



# Using Apache NiFi with Apache Pulsar for Fast Data On-Ramp

Tim Spann | Developer Advocate





**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/>



# FLiP Stack Weekly

This week in Apache Flink, Apache Pulsar, Apache NiFi, Apache Spark and open source friends.

<https://bit.ly/32dAJft>

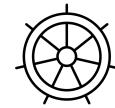
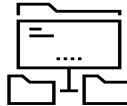


- Founded the original developers of Apache Pulsar.
- Passionate and dedicated team.
- StreamNative helps teams to capture, manage, and leverage data using Pulsar's unified messaging and streaming platform.

# The Need For Real-Time Data



**Hybrid and multi-cloud strategies** with native geo-replication



**Built for Kubernetes**  
CloudNative migrations with tools



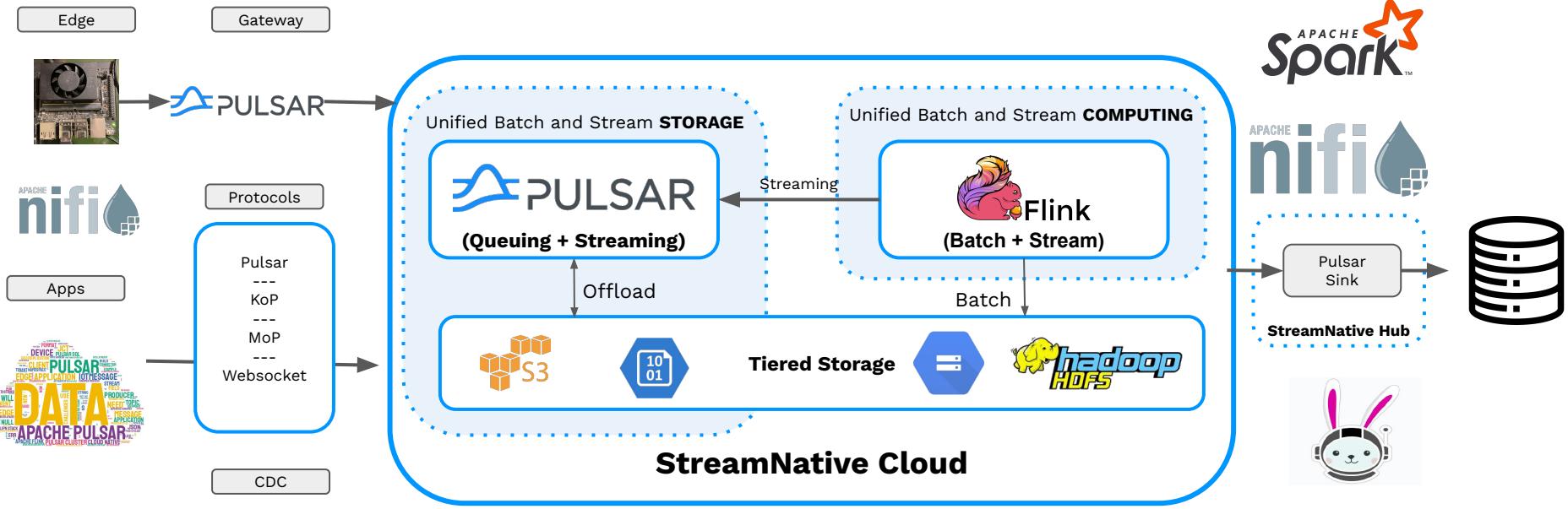
**Seamlessly build microservice architectures** with support for streaming and messaging workloads

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

# Streaming FLIPN Apps

Events <->  PULSAR <-> Events

 Flink



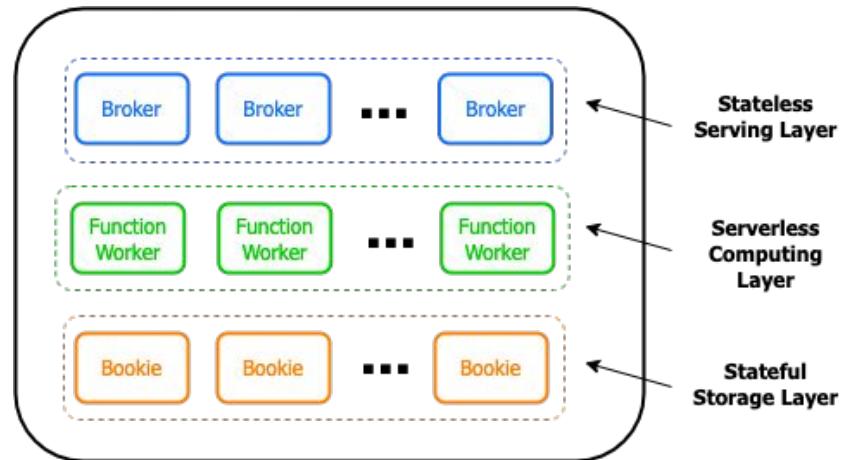
 PULSAR  
Functions

Apache Pulsar



# Apache Pulsar

- Serverless computing framework.
- Unbounded storage, multi-tiered architecture, and tiered-storage.
- Streaming & Pub/Sub messaging semantics.
- Multi-protocol support



