

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



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://github.com/tspannhw/FLIP-Stream2Clickhouse>



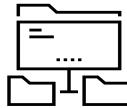


- Founded the original developers of Apache Pulsar.
- Passionate and dedicated team.
- StreamNative helps teams to capture, manage, and leverage data using Pulsar's unified messaging and streaming platform.

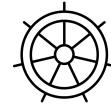
The Need For Real-Time Data



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

APACHE

PULSTAR



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

What are the Benefits of Pulsar?



Multi-Tenancy

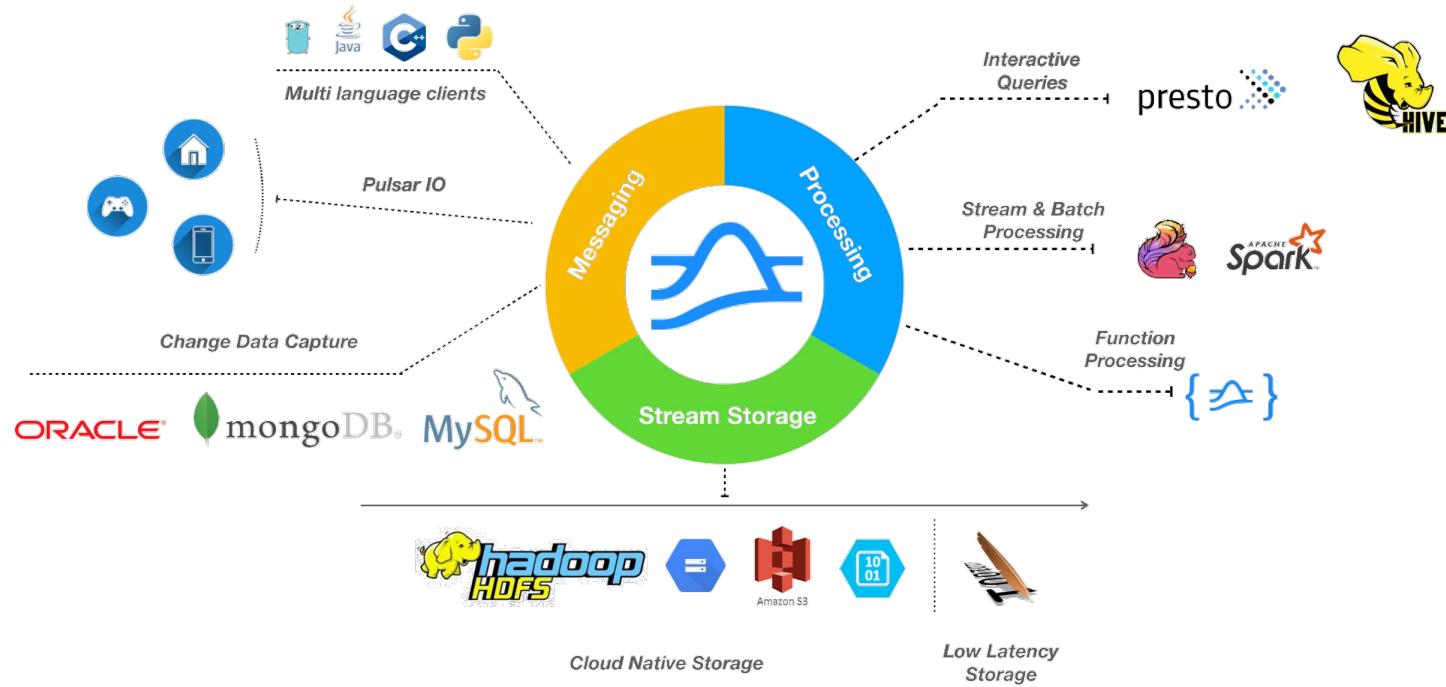
Scalability

Geo-Replication

Unified Messaging
Model

Data Durability

Apache Pulsar

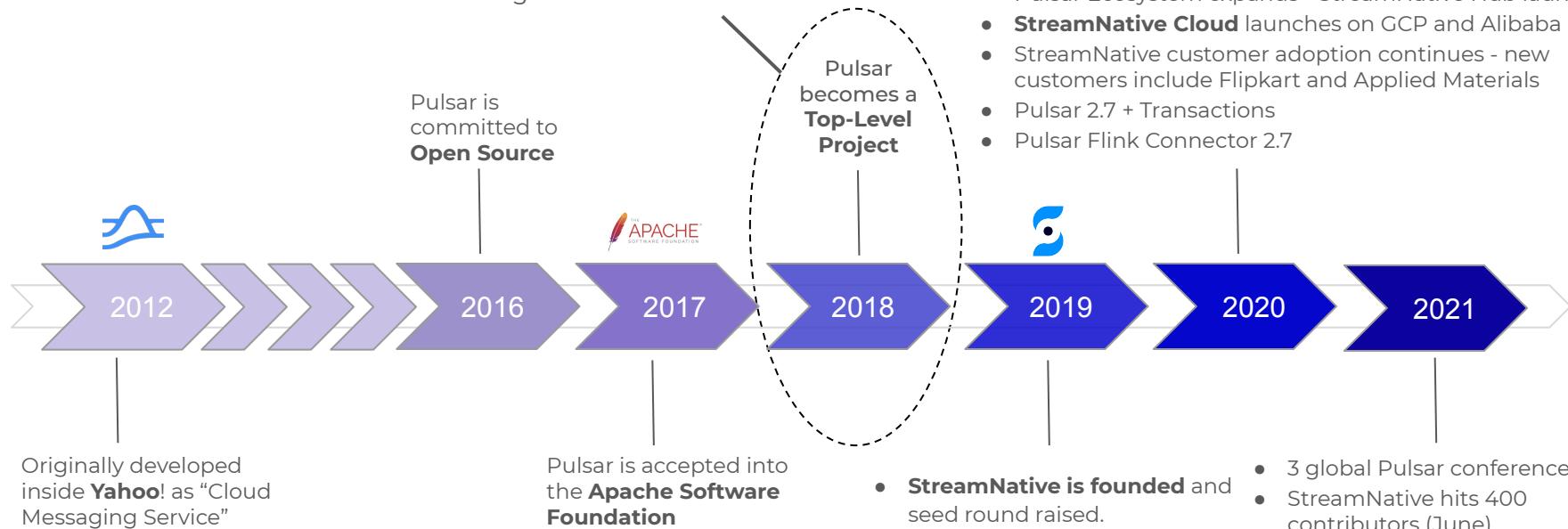


