

Supletivo **Apache NiFi**

#data-engineering



DataHackers



Pergunte o que quiser aos nossos especialistas no Slack

Data Hackers

Data Hacker (você)



Todos os tópicos

Canais

data-engineering

machine-learning

random

data-science

data-viz

english-only

opendata

ds-maringa

marketing

data-governance

data-writers

pesquisa

Mensagens Diretas

Ei, psiu!

1

#random

34895 | Assuntos gerais quando não é aplicado a um canal específico.



image.png



Marcio Ribeiro

Boa tarde pessoal.

Gostaria de compartilhar um projeto que estou desenvolvendo baseado nos dados do **CartolaFC**.

Ainda está em fase inicial e com alguns pequenos bugs.

Meu objetivo é aprender sobre as ferramentas do Google Cloud e como se comunicam.

Agora estou utilizando os dados que coletei para treinar um modelo de previsão de pontuação dos atletas. Até a rodada 15 não bons resultados, mas o que vale mesmo é o aprendizado. 😊

<http://www.marcioribeiro.tec.br/portfolio/cartolafc2020>

2 respostas



Jonas Scherer

image.png

1 resposta



Gilberto Titericz

https://www.infoq.com/news/2020/10/training-exceeds-gpt3/?utm_source=sumome&utm_medium=twitter&utm_campaign=sumome_share



Serves

Polêmica: "Don't Choose Python as Your First Programming Language" by Shubham Pathania <https://link.medium.com/h8fw0vgdDab>

12 respostas



Marlesson

E aí, qual buzzword você acha mais irritante nesse meio? Bora discutir aí nos comentários!

<https://medium.com/data-hackers/ufb-ultimate-fighter-buzzword-data-hackers-podcast-32-24e08cbd4ffe>

6 respostas

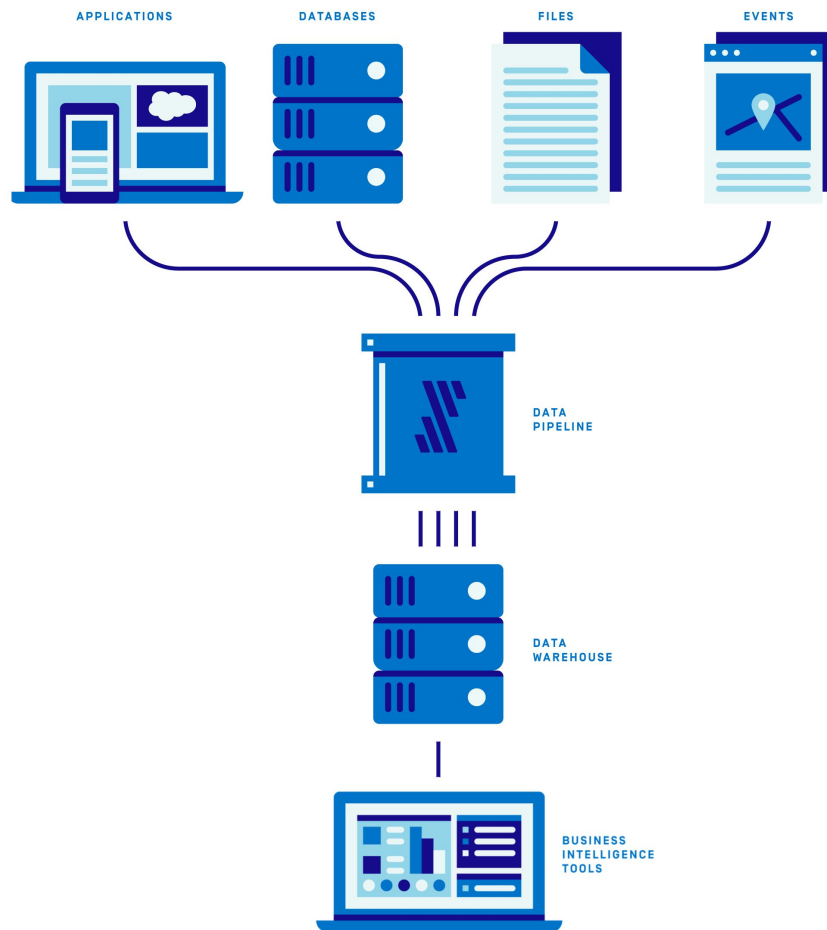


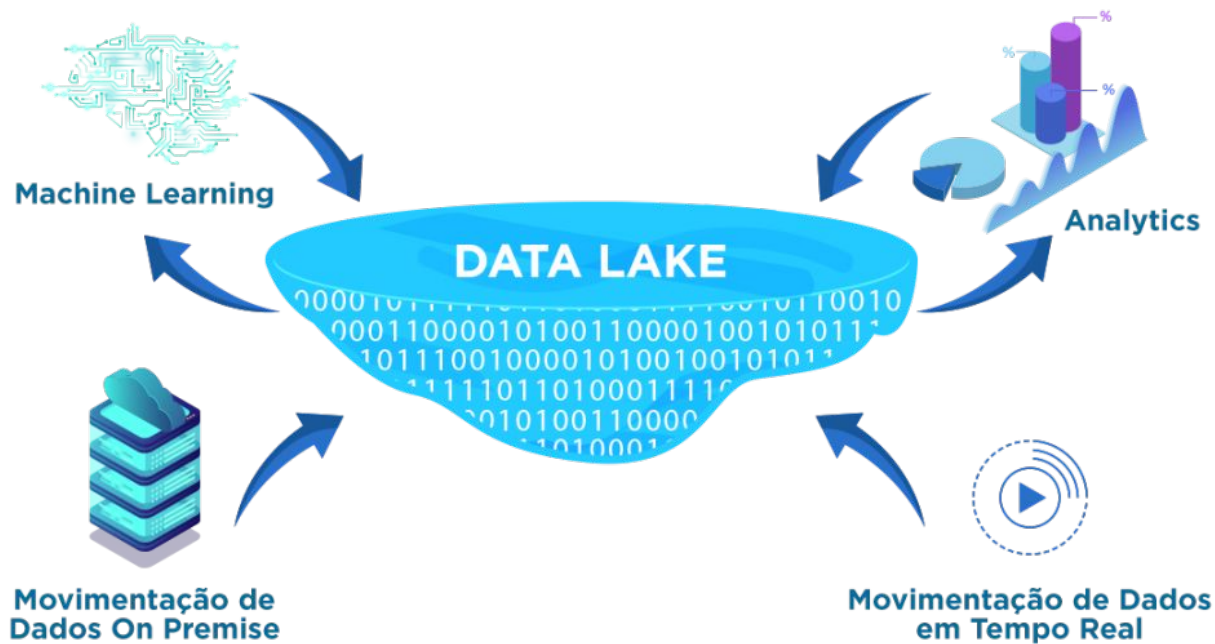
Envie uma mensagem

Apache NiFi

Roteiro

- Motivação
- Apache NiFi
- DataFlow
- Arquitetura
- Componentes
- MiNiFi
- Registry
- DEMO





Motivação

Data Lake ?





An easy to use, powerful, and reliable
system to process and distribute data.

