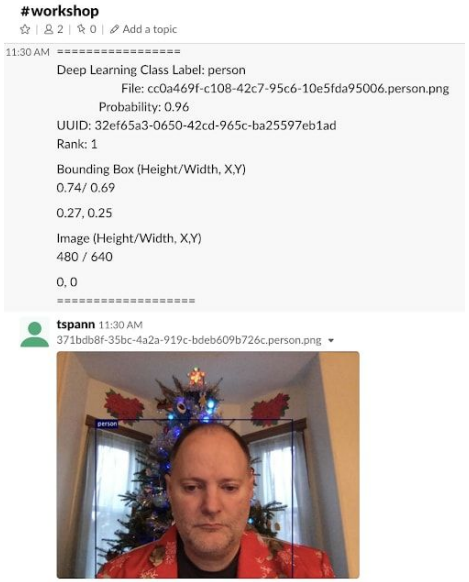
https://github.com/tspannhw/nifi-mxnetinference-processor

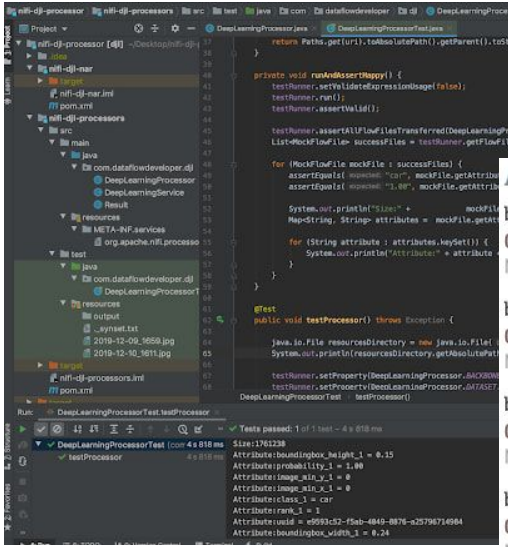
https://community.hortonworks.com/articles/229215/apache-nifi-processor-for-apache-mxnet-ssd-single.html

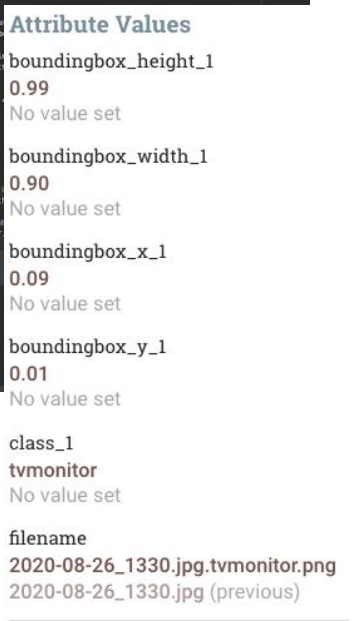
https://www.youtube.com/watch?v=Q4dSGPvqXSA

Apache MXNet Native Processor through DJL.AI for Apache NiFi

DJL







FLaNK

flankstack.dev

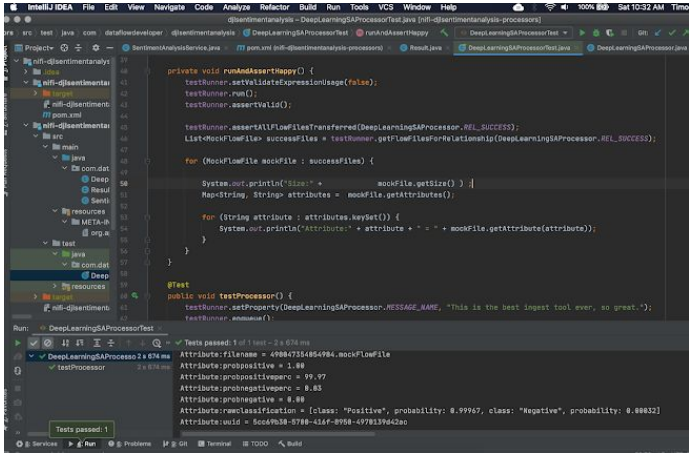
This processor uses the DJL.AI Java Interface