A Unified Messaging Platform



Key Milestones

Major increase in adoption following TLP designation in 2018

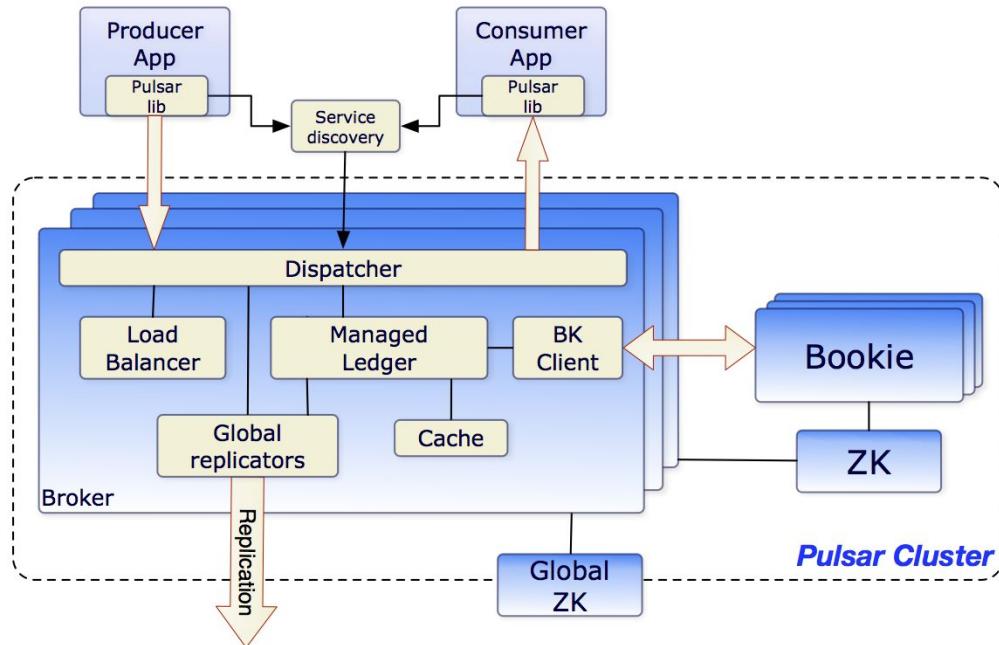


- 2 global Pulsar conferences, 80+ speakers, 1,500+ attendees
- Pulsar hits 340 contributors
- StreamNative and OVHCloud launch Kafka on Pulsar (**KoP**)
- StreamNative + China Mobile launch AMQP on Pulsar (**AoP**)
- Pulsar Ecosystem expands - StreamNative Hub launches
- **StreamNative Cloud** launches 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

- 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

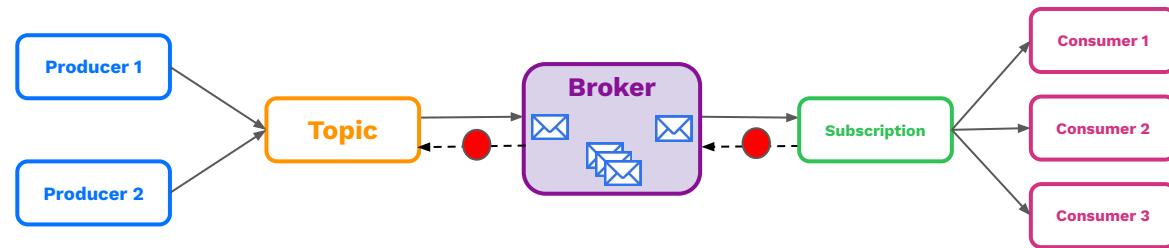
Apache Pulsar Overview

- 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
 - ...

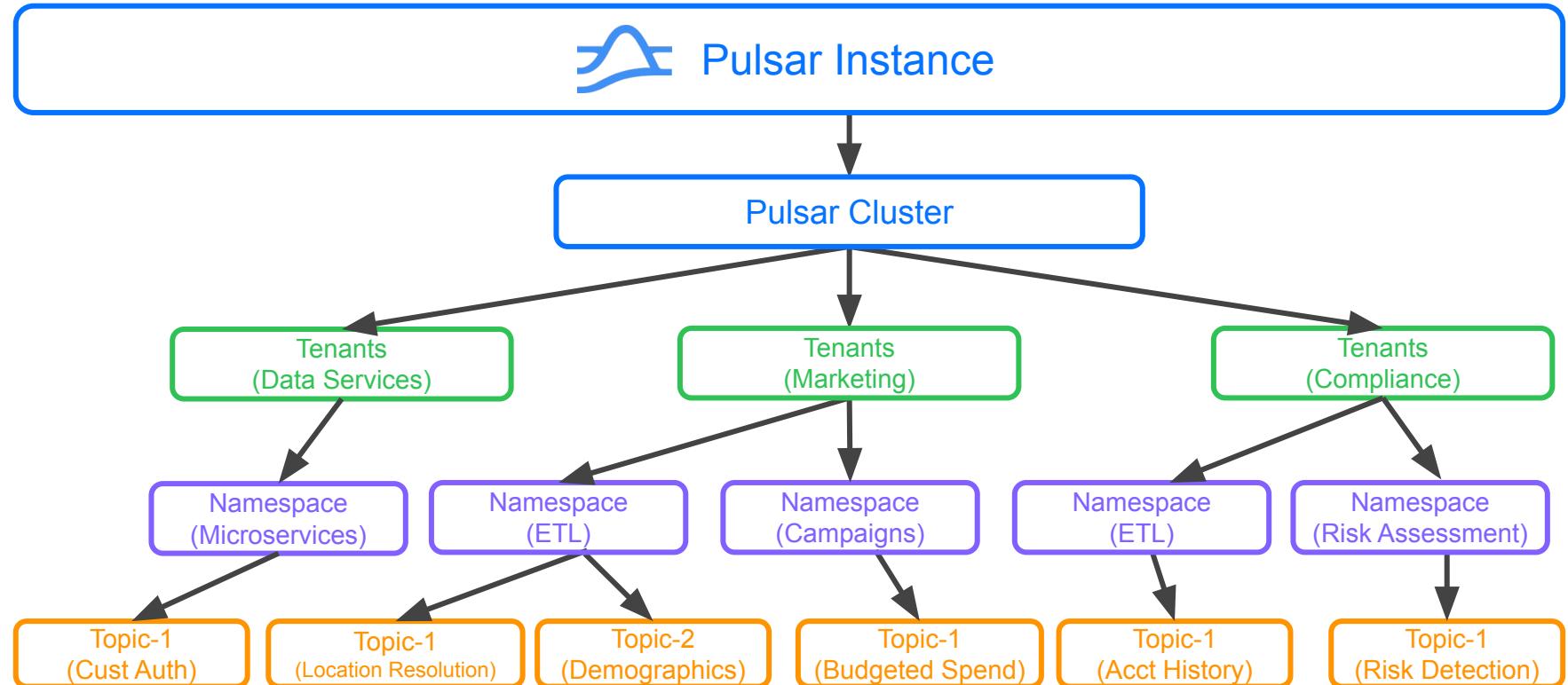


Pulsar's Publish-Subscribe model

- 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.



