Apache NiFi 101

Timothy Spann | Developer Advocate

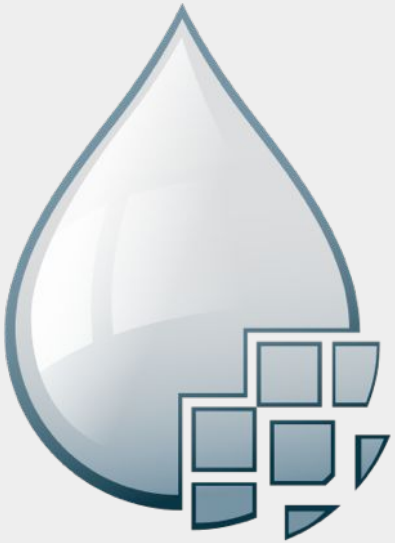
Tim Spann

Developer Advocate

DZone Zone Leader and Big Data MVB Data DJay

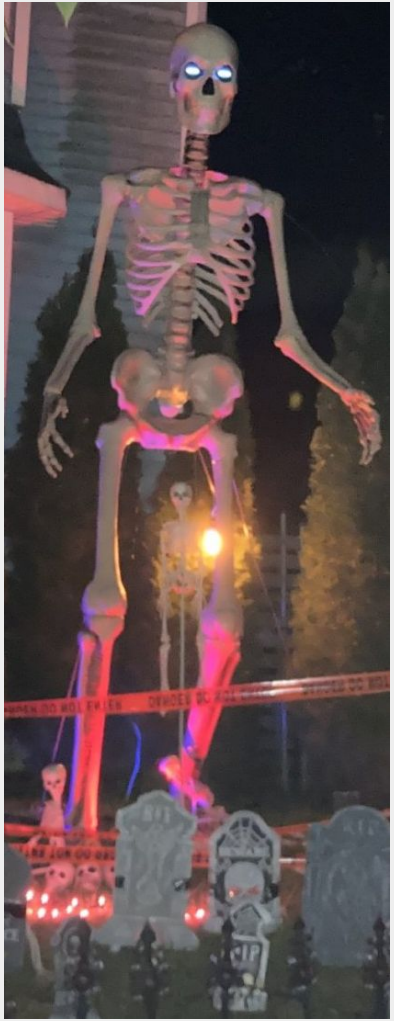
- https://www.datainmotion.dev/
- https://github.com/tspannhw/SpeakerProfile
- https://dev.to/tspannhw
- https://sessionize.com/tspann/
- https://streamnative.io/

Apache NiFi



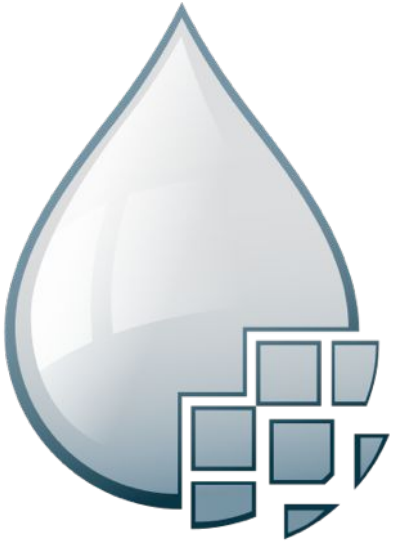
StreamNative

Don’t Be Afraid of Open Source



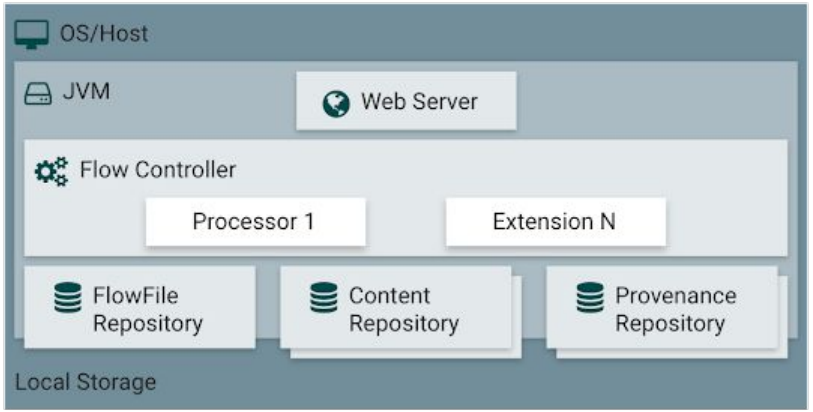
StreamNative

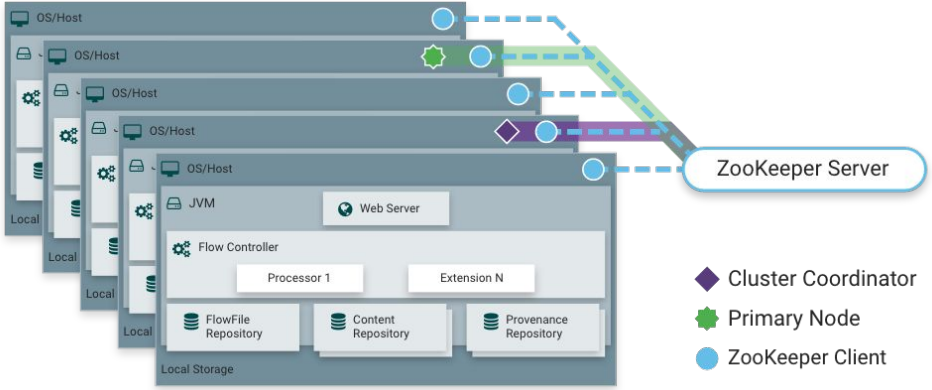
Why Apache NiFi?



