

MBA⁺

***MBA EM BIG DATA
(DATA SCIENCE)***

MBA⁺

INGESTÃO DE DADOS
NiFi

Prof. Thiago Nascimento Nogueira
tnnogueira@gmail.com

Linked-in: <https://www.linkedin.com/in/thnogueira>



NiFi

Características

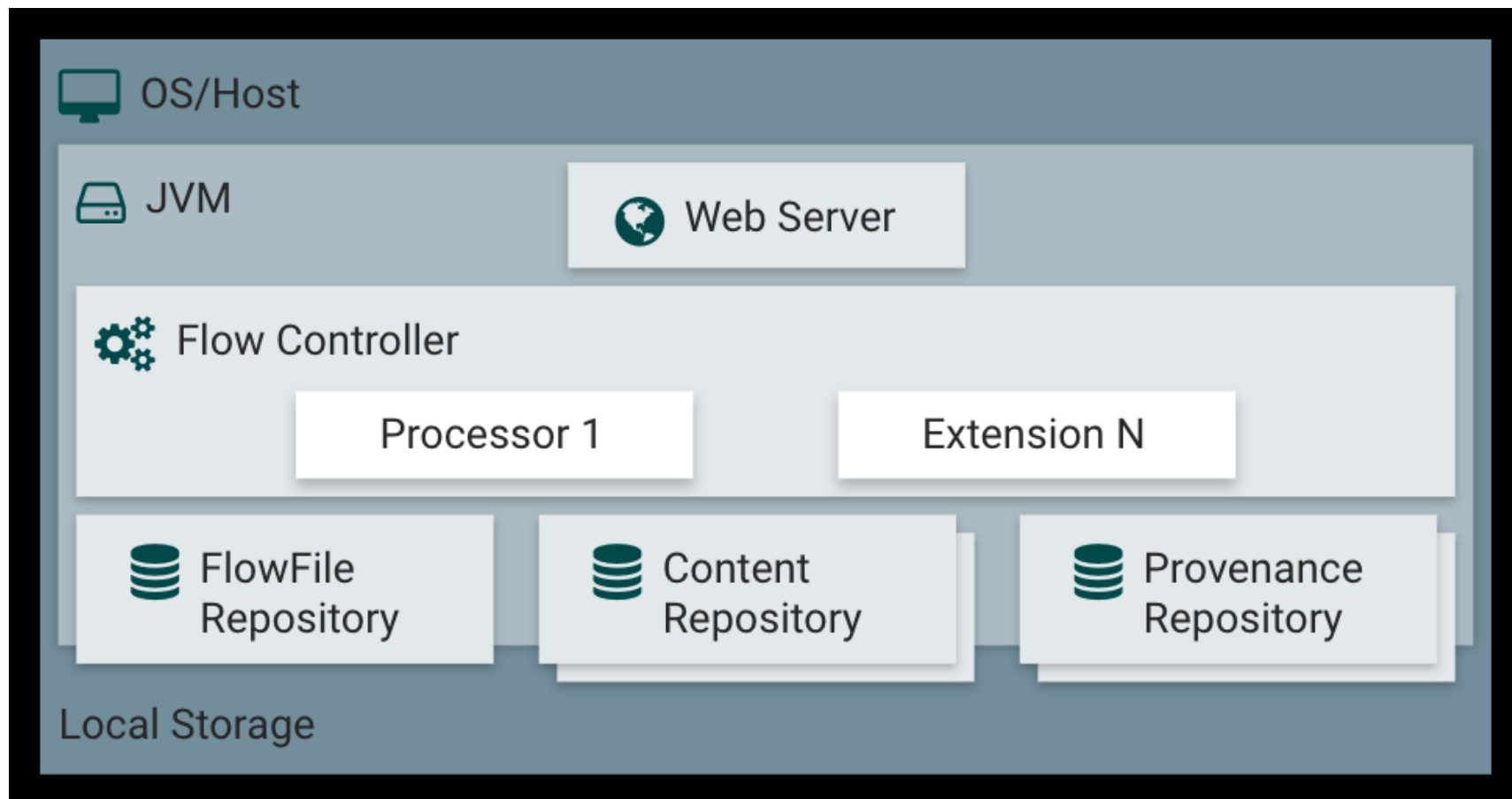


- “Canivete suíço” para ingestão de dados
- Utilizado em projetos que não envolvem necessariamente Hadoop
- Interface Web
 - Experiência integrada entre desenho, controle, feedback e monitoria
- Altamente configurável
 - Tolerante a falhas com garantia de entrega
 - Baixa latência e alto throughput
 - Priorização dinâmica
 - Fluxo pode ser alterado em tempo de execução
 - “Back pressure”

- Data Provenance (Procedência de dados)
 - Rastreamento do dataflow, do início ao fim
- Desenhado para extensibilidade
 - Permite desenvolvimento de processors customizados
 - Rápido desenvolvimento e testes efetivos
- Segurança
 - SSL, SSH, HTTPS, criptografia de conteúdo, etc..
 - Autorização Multi-tenant e autorização / política de gerenciamento interna

- **FlowFile** – Objeto que se move através do fluxo onde, para cada FlowFile, o NiFi mantém um mapeamento key/value com atributos e seu conteúdo associado, com zero ou mais bytes
- **Processadores de Fluxo** – Realizam o trabalho de fato. Realizam alguma combinação entre roteamento de dados, transformação, ou mediação entre sistemas. Possuem acesso aos atributos e conteúdos do FlowFile
- **Conexão** – Promovem as ligações entre os processadores. Funcionam como filas e permitem que vários processadores interajam de diferentes maneiras
 - Podem ser priorizadas dinamicamente
 - Possuem limites máximos de carga, permitindo controle tipo “back pressure”
- **Controlador de Fluxo** – Mantém o controle de como os processos se conectam e gerencia os recursos de cada processo
- **Grupo de Processos** – Agrupamento lógico de conjunto de processos.

- **WebServer** – Hospeda os comando NiFi baseados em http e API de controle
- **Flow Controller** – Cérebro da operação. Provê as threads para que os processos rodem e gerencia o agendamento de quando o processo recebe os recursos para executar
- **FlowFile Repository** – Lugar onde o NiFi mantém as informações sobre os estados dos FlowFiles
- **Content Repository** – Lugar o são armazenados os conteúdos do FlowFile em execução
- **Provenance Repository** – Local onde os dados de eventos de data provenance são armazenados

