



Apache NiFi 101: Introduction and Best Practices

Tim Spann | Developer Advocate



Question Everything!

Session

Regular 50 minute session

Apache NiFi 101: Introduction
and Best Practices

Primary Speaker

Fri 12:00



Feedback Link

<https://sqlb.it/?7108>

FLiP(N) Stack = Flink, Pulsar and NiFi Stack

Streaming Systems & Data Architecture Expert

Experience:

15+ years of experience with streaming technologies including Pulsar, Flink, Spark, NiFi, Kafka, Big Data, Cloud, ML, IoT and more.

Today, he helps to grow the Pulsar community sharing rich technical knowledge and experience at both global conferences and through individual conversations



Timothy Spann | Developer Advocate





Founded by 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



FLiP Stack Weekly

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

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



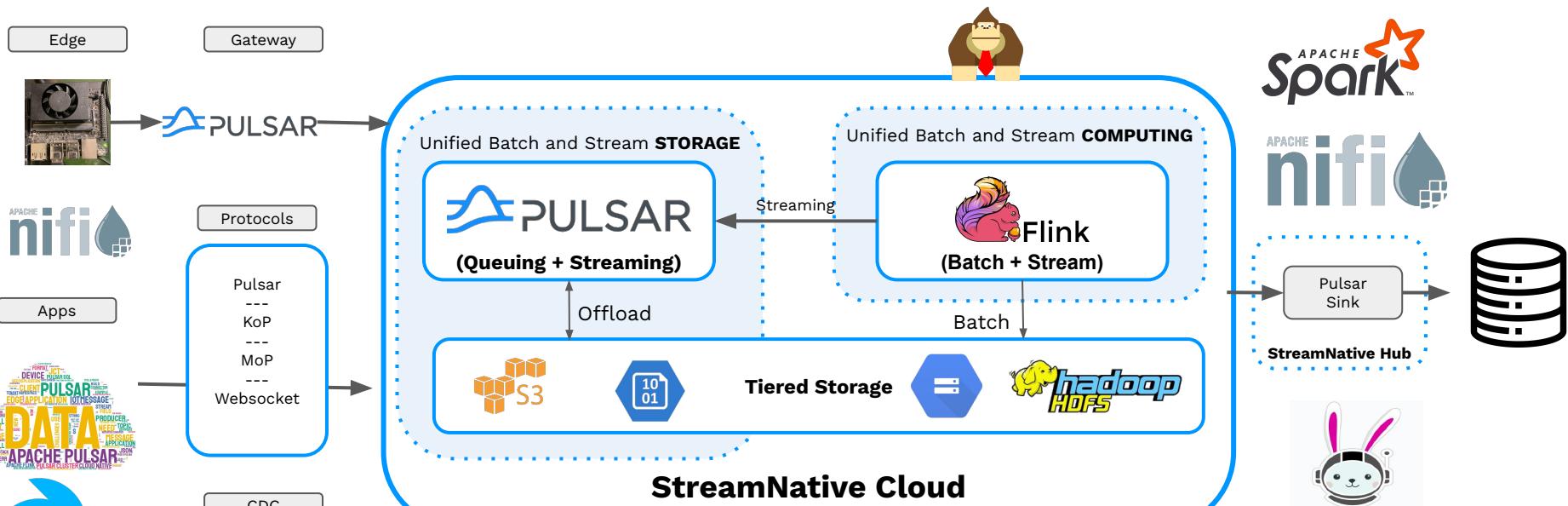
Streaming FLiPN Apps



<->



<-> Events <->



Apps



CDC

- Pulsar
-
- Kop
-
- MoP
-
- Websocket

Gateway

Protocols

Edge



Unified Batch and Stream **STORAGE**



(Queuing + Streaming)

Unified Batch and Stream **COMPUTING**



Flink
(Batch + Stream)

Streaming



Flink
(Batch + Stream)