https://github.com/tspannhw/nifi-djl-processor

https://dev.to/tspannhw/easy-deep-learning-in-apache-nifi-with-djl-2d79

DJL NiFi Processors - Sentiment Analysis

DJL



probnegative

0.99

No value set

probnegativeperc

99.44

No value set

probpositive

0.01

No value set

probpositiveperc

0.56

No value set

rawclassification

[class: "Negative", probability: 0.99440, class: "Positive", probability: 0.00559]

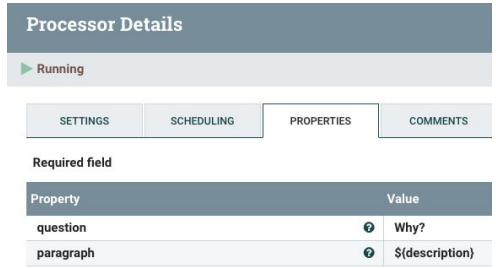
https://www.datainmotion.dev/2020/09/using-djlai-for-deep-learning-based.html

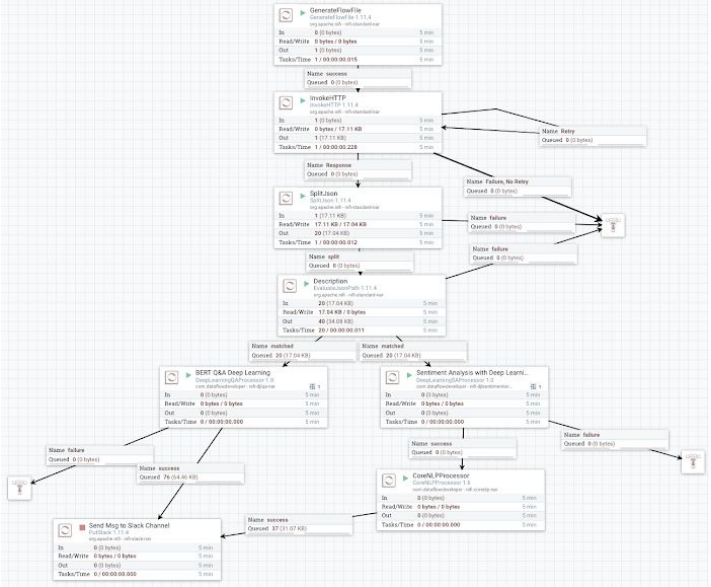
https://github.com/tspannhw/nifi-djlsentimentanalysis-processor

