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

StreamNative

Timothy Spann | Developer Advocate

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

CLOUDERA HORTONWORKS Pivotal Hewlett Packard Enterprise pwc

StreamNative

Stream Native

Stream Native

Founded by the original developers of Apache Pulsar.

Passionate and dedicated team.

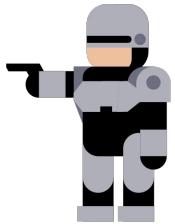
StreamNative helps teams to **capture**, **manage**, and **leverage data** using Pulsar's unified messaging and streaming platform.

streamnative.io

FLiP Stack Weekly

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

https://bit.ly/32dAJft



StreamNative

Streaming FLiPN Apps

<-> 

<-> Events <-> 

Edge

Gateway

Protocols

Apps

Pulsar

KoP

MoP

Websocket

CDC

Unified Batch and Stream **STORAGE**

Streaming

Unified Batch and Stream **COMPUTING**

Offload

Batch

Tiered Storage

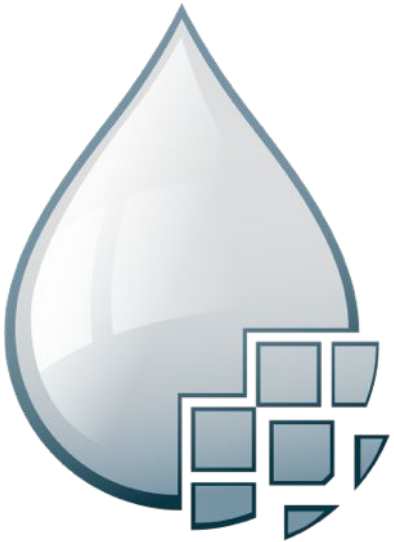
StreamNative Cloud

Pulsar
Sink

StreamNative Hub

Apache NiFi

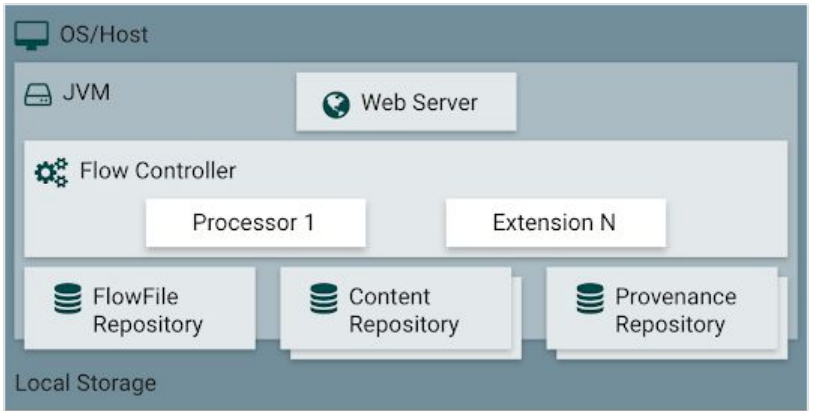
Why Apache NiFi?

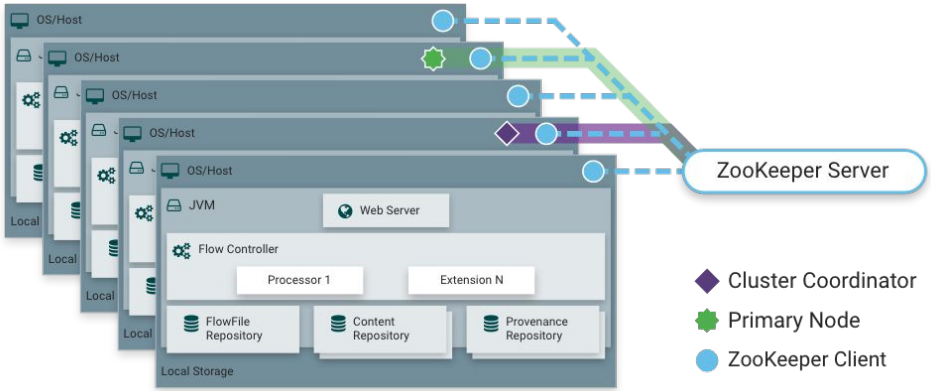


- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models

- 350+ processors
- Visual command and control
- Over a 100 sources
- Flow templates
- Pluggable/multi-role security
- Designed for extension
- Clustering
- Version Control

Architecture

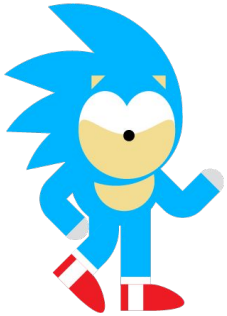




https://nifi.apache.org/docs/nifi-docs/html/overview.html

Flow File

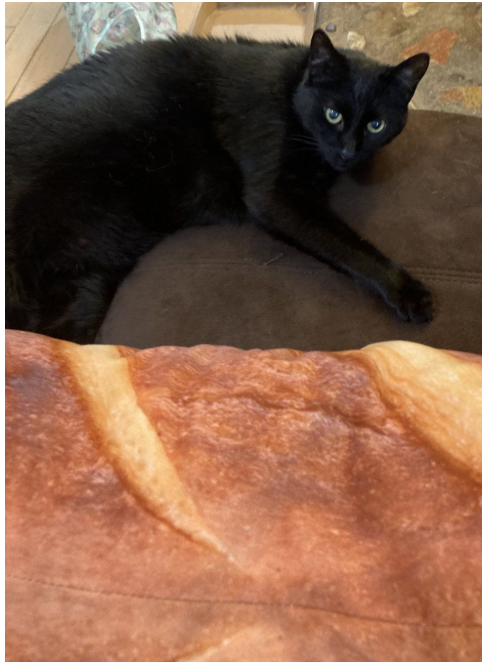
Flow Files are content and key/value pairs for attributes that are each event/message/file that has been introduced into NiFi.