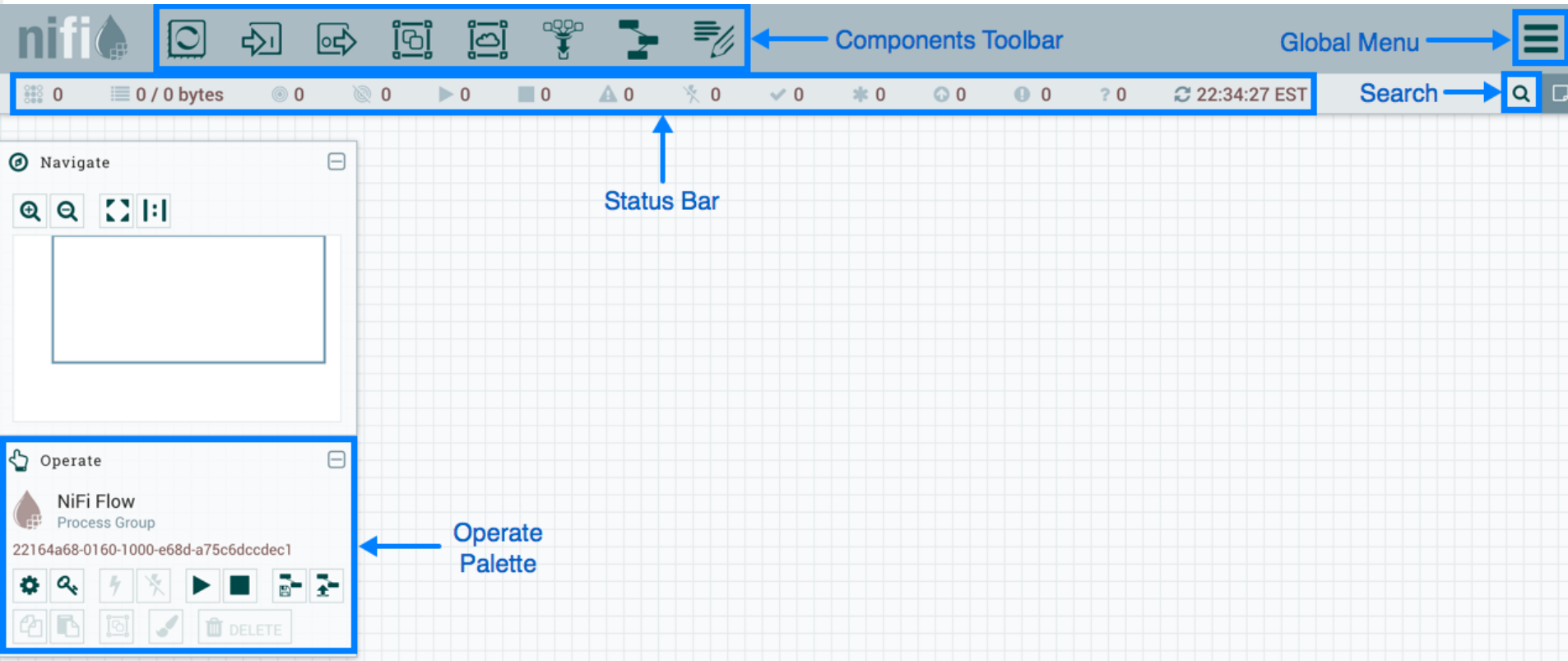


Tela Inicial

FIAP

The screenshot displays the Apache NiFi web interface. At the top, there is a header bar with the 'nifi' logo and a series of icons representing different components: a clock, a square with an arrow, a square with a circle, a square with a cloud, a square with a person, a square with a gear, and a square with a list. Below the header bar, a status bar shows various metrics: '0', '0 / 0 bytes', '0', '0', '0', '0', '0', and '19:34:16 EDT'. On the left side, there are two sidebars. The top sidebar is titled 'Navigate' and contains a search icon, a magnifying glass icon, and a list of icons. The bottom sidebar is titled 'Operate' and contains a 'NiFi Flow' section with a 'Process Group' and a UUID '64c118d3-efbb-4976-acb2-1a13f7cfe1ef'. Below this, there are several icons for actions like 'Run', 'Stop', 'Refresh', 'Delete', and 'Export'. The main area of the interface is a large grid with a light blue background and a white grid pattern. At the bottom left, there is a small label 'NiFi Flow'.

FIAP



Componentes

The screenshot displays the Apache NiFi web interface. The top navigation bar includes the NiFi logo, a **Components Toolbar** with icons for various actions, a **Status Bar** showing system metrics (0/0 bytes, 0 errors, 0 warnings, 0 info, 0 debug, 0 audit, 0 metrics, 0 logs, 0 alerts, 0 notifications, 0 help, 0 about), a **Global Menu** icon, and a **Search** bar. On the left, there is a **Navigate** sidebar and an **Operate** sidebar. The **Operate** sidebar is highlighted with a blue box and labeled **Operate Palette**; it shows the **NiFi Flow** process group and a list of actions including **Settings**, **Search**, **Run**, **Stop**, **Restart**, **Refresh**, **Export**, **Import**, and **DELETE**. On the right, a **Global Menu** is open, listing various system components: **Summary**, **Counters**, **Bulletin Board**, **Data Provenance**, **Controller Settings**, **Flow Configuration History**, **Users**, **Policies**, **Templates**, **Help**, and **About**.