BERT QA through DJL.AI for Apache NiFi

BERT QA

DJL





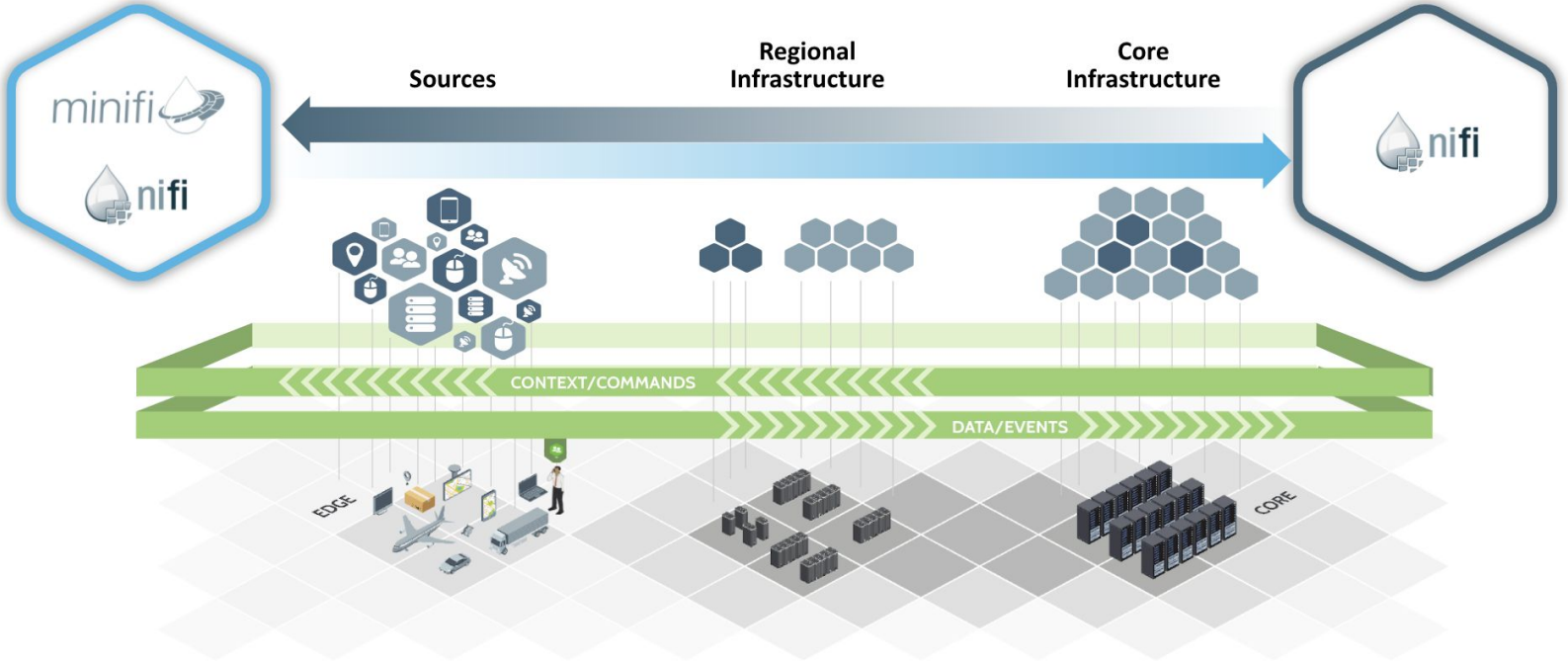
prediction
[the, data, analytics]

The pretrained model is DistilBERT model trained by HuggingFace using PyTorch.

This processor uses the DJL.AI Java Interface

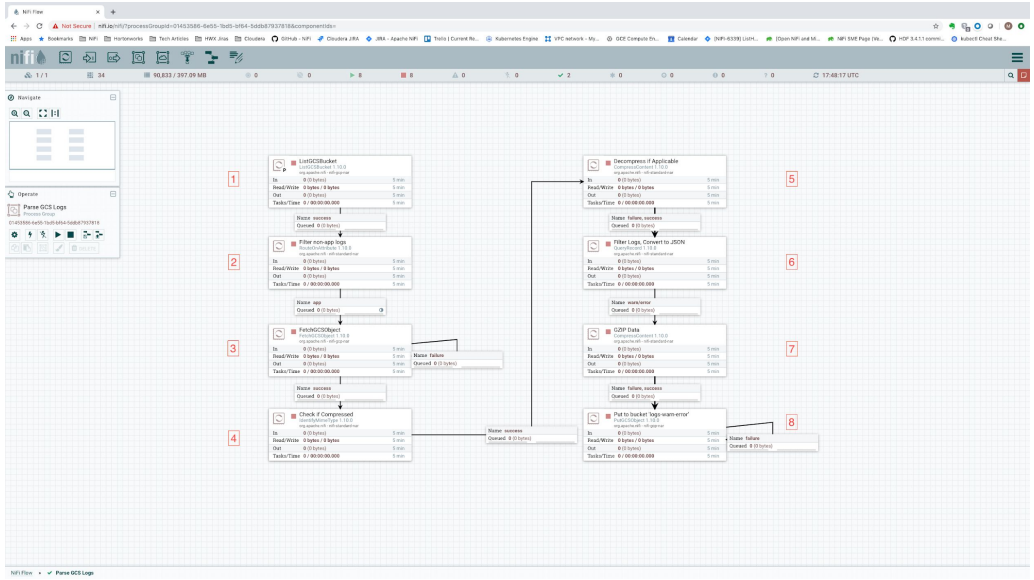
https://github.com/tspannhw/nifi-djlqa-processor
https://www.datainmotion.dev/2020/09/using-djlai-for-deep-learning-bert-q-in.html
https://dev.to/tspannhw/easy-deep-learning-in-apache-nifi-with-djl-2d79