Batch



Tiered Storage



hadoop
HDFS



StreamNative Hub

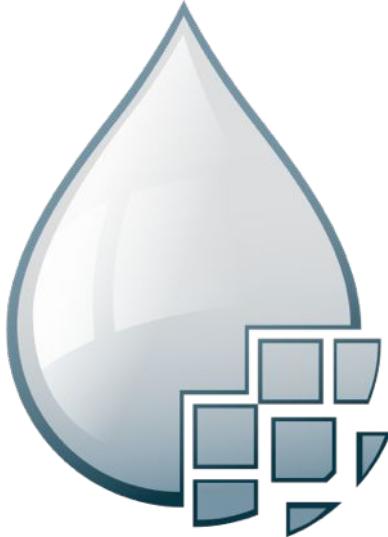


PULSAR
Functions

Apache NiFi

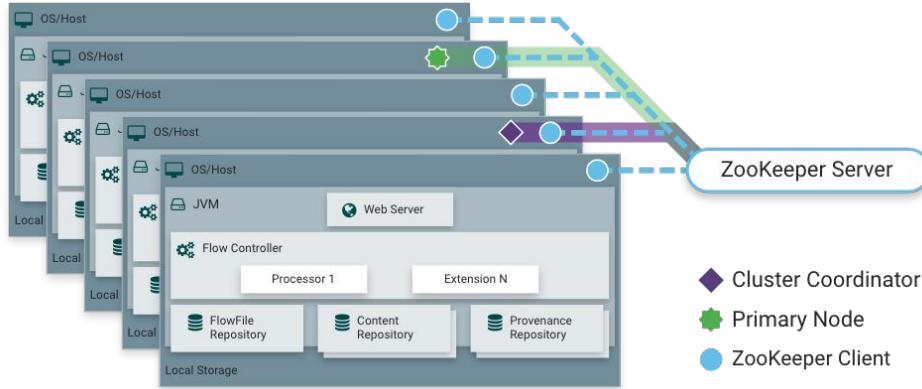
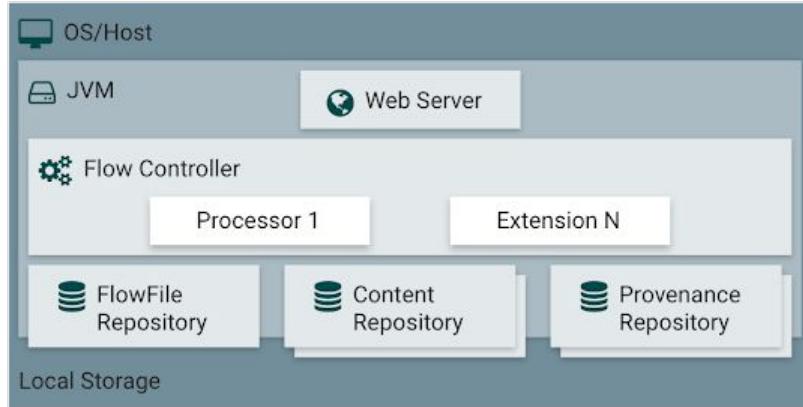


Why Apache NiFi?



- Guaranteed delivery
- Data buffering
 - Backpressure
 - Pressure release
- Prioritized queuing
- Flow specific QoS
 - Latency vs. throughput
 - Loss tolerance
- Data provenance
- Supports push and pull models
- 350+ processors
- Visual command and control
- Over a 100 sources
- Flow templates
- Pluggable/multi-role security
- Designed for extension
- Clustering
- Version Control

Architecture



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

Flow File

Flow Files are content and key/value pairs for attributes that are each event/message/file that has been introduced into NiFi.



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

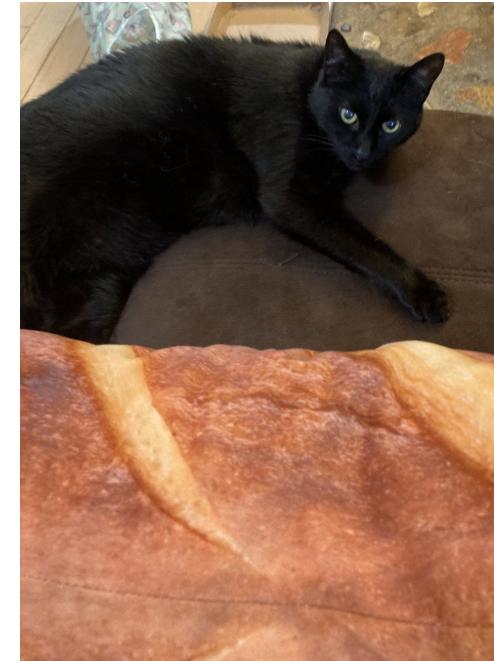


Processor

A Java component that runs in NiFi to route, process or manipulate data. You can build your own if it is not included in Standard NiFi or not in the open source.

Controller

Like a connection pools, connections, processes that ingest or work with outside data.



Connection

These link together NiFi processors.

Process Groups

Groups of processors. These are versionable and reusable components/modules.