- 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

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	6fdbae4f-6ba4-47c3-ba83-8830ec7cd3db	89 byt
12/22/2020 16:54:17.192 UTC	ATTRIBUTES_MODIFIED	1213e84d-e84d-4218-b3d0-c2588e7f901e	81 byt
12/22/2020 16:54:14.194 UTC	ATTRIBUTES_MODIFIED	378c2d53-4185-48b0-b633-7e8a74ed3718	81 byt
12/22/2020 16:54:03.297 UTC	ATTRIBUTES_MODIFIED	6f9d6fce-9d71-4cfd-8f73-b5b46ee03622	83 byt
12/22/2020 16:53:59.296 UTC	ATTRIBUTES_MODIFIED	df43c05c-5aae-44c2-9eda-c620a8148604	84 byt
12/22/2020 16:53:59.295 UTC	ATTRIBUTES_MODIFIED	4b1dbb1c-1f83-4af9-b309-2da1277cd2c6	84 byt
12/22/2020 16:53:58.296 UTC	ATTRIBUTES_MODIFIED	45fe8edd-ce55-431e-82b9-436c4af4092e	81 byt
12/22/2020 16:53:57.298 UTC	ATTRIBUTES_MODIFIED	bb07034b-63e1-4b34-b	
12/22/2020 16:53:57.297 UTC	ATTRIBUTES_MODIFIED	d1a26b1a-a7af-4e16-b	
12/22/2020 16:53:57.297 UTC	ATTRIBUTES_MODIFIED	2f9966d0-4153-41bc-a	
12/22/2020 16:53:43.753 UTC	ATTRIBUTES_MODIFIED	1ca5c744-1cb4-4fff-8b	
12/22/2020 16:53:37.747 UTC	ATTRIBUTES_MODIFIED	faf647db-96db-48cd-af	
12/22/2020 16:53:21.646 UTC	ATTRIBUTES_MODIFIED	dcf160f8-9d46-460e-99	
12/22/2020 16:53:05.515 UTC	ATTRIBUTES_MODIFIED	964d65fc-d958-460c-b7	
12/22/2020 16:52:43.374 UTC	ATTRIBUTES_MODIFIED	79fcfa9b-b16b-4fc4-8a	
12/22/2020 16:52:29.308 UTC	ATTRIBUTES_MODIFIED	3433eeb3-953c-4952-a	
12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	a166c9a7-118a-4262-9	
12/22/2020 16:52:29.307 UTC	ATTRIBUTES_MODIFIED	bd2946fd-5a89-40a7-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	578540fd-c446-471f-85	
12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	3d44c5f8-a737-4a8e-82	
12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	4dc93617-705f-424e-9	
12/22/2020 16:52:29.306 UTC	ATTRIBUTES_MODIFIED	9f1b9cd1-f384-4c11-93	