What is Apache NiFi and MiNiFi used for?

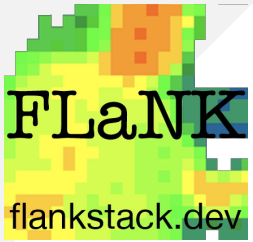


NiFi Processing Billions of Events

https://blog.cloudera.com/benchmarking-nifi-performance-and-scalability/

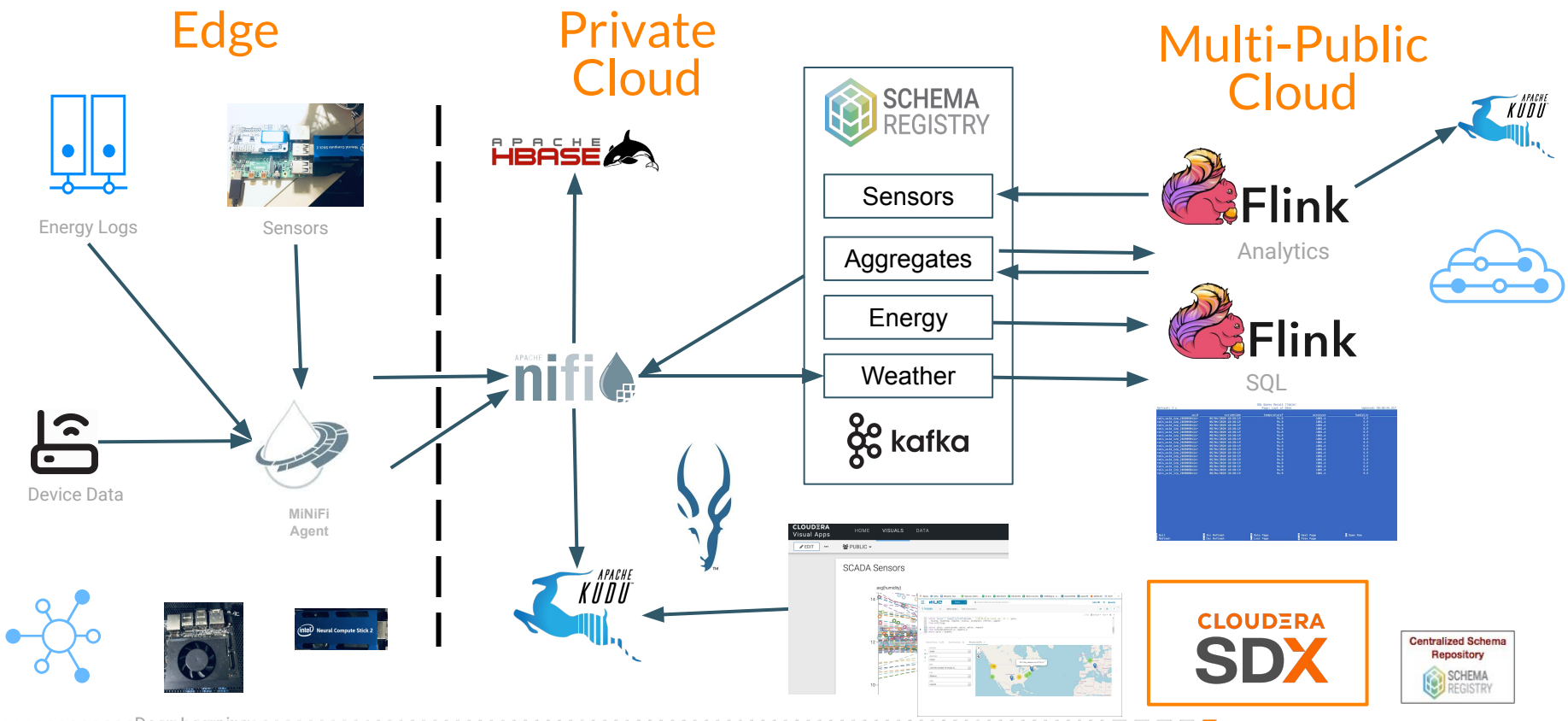


Nodes	Data rate/sec	Events/sec	Data rate/day	Events/day
1	192.5 MB	946,000	16.6 TB	81.7 Billion
5	881 MB	4.97 Million	76 TB	429.4 Billion
25	5.8 GB	26 Million	501 TB	2.25 Trillion
100	22 GB	90 Million	1.9 PB	7.8 Trillion
150	32.6 GB	141.3 Million	2.75 PB	12.2 Trillion

