Hello Hydrate! From Stream to Clickhouse with Apache Pulsar and Friends

OSA CON 2021

Stream Native

Tim Spann

Developer Advocate

DZone Zone Leader and Big Data MVB Data DJay

- https://www.datainmotion.dev/
- https://github.com/tspannhw/SpeakerProfile
- https://dev.to/tspannhw
- https://sessionize.com/tspann/

StreamNative

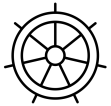
Stream Native

- Founded the original developers of Apache Pulsar.
- Passionate and dedicated team.
- StreamNative helps teams to capture, manage, and leverage data using Pulsar's unified messaging and streaming platform.

The Need For Real-Time Data

Hybrid and multi-cloud strategies with native geo-replication

Seamlessly build microservice architectures with support for streaming and messaging workloads



Built for Kubernetes CloudNative migrations with tools

360 degree customer data multi-tenancy, infinite retention, and extensive connector ecosystem

PULSAR

Apache  is an open source, cloud-native distributed messaging and streaming platform.

What are the Benefits of Pulsar?

Multi-Tenancy

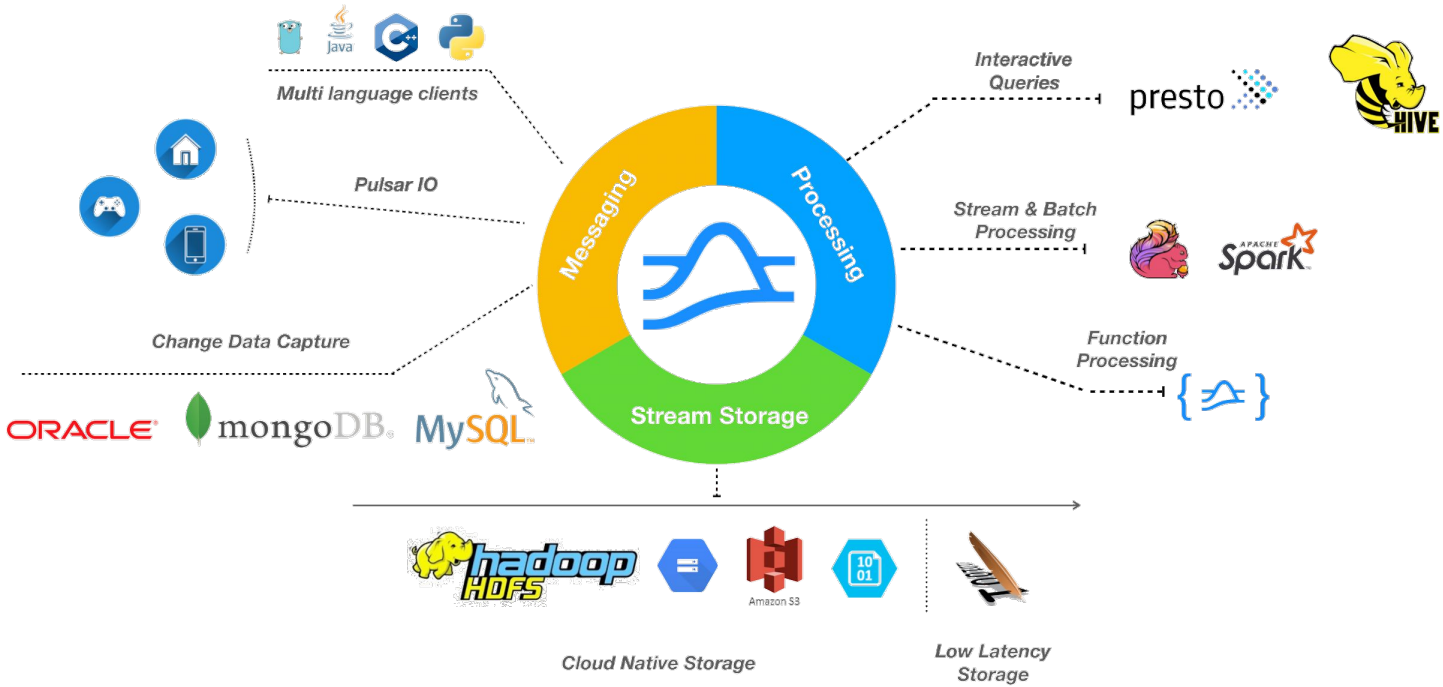
Scalability

Geo-Replication

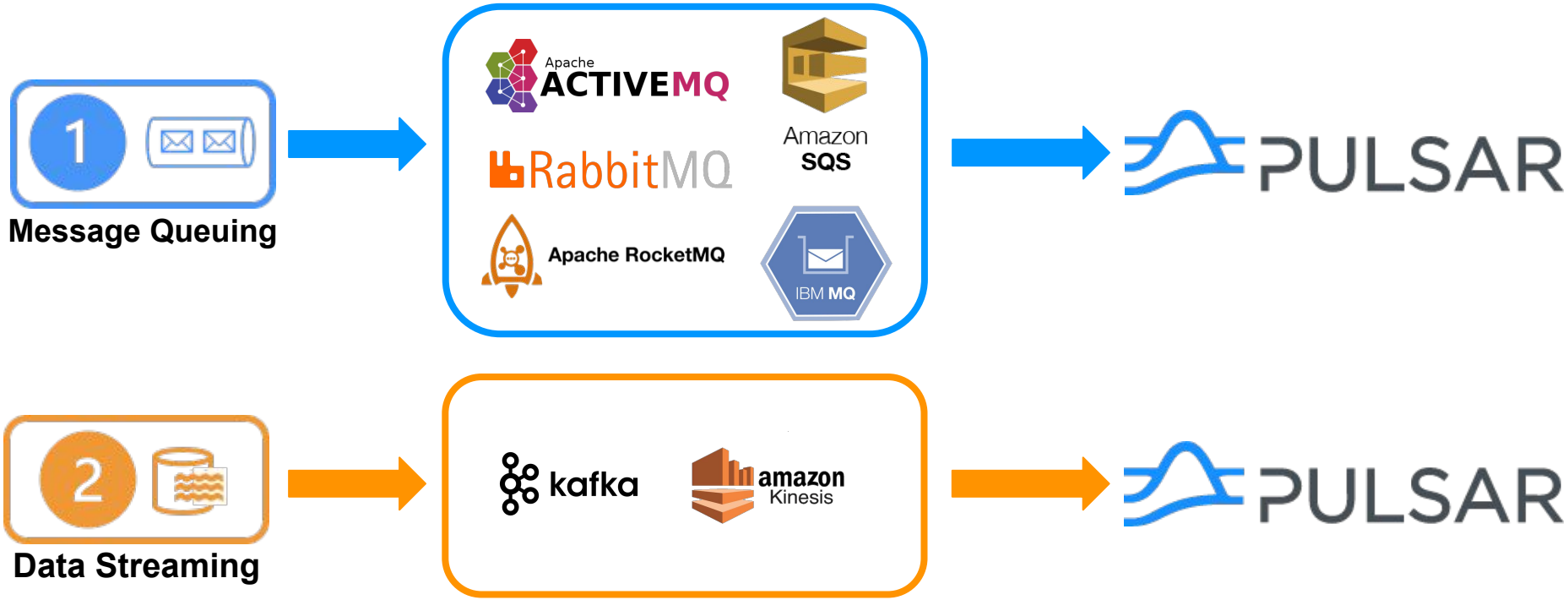
Unified Messaging Model

Data Durability

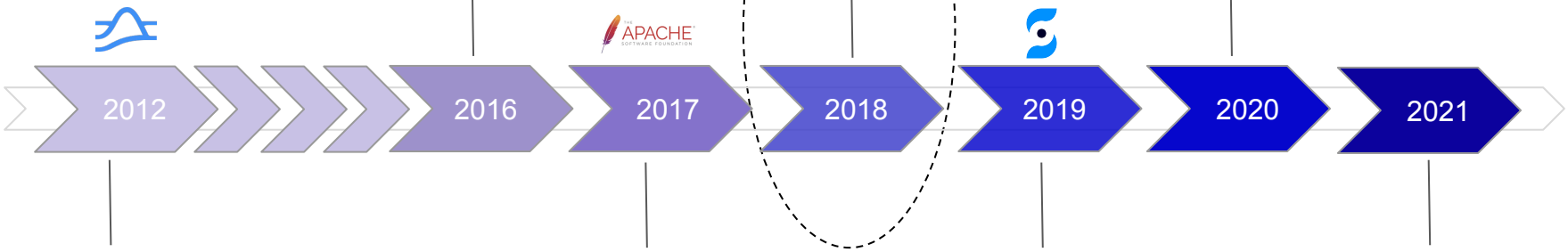
Apache Pulsar



A Unified Messaging Platform



Key Milestones



2012

Originally developed inside **Yahoo**! as “Cloud Messaging Service”

2016

Pulsar is committed to **Open Source**

2017

Pulsar is accepted into the **Apache Software Foundation**

2018

Major increase in adoption following TLP designation in 2018

Pulsar becomes a **Top-Level Project**

2019

- **StreamNative is founded** and seed round raised.
- Tencent adopts Pulsar for payment processing platform.
- BestPay adopts Pulsar for payment processing.
- Pulsar hits 200 contributors.