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.

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

Connection

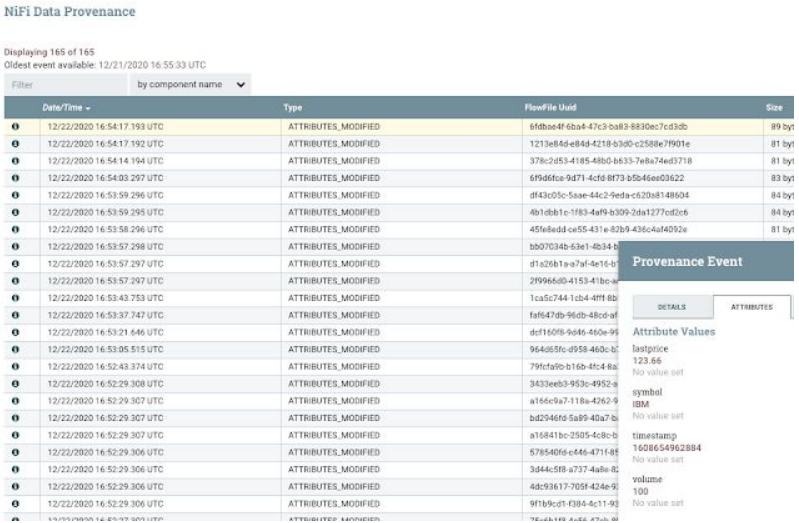
These link together NiFi processors.

Process Groups

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

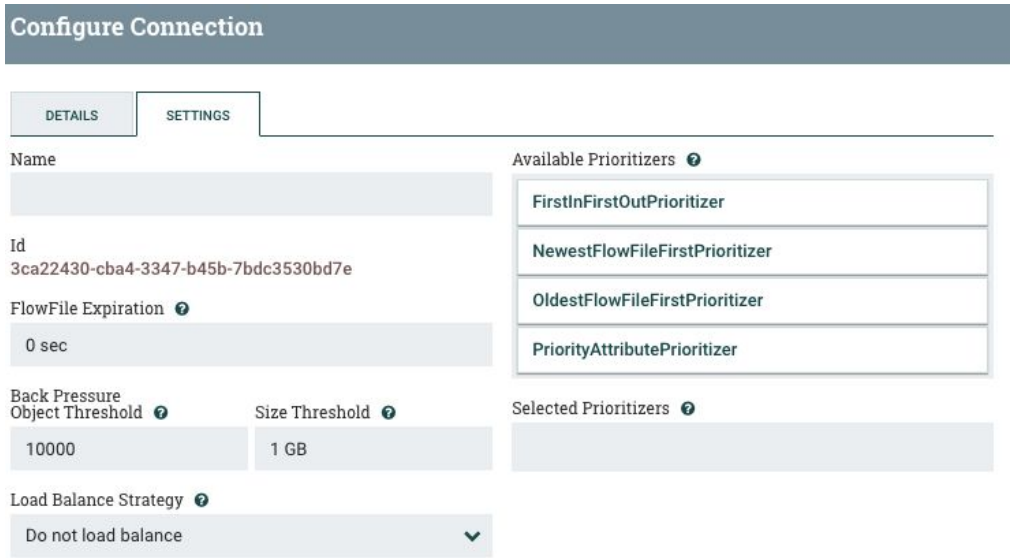
https://nifi.apache.org/docs/nifi-docs/html/overview.html

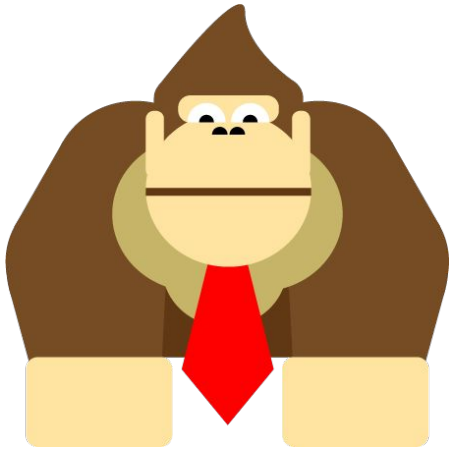
Provenance



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

Backpressure & Prioritizers





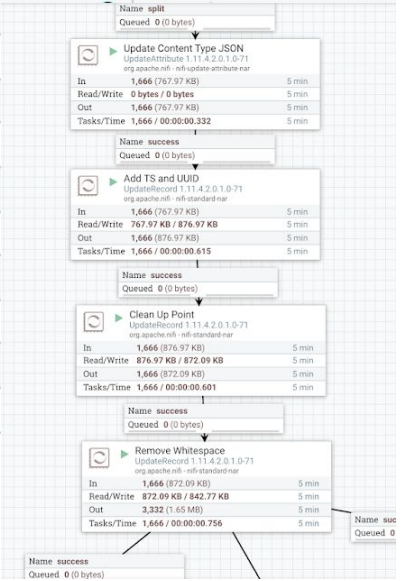
https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Record Processors

