



CONFLOSS
CONFERÊNCIA DE FREE/LIBRE E OPEN SOURCE SOFTWARE

Ambiente Livre
Big Data & Data Science

Automação de Fluxo de Dados em Cluster com Apache NiFi

Marcio Junior Vieira
CEO & Data Scientist, Ambiente Livre
Pesquisador da UNB.

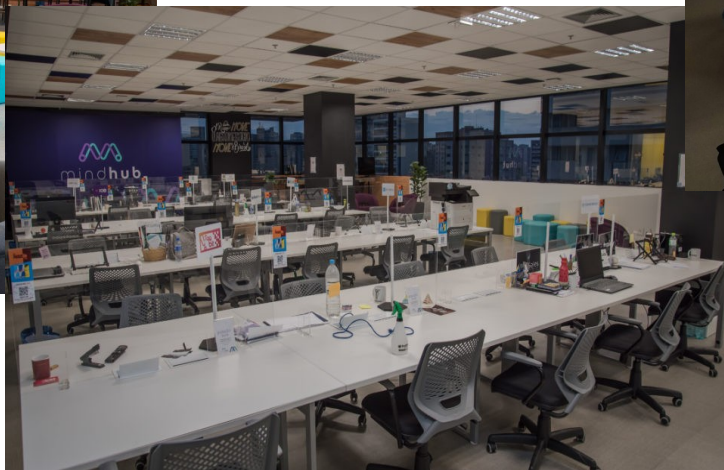
Mini-CV

- 22 anos de experiência em TI, vivência em desenvolvimento e análise de sistemas de gestão empresarial e ciência de dados.
- CEO da Ambiente Livre atuando como Cientista de Dados, Engenheiro de Dados e Arquiteto de Software.
- Professor dos MBAs em Big Data & Data Science, Inteligência Artificial e Business Intelligence e Analytics da Universidade Positivo.
- Professor do MBA Artificial Intelligence e Machine Learning da FIAP.
- Pesquisador do Laboratório de Tecnologias para Tomada de Decisão da Universidade de Brasília (Unb/Latitude).
- Trabalhando com Free Software e Open Source desde 2000 com serviços de consultoria e treinamento.
- Graduado em Tecnologia em Informática(2004) e pós-graduado em Software Livre(2005) ambos pela UFPR.
- Palestrante FLOSS em: FISL, TDC, Latinoware, Campus Party, Pentaho Day, Ticonova, PgDay e FTSL.
- Organizador Geral: Pentaho Day 2017, 2015, 2019 e apoio nas ed. 2013 e 2014.
- Data Scientist, instrutor e consultor de Big Data e Data Science com tecnologias abertas.
- Ajudou a capacitar equipes de Big Data na IBM, Accenture, Tivit, Serpro, Natura, MP, Netshoes, Embraer entre outras.
- Especialista em implantação e customização de Big Data com Hadoop, Spark, Pentaho, Cassandra e MongoDB.
- Contribuidor de projetos internacionais, tais como Pentaho, LimeSurvey, SuiteCRM e Camunda.
- Especialista em implantação e customização de ECM com Alfresco e BPM com Activiti, Flowable e Camunda.
- Certificado (Certified Pentaho Solutions) pela Hitachi Vantara (Pentaho).
- Membro da The Order Of de Bee (comunidade Alfresco para desenvolver o ecossistema Alfresco independente)
- Trabalha profissionalmente com NiFi desde 2019.

Sobre a Ambiente Livre

Open Software for Business

- Fundada em 2004 com foco em consultoria com FLOSS.
- Experts em 34 soluções para geração de negócios com Software Livre/Código Aberto.
- Atualmente estamos sediados no Hub de Inovação Mindhub em Curitiba (FAE).

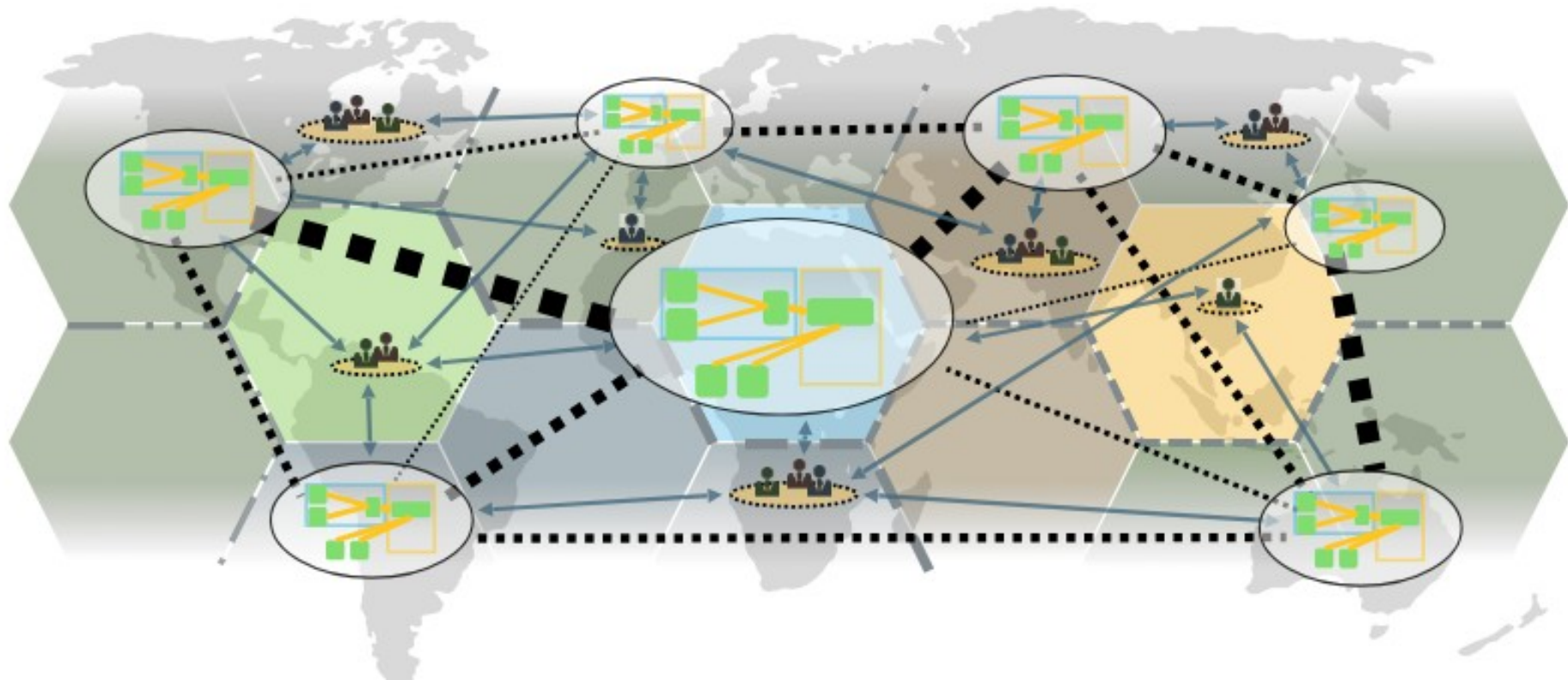


Nosso Ecossistema de Serviços

Big Data e Data Science	CRM e CMS	ECM e BPM	Business Intelligence
Análise de Dados da IoT Análise Preditiva Processamento Distribuído Banco de Dados Colunares Big Data & Data Lake Big Data Analytics Machine Learning Consultoria Treinamento Projeto	Marketing e Vendas Fidelização SAC e Pós-vendas Portais de Conteúdo Customer Relationship Management Content Management System Pesquisa de Mercado & SLA Consultoria Treinamento Projeto	Gestão de Documentos Gerenciamento de Mídias Processo de Negócio BPMN e BPMS Enterprise Content Management Records Management Business Process Management Consultoria Treinamento Projeto	Painéis de Indicadores Cubos de Análise Relatórios Gerenciais Tomada de Decisão Business Intelligence & Analytics Dashboards e OLAP Data Integration & Data Mining Consultoria Treinamento Projetos



Um problema: Transporte de dados!