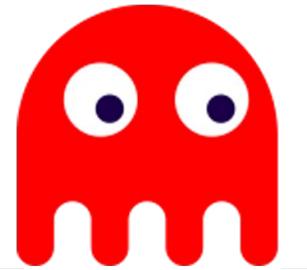
Provenance

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
1	12/22/2020 16:54:17.193 UTC	ATTRIBUTES_MODIFIED	f6fbac84f6ba4-47c3-ba03-8830ec7cd3db	89 byte
1	12/22/2020 16:54:17.192 UTC	ATTRIBUTES_MODIFIED	1233e8d4e84d-421b-b3d0-2598e7f901e	87 byte
1	12/22/2020 16:54:14.194 UTC	ATTRIBUTES_MODIFIED	37fb21b3-41b5-4bb0-b633-7eb074ad8718	81 byte
1	12/22/2020 16:54:03.297 UTC	ATTRIBUTES_MODIFIED	6f9d6fce-9d71-4cf6-b733-b5c4e0d3622	83 byte
1	12/22/2020 16:53:59.296 UTC	ATTRIBUTES_MODIFIED	d43305c5-5aae-44c2-9edc-c20a8148604	84 byte
1	12/22/2020 16:53:59.295 UTC	ATTRIBUTES_MODIFIED	4b1dbb1f-f8f3-4a9f-b309-2da1277cd7c6	84 byte
1	12/22/2020 16:53:58.296 UTC	ATTRIBUTES_MODIFIED	45fe8edd-cx55-411e-82b9-436c4a4092e	81 byte
1	12/22/2020 16:53:57.297 UTC	ATTRIBUTES_MODIFIED	b07034b-6361-4c34-b	
1	12/22/2020 16:53:57.297 UTC	ATTRIBUTES_MODIFIED	d12601a-7974-4c16-b	
1	12/22/2020 16:53:57.297 UTC	ATTRIBUTES_MODIFIED	29966d0-4153-41bc-a	
1	12/22/2020 16:53:43.753 UTC	ATTRIBUTES_MODIFIED	1ca5c744-1cb4-4fff-bb	
1	12/22/2020 16:53:37.747 UTC	ATTRIBUTES_MODIFIED	faf647db-9e65-48c0-a	
1	12/22/2020 16:53:21.646 UTC	ATTRIBUTES_MODIFIED	df1f60ff-6d65-460e-99	
1	12/22/2020 16:53:05.515 UTC	ATTRIBUTES_MODIFIED	964695fc-d953-440c-b	
1	12/22/2020 16:52:43.374 UTC	ATTRIBUTES_MODIFIED	79fcfa90-b160-4fc4-8a	
1	12/22/2020 16:52:29.308 UTC	ATTRIBUTES_MODIFIED	3433eeb9-953c-4952-a	
1	12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	a166e297-118a-4262-9	
1	12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	bd2946fd-5a99-40d7-b	
1	12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	a16841bc-2505-4c8c-b	
1	12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	5785406e-e449-471f-a	
1	12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	3d44c5fb-4737-4a9e-8	
1	12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	4dc93a17-7059-424e-9	
1	12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	9fb9d9c1-f304-4c11-93	



<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

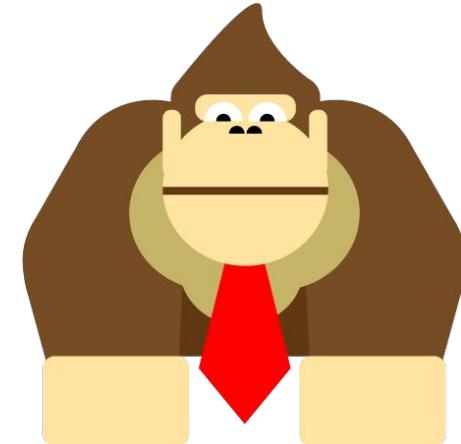
Available Prioritizers:

- FirstInFirstOutPrioritizer
- NewestFlowFileFirstPrioritizer
- OldestFlowFileFirstPrioritizer
- PriorityAttributePrioritizer

Selected Prioritizers:

Load Balance Strategy: ▾

Do not load balance



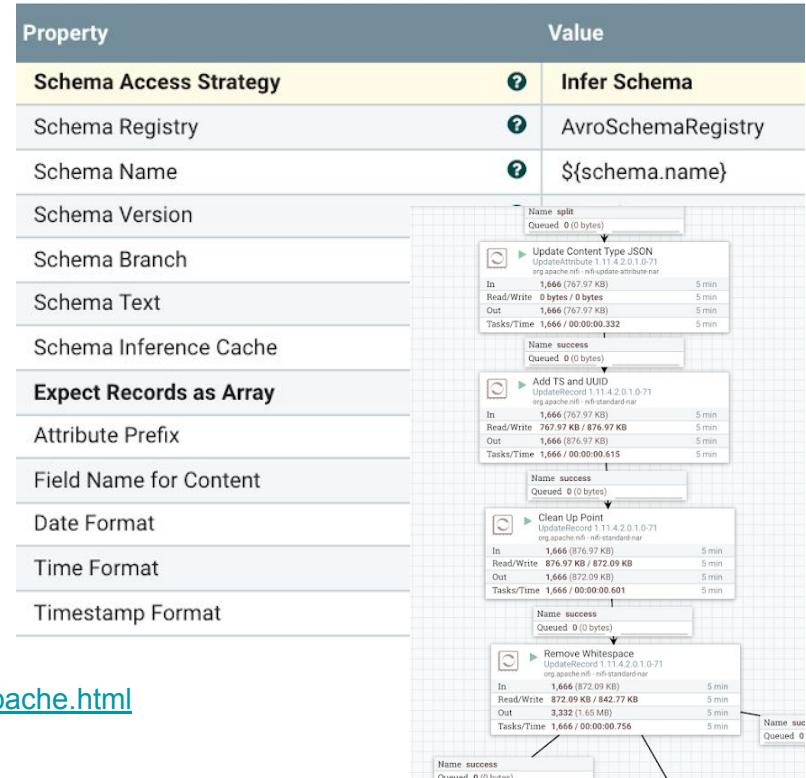
<https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html>

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

Property

Add Controller Service

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

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>

Caching

Property	Value
Record Reader	infer JsonTreeReader
Record Writer	Standard Inherit JsonRecordSetWriter
Lookup Service	KuduLookupService
Result RecordPath	No value set
Routing Strategy	Route to 'matched' or 'unmatched'
Record Result Contents	Insert Entire Record
Record Update Strategy	Use Property
setid	/setid
version	/version



<https://dev.to/tspannhw/flank-using-apache-kudu-as-a-cache-for-fda-updates-4knj>

