

Apache NiFi

21st century open-source data flows

Frank Thiele

München, 2018-05-18



Content

Introduction

Technology

Functionality

Live demo

Summary



Apache NiFi

<https://pxhere.com/en/photo/516183>

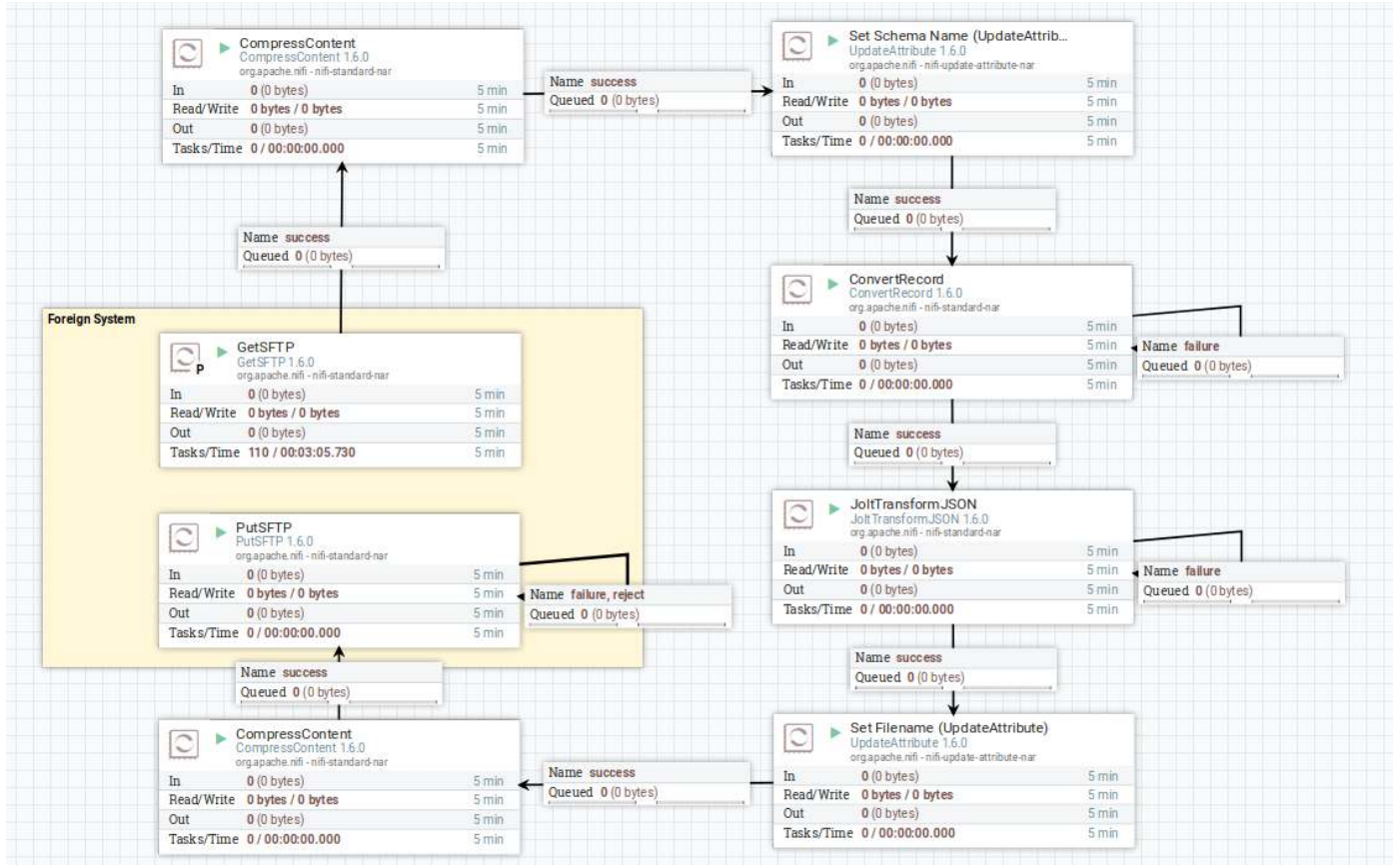
Apache NiFi

- Based on NSA project NiagaraFiles
- Automation of data flows between applications
- Available under Apache License since 2014
- Development taken over by Hortonworks (2015)
- Release v1.0.0, August 2016
- Latest v1.6.0, April 2018

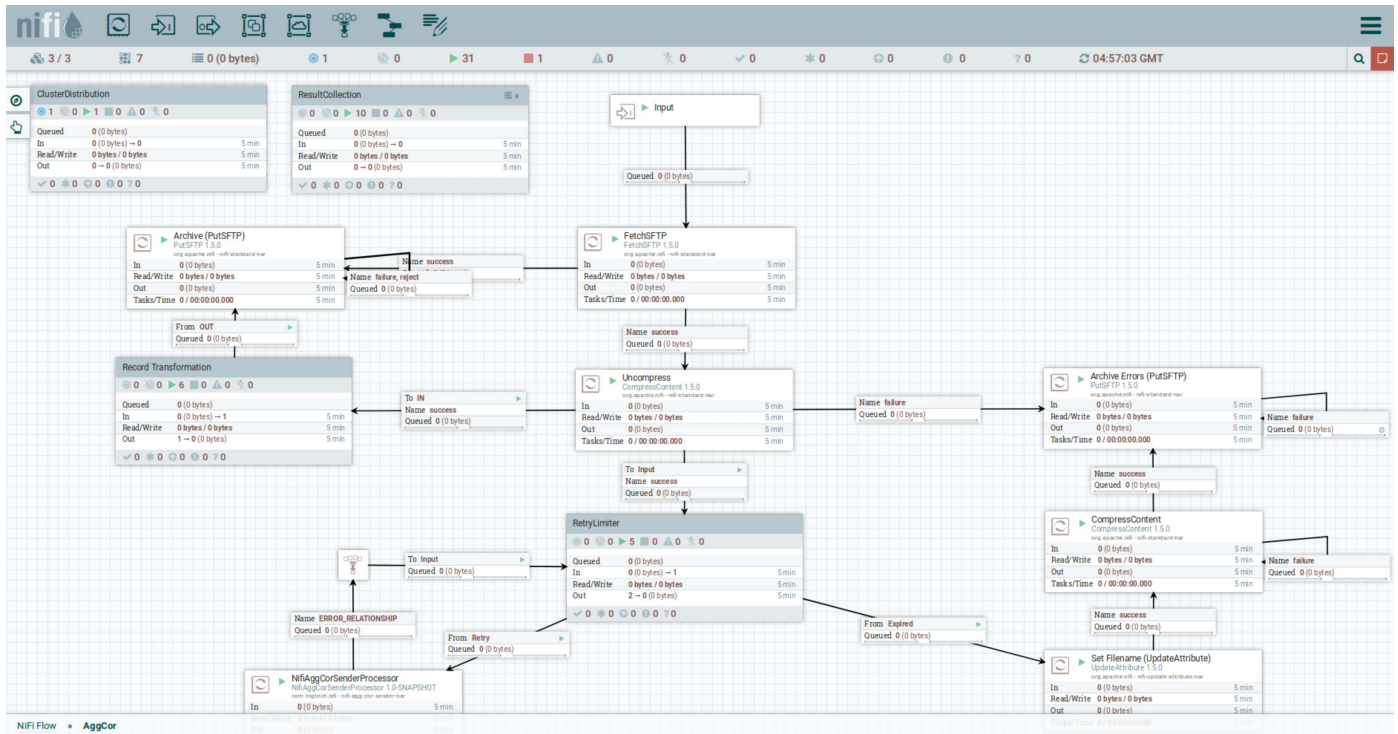
Apache NiFi – What is it?