Apache NiFi.

- É um projeto de software da Apache Software Foundation projetado para automatizar o fluxo de dados entre sistemas de software.
- O design do software é baseado no modelo de programação baseado em fluxo e oferece recursos que incluem a capacidade de operar em laptops como ou dentro de clusters.
- Segurança usando criptografia TLS.
- Extensibilidade.
- Suportar fluxos de dados altamente escaláveis e flexíveis.
- Recursos de usabilidade aprimorados como um portal que pode ser usado para visualizar e modificar o comportamento visualmente.
- Tem atualmente como principal desenvolvedora e suporte comercial a empresa Hortonworks (agora incorporada à Cloudera).



Visão de negócio.

- Automatiza a movimentação de dados entre diferentes fontes e sistemas de dados.
- Plataforma de logística de dados integrada para automatizar a movimentação de dados entre sistemas distintos.
- Fornece controle em tempo real que facilita o gerenciamento da movimentação de dados entre qualquer origem e qualquer destino.
- É agnóstico quanto à fonte de dados, suportando fontes díspares e distribuídas de diferentes formatos, esquemas, protocolos, velocidades e tamanhos, como máquinas, dispositivos de localização geográfica, fluxos de cliques, arquivos, feeds sociais, arquivos de log e vídeos, etc.
- Configurável para movimentação de dados, **semelhante a como a Fedex, UPS ou outros serviços de entrega de correio transportam pacotes.**
- E, assim como esses serviços, **permite que você rastreie seus dados em tempo real, da mesma forma que rastreia uma entrega.**



Apache NiFi.

- Baseia-se no software "NiagaraFiles" anteriormente desenvolvido pela NSA, que também dá origem a uma parte do seu nome atual – NiFi.
- Transformou-se em código aberto como parte do programa de transferência de tecnologia da NSA em 2014.
- A empresa criadora inicialmente do Software foi a Onyara Inc que foi adquirida pela Hortonworks em 2015. (*1) e hoje incorporada a Cloudera.



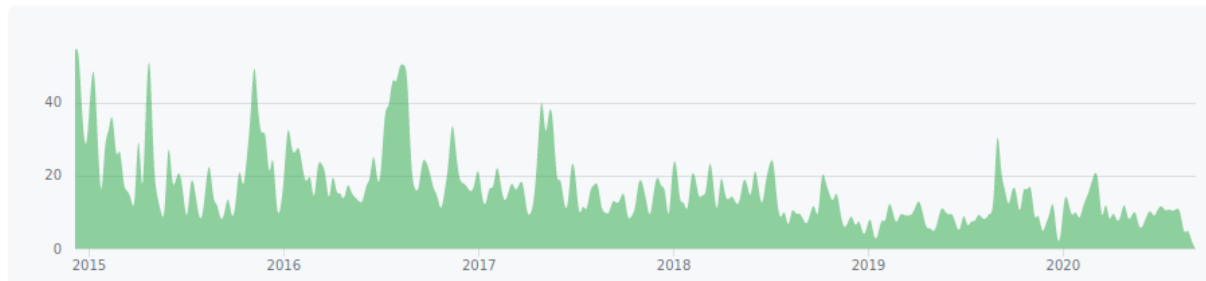
Apache NiFi.

- A última versão estável é a 1.15 lançada em 07/11/2021 (nifi-1.15.0)
- Aproximadamente uma versão a cada 3 meses.
- Mais de 370 desenvolvedores contribuidores do projeto.

Dec 7, 2014 – Sep 7, 2020

Contributions: Commits ▼

Contributions to main, excluding merge commits



Definição

- 491 Projetos Open Source.
- +7000 Committers, e com uma média de 50 novos mensais... Seja um!
- Data Science = Apache = Open Source
- **Apache é líder em Big Data e Data Science!**
- ~49 projetos da linha “Big Data” incluindo “Apache Hadoop” e “Spark”
- ~25 projetos de database incluindo “Apache Cassandra”



Patrocinadores da Apache Software Foundation.

PLATINUM SPONSORS:



LeaseWeb



Facebook



Amazon Web Services



Pineapple Fund



Verizon Media



Tencent



Google



Huawei



Comcast

GOLD SPONSORS:

Anonymous



Baidu



Bloomberg



Cloudera



Handshake



IBM



Union Investment



Workday

SILVER SPONSORS:



