

Apache NiFi 101

Timothy Spann

Developer Advocate / StreamNative

Thursday 23-Sept-2021



Agenda

Thursday 14:10 UTC

Apache NiFi 101: Introduction and Best Practices

Timothy Spann

In this talk, we will walk step by step through Apache NiFi from the first load to first application. I will include slides, articles and examples to take away as a Quick Start to utilizing Apache NiFi in your real-time dataflows. I will help you get up and running locally on your laptop, Docker or in CDP Public Cloud.

Speaker Bio

Timothy Spann

Developer Advocate @  **Stream
Native**

DZone Zone Leader and Big Data MVB
@PaasDev

<https://github.com/tspannhw>

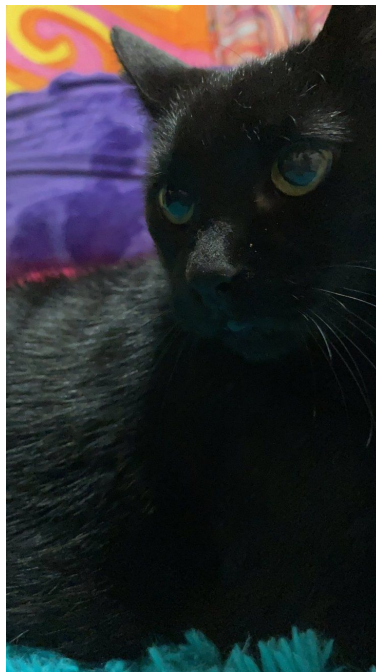
<https://www.datainmotion.dev/>

<https://github.com/tspannhw/SpeakerProfile>

<https://dev.to/tspannhw>

<https://sessionize.com/tspann/>

<https://www.slideshare.net/bunkertor>