- ...comparable to Ab Initio or Talend Data Integration
- ...or Apache Camel
- **Data Flow Management Platform**
- ...more than a classical ETL Framework

It models your data flow



... and more complex data flows



Apache NiFi – Use cases

- Workflow modeling with data flows
- Connects different technology stacks
- Mass data processing
- Centralization of complex data flows
- Accountability of data flows
- IOT, Telecommunication, Banking, ...
- ETL – Extract, Transform, Load



Highlights

- UI based flow modeling
- Provenance: Extended tracing of data
- Low latency vs. high throughput
- Queueing and QoS
- Simple horizontal and vertical scaling
- Transactional
- Security (Role concept, SSL/TLS, etc.)

<https://pxhere.com/en/photo/284658>

Content

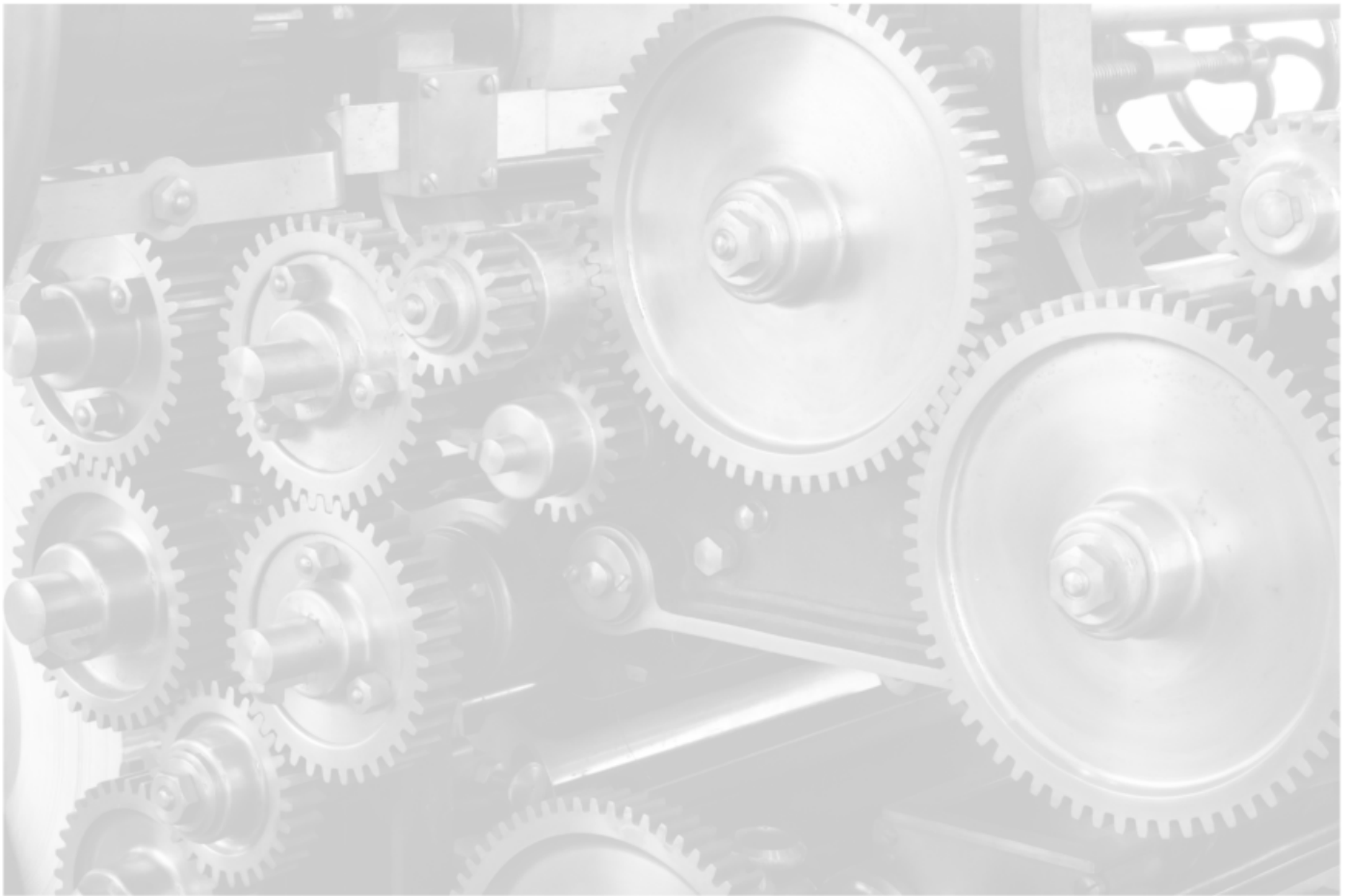
Introduction

Technology

Functionality

Live demo

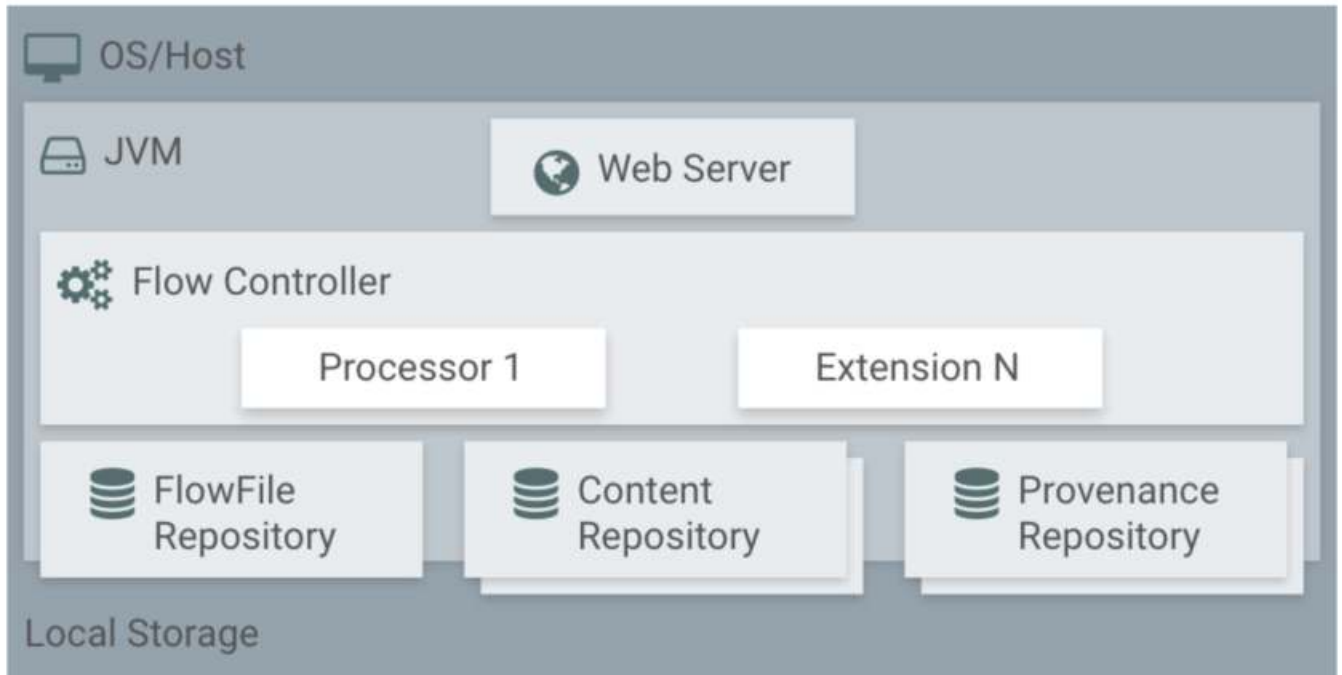
Summary



Technology

- Java 8
- Integration with current technologies, e.g.
 - Apache Software Foundation
 - HBase, Cassandra
 - Kafka, Ignite
 - Spark
 - Druid
 - Atlas
 - Couchbase, MongoDB, InfluxDB

Nifi Architecture

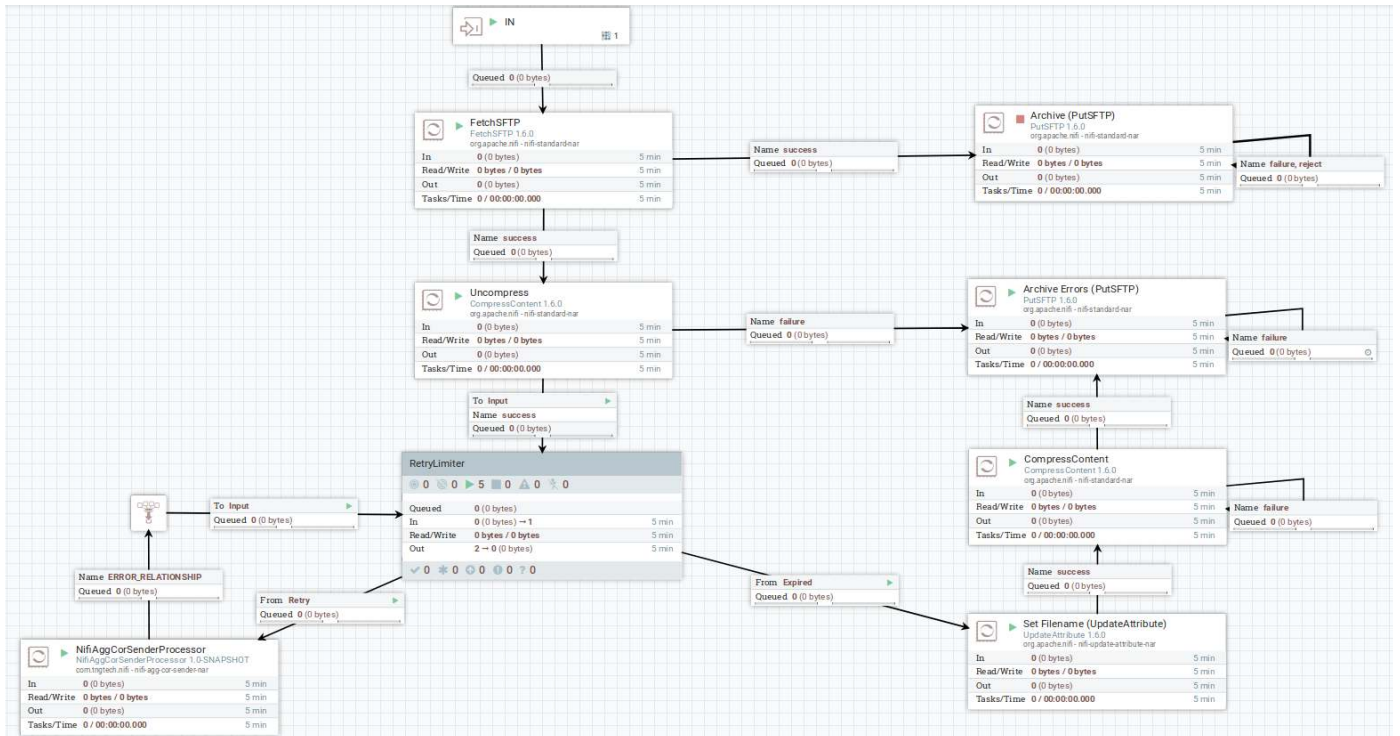


<https://nifi.apache.org/docs/nifi-docs/html/images/zero-master-node.png>

Concepts

- Flow-based programming
- FlowFiles, processors, controllers, groups and connections
- Pass-by-reference between processors
- Dedicated repositories as storage
- Happy and error paths are modeled equivalently

