Stream Native

Hail Hydrate! From Stream to Lake with Pulsar and Friends

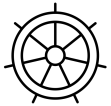
Tim Spann | Developer Advocate

https://portotechhub.com/conference-2021

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

StreamNative

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.

PULSAR

StreamNative

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

What are the Benefits of Pulsar?

Multi-Tenancy

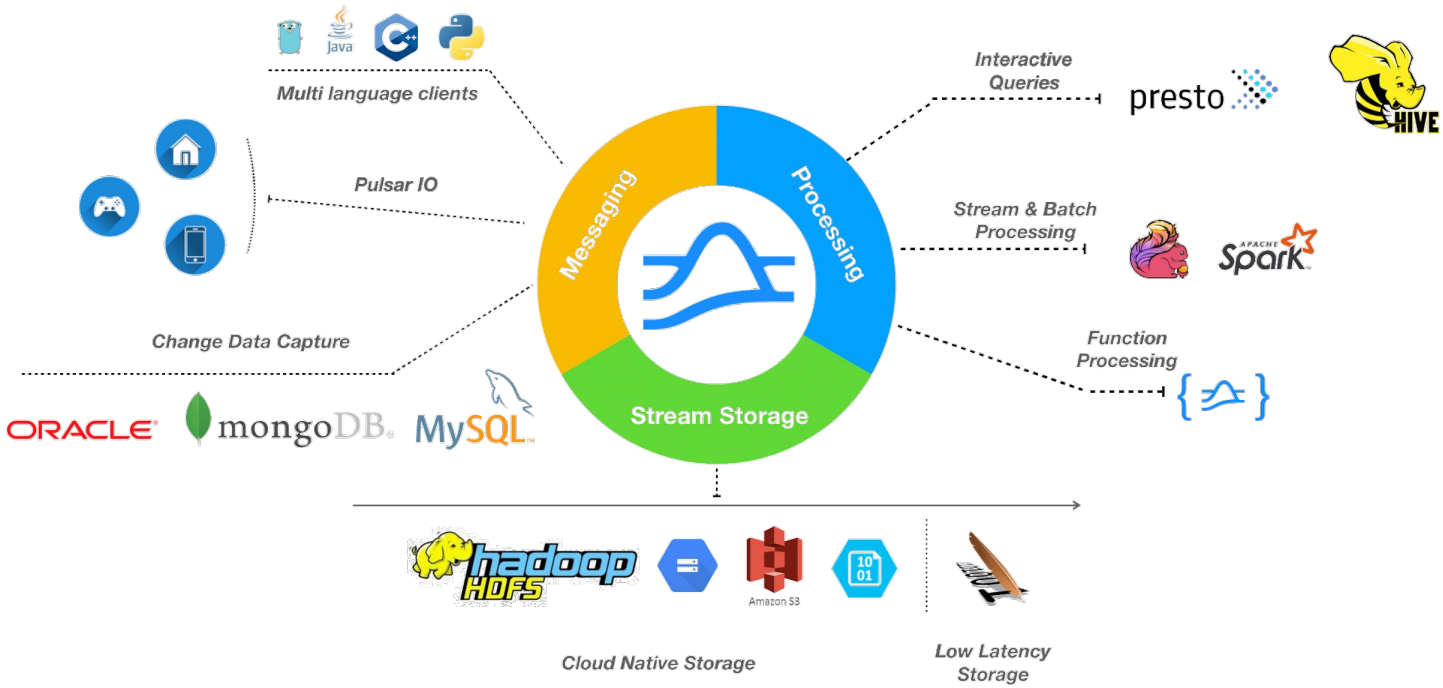
Scalability

Geo-Replication

Unified Messaging Model

Data Durability

Apache Pulsar

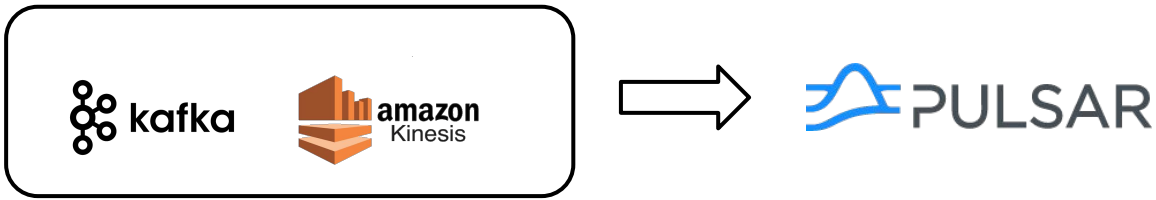


StreamNative

Top Pulsar Use Cases

#1 Message Queuing

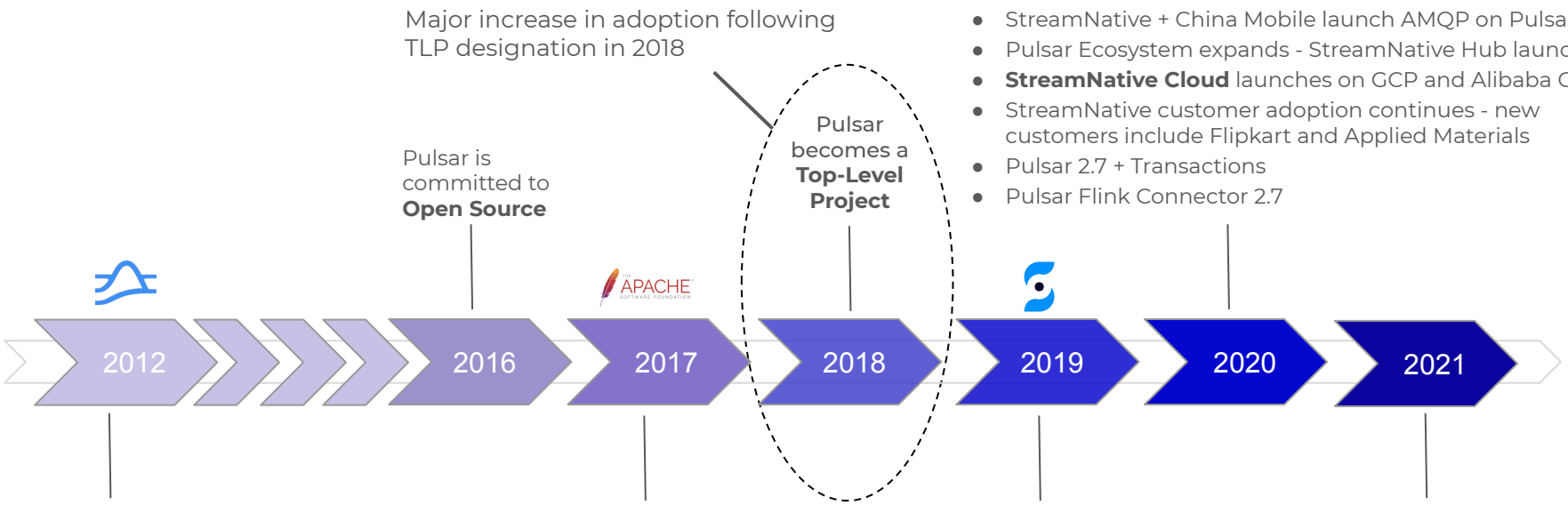
#2 Data Streaming



- Not built for the cloud
- Single tenant systems
- Monolithic architecture couples compute with storage
- Lack of geo replication support

StreamNative

Key Milestones



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

- 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

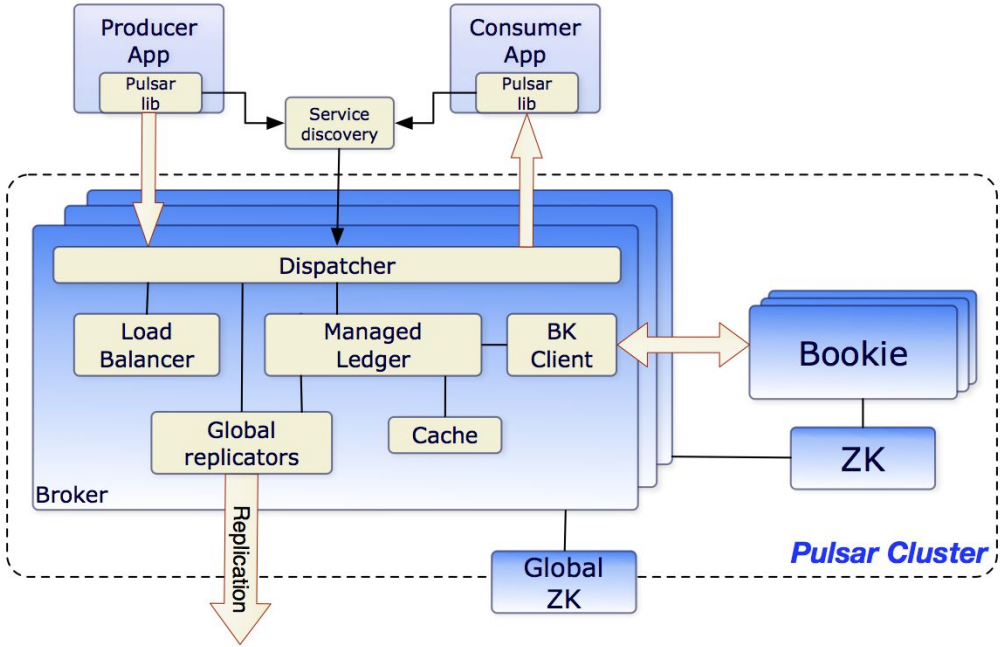
- 3 global Pulsar conferences
- StreamNative hits 400 contributors (June).
- **Pulsar surpasses Kafka** in monthly active contributors.
- Pulsar 2.8 + Exactly-Once semantics
- **StreamNative Platform** launches

StreamNative

Apache Pulsar Overview

Enable Geo-Replicated Messaging

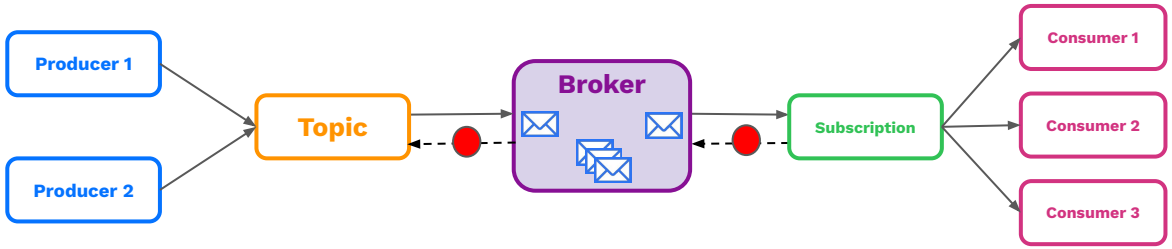
- Pub-Sub
- Geo-Replication
- Pulsar Functions
- Horizontal Scalability
- Multi-tenancy
- Tiered Persistent Storage
- Pulsar Connectors
- REST API
- CLI
- Many clients available
- Four Different Subscription Types
- Multi-Protocol Support
  - MQTT
  - AMQP
  - JMS
  - Kafka
  - ...