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

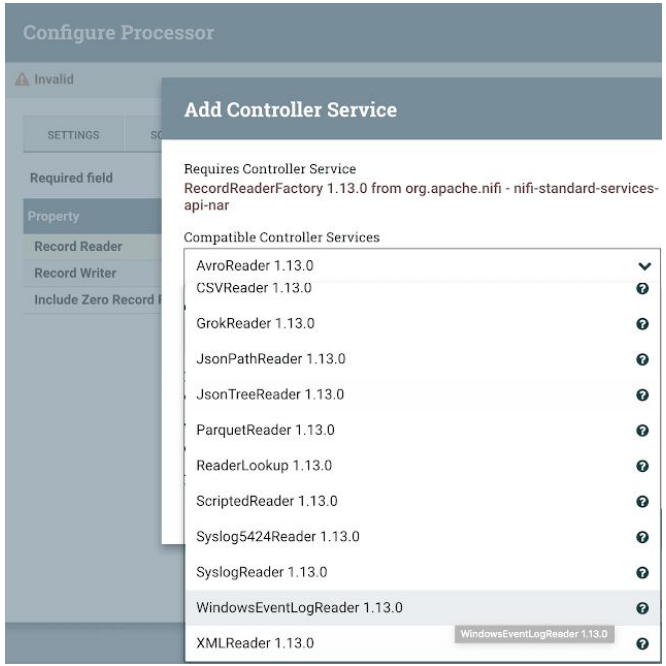
Property	Value
Record Reader	**XMLReader**
Record Writer	**JsonRecordSetWriter**
Include Zero Record FlowFiles	**false**
Cache Schema	**true**
query1	SELECT * FROM FLOWFILE

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

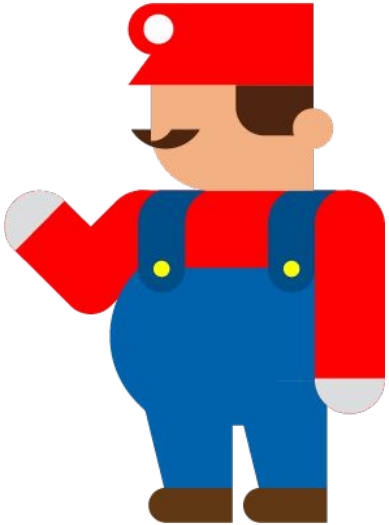
Bundle

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

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

Caching

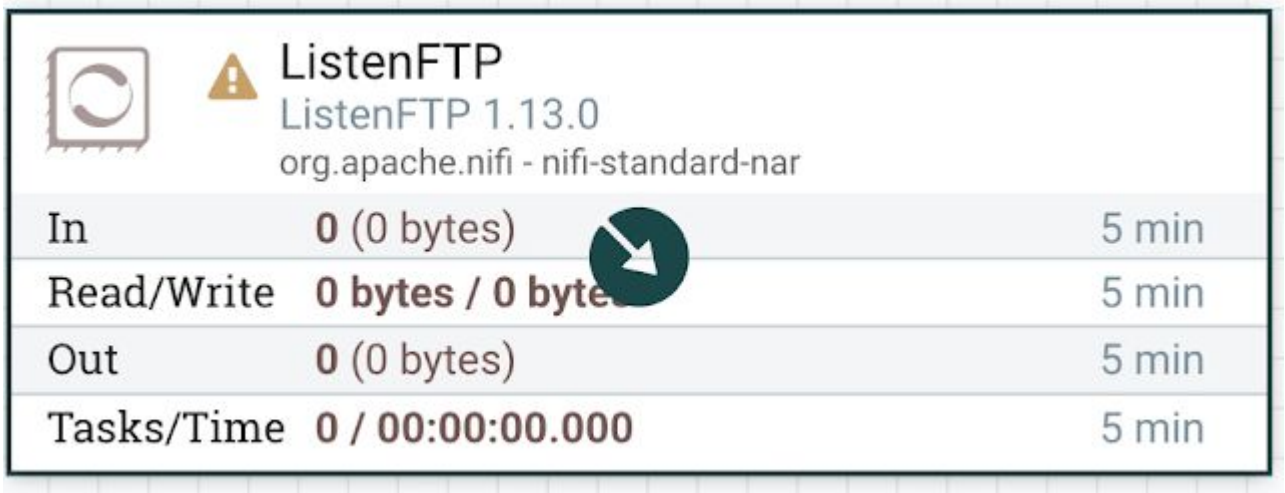
Property	Value
Record Reader	Infer JsonTreeReader
Record Writer	Standard Inherit JsonRecordSetWriter
Lookup Service	KuduLookupService
Result RecordPath	No value set
Routing Strategy	Route to 'matched' or 'unmatched'
Record Result Contents	Insert Entire Record
Record Update Strategy	Use Property
setid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



