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.

Marcio Junior Vieira

Ambiente Livre
Big Data & Data Science

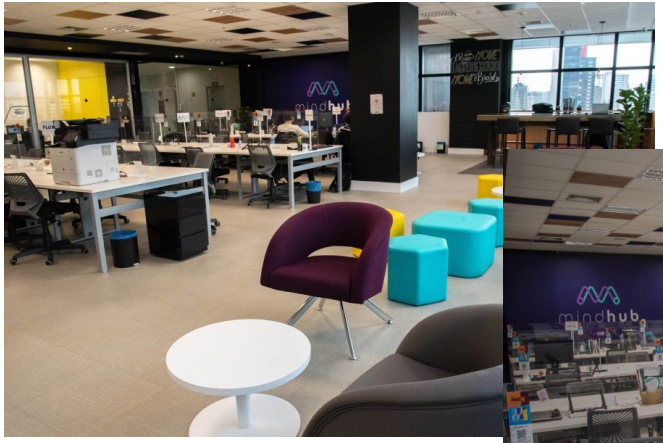
Mini-CV

- 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, Ticnova, 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

Ambiente Livre
Big Data & Data Science

Big Data e Data Science

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

CRM e CMS

Marketing e Vendas
Fidelização
SAC e Pós-vendas
Portais de Conteúdo

Customer Relationship Management
Content Management System
Pesquisa de Mercado & SLA

Consultoria | Treinamento | Projeto

ECM e BPM

Gestão de Documentos
Gerenciamento de Mídias
Processo de Negócio
BPMN e BPMS

Enterprise Content Management
Records Management
Business Process Management

Consultoria | Treinamento | Projeto

Business Intelligence

Painéis de Indicadores
Cubos de Análise
Relatórios Gerênciais
Tomada de Decisão

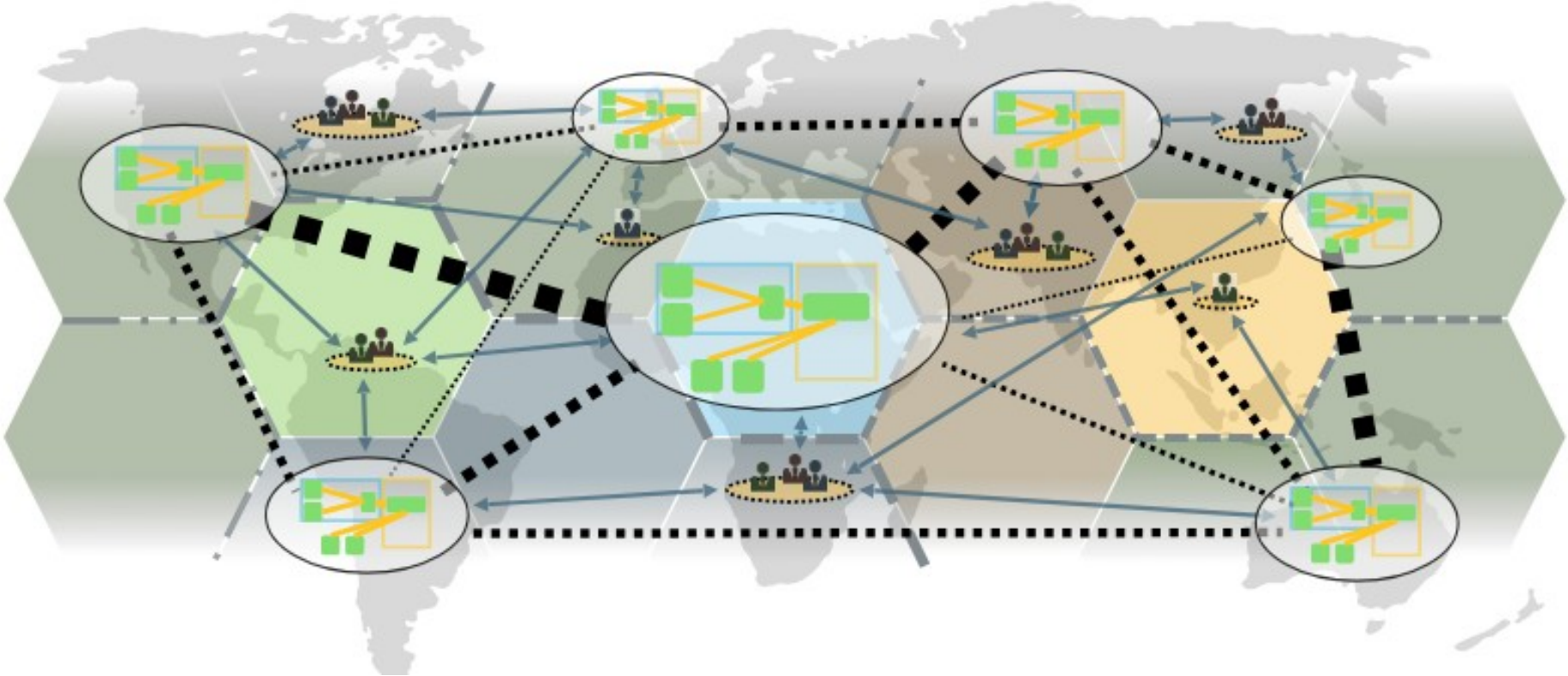
Business Intelligence & Analitycs
Dashboards e OLAP
Data Integration & Data Mining

Consultoria | Treinamento | Projetos

pentaho
Alfresco
hadoop
Apache HOP
Apache Superset
RANCHER
docker
kubernetes
LIFERAY
Enterprise. Open Source. For Life.
SCYLLA
Cassandra
suiteCRM
LPI
Spark
Apache Zeppelin
Grafana
Flowable
Activiti
LimeSurvey
PostgreSQL
the world's most advanced open source database
Apache Flink
Airflow
Metabase
zeebe
by Camunda
camunda
APACHE nifi
APACHE HBASE
mongoDB.
moodle
MariaDB
Drupal

Um problema: Transporte de dados!

Ambiente Livre
Big Data & Data Science



O Projeto Apache NiFi.

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).

Apache NiFi.

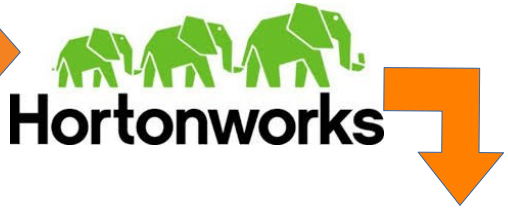
Visão de negócio.

- 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**.

Histórico do Apache NiFi.

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.