StreamNative

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.

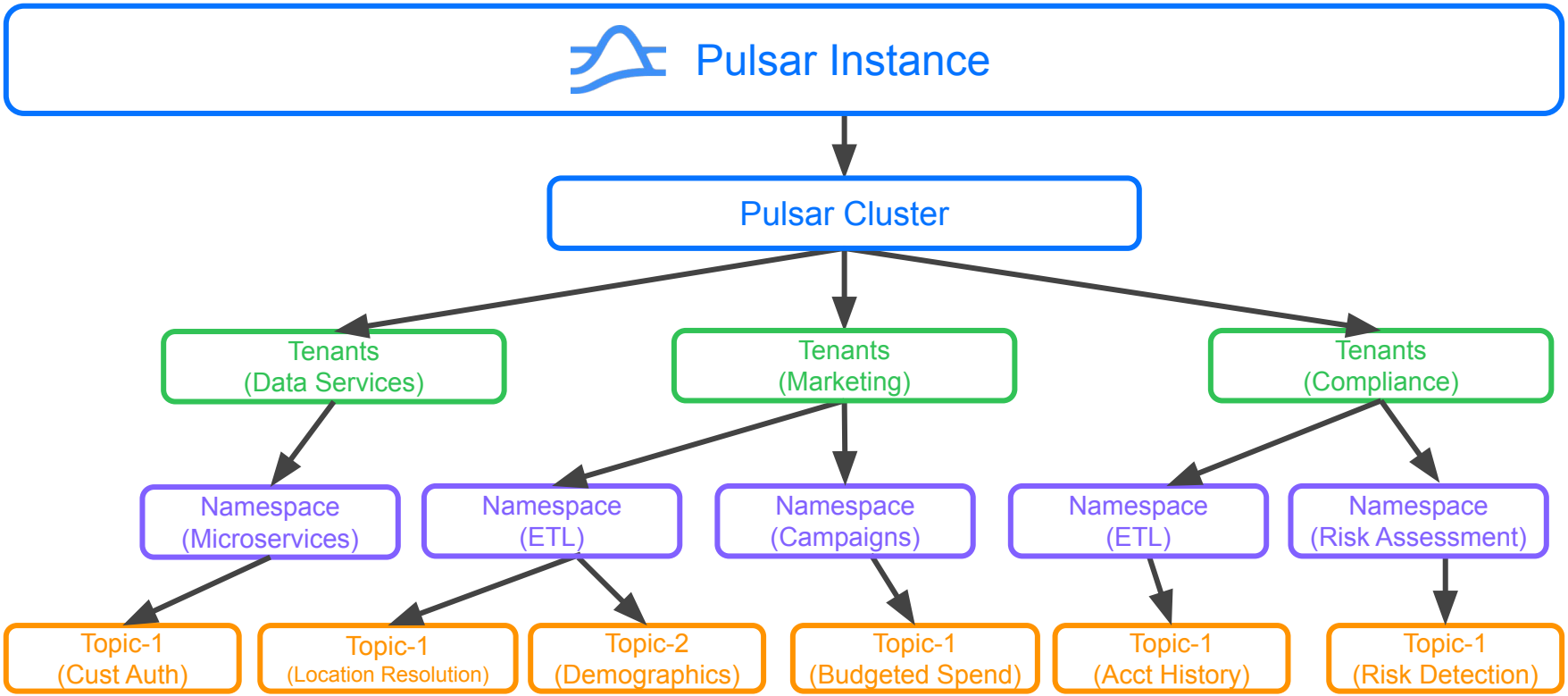


StreamNative

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)

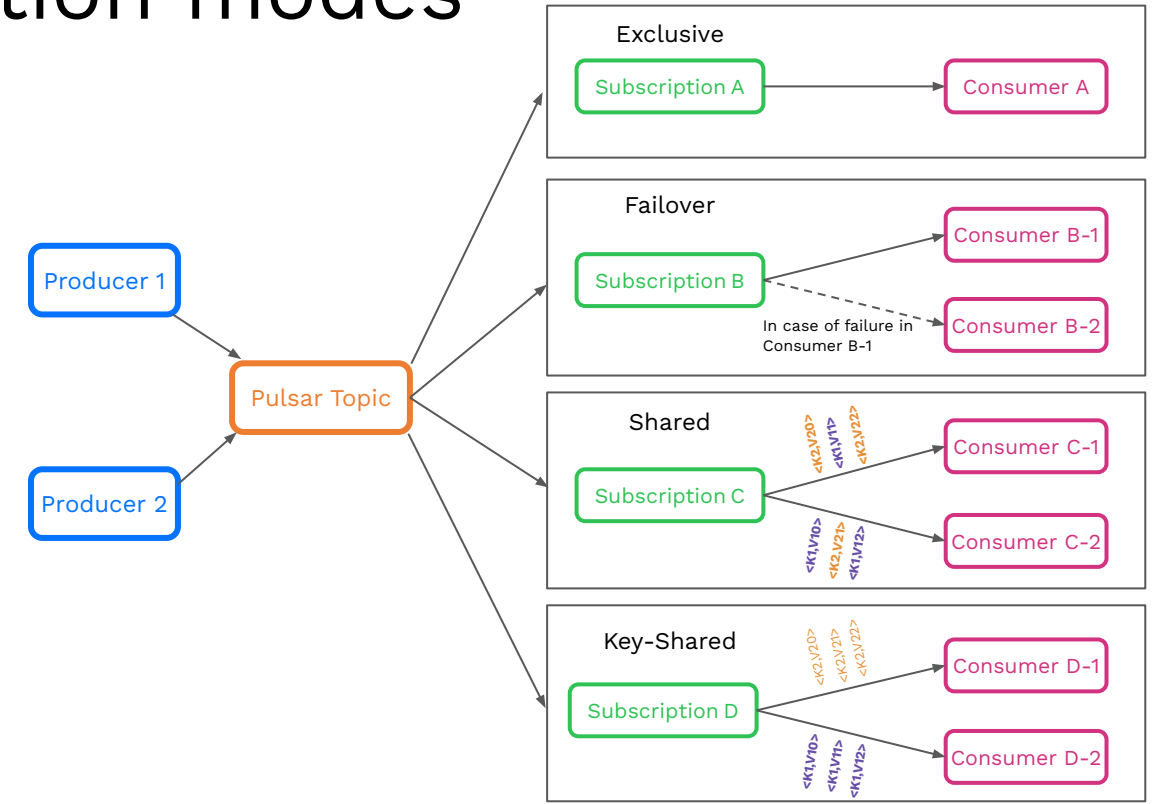
Pulsar subscription modes

Different subscription modes have different semantics:

Exclusive/Failover - guaranteed order, single active consumer

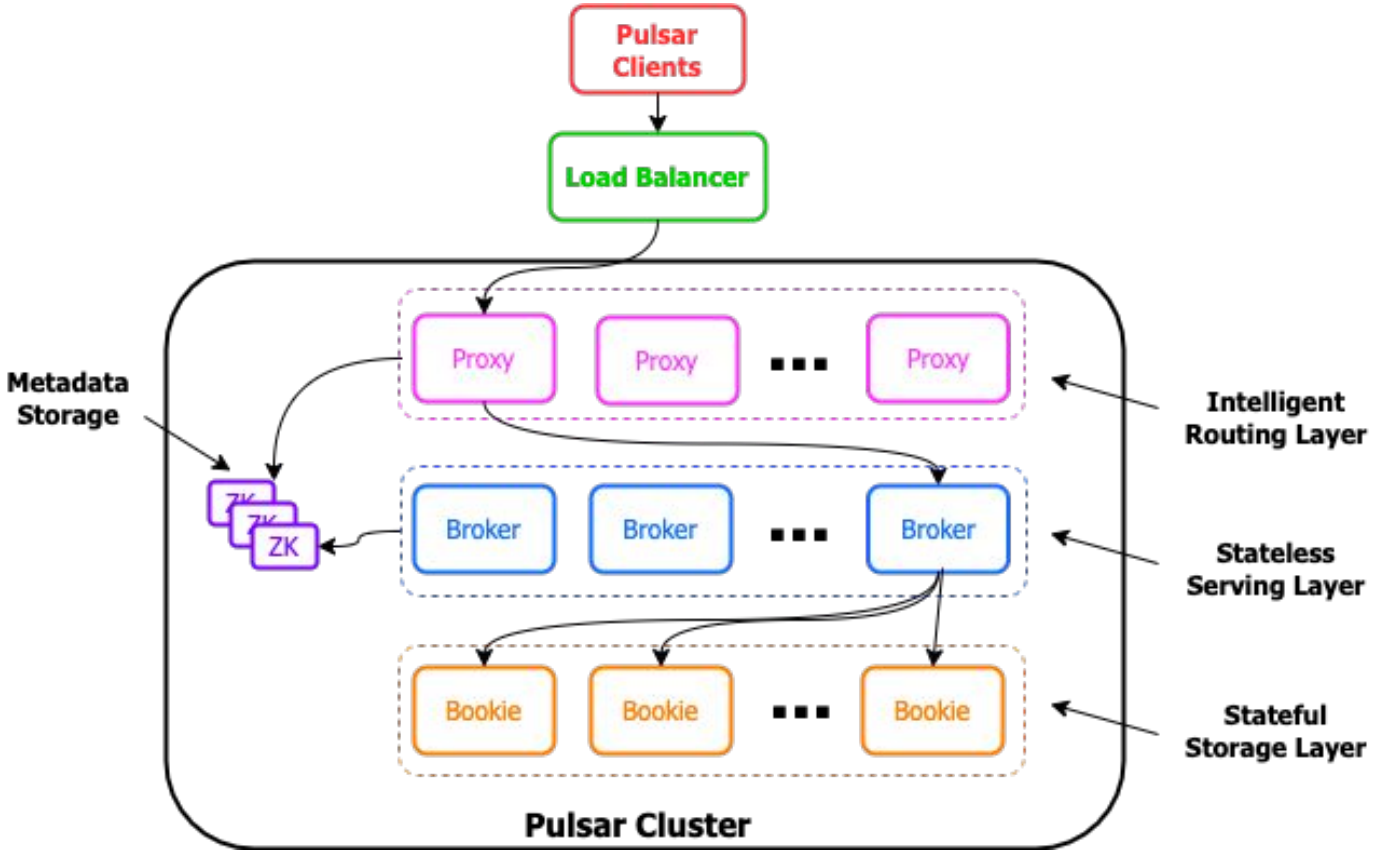
Shared - multiple active consumers, no order