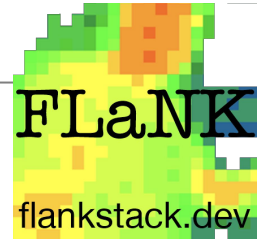


Demo

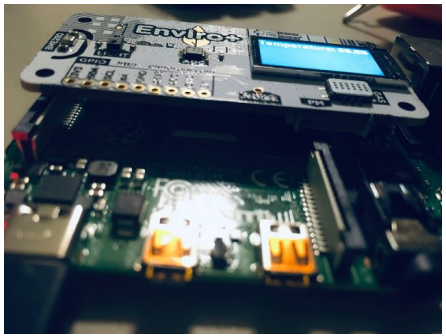
Edge AI to Cloud Streaming Pipeline



SHOW ME THE DATA

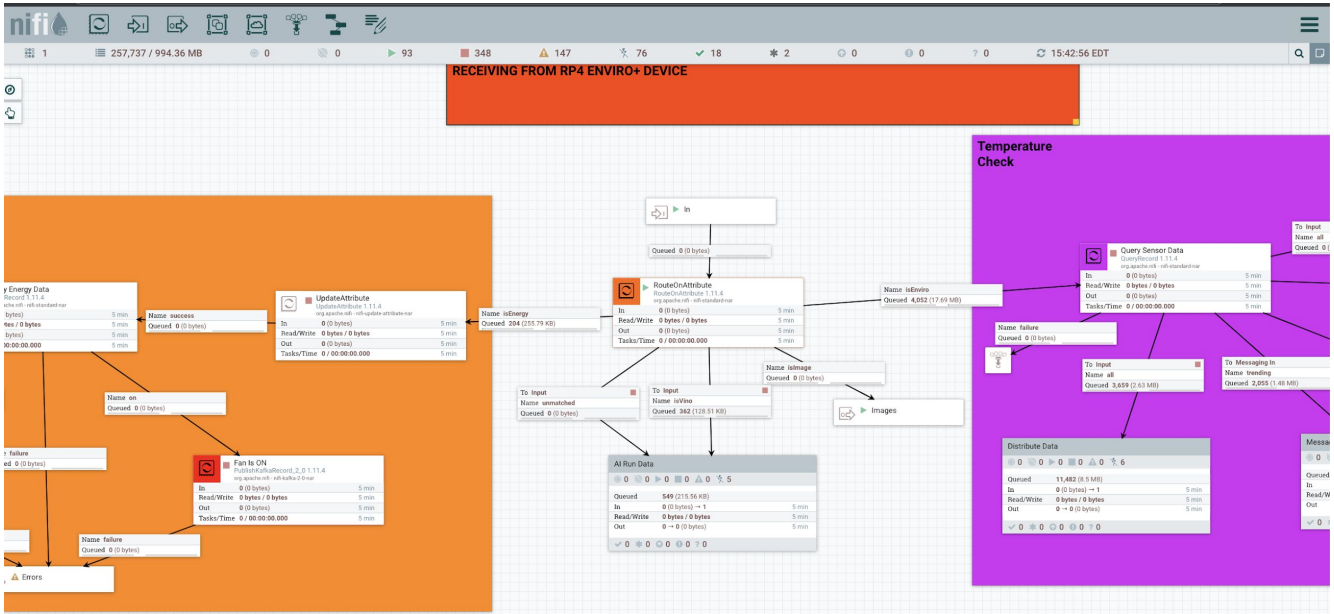


{"uuid": "rpi4_uuid_jfx_20200826203733", "amplitude100": 1.2, "amplitude500": 0.6, "amplitude1000": 0.3, "lownoise": 0.6, "midnoise": 0.2, "highnoise": 0.2, "amps": 0.3, "ipaddress": "192.168.1.76", "host": "rp4", "host_name": "rp4", "macaddress": "6e:37:12:08:63:e1", "systemtime": "08/26/2020 16:37:34", "endtime": "1598474254.75", "runtime": "28179.03", "starttime": "08/26/2020 08:47:54", "cpu": 48.3, "cpu_temp": "72.0", "diskusage": "40219.3 MB", "memory": 24.3, "id": "20200826203733_28ce9520-6832-4f80-b17d-f36c21fd8fc9", "temperature": "47.2", "adjtemp": "35.8", "adjtempf": "76.4", "temperaturef": "97.0", "pressure": 1010.0, "humidity": 8.3, "lux": 67.4, "proximity": 0, "oxidising": 77.9, "reducing": 184.6, "nh3": 144.7, "gasKO": "Oxidising: 77913.04 Ohms\nReducing: 184625.00 Ohms\nNH3: 144651.47 Ohms"}

