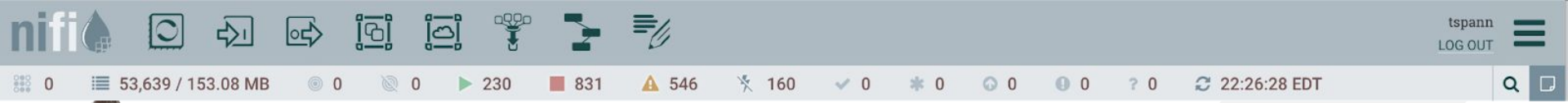
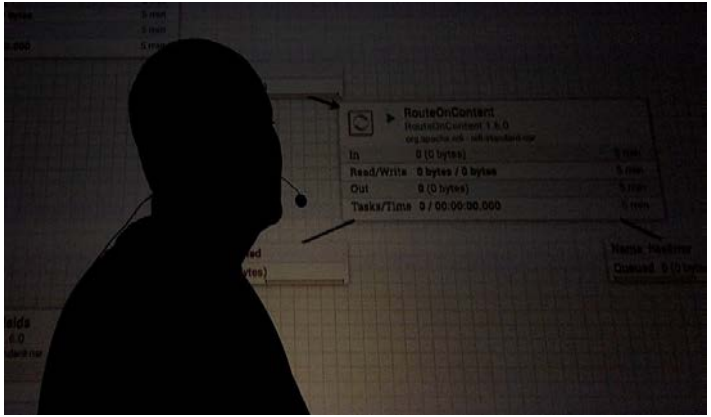


PULSAR

Using the FLiPN Pattern for Edge AI (Flink, NiFi, Pulsar)

David Kjerrumgaard, Developer Advocate - StreamNative
Tim Spann, Principal Developer Advocate - Cloudera



Tim Spann

Twitter: @**PaasDev** // Blog: **datainmotion.dev**

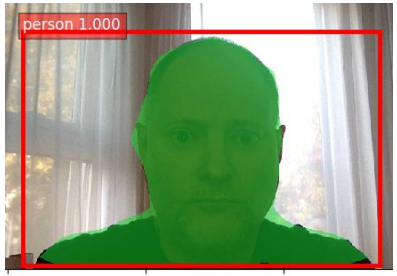
Principal Developer Advocate.

Princeton Future of Data Meetup.

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

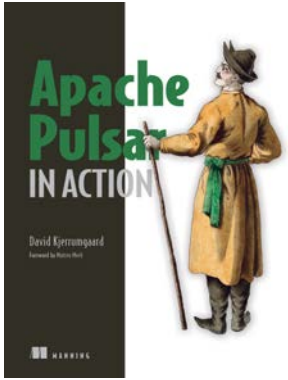
https://medium.com/@tspann

https://github.com/tspannhw



David Kjerrumgaard
Twitter: @**DavidKjerrumga1**
Developer Advocate - StreamNative
Published Author.
ex-AWS, ex-Hortonworks, ex-Splunk

https://github.com/david-streamlio



FLiPN Stack Weekly by Tim Spann

https://bit.ly/32dAJft

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

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

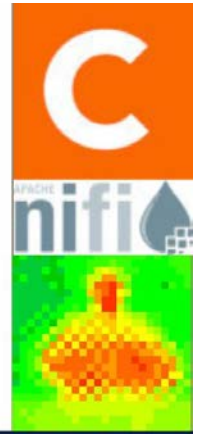
Future of Data - NYC + NJ + Philly + Virtual

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 ...

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY



ICEBERG

@PaasDev

Introduction

Overview

Examples

Apache Pulsar

Apache Flink

Apache NiFi

Demos

FLiP(N) Stack

- Apache **Fl**ink
- Apache **P**ulsar
- StreamNative's Flink Connector for Pulsar
- Apache **N**iFi
- Apache projects +++

Apache projects are the way for all streaming use cases.

