LATINOWARE 2021

Ambiente Livre

Big Data & Data Science

Apache Hop - A última geração em orquestração de dados

Marcio Junior Vieira
CEO & Data Scientist, Ambiente Livre

Marcio Junior Vieira

Ambiente Livre
Big Data & Data Science

Mini-cv

- 20 anos de experiência em TI, vivência em desenvolvimento, análise 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 BI da Universidade Positivo.
- Professor do MBA Artificial Intelligence e Machine Learning da FIAP.
- Pesquisador do Laboratório de tecnologias de 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, etc.
- Especialista em implantação e customização de Big Data com Hadoop, Spark, Pentaho, Cassandra e MongoDB.
- Contribuidor de projetos internacionais, tais como Apache Hop, 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)

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

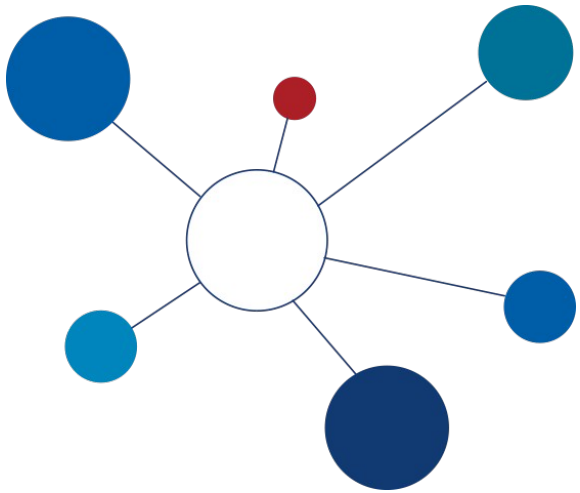
Painéis de Indicadores
Cubos de Análise
Relatórios Gerênciais
Tomada de Decisão

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

Consultoria | Treinamento | Projetos

pentaho · Alfresco · hadoop · mongoDB · Cassandra · suiteCRM · LPI · Spark · Hop · Apache Software Foundation · LIFERAY Enterprise. Open Source. For Life. · LimeSurvey · PostgreSQL the world's most advanced open source database · Apache Flink · camunda · Flowable making process personal · moodle · MariaDB · Drupal

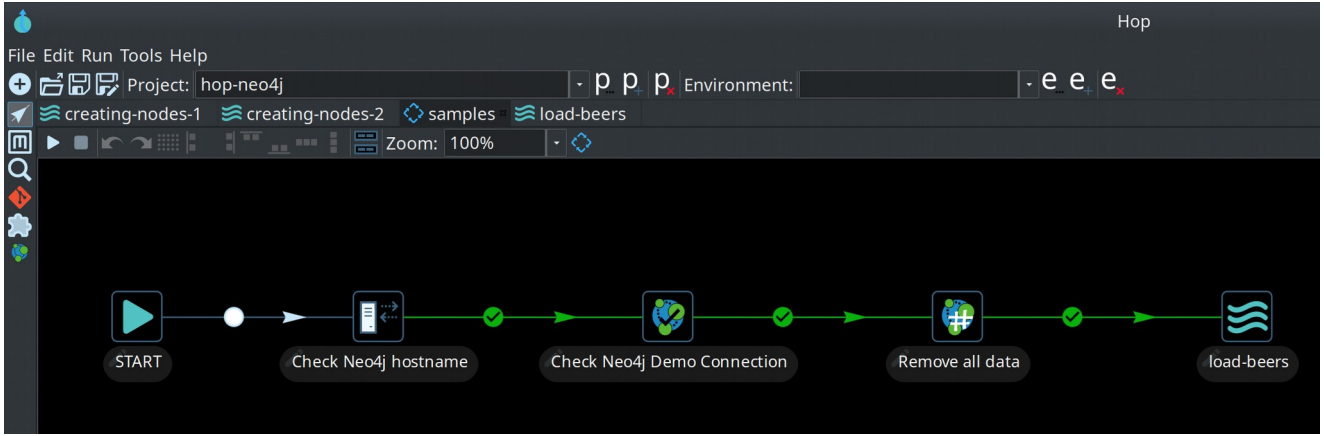
Agenda



① O que é o Apache Hop.
② Conceitos.
③ Downloads e Instalação.
④ Ferramentas.
⑤ Interface Gráfica do Apacha Hop.
⑥ Pipelines.
⑦ Workflow.