Listen FTP

Let Apache NiFi be your FTP server

 ⚠ ListenFTP
ListenFTP 1.13.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)		5 min
Read/Write	0 bytes / 0 bytes		5 min
Out	0 (0 bytes)		5 min
Tasks/Time	0 / 00:00:00.000		5 min

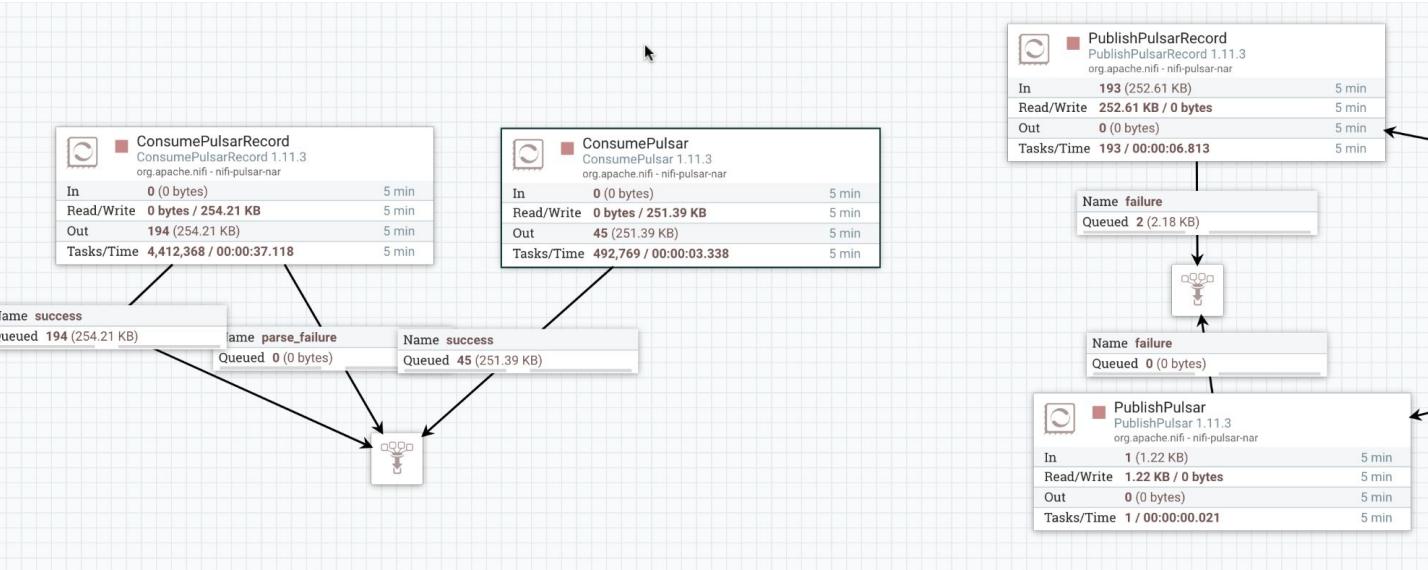
Consume MQTT

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



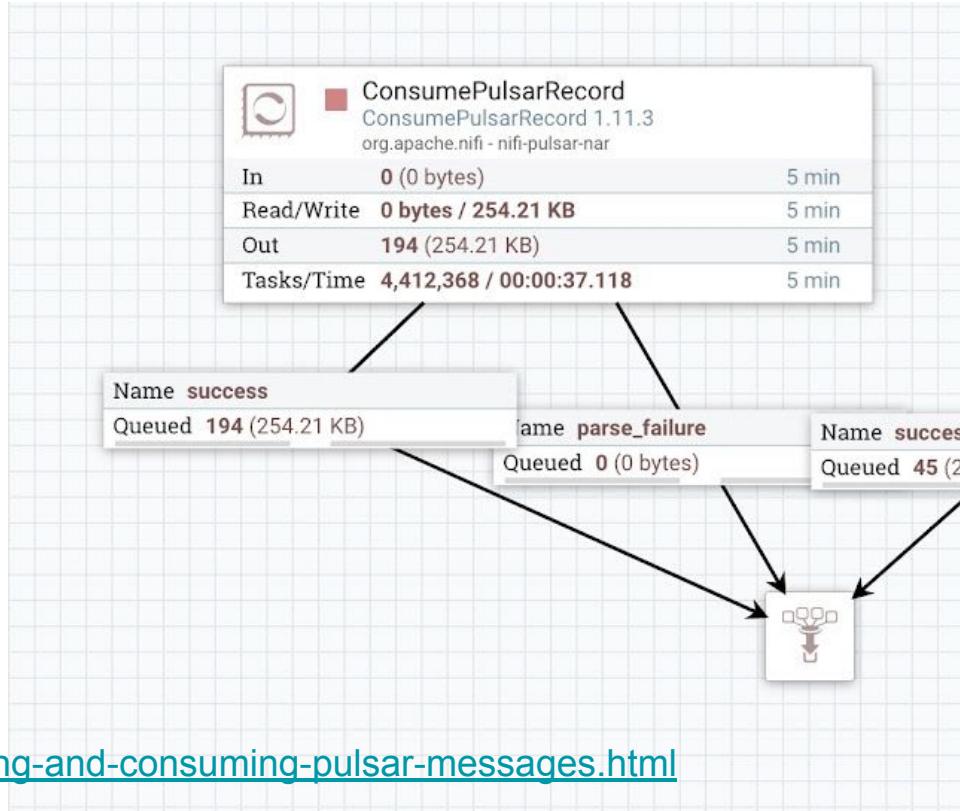
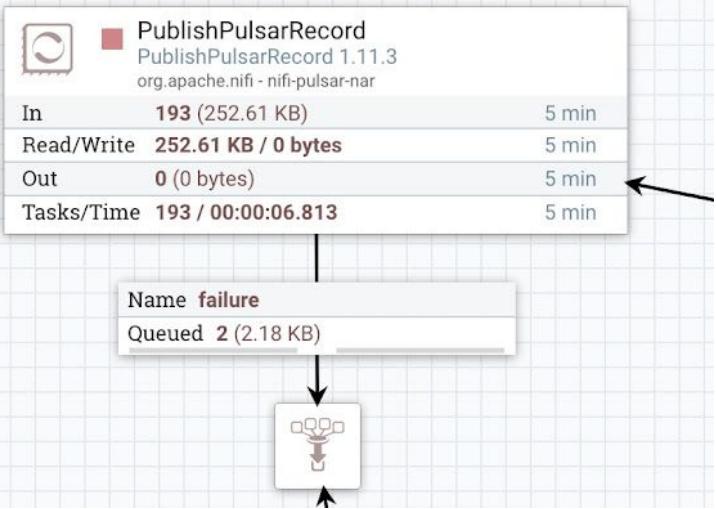
Property	ConsumeMQTT Processor	Value
Session state	?	Clean Session
MQTT Specification Version	?	AUTO
Connection Timeout (seconds)	?	30
Keep Alive Interval (seconds)	?	60
Group ID	?	No value set
Topic Filter	?	No value set
Quality of Service(QoS)	?	0 - At most once
Max Queue Size	?	No value set
Record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

