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/

StreamNative

FLiP Stack Weekly

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

https://bit.ly/32dAJft

StreamNative

Stream Native

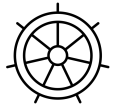
- Founded the original developers of Apache Pulsar.
- Passionate and dedicated team.
- StreamNative helps teams to capture, manage, and leverage data using Pulsar's unified messaging and streaming platform.

streamnative.io

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

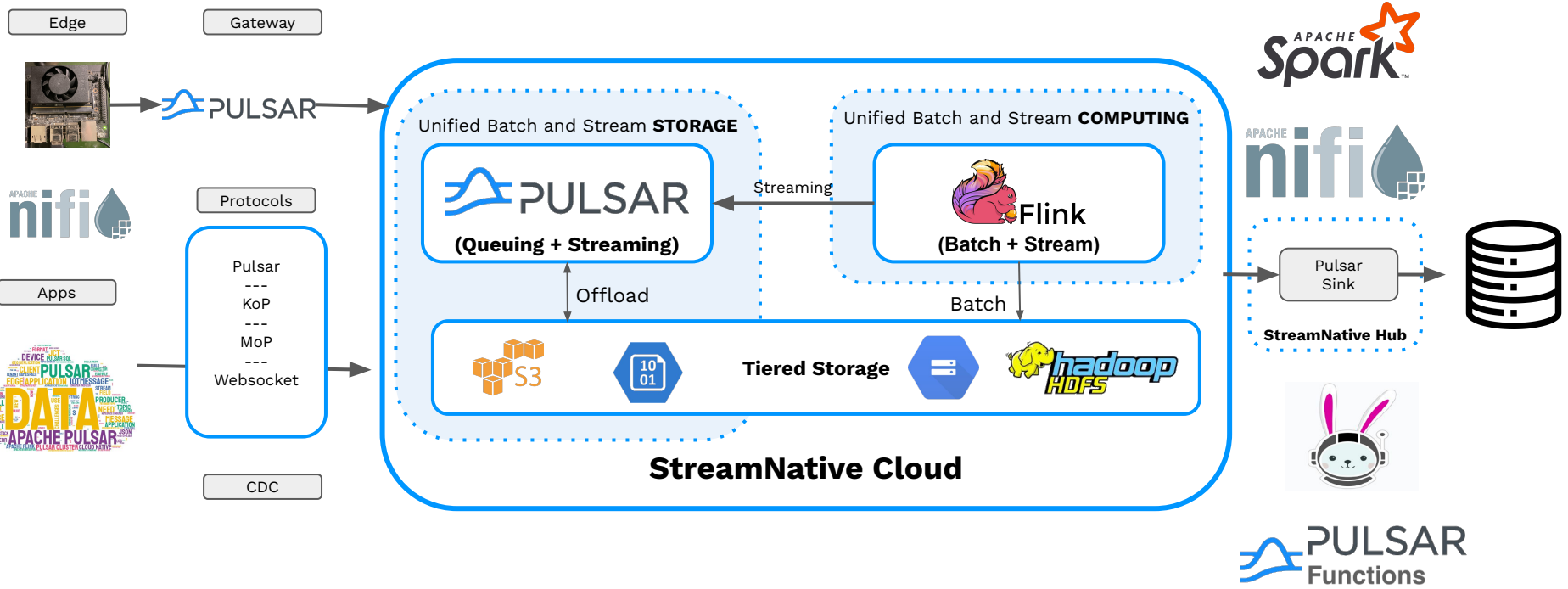


Built for Kubernetes CloudNative migrations with tools

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

Streaming FLiPN Apps

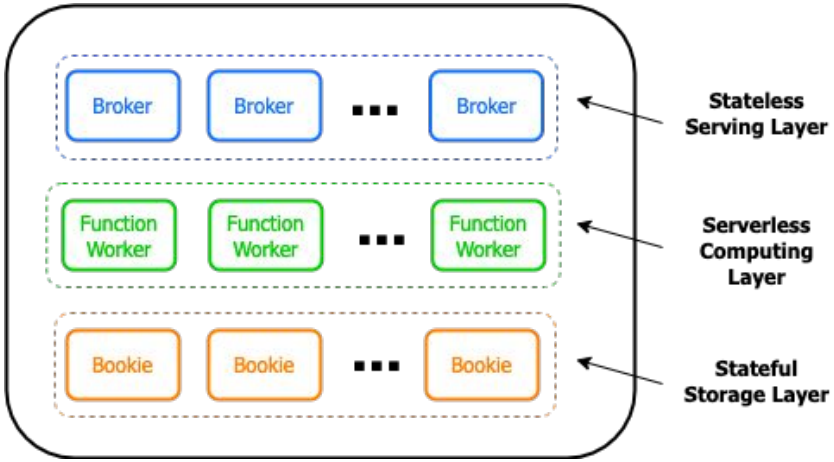
Events <-> PULSAR **<-> Events <->** Flink