2020

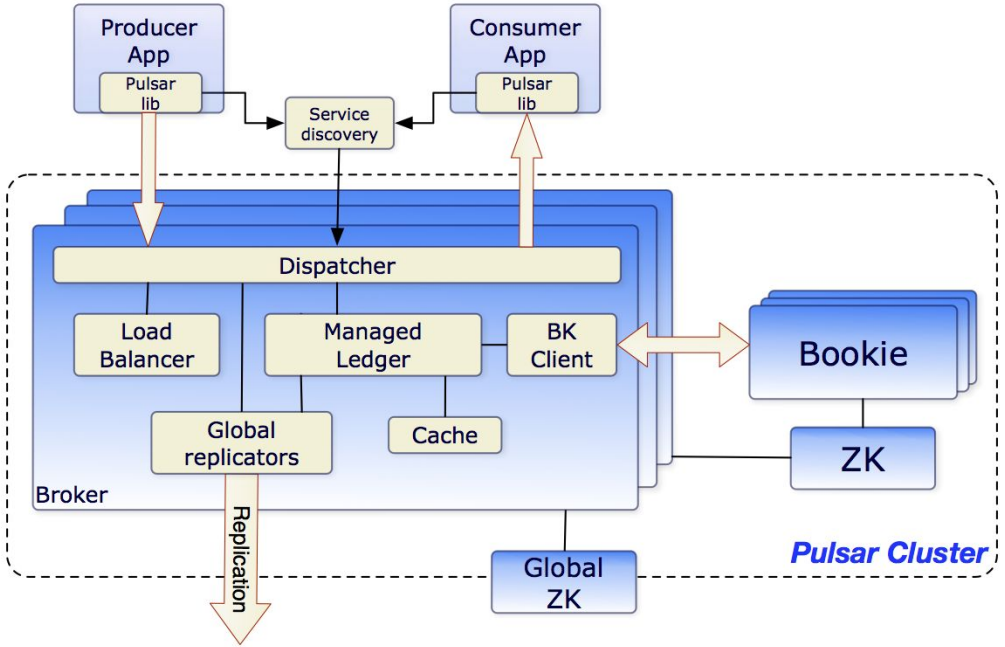
- 2 global Pulsar conferences, 80+ speakers, 1,500+ attendees
- Pulsar hits 340 contributors
- StreamNative and OVHCloud launch Kafka on Pulsar (**KoP**)
- StreamNative + China Mobile launch AMQP on Pulsar (**AoP**)
- Pulsar Ecosystem expands - StreamNative Hub launches
- **StreamNative Cloud** launches on GCP and Alibaba Cloud
- StreamNative customer adoption continues - new customers include Flipkart and Applied Materials
- Pulsar 2.7 + Transactions
- Pulsar Flink Connector 2.7

2021

- 3 global Pulsar conferences
- StreamNative hits 400 contributors (June).
- **Pulsar surpasses Kafka** in monthly active contributors.
- Pulsar 2.8 + Exactly-Once semantics
- **StreamNative Platform** launches

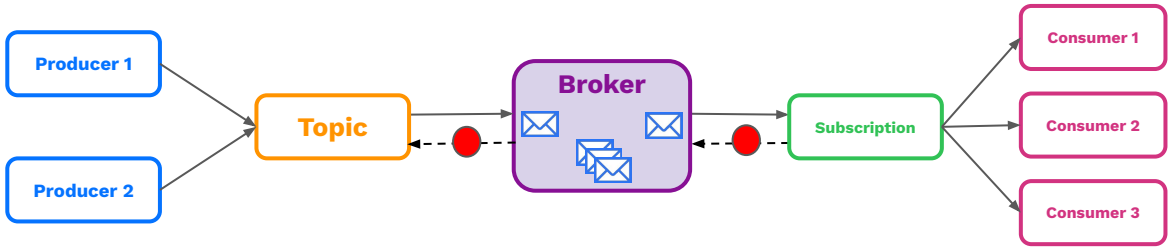
Apache Pulsar Overview

- Pub-Sub
- Geo-Replication
- Pulsar Functions
- Horizontal Scalability
- Multi-tenancy
- Tiered Persistent Storage
- Pulsar Connectors
- REST API
- CLI
- Many clients available
- Four Different Subscription Types
- Multi-Protocol Support
  - MQTT
  - AMQP
  - JMS
  - Kafka
  - ...



Pulsar's Publish-Subscribe model

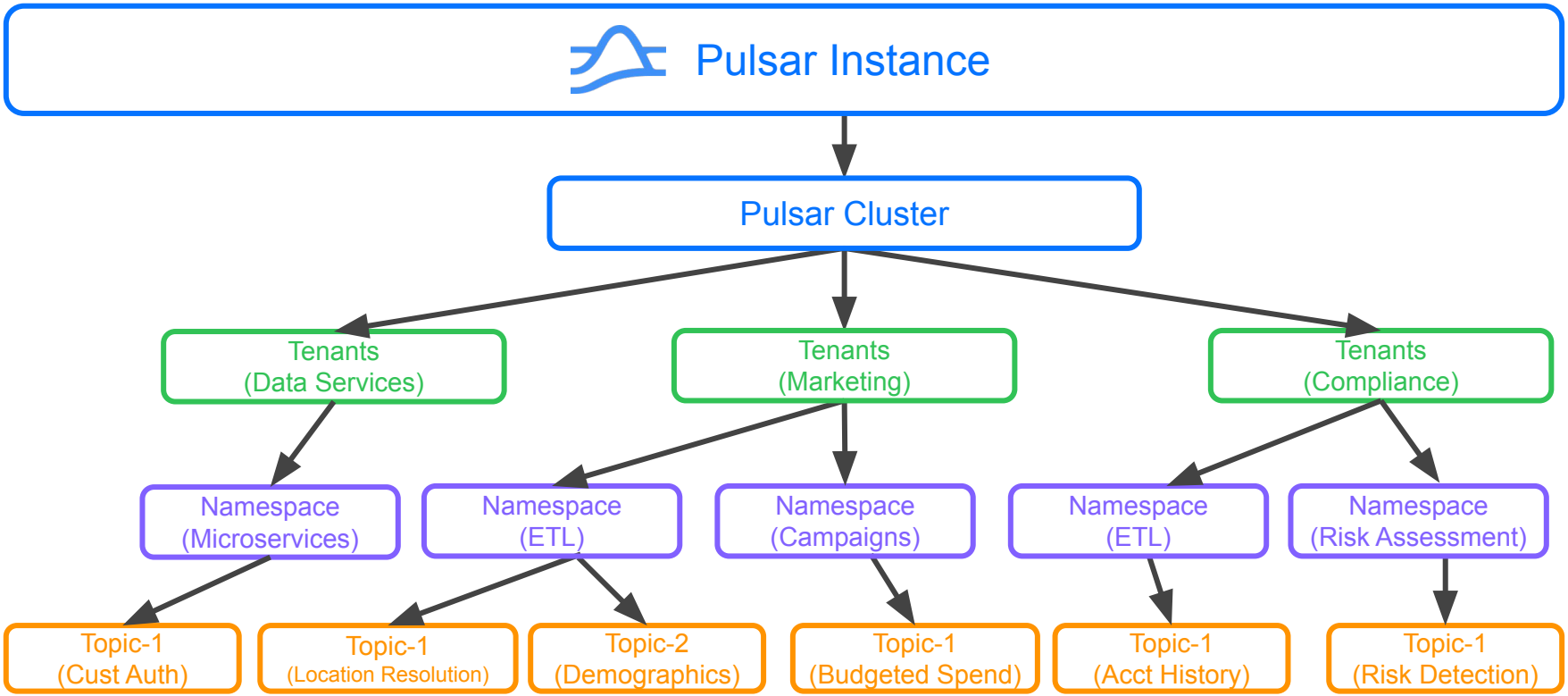
- Producers send messages.
- Topics are an ordered, named channel that producers use to transmit messages to subscribed consumers.
- Messages belong to a topic and contain an arbitrary payload.
- Brokers handle connections and routes messages between producers / consumers.
- Subscriptions are named configuration rules that determine how messages are delivered to consumers.
- Consumers receive messages.



What is the Pulsar Ecosystem?