Aetna



Alibaba Cloud Computing



Budget Direct



Capital One



Cerner



Inspur

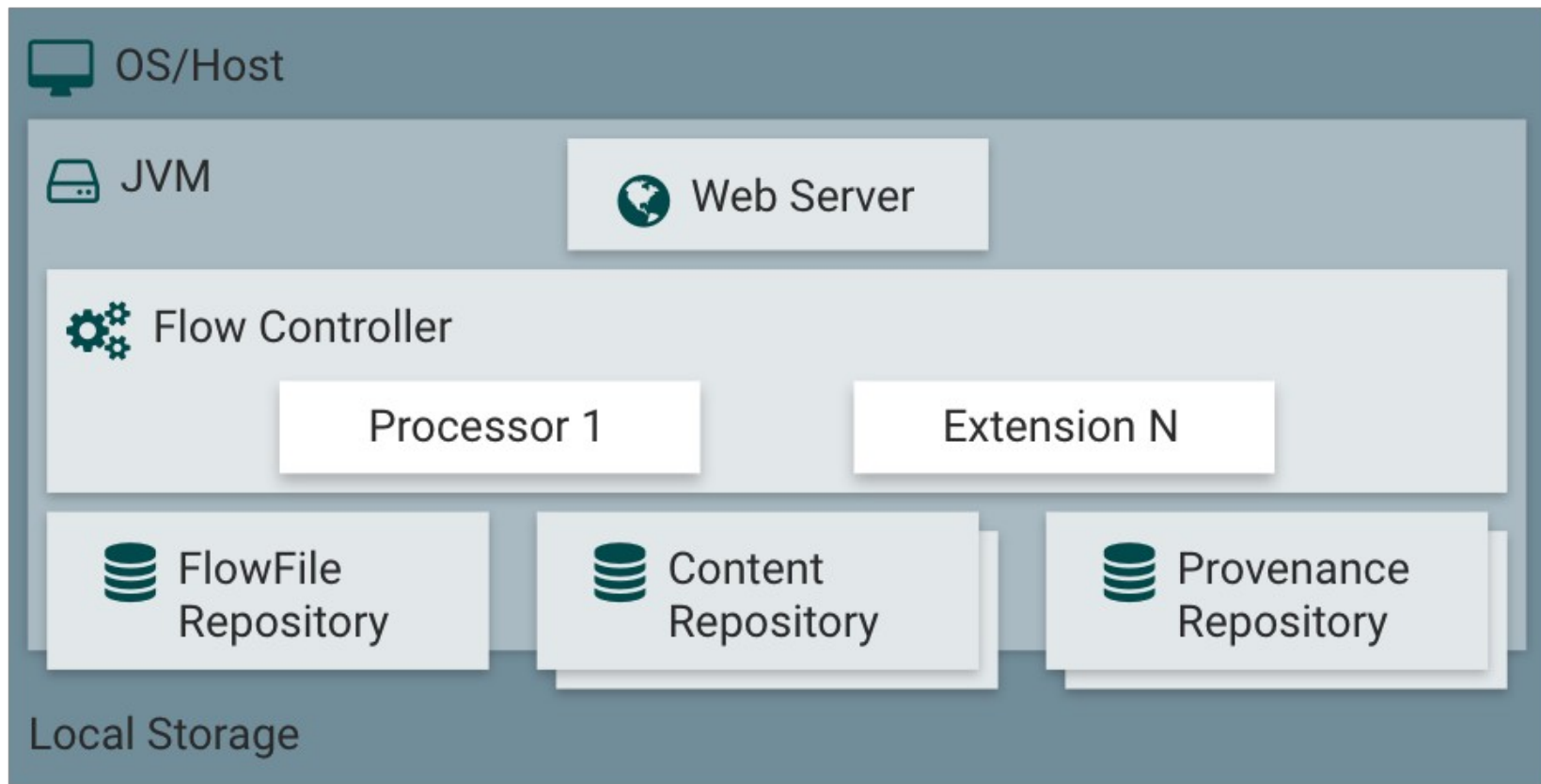


Red Hat, Inc.

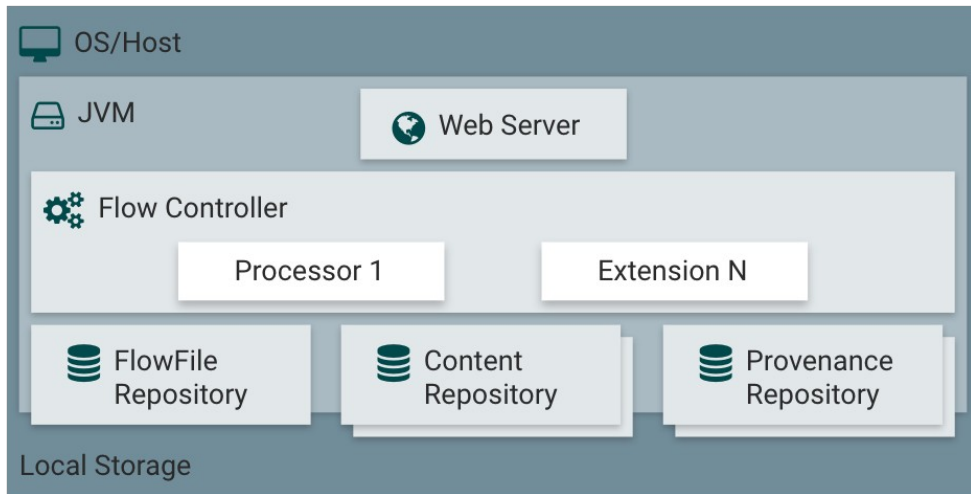


Target

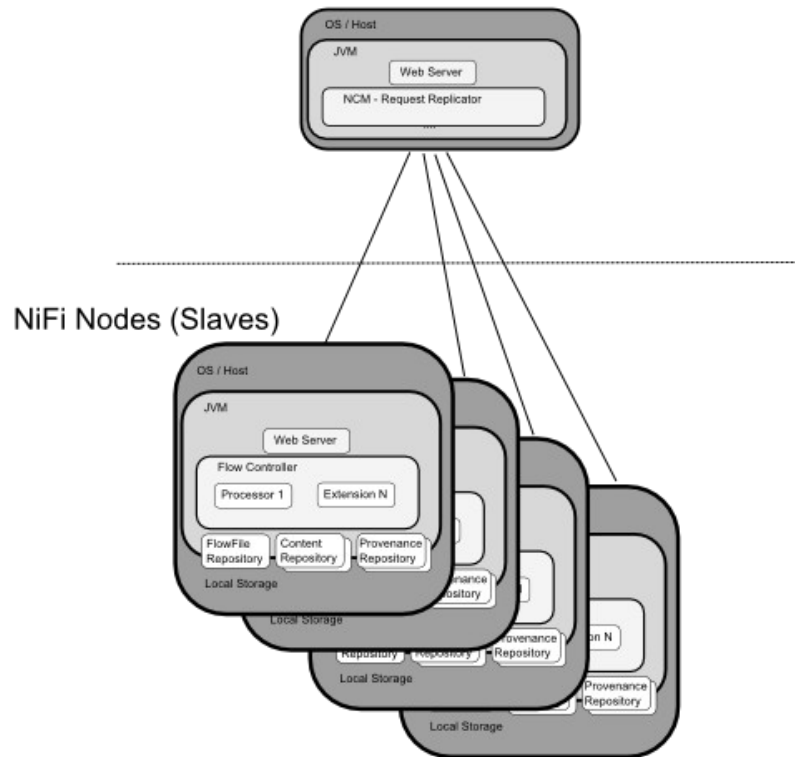
Arquitetura do Apache NiFi.



Arquitetura do Apache NiFi - Cluster



NiFi Cluster Manager (Master)



NiFi x PDI x HOP x Talend x Sqoop.

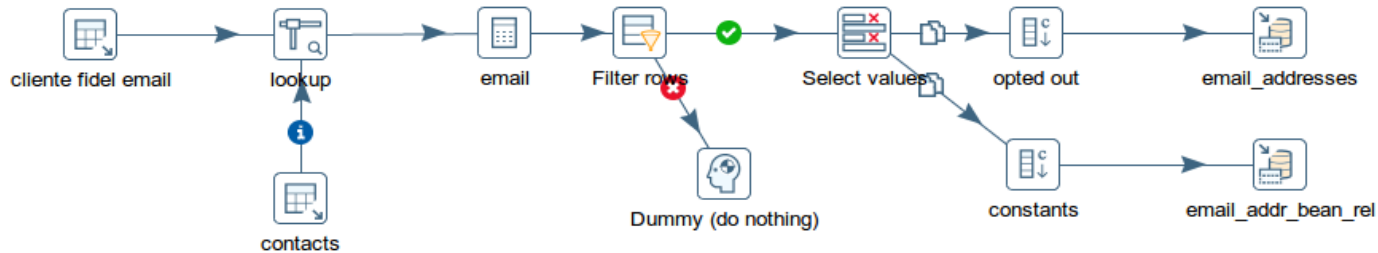
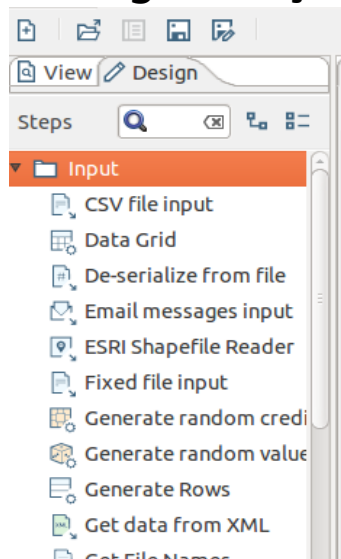
- Opções Open Source para Data Flow ou ETL ou Stream de dados!