Apache Pulsar

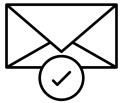
Apache Pulsar

- 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

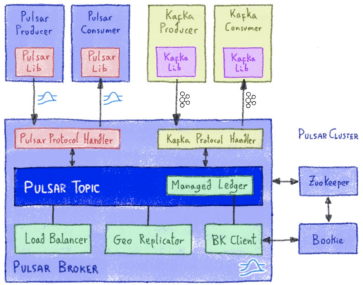
Infinite Scalability

Messages - the basic unit of Pulsar

Component	Description
Value / data payload	The data carried by the message. All Pulsar messages contain raw bytes, although message data can also conform to data schemas.
Key	Messages are optionally tagged with keys, used in partitioning and also is useful for things like topic compaction.
Properties	An optional key/value map of user-defined properties.
Producer name	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.
Sequence ID	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

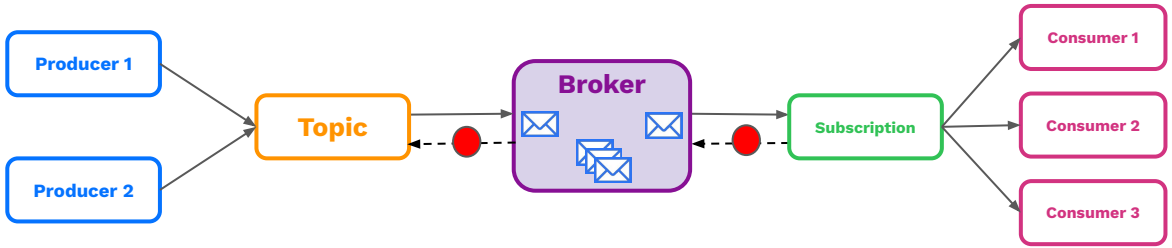
hub.streamnative.io



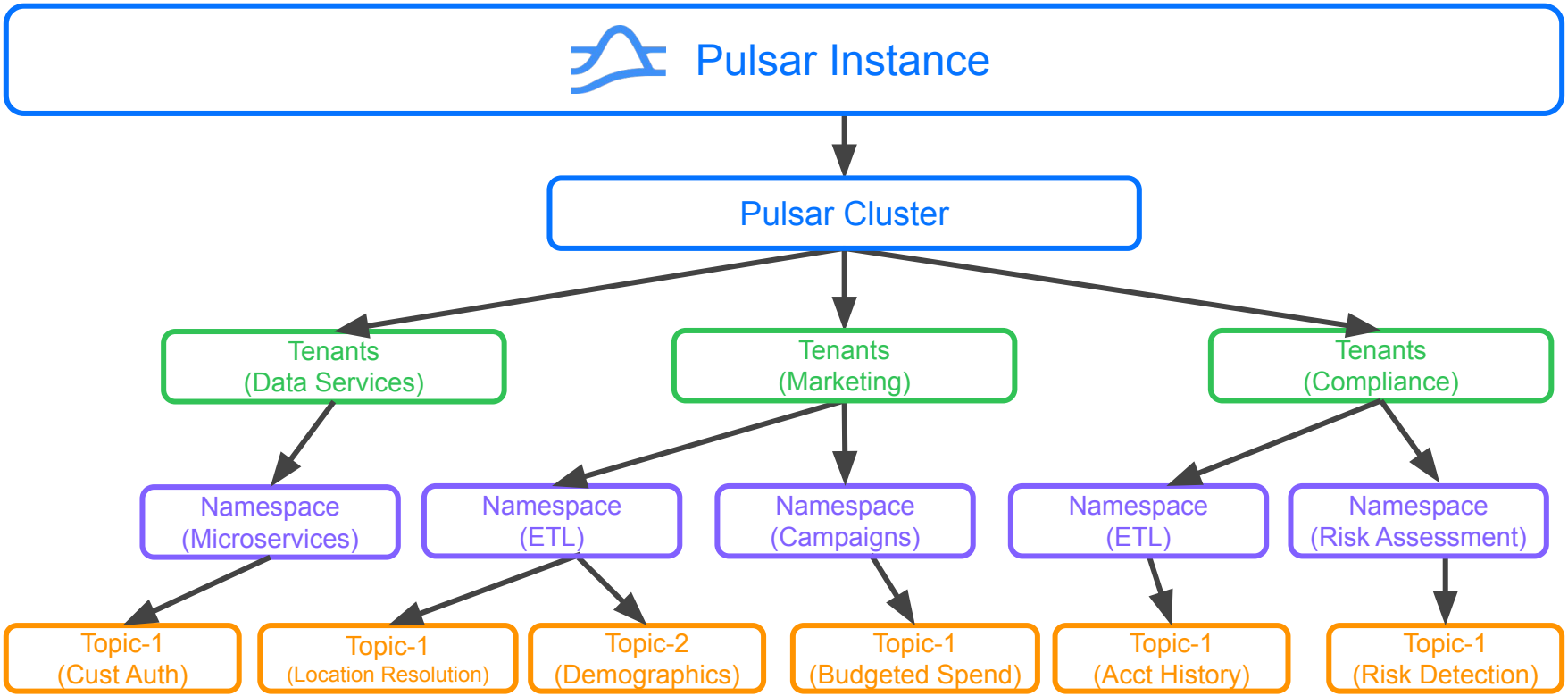
- **Functions** - Lightweight Stream Processing (Java, Python, Go)
- **Connectors** - Sources & Sinks (Cassandra, Kafka, …)
- **Protocol Handlers** - AoP (AMQP), KoP (Kafka), MoP (MQTT)
- **Processing Engines** - Flink, Spark, Presto/Trino via Pulsar SQL
- **Data Offloaders** - Tiered Storage - (S3)