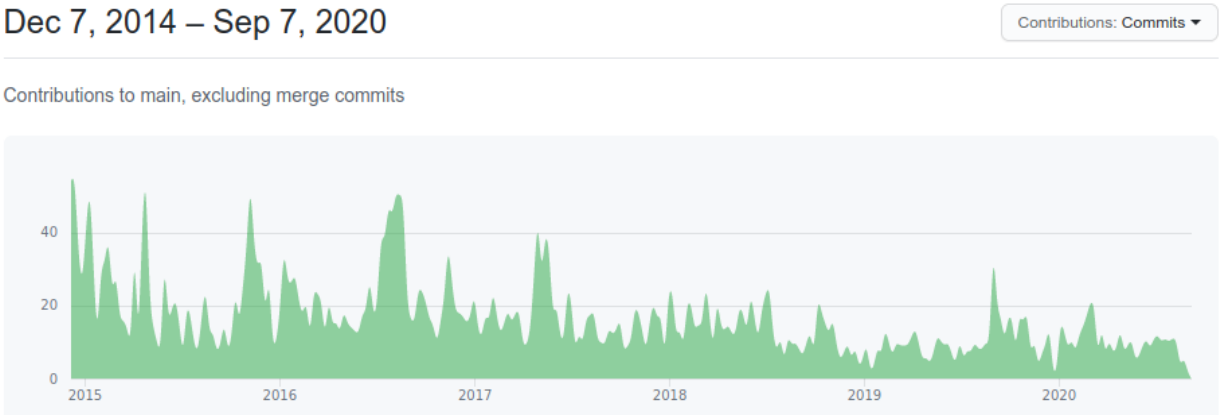


Fontes:
https://www.convergedigest.com/2015/08/hortonworks-to-acquire-onyara-for.html

Versões do Apache NiFi.

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.



Apache Software Foundation

Ambiente Livre
Big Data & Data Science

Definição

- 491 Projetos Open Source.
- +7000 Committers, e com uma média de 50 novos mensais… Seja um!
- Data Science = Apache = Open Source
- **Apache é lider 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.

Ambiente Livre
Big Data & Data Science

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

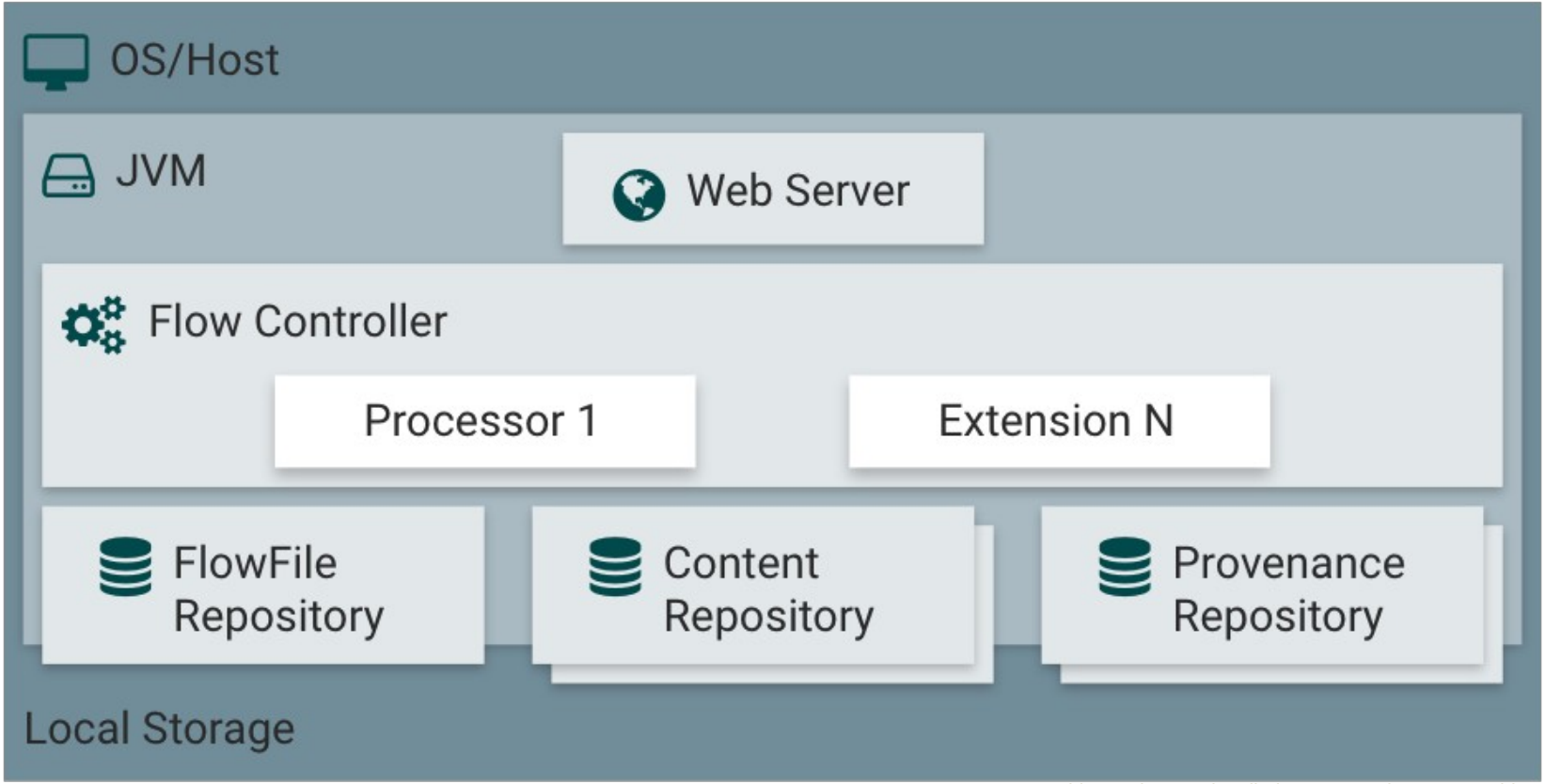
Cerner

Inspur

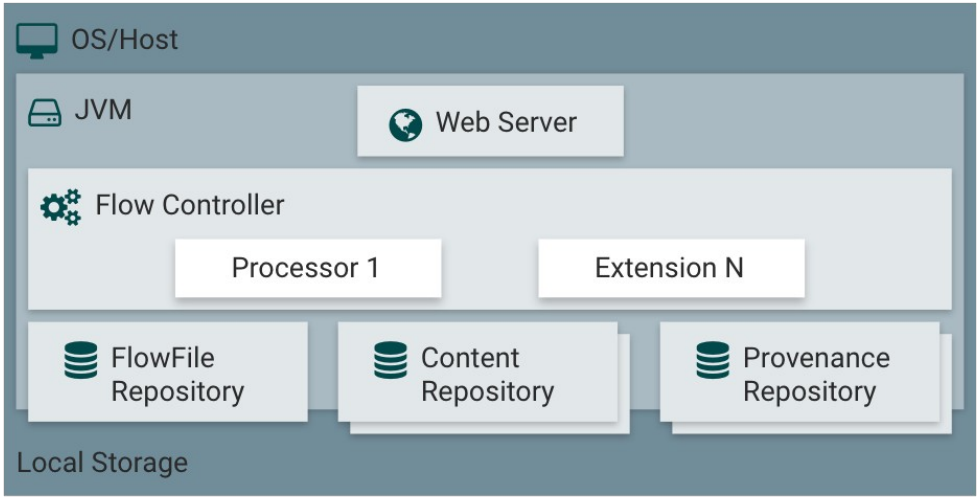

Red Hat, Inc.