Provenance Event

DETAILS | ATTRIBUTES

Attribute Values

lastprice
123.66
No value set

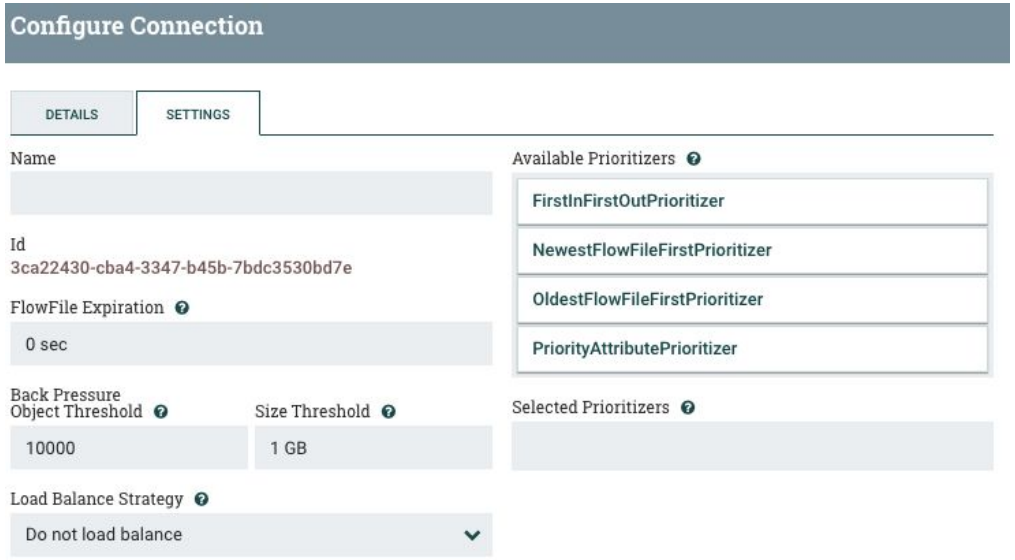
symbol
IBM
No value set

timestamp
1608654962884
No value set

volume
100
No value set

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

Backpressure & Prioritizers



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

System Diagnostics

NiFi Summary

PROCESSORS | INPUT PORTS | OUTPUT PORTS | REMOTE PROCESS GROUPS | CONNECTIONS | PROCESS GROUPS

Displaying 1,338 of 1,338

Filter by name

Name	Type	Process Group	Run Status	In (Size) 5 min	Read | Write 5 min	Out (Size) 5 min	Tasks | Time 5 min
PublishPulsar	PublishPulsar	Status Pulsar	Running (1)	61 (50.13 KB)	50.13 KB | 0 bytes	0 (0 bytes)	5,504,778 | 00:00:23.264
Acquire Satellite Data	GenerateFlowFile	Satellite Data	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000
Acquire Satellite Data	GenerateFlowFile	Satellite Data	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000
Analyze Data in Stream	QueryRecord	Fresh Food	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000
App Data	PublishKafkaRecord_2_0	Mobile Ingest	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000
App Data	PublishKafkaRecord_2_0	Mobile Ingest	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000
AttributeCleanerProcessor	AttributeCleanerProcessor	Mobile Ingest	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000
AttributeCleanerProcessor	AttributeCleanerProcessor	Mobile Ingest	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000
Attributes Grab	EvaluateJsonPath	Predict Temperature	Disabled	0 (0 bytes)	0 bytes | 0 bytes	0 (0 bytes)	0 | 00:00:00.000

Flow File

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

View as: formatted


```
[ {
  "title" : "CDP Control-plane network maintenance (low-impact)",
  "description" : "<p><small>Jul <var data-var='date'> 9</var>, <var data-var='time'>00:14</var> UTC</small><br><strong>Completed</strong> - This operation is complete.</p><p><small>Jul <var data-var=
  "pubDate" : "Thu, 09 Jul 2020 00:14:58 +0000",
  "link" : "https://status.cloudera.com/incidents/967wlx3x5rv7",
  "guid" : "https://status.cloudera.com/incidents/967wlx3x5rv7"
} ]
```