# Why Apache Pulsar?



Unified  
Messaging Platform



Guaranteed  
Message Delivery



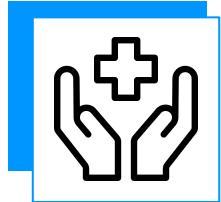
Resiliency



Infinite  
Scalability

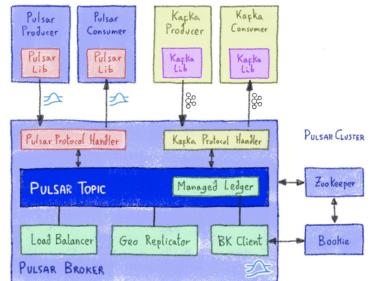
# Messages - the basic unit of Pulsar

Component	Description
<b>Value / data payload</b>	The data carried by the message. All Pulsar messages contain raw bytes, although message data can also conform to data schemas.
<b>Key</b>	Messages are optionally tagged with keys, used in partitioning and also is useful for things like topic compaction.
<b>Properties</b>	An optional key/value map of user-defined properties.
<b>Producer name</b>	The name of the producer who produces the message. If you do not specify a producer name, the default name is used. Message De-Duplication.
<b>Sequence ID</b>	Each Pulsar message belongs to an ordered sequence on its topic. The sequence ID of the message is its order in that sequence. Message De-Duplication.



# Connectivity

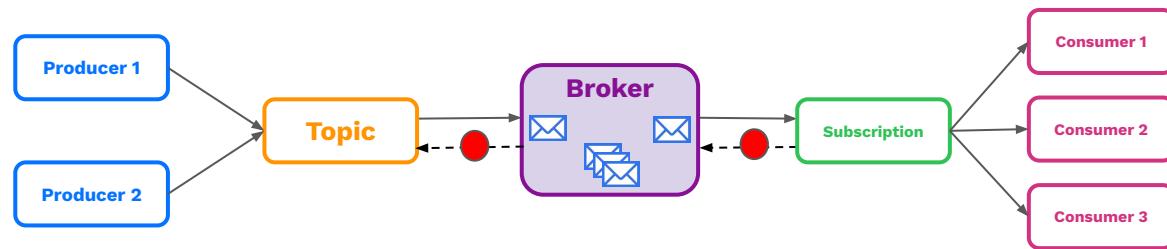
[hub.streamnative.io](http://hub.streamnative.io)



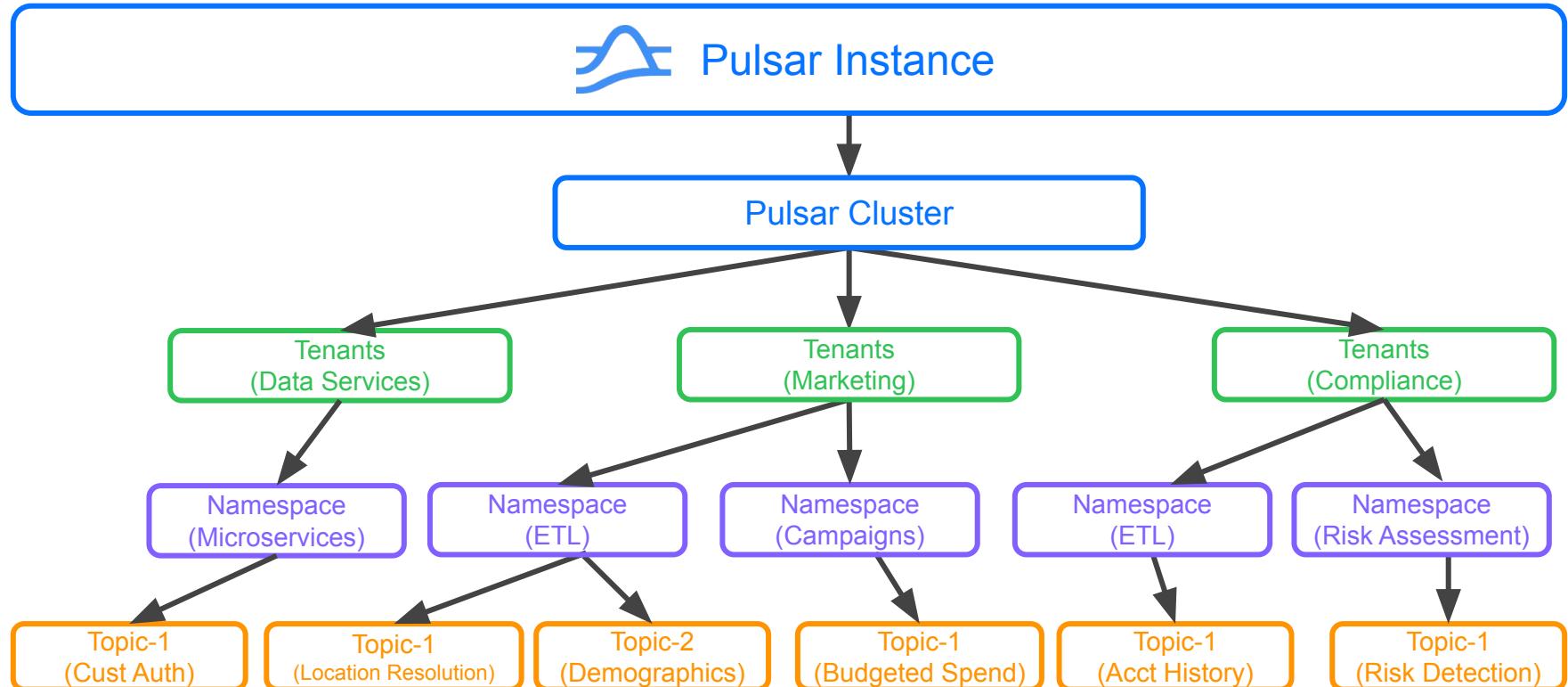
- **Functions** - Lightweight Stream Processing (Java, Python, Go)
- **Connectors** - Sources & Sinks (Cassandra, Kafka, ...)
- **Protocol Handlers** - AoP (AMQP), KoP (Kafka), MoP (MQTT)
- **Processing Engines** - Flink, Spark, Presto/Trino via Pulsar SQL
- **Data Offloaders** - Tiered Storage - (S3)