- Functions and Connectors
  - Functions: Lightweight stream processing
  - Connectors: Part of “Pulsar IO”, includes “Source” and “Sink” APIs
    - Files, Databases, Data tools, Cloud Services, etc
- Protocol Handlers
  - Allows Pulsar to handle additional protocols by an extendable API running in the broker
    - AoP (AMQP), KoP (Kafka), MoP (MQTT)

Topics


Pulsar Instance
Pulsar Cluster
Tenants (Data Services)
Tenants (Marketing)
Tenants (Compliance)
Namespace (Microservices)
Namespace (ETL)
Namespace (Campaigns)
Namespace (ETL)
Namespace (Risk Assessment)
Topic-1 (Cust Auth)
Topic-1 (Location Resolution)
Topic-2 (Demographics)
Topic-1 (Budgeted Spend)
Topic-1 (Acct History)
Topic-1 (Risk Detection)

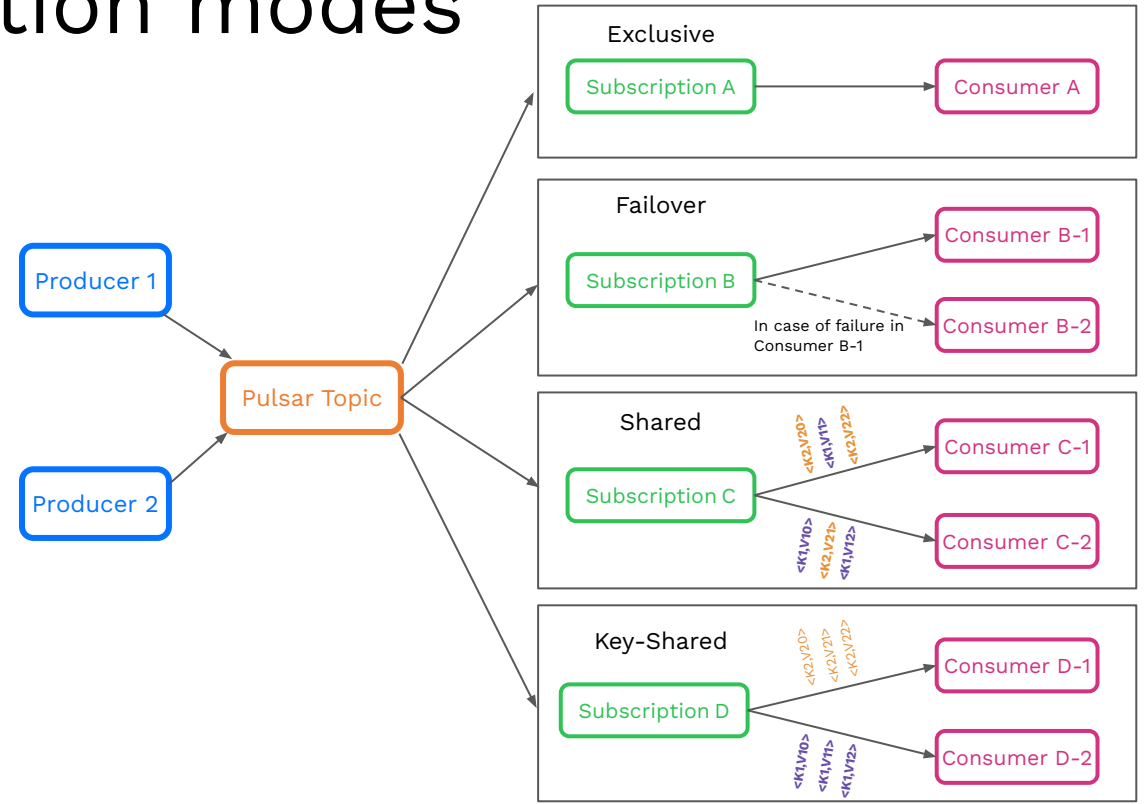
Pulsar subscription modes

Different subscription modes have different semantics:

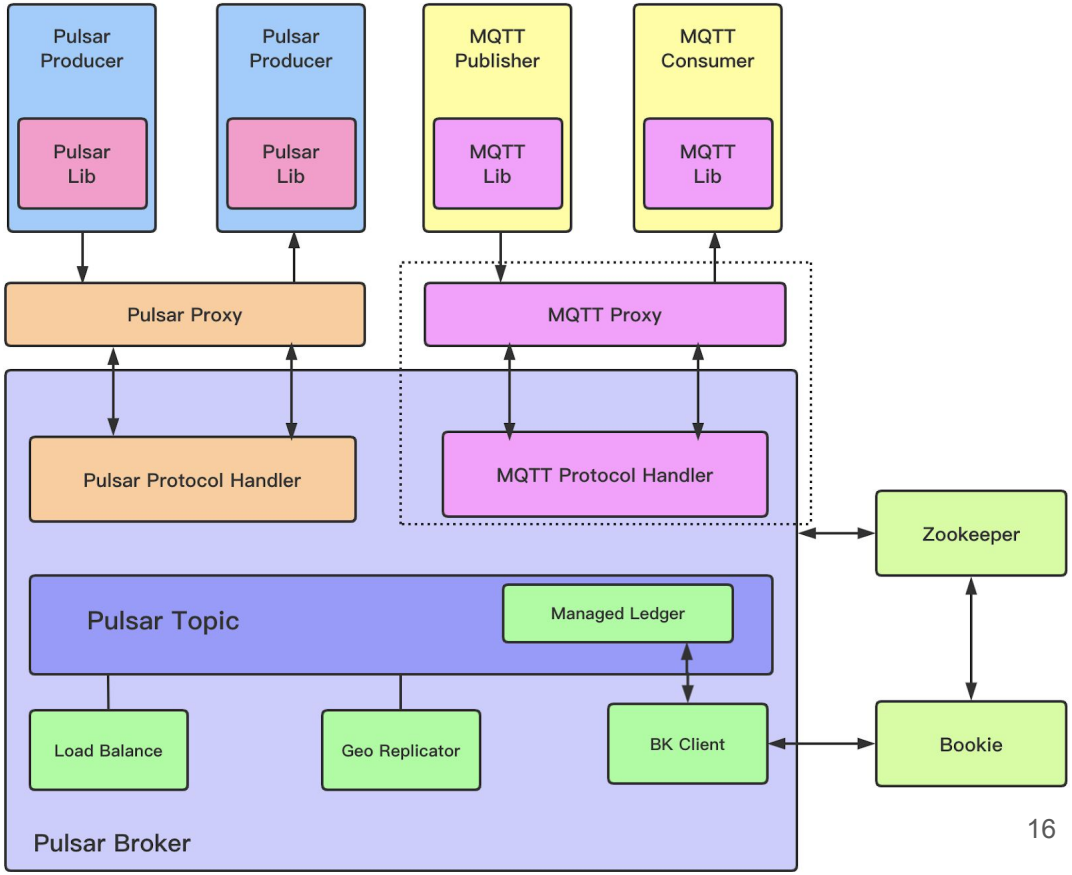
Exclusive/Failover - guaranteed order, single active consumer

Shared - multiple active consumers, no order

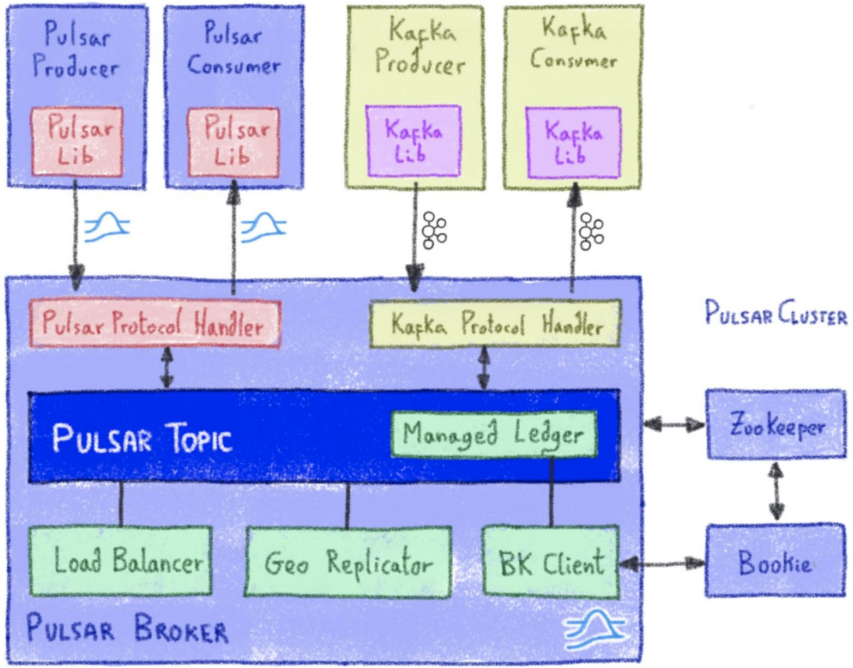
Key_Shared - multiple active consumers, order for given key



MQTT on Pulsar (MoP)



Kafka-on-Pulsar (Kop)



Moving Data In and Out of Pulsar

IO/Connectors are a simple way to integrate with external systems and move data in and out of Pulsar.