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

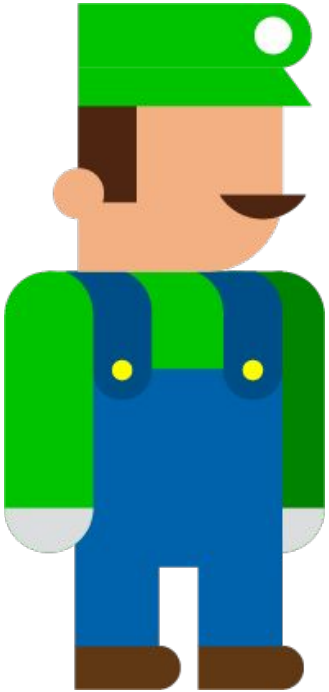
Listen FTP

Let Apache NiFi be your FTP server



Consume MQTT

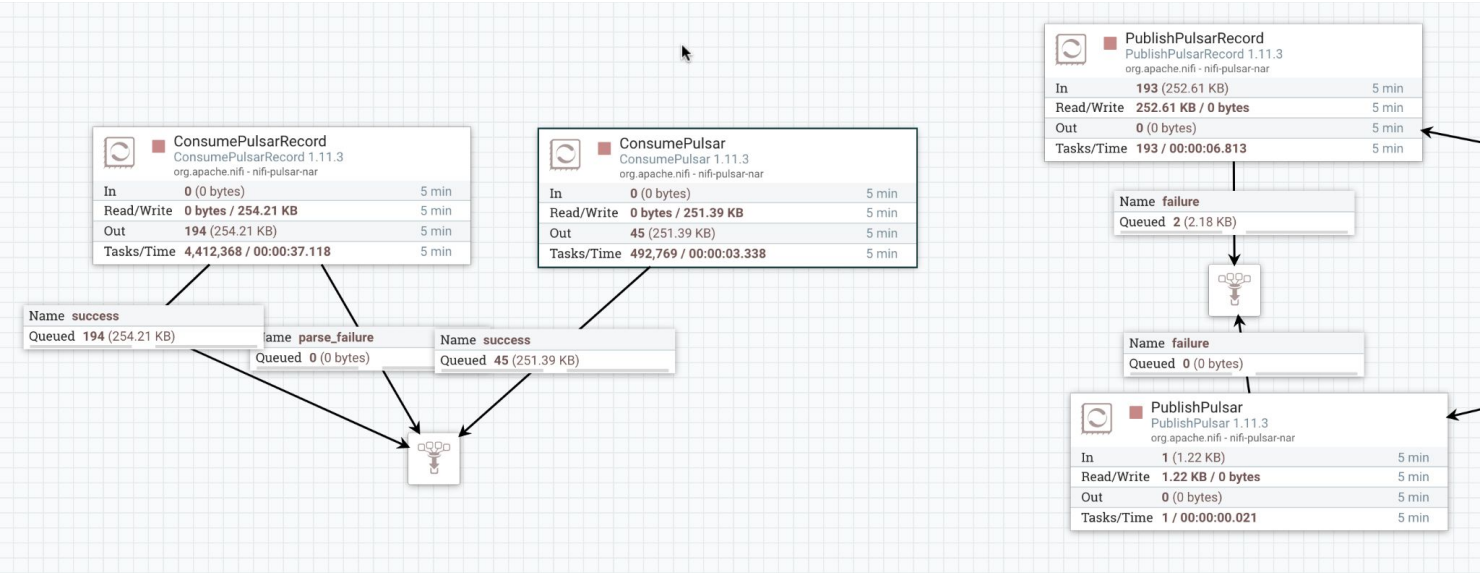
This could read from Apache Pulsar - MoP (MQTT on Pulsar)