# 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



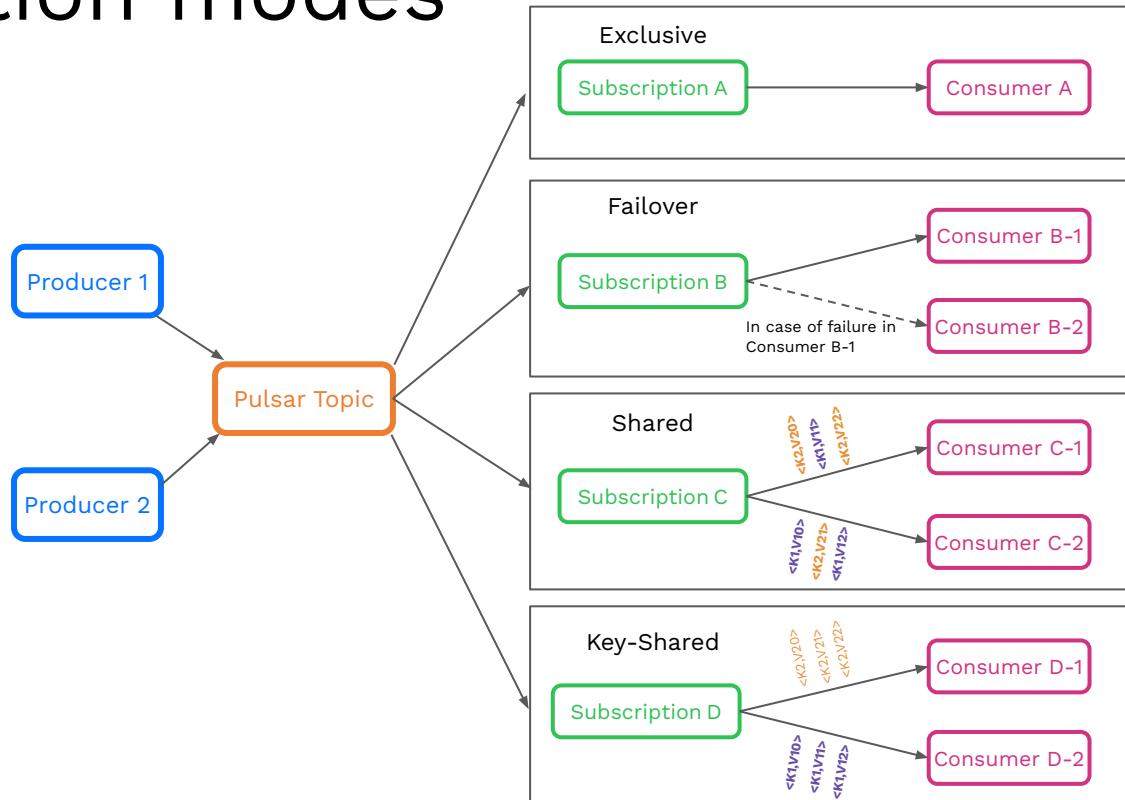
# Pulsar subscription modes

Different subscription modes have different semantics:

Exclusive/Failover - guaranteed order, single active consumer

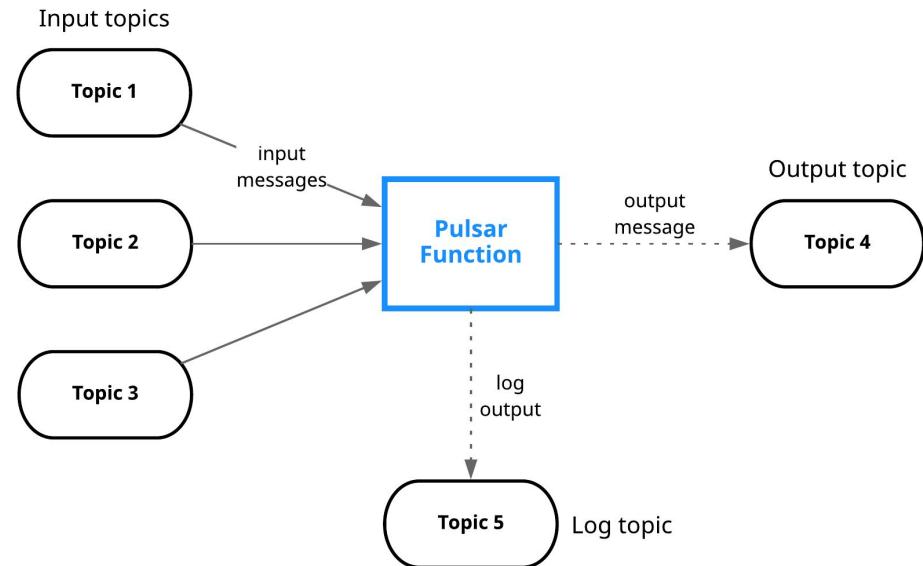
Shared - multiple active consumers, no order