FlowFile

DETAILS | ATTRIBUTES

FlowFile Details

UUID
29235844-0576-432b-9078-b477f9ff6ff5

Filename
35714351-11c8-48e0-b52c-6e295edbbb69

File Size
1.15 KB

Queue Position
No value set

Queued Duration
1 days and 03:58:04.364

Lineage Duration
1 days and 04:01:15.707

Penalized
No

Content Claim

Container
default

Section
2

Identifier
1631110638665-2

Offset
50840

Size
1.15 KB

DOWNLOAD VIEW

FlowFile

DETAILS | ATTRIBUTES

Attribute Values

invokehttp.tx.id
4e97d934-a7f5-463b-ad4a-d74ae6da4325

link
https://status.cloudera.com/incidents/967wlx3x5rv7

mime-type
application/json

mime.type
application/json

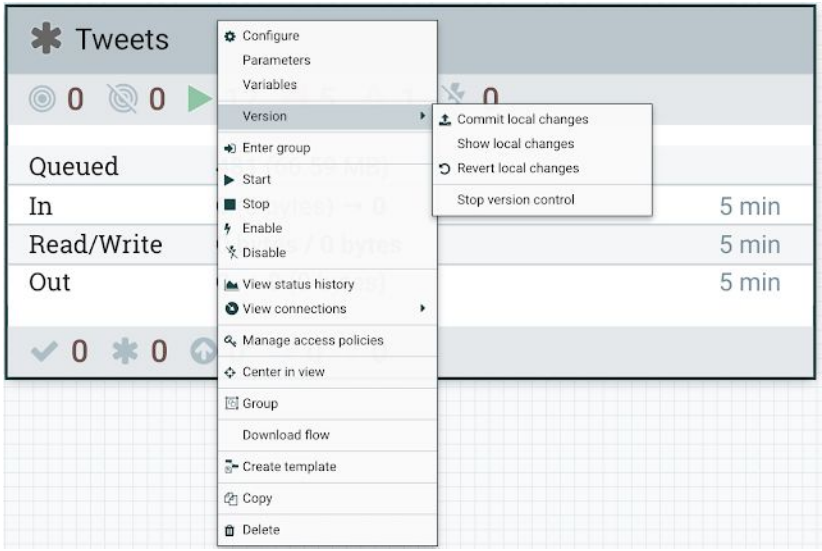
path
./

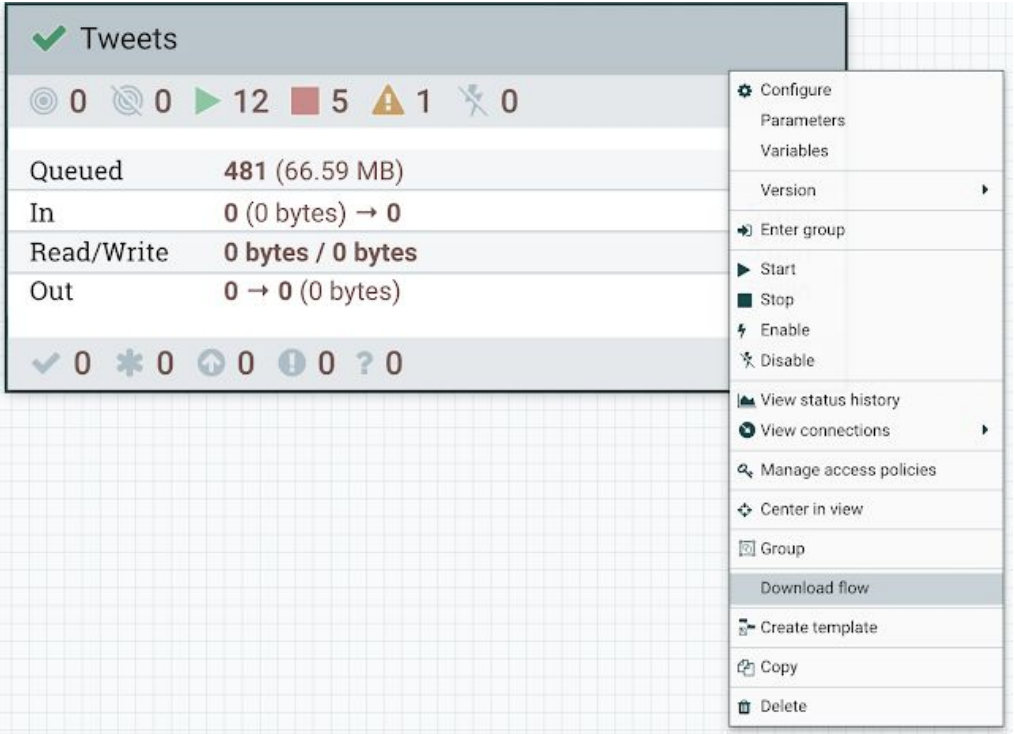
pubdate
Thu, 09 Jul 2020 00:14:58 +0000

record.count

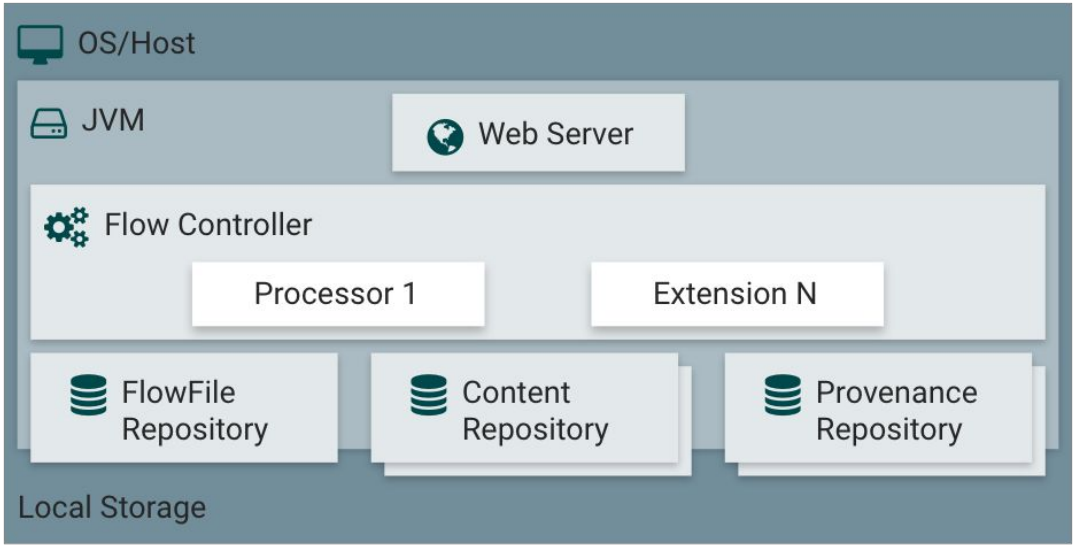
https://nifi.apache.org/docs/nifi-docs/html/overview.html

Version Control (Github and Beyond)





Repositories



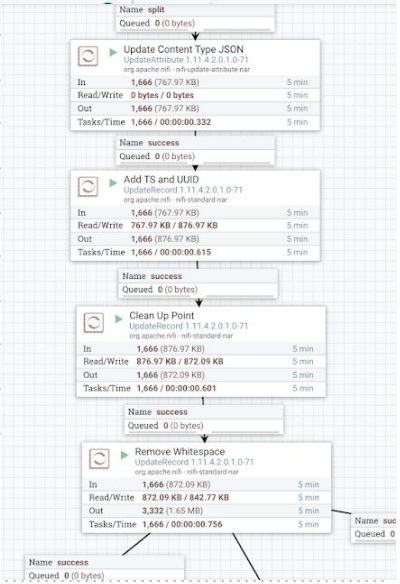
https://nifi.apache.org/docs/nifi-docs/html/nifi-in-depth.html#repositories

Record Processors

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

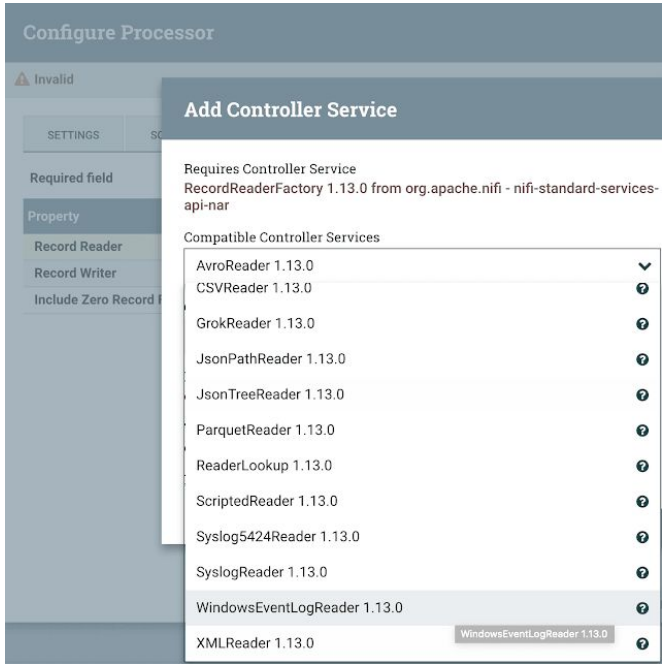
Property	Value
Record Reader	**XMLReader**
Record Writer	**JsonRecordSetWriter**
Include Zero Record FlowFiles	**false**
Cache Schema	**true**
query1	SELECT * FROM FLOWFILE