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

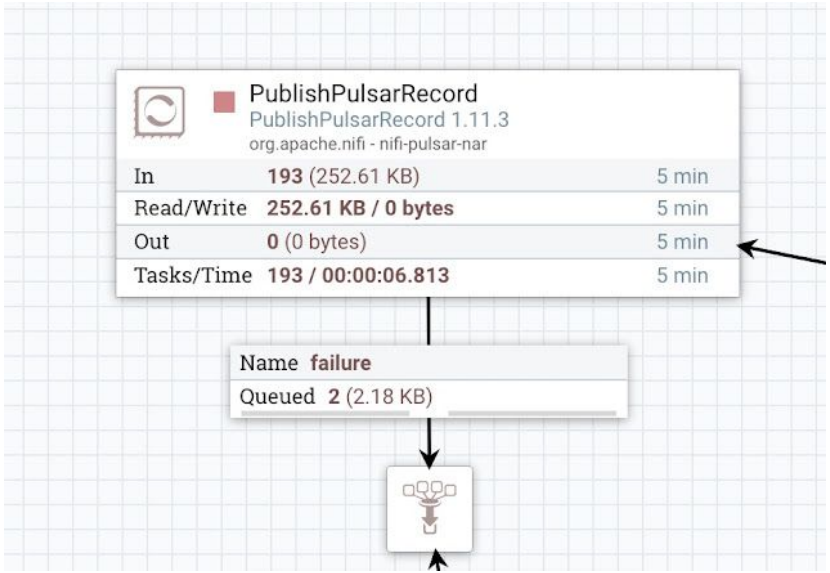
Apache NiFi Pulsar Connector

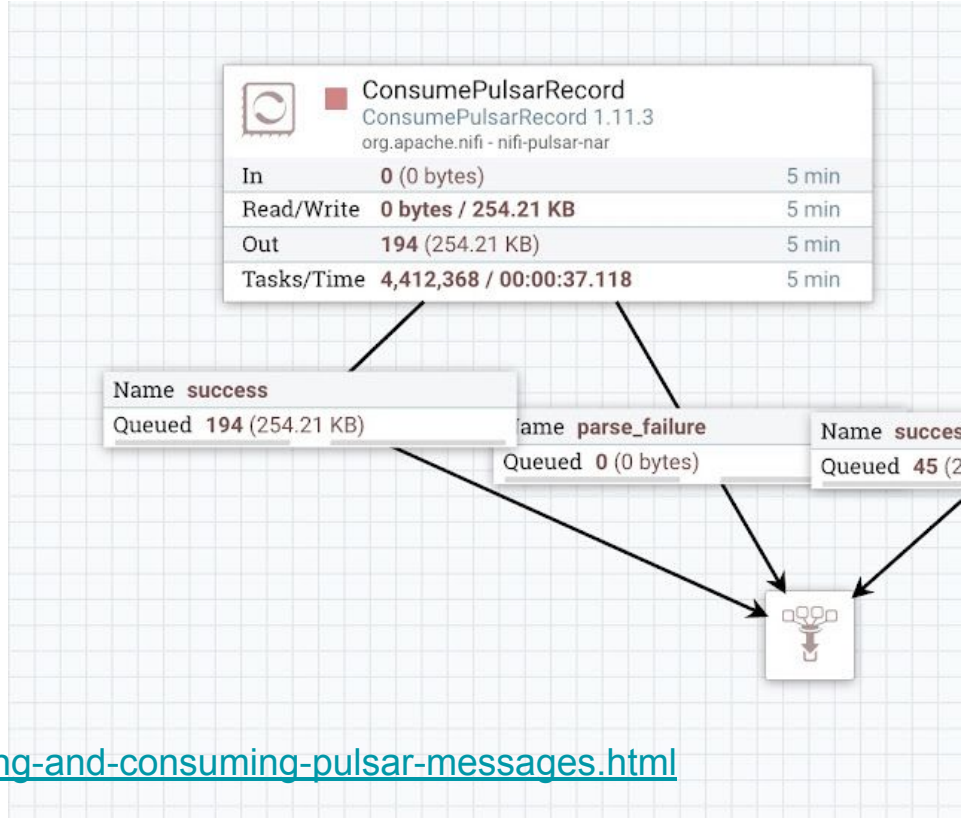


https://github.com/david-streamlio/pulsar-nifi-bundle

StreamNative

Apache NiFi Pulsar Connector





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

StreamNative

Apache NiFi Pulsar Connector

Displaying 12 of 339

pulsar

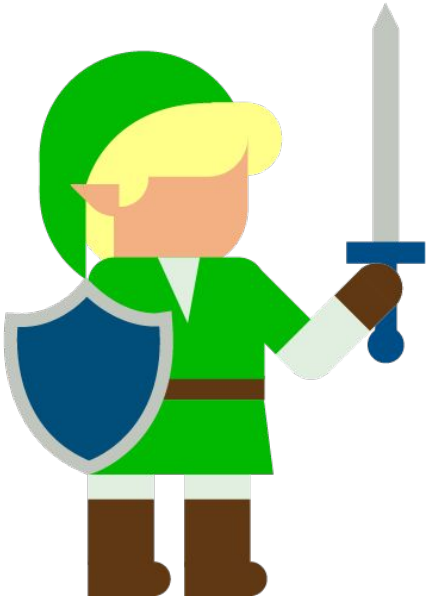
Type	Version ▲	Tags
ConsumePulsar	1.11.0	PubSub, Consume, Ingest, Get, I...
ConsumePulsarRecord	1.11.0	PubSub, Consume, Ingest, Get, ...
PublishPulsar	1.11.0	PubSub, Message, Pulsar, Apac...
PublishPulsarRecord	1.11.0	PubSub, 1.0, Message, csv, json...
ConsumePulsar	1.11.3	PubSub, Consume, Ingest, Get,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