Topics



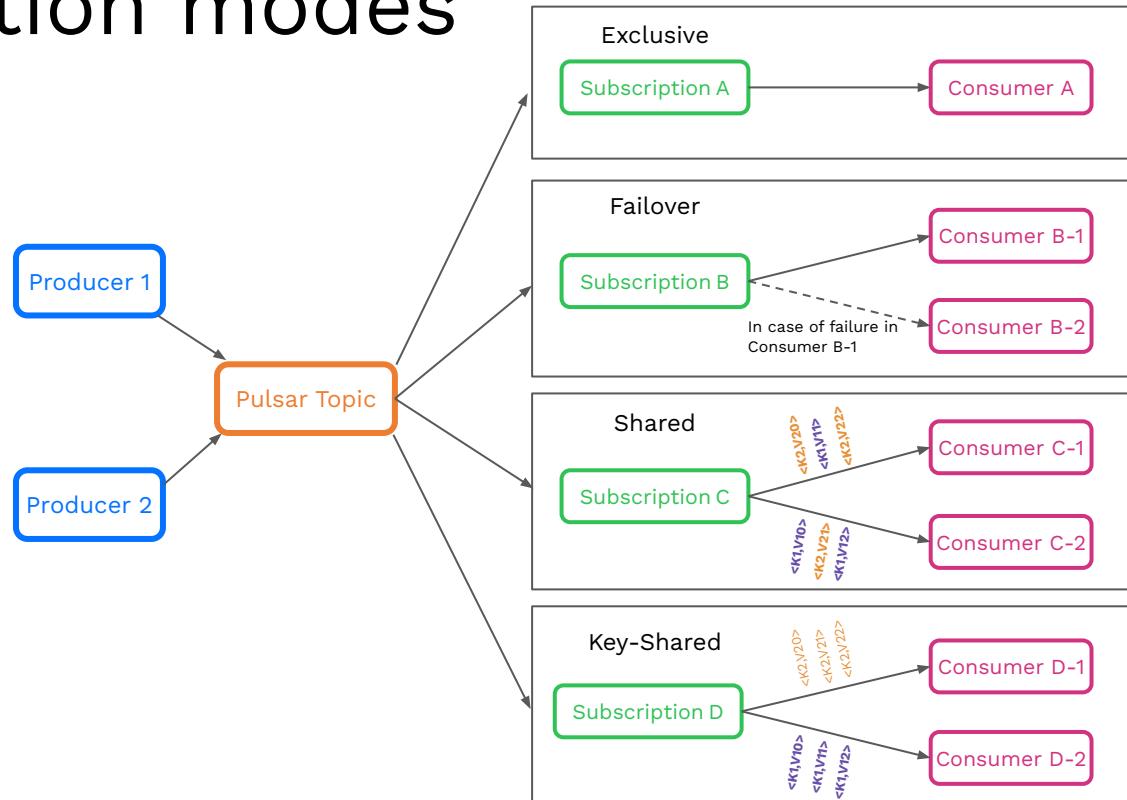
Pulsar subscription modes

Different subscription modes have different semantics:

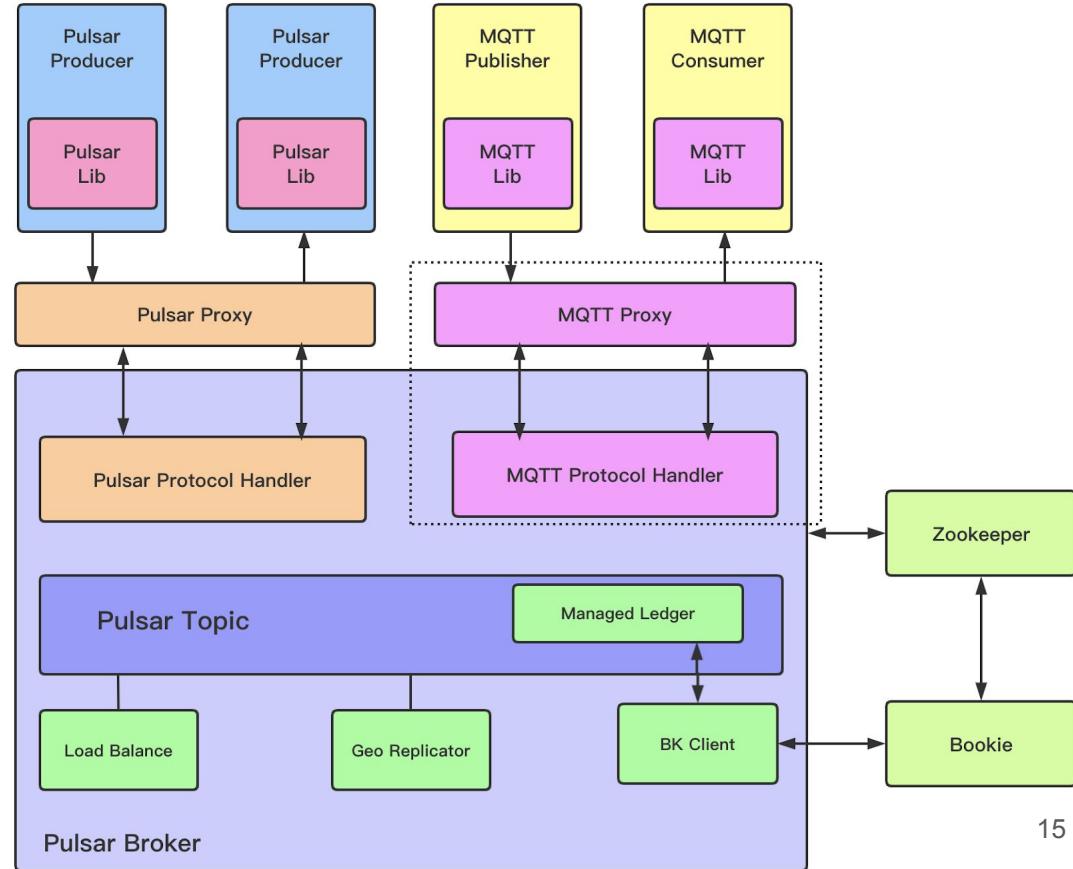
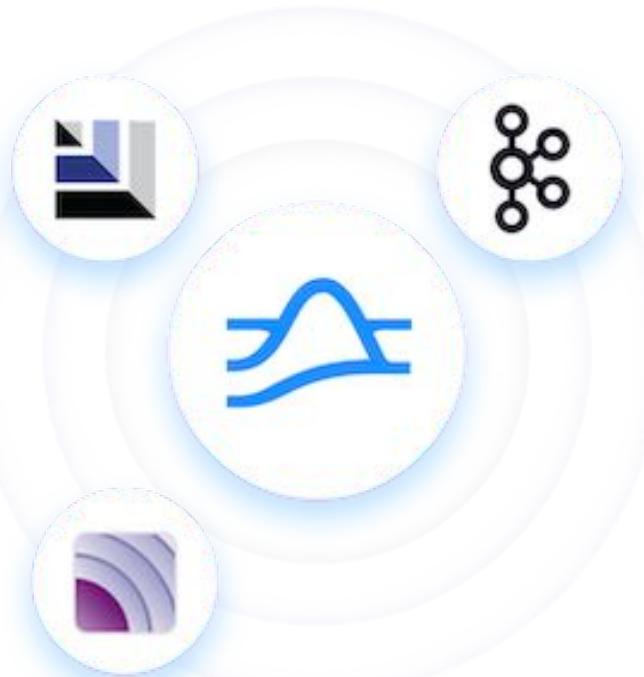
Exclusive/Failover - guaranteed order, single active consumer