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 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)

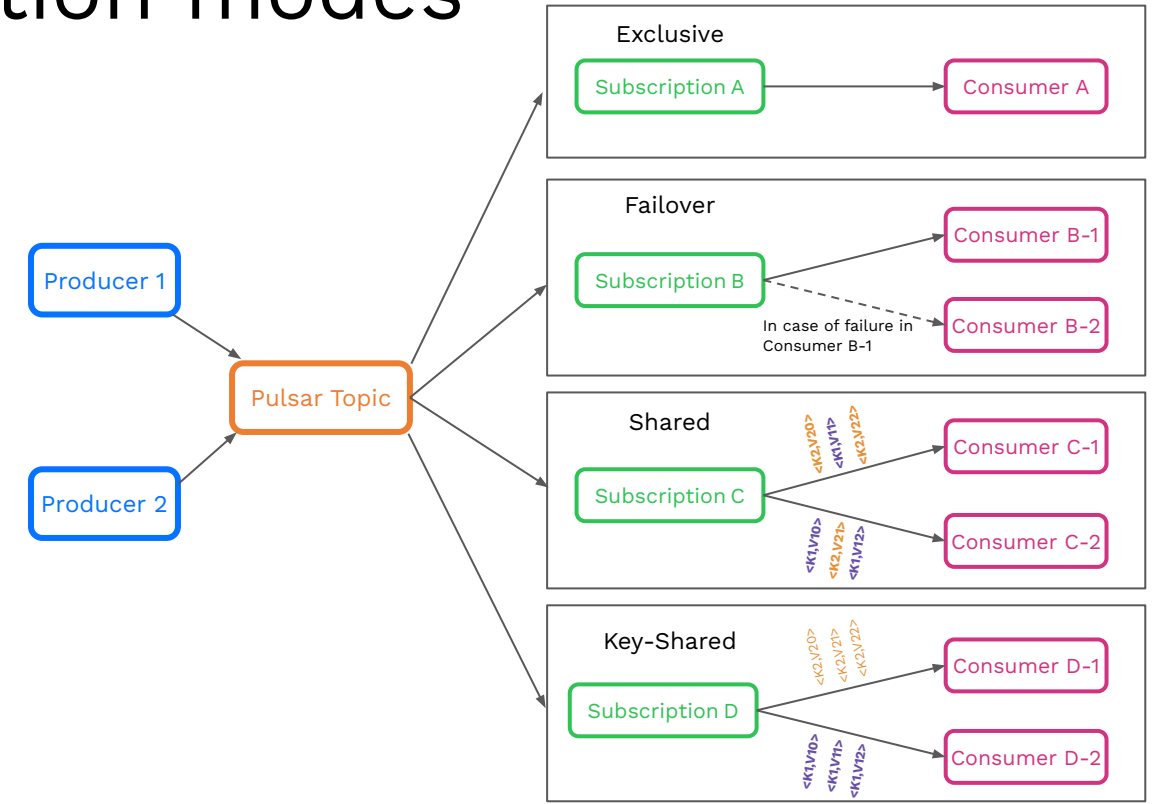
Pulsar subscription modes

Different subscription modes have different semantics:

Exclusive/Failover - guaranteed order, single active consumer

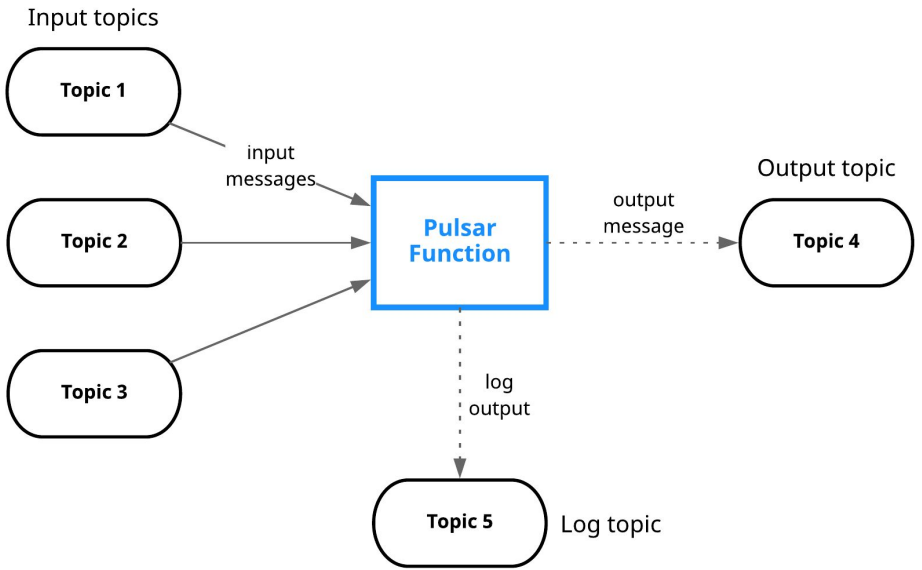
Shared - multiple active consumers, no order