Flow-based programming



FlowFiles

- The data bucket of NiFi
- Meta data container
 - UUID
 - Attributes
 - Content link
- Best Practice: Modify and read attributes, not content
- Transferred from one processor to another – by reference



Repositories

<https://pxhere.com/en/photo/516183>

Repositories – FlowFile

- Storage for meta data
- Write-ahead log for data consistency/persistence
- Snapshots/Checkpoints for restoring
- Copy-on-write
- Swapping supported

Repositories – Content

- Storage for actual data of a FlowFile
- Can be scaled over partitions and cluster nodes
- Content is immutable (write-once, copy-on-write)

Repositories – Provenance

- Storage for history of the FlowFiles
- Can be scaled over partitions and cluster nodes
- Any FlowFile can be viewed at any point in time
- Apache Lucene Index

Content

Introduction

Technology

Functionality

Live demo

Summary



Integration

- Web-based UI for monitoring and configuration
- REST API
- Batch vs. event processing
- Automatic zero-master clustering
- Apache Ambari integration
- Commodity hardware

<https://pxhere.com/en/photo/1270987>


Flexibility

- High number of available processors (ca. 250)
- Freely programmable processors for customizations
- Many programming languages supported in addition to Java (e.g. Python, Lua, JS)
- Template support
- Configuration changes of the stream without downtime
- NiFi Registry for configuration management