StreamNative

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

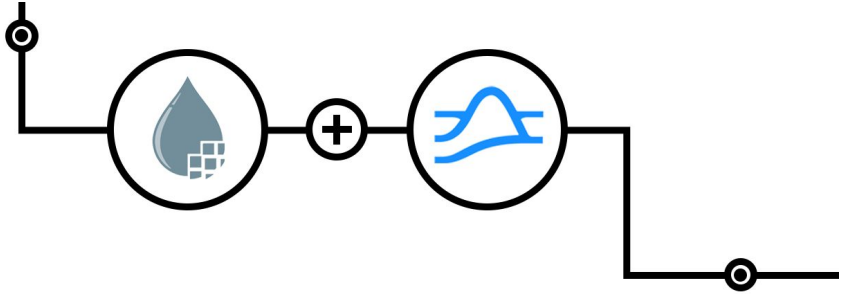
StreamNative

Apache NiFi Pulsar Connector

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

StreamNative

Apache NiFi Pulsar Connector



CLOUDERA + Stream Native

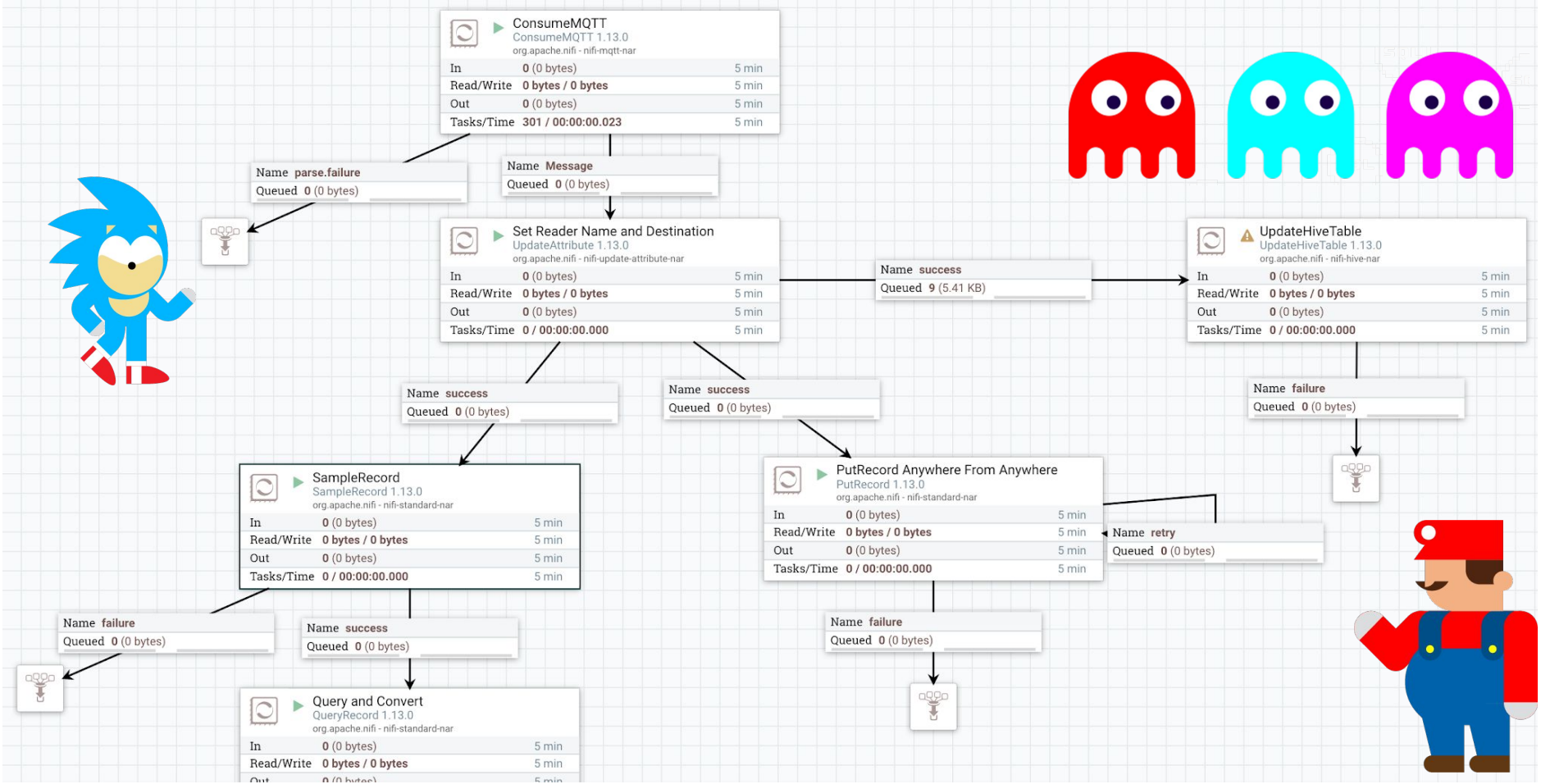
Announcing the Integration of
Apache NiFi and Apache Pulsar