FLiPN Pattern for Edge Data Engineers - Edge AI / IIoT

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

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

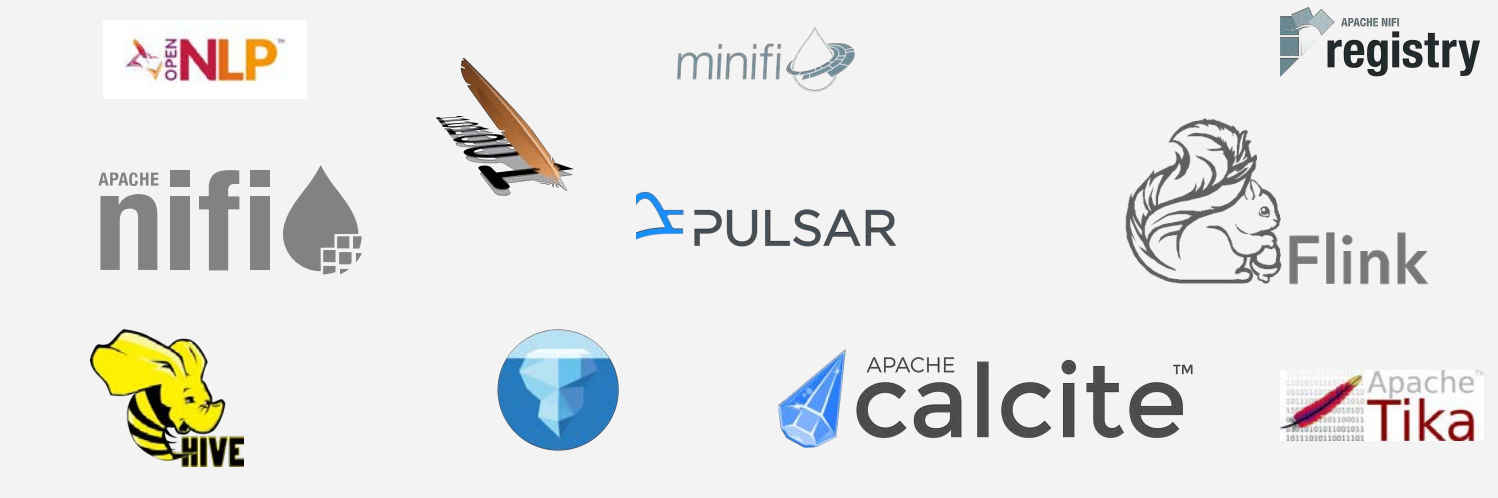
CAT

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

AI / Deep Learning / ML / DS

- Can run in Apache NiFi
- Can run in Apache Pulsar Functions
- Can run in Apache Flink
- Can run in Apache NiFi - MiNiFi Agents

Apache Tools and Frameworks Used



APACHE FLINK

3B+

3B+ data points daily streaming in from 25 million customers running real time machine learning prediction

USE CASE

Streaming real-time data pipelines that need to handle complex stream or batch data event processing, analytics, and/or support event-driven applications

TECHNOLOGY

Flink performs compute at in-memory speed at any scale

Flink parses SQL using Apache Calcite, which supports standard ANSI SQL

Flink runs standalone, on YARN, and has a K8s Operator

APPLICATION

Comcast a global media uses Flink for operationalizing machine learning models and near-real-time event stream processing

Flink helps deliver a personalized, contextual interaction reducing time to support resolutions saving millions of dollars per year

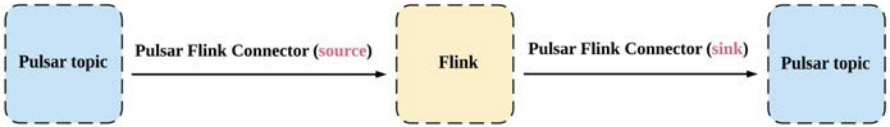
CONSIDERATION

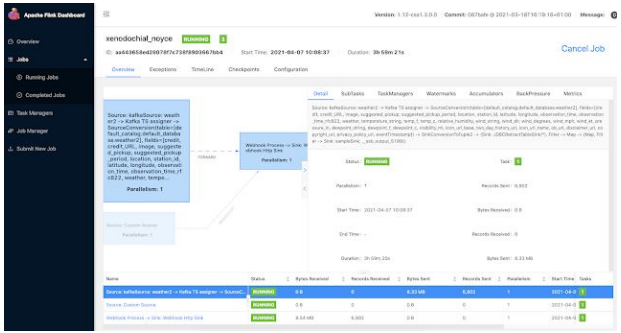
Data Freshness SLAs

Flink can read and write from Hive data

Review requirements for fault tolerance, resilience, and HA

Why Apache Flink?