Shared - multiple active consumers, no order

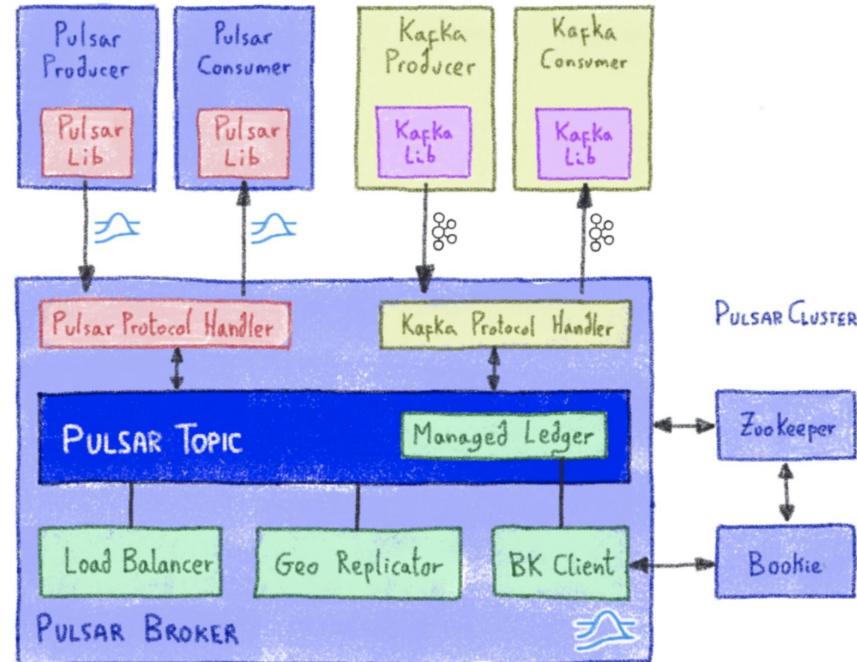
Key_Shared - multiple active consumers, order for given key

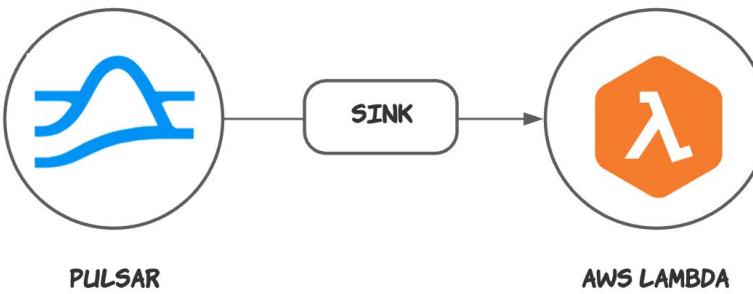
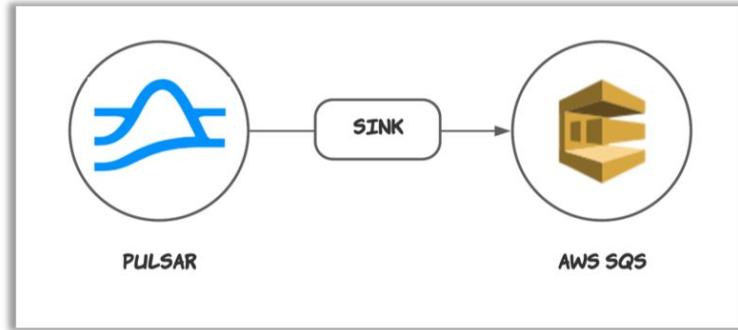
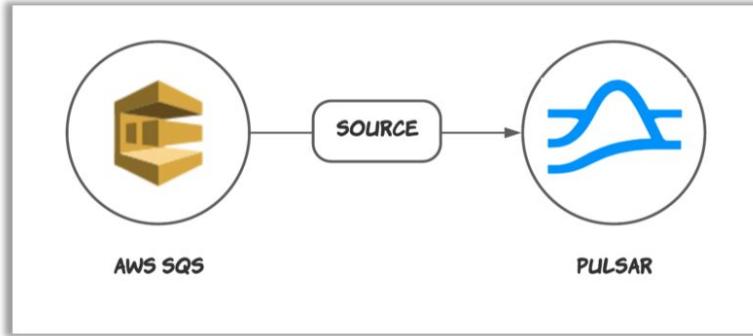


MQTT on Pulsar (MoP)



Kafka-on-Pulsar (Kop)





Moving Data 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

Cluster Launch Wizard

1 Resources Configuration

2 High Availability Configuration

3 Connection Configuration

4 Review & Launch

Resources Configuration

Name *
streamnative
Cluster name tag will be used in ClickHouse configuration and it may contain only lowercase letters [a-z], numbers [0-9] and hyphen [-]