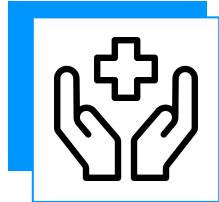
Key\_Shared - multiple active consumers, order for given key



# What are Pulsar Functions?

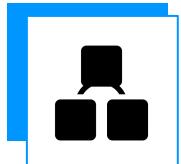
- Lambda-style functions that use Pulsar as the message bus.
- Handles producer/consumer setup
- Applies user supplied business logic against consumed message.





# Benefits of Pulsar Functions

- Allow you to focus on the business logic.
- Eliminates boilerplate code.
- Handles message consumption and publication
- No need for another processing framework.
- Can be scaled up independently



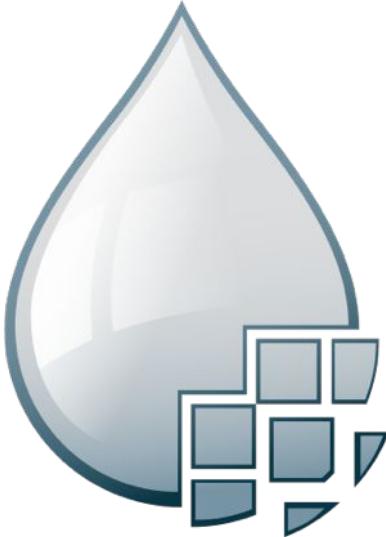
# Use cases

- Unified Messaging Platform
- AdTech
- Fraud Detection
- Connected Car
- IoT Analytics
- Microservices Development

# Apache NiFi

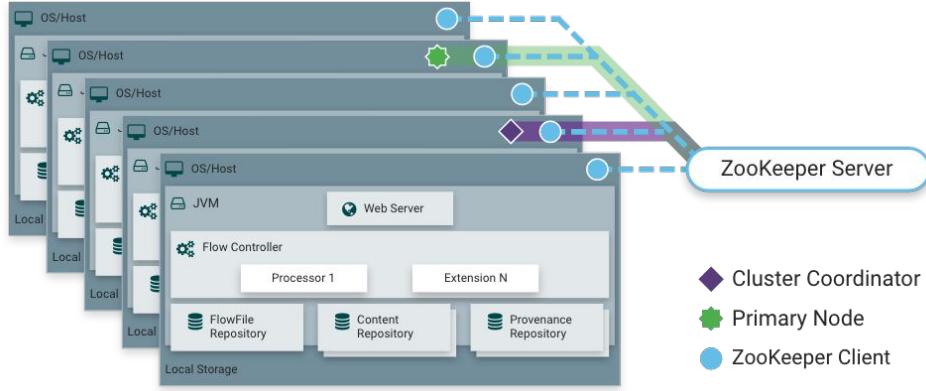
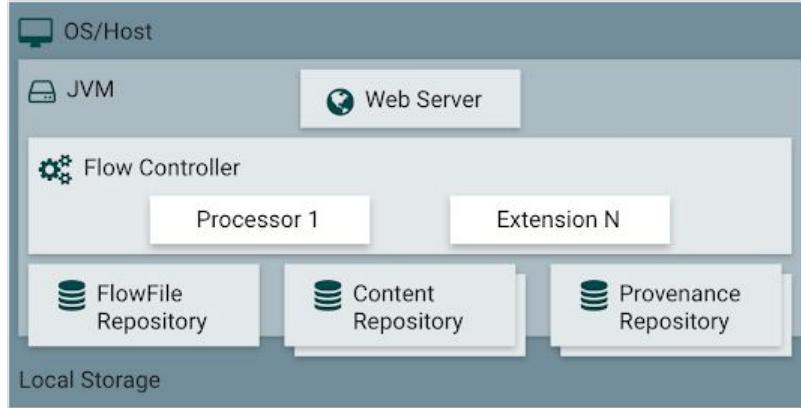