Apache NiFi

DataFlow

<https://nifi.apache.org/>
<https://hub.docker.com/r/apache/nifi/>



Apache NiFi

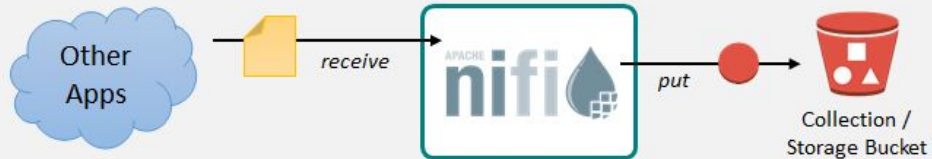
Data Flow



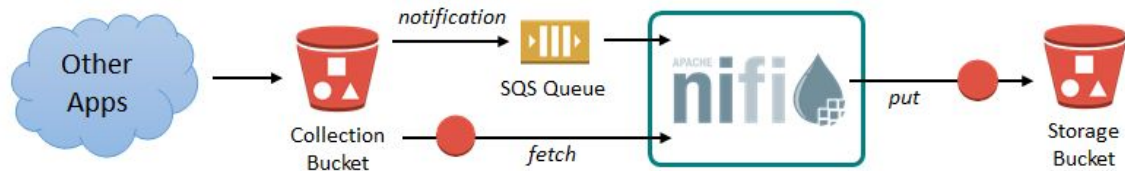
Apache NiFi

Data Flow - S3

**NiFi as
Gateway to S3**

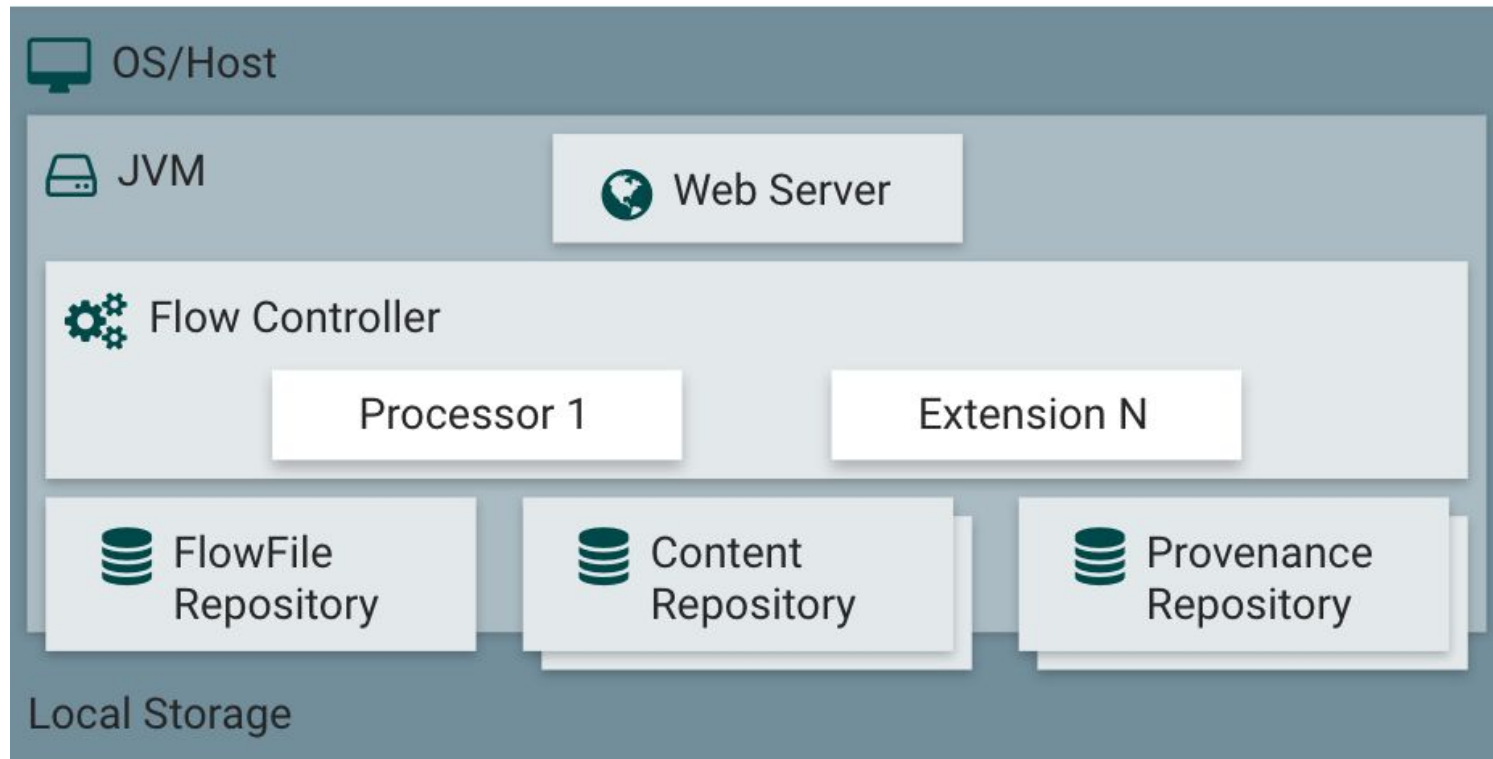


**NiFi Processing S3
Event Notifications**



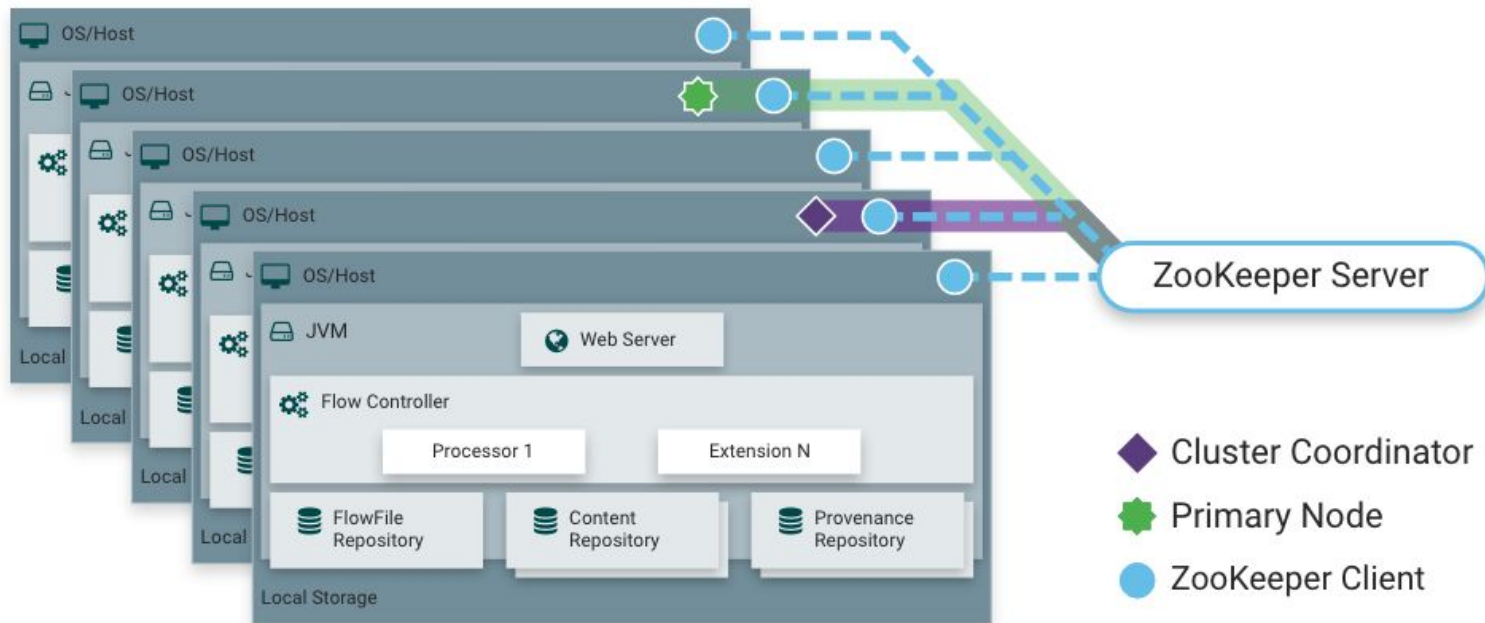
Apache NiFi

Arquitetura



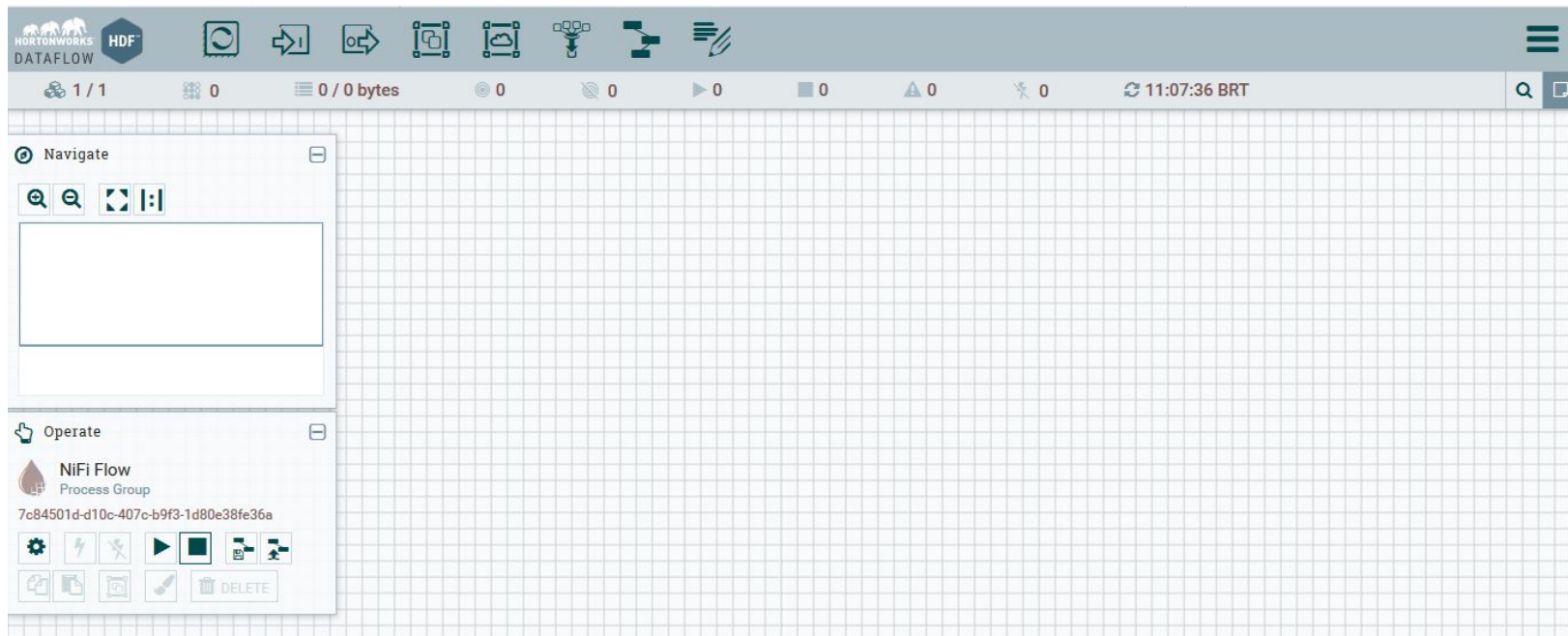
Apache NiFi

Arquitetura



Apache NiFi

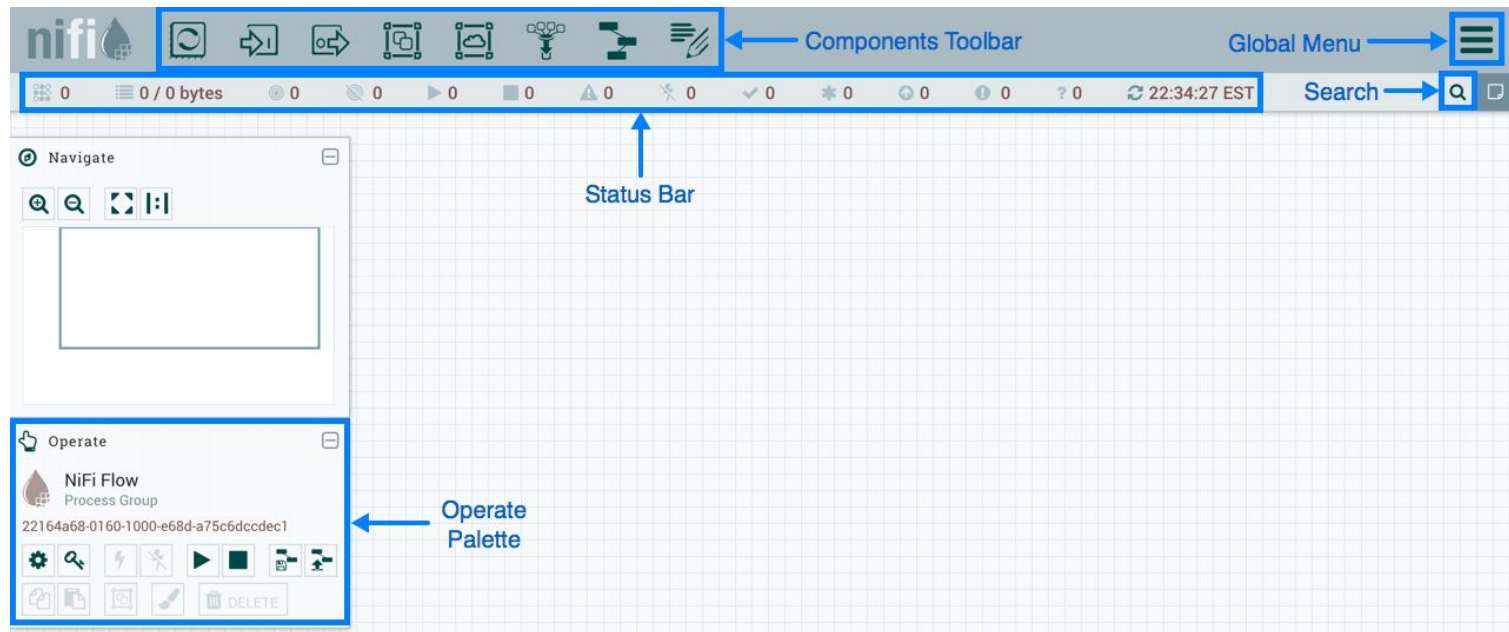
Core concepts - UI



Apache NiFi

Core concepts - UI

Core concepts - UI



Apache NiFi

Core concepts - UI

The screenshot displays the Apache NiFi user interface. At the top is a toolbar with various icons for navigation and actions. Below the toolbar is a status bar showing metrics like '0 / 0 bytes' and a timestamp '22:51:04 EST'. The main workspace is a grid where a 'Final Process Group' is being edited. A blue box highlights the 'Navigate' palette on the left, which contains a 'Bird's Eye View' of the process group. Another blue box highlights the 'Breadcrumbs' at the bottom, showing the path: 'NiFi Flow » Process Group A » Inner Group » Another Process Group'. A third blue box highlights the 'Final Process Group' summary table on the right.