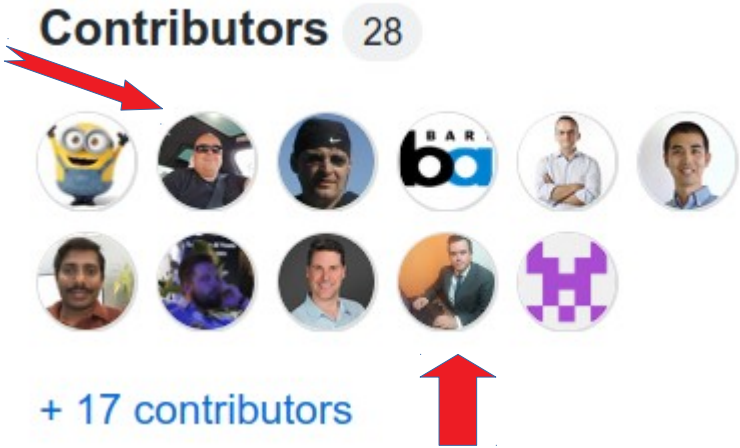
Apache Hop

- 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.



Background

- Iniciativa de uma Comunidade (junto com Matt Casters)
- Fork do Kettle/Pentaho/PDI 8.2 + WebSpoon + Patches + plugins… (Somados são mais de 20 anos de desenvolvimento de software)
- Interface gráfica renovada.
- Back-end de metadados novo.
- Maior simplicidade.
- Muitos códigos refatorados.
- Licença Apache v2.0



Apache Hop

Ambiente Livre
Big Data & Data Science

Apache Incubation

- Versão 1.0 lancada após 2 anos de refatoração e implementações.
- O projeto está encubado na Fundação Apache.
- Fontes em https://github.com/apache/incubator-hop
- Website: https://hop.apache.org

Fundação Apache

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”

Quem Patrocina a Fundação?

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

Apache Hop - Funcionalidades

Ambiente Livre
Big Data & Data Science

Funcionalidades Tradicionais

- Usadas em projetos de data warehouse e data lakes.
- Usado para projetos de engenharia de dados.

Funcionalidades Adicionais

- Migração de dados entre aplicações/banco de dados
- Exportar dados de banco de dados para arquivos texto
- Carregar massivamente dados em banco de dados
- Data Cleansing – disciplina de qualidade/limpeza de dados de data warehouse
- Integração de aplicações.
- Gerenciamento de Filesystem (File management)

Apache Hop – Arquitetura e Funcionalidade

Ambiente Livre
Big Data & Data Science

Funcionalidades

- Geração de Metadata: Sem Codificação (Low-code).
- Modular, plugável ou embutido.
- Rápido Start!
- Apache Beam com suporte a Apache Spark, Flink e GCP Dataflow.
- Pronto para uso com ferramentas simples.
- Testes Integrados.
- VFS File Systems (Local, AWS S3,
  Azure Blob, DropBox,Google Cloud Storag
  Google Driver, HDFS, FTP, WebDav, RAM.