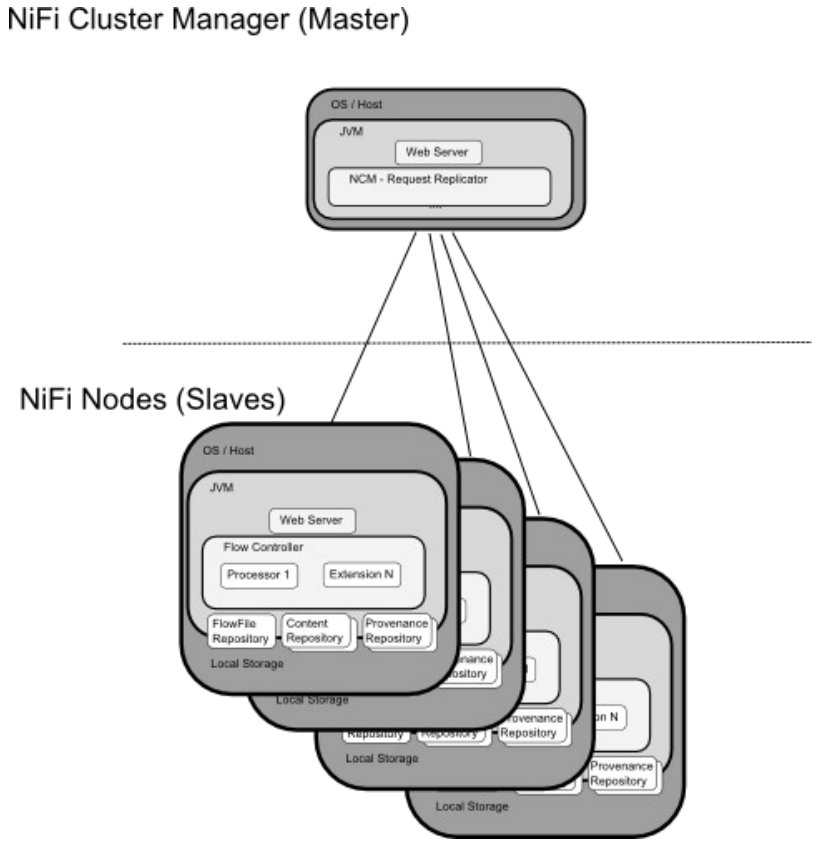
Target

Arquitetura do Apache NiFi.



Arquitetura do Apache NiFi. - Cluster



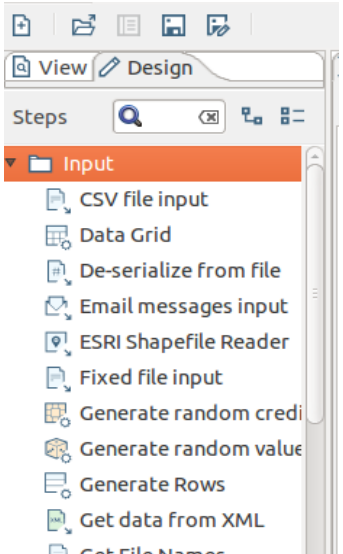


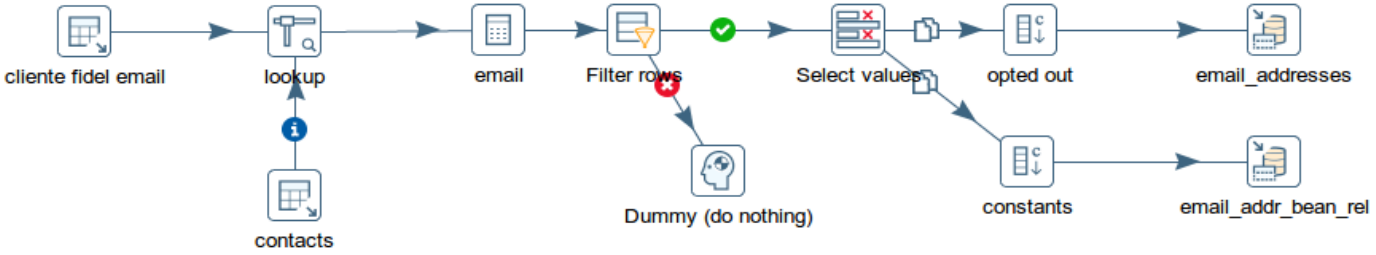
NiFi x PDI x HOP x Talend x Sqoop.

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

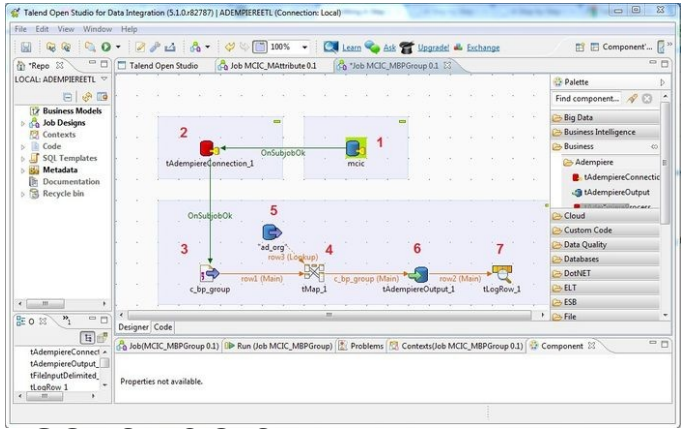
Pentaho Data Integration (PDI)

- 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





Talend



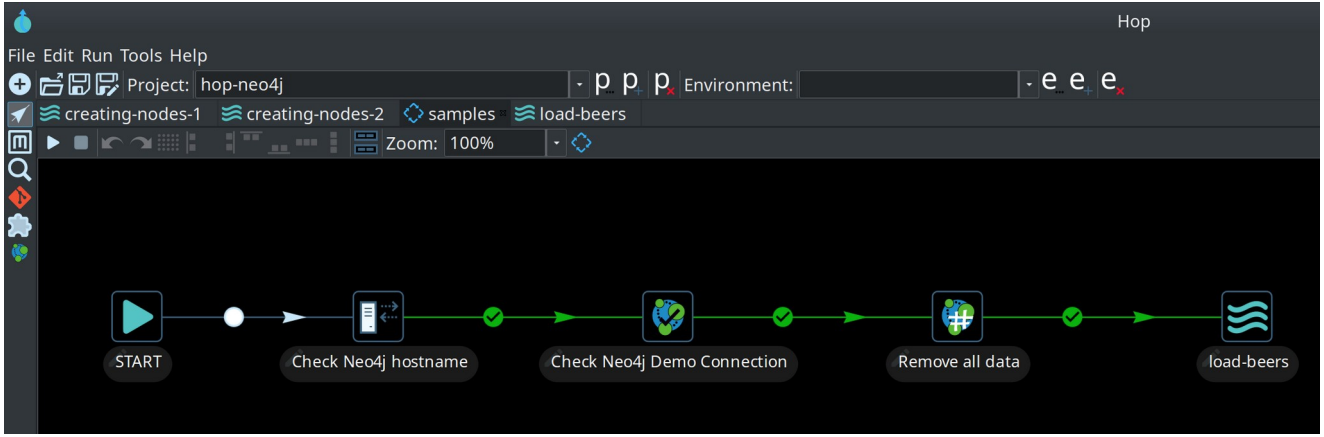
- Muito Similar ao Pentaho Data Integration.
- Funcionalidades interessantes como detecção de registros duplicado, governança de data quality não existem na versão Community.