Navigate Palette

Bird's Eye View

Breadcrumbs

Final Process Group		
Queued	0 (0 bytes)	
In	0 (0 bytes) → 0	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 → 0 (0 bytes)	5 min



Apache NiFi

Core concepts - Processors

Add Processor

Source

all groups

amazon attributes
avro aws consume
csv database fetch
files get hadoop
ingest input insert
json listen logs
message put
remote restricted
source sql text
update

Displaying 219 of 219

Filter

Type	Version	Tags
AttributeRollingWindow	1.2.0	rolling, data science, Attribute Expression Language, st...
AttributesToJSON	1.2.0	flowfile, json, attributes
Base64EncodeContent	1.2.0	encode, base64
CaptureChangeMySQL	1.2.0	cdc, jdbc, mysql, sql
CompareFuzzyHash	1.2.0	fuzzy-hashing, hashing, cyber-security
CompressContent	1.2.0	lzma, decompress, compress, snappy framed, gzip, sna...
ConnectWebSocket	1.2.0	subscribe, consume, listen, WebSocket
ConsumeAMQP	1.2.0	receive, amqp, rabbit, get, consume, message
ConsumeEWS	1.2.0	EWS, Exchange, Email, Consume, Ingest, Message, Get,...
ConsumeIMAP	1.2.0	Imap, Email, Consume, Ingest, Message, Get, Ingress
ConsumeJMS	1.2.0	jms, receive, get, consume, message
ConsumeKafka	1.2.0	PubSub, Consume, Inqest, Get, Kafka, Ingress, Topic, 0...

AttributeRollingWindow 1.2.0 org.apache.nifi - nifi-stateful-analysis-nar

Track a Rolling Window based on evaluating an Expression Language expression on each FlowFile and add that value to the processor's state. Each FlowFile will be emitted with the count of FlowFiles and total aggregate value of values processed in the current time window.

CANCEL

ADD



Core concepts - Processors



DATA INGESTION PROCESSORS

- | | | |
|---------------------|---------------|------------------------|
| X GenerateFlowFiles | X GetJMSTopic | X GetKafka |
| X GetFile | X GetHTTP | X QueryDatabaseTable |
| X GetFTP | X ListenHTTP | X GetMongo |
| X GetSFTP | X ListenUDP | X GetTwitter |
| X GetJMSQueue | X GetHDFS | X ListHDFS / FetchHDFS |



Apache NiFi

Core concepts - Processors



DATA TRANSFORMATION PROCESSORS

- | | | |
|--------------------|-----------------------|---------------------|
| X ConvertRecord | X ReplaceText | X EncryptContent |
| X UpdateRecord | X CompressContent | X TransformXml |
| X ConvertJSONToSQL | X ConvertCharacterSet | X JoltTransformJSON |



Core concepts - Processors



DATA EGRESS / SENDING DATA PROCESSORS

X PutEmail

X PutFile

X PutFTP

X PutSFTP

X PutJMS

X PutSQL

X PutKafka

X PutMongo

X PutHDFS



Apache NiFi

Core concepts - Processors



ROUTING AND MEDIATION PROCESSORS

- | | | |
|-------------------|--------------------|---------------|
| X ControlRate | X RouteOnAttribute | X ScanContent |
| X DetectDuplicate | X RouteOnContent | X ValidateXml |
| X DistributeLoad | X ScanAttribute | X ValidateCSV |