Key_Shared - multiple active consumers, order for given key

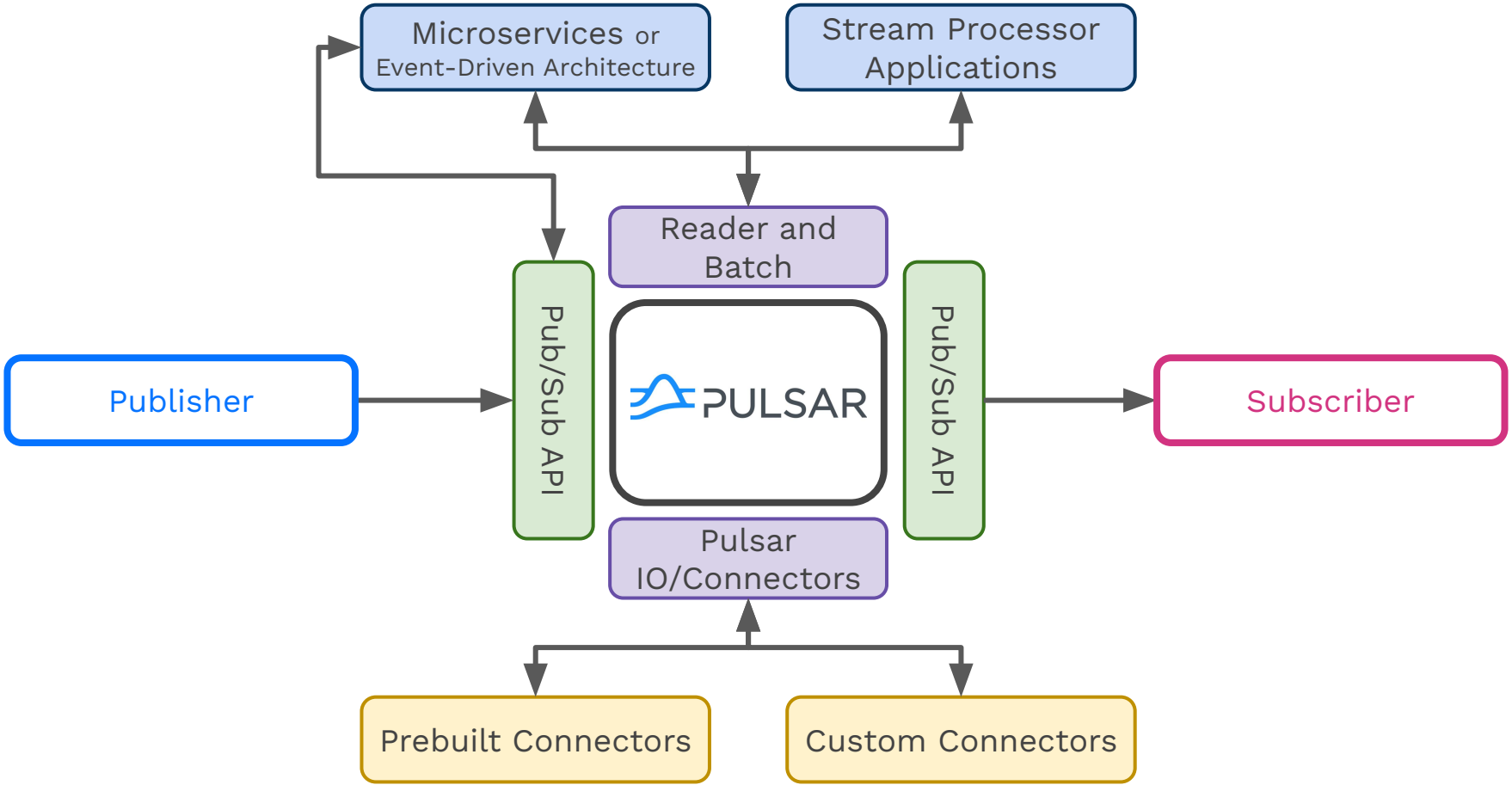


StreamNative

Multi-Tiered Architecture



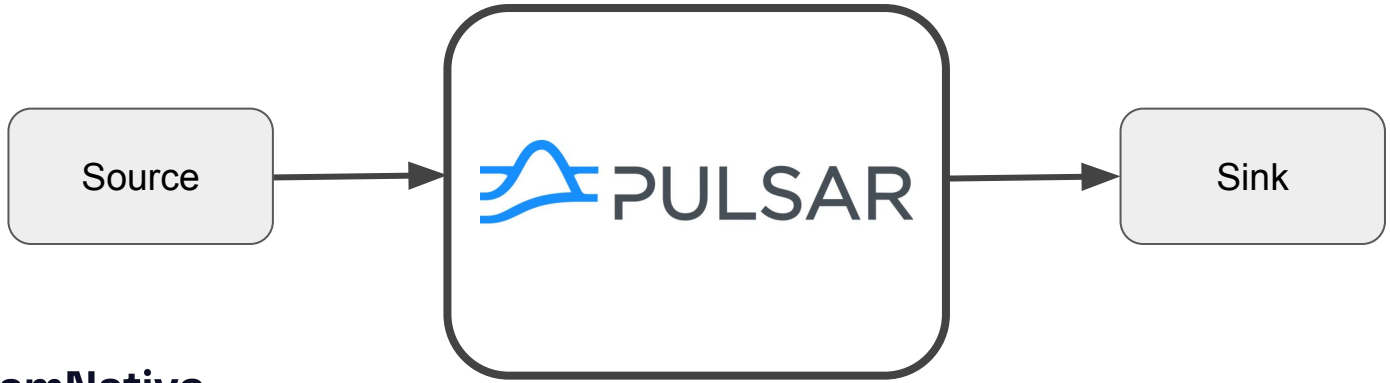
StreamNative


Microservices or
Event-Driven Architecture
Stream Processor
Applications
Reader and
Batch
Publisher
Pub/Sub API
PULSAR
Pub/Sub API
Subscriber
Pulsar
IO/Connectors
Prebuilt Connectors
Custom Connectors

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. https://pulsar.apache.org/docs/en/io-jdbc-sink/

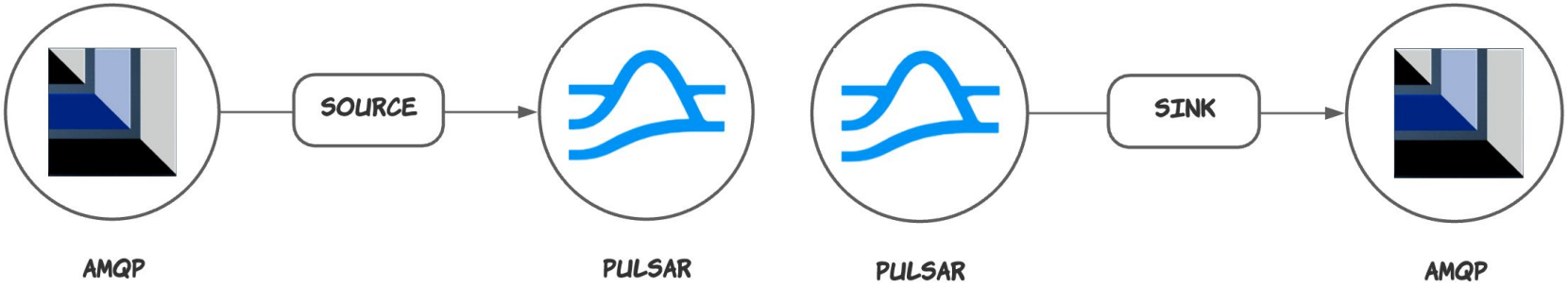
- Built on top of Pulsar Functions
- Built-in connectors - hub.streamnative.io



AMQP / RabbitMQ Protocol

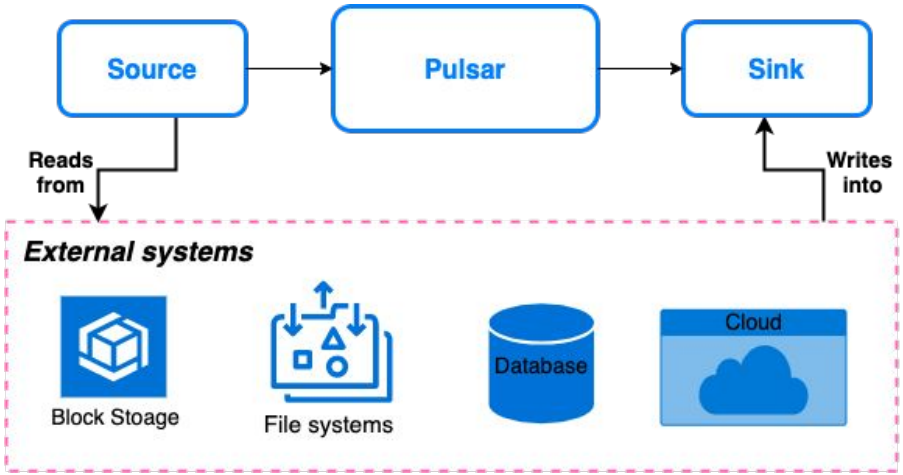
AMQP on Pulsar (AoP) https://github.com/streamnative/aop

https://hub.streamnative.io/connectors/amqp-1-0-sink/
https://hub.streamnative.io/connectors/amqp-1-0-source