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 Solution

APP Layer

Application Messaging



Micro Service

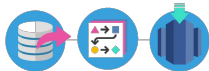


Payment



Notification

Data Pipelines



ETL

Real-time Contextual Analytics



Dashboard



Risk Control



Auditing

StreamNative Platform

Computing Layer



Storage Layer



Apache BookKeeper™

Tiered Storage



Stream Native Platform

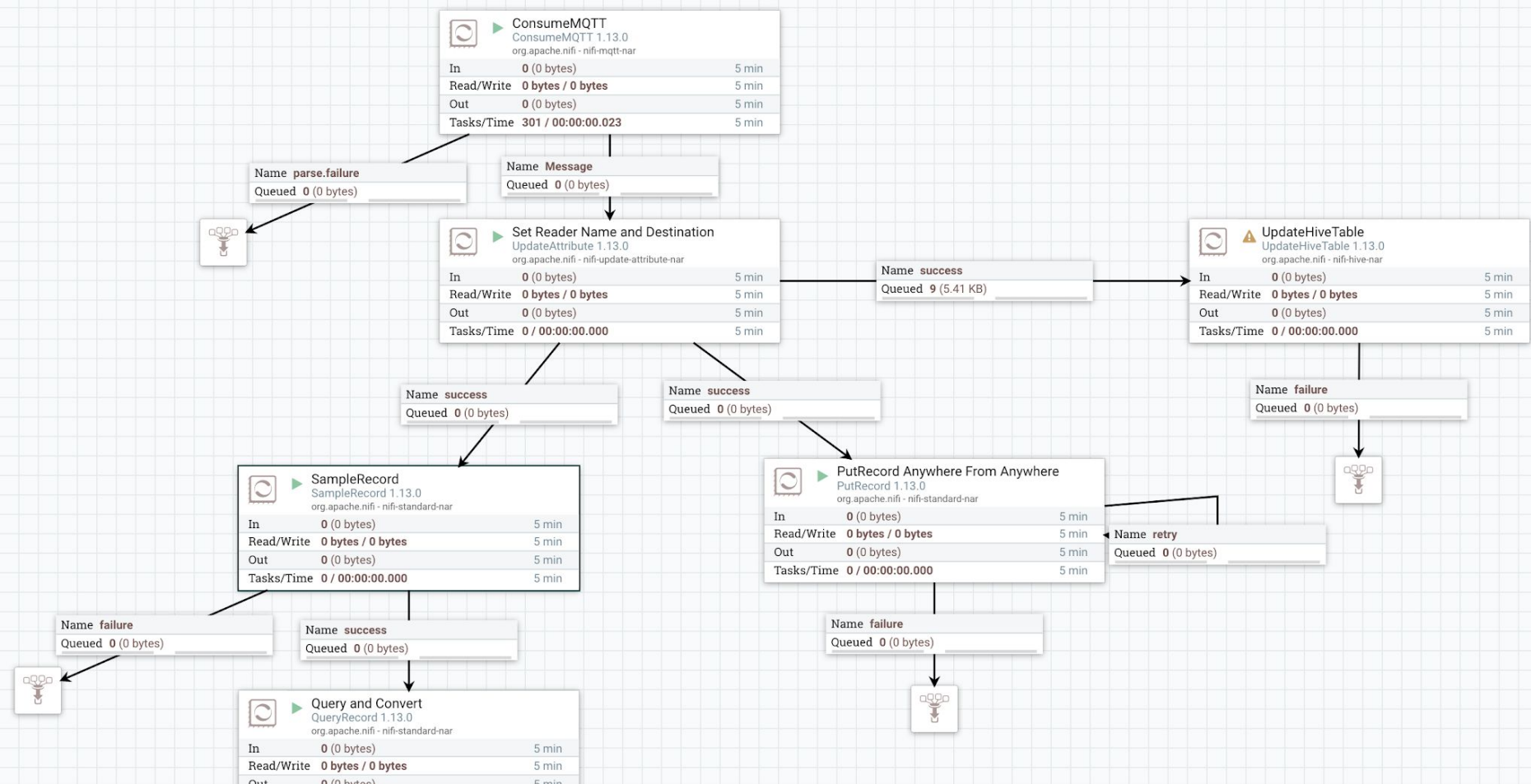
IaaS Layer



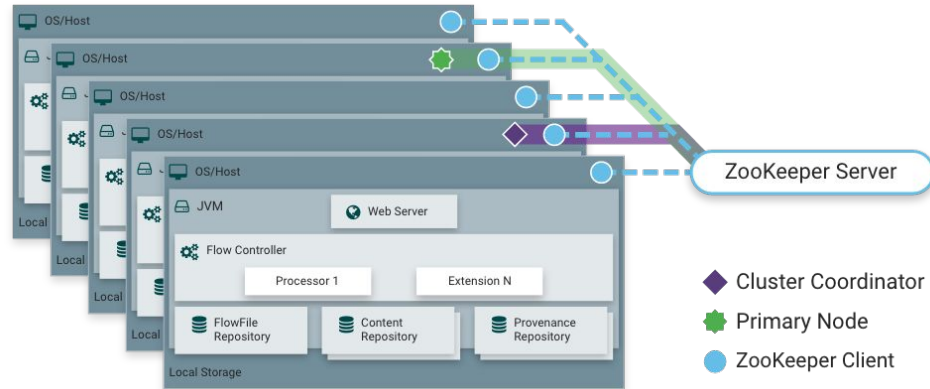
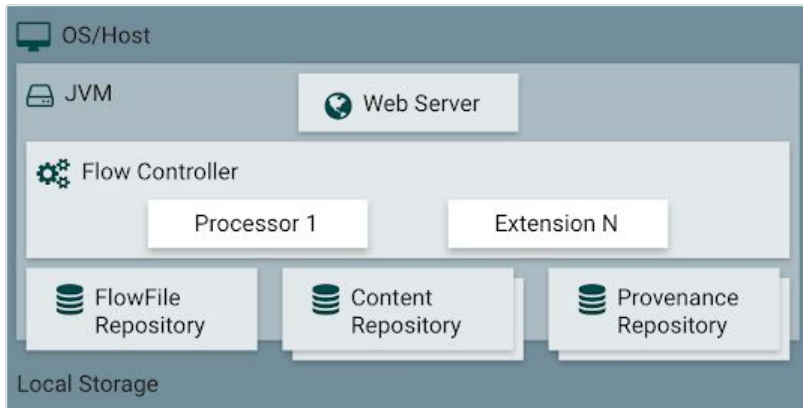
kubernetes



Stream Native Cloud



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>



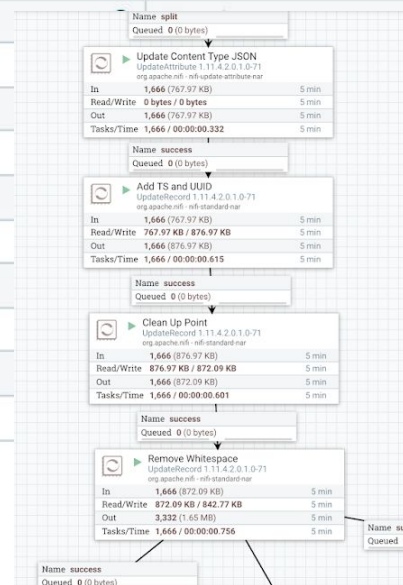
Record Processors



- 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



Configure Processor

Invalid

SETTINGS SCHEMAS

Required field

Property

Record Reader

Record Writer

Include Zero Record

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	?