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

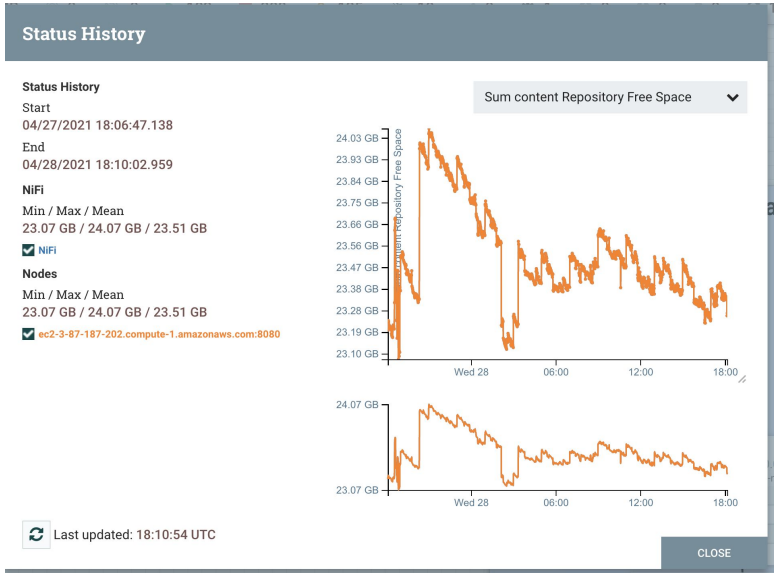
https://www.datainmotion.dev/2019/03/advanced-xml-processing-with-apache.html

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

Metrics, Status, Charts



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

DevOps

nifi-toolkit/bin/cli.sh nifi list-param-contexts -u http://edge2ai-1.dim.local:8080
nifi-toolkit/bin/cli.sh nifi pg-list -u http://edge2ai-1.dim.local:8080
nifi-toolkit/bin/cli.sh nifi pg-set-param-context -u http://edge2ai-1.dim.local:8080 ...

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 http://edge2ai-1.dim.local:8080 --processGroupId root
nifi pg-start -u http://edge2ai-1.dim.local:8080 -pgid LOOKTHISUP
nifi list-param-contexts -u http://edge2ai-1.dim.local:8080 -verbose
nifi create-reporting-task -u http://edge2ai-1.dim.local:8080 -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

Consume MQTT

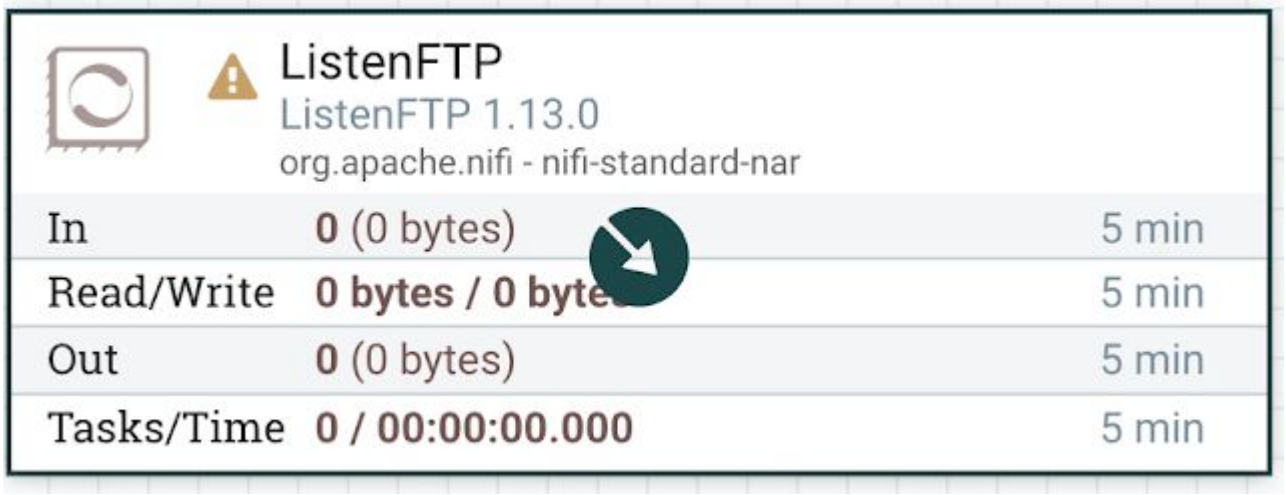
This could read from Apache Pulsar - MoP (MQTT on Pulsar)