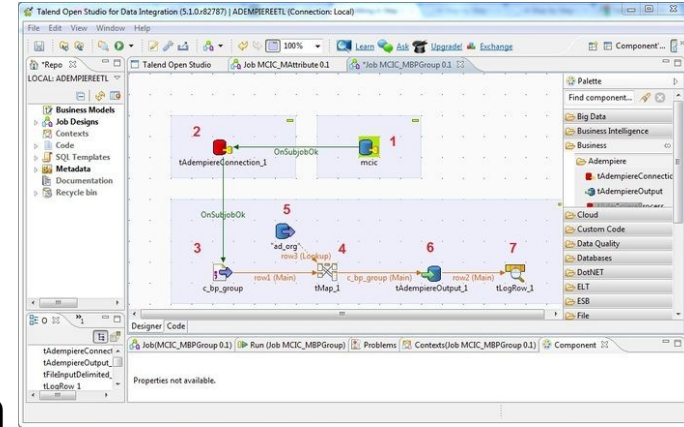
APACHE SQOOP



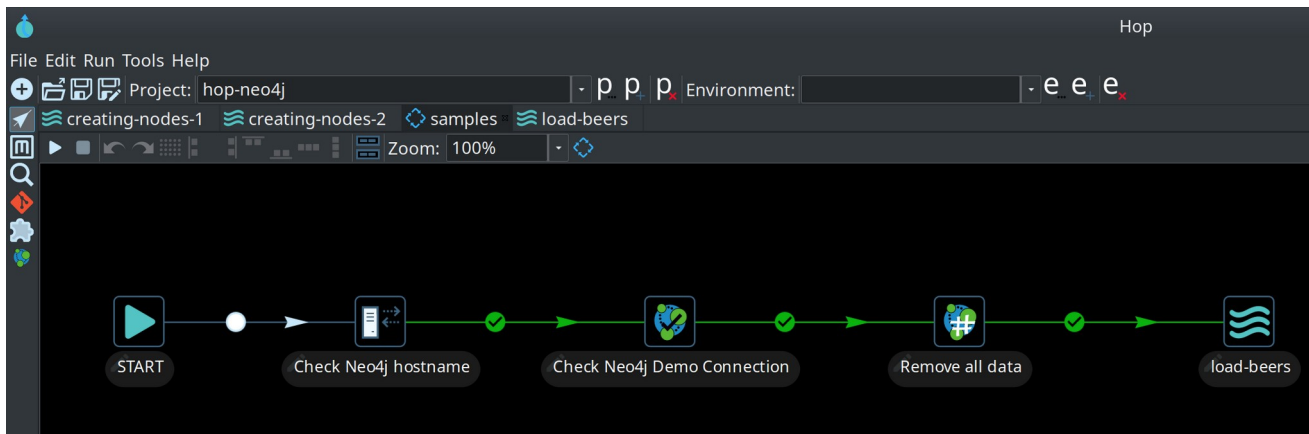
- Processa em Paralelo (Também roda em Cluster Spark)
- Acessar dados diretamente (se necessário sem DW)
- Permite publicar dados diretamente em Reports, Ad-Hoc Reports e Dashboards.
- “Programação e Fluxo Visual” com aproximadamente 350 steps



- Muito Similar ao Pentaho Data Integration.
- Funcionalidades interessantes como detecção de registros duplicado, governança de data quality não existem na versão Community.
- Algumas funcionalidades que só existem na EE podem ser supridas por integrações porém funcionalidade.
- Abaixo a lista completa (na visão da Talend) das comparações , alguns recursos que não existem na community podem ser suprimidos por ferramentas acessórios open source.
- **Recentemente comprado pela Qlik e Descontinuando a CE**
<https://www.talend.com/products/mdm/mdm-compare-all/>



- Acronimo de: **H**op **O**rquestration **P**lataform
- Orquestração
 - **Dados** – Pipelines e Workflows.
 - **Metadata** – Edição, Manuseio e gerenciamento.
 - **Insights**: Execução e tratamento de dados, log do processo.
- - **Configurações**: Manuseio de ecossistemas complexos.



- Apache Sqoop – Uma abordagem prática para a unificação de dados.
- Abreviação de “SQL para Hadoop”

Objetivo

- executar a transferência eficiente e bidirecional de dados entre o Hadoop e diversos serviços de armazenamento externo de dados estruturados



Usuários Apache Nifi no Brasil

3.262 no Mundo



americanas.com



Confidencial

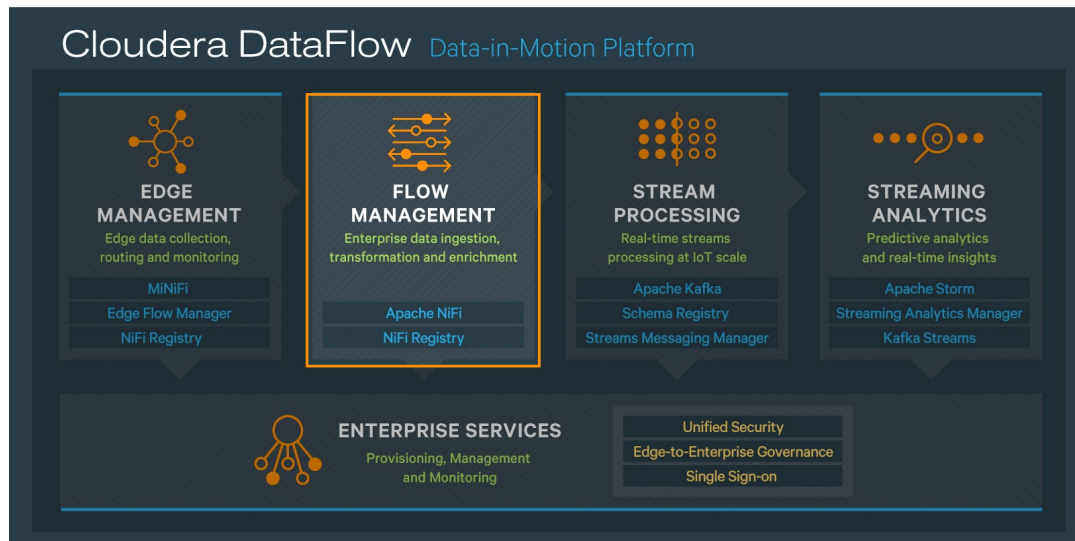
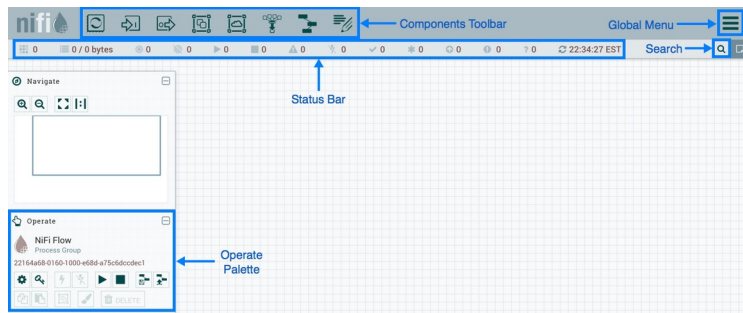