# Why Apache NiFi?



- 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



- ◆ Cluster Coordinator
- ◆ Primary Node
- ZooKeeper Client

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

# Provenance

## NiFi Data Provenance

Displaying 165 of 165  
Oldest event available: 12/21/2020 16:55:33 UTC

Filter by component name ▾

Date/Time	Type	Flowfile Uuid	Size
12/22/2020 16:54:17.193 UTC	ATTRIBUTES_MODIFIED	6fbbae84f6ba4-47c3-ba03-8830ec7cd3db	89 byte
12/22/2020 16:54:17.192 UTC	ATTRIBUTES_MODIFIED	1233e8d4e84d-4218-b3d0-2598e7f901e	87 byte
12/22/2020 16:54:14.194 UTC	ATTRIBUTES_MODIFIED	37fbca2153-4185-4460-b633-7eb074ad5718	81 byte
12/22/2020 16:54:05.297 UTC	ATTRIBUTES_MODIFIED	699d6fc9-97f1-4cf6-b733-b5c46e05622	83 byte
12/22/2020 16:53:59.296 UTC	ATTRIBUTES_MODIFIED	d43c05c5-5aae-44c2-9edc-c20a8148604	84 byte
12/22/2020 16:53:59.295 UTC	ATTRIBUTES_MODIFIED	4b1dbb1c-1f83-4a93-b309-2da1277cd7c6	84 byte
12/22/2020 16:53:58.296 UTC	ATTRIBUTES_MODIFIED	45fe8edd-cx55-431e-82b9-436c4a4092e	81 byte
12/22/2020 16:53:57.298 UTC	ATTRIBUTES_MODIFIED	b07034b-6361-4c34-b	
12/22/2020 16:53:57.297 UTC	ATTRIBUTES_MODIFIED	d12601a-c793-4c16-b	
12/22/2020 16:53:57.297 UTC	ATTRIBUTES_MODIFIED	29966d0-4153-41bc-a	
12/22/2020 16:53:43.753 UTC	ATTRIBUTES_MODIFIED	1ca5c744-1cb4-4ff1-bb	
12/22/2020 16:53:37.747 UTC	ATTRIBUTES_MODIFIED	faf647db-9e65-48c0-a	
12/22/2020 16:53:21.646 UTC	ATTRIBUTES_MODIFIED	df1f60ff-6d65-460e-99	
12/22/2020 16:53:05.515 UTC	ATTRIBUTES_MODIFIED	964693fc-d953-440c-b	
12/22/2020 16:52:43.374 UTC	ATTRIBUTES_MODIFIED	79fcfa90-b160-4fc4-8a	
12/22/2020 16:52:29.308 UTC	ATTRIBUTES_MODIFIED	3453eeb3-953c-4952-a	
12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	a166e2a7-118a-4262-9	
12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	bd2946fd-5a93-42d7-b	
12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	a16841bc-2505-4c8c-b	
12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	578540f6-e449-471f-a	
12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	3d44c5fb-a737-4a9e-82	
12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	4dc93a17-7059-424e-9	
12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	9fb9dc1-f304-4c11-93	

## Provenance Event

DETAILS ATTRIBUTES

### Attribute Values

lastprice	123.66
No value set	
symbol	
IBM	No value set
timestamp	1608654962884
No value set	
volume	100
No value set	

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

# Backpressure & Prioritizers

Configure Connection

DETAILS SETTINGS

Name

Id 3ca22430-cba4-3347-b45b-7bdc3530bd7e

FlowFile Expiration  0 sec

Back Pressure Object Threshold  10000

Size Threshold  1 GB

Available Prioritizers  FirstInFirstOutPrioritizer  
 NewestFlowFileFirstPrioritizer  
 OldestFlowFileFirstPrioritizer  
 PriorityAttributePrioritizer

Selected Prioritizers

Load Balance Strategy  Do not load balance

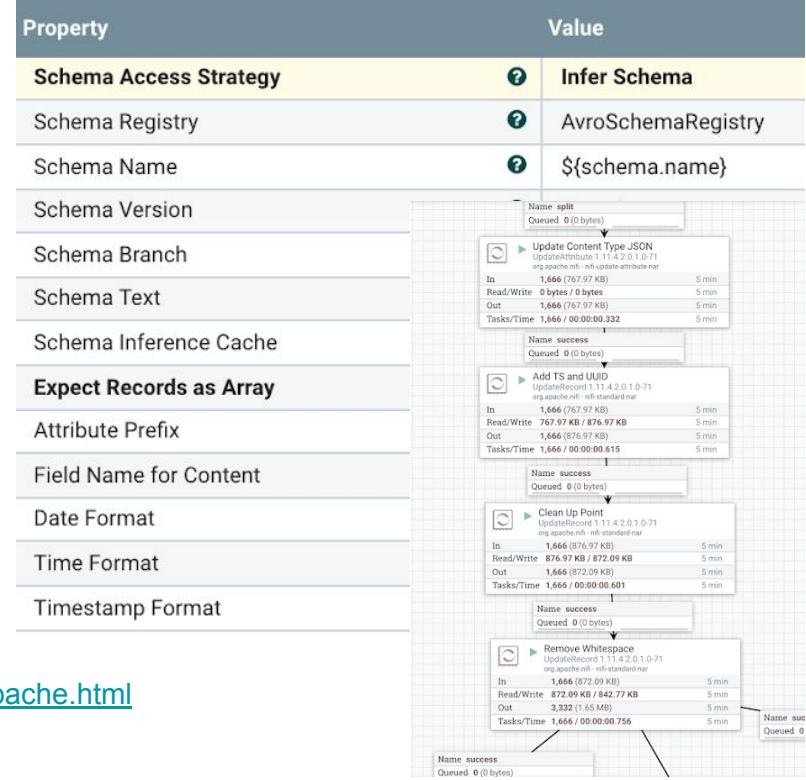
<https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html>

# 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

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



# Record Processors



Configure Processor

⚠ Invalid

SETTINGS    SCHEMAS

Required field

Add Controller Service

Property

Record Reader

Record Destination S...