- Built on top of Pulsar Functions
- Built-in connectors - hub.streamnative.io

Moving Data Out of Pulsar to Clickhouse

IO/Connectors are a simple way to integrate with external systems and move data in and out of Pulsar. https://pulsar.apache.org/docs/en/io-jdbc-sink/

- Built on top of Pulsar Functions
- Built-in connectors - hub.streamnative.io

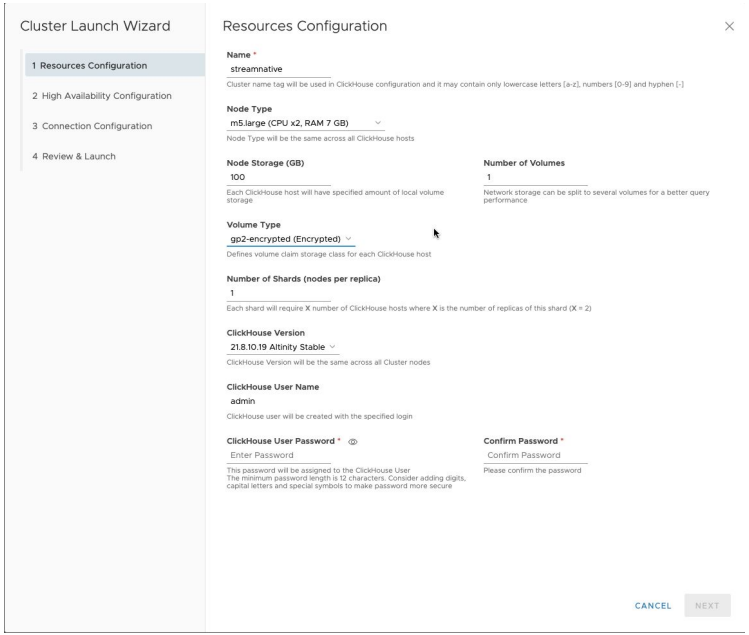
Streaming Events into Altinity Cloud

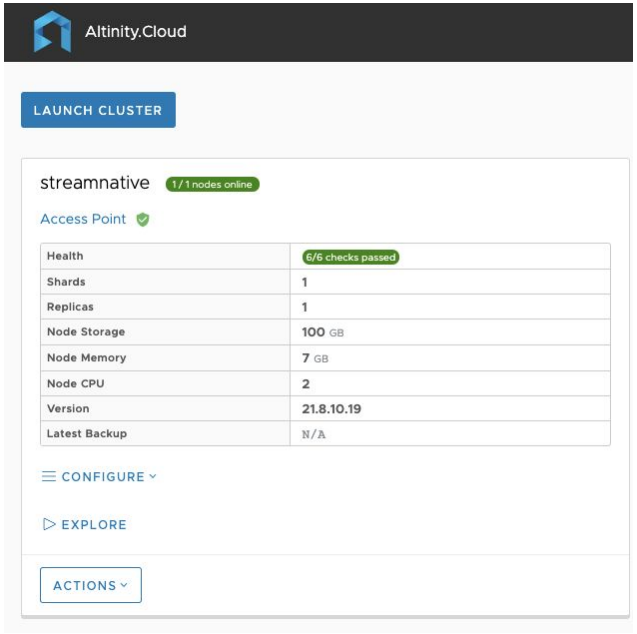
```
CREATE TABLE iotjetsonjson_local
(
       uuid String,  camera String,      ipaddress String,    networktime String,          top1pct String,
       top1 String,  cputemp String,     gputemp String,         gputempf String,
       cputempf String,     runtime String,
       host String,  filename String,    host_name String,          macaddress String,
       te String,    systemtime String,  cpu String,         diskusage String,
       memory String,       imageinput String
)
ENGINE = MergeTree()
  PARTITION BY uuid
  ORDER BY (uuid);

CREATE TABLE iotjetsonjson ON CLUSTER '{cluster}' AS iotjetsonjson_local
ENGINE = Distributed('{cluster}', default, iotjetsonjson_local, rand());
```

https://docs.altinity.com/altinitycloud/quickstartguide/connectclient/

pulsar-io-jdbc-clickhouse-2.8.0.nar

Build Cluster in Altinity Cloud





https://docs.altinity.com

Run A Cluster in Altinity Cloud

osacon2021 overview

Type	**Kubernetes**
ClickHouse Clusters	1
ClickHouse Nodes	1
Zookeeper Clusters	1
Zookeeper Nodes	1
Node Types	6

Kubernetes Namespace	**osacon2021**
Kubernetes Nodes	**3**
Availability Zones	**us-east-1a, us-east-1b, us-east-1c**
CH Operator	Version 0.15.0
Monitoring	View in Grafana