- 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/

Apache Hop

- 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

- 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

Ambiente Livre
Big Data & Data Science

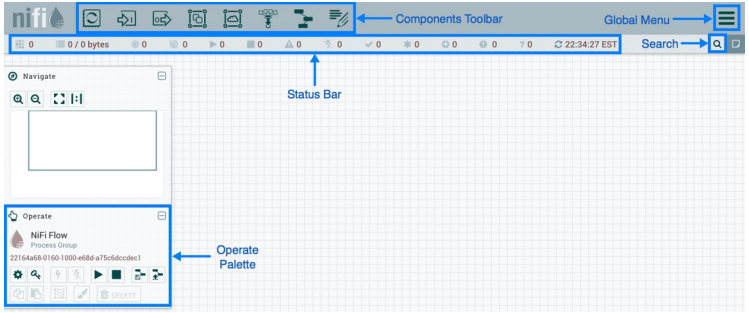
Confidêncial

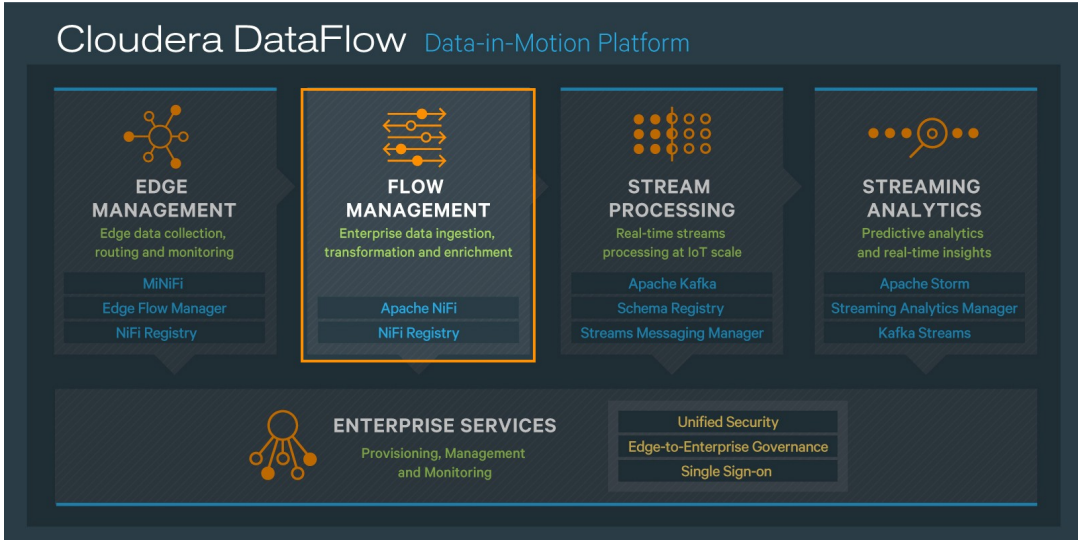
Todos Usuários

Terminologias.