Apache NiFi

Core concepts - Processors



DATABASE ACCESS PROCESSORS

X ConvertJSONToSQL

X ExecuteSQL

X PutSQL

X SelectHiveQL

X PutHiveQL

X ListDatabaseTables



Apache NiFi

Core concepts - Processors



ATTRIBUTE EXTRACTION PROCESSORS

X EvaluateJsonPath

X EvaluateXPath

X EvaluateXQuery

X ExtractText

X HashAttribute

X HashContent

X IdentifyMimeType

X UpdateAttribute

X LogAttribute



Apache NiFi

Core concepts - Processors



SYSTEM INTERACTION PROCESSORS

- X ExecuteProcess
- X ExecuteStreamCommand



Apache NiFi

Core concepts - Processors



SPLITTING AND AGGREGATION PROCESSORS

X SplitText

X SplitJson

X SplitXml

X SplitRecord

X SplitContent

X UnpackContent

X SegmentContent

X MergeContent

X QueryRecord



Apache NiFi

Core concepts - Processors



HTTP AND UDP PROCESSORS

X GetHTTP

X ListenHTTP

X InvokeHTTP

X PostHTTP

X HandleHttpRequest

X HandleHttpResponse

X Listen UDP

X PutUDP

X ListenUDPRecord



Apache NiFi

Core concepts - Processors



AMAZON WEB SERVICES PROCESSORS

X FetchS3Object

X PutS3Object

X PutSNS

X GetSQS

X PutSQS

X DeleteSQS

X GetDynamoDB

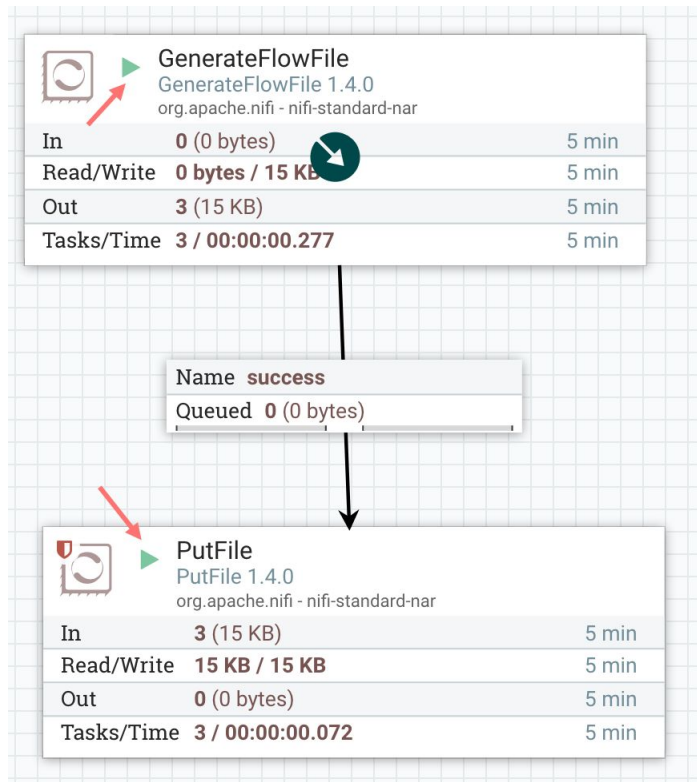
X PutDynamoDB

X PutLambda



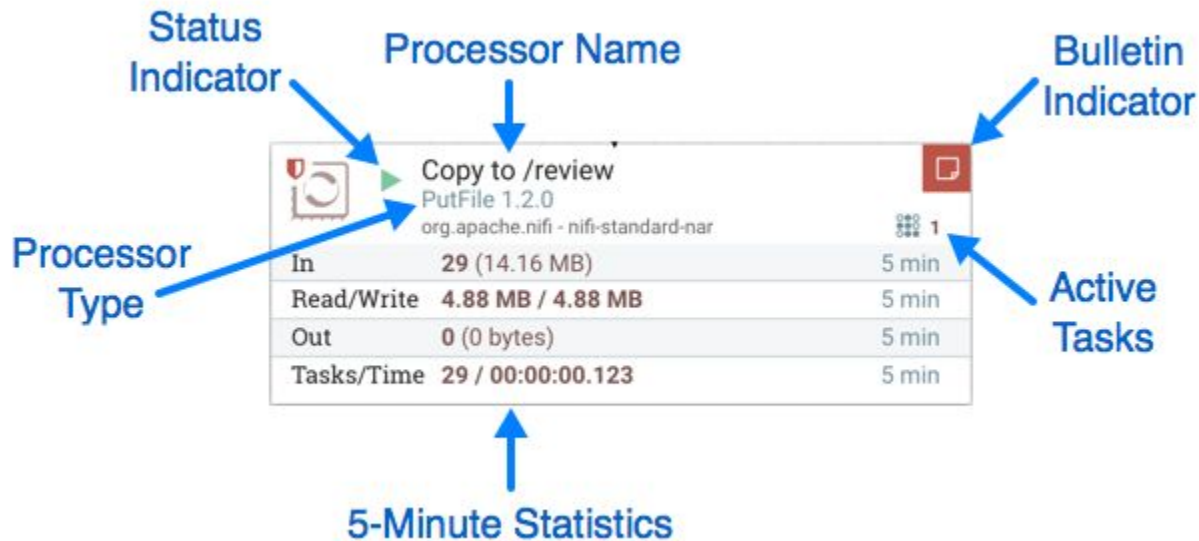
Apache NiFi

Core concepts - Processors



Apache NiFi

Core concepts - Processors



Apache NiFi

Core concepts - Processors

The screenshot displays the Apache NiFi web interface. At the top, a toolbar contains various icons for navigation and actions. Below the toolbar, a status bar shows metrics: 0 flows, 0/0 bytes, 0 processors, 0 connections, 0 queues, 2 alerts, 0 errors, and the current time 14:31:49 EST. The main workspace is a grid where two processors are visible: **GenerateFlowFile** and **LogAttribute**. A context menu is open over the **GenerateFlowFile** processor, listing actions: **Configure**, **Status History**, **Data provenance**, **Upstream connections**, **Downstream connections**, **Usage**, **Change color**, **Center in view**, **Copy**, and **Delete**. On the left, a **Navigate** panel shows a search bar and a list of flows. Below it, an **Operate** panel shows the details of the selected **GenerateFlowFile** processor, including its ID **a39b072f-0159-1000-dad3-2bdab60c82f0** and a set of icons for configuration, deletion, and other actions.

GenerateFlowFile Processor Details:

Metric	Value
In	0 (0 bytes)
Read/Write	0 bytes / 0 bytes
Out	0 (0 bytes)
Tasks/Time	0 / 00:00:00.000

LogAttribute Processor Details:

Metric	Value
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



Apache NiFi

Core concepts - Processors

Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Scheduling Strategy ?

Timer driven ▼

Concurrent Tasks ?

1

Execution ?

All nodes ▼

Run Duration ?

0ms 25ms 50ms 100ms 250ms 500ms 1s 2s

Lower latencyHigher throughput

Run Schedule ?

0 sec

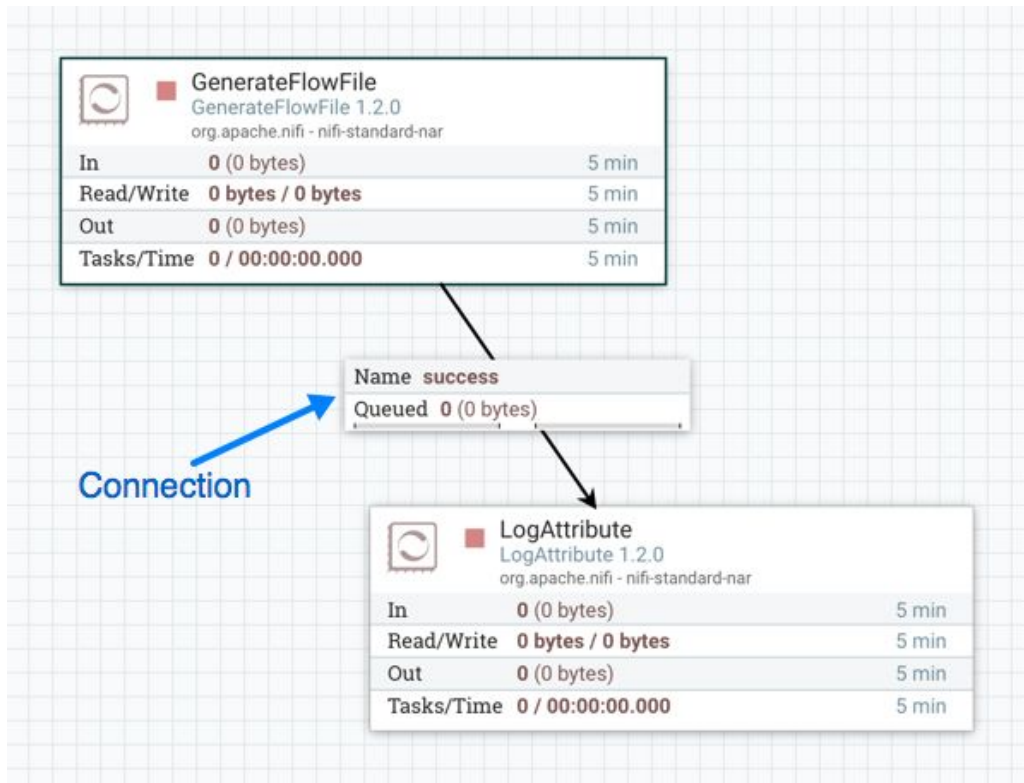
CANCEL

APPLY



Apache NiFi

Core concepts - Connection



Apache NiFi

Core concepts - Connection

Configure Connection

DETAILS

SETTINGS

Name

Id

208bf461-0166-1000-ffff-ffffefb88d70

FlowFile Expiration

0 sec

Back Pressure

Object Threshold

10000

Size Threshold

1 GB

Load Balance Strategy

Partition by attribute

Attribute Name

myAttribute

Load Balance Compression

Do not compress

Available Prioritizers

FirstInFirstOutPrioritizer

NewestFlowFileFirstPrioritizer

OldestFlowFileFirstPrioritizer

PriorityAttributePrioritizer

Selected Prioritizers

CANCEL

APPLY



Apache NiFi

Core concepts - Controller Service

The screenshot shows the 'NiFi Flow Configuration' dialog box in the Apache NiFi web console. The 'CONTROLLER SERVICES' tab is selected. A table lists the configured controller services. In this case, there is one service: 'DBCPConnectionPool' of type 'DBCPConnectionPool 1.4.0' and bundle 'org.apache.nifi - nifi-dbcp-service-nar'. It is in an 'Enabled' state and has a scope of 'NiFi Flow'. The interface also shows a 'Last updated' timestamp and a note about the service's availability to descendant processors.