Key_Shared - multiple active consumers, order for given key



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

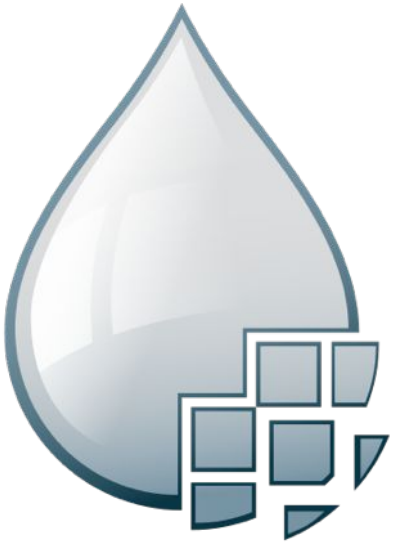
- Allow you to focus on the business logic.
- Eliminates boilerplate code.
- Handles message consumption and publication
- No need for another processing framework.
- Can be scaled up independently

Use cases

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

Apache NiFi

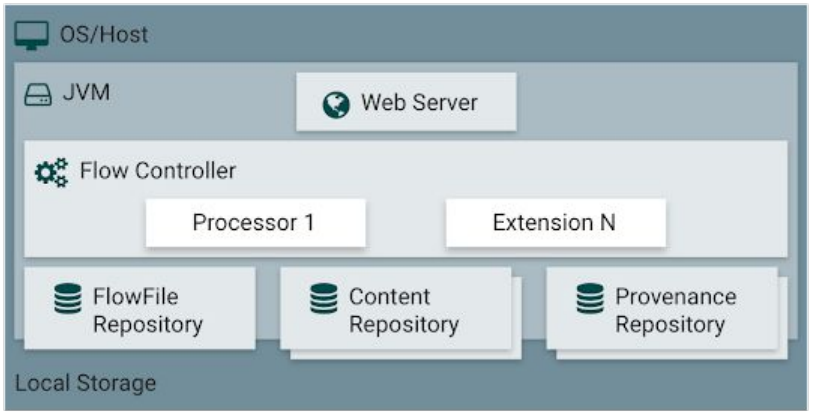
Why Apache NiFi?