FINTECH
FINANCIAL TECHNOLOGY

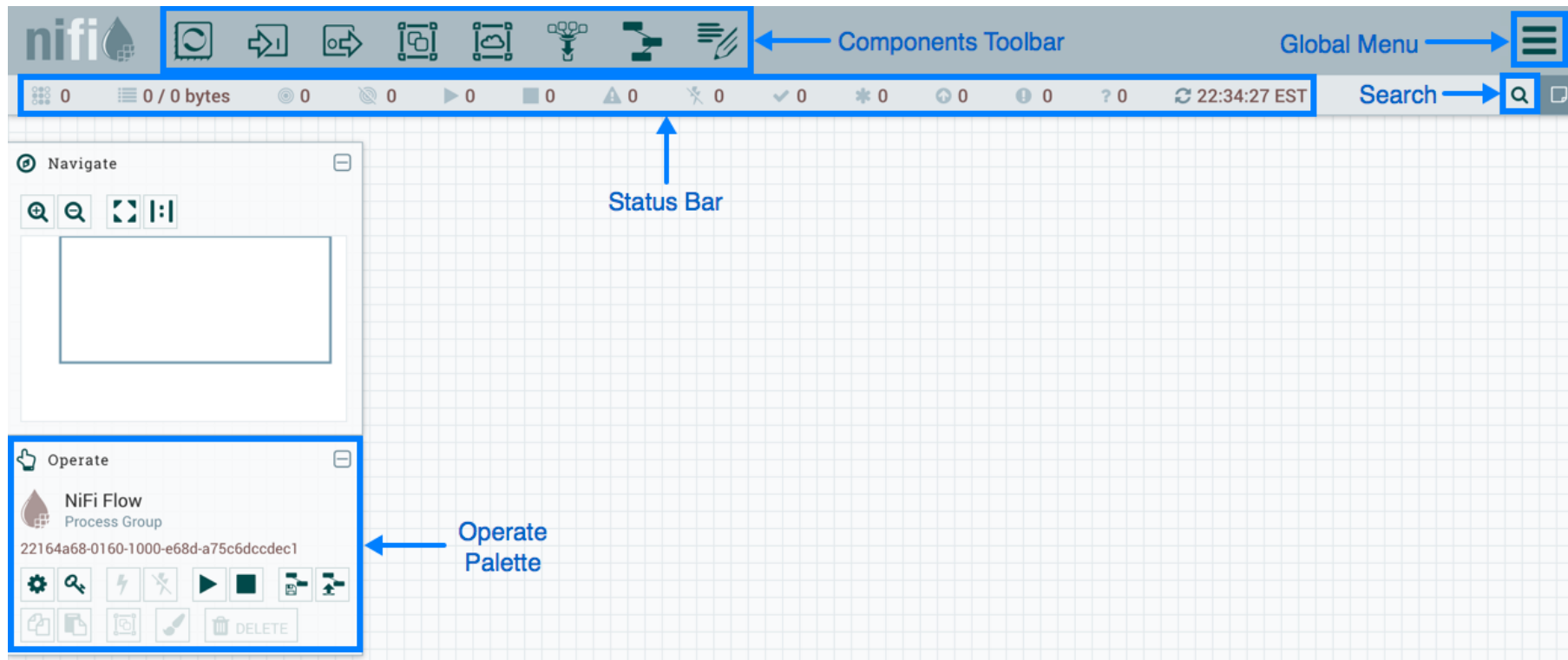
Todos Usuários
CLOUDERA

DataFlow Manager

- Um **DataFlow Manager (DFM)** é um usuário NiFi que tem permissão para adicionar, remover e modificar componentes de um fluxo de dados NiFi.



Interface gráfica do NiFi.



Processador.

- São os componentes responsáveis por processar os dados.
- Existem diversos tipos de processador.
- Existem algumas tags para agrupá-los
- São 288 processadores nativos (ver.1.12)

Add Processor

Source: all groups ▼

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, Ingest, 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.

CANCELADD

Configuração do Processador

- Cada processador tem propriedades específicas.
- Podem ser agendadas as formas de execução.
- Podem ser adicionados comentários.
- Os processador podem ficar em execução ou parados.

Processor Details

▶ Running

⚙ STOP & CONFIGURE

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Name

GetFile CSV Cities

Id

6b54b505-0174-1000-ac06-6307450a3aca

Type

GetFile 1.12.0

Bundle

org.apache.nifi - nifi-standard-nar

Penalty Duration ⓘ

30 sec

Yield Duration ⓘ

1 sec

Bulletin Level ⓘ

WARN

Automatically Terminate Relationships ⓘ

success

All files are routed to success

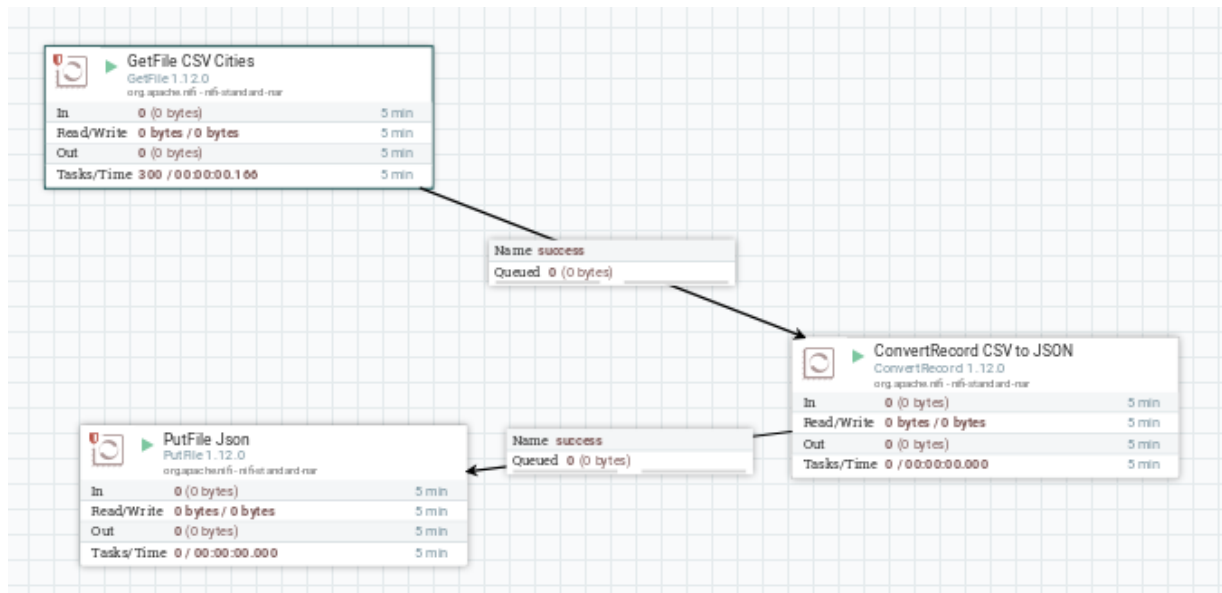
Grupos de Processador.

- Os grupos de processadores são formas de organizar seus processadores.
- É uma divisão lógica do Apache NiFi.
- O primeiro acesso ao NiFi é aberto o Process Group principal.



Conexões

- São responsáveis pelo direcionamento do fluxo do dados.
- São relacionados por possíveis caminhos como a serem seguidos:
 - * Sucesso.
 - * Falha.
 - * Outros dependendo do Processor.



Configuração do Processador

- Processados pelo Nifi e são mantidos em um mapa hash na memória da JVM.
- Isso o torna muito eficiente para processá-los.
- O mesmo pode gravar cache em disco.
- O Repositório FlowFile é um "Log Write-Ahead" (ou registro de dados) dos metadados de cada um dos FlowFiles que existem atualmente no sistema.
- Um FlowFile pode ter diversos formatos.
- O Tamanho pode ser configurado
- O Número de linhas pode ser configurado.
- O conteúdo pode ser criptografado.