Include Zero Record I...

Requires Controller Service  
RecordReaderFactory 1.13.0 from org.apache.nifi - nifi-standard-services-api-nar

Compatible Controller Services

- AvroReader 1.13.0
- CSVReader 1.13.0
- GrokReader 1.13.0
- JsonPathReader 1.13.0
- JsonTreeReader 1.13.0
- ParquetReader 1.13.0
- ReaderLookup 1.13.0
- ScriptedReader 1.13.0
- Syslog5424Reader 1.13.0
- SyslogReader 1.13.0
- WindowsEventLogReader 1.13.0
- XMLReader 1.13.0

Property

Record Reader

Record Destination S...

Include Zero Record I...

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>

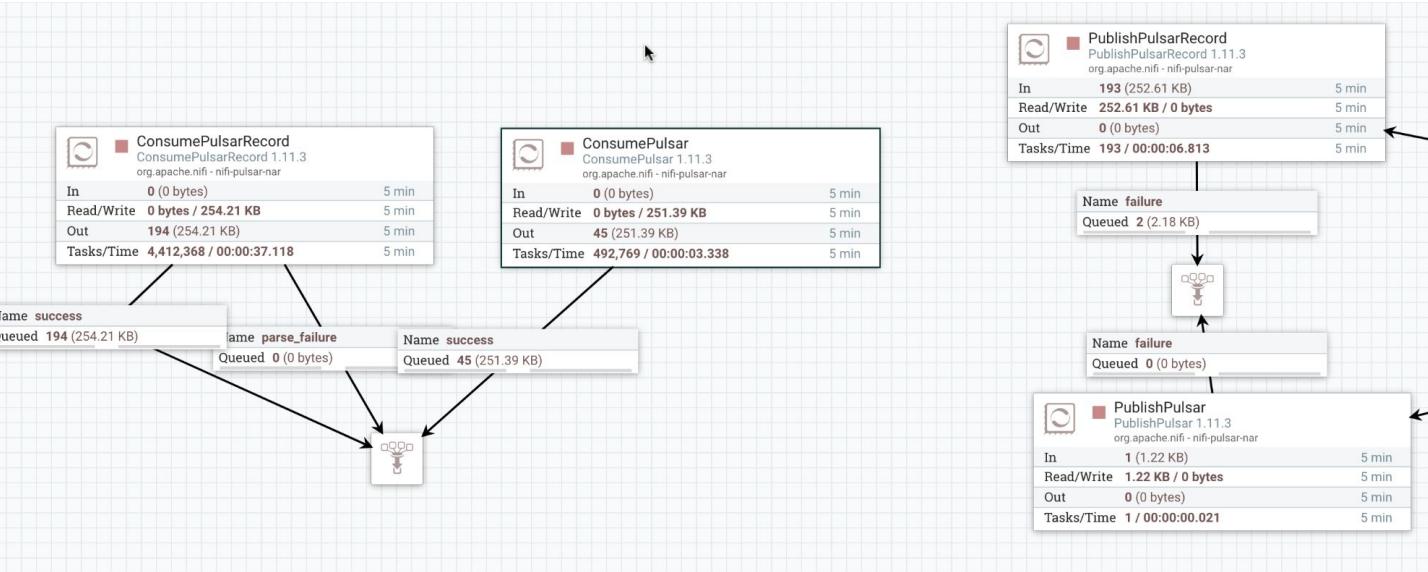
# Consume MQTT



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

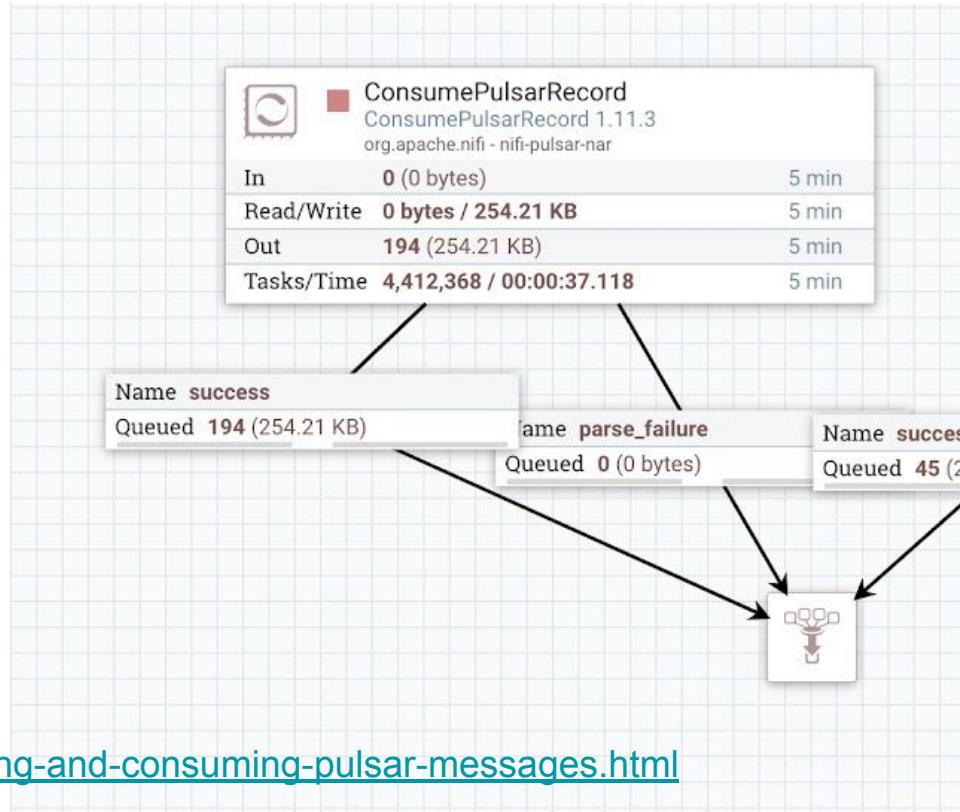
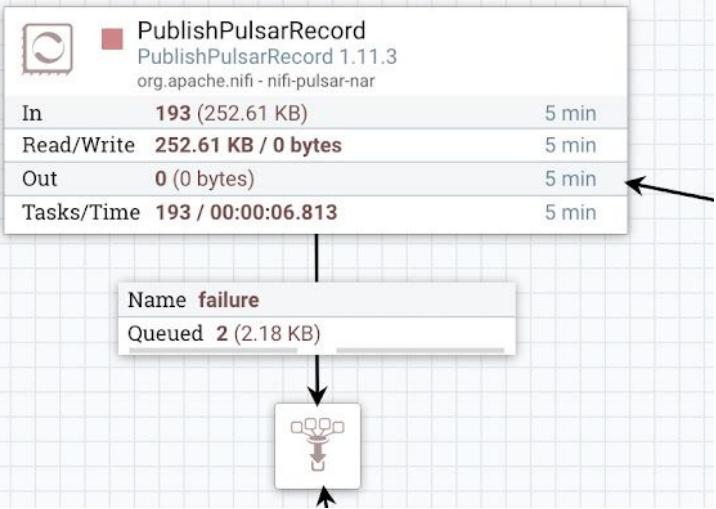
Property	ConsumeMQTT Processor	Value
Session state	?	Clean Session
MQTT Specification Version	?	AUTO
Connection Timeout (seconds)	?	30
Keep Alive Interval (seconds)	?	60
Group ID	?	No value set
Topic Filter	?	No value set
Quality of Service(QoS)	?	0 - At most once
Max Queue Size	?	No value set
Record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

# Apache NiFi Pulsar Connector



<https://github.com/david-streamlio/pulsar-nifi-bundle>

# Apache NiFi Pulsar Connector



<https://www.datainmotion.dev/2021/11/producing-and-consuming-pulsar-messages.html>

# Apache NiFi Pulsar Connector

Displaying 12 of 339

pulsar

Type	Version ▲	Tags
ConsumePulsar	1.11.0	PubSub, Consume, ingest, Get, I...
ConsumePulsarRecord	1.11.0	PubSub, Consume, Ingest, Get, ...
PublishPulsar	1.11.0	PubSub, Message, Pulsar, Apac...
PublishPulsarRecord	1.11.0	PubSub, 1.0, Message, csv, json...