Use Azure BlobStore offloader with Pulsar

https://pulsar.apache.org/docs/en/tiered-storage-azure/



StreamNative

Apache Pulsar - Other Sinks

mongoDB

AWS Lambda

redis

AWS S3

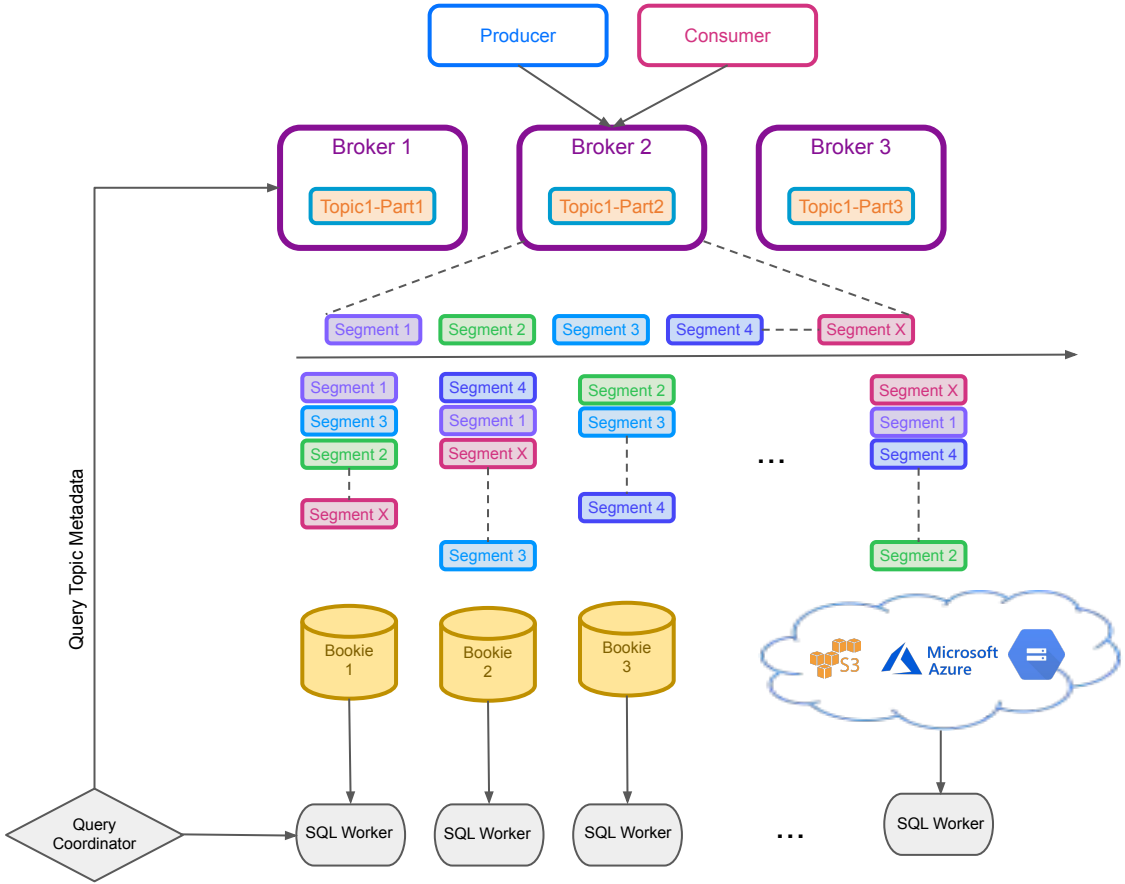
GCS

https://hub.streamnative.io/connectors/cloud-storage-sink/2.5.1/

StreamNative

Pulsar SQL

Presto/Trino workers can read segments directly from bookies (or offloaded storage) in parallel.



trino

StreamNative

Query Your Topics with Pulsar SQL (Trino)

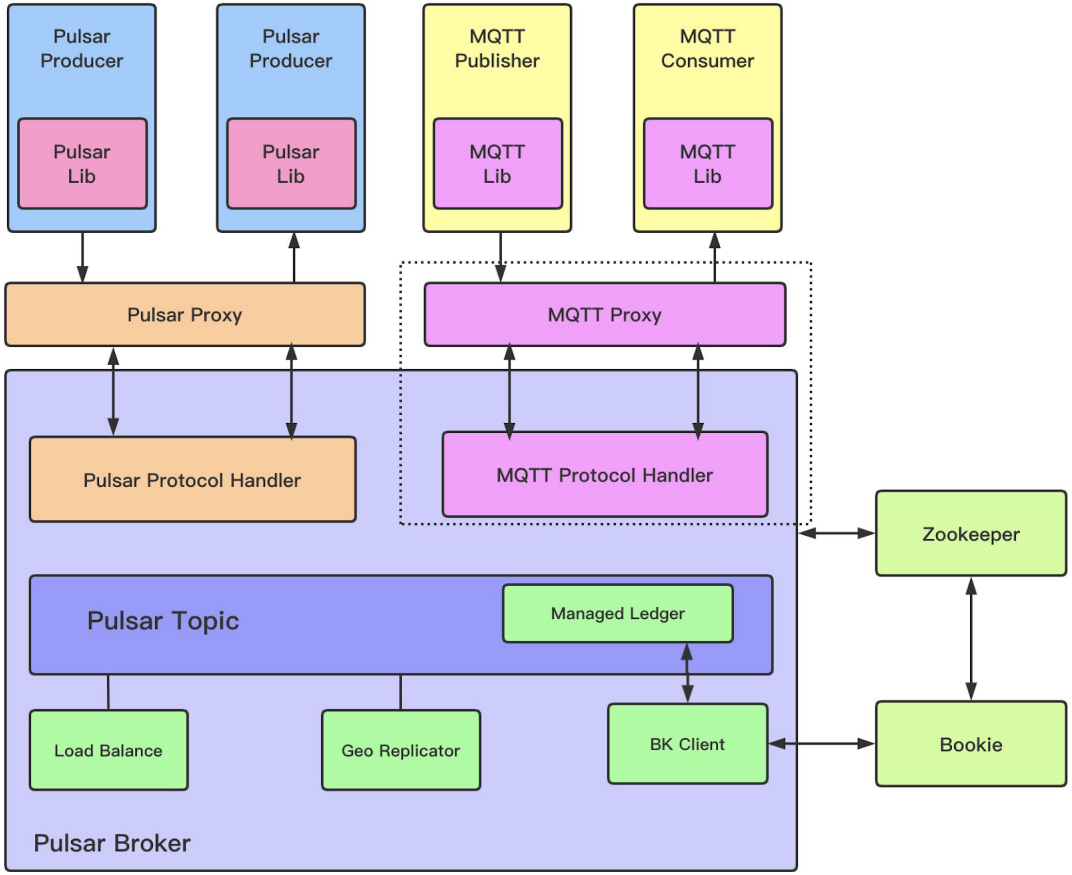
```
presto> select camera, cpu, cputempf, gputempf, memory, top1, top1pct, uuid, __publish_time__, __message_id__, __key__ from  pulsar."public/default".iotjetsonjson;
   camera    | cpu  | cputempf | gputempf | memory |       top1       |    top1pct     |                uuid                |     __publish_time__     | __message_id__ |
-------------+------+----------+----------+--------+------------------+----------------+------------------------------------+--------------------------+----------------+
 /dev/video0 |  8.7 | 82       | 82       |   33.5 | microphone, mike | 18.85986328125 | xav_uuid_video0_lgl_20211001183019 | 2021-10-01 14:30:30.657 | (564,3,0)      |
 /dev/video0 |  8.7 | 82       | 82       |   33.6 | microphone, mike | 19.22607421875 | xav_uuid_video0_kpt_20211001183033 | 2021-10-01 14:30:44.380 | (564,4,0)      |
 /dev/video0 | 12.0 | 80       | 81       |   33.5 | microphone, mike | 12.53662109375 | xav_uuid_video0_gzd_20211001182930 | 2021-10-01 14:29:48.756 | (564,0,0)      |
 /dev/video0 |  8.5 | 82       | 82       |   33.6 | microphone, mike |        14.0625 | xav_uuid_video0_wlw_20211001182951 | 2021-10-01 14:30:02.919 | (564,1,0)      |
 /dev/video0 |  8.5 | 82       | 82       |   33.5 | microphone, mike |     29.8828125 | xav_uuid_video0_ulq_20211001183005 | 2021-10-01 14:30:16.787 | (564,2,0)      |
(5 rows)
(END)
```

```
presto> show tables in pulsar."public/default";
     Table
----------------
 generator_test
 iotjetsonjson
 mqtt-2
(3 rows)

Query 20211001_054538_00008_s8x23, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
0:00 [3 rows, 105B] [14 rows/s, 493B/s]
```

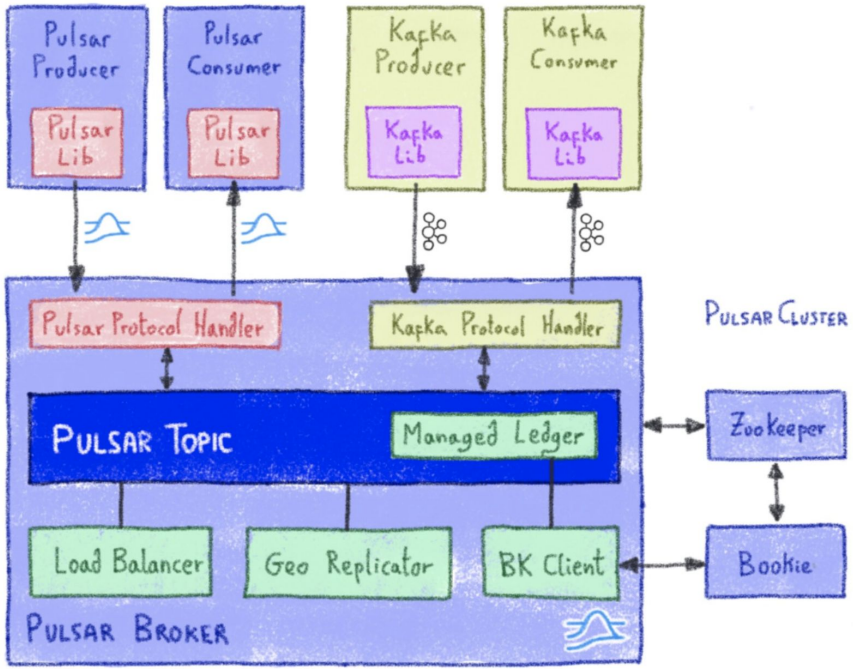
StreamNative

MQTT on Pulsar (MoP)