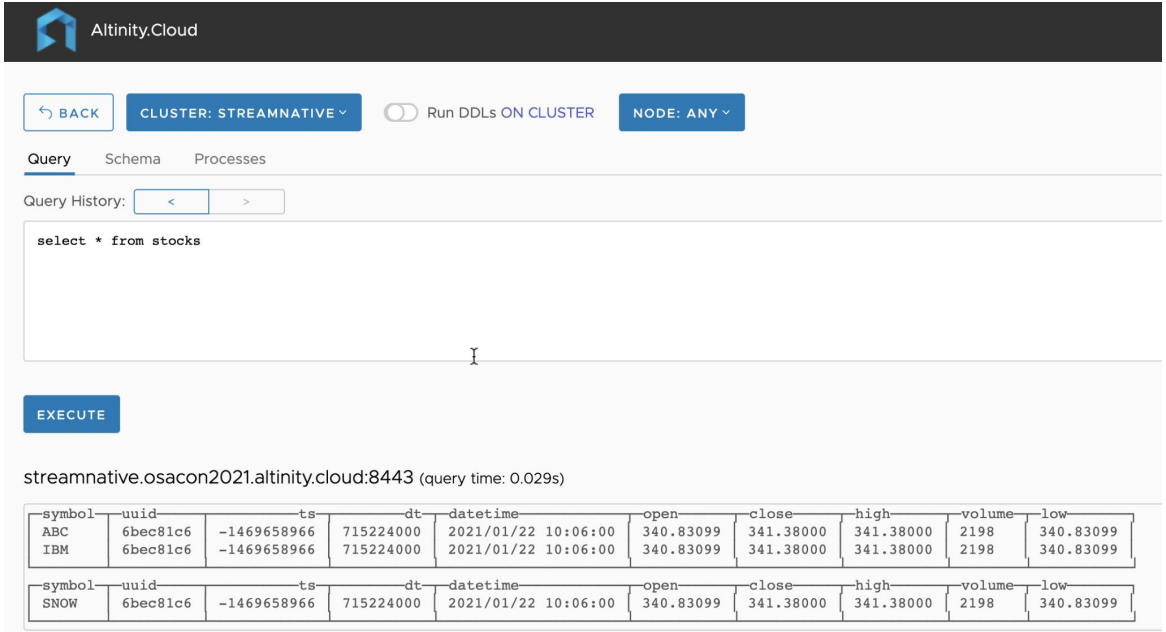
memory

cpu

https://docs.altinity.com

OSACON 2021

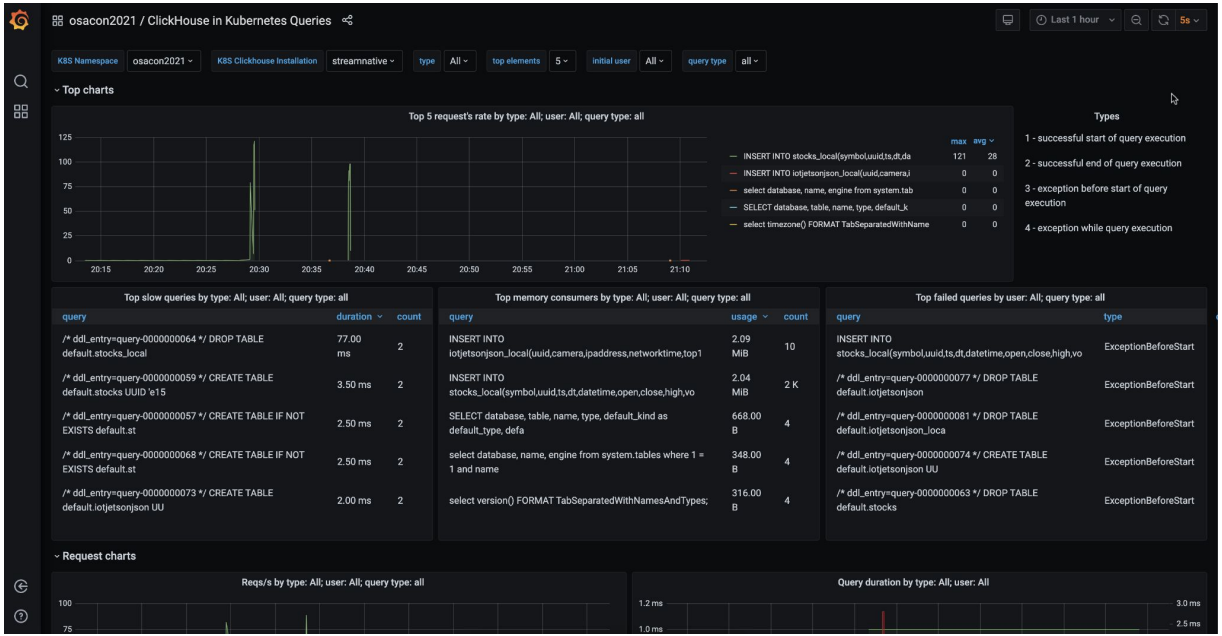
Run A Cluster in Altinity Cloud



https://docs.altinity.com

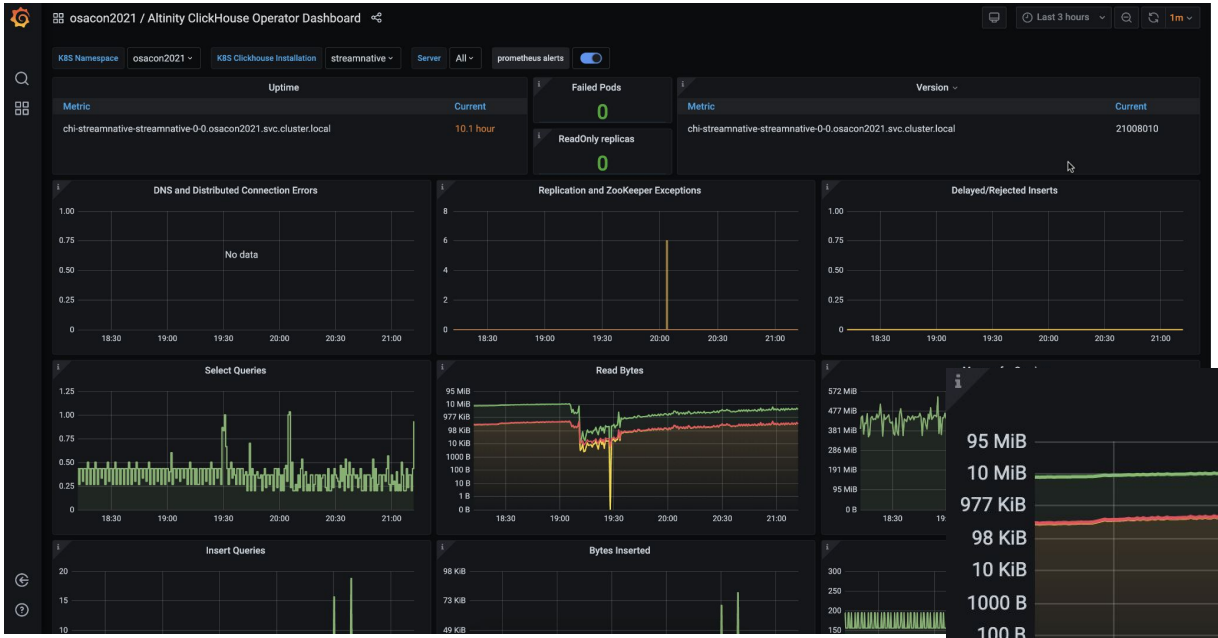
OSA CON 2021

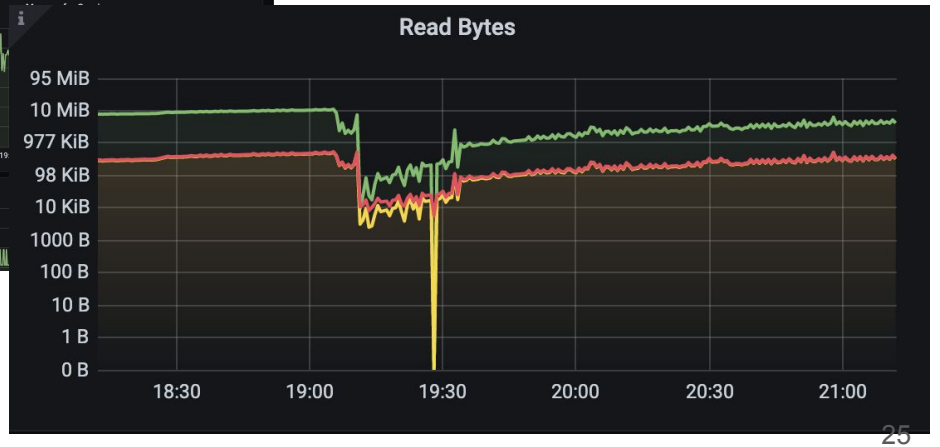
Monitor The Cluster in Altinity Cloud



https://docs.altinity.com

Monitor The Cluster in Altinity Cloud



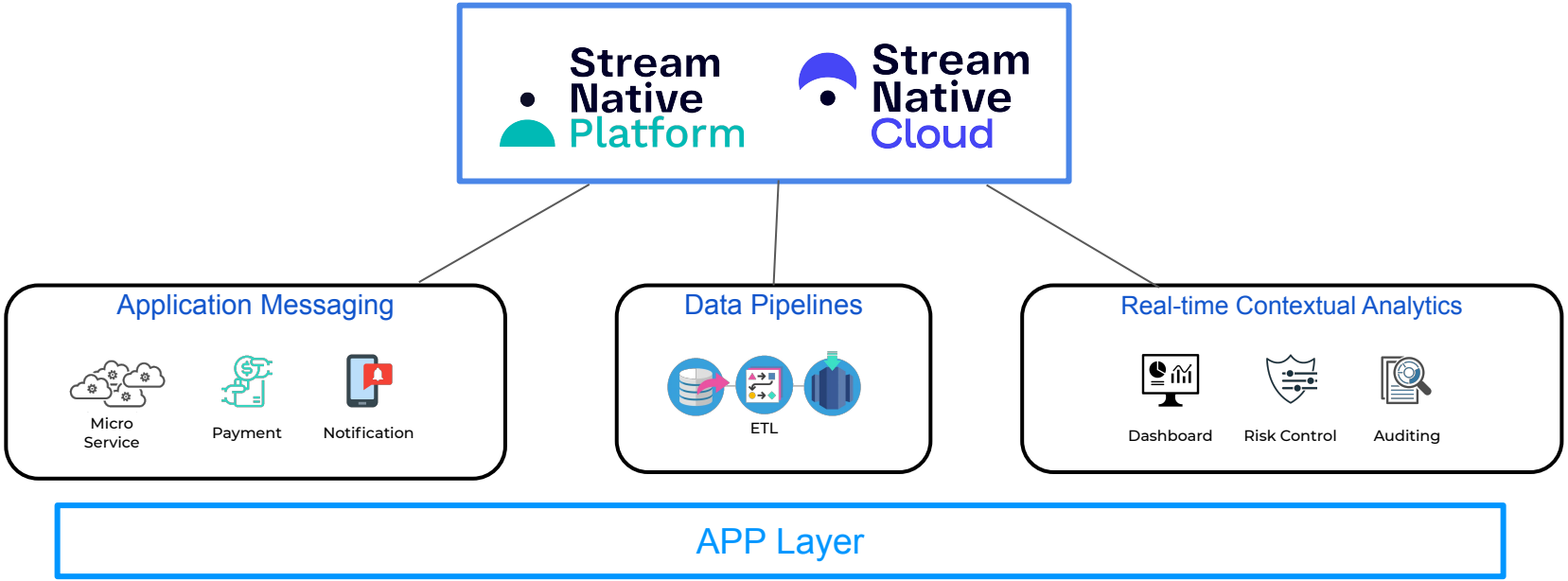


https://docs.altinity.com

StreamNative
Cloud

Stream
Native
Cloud

How Companies Use StreamNative today



StreamNative Cloud

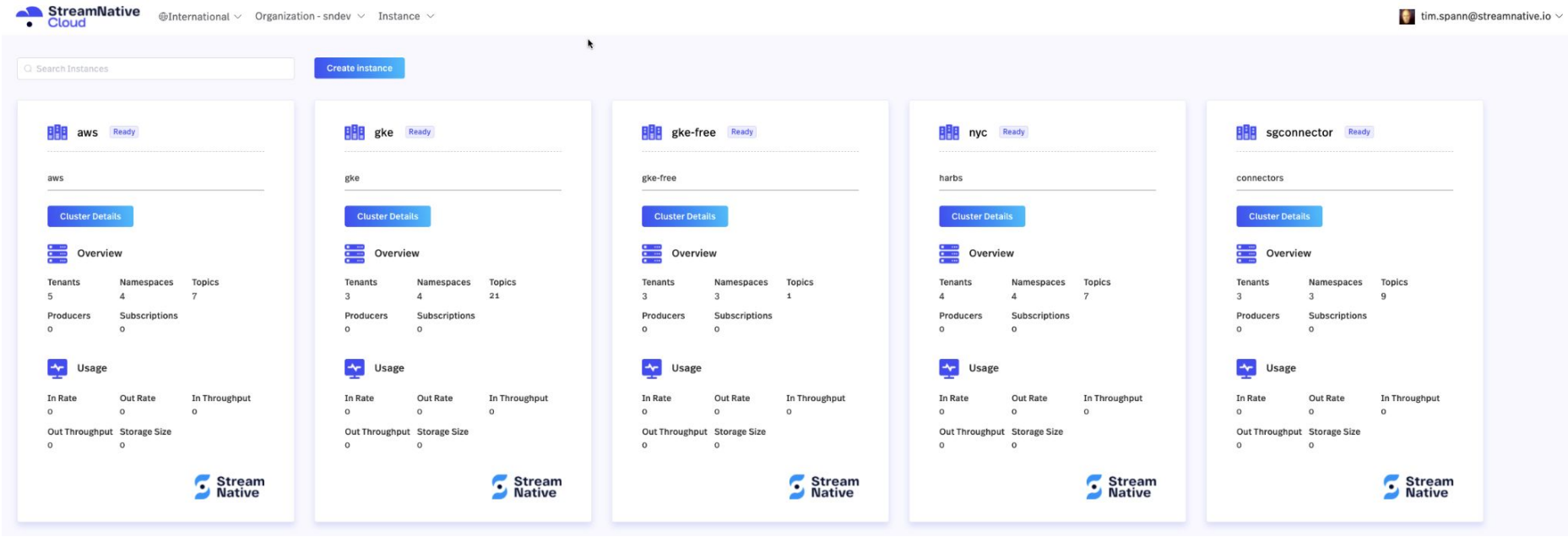
Powered by Apache Pulsar, StreamNative provides a cloud-native, real-time messaging and streaming platform to support multi-cloud and hybrid cloud strategies.