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

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

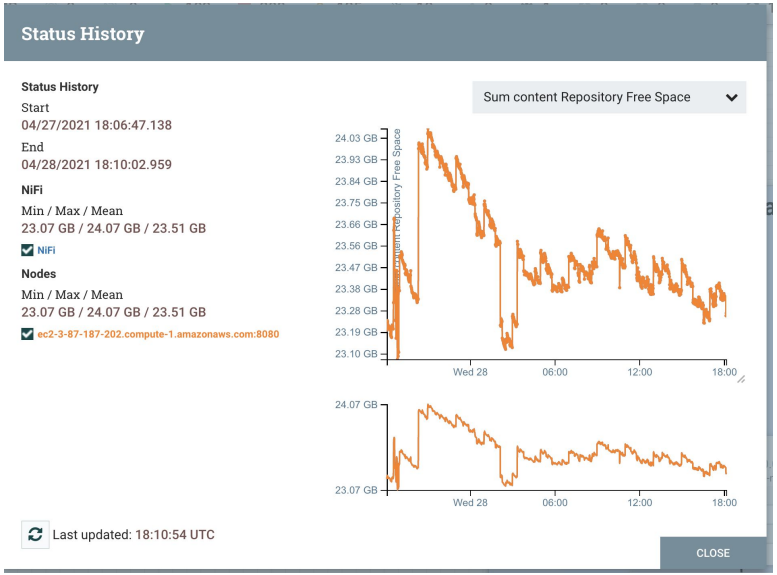
https://t.co/TbcYhdUPVn

StreamNative



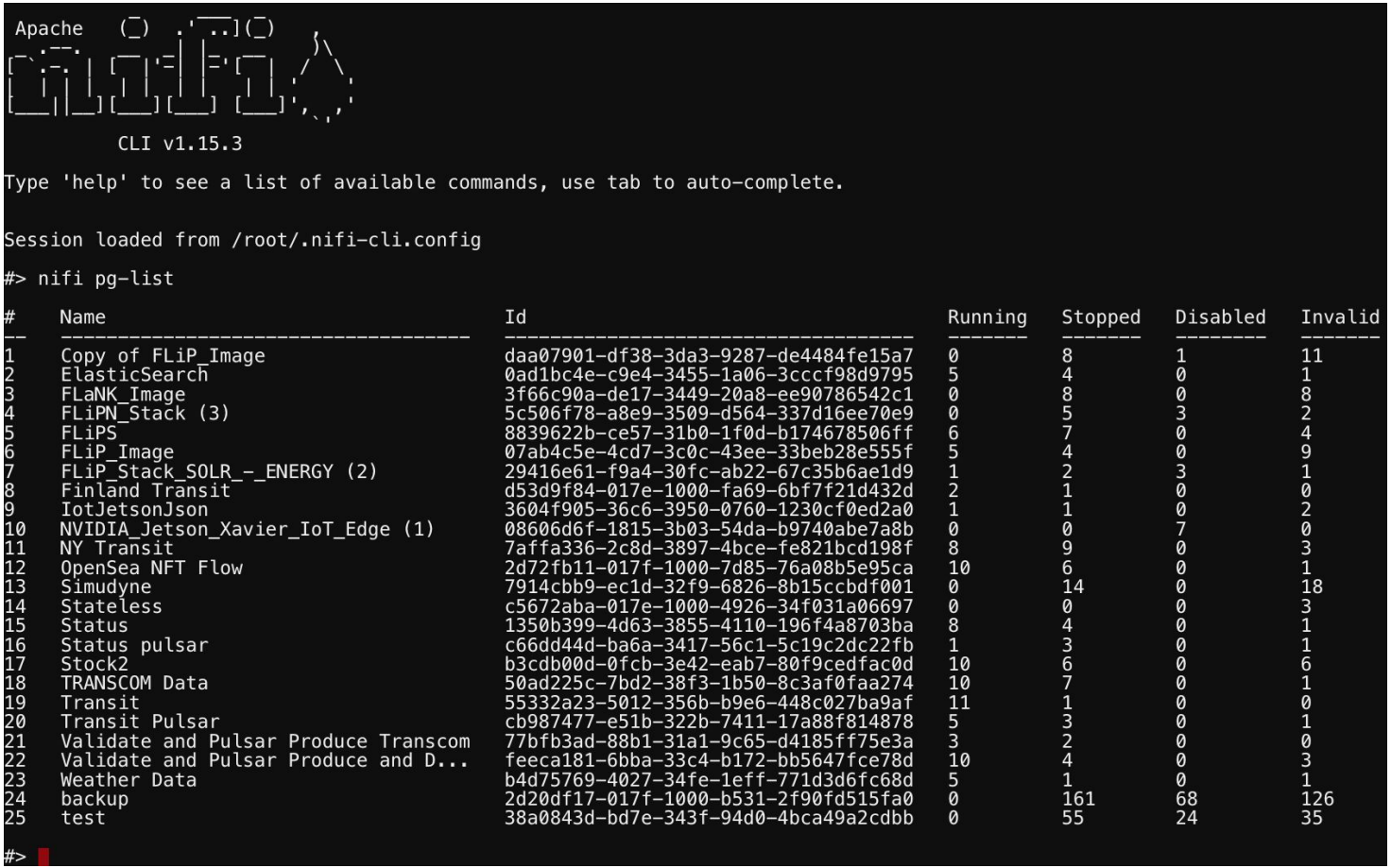
StreamNative

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=93498Dfdjfhujdhure8d8hfd84j3n43jd

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

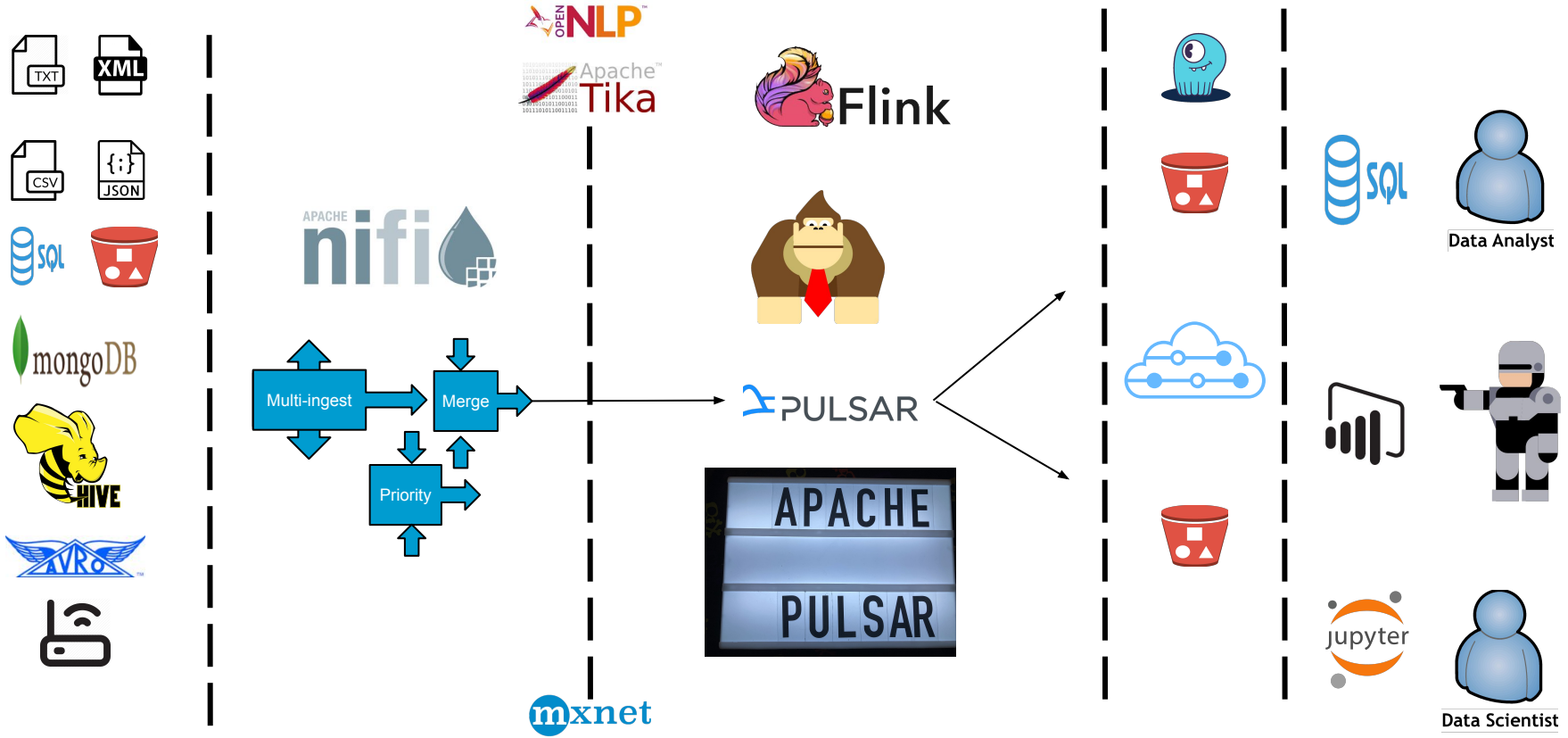
https://www.datainmotion.dev/2021/01/automating-starting-services-in-apache.html
https://nipyapi.readthedocs.io/en/latest/