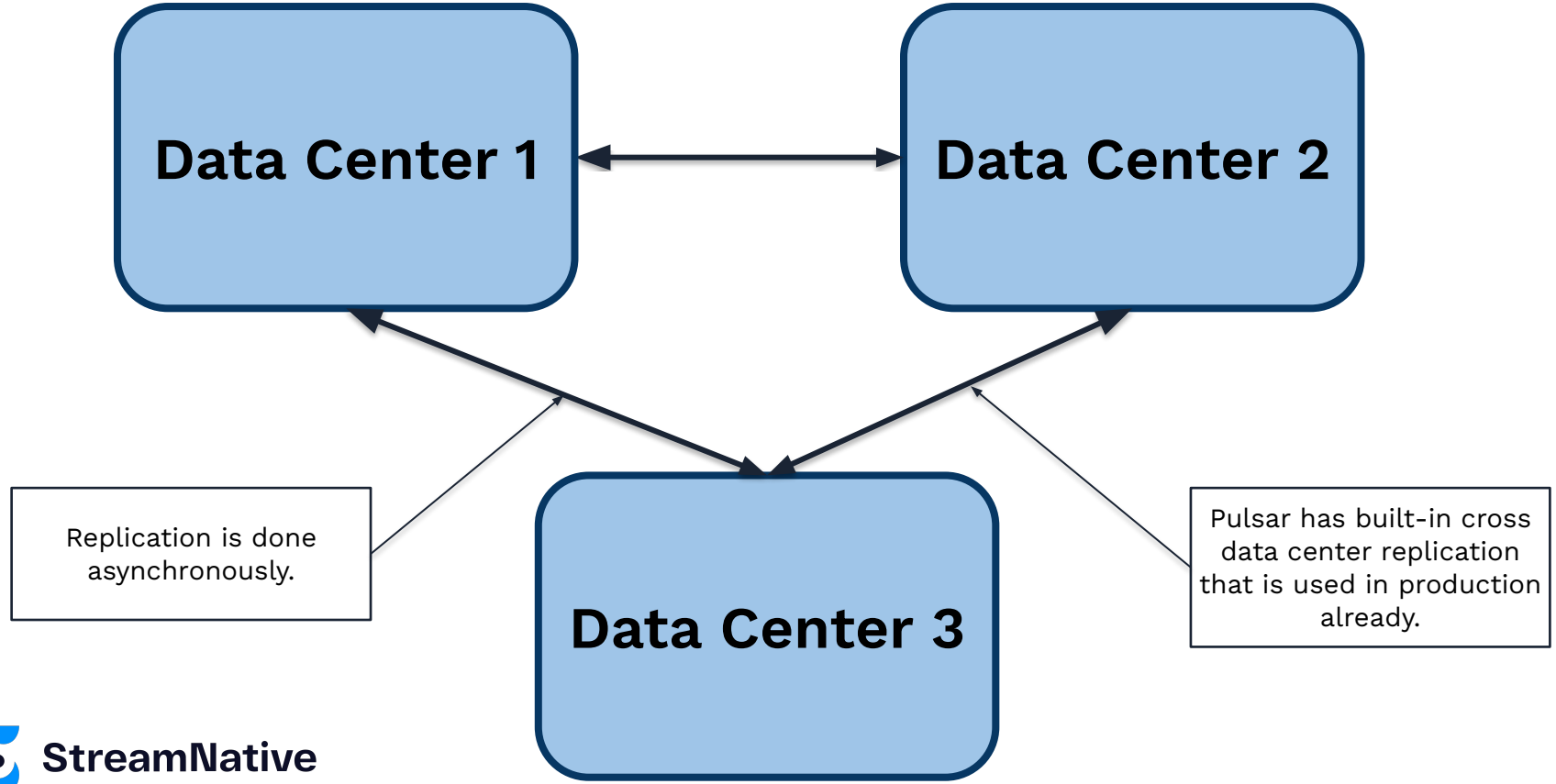
StreamNative

Kafka-on-Pulsar (Kop)

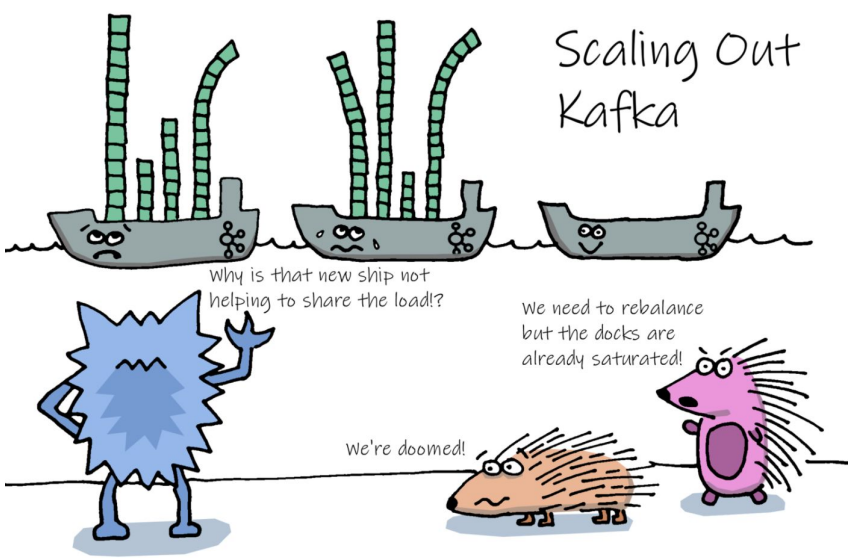


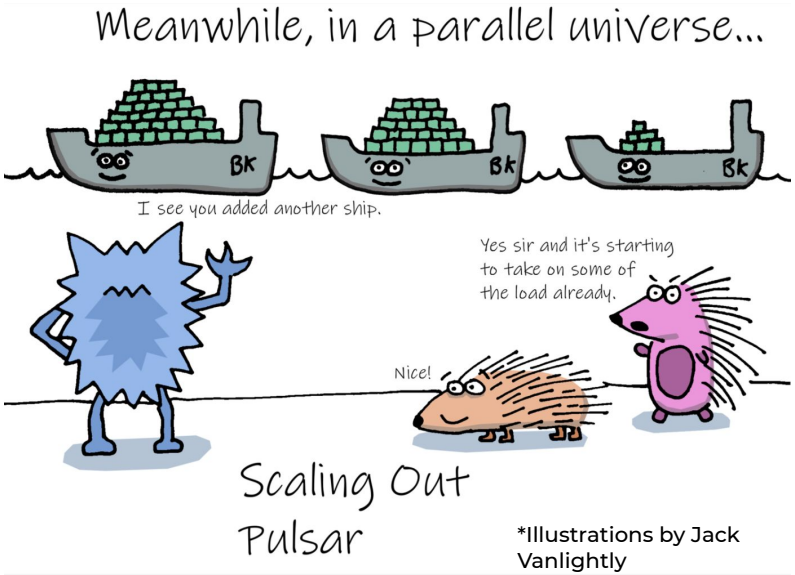
StreamNative

Geo Replication



StreamNative





Pulsar is built for easy scale-out.

StreamNative

StreamNative Cloud

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.