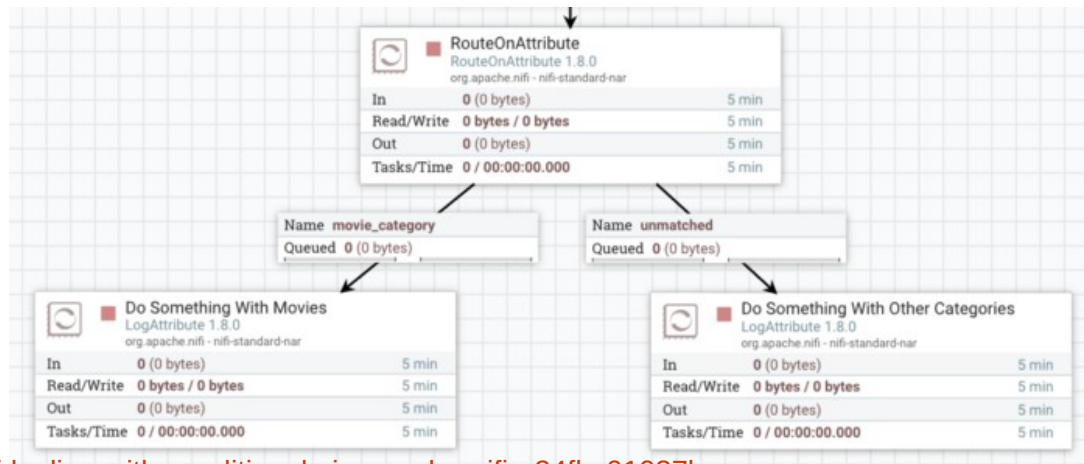


- **Types**
 - Events
 - Objects
 - Files
 - Messages
 - Media
- **Formats**
 - JSON
 - Avro
 - Text
 - Mp4
 - Proprietary
- **Sizes**
 - Bytes to GBs

Processador de Condição.

- Permite verificar o dado e tomar decisão do roteamento do fluxo.
- Utiliza da NiFi Expression Language.

```
${category:equals('movies')}
```



<https://medium.com/mpharma-tech-blog/dealing-with-conditionals-in-apache-nifi-e24fbc01037b>

Processador de Condição.

- Quando processamos os FlowFiles eles gera atributos.
- O atributos gerados podem ser acessados pela NiFi EL.
- Exemplos:

```
${category:equals('movies')}
```

```
/opt/hacks/${filename}
```

```
${filename:toUpper()}
```

```
${"my attribute"}
```

Apache Nifi Expression Language Cheat Sheet

Reserved Characters	Logic Operators	Text Search
<p>If these characters are present in attribute names they need to be quoted</p> <p><code>\$ { } () [] , : ; / * ' (space) \t \r \n</code></p> <p>Ex. <code>\${'a:attribute name'}</code> <code>\${"a:attribute name"}</code></p>	<p><code>isNull()</code> <code>\${filename:isNull()}</code></p> <p><code>notNull()</code> <code>\${filename:notNull()}</code></p> <p><code>isEmpty()</code> <code>\${literal('') :isEmpty()}</code></p> <p><code>equals(string)</code> <code>\${filename :equals('value')}</code></p> <p><code>equalsIgnoreCase (string)</code> <code>\${filename :equalsIgnoreCase('v')}</code></p> <p><code>gt(number)</code> <code>\${fileSize:gt(64)}</code></p> <p><code>ge(number)</code> <code>\${fileSize:ge(64)}</code></p> <p><code>lt(number)</code> <code>\${fileSize:lt(64)}</code></p> <p><code>le(number)</code> <code>\${fileSize:le(64)}</code></p> <p><code>and(bool)</code> <code>\${fileSize:gt(1) :and({fileSize:lt(4)})}</code></p> <p><code>or(bool)</code> <code>\${fileSize:lt(1) :or({fileSize:gt(4)})}</code></p> <p><code>not()</code> <code>\${filename :endsWith('sv'):not()}}</code></p> <p><code>ifElse ('true val', 'falseval')</code> <code>\${filename :endsWith('csv') :ifElse('is csv', 'is not csv')}</code></p>	<p><code>filename:equals('fizz buzz bazz.txt')</code></p> <p><code>startsWith (string)</code> <code>\${filename :startsWith('fizz')}</code></p> <p><code>endsWith (string)</code> <code>\${filename :endsWith('txt')}</code></p> <p><code>Contains (string)</code> <code>\${filename :contains('buzz')}</code></p> <p><code>in(string, string...)</code> <code>\${literal('NO') :in('NO', 'NOT')}</code></p> <p><code>indexOf(string)</code> <code>\${filename :indexOf('buzz')} == 5</code></p> <p><code>lastIndexOf (string)</code> <code>\${filename :lastIndexOf('z')} == 13</code></p> <p><code>find(regex)</code> <code>\${filename:find('.*zz')}</code></p> <p><code>matches(regex)</code> <code>\${filename :matches('fizz.*txt')}</code></p> <p><code>jsonPath(path)</code> <code>\${theJson :jsonPath('\$.attribute')}</code></p>
Type Conversion		
<p>Coerces from one format to another</p> <p><code>toString()</code> <code>\${literal(2):toString() :equals('2')}</code></p> <p><code>toNumber()</code> <code>\${literal('2'):toNumber() :equals(2)}</code></p> <p><code>toDecimal()</code> <code>\${filesize:toDecimal()}</code></p>		
Mathematical		
<p><code>plus()</code> <code>\${fileSize:plus(10)}</code></p> <p><code>minus()</code> <code>\${fileSize:minus(10)}</code></p> <p><code>multiply()</code> <code>\${fileSize:multiply(10)}</code></p> <p><code>divide()</code> <code>\${fileSize:divide(10)}</code></p> <p><code>mod()</code> <code>\${fileSize:mod(10)}</code></p>		
Utilities		
		<p>These subjectless functions provide useful utilities.</p> <p><code>ip()</code> local ip</p> <p><code>hostname(bool)</code> <code>\${hostname(true)}</code> fully qualified hostname</p>

Services

- Os controllers services são divididos por grupos.
- Atualmente são mais de 80 Controllers Services.

Add Controller Service

Source

all groups

Displaying 89 of 89

Filter

Type ^	Version	Tags
ADLSCredentialsController...	1.12.0	cloud, credentials, adls, storag...
AWSCredentialsProviderCo...	1.12.0	credentials, provider, aws
AvroReader	1.12.0	comma, reader, record, values, ...
AvroRecordSetWriter	1.12.0	result, set, record, serializer, rec...
AvroSchemaRegistry	1.12.0	schema, registry, csv, json, avro
AzureStorageCredentialsCo...	1.12.0	cloud, blob, credentials, storag...
AzureStorageCredentialsCo...	1.12.0	cloud, blob, credentials, storag...
CSVReader	1.12.0	comma, reader, csv, record, val...
CSVRecordLookupService	1.12.0	lookup, cache, csv, record, relo...
CSVRecordSetWriter	1.12.0	result, set, tab, tsv, csv, record, ...
CassandraSessionProvider	1.12.0	database, pooling, cassandra, ...
ConfluentSchemaRegistry	1.12.0	schema, registry, confluent, kaf...