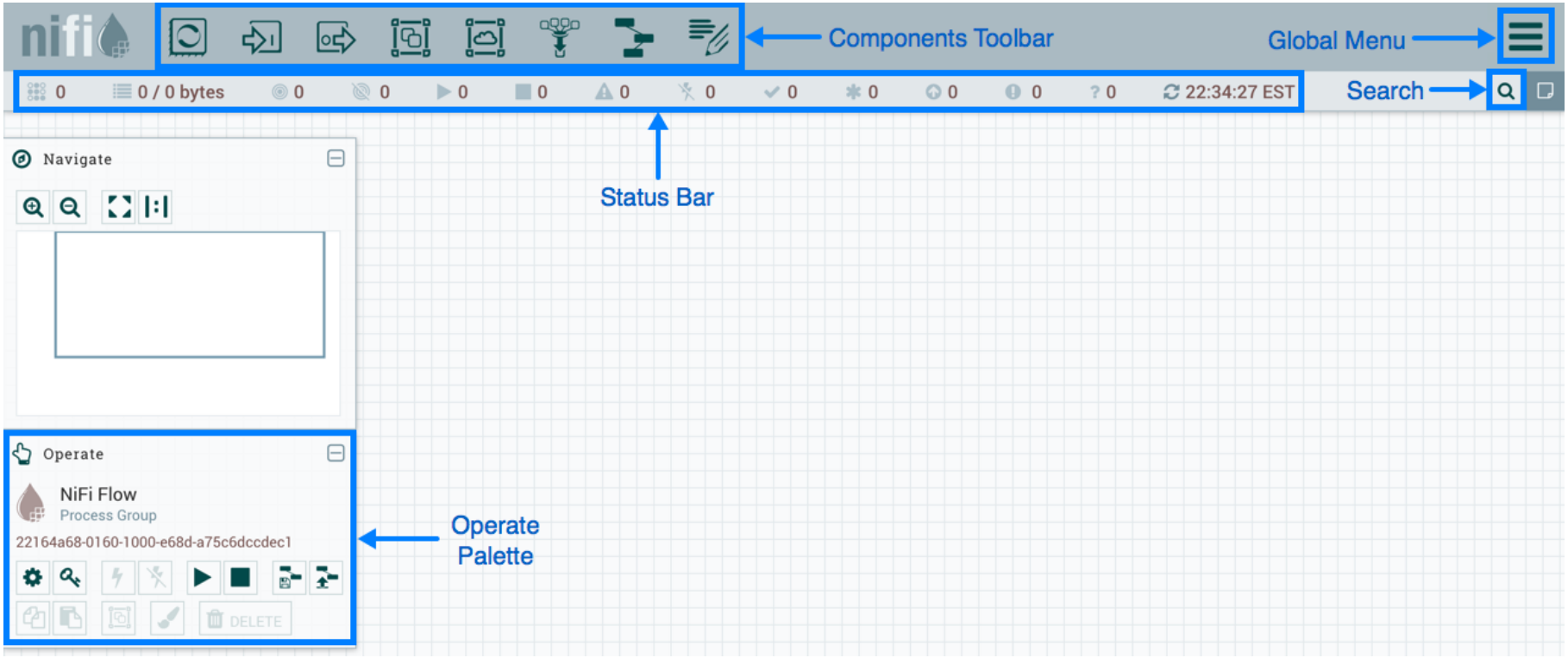
DataFlow Manager

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





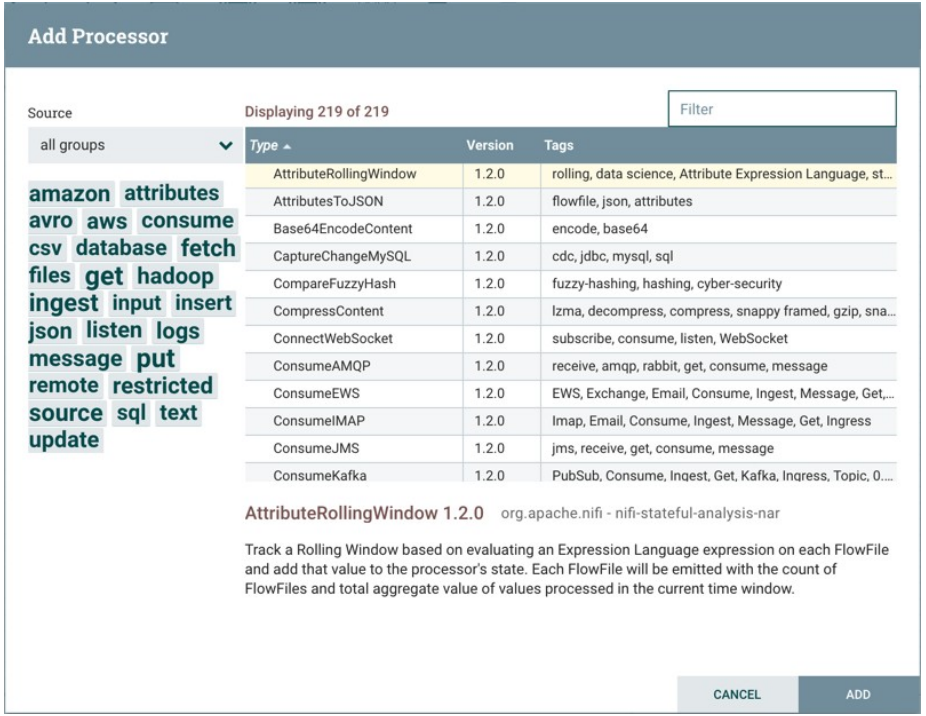
Interface gráfica do NiFi.



Processor

Processador.

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

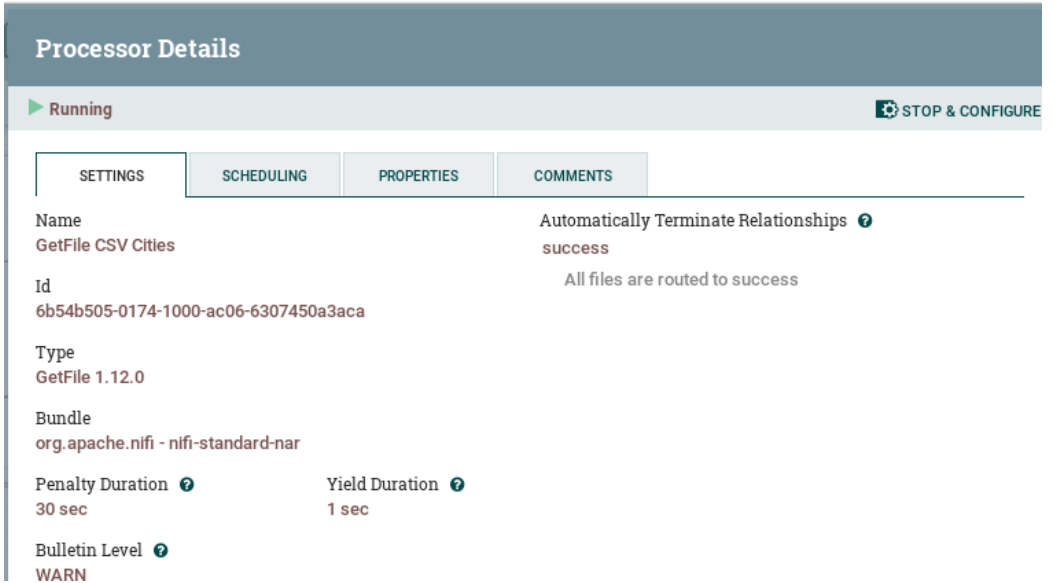


Processor - Configuration

Ambiente Livre
Big Data & Data Science

Configuração do Processador

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



Processor Groups

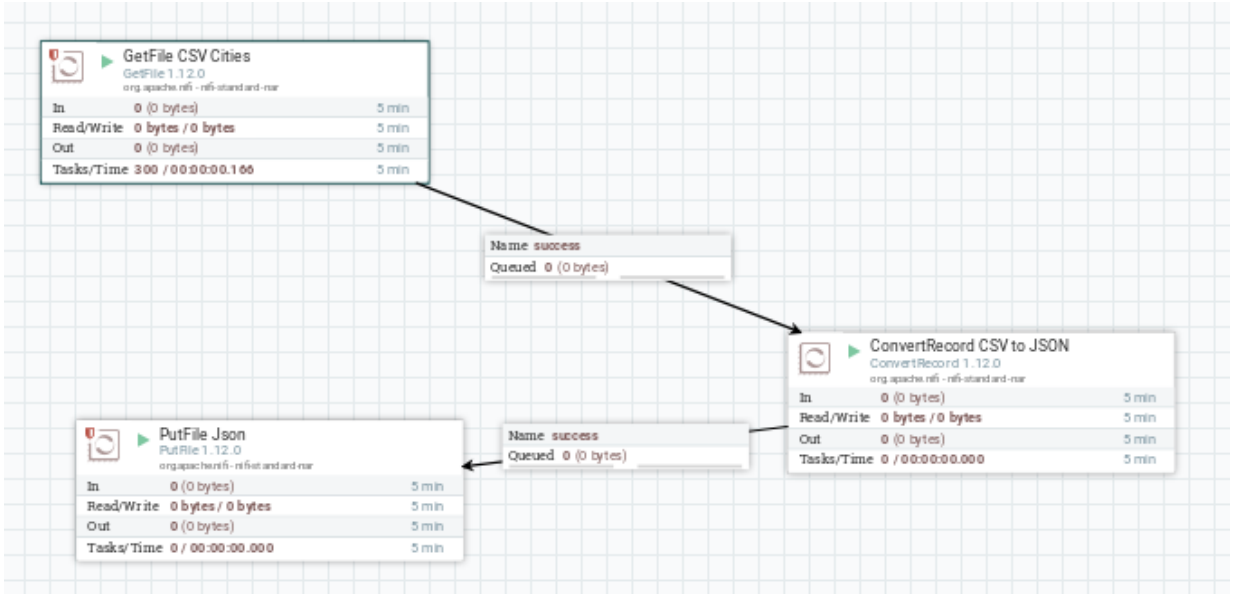
Grupos de Processador.