NiFi Registry

- Since v1.5, so January 2018
- Configuration management

NiFi Registry UI

 NiFi Registry / test ▾ / All ▾ anonymous ? ⚙

Sort by: Name (a - z) ▾

AggCorFlow - Test Flow

VERSIONS
2

⬆

DESCRIPTION

No description specified

CHANGE LOG

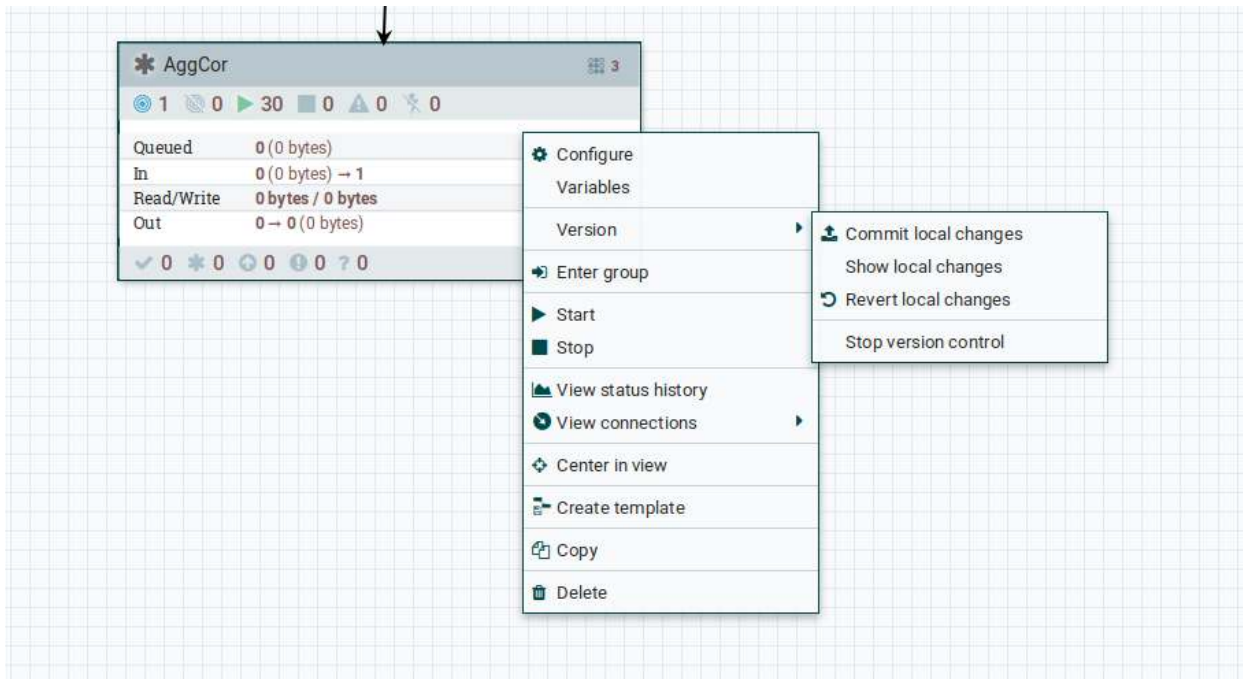
Version 2 - 2 minutes ago
by anonymous

Adapted create directory flag
May-10-2018 at 12:47 PM

Version 1 - a day ago
by anonymous

ACTIONS ▾

NiFi Registry integration



Further advantages

- Open source (active development by Hortonworks)
- Paradigm shift from file and content to events and attributes
- Advanced backpressure control
- Multi-tenancy
- Unit and integration tests
- Reduced I/O overhead

Content

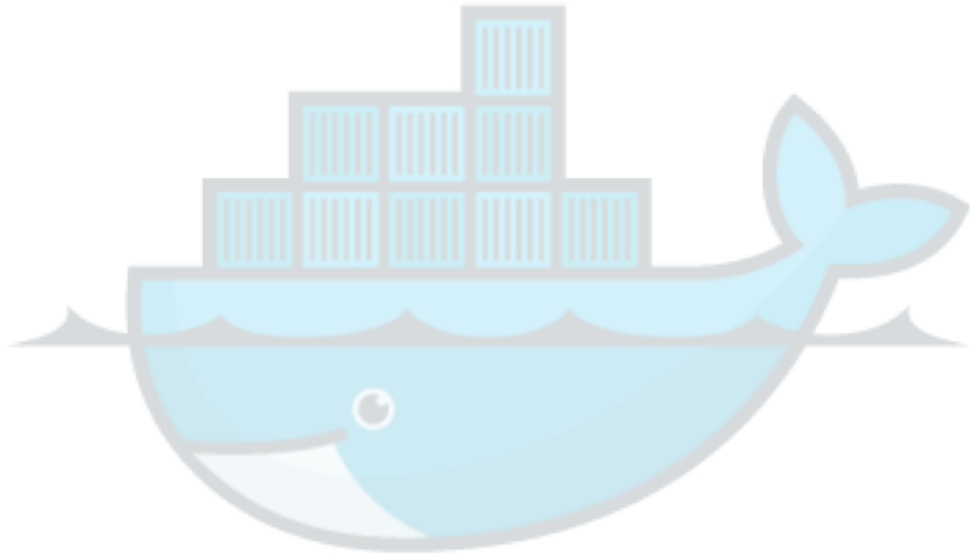
Introduction

Technology

Functionality

Live demo

Summary



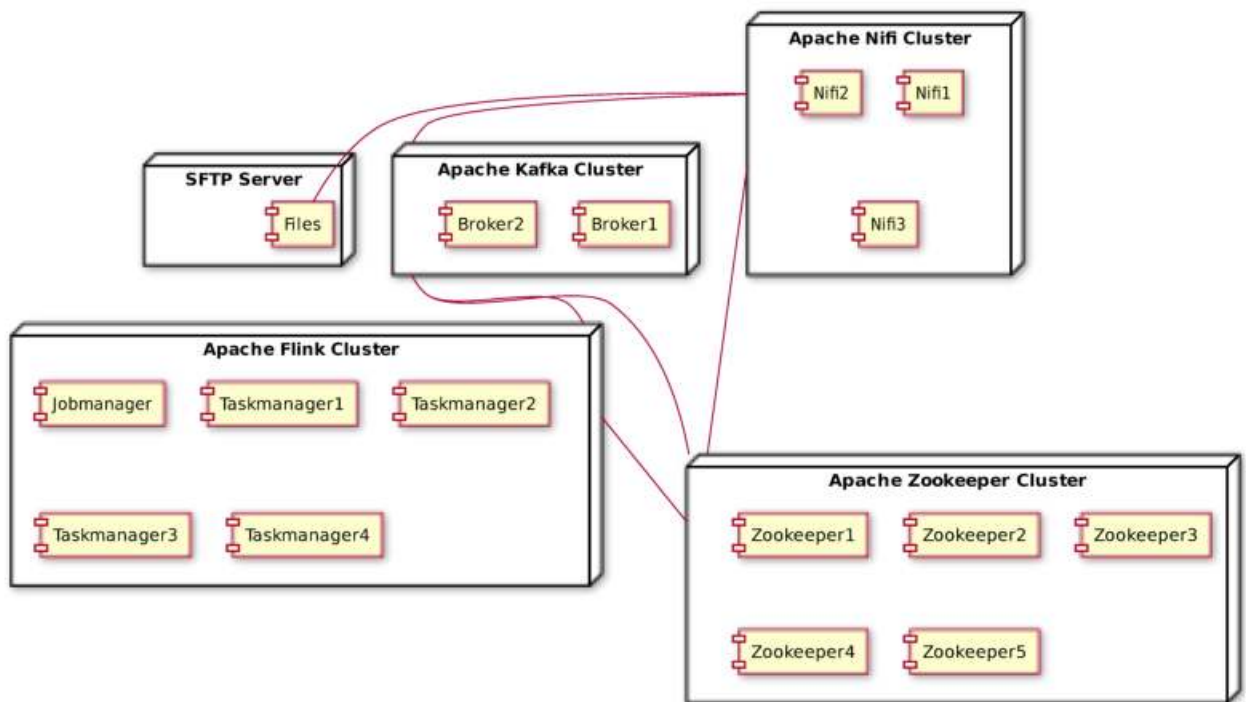
docker

Docker

- Dockerized with Docker Swarm
- Adapted and linked configuration files
- NAR files linked, too