ConsumePulsar	1.11.3	PubSub, Consume, Ingest, Get, I...
ConsumePulsarRecord	1.11.3	PubSub, Consume, Ingest, Get, ...
PublishPulsar	1.11.3	PubSub, Message, Pulsar, Apac...
PublishPulsarRecord	1.11.3	PubSub, 1.0, Message, csv, json...
ConsumePulsar	1.14.0	PubSub, Consume, Ingest, Get, I...
ConsumePulsarRecord	1.14.0	PubSub, Consume, Ingest, Get, ...
PublishPulsar	1.14.0	PubSub, Message, Pulsar, Apac...
PublishPulsarRecord	1.14.0	PubSub, 1.0, Message, csv, json...

**ConsumePulsar 1.11.3** org.apache.nifi - nifi-pulsar-nar

Consumes messages from Apache Pulsar. The complementary NiFi processor for sending messages is PublishPulsar.

# Apache NiFi Pulsar Connector

Controller Service Details

SETTINGS PROPERTIES COMMENTS

Required field

Property	Value
Pulsar Service URL	pulsar+ssl://gke.sndev.snio.cloud:6651
Pulsar Client Authentication Service	PulsarClientOauthAuthenticationService14sn →
Maximum concurrent lookup-requests	5000
Maximum connects per Pulsar broker	1
I/O Threads	1
Keep Alive interval	30 sec
Listener Threads	1
Maximum lookup requests	50000 
Maximum rejected requests per connection	50
Operation Timeout	30 sec
Stats interval	60 sec
Allow TLS Insecure Connection	false
Enable TLS Hostname Verification	false
Use TCP no-delay flag	false

# Apache NiFi Pulsar Connector

## Controller Service Details

SETTINGS	PROPERTIES	COMMENTS
----------	------------	----------

Required field

Property	Value
Audience	urn:sn:pulsar:sndev:gke
Issuer URL	<a href="https://auth.streamnative.cloud">https://auth.streamnative.cloud</a>
Private key file	file:///Users/tspann/Documents/servers/services/apache-pulsar-2.8.0/sndev-tspann.json
Trusted Certificate Filename	?