Components Toolbar

Status Bar

Global Menu

Search

Navigate

Operate

NiFi Flow
Process Group
22164a68-0160-1000-e68d-a75c6dccdec1

Operate Palette

- Summary
- Counters
- Bulletin Board
- Data Provenance
- Controller Settings
- Flow Configuration History
- Users
- Policies
- Templates
- Help
- About

Adicionando um Processor



FIAP

Add Processor

Source

Displaying 219 of 219

Filter

all groups



Type

Version

Tags

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

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

Criando uma conexão



FIAP

Create Connection

DETAILS

SETTINGS

Name

Id

No value set

FlowFile Expiration ?

0 sec

Back Pressure Object Threshold ?

10000

Back Pressure Data Size Threshold ?

1 GB

Available Prioritizers ?

FirstInFirstOutPrioritizer

NewestFlowFileFirstPrioritizer

OldestFlowFileFirstPrioritizer

PriorityAttributePrioritizer

Selected Prioritizers ?

CANCEL

ADD

Data Provenance



NiFi Data Provenance

Displaying 193 of 193
Oldest event available: 07/18/2016 11:09:24 EDT

Showing the most recent events.

Filter	by component n...							
Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type			
07/18/2016 11:09:57.574 EDT	CONTENT_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	2.91 KB	ExecuteSQL	ExecuteSQL			
07/18/2016 11:10:04.224 EDT	ATTRIBUTES_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	2.91 KB	UpdateAttribute	UpdateAttribute			
07/18/2016 11:10:13.713 EDT	CONTENT_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	5.76 KB	ConvertAvroToORC	ConvertAvroToORC			
07/18/2016 11:10:40.352 EDT	SEND	0fe67e3c-9408-4eff-8772-252f1f4b399f	5.76 KB	PutHDFS	PutHDFS			
07/18/2016 11:10:40.352 EDT	ATTRIBUTES_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	5.76 KB	PutHDFS	PutHDFS			
07/18/2016 11:11:25.506 EDT	CONTENT_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	422 bytes	ReplaceText	ReplaceText			
07/18/2016 11:11:48.435 EDT	SEND	0fe67e3c-9408-4eff-8772-252f1f4b399f	422 bytes	PutHiveQL	PutHiveQL			
07/18/2016 11:11:48.871 EDT	DROP	0fe67e3c-9408-4eff-8772-252f1f4b399f	422 bytes	PutHiveQL	PutHiveQL			
07/18/2016 11:09:57.541 EDT	CONTENT_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	2.88 KB	ExecuteSQL	ExecuteSQL			
07/18/2016 11:10:04.224 EDT	ATTRIBUTES_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	2.88 KB	UpdateAttribute	UpdateAttribute			
07/18/2016 11:10:13.657 EDT	CONTENT_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	5.77 KB	ConvertAvroToORC	ConvertAvroToORC			
07/18/2016 11:10:40.162 EDT	SEND	1b105de2-eb64-4ff1-b465-4fd8844da437	5.77 KB	PutHDFS	PutHDFS			
07/18/2016 11:10:40.163 EDT	ATTRIBUTES_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	5.77 KB	PutHDFS	PutHDFS			
07/18/2016 11:11:25.505 EDT	CONTENT_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	422 bytes	ReplaceText	ReplaceText			
07/18/2016 11:11:46.354 EDT	SEND	1b105de2-eb64-4ff1-b465-4fd8844da437	422 bytes	PutHiveQL	PutHiveQL			
07/18/2016 11:11:48.871 EDT	DROP	1b105de2-eb64-4ff1-b465-4fd8844da437	422 bytes	PutHiveQL	PutHiveQL			
07/18/2016 11:09:57.567 EDT	CONTENT_MODIFIED	213f984e-e355-42a6-8632-78e1bc37bacf	2.9 KB	ExecuteSQL	ExecuteSQL			
07/18/2016 11:10:04.224 EDT	ATTRIBUTES_MODIFIED	213f984e-e355-42a6-8632-78e1bc37bacf	2.9 KB	UpdateAttribute	UpdateAttribute			
07/18/2016 11:10:13.699 EDT	CONTENT_MODIFIED	213f984e-e355-42a6-8632-78e1bc37bacf	5.76 KB	ConvertAvroToORC	ConvertAvroToORC			
07/18/2016 11:10:40.352 EDT	SEND	213f984e-e355-42a6-8632-78e1bc37bacf	5.76 KB	PutHDFS	PutHDFS			