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

Connections

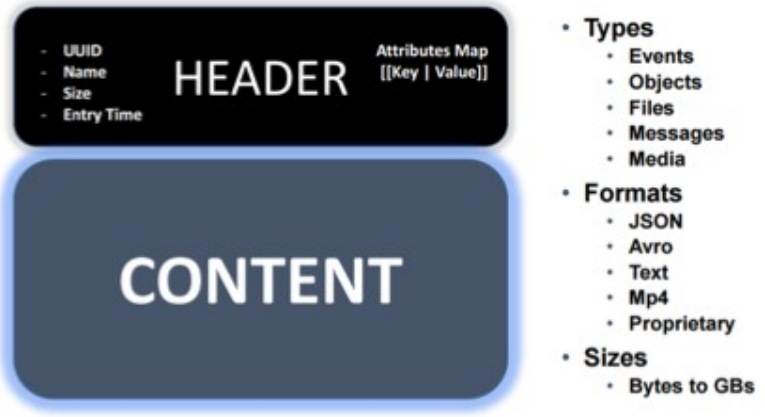
Conexões

- 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.

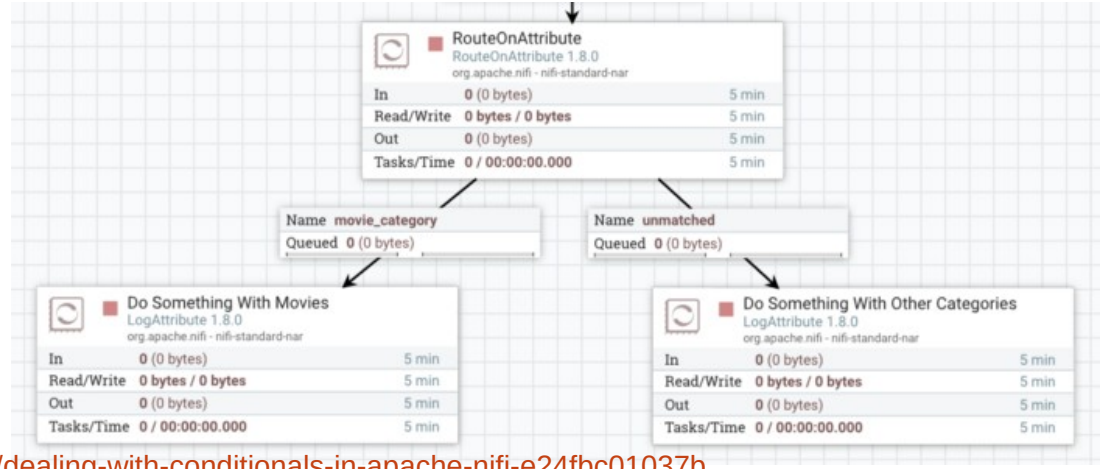


Condicional Processor

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

Ambiente Livre
Big Data & Data Science

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

If these characters are present in attribute names they need to be quoted

$ | { } ( ) [ ] , : ; / * '
(space) \t \r \n

Ex. ${'a:attribute name'}
${"a:attribute name"}

Type Conversion

Coerces from one format to another

toString()	${literal(2):toString() :equals('2')}
toNumber()	${literal('2'):toNumber() :equals(2) }
toDecimal()	${filesize:toDecimal()}

Mathematical

plus()	${fileSize:plus(10)}
minus()	${fileSize:minus(10)}
multiply()	${fileSize:multiply(10)}
divide()	${fileSize:divide(10)}
mod()	${fileSize:mod(10)}

Logic Operators

isNull()	${filename:isNull()}
notNull()	${filename:notNull()}
isEmpty()	${literal('') :isEmpty()}
equals(string)	${filename :equals('value')}
equalsIgnoreCase (string)	${filename :equalsIgnoreCase('v')}
gt(number)	${fileSize:gt(64)}
ge(number)	${fileSize:ge(64)}
lt(number)	${fileSize:lt(64)}
le(number)	${fileSize:le(64)}
and(bool)	${fileSize:gt(1) :and(${fileSize:lt(4)})}
or(bool)	${fileSize:lt(1) :or(${fileSize:gt(4)})}
not()	${filename :endsWith('sv'):not()}
ifElse ('true val', 'falseval')	${filename :endsWith('csv') :ifElse('is csv', 'is not csv')}

Text Search

filename:equals('fizz buzz bazz.txt')

startsWith (string)	${filename :startsWith('fizz')}
endsWith (string)	${filename :endsWith('txt')}
Contains (string)	${filename :contains('buzz')}
in(string, string...)	${literal('NO') :in('NO','NOT')}
indexOf(string)	${filename :indexOf('buzz')} == 5
lastIndexOf (string)	${filename :lastIndexOf('z')} == 13
find(regex)	${filename:find('.*zz')}
matches(regex)	${filename :matches('fizz.*txt')}
jsonPath(path)	${theJson :jsonPath('$.attribute')}

Utilities

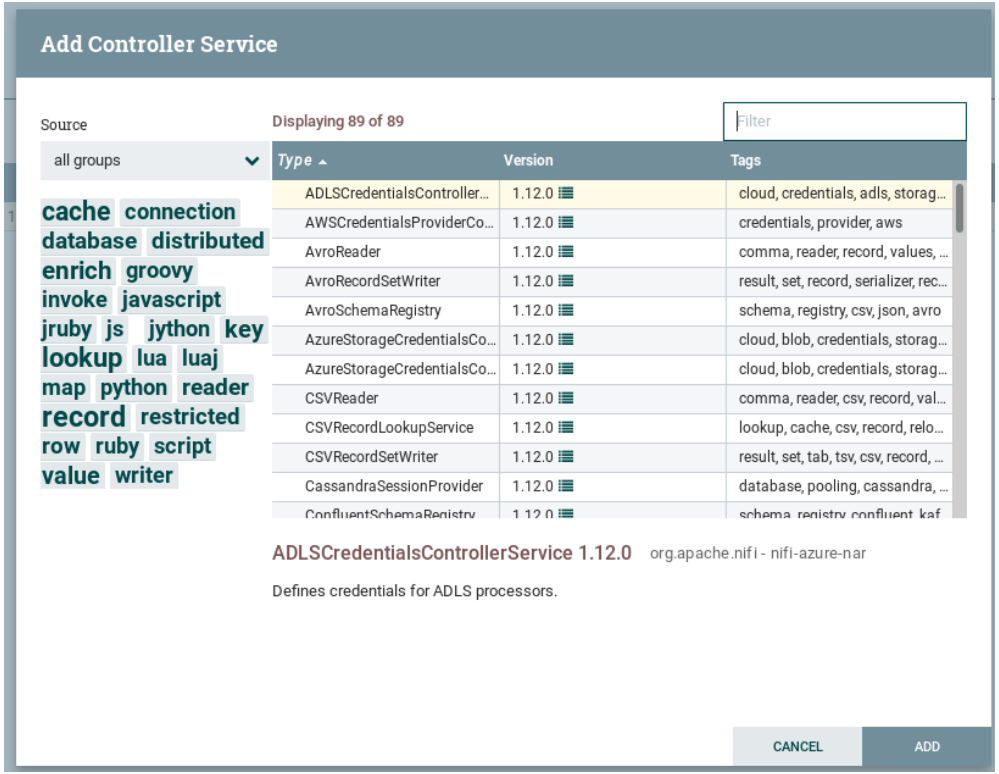
These subjectless functions provide useful utilities.