Node Type
m5.large (CPU x2, RAM 7 GB)
Node Type will be the same across all ClickHouse hosts

Node Storage (GB)
100
Each ClickHouse host will have specified amount of local volume storage

Number of Volumes
1
Network storage can be split to several volumes for a better query performance

Volume Type
gp2-encrypted (Encrypted) 

Defines volume claim storage class for each ClickHouse host

Number of Shards (nodes per replica)
1
Each shard will require X number of ClickHouse hosts where X is the number of replicas of this shard (X = 2)

ClickHouse Version
21.8.10.19 Altinity Stable 

ClickHouse Version will be the same across all Cluster nodes

ClickHouse User Name
admin
ClickHouse user will be created with the specified login

ClickHouse User Password *  
Enter Password
This password will be assigned to the ClickHouse User. The minimum password length is 12 characters. Consider adding digits, capital letters and special symbols to make password more secure.

Confirm Password *
Confirm Password
Please confirm the password

CANCEL  NEXT

Altinity.Cloud

LAUNCH CLUSTER

streamnative 1/1 nodes online

Access Point 

Health	6/6 checks passed
Shards	1
Replicas	1
Node Storage	100 GB
Node Memory	7 GB
Node CPU	2
Version	21.8.10.19
Latest Backup	N/A

CONFIGURE 

EXPLORE 

ACTIONS 

<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

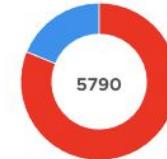
Monitoring View in Grafana

memory



Used
Free

cpu



Used
Free

<https://docs.altinity.com>

Run A Cluster in Altinity Cloud

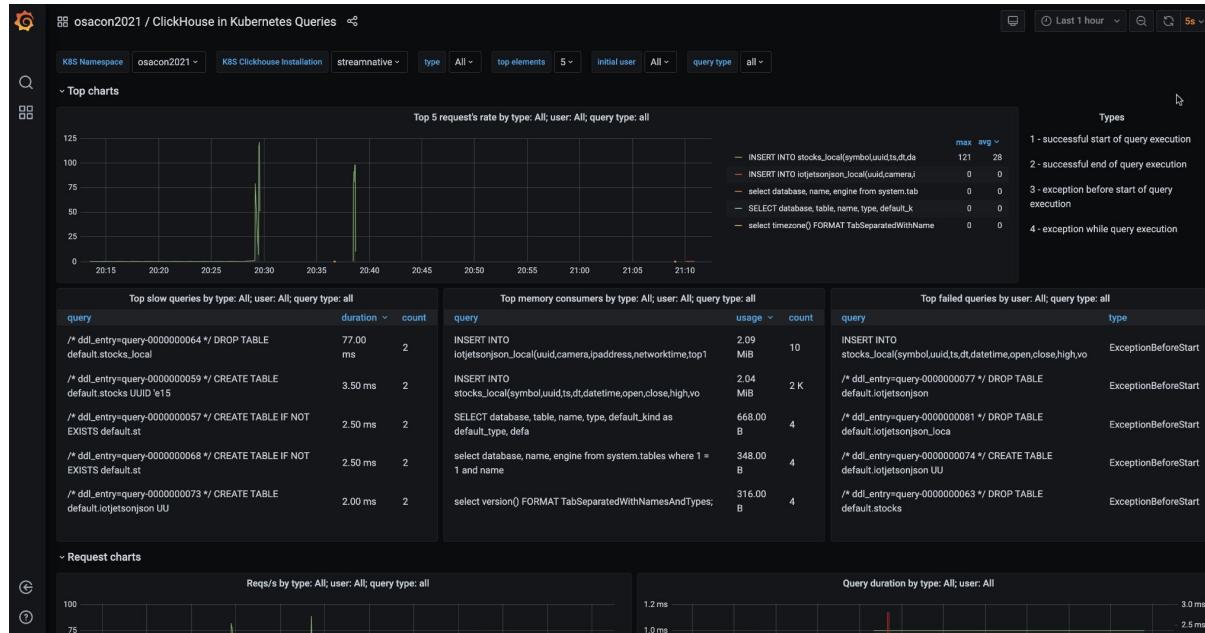
The screenshot shows the Altinity Cloud web interface. At the top, there's a dark header with the Altinity Cloud logo and a navigation bar with 'BACK', 'CLUSTER: STREAMNATIVE', 'Run DDLs ON CLUSTER', and 'NODE: ANY'. Below the header, there are tabs for 'Query' (which is selected), 'Schema', and 'Processes'. Under the 'Query' tab, there's a 'Query History' section with a back/forward button and a text input field containing the query: 'select * from stocks'. Below the input field is a large empty area where the results will be displayed. At the bottom left is a blue 'EXECUTE' button. After executing the query, the results are shown in two tables:

symbol	uuid	ts	dt	datetime	open	close	high	volume	low
ABC	6bec81c6	-1469658966	715224000	2021/01/22 10:06:00	340.83099	341.38000	341.38000	2198	340.83099
IBM	6bec81c6	-1469658966	715224000	2021/01/22 10:06:00	340.83099	341.38000	341.38000	2198	340.83099

symbol	uuid	ts	dt	datetime	open	close	high	volume	low
SNOW	6bec81c6	-1469658966	715224000	2021/01/22 10:06:00	340.83099	341.38000	341.38000	2198	340.83099