- Unified computing engine
- Batch processing is a special case of stream processing
- Stateful processing
- Massive Scalability
- Flink SQL for queries, inserts against Pulsar Topics
- Streaming Analytics
- Continuous SQL
- Continuous ETL
- Complex Event Processing
- Standard SQL Powered by Apache Calcite

APACHE PULSAR

10 PB

10 Petabytes of data ingested daily from customers running real time cyber security detection platform

Pulsar

USE CASE

A horizontally scalable streaming platform that supports real-time data pipelines to perform event processing, analytics, and support event-driven applications.

TECHNOLOGY

Pulsar is a durable streaming platform with infinite scale.

Pulsar supports a variety of messaging protocols including, Kafka, MQTT, AMQP, etc.

Pulsar is designed to run in the cloud, and has a K8s Operator

APPLICATION

Splunk a global observability platform uses Pulsar to detect cyber-security threats in near-real-time.

Pulsar helps ingest data from a variety of sources and protocols with low latency and TCO. Saving millions of dollars per year in infrastructure cost.

CONSIDERATION

Ability to scale dynamically, with zero data movement

Fault tolerance, resiliency, and geo-replication

Apache Flink can easily read from and write data to Pulsar.

Why Apache Pulsar?

Everything Kafka can do, but better and more!

Messaging and streaming

- Message retention
- Built-in tiered storage
- Built-in stream processing
- Queuing semantics
- Dead letter queues
- Scheduled and delayed delivery
- Multi-protocol support

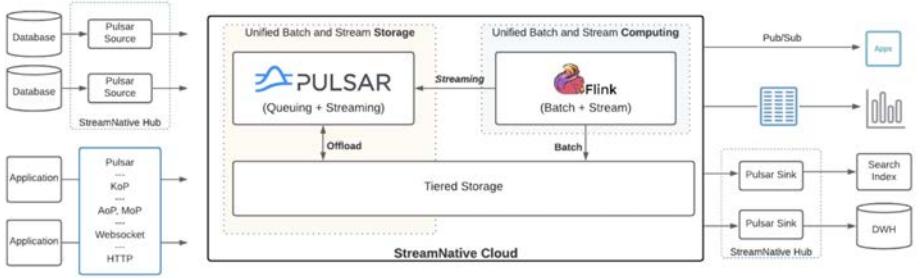
Performance and scaling

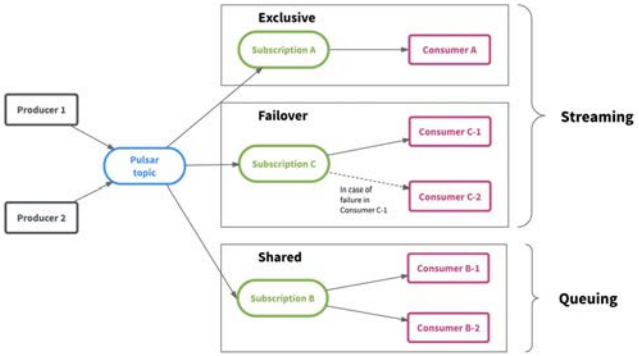
- Elastically scalable
- ***Rebalance-free*** scaling
- Up to **10 million** of topics

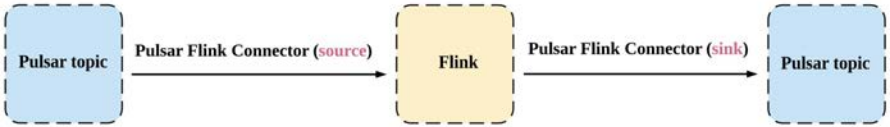
Management features

- Geo-replication
- Multi-tenancy
- Schema management
- End-to-end encryption

Flink + Pulsar (FLiP)