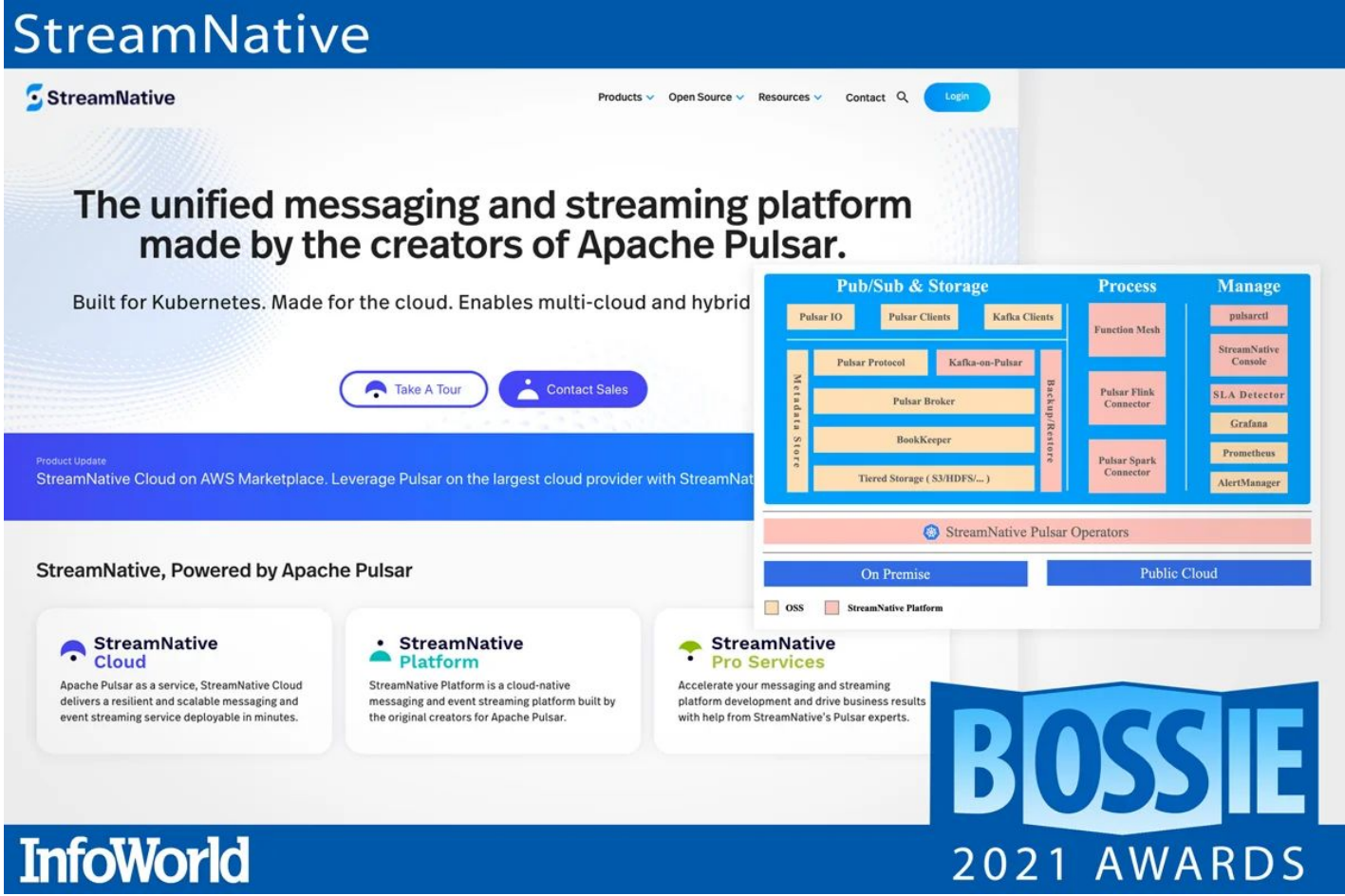
Cloud Native

Built for Containers

Flink SQL

StreamNative

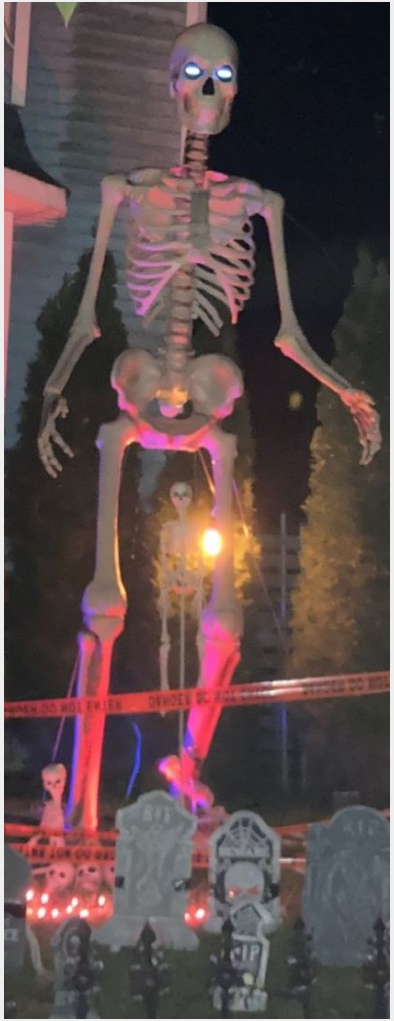
StreamNative



StreamNative

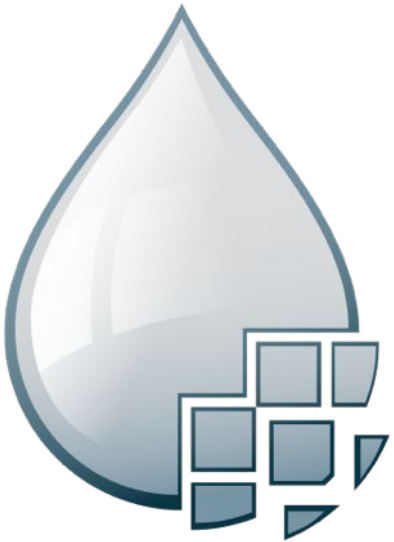
Apache NiFi

Don’t Be Afraid of Open Source



StreamNative

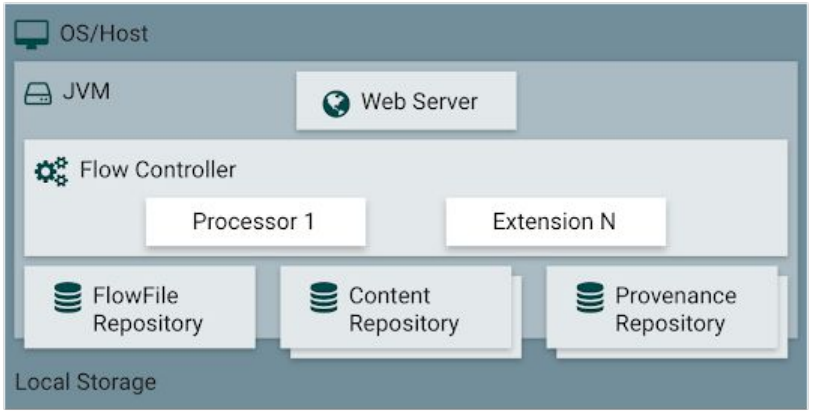
Why Apache NiFi?

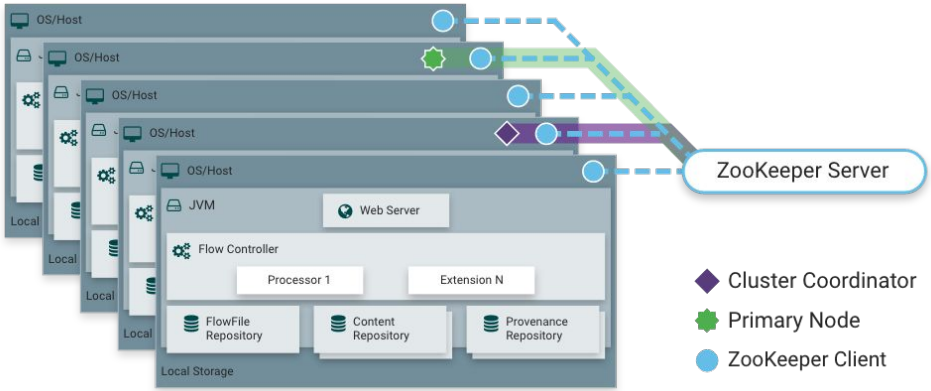


- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models

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

Architecture

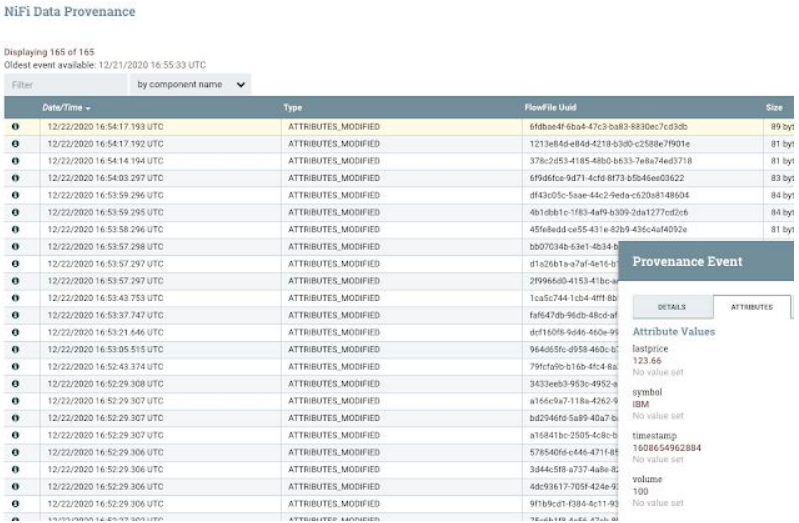




https://nifi.apache.org/docs/nifi-docs/html/overview.html

StreamNative

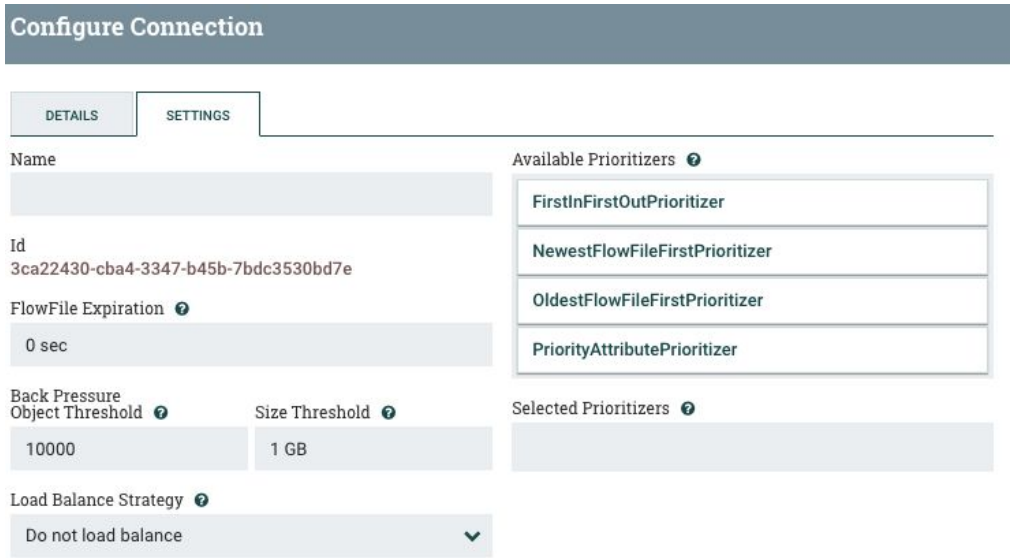
Provenance



https://www.datainmotion.dev/2021/01/automating-starting-services-in-apache.html

StreamNative

Backpressure & Prioritizers



https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

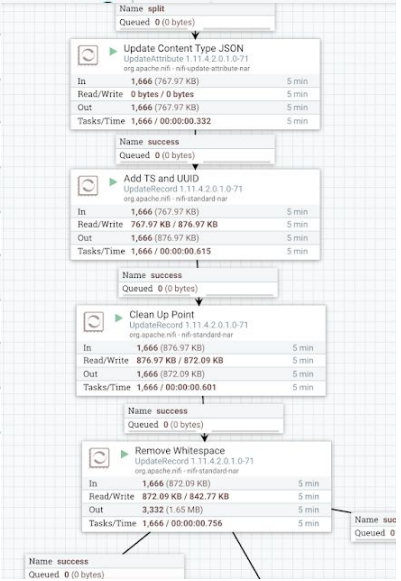
StreamNative

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

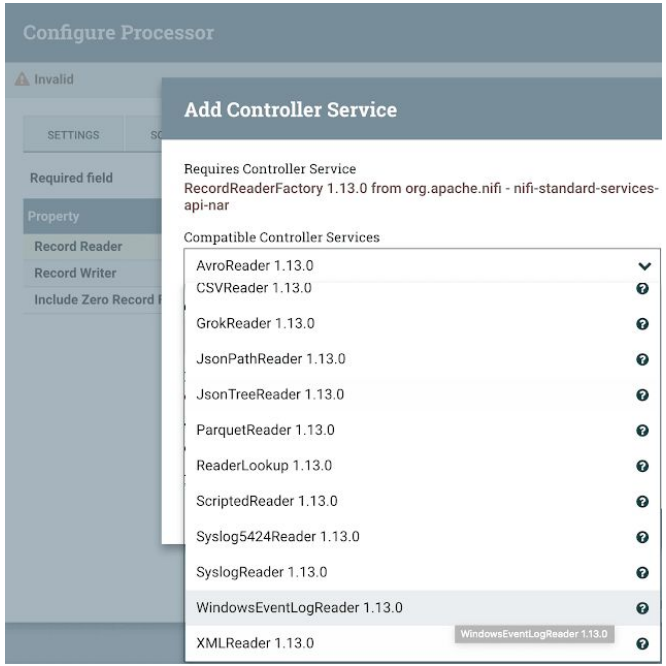
Property	Value
Schema Access Strategy	**Infer Schema**
Schema Registry	AvroSchemaRegistry
Schema Name	${schema.name}
Schema Version	
Schema Branch	
Schema Text	
Schema Inference Cache	
Expect Records as Array	
Attribute Prefix	
Field Name for Content	
Date Format	
Time Format	
Timestamp Format	