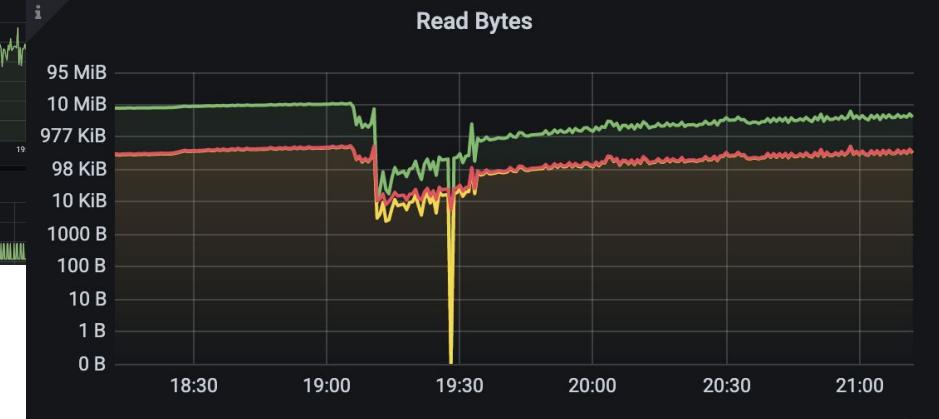
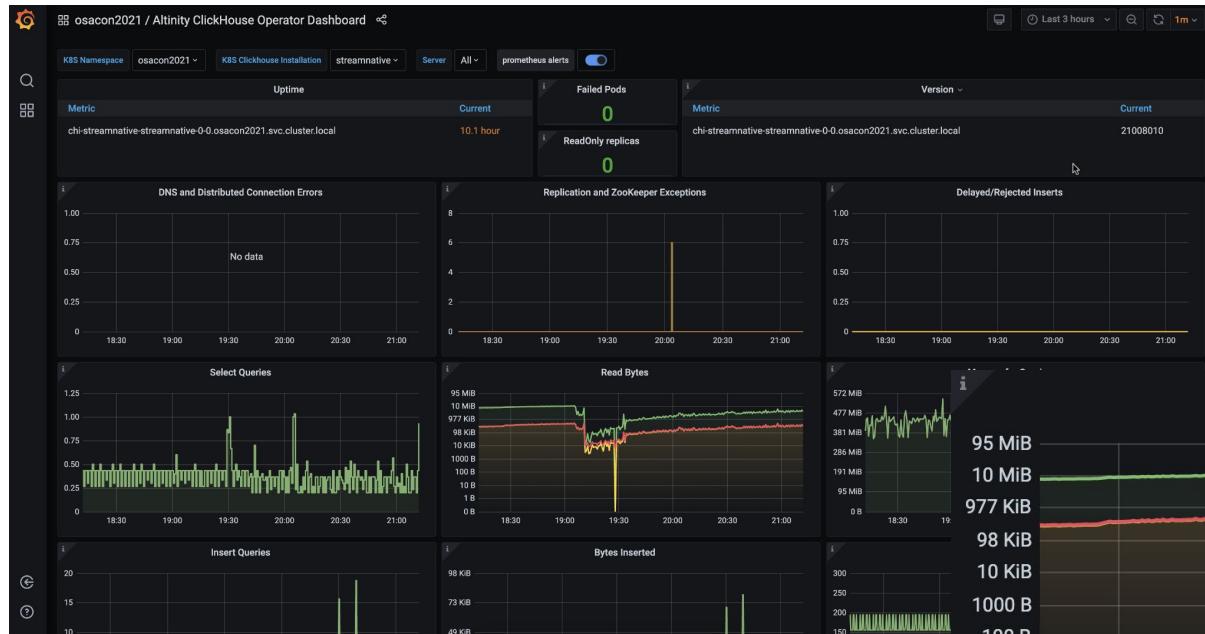
<https://docs.altinity.com>

Monitor The Cluster in Altinity Cloud



<https://docs.altinity.com>

Monitor The Cluster in Altinity Cloud

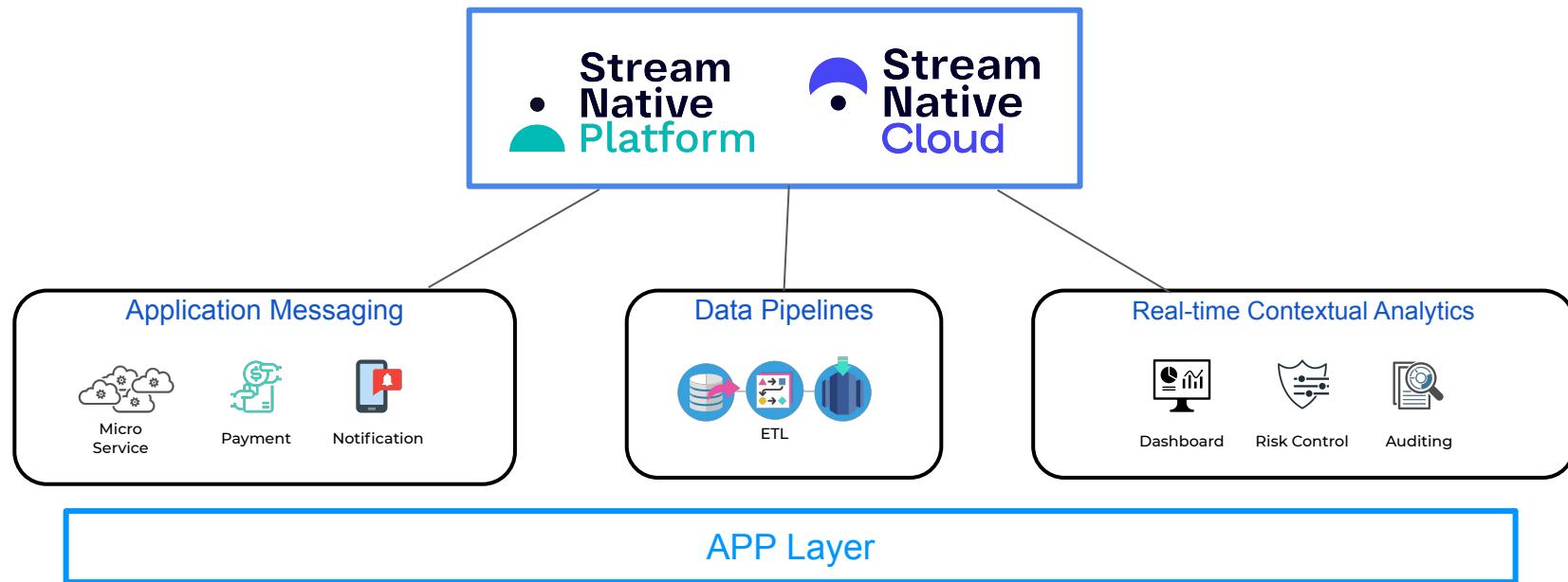


<https://docs.altinity.com>

StreamNative
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



kubernetes

Built for Containers



Flink

Flink SQL

StreamNative Cloud International ▾ Organization - sdev ▾ Instance ▾ tim.spann@streamnative.io ▾

Search Instances Create instance

 aws Ready

aws

[Cluster Details](#)

[Overview](#)

Tenants	Namespaces	Topics
5	4	7

Producers	Subscriptions
0	0

 Usage

In Rate	Out Rate	In Throughput
0	0	0

Out Throughput	Storage Size
0	0



 gke Ready

gke

[Cluster Details](#)

[Overview](#)