<https://www.flickr.com/photos/xmodulo/14098888813>

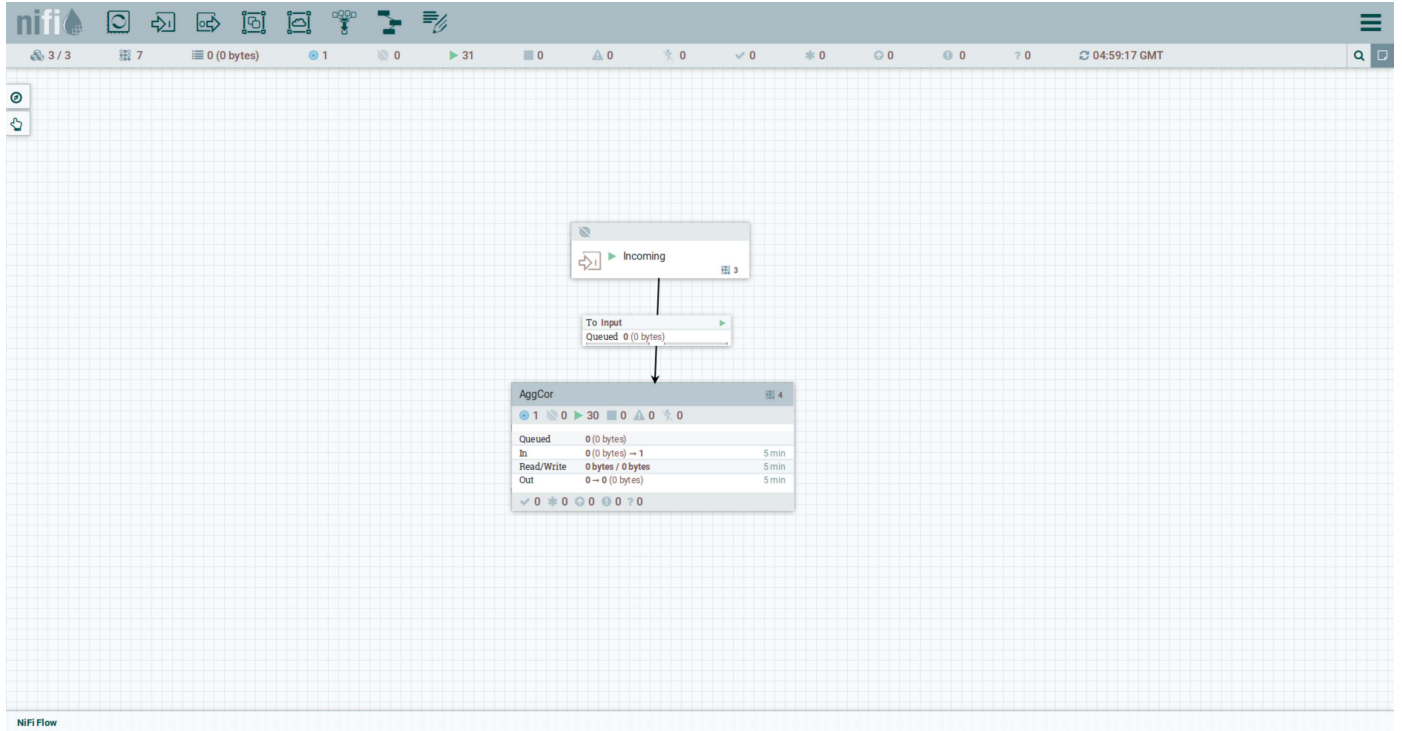
Setup



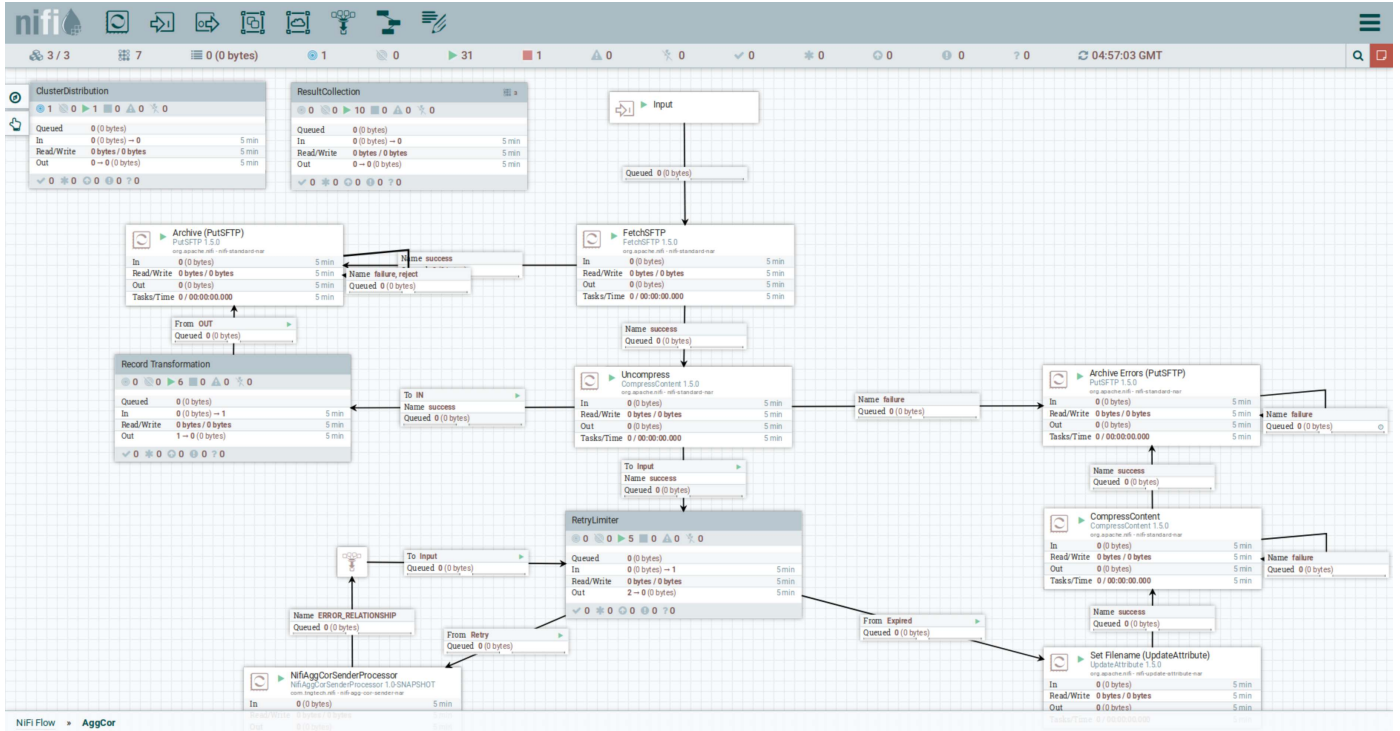
Live demo

- UI Overview
- Stream example
- Feature examples
 - Data provenance
 - Back pressure
 - Error handling
 - ETL with Jolt

UI overview



Stream example



Processor configuration

The screenshot shows the Apache NiFi web interface with a 'Processor Details' dialog box open for configuring a 'Set Filename' processor. The dialog has four tabs: SETTINGS, SCHEDULING, PROPERTIES, and COMMENTS. The PROPERTIES tab is active, displaying a table of properties. Below the table are 'ADVANCED' and 'OK' buttons. In the background, a flow diagram is visible, showing an 'IN' connector, a 'Name success' processor, and a 'Set Filename (UpdateAttribute)' processor.