NiFi Flow Configuration

GENERAL CONTROLLER SERVICES

Name	Type	Bundle	State	Scope
DBCPConnectionPool	DBCPConnectionPool 1.4.0	org.apache.nifi - nifi-dbcp-service-nar	Enabled	NiFi Flow

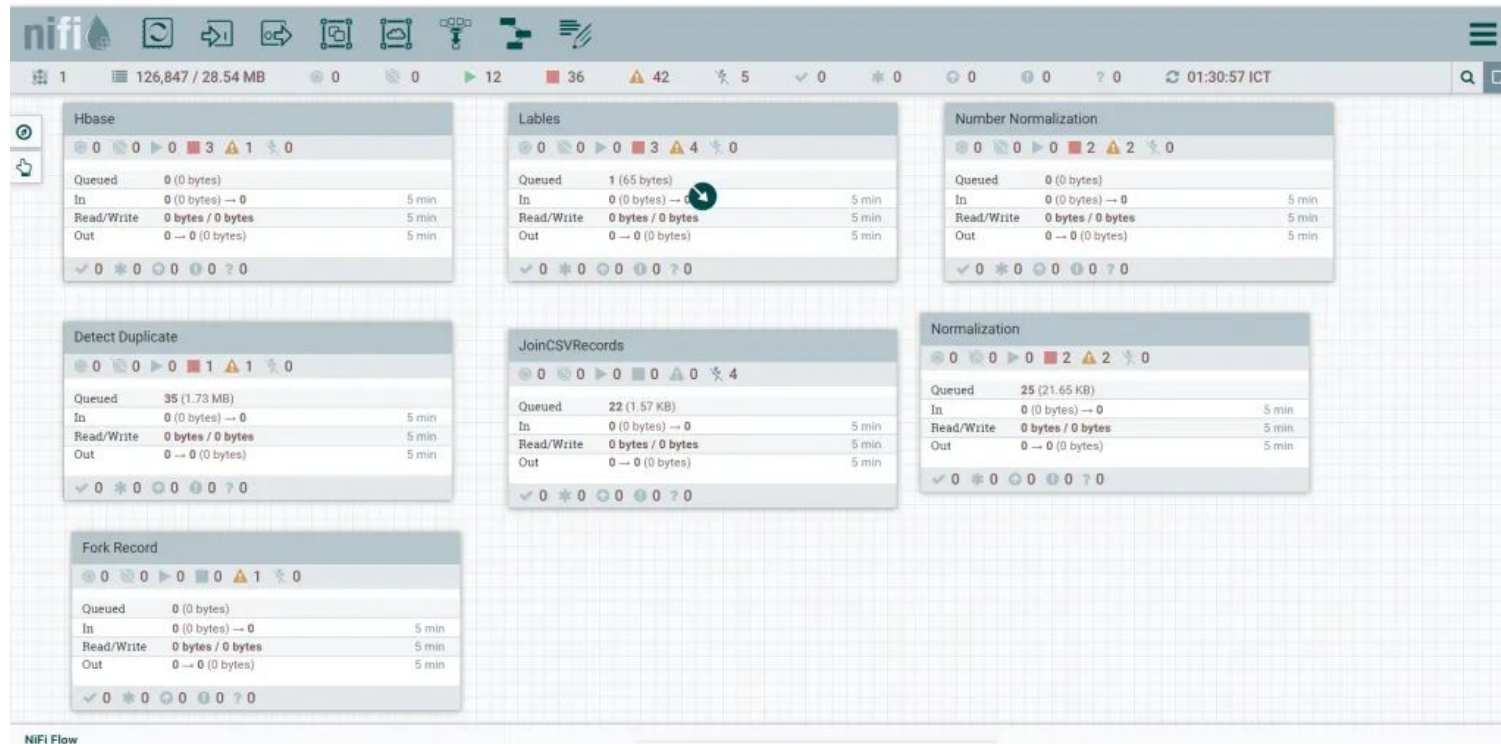
Last updated: 12:08:48 EDT

Listed services are available to all descendant Processors and services of this Process Group.



Apache NiFi

Core concepts - Process Group



Apache NiFi

Core concepts - Process Group

The diagram illustrates the components of an Apache NiFi Process Group card. The card is titled "Process Group ABC" and shows 2 active tasks. It includes a row of component counts (1, 1, 2, 9, 4, 0) and a 5-minute statistics table. At the bottom, it displays version state counts (0, 0, 0, 0, 0, 0). Annotations point to various parts of the card:

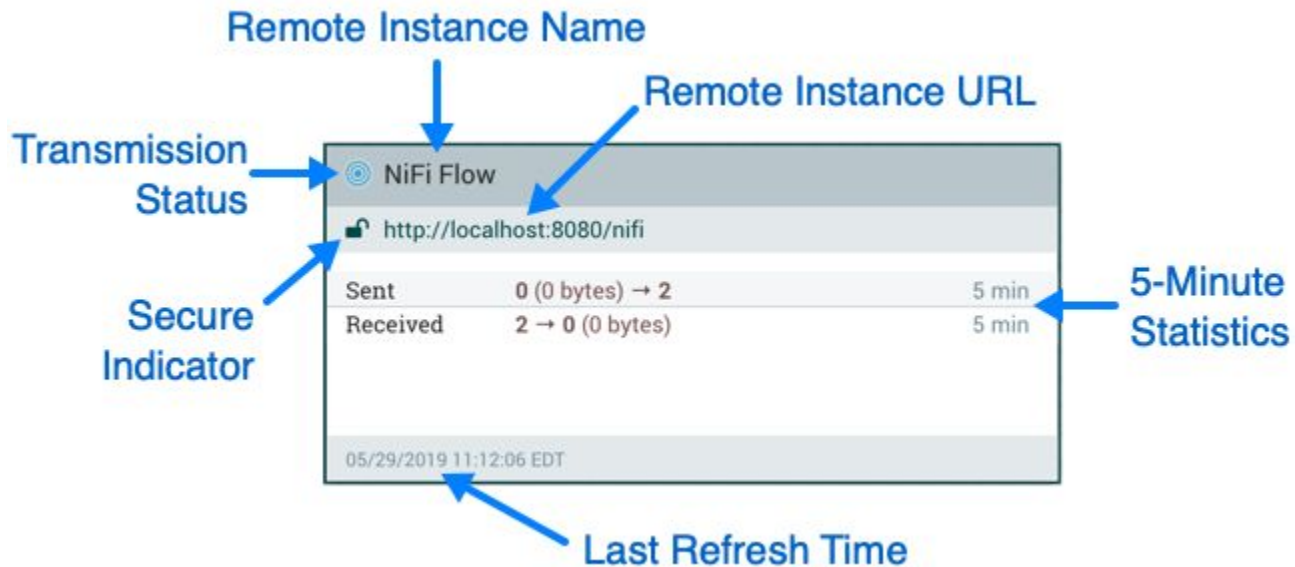
- Name**: Points to the title "Process Group ABC".
- Active Tasks**: Points to the "2" next to the task icon.
- Component Counts**: Points to the row of icons and numbers (1, 1, 2, 9, 4, 0).
- 5-Minute Statistics**: Points to the statistics table.
- Bulletin Indicator**: Points to the red square icon.
- Version State Counts**: Points to the row of icons and numbers (0, 0, 0, 0, 0, 0).
- Hover over to see Comments**: Points to the bottom right corner of the card.