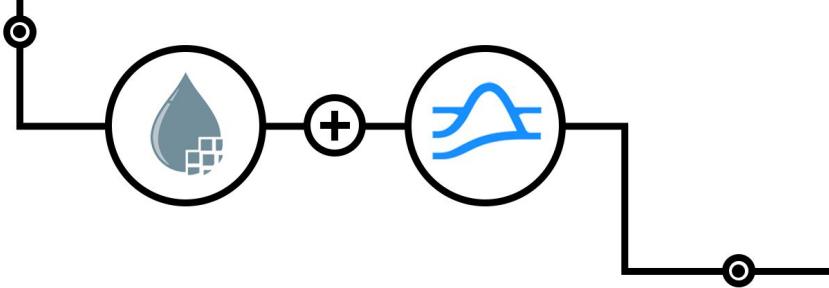
COMMENTS

Required field

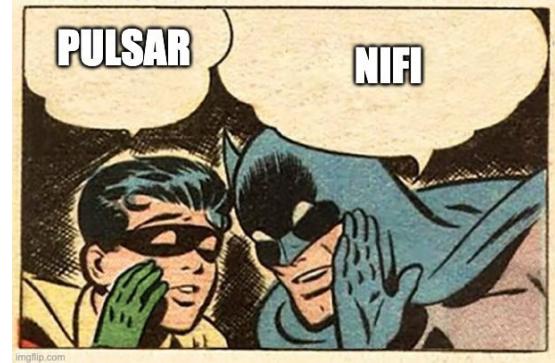
Property	Value
Audience	urn:sn:pulsar:sndev:gke
Issuer URL	https://auth.streamnative.cloud
Private key file	file:///Users/tspann/Documents/servers/services/apache-pulsar-2.8.0/sndev-tspann.json
Trusted Certificate Filename	?

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

Apache NiFi Pulsar Connector



CLOUDERA + Stream Native



Announcing the Integration of
Apache NiFi and Apache Pulsar

<https://streamnative.io/apache-nifi-connector/>

<https://github.com/tspannhw/awesome-nifi-pulsar>

<https://t.co/TbcYhdUPVn>



▶ ConsumeMQTT
ConsumeMQTT 1.13.0
org.apache.nifi - nifi-mqtt-nar

In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 301 / 00:00:00.023 5 min

Name parse.failure
Queued 0 (0 bytes)

Name Message
Queued 0 (0 bytes)

▶ Set Reader Name and Destination
UpdateAttribute 1.13.0
org.apache.nifi - nifi-update-attribute-nar

In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

Name success
Queued 9 (5.41 KB)

▶ SampleRecord
SampleRecord 1.13.0
org.apache.nifi - nifi-standard-nar

In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name failure
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

▶ Query and Convert
QueryRecord 1.13.0
org.apache.nifi - nifi-standard-nar

In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min



▶ UpdateHiveTable
UpdateHiveTable 1.13.0
org.apache.nifi - nifi-hive-nar

In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name failure
Queued 0 (0 bytes)



▶ PutRecord Anywhere From Anywhere
PutRecord 1.13.0
org.apache.nifi - nifi-standard-nar