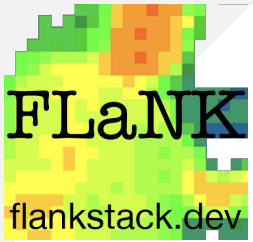


WHERE DID THAT DATA COME FROM?

BME280 - temperature, pressure, humidity sensor
LTR-559 - light and proximity sensor
MICS6814 - analog gas sensor
ADS1015 ADC
MEMS - microphone
0.96-inch, 160 x 80 color LCD



FLaNK
flankstack.dev

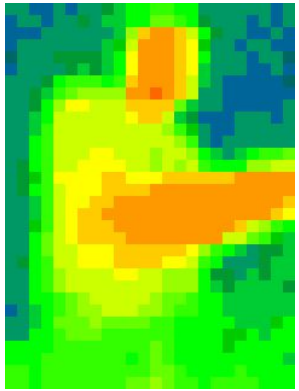


Learn More

DEMO SOURCE CODE

- https://github.com/tspannhw/FlinkForwardGlobal2020
- https://github.com/tspannhw/ApacheConAtHome2020
- https://github.com/tspannhw/minifi-xaviernx
- https://github.com/tspannhw/minifi-jetson-nano
- https://github.com/tspannhw/minifi-enviroplus
- https://github.com/tspannhw/EverythingApacheNiFi

The code, build scripts, schemas, table DDL, Flink SQL, Kafka Connect configuration, NiFi flows, HBase tables, Kudu tables, Hive tables, HDFS directories, alerts, images, HTML, docs, links and all the goodies are here. Please **fork** and contribute.



THANK YOU