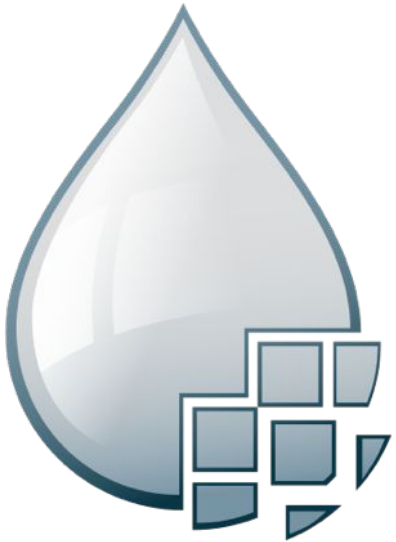






https://flink.apache.org/2019/05/03/pulsar-flink.html
https://github.com/streamnative/pulsar-flink
https://streamnative.io/en/blog/release/2021-04-20-flink-sql-on-streamnative-cloud

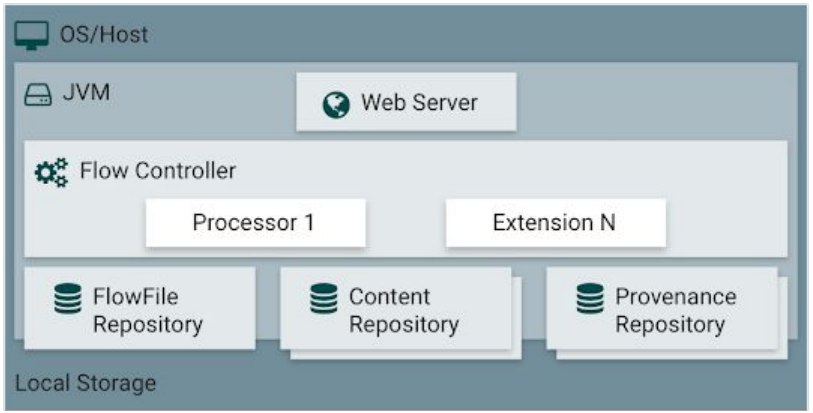
Why Apache NiFi?

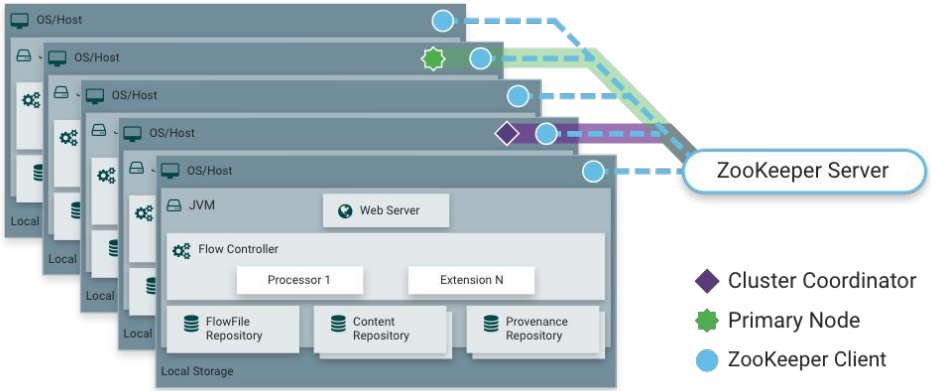


- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models

- Hundreds of processors
- Visual command and control
- Over a sixty sources
- Flow templates
- Pluggable/multi-role security
- Designed for extension
- Clustering
- Version Control

Architecture





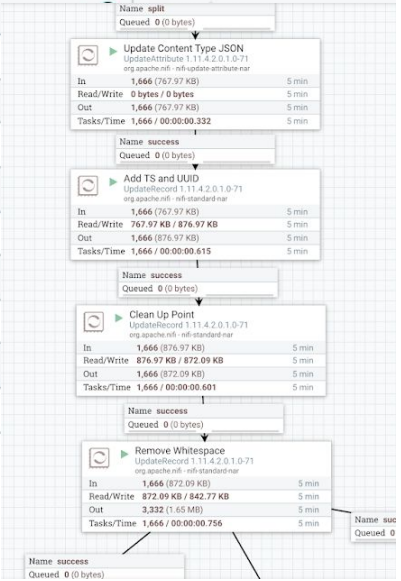
https://nifi.apache.org/docs/nifi-docs/html/overview.html