Cloud Native

Built for Containers

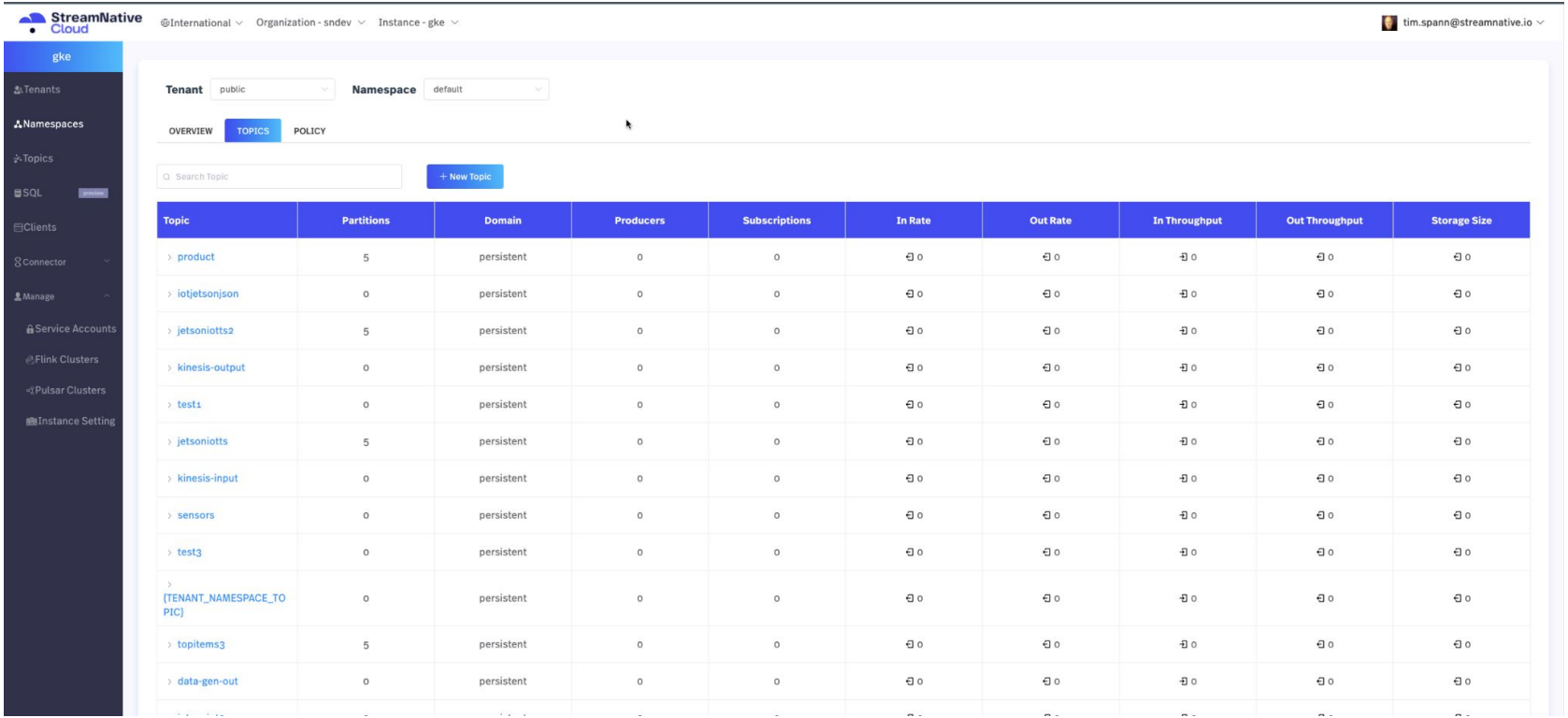
Flink SQL

Stream Native Cloud

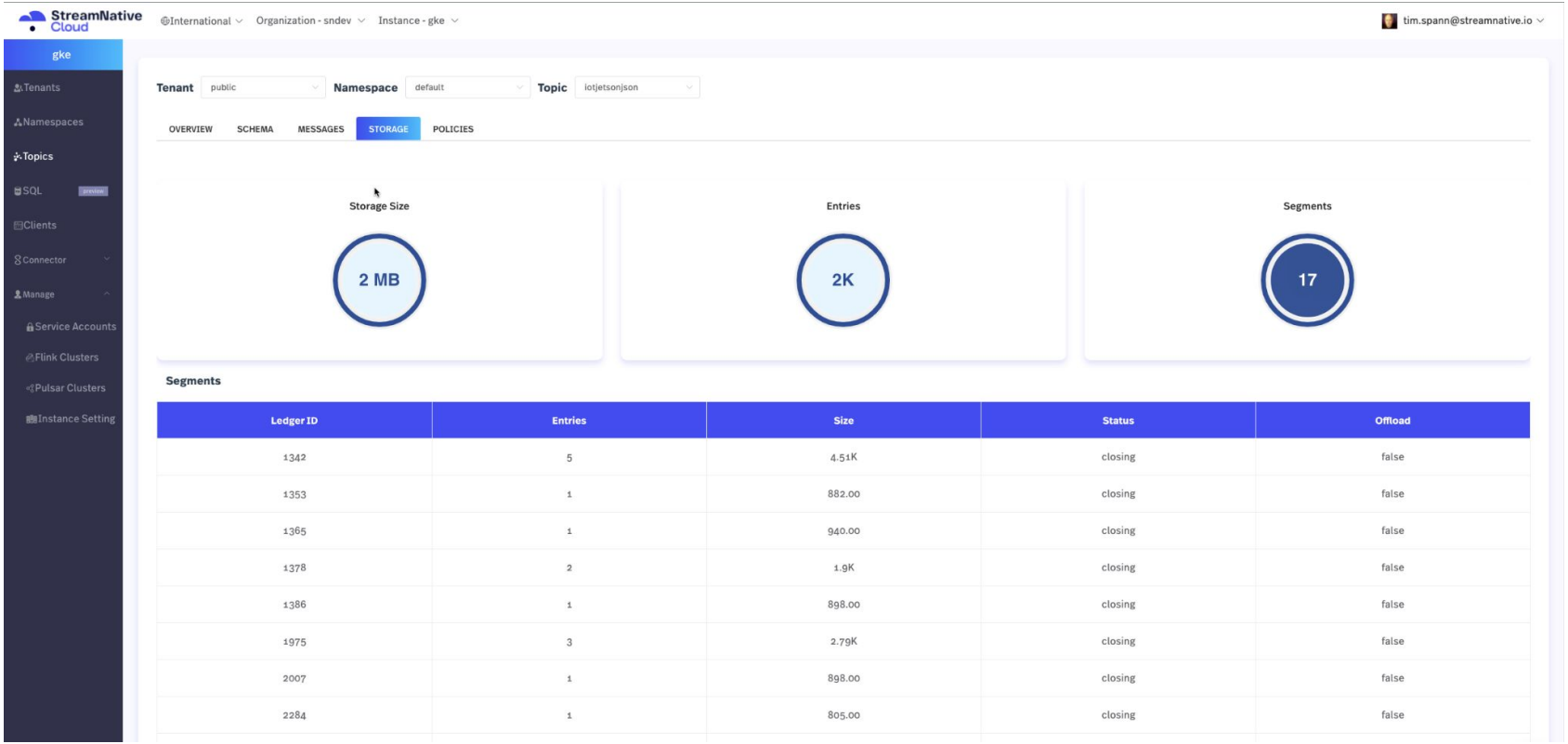


OSA CON 2021

StreamNative Cloud



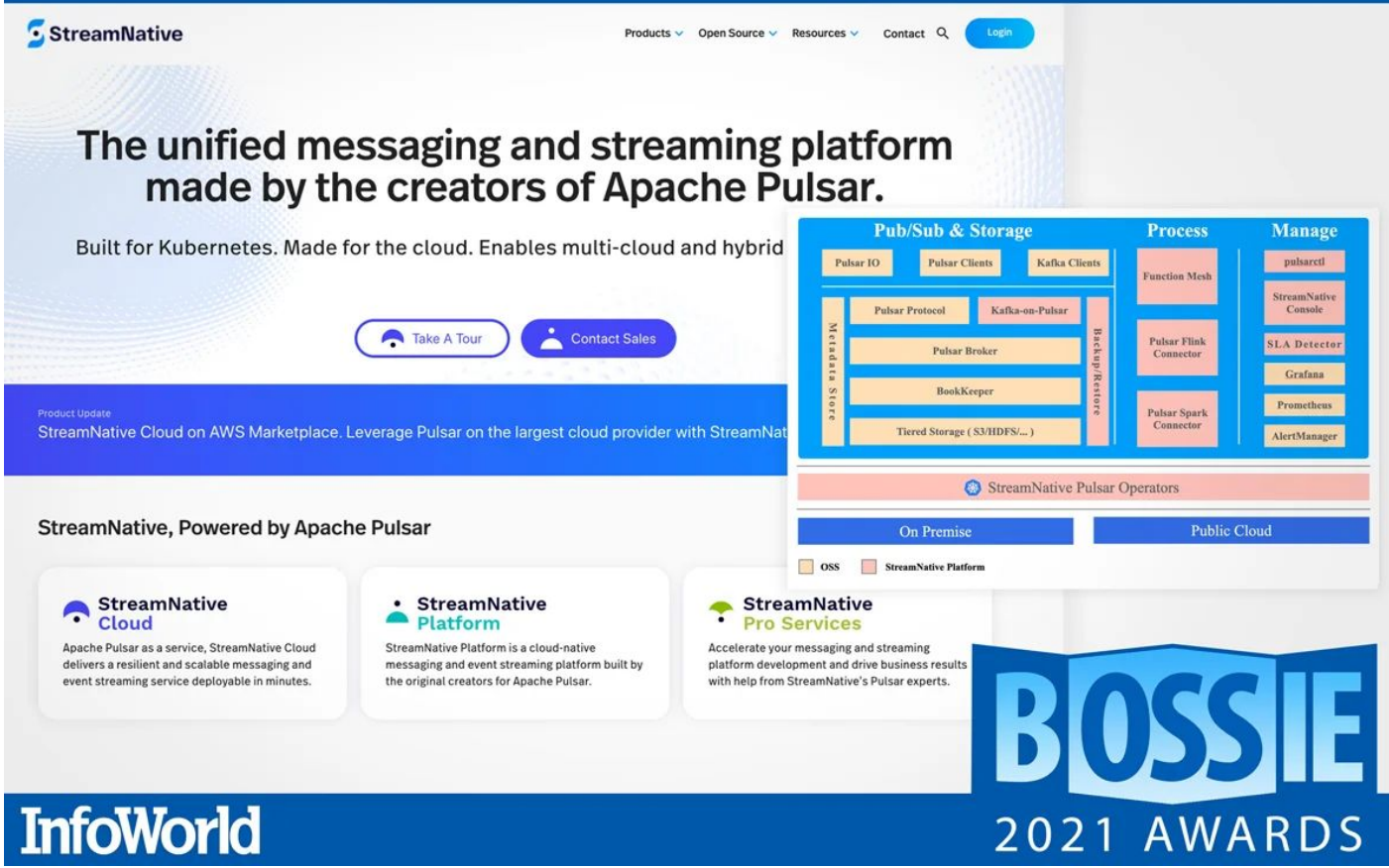
OSA CON 2021



Stream Native Cloud

OSA CON 2021

StreamNative



InfoWorld