Apache Hop – Conceitos

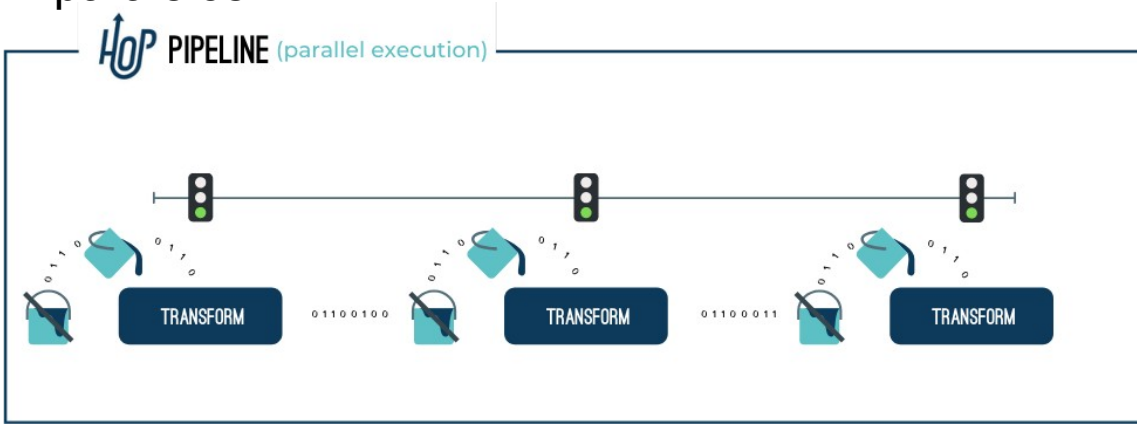
Ambiente Livre
Big Data & Data Science

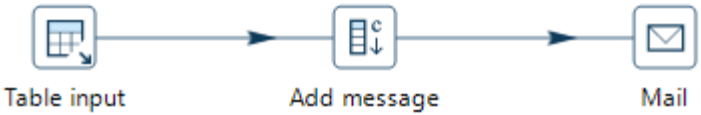
Principios

- Fácil.
- Rápido.
- Transparente.
- Inovador.
- Implementa Melhores Praticas

Pipelenes

- Coleção de transformações de dados connectadas.
- Execução de todos pipelines em paralelos
- Desine gráfico.

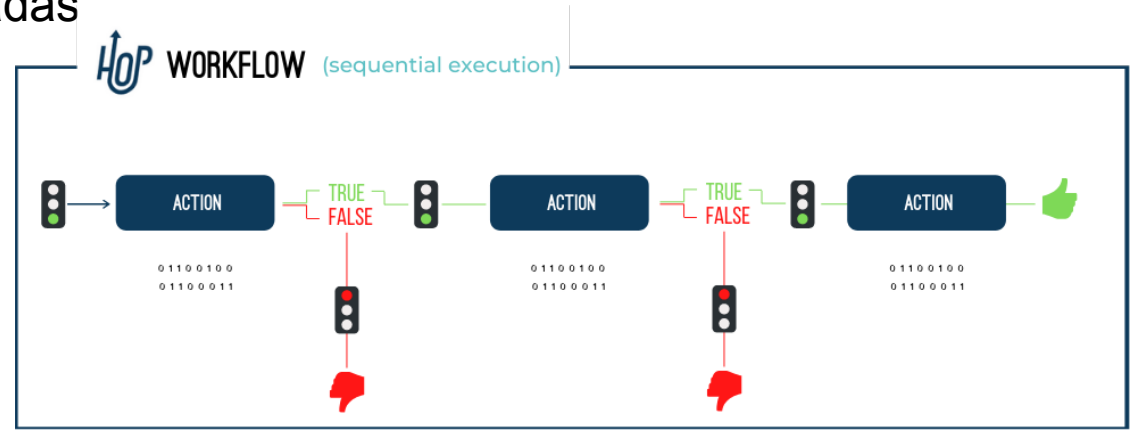


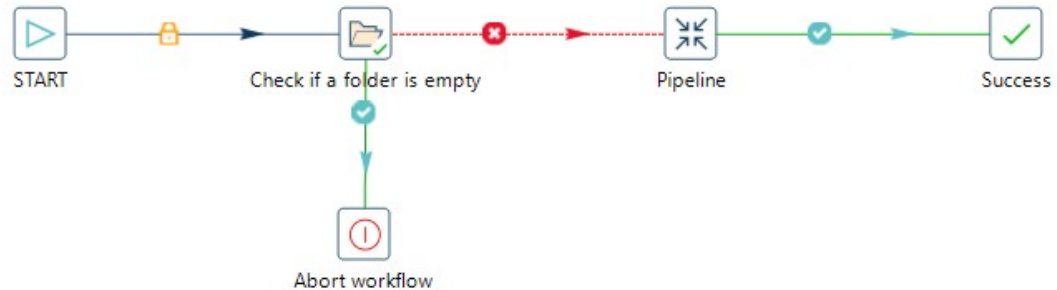


Apache Hop – Workflow

Pipelenes

- Coleção de Ações connectadas
- Executam em sequencial.
- Desine gráfico.