Record Processors

- XML, CSV, JSON, AVRO and more
- Schemas or Inferred Schemas
- Easily convert between them
- Support SQL with Apache Calcite

Property	Value
Record Reader	**XMLReader**
Record Writer	**JsonRecordSetWriter**
Include Zero Record FlowFiles	**false**
Cache Schema	**true**
query1	SELECT * FROM FLOWFILE

Property	Value
Schema Access Strategy	**Infer Schema**
Schema Registry	AvroSchemaRegistry
Schema Name	${schema.name}
Schema Version	
Schema Branch	
Schema Text	
Schema Inference Cache	
Expect Records as Array	
Attribute Prefix	
Field Name for Content	
Date Format	
Time Format	
Timestamp Format	

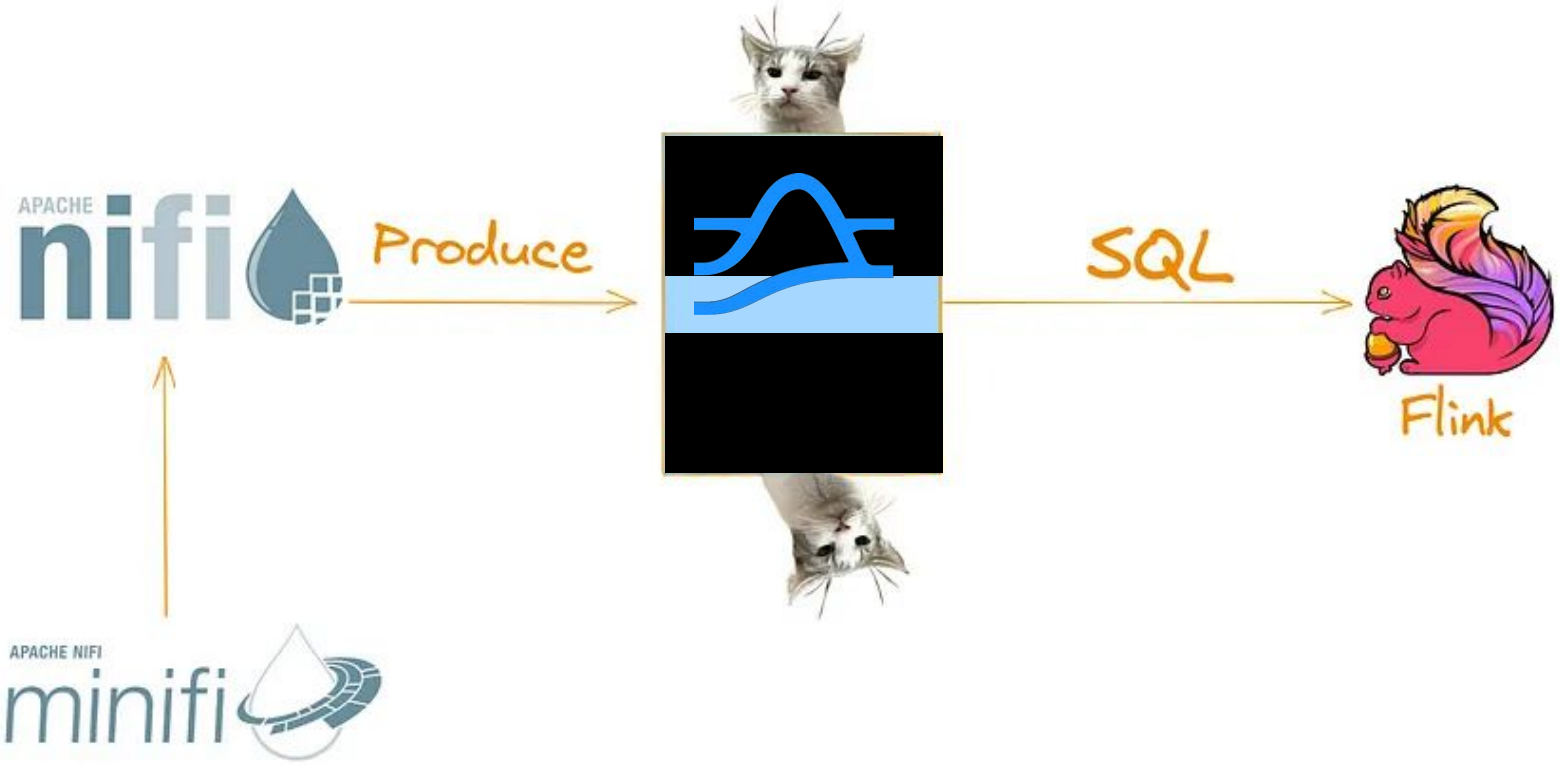


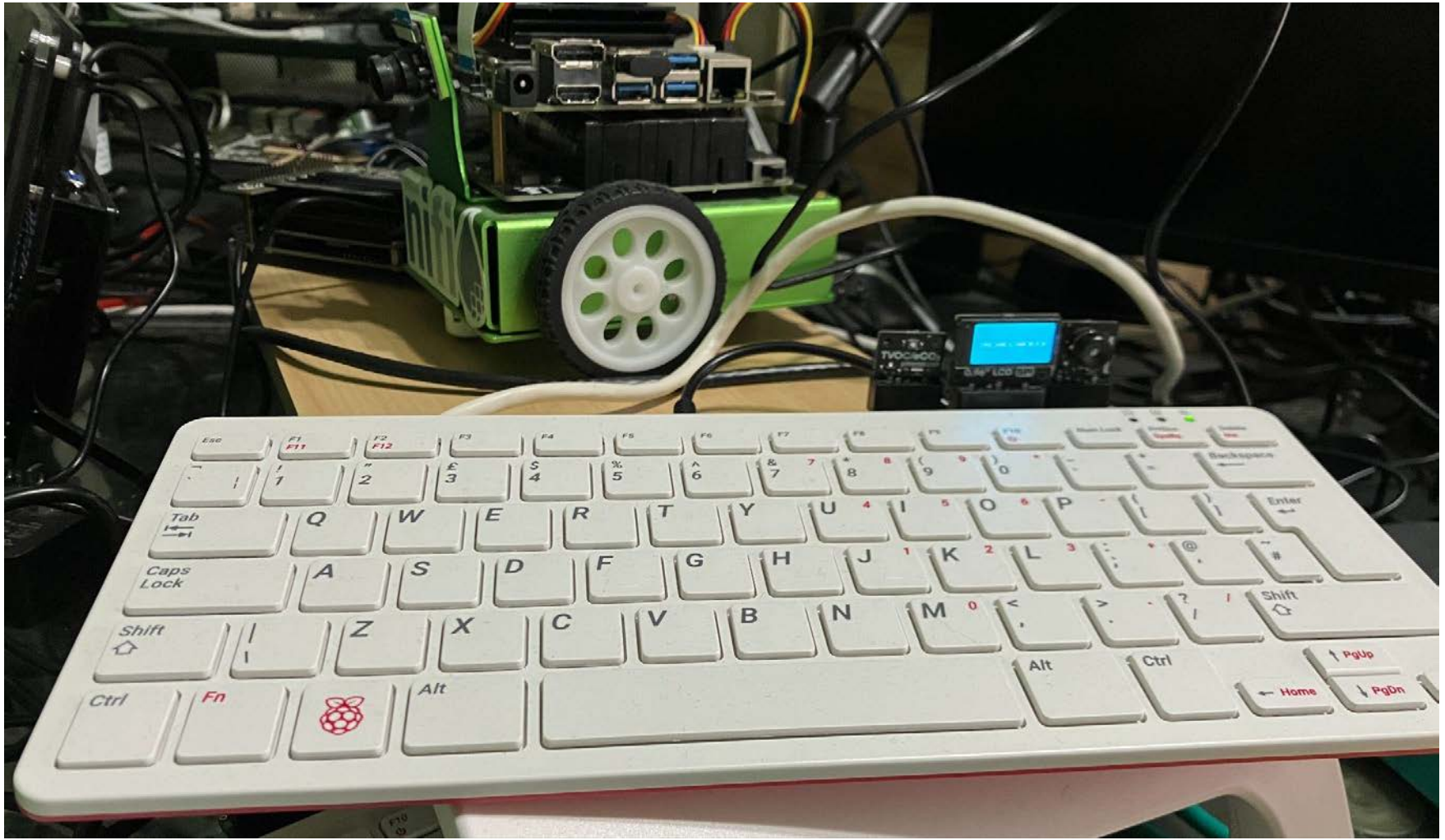
https://www.datainmotion.dev/2019/03/advanced-xml-processing-with-apache.html