Tenants	Namespaces	Topics
3	4	21

Producers	Subscriptions
0	0

 Usage

In Rate	Out Rate	In Throughput
0	0	0

Out Throughput	Storage Size
0	0



 gke-free Ready

gke-free

[Cluster Details](#)

[Overview](#)

Tenants	Namespaces	Topics
3	3	1

Producers	Subscriptions
0	0

 Usage

In Rate	Out Rate	In Throughput
0	0	0

Out Throughput	Storage Size
0	0



 nyc Ready

nyc

[Cluster Details](#)

[Overview](#)

Tenants	Namespaces	Topics
4	4	7

Producers	Subscriptions
0	0

 Usage

In Rate	Out Rate	In Throughput
0	0	0

Out Throughput	Storage Size
0	0



 sgconnector Ready

sgconnector

[Cluster Details](#)

[Overview](#)

Tenants	Namespaces	Topics
3	3	9

Producers	Subscriptions
0	0

 Usage

In Rate	Out Rate	In Throughput
0	0	0

Out Throughput	Storage Size
0	0



StreamNative Cloud @International ▾ Organization - sndev ▾ Instance - gke ▾ tim.spann@streamnative.io ▾

gke

- Tenants
- Namespaces
- Topics
- SQL
- Clients
- Connector
- Manage
- Service Accounts
- Flink Clusters
- Pulsar Clusters
- Instance Setting

Tenant: public Namespace: default

OVERVIEW TOPICS POLICY

+ New Topic

Topic	Partitions	Domain	Producers	Subscriptions	In Rate	Out Rate	In Throughput	Out Throughput	Storage Size
> product	5	persistent	0	0	0	0	0	0	0
> iotjetsonjson	0	persistent	0	0	0	0	0	0	0
> jetsoniottss2	5	persistent	0	0	0	0	0	0	0
> kinesis-output	0	persistent	0	0	0	0	0	0	0
> test1	0	persistent	0	0	0	0	0	0	0
> jetsoniotts	5	persistent	0	0	0	0	0	0	0
> kinesis-input	0	persistent	0	0	0	0	0	0	0
> sensors	0	persistent	0	0	0	0	0	0	0
> test3	0	persistent	0	0	0	0	0	0	0
> [TENANT_NAMESPACE_TO_PIC]	0	persistent	0	0	0	0	0	0	0
> topitems3	5	persistent	0	0	0	0	0	0	0
> data-gen-out	0	persistent	0	0	0	0	0	0	0

StreamNative Cloud International Organization - sndev Instance - gke tim.spann@streamnative.io

gke

- Tenants
- Namespaces
- Topics
 - SQL
 - Clients
 - Connector
 - Manage
 - Service Accounts
 - Flink Clusters
 - Pulsar Clusters
- Instance Setting

Tenant: public Namespace: default Topic: iotjetsonjson

OVERVIEW SCHEMA MESSAGES STORAGE POLICIES

Storage Size: 2 MB

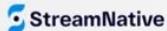
Entries: 2K

Segments: 17

Segments

Ledger ID	Entries	Size	Status	Offload
1342	5	4.51K	closing	false
1353	1	882.00	closing	false
1365	1	940.00	closing	false
1378	2	1.9K	closing	false
1386	1	898.00	closing	false
1975	3	2.79K	closing	false
2007	1	898.00	closing	false
2284	1	805.00	closing	false

StreamNative



Products ▾ Open Source ▾ Resources ▾ Contact 🔎 Login

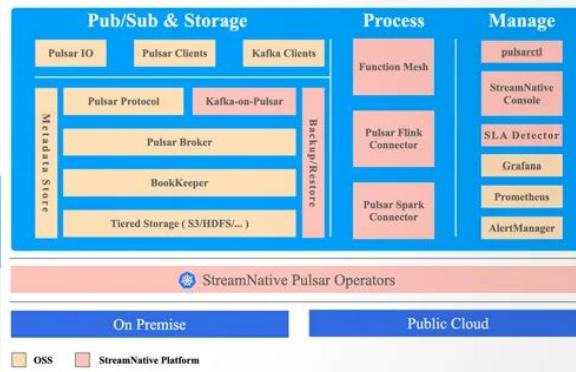
The unified messaging and streaming platform made by the creators of Apache Pulsar.

Built for Kubernetes. Made for the cloud. Enables multi-cloud and hybrid.

[Take A Tour](#) [Contact Sales](#)

Product Update

StreamNative Cloud on AWS Marketplace. Leverage Pulsar on the largest cloud provider with StreamNative Cloud.



StreamNative, Powered by Apache Pulsar



StreamNative Cloud
Apache Pulsar as a service, StreamNative Cloud delivers a resilient and scalable messaging and event streaming service deployable in minutes.



StreamNative Platform
StreamNative Platform is a cloud-native messaging and event streaming platform built by the original creators for Apache Pulsar.



StreamNative Pro Services
Accelerate your messaging and streaming platform development and drive business results with help from StreamNative's Pulsar experts.