Process Group ABC	
Queued	26 (12.7 MB)
In	8 (800 KB) → 2 5 min
Read/Write	14.72 MB / 14.8 MB 5 min
Out	3 → 16 (78.57 KB) 5 min



Apache NiFi

Core concepts - Remote Process Group



Apache NiFi

Parameter Context

Configure Controller Service

SETTINGS

PROPERTIES

COMMENTS

Required field

+

Property		Value	
Database Connection URL	?	No value set	
Database Driver Class Name	?	No value set	
Database Driver Location(s)	?	No value set	
Database User	?	No value set	
Password	?	No value set	
Max Wait Time	?	500 millis	
Max Total Connections	?	8	
Validation query	?	No value set	













Apache NiFi

Parameter Context

Update Parameter Context

SETTINGS

PARAMETERS

Name ▲	Value	
JKS	JKS	 
TLS	TLS	 
broker	<input type="text" value=":9092"/>	 
truststoreFile	/Library/Java/JavaVirtualMachines/...	 
truststorePass	Sensitive value set	 



Parameter
JKS

Referencing Components ?

▼ **Stock to Kafka** (1)
Referencing Processors
None

Referencing Controller Services
 [sslStocks](#)

Unauthorized Referencing Components
None



Apache NiFi

Parameter Context

Controller Service Details

SETTINGS

PROPERTIES

COMMENTS

Required field

Property		Value	
Database Connection URL	?	<code>{db-con-url}</code>	→
Database Driver Class Name	?	<code>{mysql-jdbc-driver-name}</code>	→
Database Driver Location(s)	?	<code>{jdbc-jar-path}</code>	→
Kerberos Credentials Service	?	No value set	
Database User	?	<code>{db-user}</code>	→
Password	?	Sensitive value set	
Max Wait Time	?	500 millis	
Max Total Connections	?	8	
Validation query	?	No value set	