WindowsEventLogReader 1.13.0

Required field

Property

Record Reader

Record Destination Service

Include Zero Record

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>

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	6f8bce4f-6ba4-47c9-ba83-8830ec7cd3db	89 byt
12/22/2020 16:54:17:192 UTC	ATTRIBUTES_MODIFIED	1213e84d-e848-4218-b3d0-c2588e7f991e	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-9d71-4cf6-8f73-b5b46ee03622	83 byt
12/22/2020 16:53:59:296 UTC	ATTRIBUTES_MODIFIED	d143c05c-5aae-44c2-9eda-c620a8148604	84 byt
12/22/2020 16:53:59:295 UTC	ATTRIBUTES_MODIFIED	4b1dbb1e-1f83-4a9b-309-3da1277edc6	84 byt
12/22/2020 16:53:58:296 UTC	ATTRIBUTES_MODIFIED	45f8edd4-ca55-431e-82b9-436c4af4052e	81 byt
12/22/2020 16:53:57:298 UTC	ATTRIBUTES_MODIFIED	bb07034b-63e1-4b34-b0e1-4b34-b0e1	
12/22/2020 16:53:57:297 UTC	ATTRIBUTES_MODIFIED	d1a2b81e-a7af-4e16-b0e1-4b34-b0e1	
12/22/2020 16:53:57:297 UTC	ATTRIBUTES_MODIFIED	2f9966d0-4153-41bc-b0e1-4b34-b0e1	
12/22/2020 16:53:43:753 UTC	ATTRIBUTES_MODIFIED	1ca5c744-1cb4-4ff1-b0e1-4b34-b0e1	
12/22/2020 16:53:37:747 UTC	ATTRIBUTES_MODIFIED	fa5647db-96db-48cd-b0e1-4b34-b0e1	
12/22/2020 16:53:21:446 UTC	ATTRIBUTES_MODIFIED	dcf1609b-9648-460e-b0e1-4b34-b0e1	
12/22/2020 16:53:05:515 UTC	ATTRIBUTES_MODIFIED	964a05fc-d958-460e-b0e1-4b34-b0e1	
12/22/2020 16:52:43:374 UTC	ATTRIBUTES_MODIFIED	79fcb9b0-b16b-4fc4-8a-b0e1-4b34-b0e1	
12/22/2020 16:52:29:308 UTC	ATTRIBUTES_MODIFIED	3433eeb3-953c-4952-a0e1-4b34-b0e1	
12/22/2020 16:52:29:307 UTC	ATTRIBUTES_MODIFIED	a166c9a7-11ba-4262-90e1-4b34-b0e1	
12/22/2020 16:52:29:307 UTC	ATTRIBUTES_MODIFIED	bd2946f9-5a89-40a7-b0e1-4b34-b0e1	
12/22/2020 16:52:29:307 UTC	ATTRIBUTES_MODIFIED	a16841bc-2505-4c8c-b0e1-4b34-b0e1	
12/22/2020 16:52:29:306 UTC	ATTRIBUTES_MODIFIED	578540fe-c446-471f-b0e1-4b34-b0e1	
12/22/2020 16:52:29:306 UTC	ATTRIBUTES_MODIFIED	3d44c5f8-a737-4a8e-b0e1-4b34-b0e1	
12/22/2020 16:52:29:306 UTC	ATTRIBUTES_MODIFIED	48c93617-705f-424e-90e1-4b34-b0e1	
12/22/2020 16:52:29:306 UTC	ATTRIBUTES_MODIFIED	9f1b9cd1-f384-4c11-90e1-4b34-b0e1	
12/22/2020 16:52:29:305 UTC	ATTRIBUTES_MODIFIED	7b2a18b0-a2d8-47ab-b0e1-4b34-b0e1	