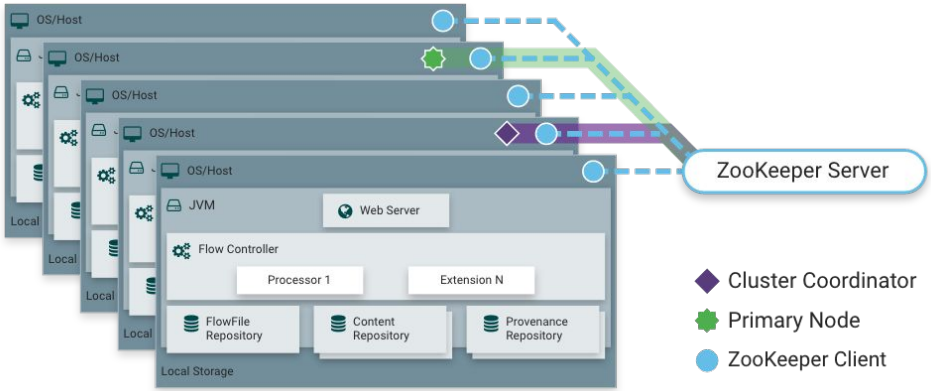


- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models

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

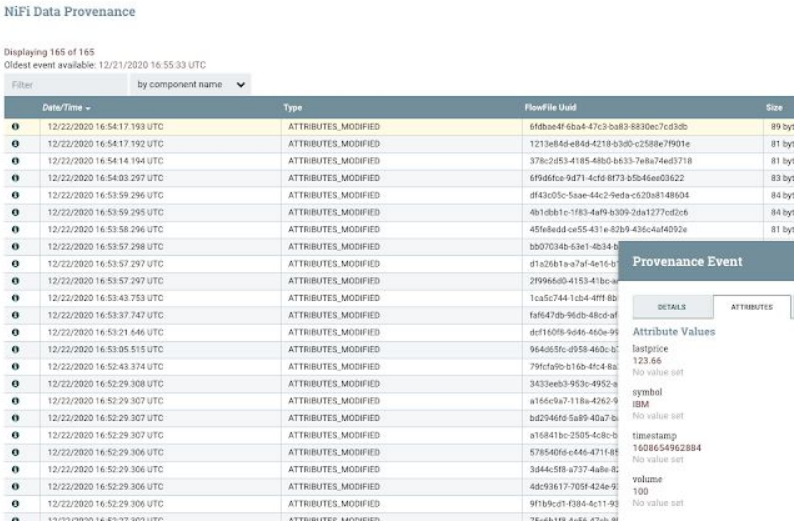
Architecture





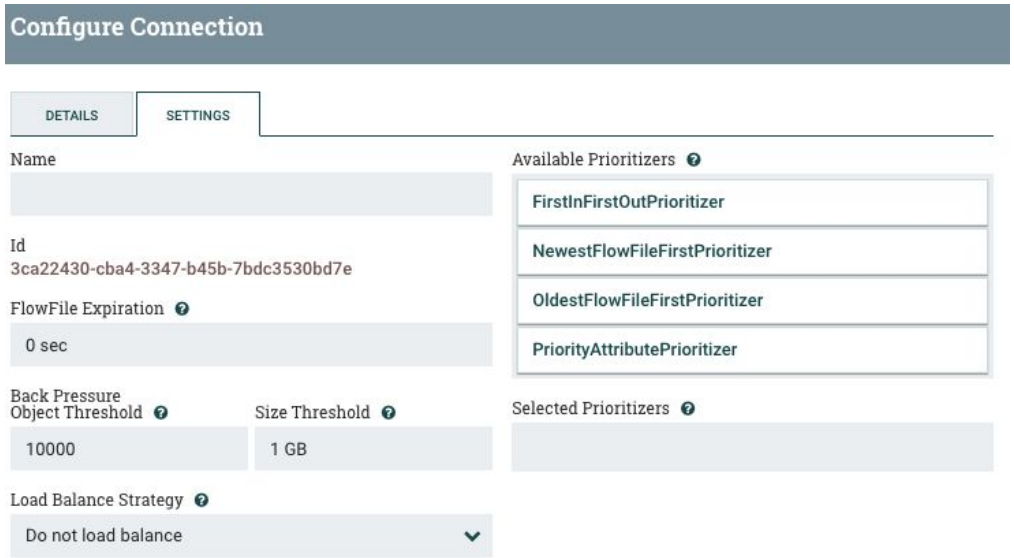
https://nifi.apache.org/docs/nifi-docs/html/overview.html

Provenance



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

Backpressure & Prioritizers



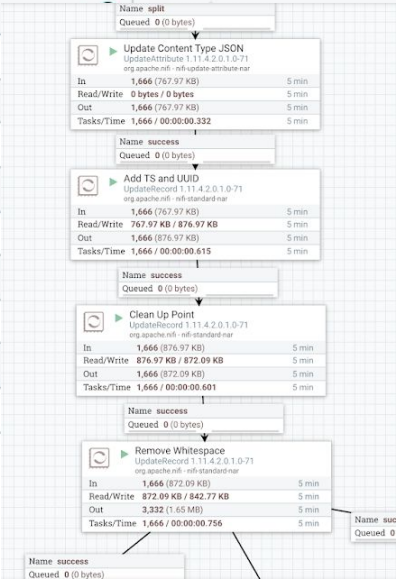
https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Record Processors

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

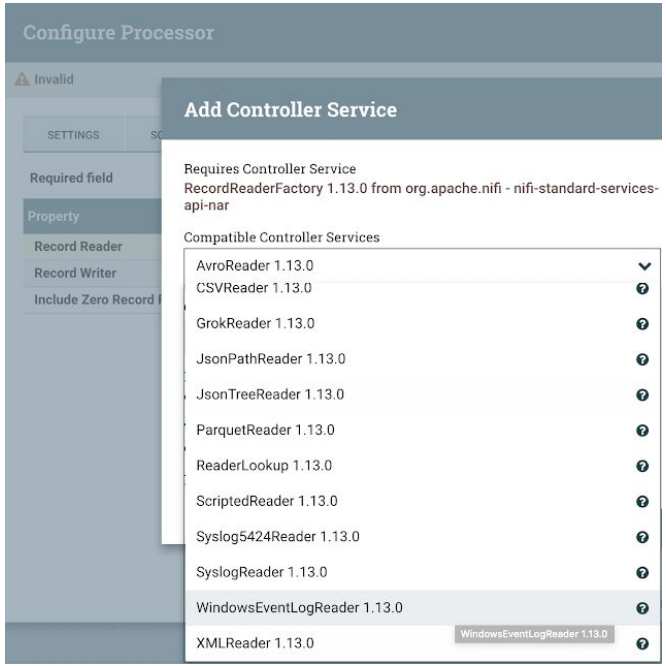
Property	Value
Record Reader	**XMLReader**
Record Writer	**JsonRecordSetWriter**
Include Zero Record FlowFiles	**false**
Cache Schema	**true**
query1	SELECT * FROM FLOWFILE