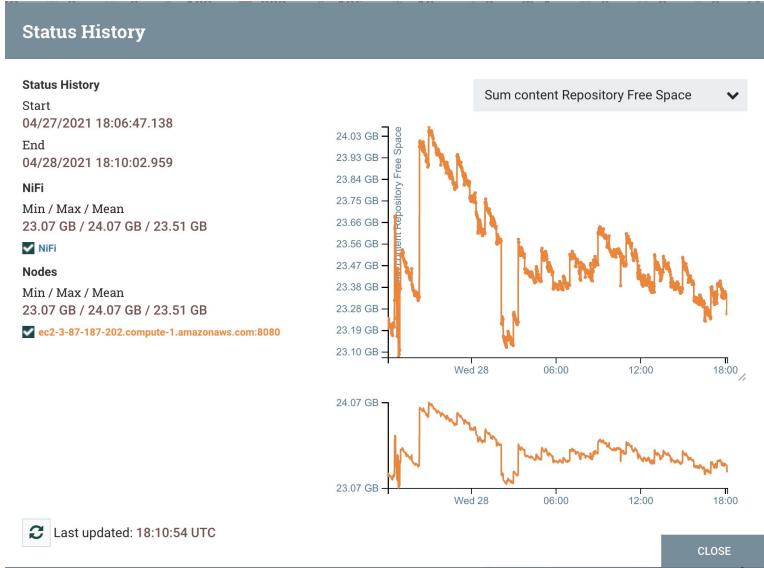
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name failure
Queued 0 (0 bytes)

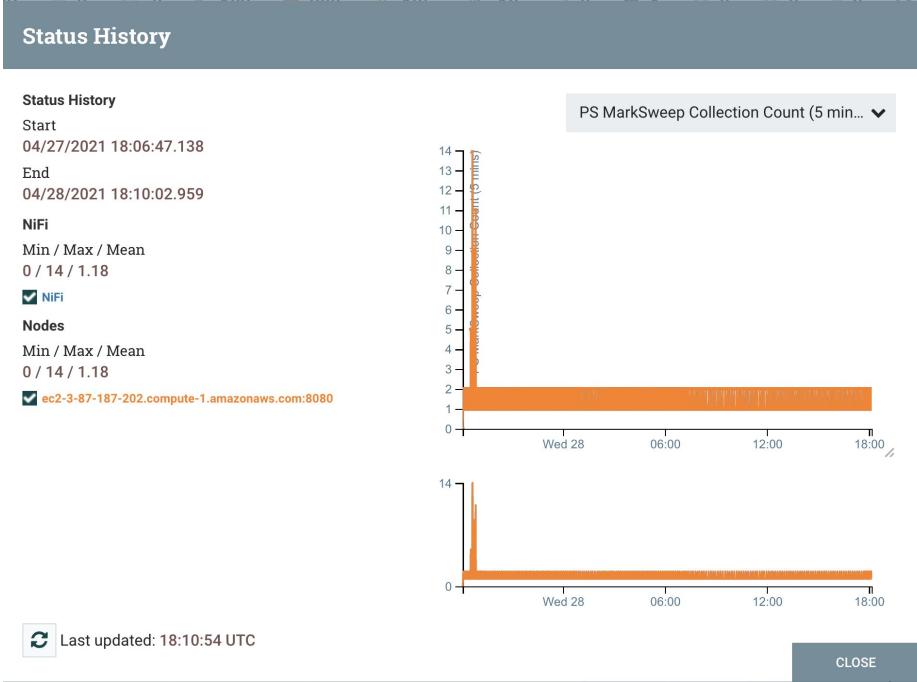


StreamNative

Metrics, Status, Charts



<https://www.clouddataops.dev/data-flow-experience>



DevOps on Apache NiFi 1.15.3

Toolkit Setup on Apache NiFi 1.15.3

Download NiFi Toolkit

Copy keystore and truststore information from your NiFi conf/nifi.properties

Create a nifi.properties file linked to the cli.sh

baseUrl=https://nvidia-desktop:8443

keystore=/home/nvidia/nvme/nifi-1.15.3/conf/keystore.p12

keystoreType=PKCS12

keystorePasswd=5325343412efaab3123c6892d93

keyPasswd=53134eee99da9dbe9349123aa17c6892d93

truststore=/home/nvidia/nvme/nifi-1.15.3/conf/truststore.p12

truststoreType=PKCS12

truststorePasswd=93498Dfdjfhujhure8d8hfd84j3n43jd

DevOps

nifi-toolkit/bin/cli.sh nifi list-param-contexts
nifi-toolkit/bin/cli.sh nifi pg-list
nifi-toolkit/bin/cli.sh nifi pg-set-param-context ...

Or

nifi-toolkit/bin/cli.sh
nifi pg-list

<https://www.datainmotion.dev/2021/01/automating-starting-services-in-apache.html>
<https://nipyapi.readthedocs.io/en/latest/>