ConsumeMQTT
Processor

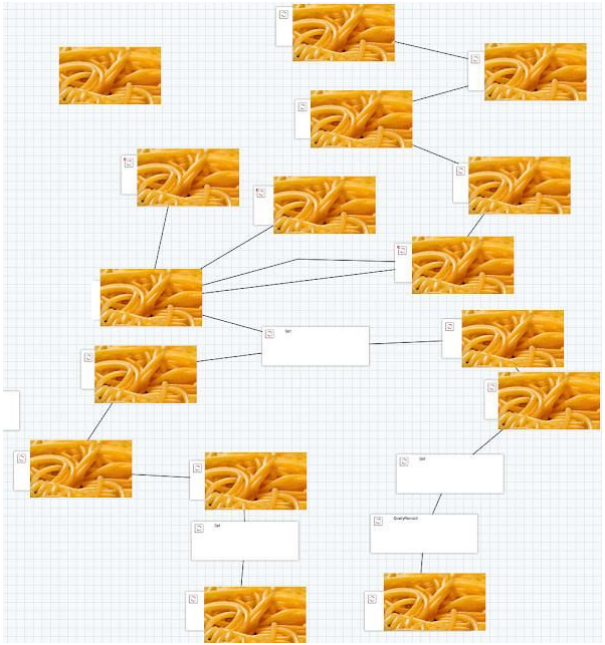
Property	Value
Session state	**Clean Session**
MQTT Specification Version	**AUTO**
Connection Timeout (seconds)	30
Keep Alive Interval (seconds)	60
Group ID	No value set
Topic Filter	**No value set**
Quality of Service(QoS)	**0 - At most once**
Max Queue Size	**No value set**
Record Reader	No value set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

Listen FTP

Let Apache NiFi be your FTP server



No More Spaghetti Flows - DO NOT

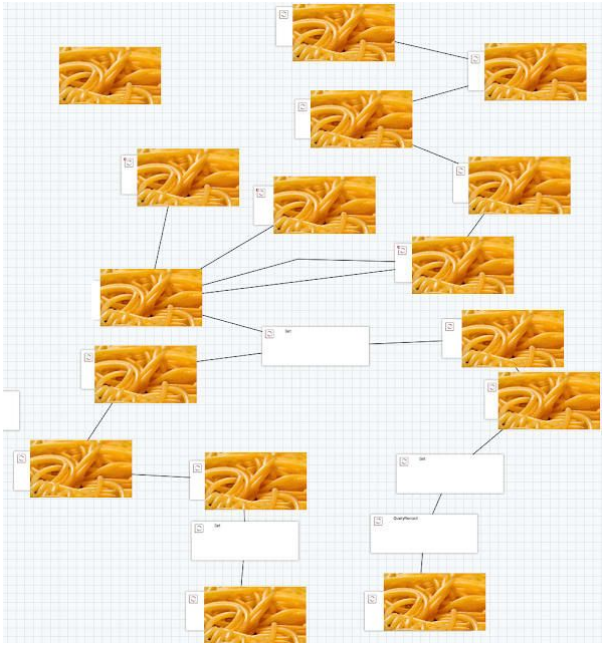


https://dev.to/tspannhw/no-more-spaghetti-flows-2emd

Do Not

- Do not Put 1,000 Flows on one workspace.
- If your flow has hundreds of steps, this is a Flow Smell. Investigate why.
- Do not Use ExecuteProcess, ExecuteScripts or a lot of Groovy scripts as a default, look for existing processors
- Do not Use Random Custom Processors you find that have no documentation or are unknown.
- Do not forget to upgrade, if you are running anything before Apache NiFi 1.14, upgrade now!
- Do not run on default 512M RAM.
- Do not run one node and think you have a highly available cluster.
- Do not split a file with millions of records to individual records in one shot without checking available space/memory and back pressure.
- Use Split processors only as an absolute last resort. Many processors are designed to work on FlowFiles that contain many records or many lines of text. Keeping the FlowFiles together instead of splitting them apart can often yield performance that is improved by 1-2 orders of magnitude.

No More Spaghetti Flows - DO



https://dev.to/tspannhw/no-more-spaghetti-flows-2emd

Do

- Reduce, Reuse, Recycle. Use Parameters to reuse common modules.
- Put flows, reusable chunks (write to Slack, Database, Kafka) into separate Process Groups.
- Write custom processors if you need new or specialized features
- Use Record Processors everywhere
- Read the Docs!
- Use the NiFi Registry for version control.
- Use NiFi CLI and DevOps for Migrations.
- Walk through your flow and make sure you understand every step and it’s easy to read and follow. Is every processor used? Are there dead ends?
- Do run Zookeeper on different nodes from Apache NiFi.
- Use routing based on content and attributes to allow one flow to handle multiple nearly identical flows is better than deploying the same flow many times with tweaks to parameters in same cluster.
- Use the correct driver for your database. There's usually a couple different JDBC drivers.

Apache MXNet Native Processor through DJL.AI for Apache NiFi

DJL

#workshop

11:30 AM ==========
Deep Learning Class Label: person
File: cc0a469f-c108-42c7-95c6-10e5fda95006.person.png
Probability: 0.96
UUID: 32ef65a3-0650-42cd-965c-ba25597eb1ad
Rank: 1