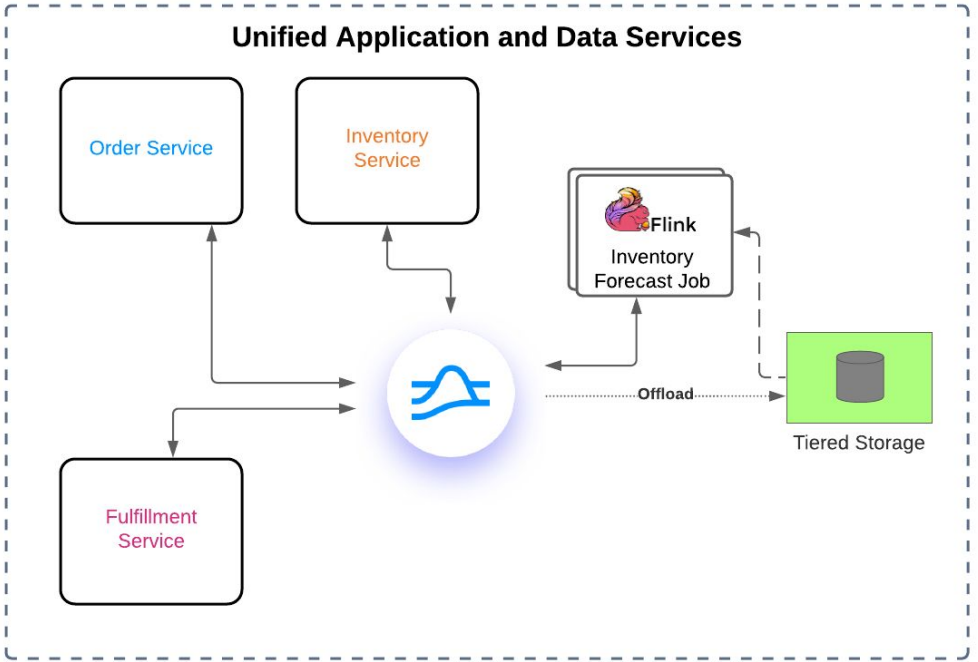
BOSSIE 2021 AWARDS

OSACON 2021

Best Practice Architectures

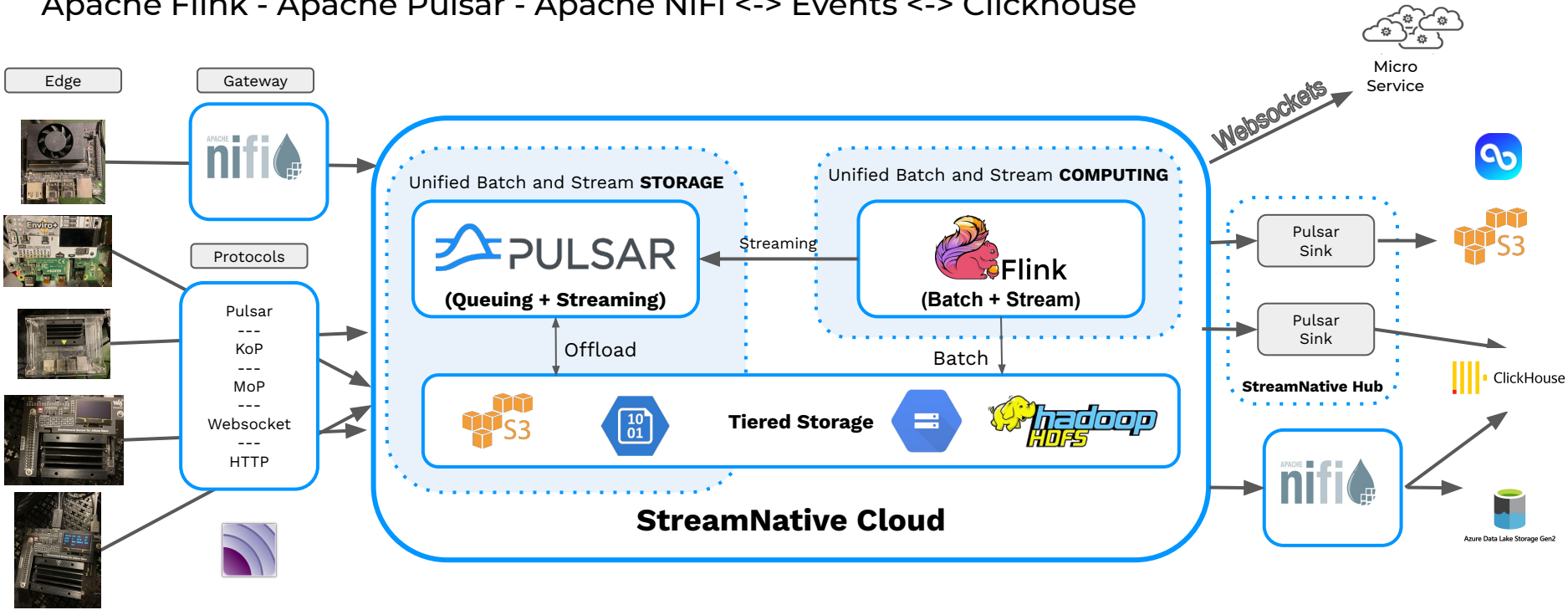
Example: E-Commerce with Pulsar

- Unified storage with access to underlying data
- Native tiered storage
- Single system to exchange data
- Teams share toolset