DETAILS

ATTRIBUTES

Attribute Values

isprince

123.66

No value set

symbol

IBM

No value set

timestamp

1608565962884

No value set

volume

100

No value set

Provenance Event

DETAILS

ATTRIBUTES

Attribute Values

lastprice

123.66

No value set

symbol

IBM

No value set

timestamp

1608654962884

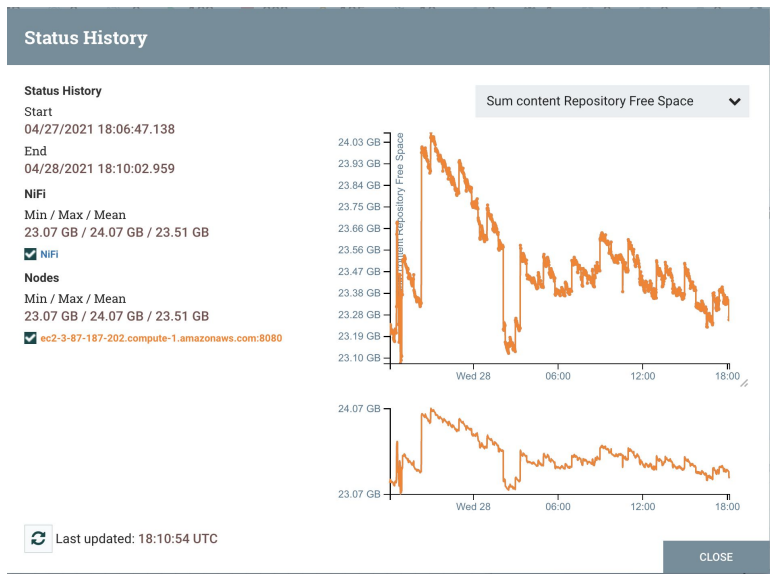
No value set

volume

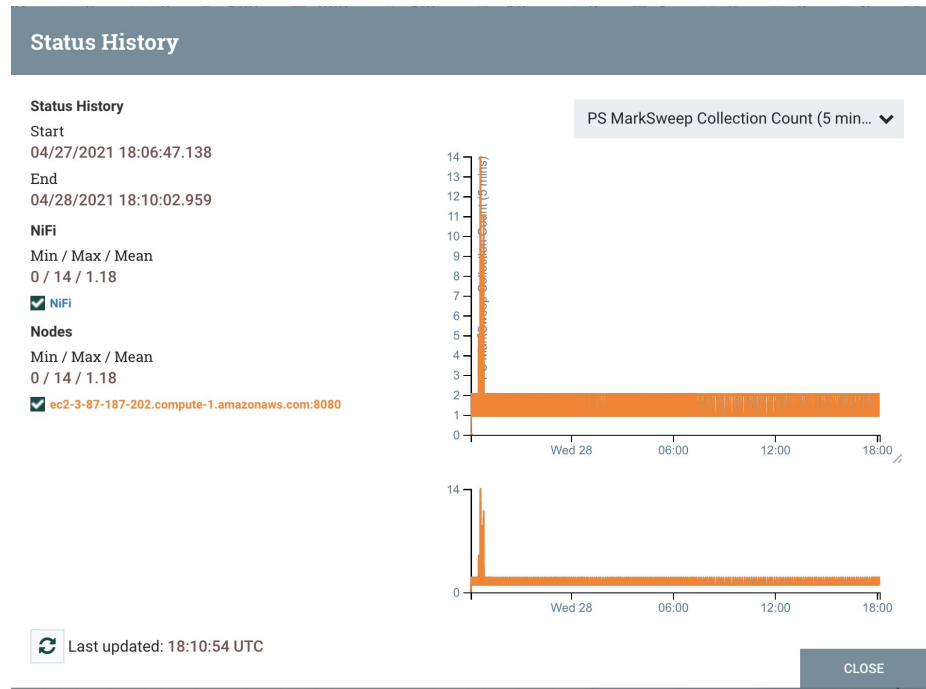
100

No value set

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)