DevOps

nifi pg-list

nifi pg-status

nifi pg-get-services

nifi pg-enable-services -u https://nvidia-desktop:8443 --processGroupId **root**

nifi pg-start -u http://edge2ai-1.dim.local:8080 -pgid LOOKTHISUP

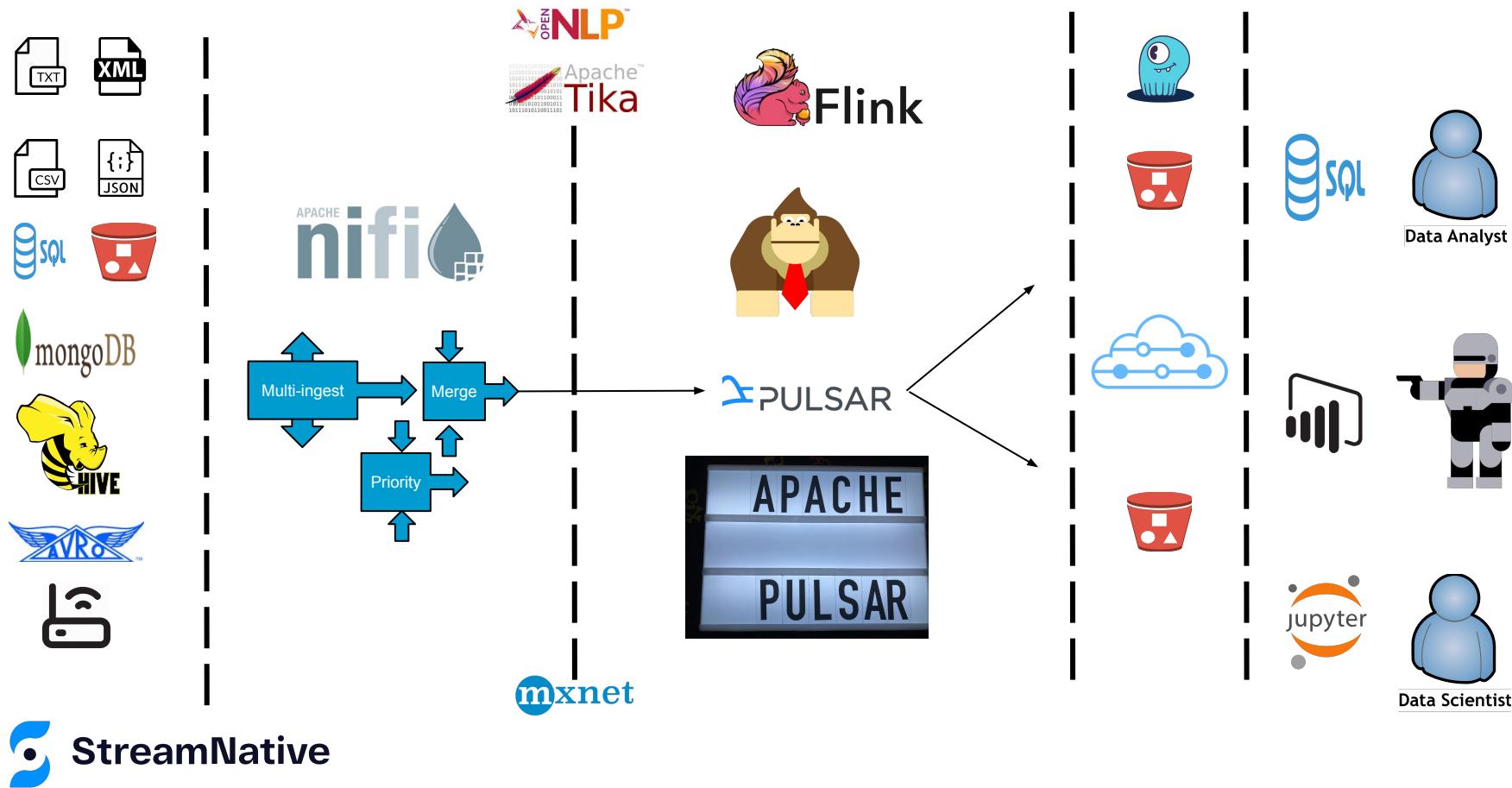
nifi list-param-contexts -u https://nvidia-desktop:8443 -verbose

nifi create-reporting-task -u https://nvidia-desktop:8443 -verbose -i

<https://dev.to/tspannhw/automating-starting-services-in-apache-nifi-and-applying-parameters-5h4n>

<https://github.com/tspannhw/ApacheConAtHome2020/blob/main/scripts/setupnifi.sh>

All Data - Anytime - Anywhere - Multi-Cloud - Multi-Protocol



Apache Pulsar



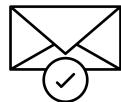


Apache Pulsar is a Cloud-Native
Messaging and Event-Streaming Platform.

Why Apache Pulsar?



**Unified
Messaging
Platform**



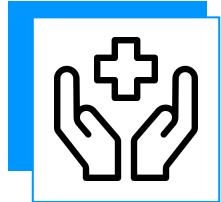
**Guaranteed
Message
Delivery**



Resiliency

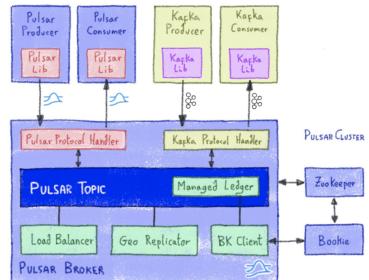


**Infinite
Scalability**

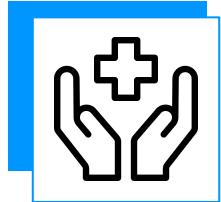


Connectivity

hub.streamnative.io



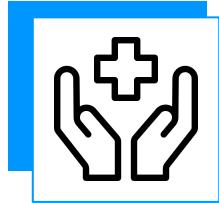
- **Functions** - Lightweight Stream Processing (Java, Python, Go)
- **Connectors** - Sources & Sinks (Cassandra, Kafka, ...)
- **Protocol Handlers** - AoP (AMQP), KoP (Kafka), MoP (MQTT)
- **Processing Engines** - Flink, Spark, Presto/Trino via Pulsar SQL, NiFi
- **Data Offloaders** - Tiered Storage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