Processor Details

SETTINGS SCHEDULING **PROPERTIES** COMMENTS

Required field

Property	Value
Jolt Transformation DSL	Chain
Custom Transformation Class Name	No value set
Custom Module Directory	No value set
Jolt Specification	[{"operation": "modify-default-beta", "spec": {"*": {"concatEX...
Transform Cache Size	1

ADVANCED OK

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

Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

Set Filename (UpdateAttribute)
UpdateAttribute 1.5 0

NiFi Flow » AggCor » Record Transformation

Data provenance

NiFi Data Provenance

Displaying 1,000 of 1,000

Oldest event available: 01/17/2018 04:30:25 GMT

Showing the most recent 1,000 of 1,000+ events, please refine the search.

Filterby component name

Date/Time	Type	FlowFile UUID	Size	Component Name	Component Type	Node	
01/17/2018 05:04:28.062 GMT	DROP	493f1f48-628f-44c1-94fa-d8e6eb86db61	71 bytes	Archive Errors (PutSFTP)	PutSFTP	nifi3.8080	
01/17/2018 05:04:28.062 GMT	SEND	493f1f48-628f-44c1-94fa-d8e6eb86db61	71 bytes	Archive Errors (PutSFTP)	PutSFTP	nifi3.8080	
01/17/2018 05:04:27.543 GMT	DROP	9d43812c-1659-495f-8798-e3c3b87ad16d	104 bytes	Archive (PutSFTP)	PutSFTP	nifi3.8080	
01/17/2018 05:04:27.543 GMT	SEND	9d43812c-1659-495f-8798-e3c3b87ad16d	104 bytes	Archive (PutSFTP)	PutSFTP	nifi3.8080	
01/17/2018 05:04:27.541 GMT	DROP	30ffe81d-7a24-4d40-8f21-0ee4c1f99ba2	73 bytes	Archive (PutSFTP)	PutSFTP	nifi3.8080	
01/17/2018 05:04:27.541 GMT	SEND	30ffe81d-7a24-4d40-8f21-0ee4c1f99ba2	73 bytes	Archive (PutSFTP)	PutSFTP	nifi3.8080	
01/17/2018 05:04:27.349 GMT	CONTENT_MODIFIED	493f1f48-628f-44c1-94fa-d8e6eb86db61	71 bytes	CompressContent	CompressContent	nifi3.8080	
01/17/2018 05:04:27.339 GMT	ATTRIBUTES_MODIFIED	493f1f48-628f-44c1-94fa-d8e6eb86db61	71 bytes	Set Filename (UpdateAttribute)	UpdateAttribute	nifi3.8080	
01/17/2018 05:04:27.322 GMT	ROUTE	493f1f48-628f-44c1-94fa-d8e6eb86db61	71 bytes	RouteOnAttribute	RouteOnAttribute	nifi3.8080	
01/17/2018 05:04:27.313 GMT	ATTRIBUTES_MODIFIED	493f1f48-628f-44c1-94fa-d8e6eb86db61	71 bytes	UpdateAttribute	UpdateAttribute	nifi3.8080	
01/17/2018 05:04:27.304 GMT	DROP	3b5f4428-04b1-483c-a87-04fa911eb19	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.304 GMT	FORK	3b5f4428-04b1-483c-a87-04fa911eb19	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.304 GMT	SEND	3b5f4428-04b1-483c-a87-04fa911eb19	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.290 GMT	ROUTE	3b5f4428-04b1-483c-a87-04fa911eb19	71 bytes	RouteOnAttribute	RouteOnAttribute	nifi3.8080	
01/17/2018 05:04:27.281 GMT	ATTRIBUTES_MODIFIED	3b5f4428-04b1-483c-a87-04fa911eb19	71 bytes	UpdateAttribute	UpdateAttribute	nifi3.8080	
01/17/2018 05:04:27.263 GMT	DROP	15fcd636-5905-489d-bab7-12134fb069dc	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.262 GMT	FORK	15fcd636-5905-489d-bab7-12134fb069dc	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.262 GMT	SEND	15fcd636-5905-489d-bab7-12134fb069dc	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.247 GMT	ROUTE	15fcd636-5905-489d-bab7-12134fb069dc	71 bytes	RouteOnAttribute	RouteOnAttribute	nifi3.8080	
01/17/2018 05:04:27.240 GMT	ATTRIBUTES_MODIFIED	15fcd636-5905-489d-bab7-12134fb069dc	71 bytes	UpdateAttribute	UpdateAttribute	nifi3.8080	
01/17/2018 05:04:27.232 GMT	DROP	b8f1a8a6-6415-427f-8e68-f950f7281941	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.232 GMT	FORK	b8f1a8a6-6415-427f-8e68-f950f7281941	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.232 GMT	SEND	b8f1a8a6-6415-427f-8e68-f950f7281941	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.215 GMT	ROUTE	b8f1a8a6-6415-427f-8e68-f950f7281941	71 bytes	RouteOnAttribute	RouteOnAttribute	nifi3.8080	
01/17/2018 05:04:27.209 GMT	ATTRIBUTES_MODIFIED	b8f1a8a6-6415-427f-8e68-f950f7281941	71 bytes	UpdateAttribute	UpdateAttribute	nifi3.8080	
01/17/2018 05:04:27.200 GMT	DROP	b29912f9-257b-4b58-a885-797e286da94b	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.200 GMT	FORK	b29912f9-257b-4b58-a885-797e286da94b	71 bytes	NifiAggCorSenderProcessor	NifiAggCorSenderProcessor	nifi3.8080	
01/17/2018 05:04:27.200 GMT	SEND	b29912f9-257b-4b58-a885-797e286da94b	71 bytes	NifiAaaCorSenderProcessor	NifiAaaCorSenderProcessor	nifi3.8080	