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

StreamNative

Record Processors



RecordSinkService 1.13.0 from org.apache.nifi - nifi-standard-services-api-nar

Compatible Controller Services

RecordSinkServiceLookup 1.13.0

Controller Service Name

RecordSinkServiceLookup

Bundle

org.apache.nifi - nifi-record-sink-service-nar

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

StreamNative

Consume MQTT

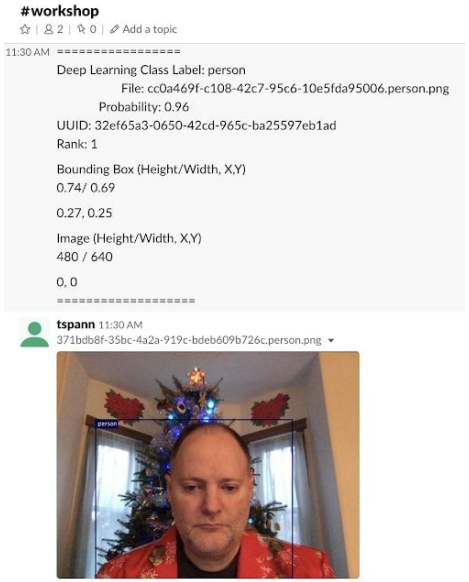
This could read from Apache Pulsar - MoP (MQTT on Pulsar)

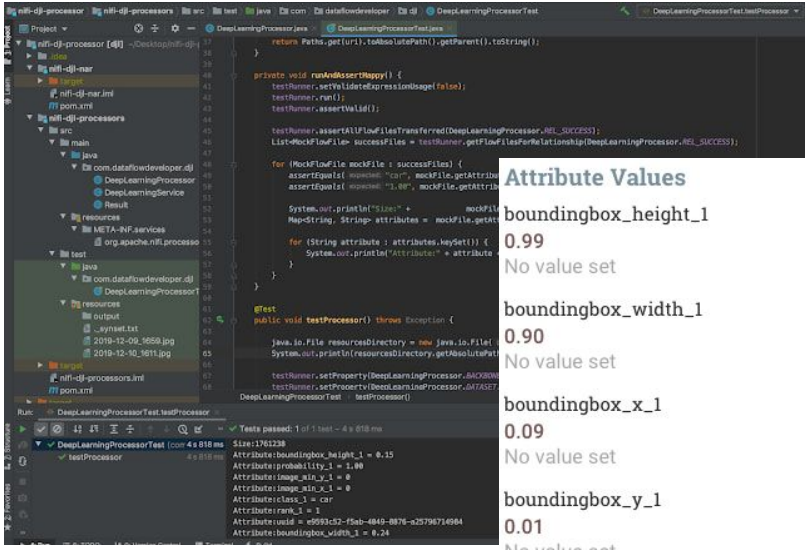
Property	ConsumeMQTT Processor	Value
Session state		**Clean Session**
MQTT Specification Version		**AUTO**
Connection Timeout (seconds)		30
Keep Alive Interval (seconds)		60
Group ID		No value set
Topic Filter		No value set
Quality of Service(QoS)		**0 - At most once**
Max Queue Size		No value set
Record Reader		No value set
Record Writer		No value set
Add attributes as fields		**true**
Message Demarcator		No value set

StreamNative

Apache MXNet Native Processor through DJL.AI for Apache NiFi

DJL





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

StreamNative

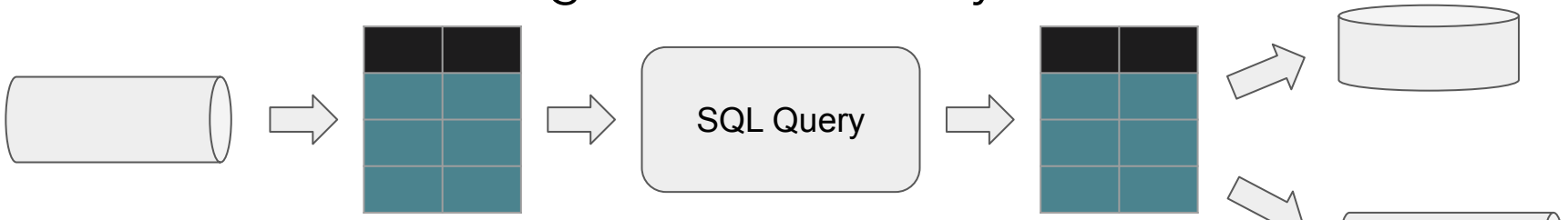
Apache Flink

Flink

https://flink.apache.org/2021/01/07/pulsar-flink-connector-270.html

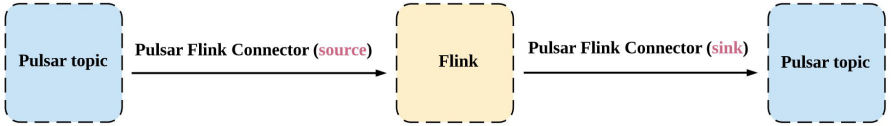
StreamNative

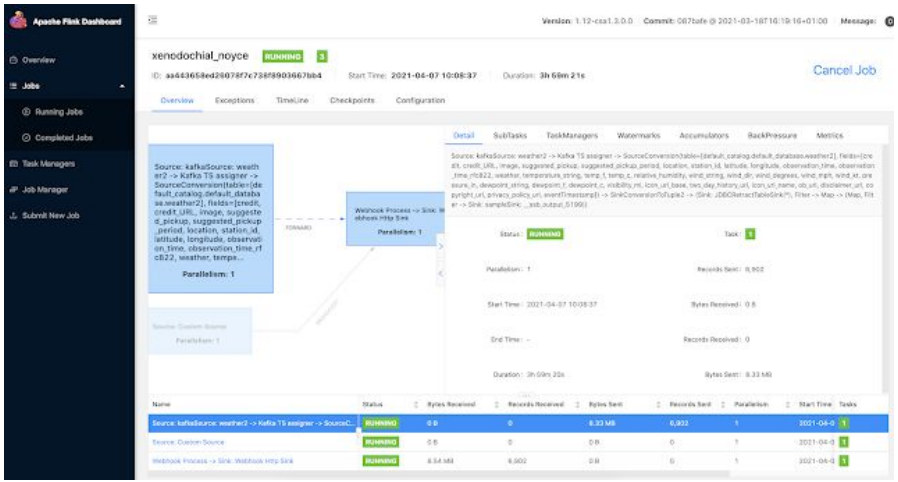
SQL / Table API: Running The Same Query On Streams

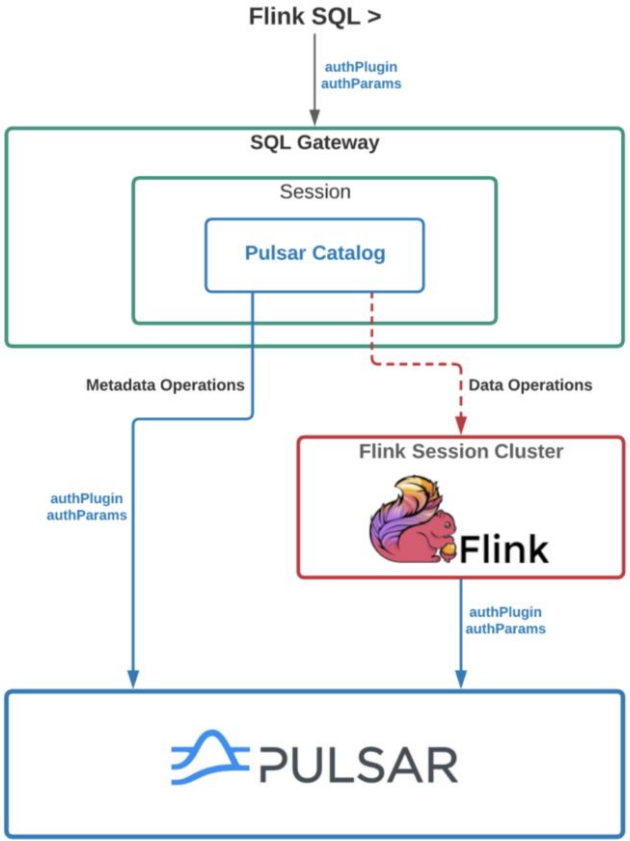



```
SELECT
      room,
      TUMBLE_END(rowtime, INTERVAL ‘1’ HOUR),
      AVG(temperature)
FROM
      sensors
GROUP BY
      TUMBLE(rowtime, INTERVAL ‘1’ HOUR), room
```