ip()	local ip
hostname(bool)	${hostname(true)} *fully qualified hostname*

Controller Services

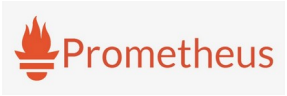
Services

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

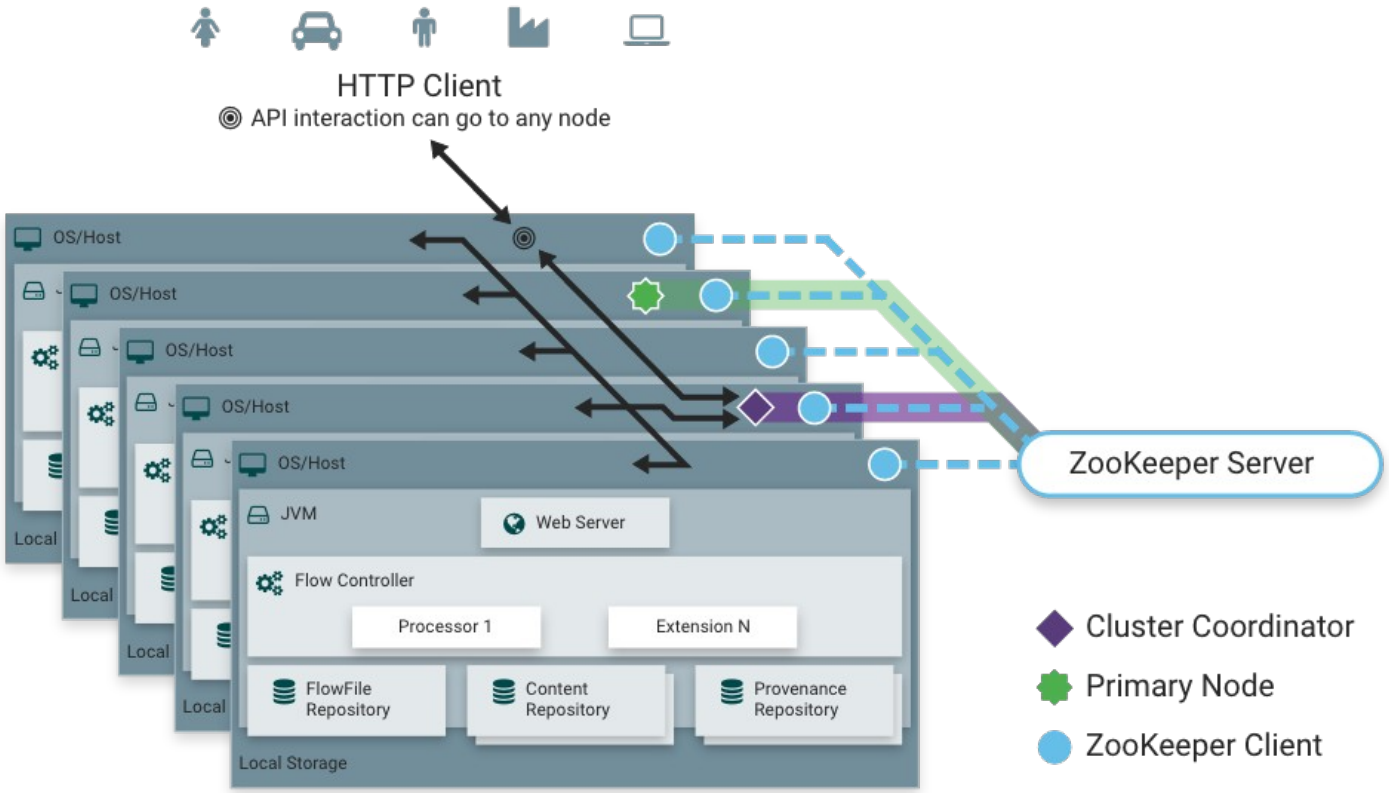


Apache NiFi - Controller Services Types

Ambiente Livre
Big Data & Data Science

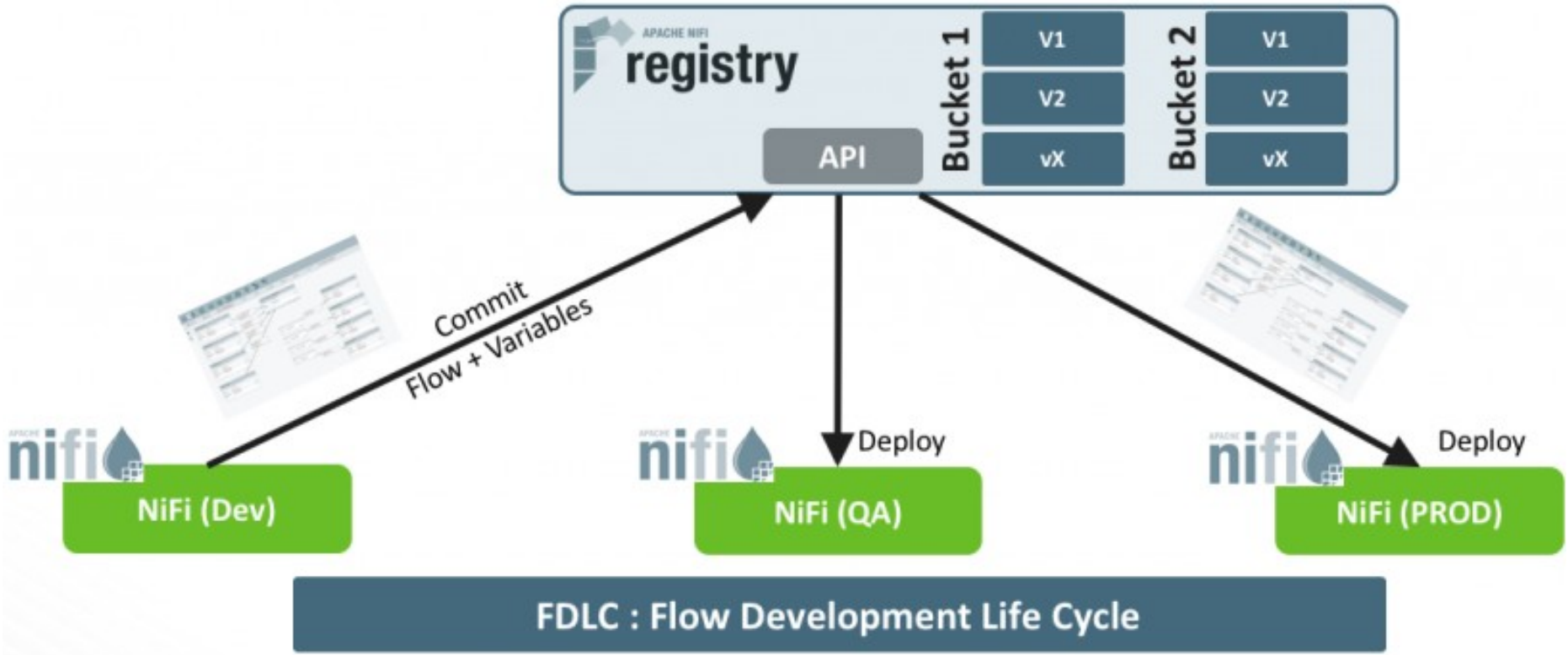