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



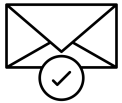
StreamNative

Apache Pulsar

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

Why Apache Pulsar?

Unified Messaging Platform



Guaranteed Message Delivery

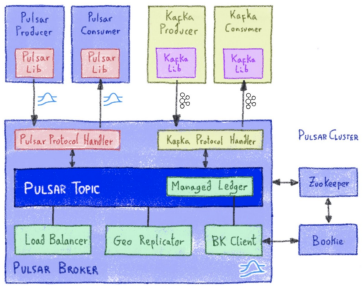
Resiliency

Infinite Scalability

StreamNative

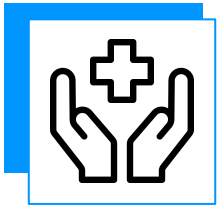
Connectivity

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)

StreamNative



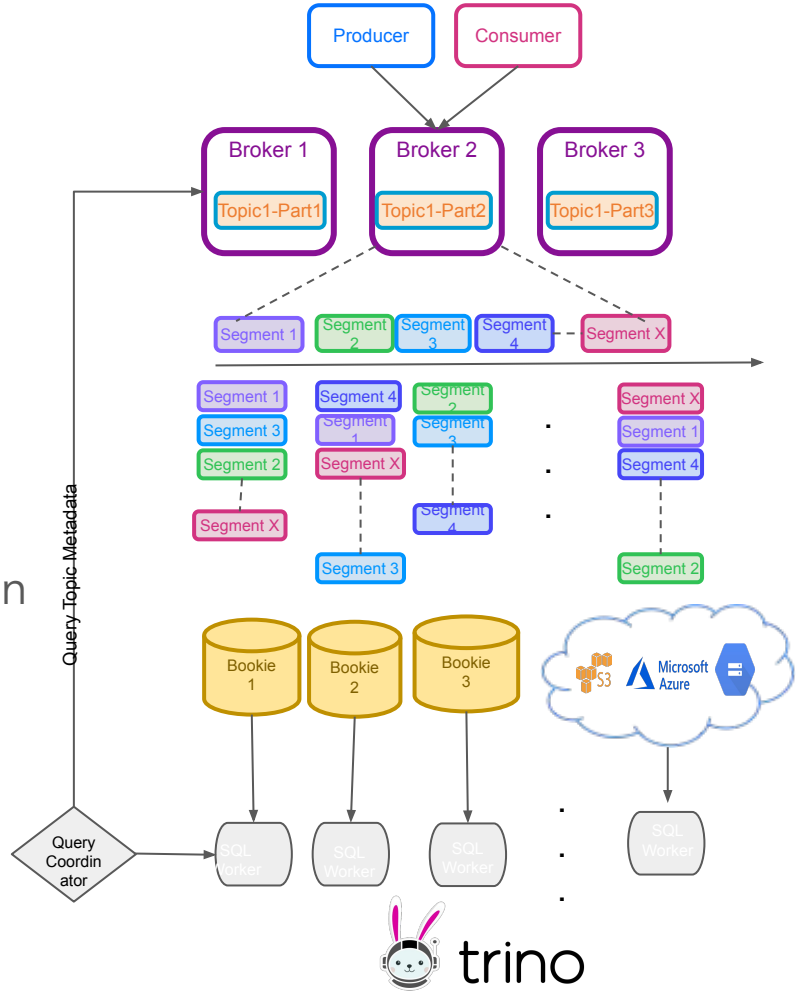
Serverless Event Streaming Framework

- 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.



StreamNative

Question Everything!

Session

Regular 50 minute session
Apache NiFi 101: Introduction and Best Practices
Primary Speaker
Fri 12:00

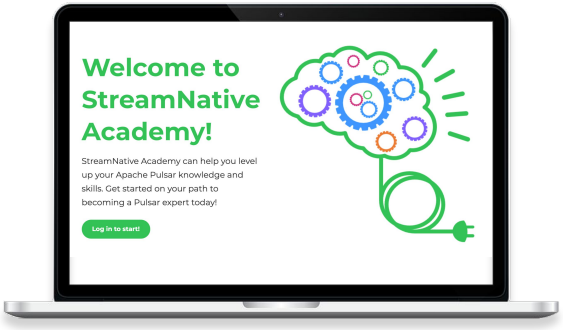
Feedback Link

https://sqlb.it/?7108

StreamNative

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

StreamNative