Apache Hop – Projetos e Ambiente

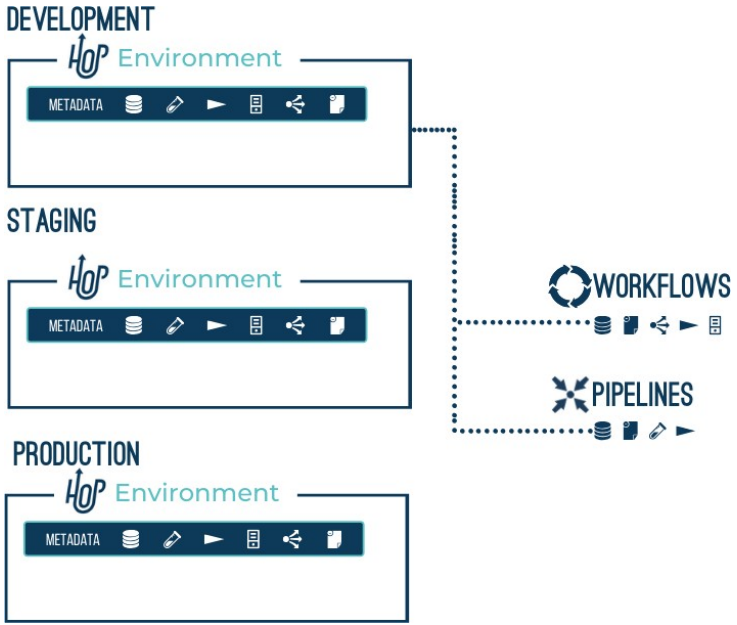
Ambiente Livre
Big Data & Data Science

Projetos

- Agrupamento de pipelines e workflow e configurações.

Ambiente

- Instâncias de projetos
  (DEV, PROD, HOMOLOG, etc)