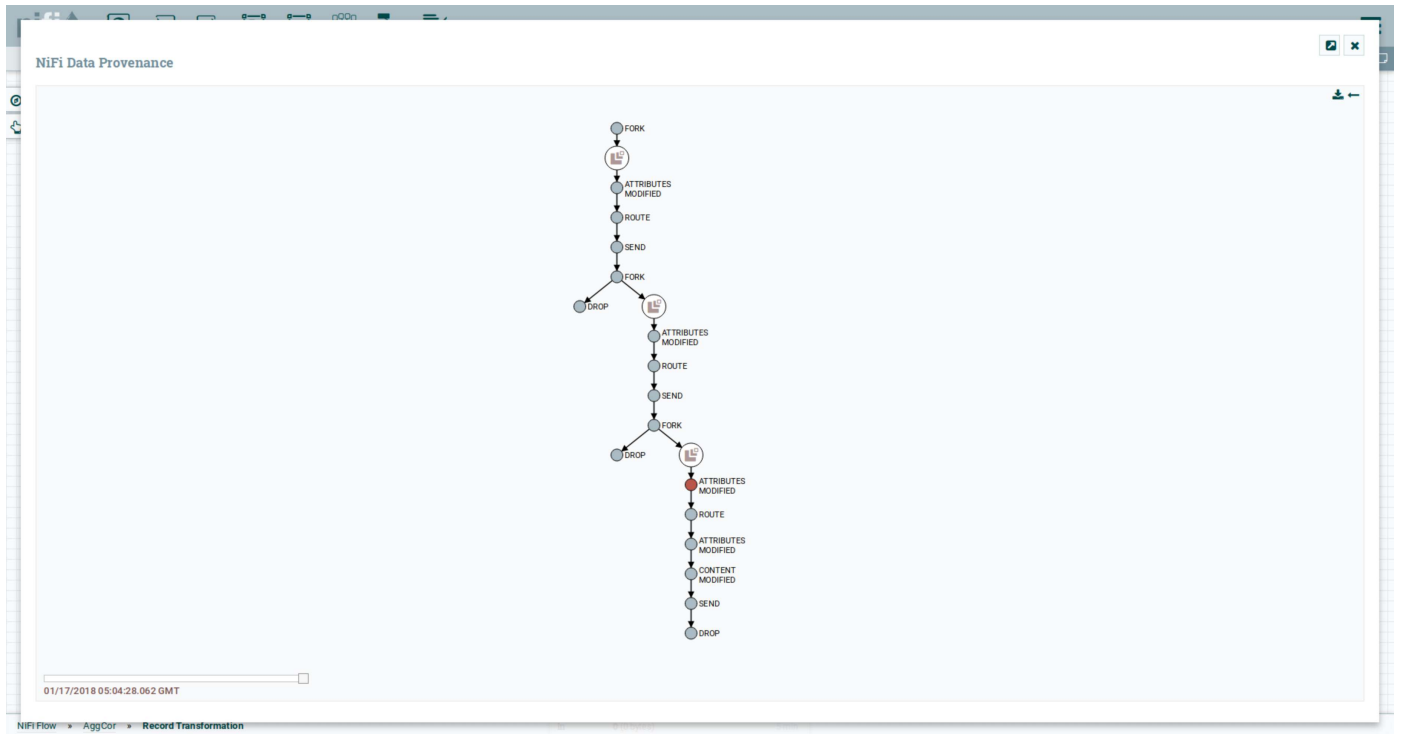
Last updated: 05:04:47 GMT

Nifi Flow

AggCor

Record Transformation

Data provenance – Graph



Data provenance – Details

The image shows the NiFi Data Provenance interface. A central dialog box titled "Provenance Event" is open, displaying details for a specific event. The dialog has three tabs: "DETAILS", "ATTRIBUTES", and "CONTENT". The "DETAILS" tab is selected, showing the following information:

Field	Value
Time	01/17/2018 05:04:27.313 GMT
Event Duration	No value set
Lineage Duration	00:00:00.959
Type	ATTRIBUTES_MODIFIED
FlowFile UUID	493f1f48-628f-44c1-94fa-d8e6eb866b61
File Size	71 bytes
Component Id	36b02f3f-015e-1000-0000-000069ea5389
Component Name	UpdateAttribute
Component Type	

On the right side of the dialog, there are two sections:

- Parent FlowFiles (0)**: No parents
- Child FlowFiles (0)**: No children

The dialog has an "OK" button at the bottom right. In the background, a flow diagram is visible with a "FORK" node at the top and "SEND" and "DROP" nodes at the bottom. A timeline at the bottom left shows the date "01/17/2018 05:04:28.062 GMT". The bottom status bar indicates the current flow is "NiFi Flow" with components "AggCor" and "Record Transformation".

Data provenance – Attributes

The screenshot displays the NiFi Data Provenance interface. A central dialog box titled "Provenance Event" is open, showing the "ATTRIBUTES" tab. The dialog box is overlaid on a flow diagram that includes a "FORK" node at the top and "SEND" and "DROP" nodes at the bottom. The "ATTRIBUTES" tab lists the following data:

Attribute	Value
file size	73
filename	testdata.1516087831.5394000.cdr
path	files/IN
RetryCount	10
	9 (previous)
RouteOnAttribute Route	unmatched
s2s address	demo_nifi2.1.vbw4orzhrflm4nvdseufmtbm.demo_webnet:48544
s2s host	demo_nifi2.1.vbw4orzhrflm4nvdseufmtbm.demo_webnet
sftp listing user	user

At the bottom of the dialog box is an "OK" button. The background flow diagram also shows a timestamp "01/17/2018 05:04:28.062 GMT" and a breadcrumb trail "NiFi Flow > AggCor > Record Transformation".