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

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

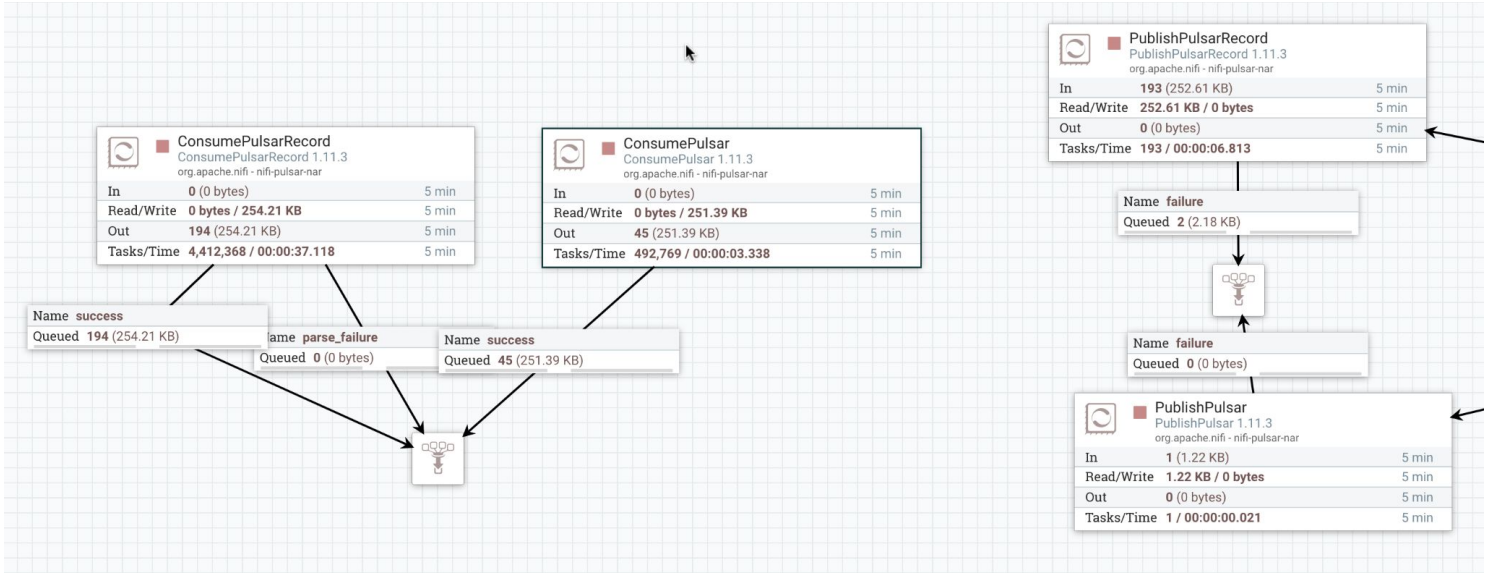
Consume MQTT

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

ConsumeMQTT Processor

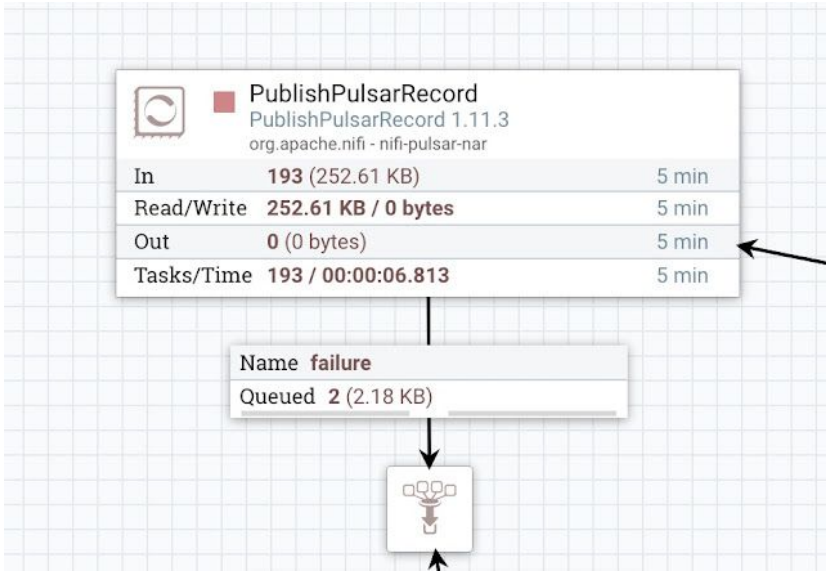
Property	Value
Session state	**Clean Session**
MQTT Specification Version	**AUTO**
Connection Timeout (seconds)	30
Keep Alive Interval (seconds)	60
Group ID	No value set
Topic Filter	No value set
Quality of Service(QoS)	**0 - At most once**
Max Queue Size	No value set
Record Reader	No value set
Record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