Using NVIDIA Jetson Devices With Pulsar

https://dev.to/tspannhw/unboxing-the-most-amazing-edge-ai-device-part-1-of-3-nvidia-jetson-xavier-nx-595k
https://github.com/tspannhw/minifi-xaviernx/
https://github.com/tspannhw/minifi-jetson-nano
https://github.com/tspannhw/Flip-iot
https://www.datainmotion.dev/2020/10/flank-streaming-edgeai-on-new-nvidia.html

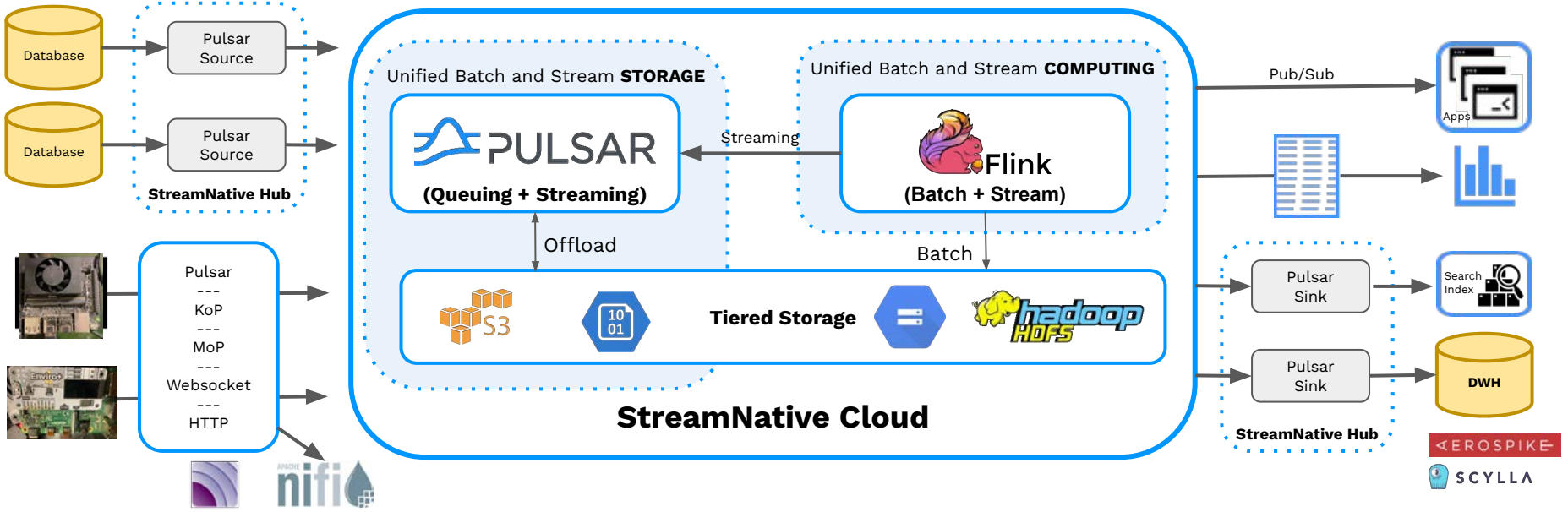
Demo Walkthrough




Esc
F1
F11
F2
F12
Backspace
Tab
Q
W
E
R
T
Y
U
I
O
P
Enter
Caps Lock
A
S
D
F
G
H
J
K
L
Shift
Z
X
C
V
B
N
M
Shift
Ctrl
Fn
Alt
Alt
Ctrl
PgUp
Home
PgDn

End-to-End Streaming FLiPN Edge AI Application

Apache Flink - Apache Pulsar - Apache NiFi <-> Devices - GPU/TPU - Python/Go/Java



All Data - Anytime - Anywhere - Multi-Cloud - Multi-Protocol