Serverless Event Streaming Framework

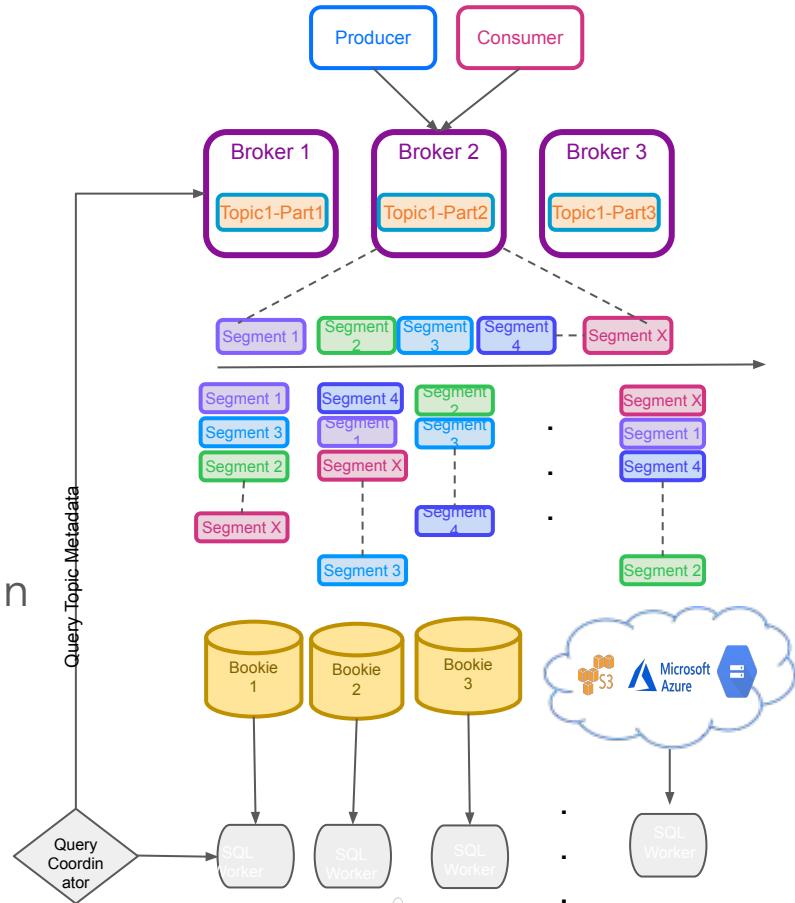
- Lightweight computation similar to AWS Lambda.
- Specifically designed to use Apache Pulsar as a message bus.
- Function runtime can be located within Pulsar Broker.
- Java, Go, Python

<https://streamnative.io/blog/engineering/2021-11-10-streaming-data-pipelines-with-pulsar-io/>



Pulsar SQL

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



Question Everything!

Session

Regular 50 minute session

**Apache NiFi 101: Introduction
and Best Practices**

Primary Speaker

Fri 12:00

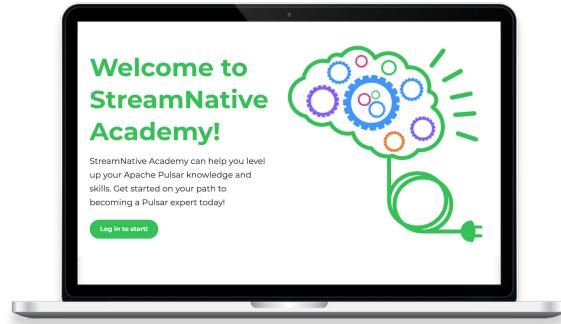


Feedback Link

<https://sqlb.it/?7108>

Deeper Content

- <https://www.datainmotion.dev/2021/11/producing-and-consuming-pulsar-messages.html>
- <https://www.datainmotion.dev/2020/06/no-more-spaghetti-flows.html>
- <https://github.com/tspannhw/EverythingApacheNiFi>
- <https://www.datainmotion.dev/2019/03/apache-nifi-101.html>
- <https://nifi.apache.org/docs/nifi-docs/html/nifi-in-depth.html>
- <https://pierrevillard.com/best-of-nifi/>
- <https://blogs.apache.org/nifi/>
- <https://www.nifi.rocks/documents/nifi-expression-language-cheat-sheet.pdf>
- <https://dev.to/tspannhw/new-features-of-apache-nifi-1-13-0-45ln>
- <https://dev.to/tspannhw/tracking-satellites-with-apache-nifi-44o7>
- <https://www.datainmotion.dev/2021/01/flank-using-apache-kudu-as-cache-for.html>
- <https://www.datainmotion.dev/2020/12/basic-understanding-of-cloudera-flow.html>
- <https://bryanbende.com/development/2021/11/10/apache-nifi-stateless>



Now Available

On-Demand Pulsar Training

Academy.StreamNative.io