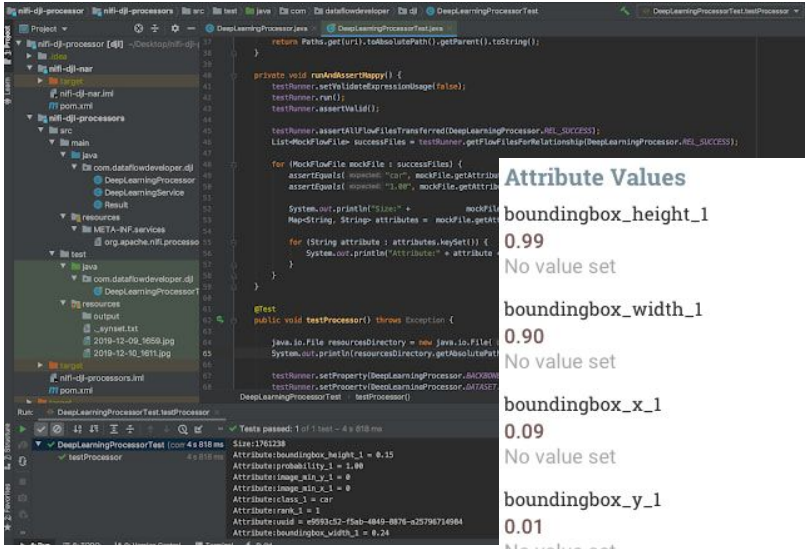
Bounding Box (Height/Width, X,Y)
0.74/ 0.69

0.27, 0.25

Image (Height/Width, X,Y)
480 / 640

0, 0
==========

tspann 11:30 AM
371bdb8f-35bc-4a2a-919c-bdeb609b726c.person.png



Attribute Values

boundingbox_height_1
0.99
No value set

boundingbox_width_1
0.90
No value set

boundingbox_x_1
0.09
No value set

boundingbox_y_1
0.01
No value set

class_1
tvmonitor
No value set

filename
2020-08-26_1330.jpg.tvmonitor.png
2020-08-26_1330.jpg (previous)

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

PULSAR

StreamNative

What are the Benefits of Pulsar?