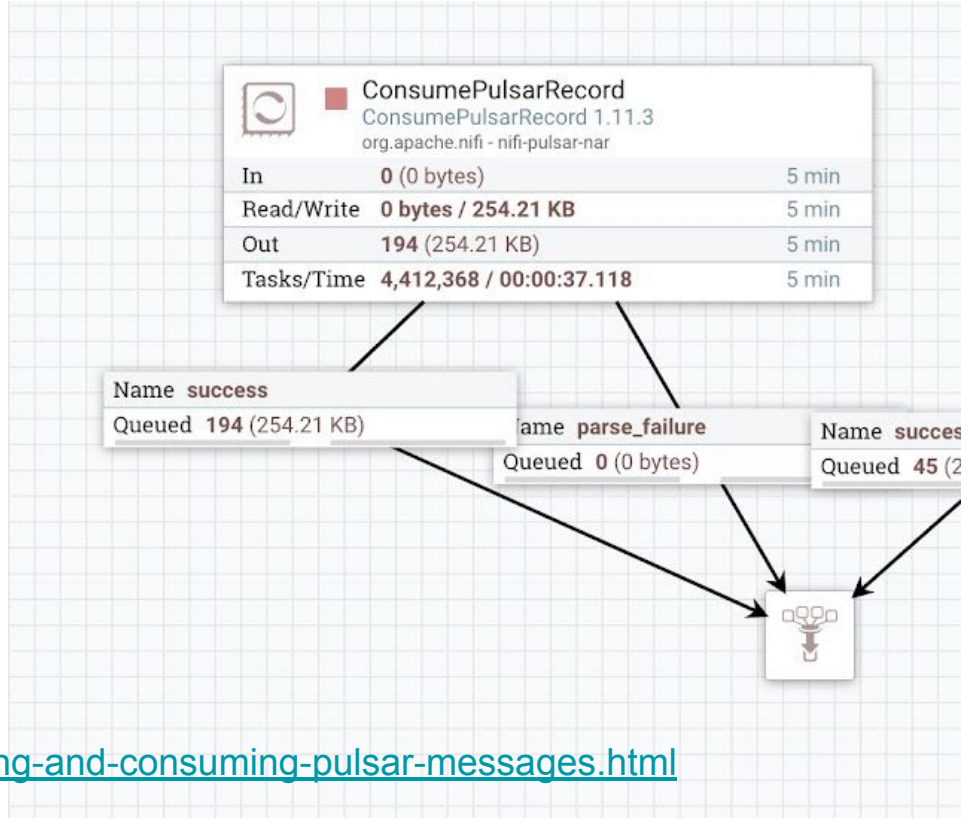
Apache NiFi Pulsar Connector



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

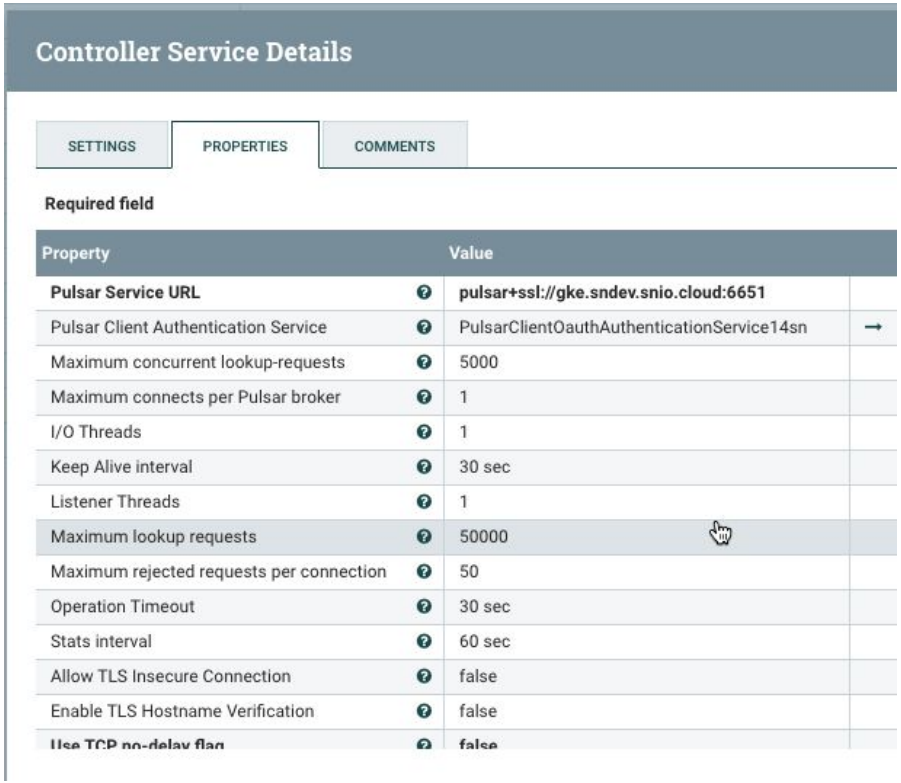
pulsar

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



Apache NiFi Pulsar Connector

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

StreamNative
Cloud

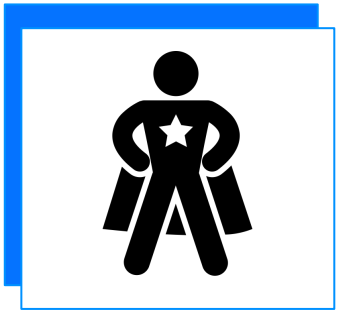
StreamNative Cloud

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

Built for Containers

Flink SQL



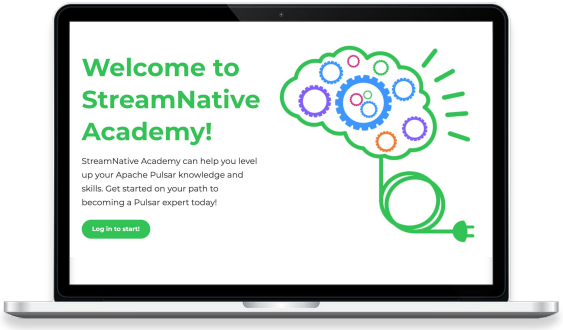
StreamNative Ambassador Program 2022