Apache NiFi

Sub-projects



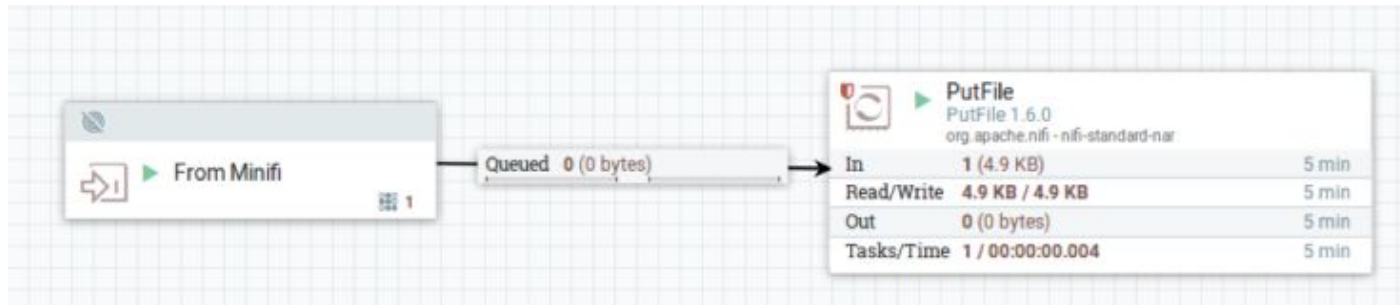
Apache NiFi

MiniFi



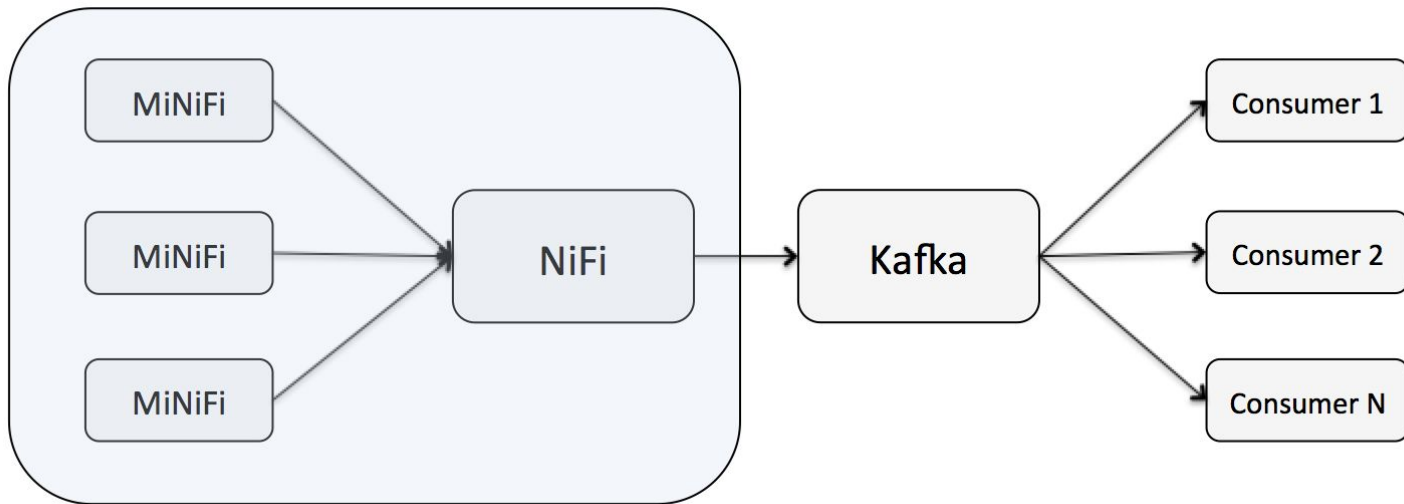
Apache NiFi

MiniFi



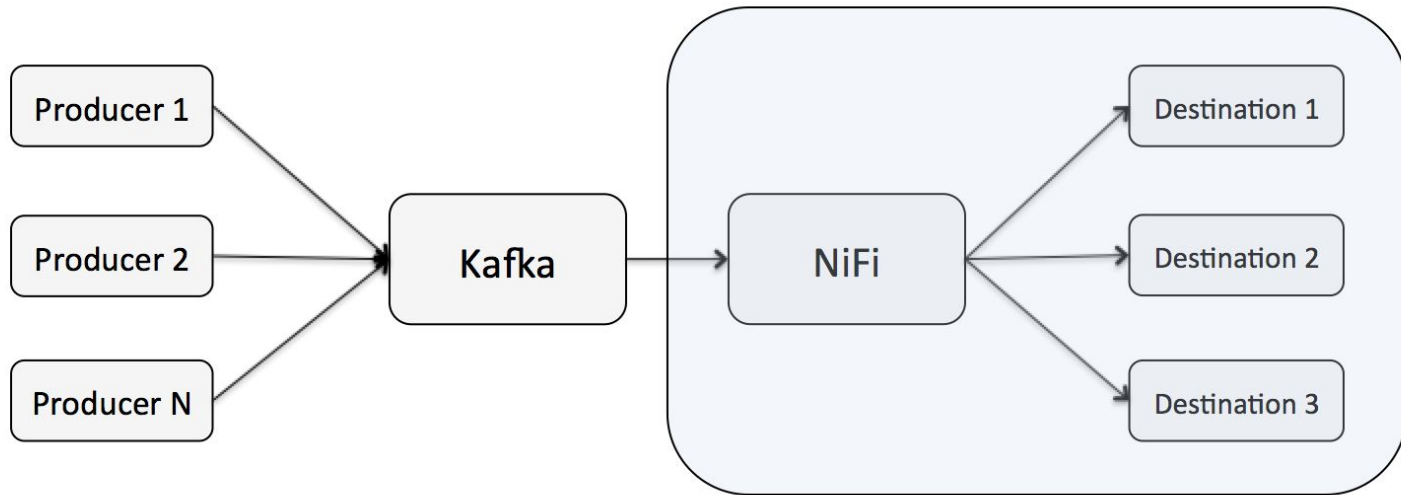
Apache NiFi

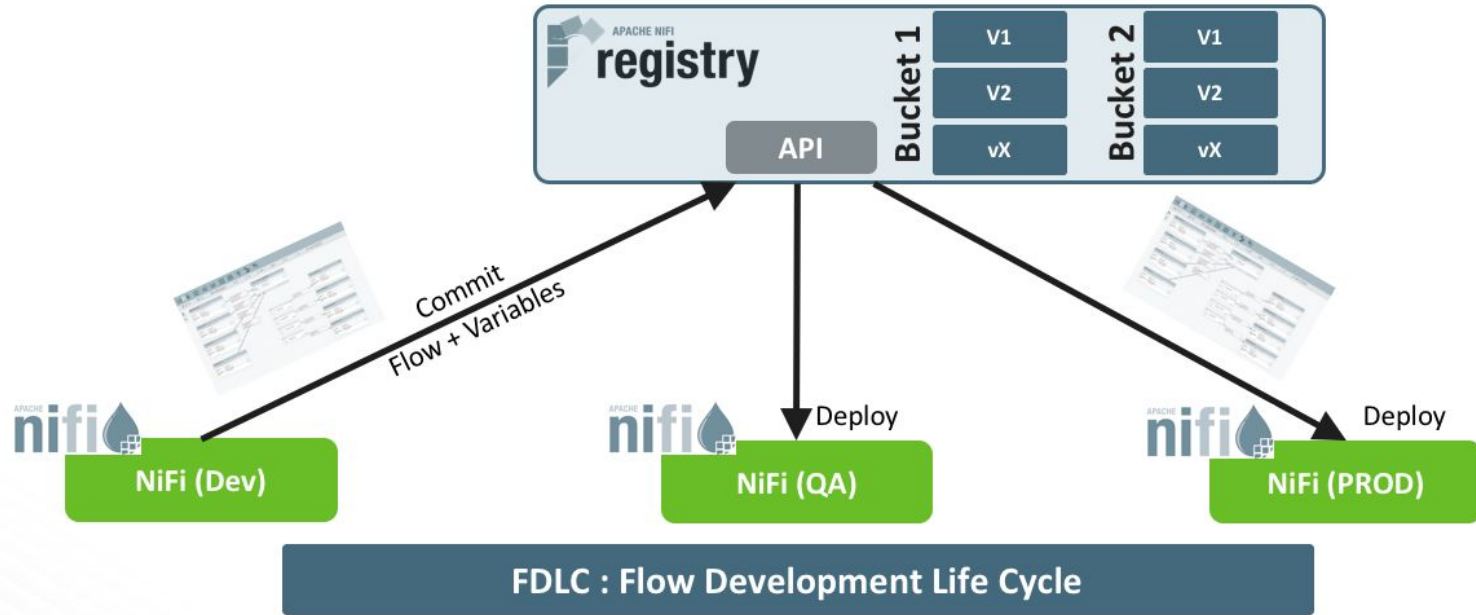
MiniFi



Apache NiFi

MiniFi





Apache NiFi Registry

Versionamento



Apache NiFi Registry

The screenshot displays the Apache NiFi Registry web interface. At the top, a dark header bar contains the NiFi logo, the text "NiFi Registry / All", and user information "registry_user" with a "LOGOUT" link. Below the header, a search bar and a "Sort by: Name (a - z)" dropdown are visible. The main content area shows a list of flows. The first flow, "Flow 1 - Bucket 1", is expanded to show its details. This detail view includes a "DESCRIPTION" section with the text "Description 1", a "CHANGE LOG" section showing "Version 1 - an hour ago by nobel", and a "Comments 1" section with a comment from "Dec-26-2017 at 11:22 PM". To the right of the description, an "ACTIONS" dropdown menu is open, showing a "delete" option. The second flow, "Flow 2 - Bucket 2", is partially visible below the first one.

Flow Name	Bucket	Flow Type	VERSIONS
Flow 1 - Bucket 1	Flow		1
Flow 2 - Bucket 2	Flow		2

Flow 1 - Bucket 1 Details:

- DESCRIPTION:** Description 1
- CHANGE LOG:**
 - Version 1 - an hour ago by nobel
- Comments 1:** Dec-26-2017 at 11:22 PM



Apache NiFi

DEMO



Hello World



Apache NiFi

DEMO 2



ETL com enriquecimento



Apache NiFi

DEMO 3



ETL Mongo -> Postgres



Apache NiFi

Boas práticas

Summary, Take-aways, Best Practices

- Use process groups. Version them in NiFi Registry
- Get creative with site-to-site and remote process groups & ports
- Use Registry buckets to organize your flows across teams or domain
- Use labels and comments for collaboration
- For CD or automated deployments, avoid changes in adjacent versions that would require manual intervention, such as removing connects
- Automation:
 - See what's possible with NiFi CLI
 - [NiPyAPI](#) Python module
 - NiFi and NiFi Registry REST API



Apache NiFi

Onde aprender?

<https://www.udemy.com/course/apache-nifi-zero-to-hero>

<https://www.udemy.com/course/apache-nifi-the-beginner-guide/?referralCode=3FCCF7F1DF404032EE74>

<https://github.com/hortonworks-gallery/nifi-templates>

<https://github.com/jfrazee/awesome-nifi>

<https://pierrebillard.com/>

https://www.youtube.com/watch?v=NC3llbjrKeA&list=PLpe6CiYAs7vTS4LG7kkvuPZEzEobvw_IR

<https://www.youtube.com/watch?v=VVnFt54jUQ8>

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



Referências

<https://www.udemy.com/course/apache-nifi-zero-to-hero>

<https://bryanbende.com/development/2016/09/15/apache-nifi-and-apache-kafka>

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

<https://arb.ahgora.com.br/rodrigo.tripodi/nifi-poc/-/blob/master/docker-compose.yml>

<https://community.cloudera.com/t5/Community-Articles/Storing-Apache-NiFi-Versioned-Flows-in-a-Git-Repository/ta-p/248713>

<https://pierrevillard.com/2018/04/09/automate-workflow-deployment-in-apache-nifi-with-the-nifi-registry/comment-page-1/>

<https://www.youtube.com/watch?v=jATV-0Oxg7o&list=PLpe6CiYAs7vQpGPQaUO4FSb7hGw62c-HU&index=5>

<https://www.youtube.com/watch?v=fb0JdkwaiM>

https://www.youtube.com/watch?v=NC3llbjrKeA&list=PLpe6CiYAs7vTS4LG7kkvuPZEzEobvw_IR





Dúvidas e sugestões



Supletivo **Apache NiFi**

#data-engineering

Rodrigo Carneiro
teoria@gmail.com
linkedin.com/in/rodrigoteoria



DataHackers