Nifi Cluster



Nifi Registry e Flow Life Cyclo

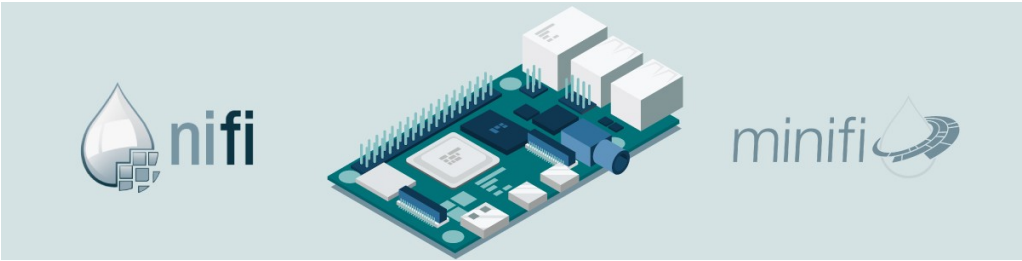
Ambiente Livre
Big Data & Data Science

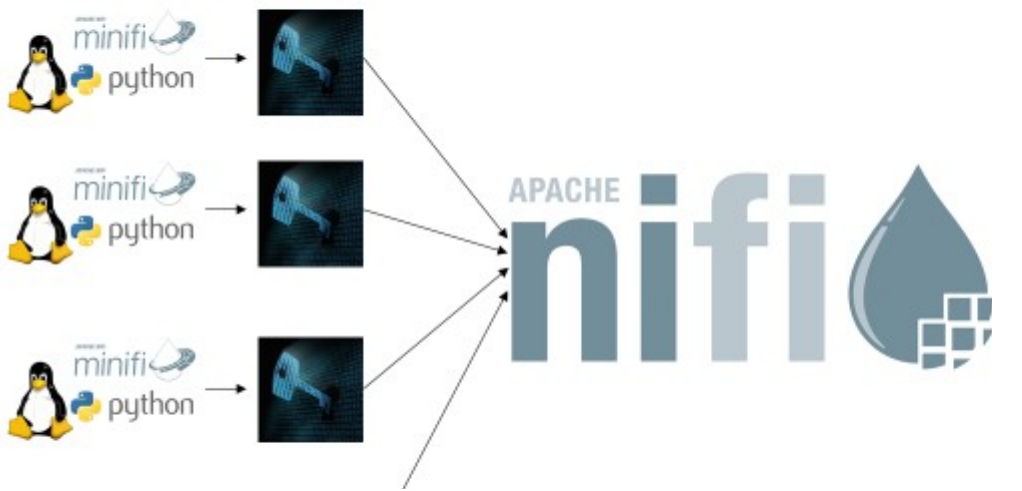


Minifi

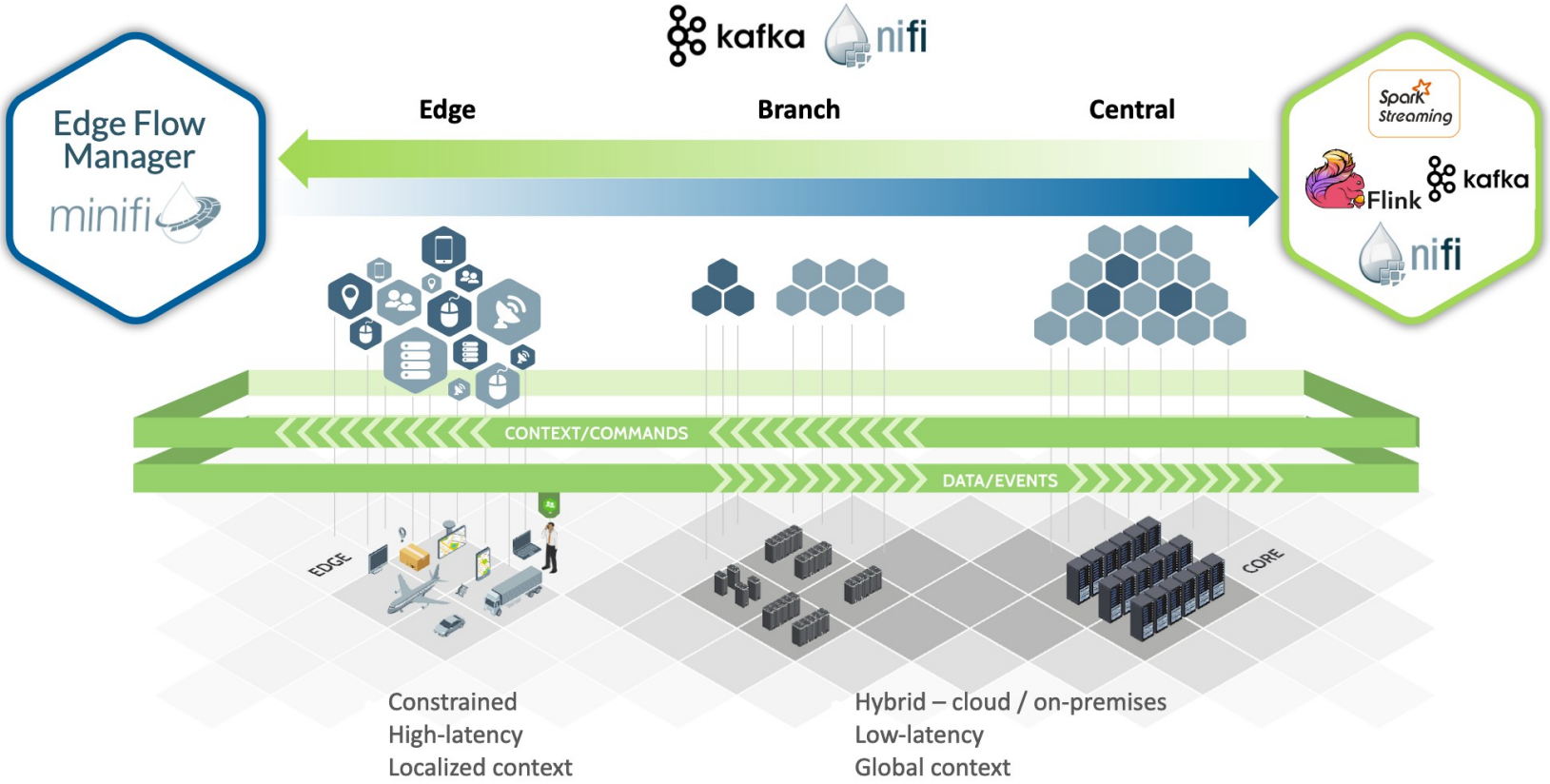
Conceito

- 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

Ambiente Livre
Big Data & Data Science

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