End-to-End Streaming FLiP(N) Apps

Apache Flink - Apache Pulsar - Apache NiFi <-> Events <-> Clickhouse



OSACON 2021

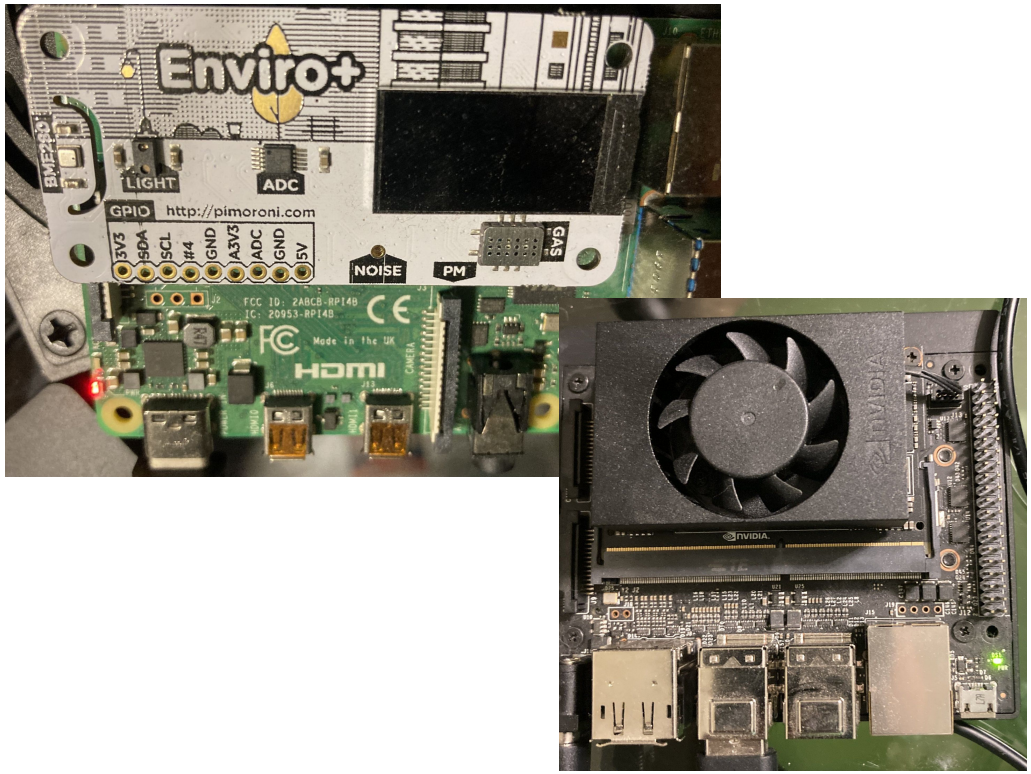
Demo

https://github.com/tspannhw/FLiP-Stream2Clickhouse/

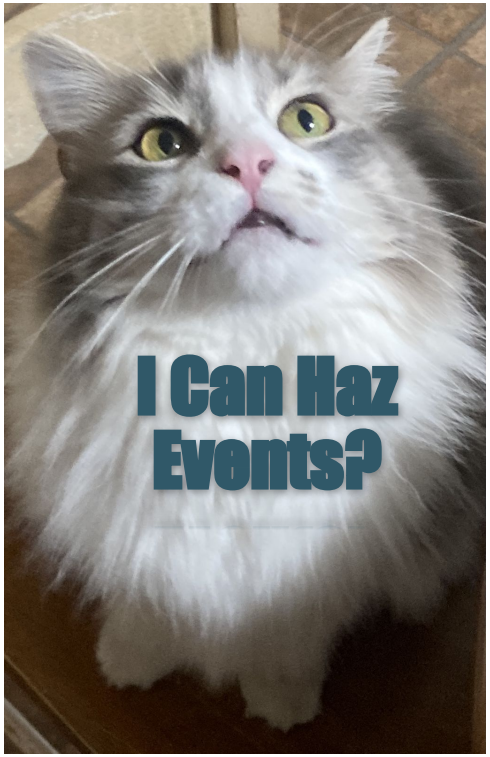
IoT Data

IoT Ingestion: High-volume streaming sources, sensors, multiple message formats, diverse protocols and multi-vendor devices creates data ingestion challenges.

Other Sources: Transit data, news, twitter, status feeds, REST data, stock data and more.



REST JSON “stonks” Events



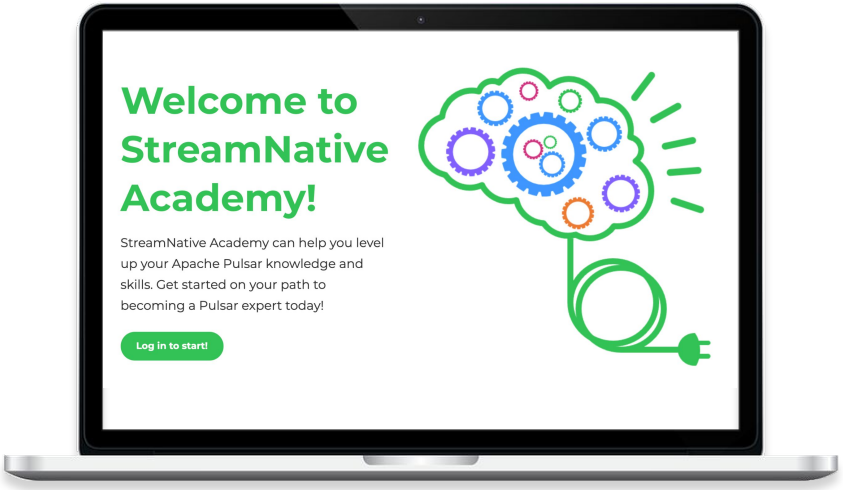
{"symbol":"STREAM",
"uuid":"10640832-f139-4b82-8780-e3ad37b3d0ce",
"ts":1618529574078,
"dt":1612098900000,
"datetime":"2021/01/31 08:15:00",
"open":"12.24500",
"close":"12.25500",
"high":"12.25500",
"volume":"12353",
"low":"12.24500"}

Wrap-Up

Now Available

On-Demand Pulsar Training

Academy.StreamNative.io



We're Hiring

streamnative.io/careers/

Connect with the Community & Stay Up-To-Date

- Join the Pulsar Slack channel - *Apache-Pulsar.slack.com*
- Follow @streamnativeio and @apache_pulsar on Twitter
- Subscribe to Monthly Pulsar Newsletter for major news, events, project updates, and resources in the Pulsar community

OSA CON 2021

Interested In Learning More?

Resources

[Flink SQL Cookbook]

[The Github Source for Flink SQL Demo]

[The GitHub Source for Demo]

Free eBooks

[Manning's Apache Pulsar in Action]

[O'Reilly Book]

Upcoming Events

[11/8] [PASS Data Community]

[11/18] [Developer Week Austin]

[11/19] [Porto Tech Hub Con]

[12/3] [Data Science Camp]

OSA CON 2021

Deeper Content

- https://www.datainmotion.dev/2020/04/building-search-indexes-with-apache.html
- https://github.com/tspannhw/nifi-solr-example
- https://github.com/streamnative/pulsar-flink
- https://www.linkedin.com/pulse/2021-schedule-tim-spann/
- https://github.com/tspannhw/SpeakerProfile/blob/main/2021/talks/20210729_HailHydrate!FromStreamtoLake_TimSpann.pdf
- https://streamnative.io/en/blog/release/2021-04-20-flink-sql-on-streamnative-cloud
- https://docs.streamnative.io/cloud/stable/compute/flink-sql

 **@PaasDev**  **timothyspann**

https://www.pulsardeveloper.com/

OSA CON 2021

Let’s Keep in Touch!

Tim Spann

Developer Advocate

@PassDev

https://www.linkedin.com/in/timothyspann

https://github.com/tspannhw

Questions

Thank You!