ADLSCredentialsControllerService 1.12.0 org.apache.nifi - nifi-azure-nar
Defines credentials for ADLS processors.

CANCEL

ADD

Apache NiFi - Controller Services Types



Microsoft Azure
Blob Storage



Cassandra



Couchbase



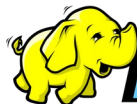
{JSON}
JavaScript Object Notation



elasticsearch



CONFLUENT



hadoop



redis



Google Cloud Platform



HIVE



Microsoft Azure

{REST:API}

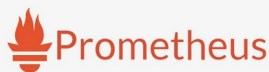
Java Message
Service



jetty://



APACHE
kafka®

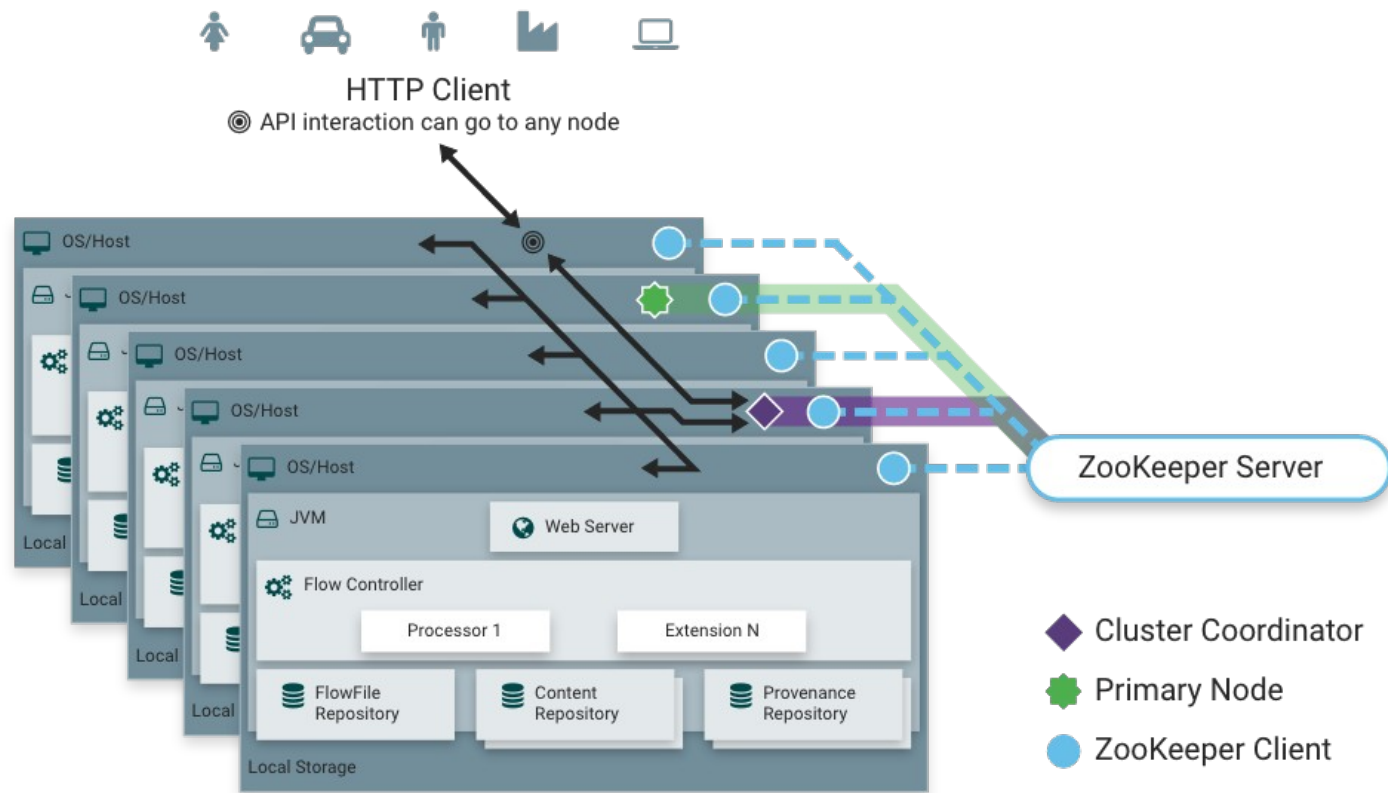


mongoDB®

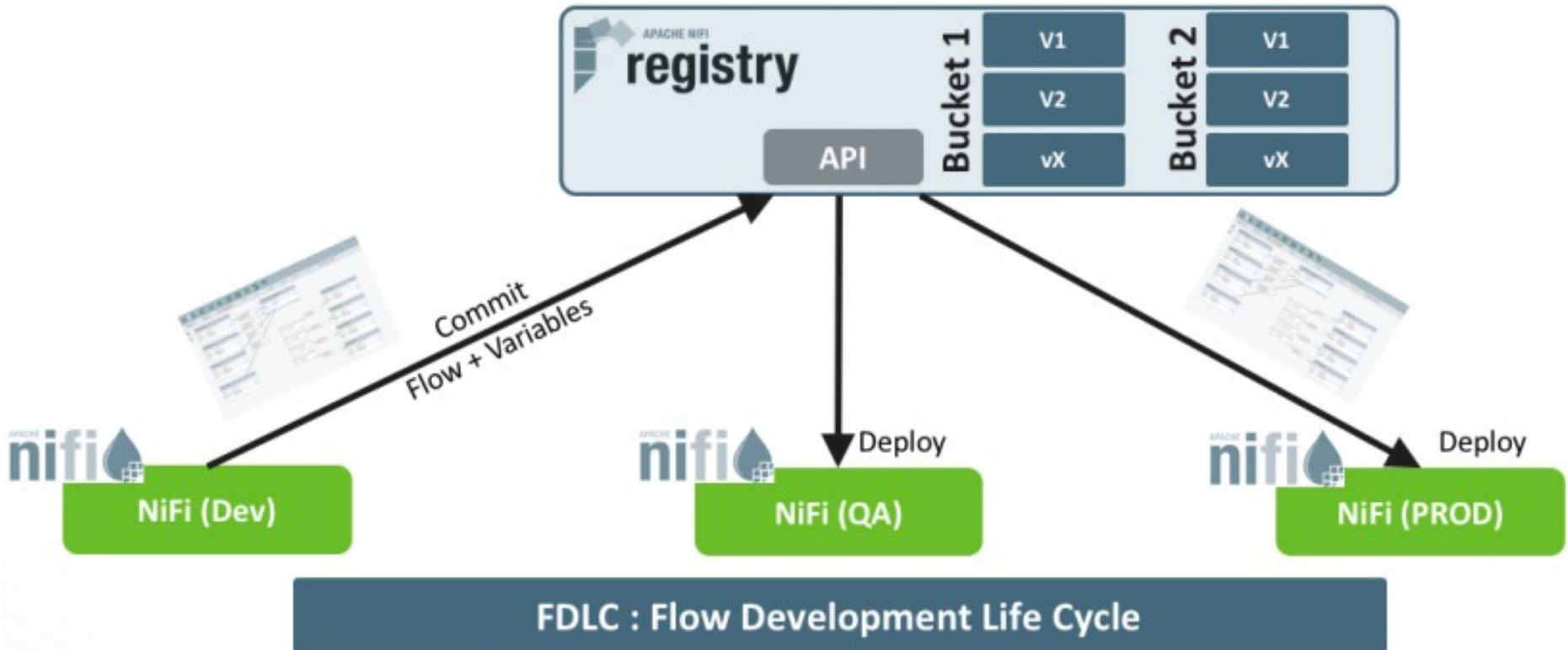


Parquet

Nifi Cluster

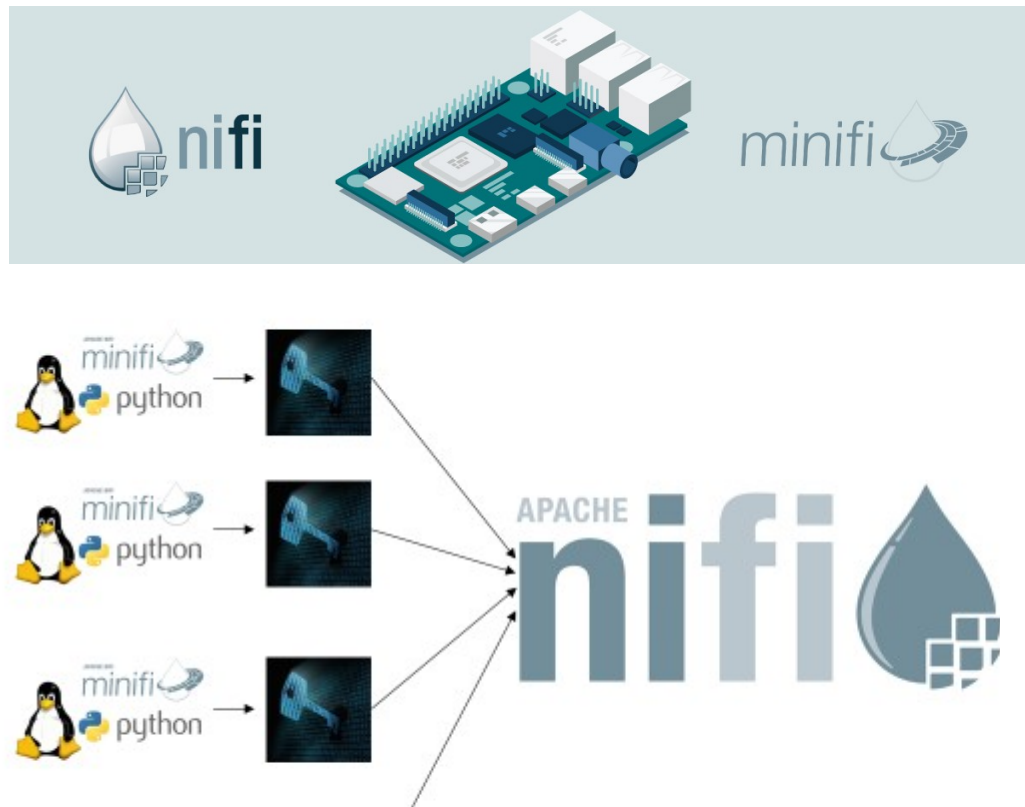