Multi-Tenancy

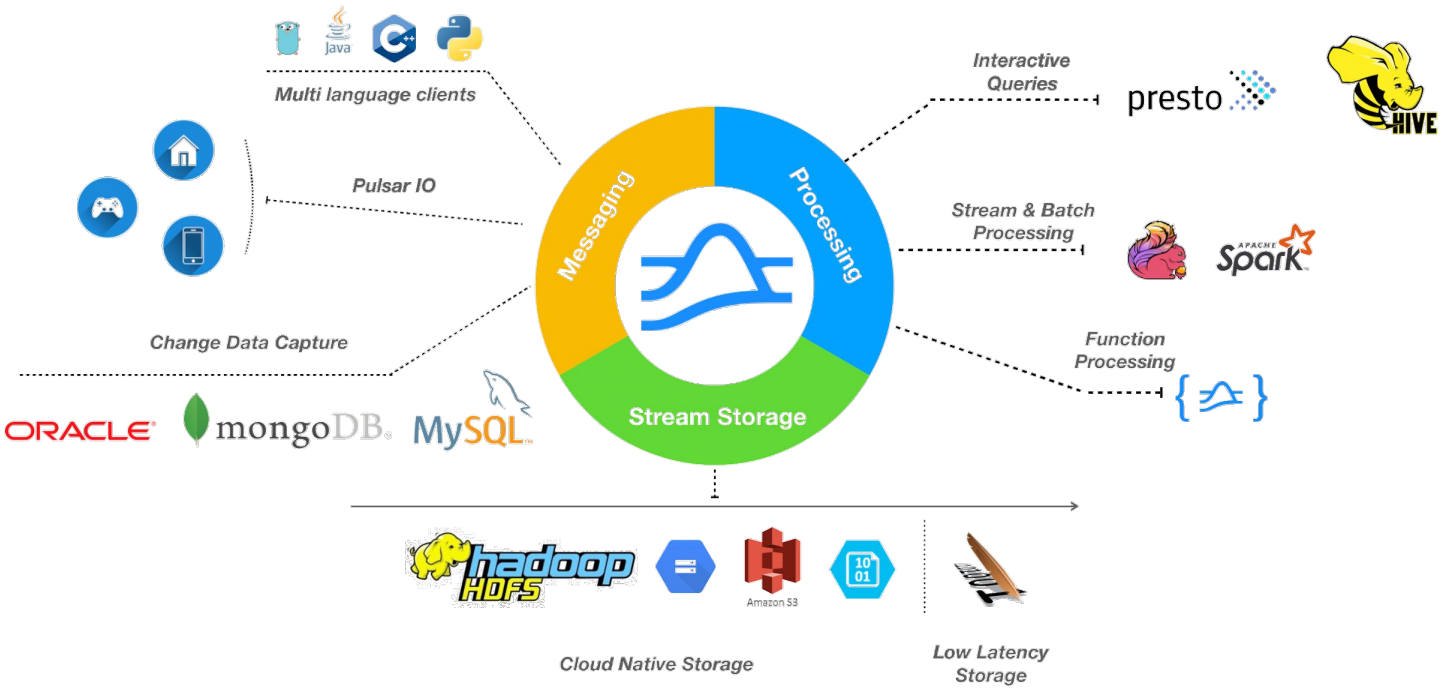
Scalability

Geo-Replication

Unified Messaging Model

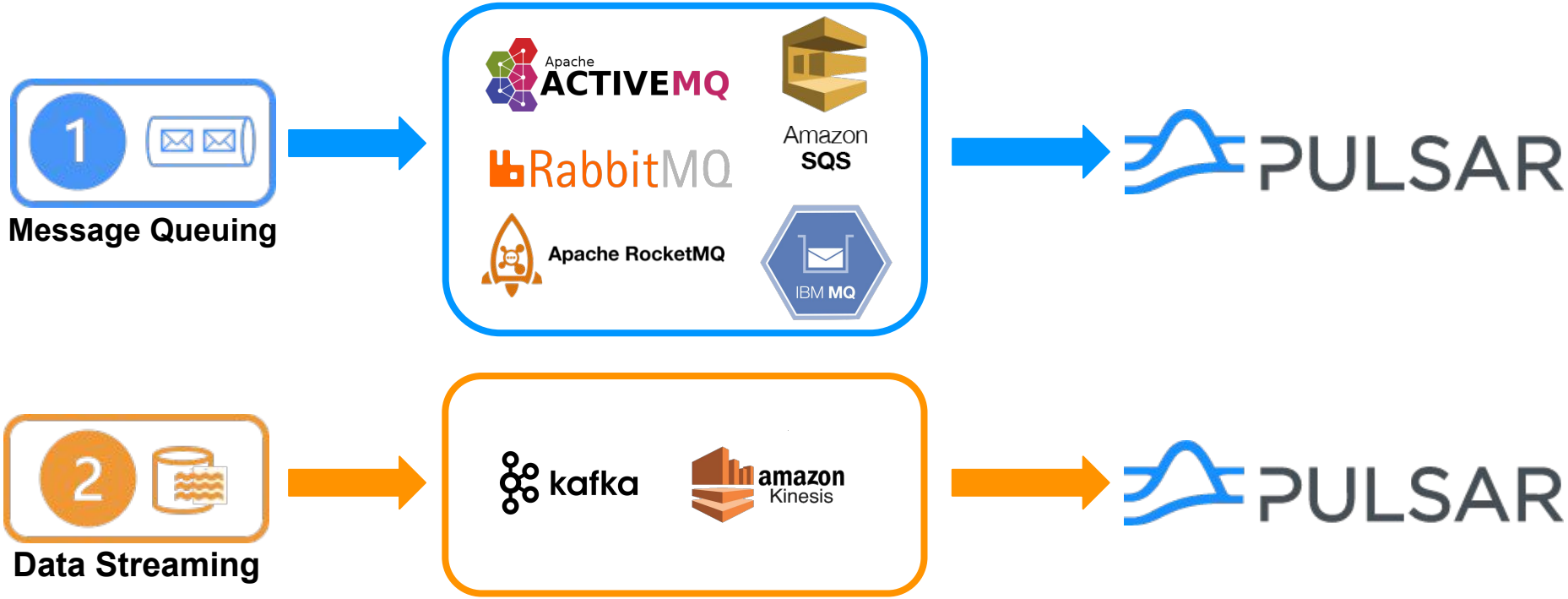
Data Durability

Apache Pulsar

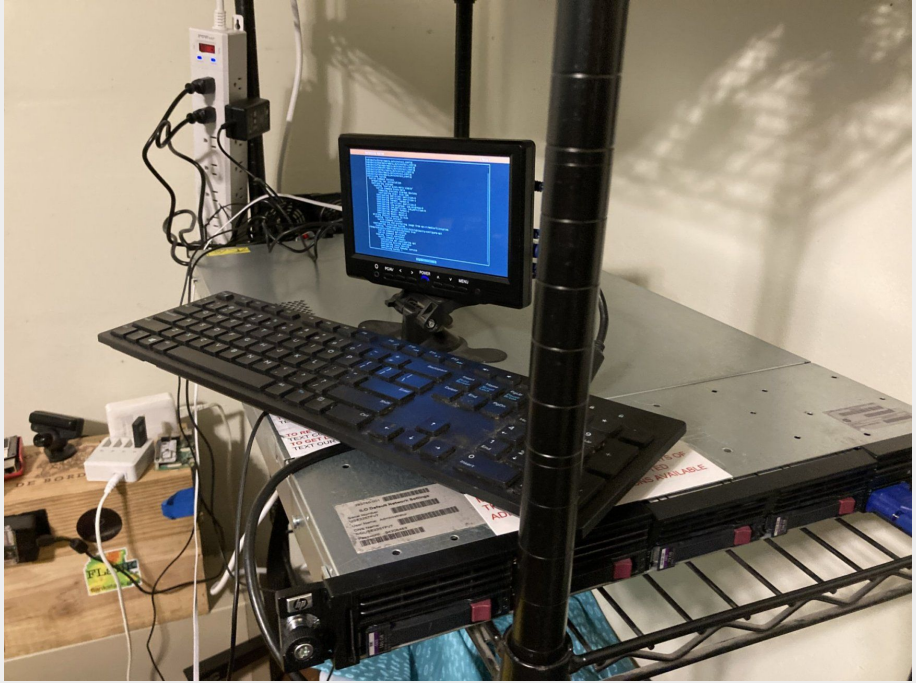


StreamNative

A Unified Messaging Platform



Demo



StreamNative

Wrap-Up

StreamNative

Stream Native

Founded by the original developers of Apache Pulsar and Apache BookKeeper, StreamNative builds a cloud-native event streaming platform that enables enterprises to easily access data as real-time event streams.

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

Platform Engineer [Remote]	San Francisco
Platform Engineer (Flink/Spark) [Remote]	San Francisco
Product Engineer - Cloud [Remote]	San Francisco
Platform Engineer (Flink/Spark) [Remote]	San Francisco
Product Engineer - Cloud [Remote]	San Francisco
Sr. Product Manager [Remote]	San Francisco

We're Hiring

streamnative.io/careers/

Let’s Keep in Touch!

Speaker Name

Speaker title

@PassDev

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

https://github.com/tspannhw

StreamNative