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

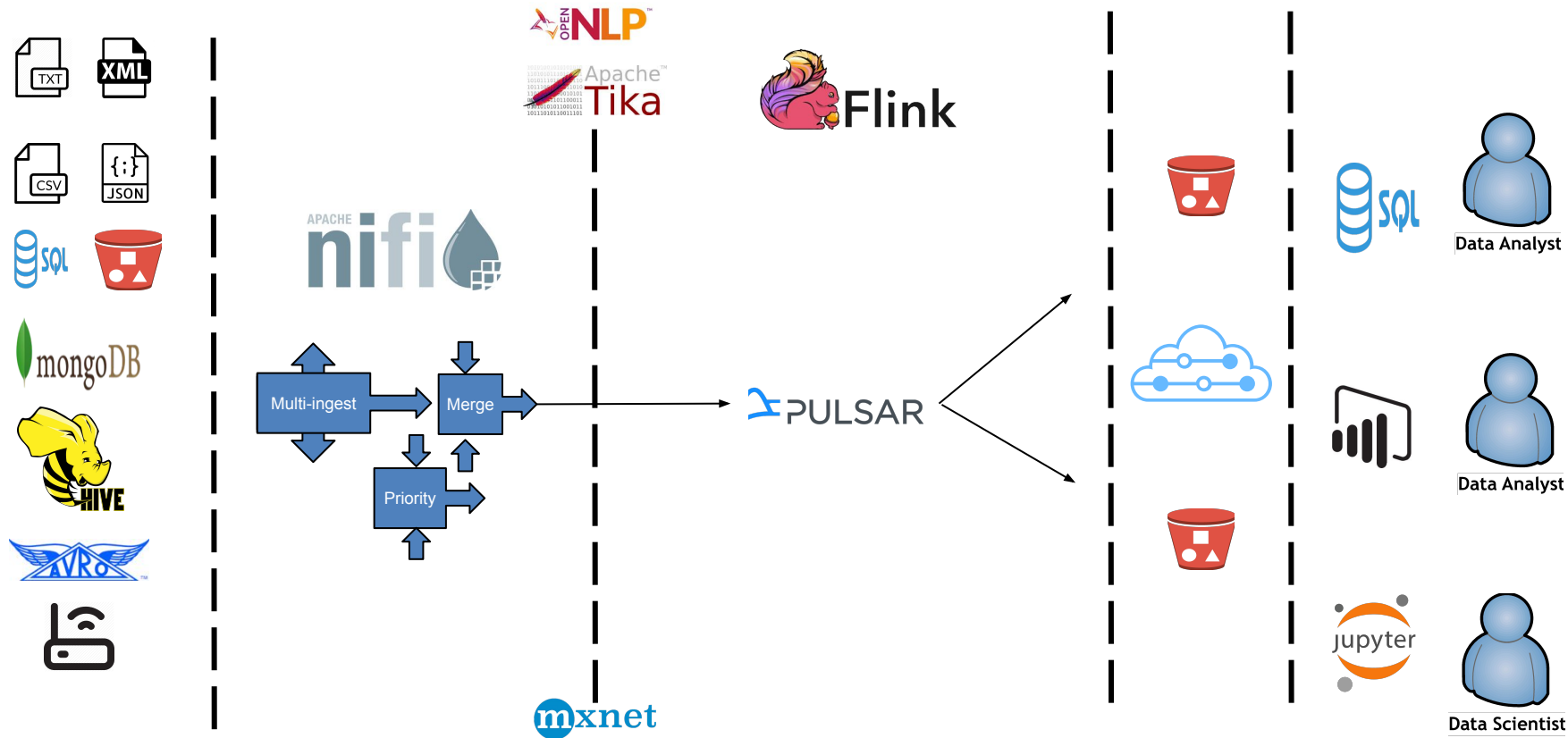
Listen FTP



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

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



Deeper Content



@PaasDev

- <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://pierrebillard.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://datainmotion.dev/>



timothyspann

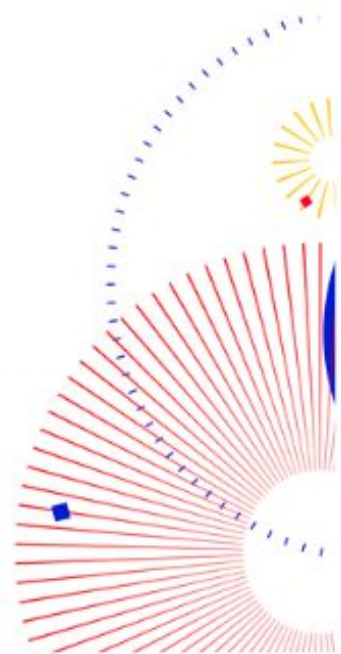


Pulsar Summit Europe

October 6, 2021

Pulsar Summit Asia

November 20-21, 2021



Contact us at partners@pulsar-summit.org to become a sponsor or partner