Last updated: 11:28:49 EDT

Data Provenance

FIAP

Provenance Event

DETAILS

ATTRIBUTES

CONTENT

Time

07/29/2016 00:58:44.829 UTC

Event Duration

No value set

Lineage Duration

00:00:00.203

Type

ATTRIBUTES_MODIFIED

FlowFile Uuid

62d2161f-0b2a-4b2a-a552-ab617bef3811

File Size

1.1 KB

Component Id

7bba4f68-2861-3a12-aac6-60f12e11e215

Component Name

EvaluateJsonPath

Component Type

Parent FlowFiles (0)

No parents

Child FlowFiles (0)

No children

OK

Data Provenance

Provenance Event

DETAILS

ATTRIBUTES

CONTENT

Attribute Values

eventType

ATTRIBUTES_MODIFIED

No value previously set

filename

6320498487869637

newSize

1119

No value previously set

oldSize

1119

No value previously set

path

./

reporting.task.transaction.id

fc9fad99-89f0-4978-a3aa-571bb8b8851b

uuid

62d2161f-0b2a-4b2a-a552-ab617bef3811

☐ Show modified attributes only

OK

Provenance Event

- DETAILS
- ATTRIBUTES
- CONTENT

Input Claim

Container
default

Section
918

Identifier
1469753924663-275350

Offset
108834

Size
1.1 KB

 DOWNLOAD

 VIEW

Output Claim

Container
default

Section
918

Identifier
1469753924663-275350

Offset
108834

Size
1.1 KB

 DOWNLOAD

 VIEW

Replay

Connection Id
88970033-a406-33a2-b679-711d04de4a35

Referências

Gregor Hohpe. Enterprise Integration Patterns [online]. Retrieved: 27 Dec 2014, from: <http://www.enterpriseintegrationpatterns.com>

Wikipedia. Service Oriented Architecture [online]. Retrieved: 27 Dec 2014, from: http://en.wikipedia.org/wiki/Service-oriented_architecture

Eric Savitz. Welcome to the API Economy [online]. Forbes.com. Retrieved: 27 Dec 2014, from: <http://www.forbes.com/sites/ciocentral/2012/08/29/welcome-to-the-api-economy>

Adam Duvander. The rise of the API economy and consumer-led ecosystems [online]. thenextweb.com. Retrieved: 27 Dec 2014, from: <http://thenextweb.com/dd/2014/03/28/api-economy>

Wikipedia. Internet of Things [online]. Retrieved: 27 Dec 2014, from: http://en.wikipedia.org/wiki/Internet_of_Things

Wikipedia. Big Data [online]. Retrieved: 27 Dec 2014, from: http://en.wikipedia.org/wiki/Big_data

Wikipedia. Flow Based Programming [online]. Retrieved: 28 Dec 2014, from: http://en.wikipedia.org/wiki/Flow-based_programming#Concepts

Matt Welsh. Berkeley. SEDA: An Architecture for Well-Conditioned, Scalable Internet Services [online]. Retrieved: 18 Jan 2018, from: <http://www.mdw.la/papers/seda-sosp01.pdf>