Data provenance - Content 1

NiFi Data Provenance

01/17/2018 05:04:28.062 GMT

NiFi Flow » AggCorr » Record Transformation

Provenance Event

DETAILS ATTRIBUTES CONTENT

default

Section 1

Identifier 1516165260100-1

Offset 2709167

Size 71 bytes

DOWNLOAD VIEW

Replay

Connection Id 36b04e44-015e-1000-ffff-ffffa6854eda

REPLAY

OK

default

Section 1

Identifier 1516165260100-1

Offset 2709167

Size 71 bytes

DOWNLOAD VIEW

FORK

BEND

DROP

Data provenance – Content 2



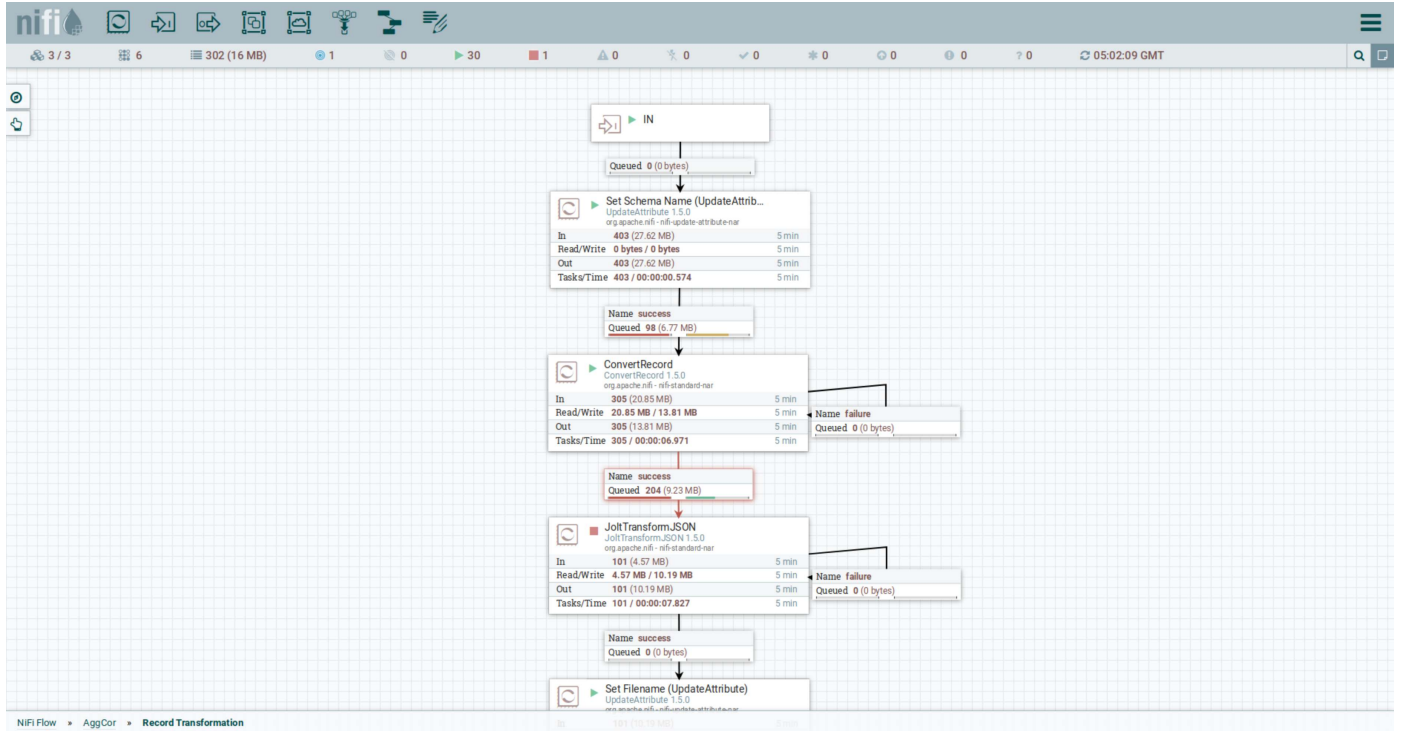
View as: original

Filename: testdata.1516087831.5394000.cdr

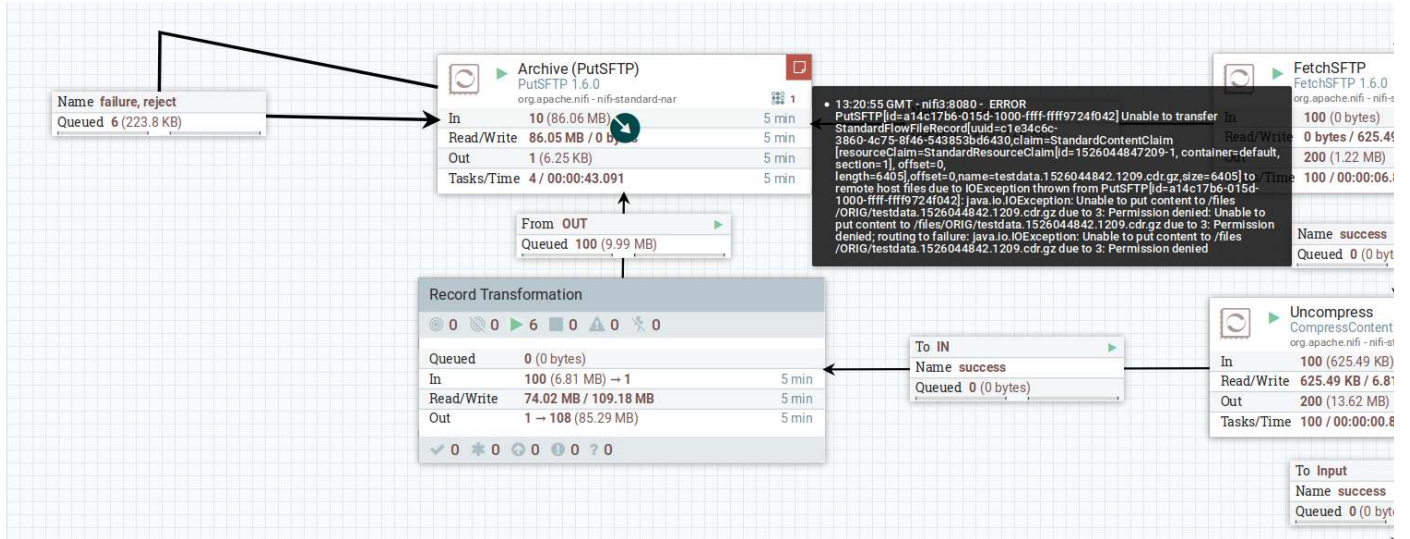
Content Type: text/plain

```
1 0,2696500,C,20180116 08:30:31 831,1,0,49176123456789,49169123456789,100
```

Back pressure



Error handling



ETL with JOLT

The screenshot displays the Jolt Transformation DSL interface. At the top, the title "Jolt Transformation DSL" is shown. On the right, a status bar indicates "Specification is Valid" with a "VALIDATE" button. Below the title, a dropdown menu is set to "Chain".

The main section is titled "Jolt Specification" and contains a JSON configuration:

```
1 [{
2   "operation": "modify-default-beta",
3   "spec": {
4     "++": {
5       "concatExampleField": "=concat(@(1,sessionId),'-',@(1,id))"
6     }
7   }
8 }, {
9   "operation": "default",
10  "spec": {
11    "++": {
12      "defaultFieldValue": 123
13    }
14  }
15 }]
```

Below the specification, there are two panels: "JSON Input" and "JSON Output". A "TRANSFORM" button is located between them.

JSON Input:

```
1 [{
2   "id": 0,
3   "sessionId": 2696500,
4   "recordType": "A"
5 }, {
6   "id": 1,
7   "sessionId": 2696500,
8   "recordType": "B"
9 }]
```

JSON Output:

```
1 [{
2   "id": 0,
3   "sessionId": 2696500,
4   "recordType": "A",
5   "concatExampleField": "2696500-0",
6   "defaultFieldValue": 123
7 }, {
8   "id": 1,
9   "sessionId": 2696500,
10  "recordType": "B",
11  "concatExampleField": "2696500-1",
12  "defaultFieldValue": 123
13 }]
```

At the bottom left, the breadcrumb "NIFI Flow > AggCor > Record Transformation" is visible.

Content

Introduction

Technology

Functionality

Live demo

Summary

Apache NiFi – Why again?

- Used in many productive systems (e.g. NSA, Slovak Telekom, ...)
- Solves many things already (UI, Tracing, Transactional, ...)
- Easily customizable
- Open Source
 - Reduce your OPEX
 - Can be extended by YOU
 - Is extended and fixed by Community
- It scales
- Cloud ready



The value of Apache NiFi

- CLOC returns about 2 million lines of code
- SLOCCount
 - Total physical source lines of code: 1,2M
 - Development effort estimate: 355 PY
 - Total estimated cost to develop: \$48M
- Of course, you might not use everything
- But even 10% of it are: \$5M !
- So it's better to not reinvent the wheel

<https://pxhere.com/en/photo/926209>

Apache NiFi – Use cases

- Workflow modeling with data flows
- Reduce latency of your data
- Centralization of complex data flows
- Big Data and BI data flows
- Integration of new/different technologies
- Accountability and lineage
- Complex Event Processing*
- ETL*

Contact

- Frank Thiele (frank.thiele@tngtech.com)