Apache Hop – Ferramentas

Ambiente Livre
Big Data & Data Science

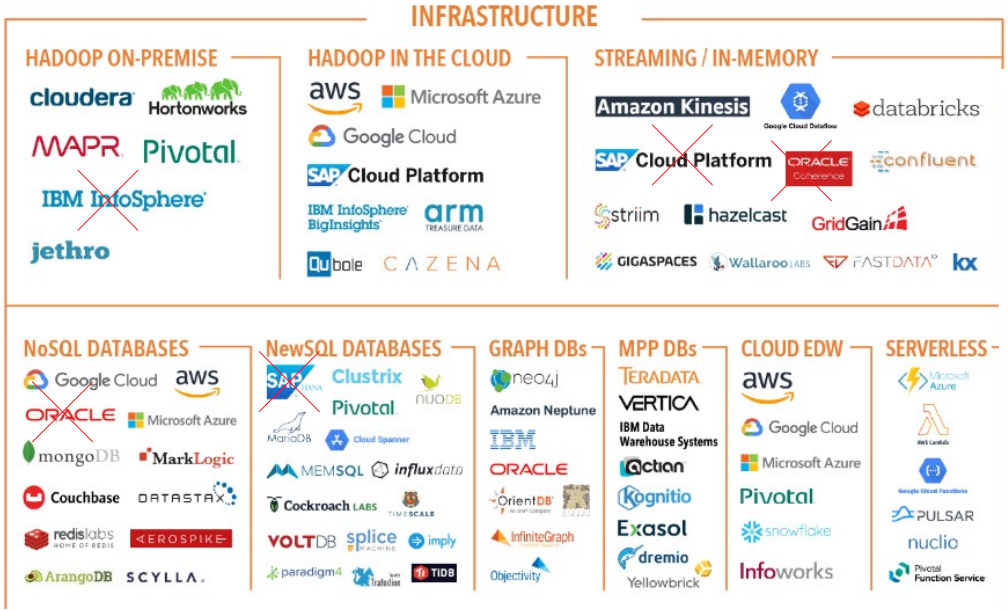
Tools

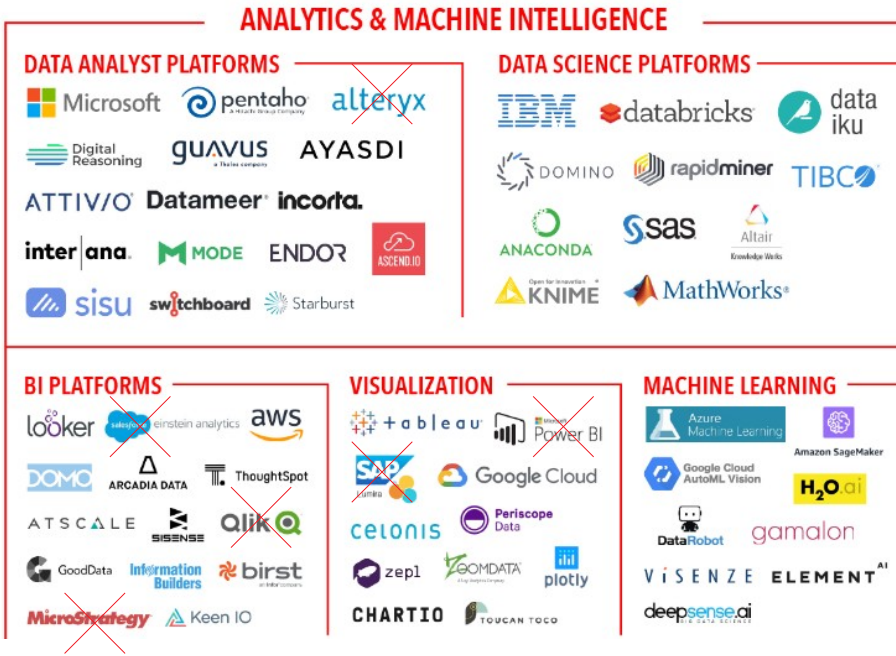
- **Hop GUI**: Desenvolvimento e Design de workflow e pipelines.
- **Hop Conf**: Gerenciamento de variaveis e aspectos configuraveis.
- **Hop Encrypt**: Controle criptografados de dados de dados sensíveis (senhas, etc)
- **Hop Run**: executa pipelines e workflow por linha de comando e agendamento (cron, apache Airflow, schedules windows, etc).
- **Hop Search**: Busca em metadados do projeto.
- **Hop Server:** Interface web para gerenciamento de pipelines e workflows.
- **Hop Translator**: interface para não-tecnicos internacionalizar a ferramenta.

Tecnologias Open Source / Free Software (FLOSS)

Fonte: Big_Data_Landscape
http://mattturck.com

Tecnologias Baseadas em FLOSS





Nem todas “marcas” são baseadas em Open Source, mas a maioria...

Fonte: Big_Data_Landscape
http://mattturck.com

Apache Hop - Instalação

Ambiente Livre
Big Data & Data Science

Pré-requisitos

- Java Runtime 8 +

Instalação

- Download
- unzip

Referências

Ambiente Livre
Big Data & Data Science

Web

- https://hop.apache.org/

Agradecimentos

Ambiente Livre
Big Data & Data Science

- Agradeço a Latinoware pelo convite!
- Obrigado aos palestrantes e congressistas!

Obrigado

Ambiente Livre
Big Data & Data Science

Marcio Junior Vieira
marcio@ambientelivre.com.br
@marviojvieira @ambientelivre
@ambientelivreopensoftware
https://www.linkedin.com/in/mvieira1/

Slide da Palestra será publicada em:
Linkedin….: https://www.linkedin.com/in/mvieira1/