Nifi Registry e Flow Life Cycle

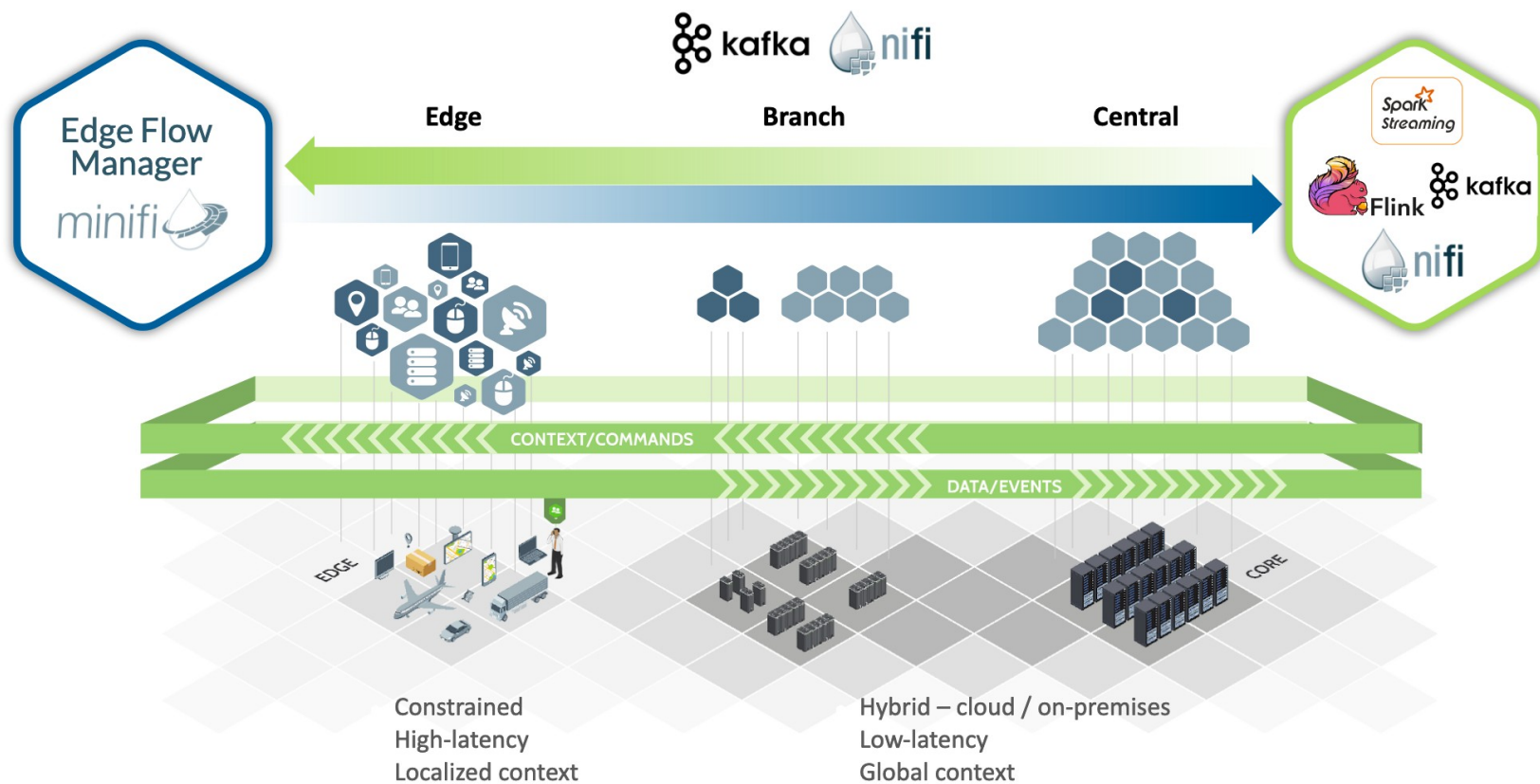


Conceito

- Versão leve do Apache NiFi, projetada para coleta eficiente de dados em dispositivos de borda ou ambientes com recursos limitados.
- Permite a transmissão segura e em tempo real de dados para um servidor NiFi centralizado.
- **Suporte Flexível:** Oferece configuração em Java, C++, Python, etc, adaptando-se a diferentes casos de uso.
- Ideal para **IoT**.



Minifi e Case



Obrigado



Marcio Junior Vieira

marcio@ambientelivre.com.br

@marviojvieira @ambientelivre

<https://www.linkedin.com/in/mvieira1/>

Blog: <http://blogs.ambientelivre.com.br/marcio/>

Slides:

<https://github.com/ambientelivre/labs>