Learn More

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

Start Survey

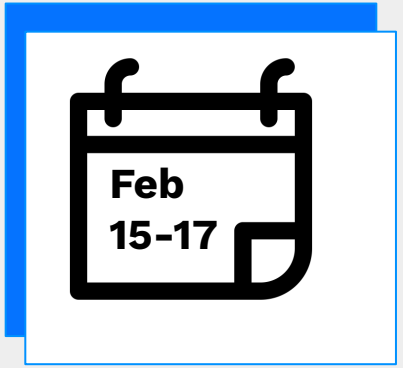
StreamNative



Now Available

On-Demand Pulsar Training

Academy.StreamNative.io



Live 3-day

Developers Training

Save Your Spot!

Times:

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

Let’s Keep in Touch!

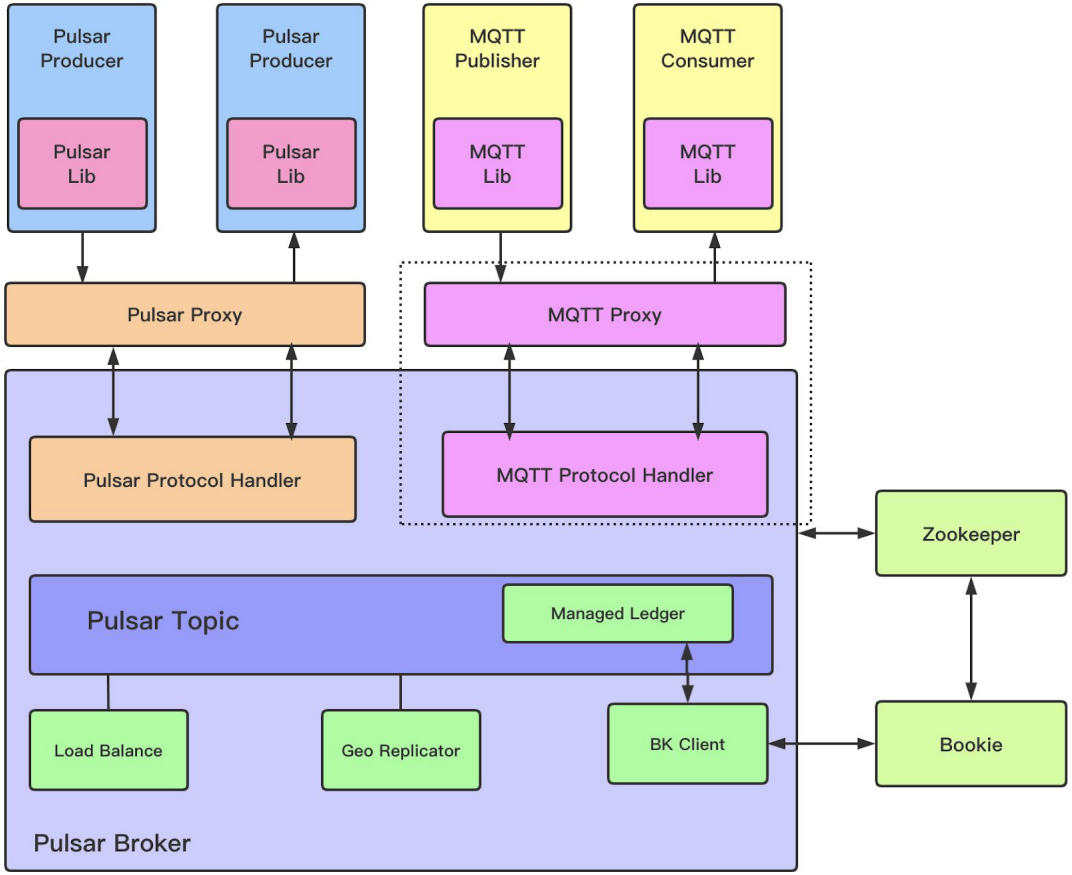
Tim Spann
Developer Advocate

@PaaSDev

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

https://github.com/tspannhw

MQTT on Pulsar (MoP)



Kafka-on-Pulsar (Kop)