<https://github.com/david-streamlio/pulsar-nifi-bundle/releases/tag/v1.14.0>

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



**kubernetes**

Built for Containers



**Flink**

Flink SQL



## StreamNative Ambassador Program 2022

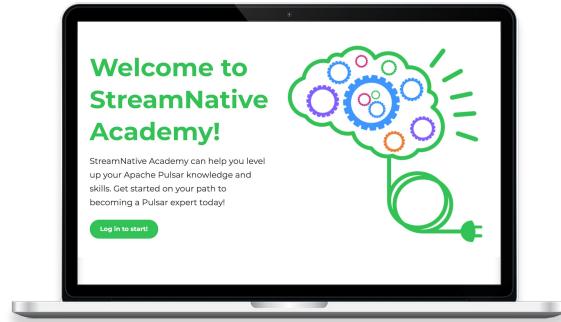
[Learn More](#)



**Take the 2022  
Apache Pulsar™  
User Survey!**

Tell us about your Pulsar experience  
and what improvements you would  
like to see!

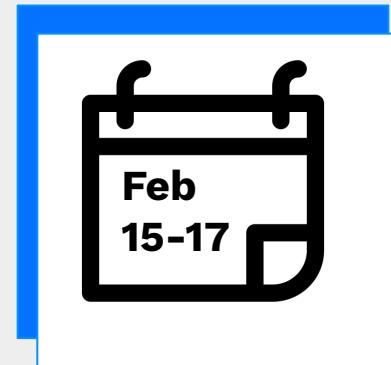
[Start Survey](#)



**Now Available**

# On-Demand Pulsar Training

[Academy.StreamNative.io](https://Academy.StreamNative.io)



## Live 3-day Developers Training

[Save Your Spot!](#)

### Times:

- Europe: 3:00 PM CET - 7:00 PM CET
- Eastern Time: 9:00 AM - 1:00 PM EST
- Pacific Time: 6:00 AM - 10 AM PST

# Let's Keep in Touch!



**Tim Spann**

Developer Advocate



@[PaaSDev](https://twitter.com/PaaSDev)

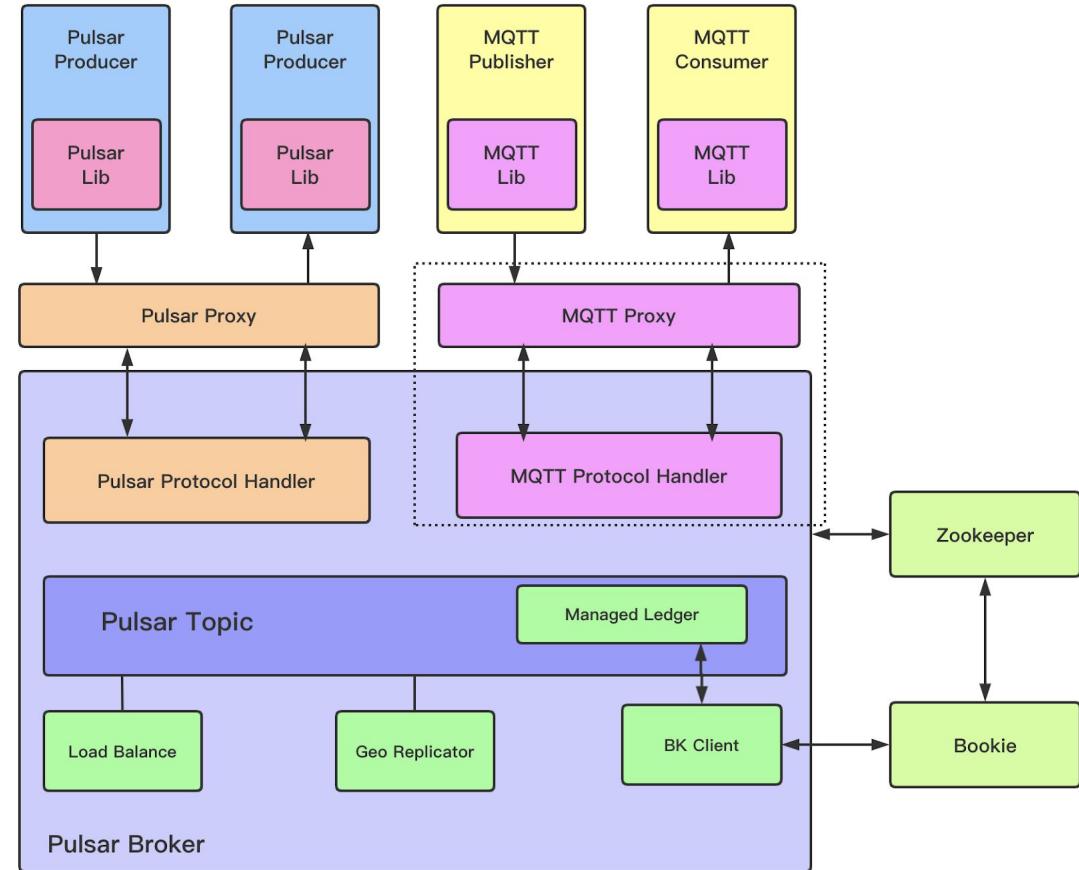


<https://www.linkedin.com/in/timothyspann>



<https://github.com/tspannhw>

# MQTT on Pulsar (MoP)



# Kafka-on-Pulsar (Kop)