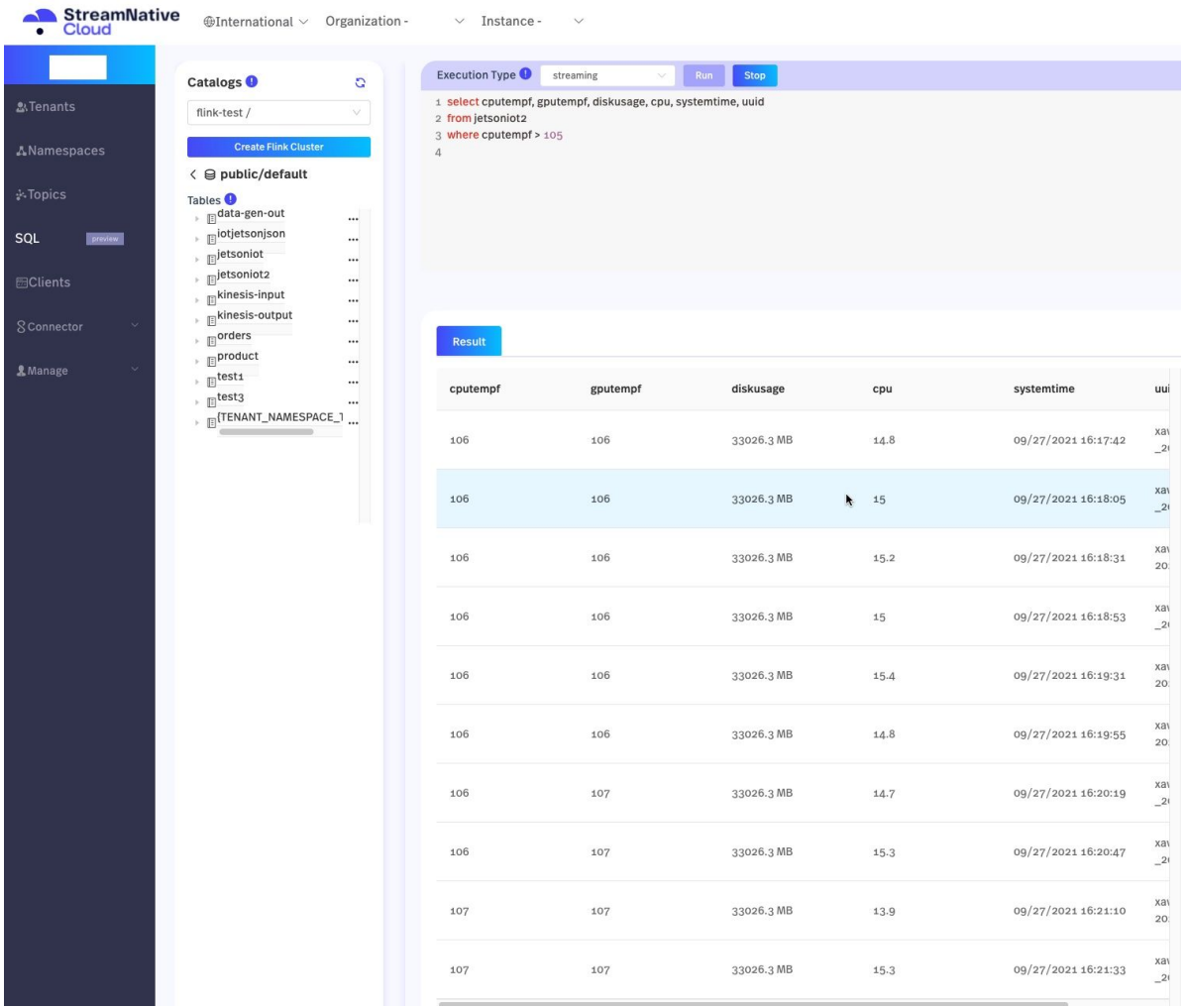
Flink SQL To Pulsar Catalog







StreamNative

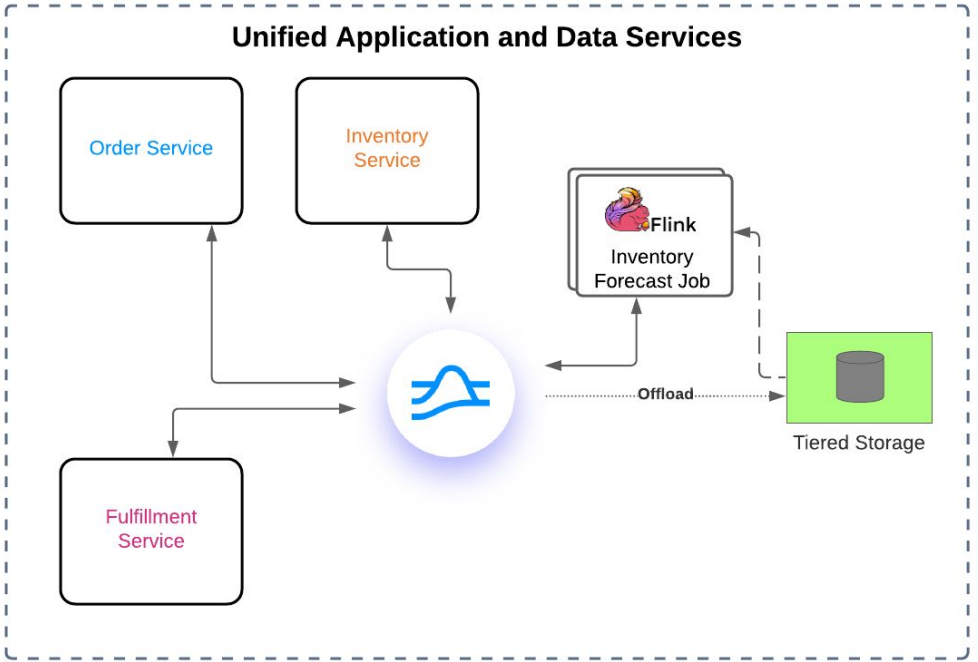


StreamNative

Best Practice Architectures

Example: E-Commerce with Pulsar

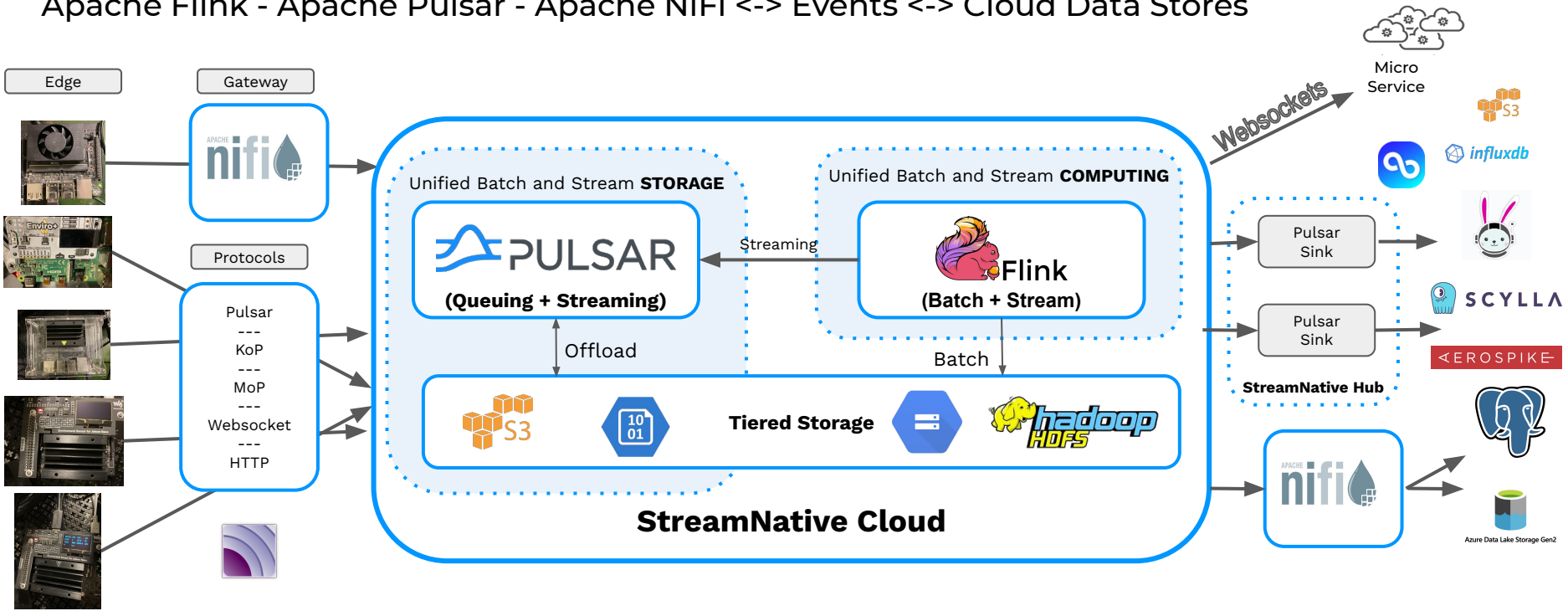
- Unified storage with access to underlying data
- Native tiered storage
- Single system to exchange data
- Teams share toolset



StreamNative

End-to-End Streaming FLiP(N) Apps

Apache Flink - Apache Pulsar - Apache NiFi <-> Events <-> Cloud Data Stores



StreamNative

Demo

StreamNative

Ingesting IoT Data via Java Pulsar

```java
UUID uuidKey = UUID.randomUUID();
String pulsarKey = uuidKey.toString();
String OS = System.getProperty("os.name").toLowerCase();
String message = "" + jct.message;
IoTMessage iotMessage = parseMessage("" + jct.message);
String topic = DEFAULT_TOPIC;
if ( jct.topic != null && jct.topic.trim().length()>0) {
    topic = jct.topic.trim();
}
ProducerBuilder<IoTMessage> producerBuilder = client.newProducer(JSONSchema.of(IoTMessage.class))
    .topic(topic)
    .producerName("jetson")
    .sendTimeout(5, TimeUnit.SECONDS);

Producer<IoTMessage> producer = producerBuilder.create();

MessageId msgID = producer.newMessage()
                .key(iotMessage.getUuid())
                .value(iotMessage)
                .send();
```

https://github.com/tspannhw/StreamingAnalyticsUsingFlinkSQL/

Ingesting IoT Data via Java Pulsar

```
private static IoTMessage parseMessage(String message) {

    IoTMessage iotMessage = null;

    try {
        if ( message != null && message.trim().length() > 0) {
            ObjectMapper mapper = new ObjectMapper();
            iotMessage = mapper.readValue(message, IoTMessage.class);
            mapper = null;
        }
    }
    catch(Throwable t) {
        t.printStackTrace();
    }

    if (iotMessage == null) {
        iotMessage = new IoTMessage();
    }
    return iotMessage;
}
```

StreamNative

MQTT from Python

```
pip3 install paho-mqtt

import paho.mqtt.client as mqtt
client = mqtt.Client("rpi4-iot")

row = { }
row['gasKO'] = str(readings)
json_string = json.dumps(row)
json_string = json_string.strip()
client.connect("pulsar-server.com", 1883, 180)
client.publish("persistent://public/default/mqtt-2", payload=json_string,
                    qos=0, retain=True)
```

StreamNative

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
https://github.com/tspannhw/FLiP-Mobile/blob/30bcc1ec98fc31e039b51a06180d98545c1e0542/python3/enviro.py

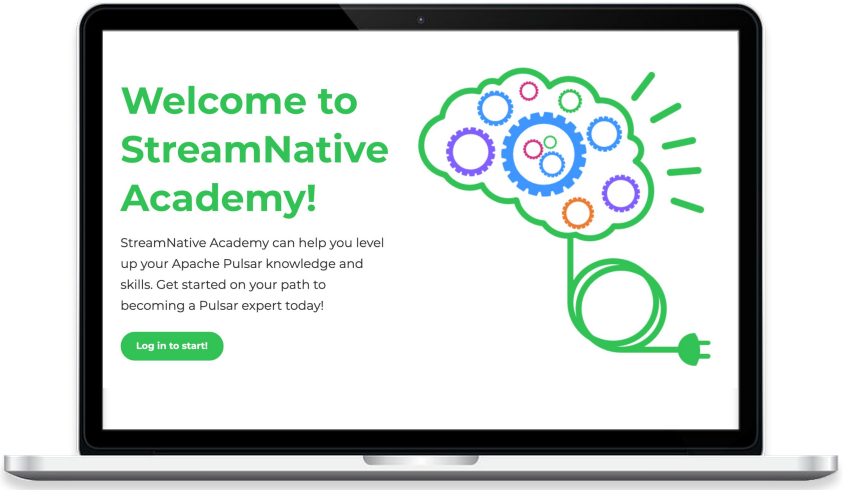
StreamNative

Wrap-Up

Now Available

On-Demand Pulsar Training

Academy.StreamNative.io



StreamNative

We're Hiring

streamnative.io/careers/

StreamNative

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

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]

StreamNative

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/

Let’s Keep in Touch!

Tim Spann

Developer Advocate

@PassDev

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

https://github.com/tspannhw

StreamNative

Questions