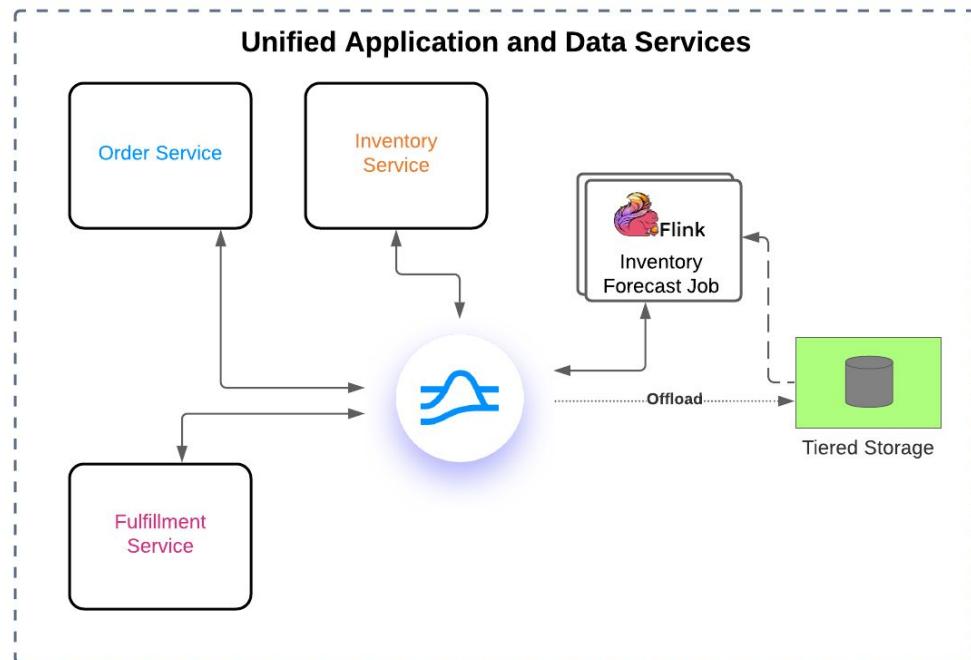
BOSSIE
2021 AWARDS

InfoWorld

Best Practice Architectures

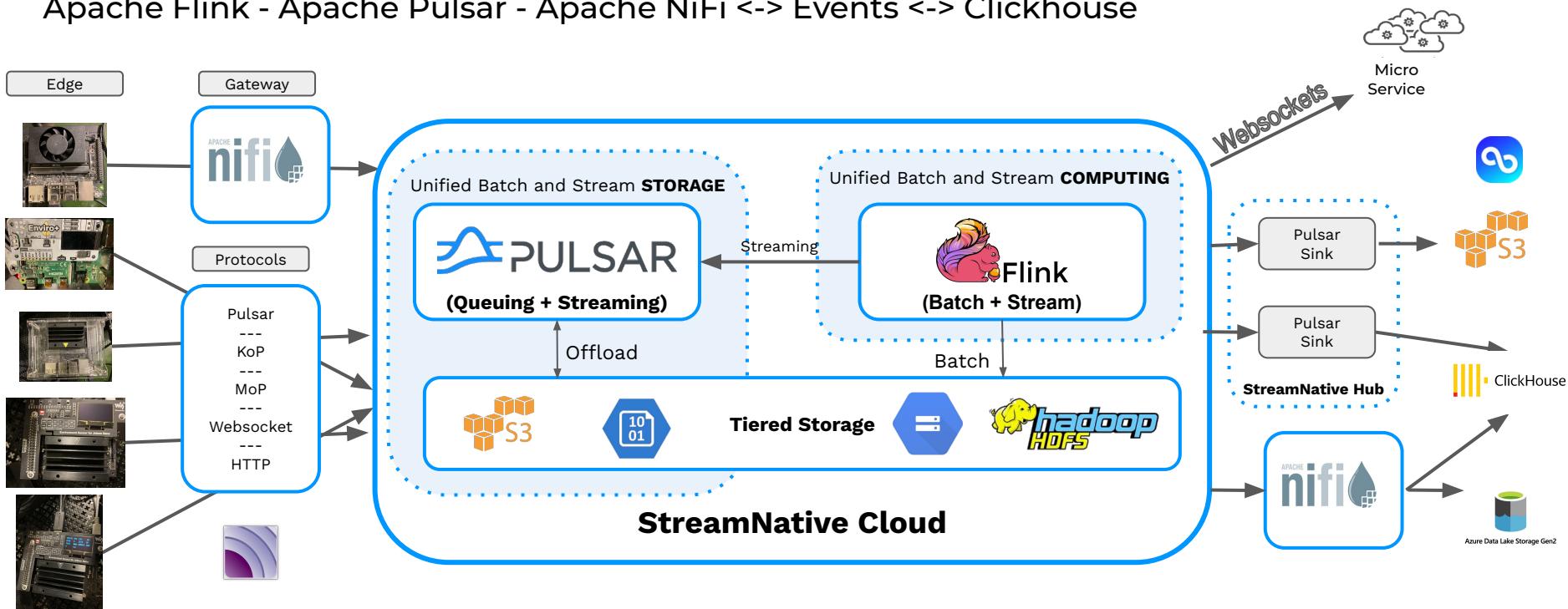
Example: E-Commerce with Pulsar

- 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



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



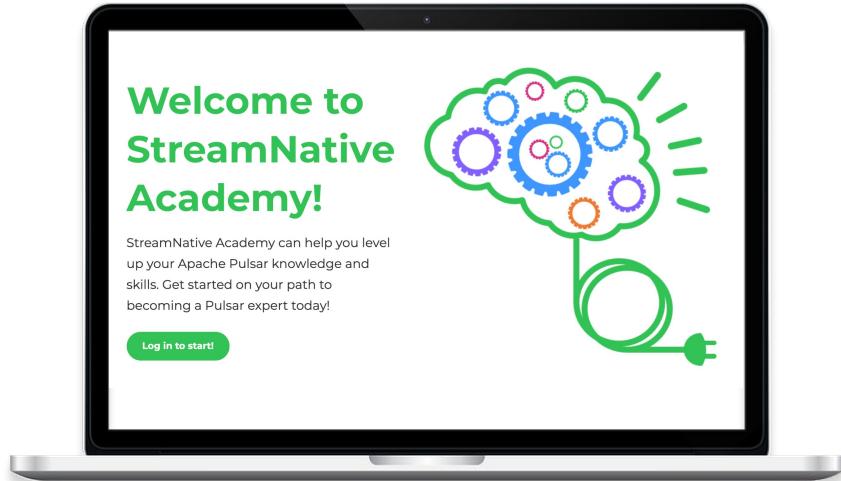
```
{"symbol":"STREAM",
"uuid":"10640832-f139-4b82-8780-e3ad37b3d0
ce",
"ts":1618529574078,
"dt":1612098900000,
"datetime":"2021/01/31 08:15:00",
"open":12.24500,
"close":12.25500,
"high":12.25500,
"volume":12353,
"low":12.24500}
```



Wrap-Up

Now Available On-Demand Pulsar Training

Academy.StreamNative.io



Platform Engineer [Remote]

📍 San Francisco

Platform Engineer (Flink/Spark) [Remote]

📍 San Francisco

Product Engineer - Cloud [Remote]

📍 San Francisco

Platform Engineer (Flink/Spark) [Remote]

📍 San Francisco

Product Engineer - Cloud [Remote]

📍 San Francisco

Sr. Product Manager [Remote]

📍 San Francisco

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](https://twitter.com/streamnativeio) and [@apache pulsar](https://twitter.com/apache_pulsar) on Twitter
- [Subscribe](#) to Monthly Pulsar Newsletter for major news, events, project updates, and resources in the Pulsar community

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](#)

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



@PaasDev



timothyspann

<https://www.pulsardeveloper.com/>

Let's Keep in Touch!



Tim Spann

Developer Advocate



@PassDev



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



<https://github.com/tspannhw>

Questions



Thank You!