Apache Hop

Fernando Anselmo
http://fernandoanselmo.orgfree.com/wordpress/

Versão 1.0 em 5 de junho de 2025

Resumo

Apache Hop (*Hop Orchestration Platform*) é uma plataforma moderna de código aberto para engenharia e orquestração de dados, projetada para tornar os processos de integração de dados mais acessíveis, flexíveis e reutilizáveis. Desenvolvido com foco em produtividade e usabilidade, o **Apache Hop** permite criar pipelines e *workflows* de dados de forma visual e intuitiva, sem depender necessariamente de programação.

1 Introdução

Ao contrário de outras ferramentas de ETL (*Extract, Transform, Load*), o **Apache Hop** adota uma arquitetura orientada a metadados, isso significa que toda a lógica de transformação e orquestração dos dados é descrita por meio de definições reutilizáveis e portáveis. Isso permite que você defina como deseja que os dados sejam processados, enquanto a plataforma cuida do trabalho pesado da execução.

Figura 1: *Logo do Apache Hop*

Uma das grandes vantagens do **Apache Hop** é sua capacidade de projetar *pipelines* uma única vez e executá-los em diferentes ambientes — locais, em nuvem, ou em frameworks como **Apache Spark**, **Apache Flink** ou **Google Dataflow** - por meio das chamadas configurações de tempo de execução, um tipo especial de metadado que abstrai o ambiente de execução.

Além disso, o **Apache Hop** centraliza os processos em uma plataforma única e gerenciável, oferece recursos avançados de controle de qualidade, persistência e rastreabilidade dos dados, contribuí para a confiabilidade e a governança da informação.

A plataforma é desenvolvida por uma comunidade aberta, colaborativa e acolhedora, sob a governança da **Apache Software Foundation**. Todos são convidados a participar: seja tirar dúvidas, relatar

problemas, propor novos recursos, contribuir com código ou documentação, auxiliar nos testes de versões ou melhorar o site oficial.

Apache Hop é uma solução robusta, extensível e preparada para os desafios modernos da engenharia de dados, ideal para organizações que buscam eficiência, automação e escalabilidade em seus fluxos de dados.

2 Apache Hop e Pentaho

Apache Hop é uma poderosa ferramenta de integração e orquestração de dados de código aberto, criada como um fork evoluído do **Pentaho Data Integration (PDI)**, também conhecido como Kettle. Embora compartilhe raízes com o PDI, **Apache Hop** foi totalmente reestruturado e modernizado para atender às necessidades atuais de engenharia de dados com mais desempenho, modularidade e escalabilidade.

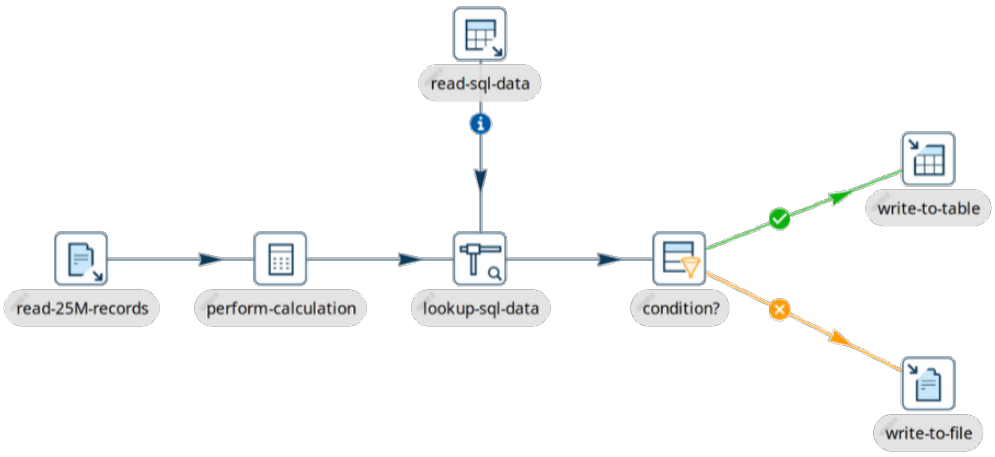

Figura 2: *Exemplo de uma Pipeline no Apache Hop*

Com uma interface de desenvolvimento visual, **Apache Hop** permite que Engenheiros e Arquitetos de Dados construam *pipelines* e *workflows* complexos de forma intuitiva, sem a necessidade de escrever código, embora isso continue sendo uma opção para os usuários mais avançados. Essa abordagem visual acelera o desenvolvimento, reduz erros e facilita a colaboração entre equipes técnicas e de negócio.

Além disso, **Apache Hop** introduz conceitos modernos, como metadados reutilizáveis, configurações de tempo de execução e suporte a múltiplos motores, tornando-se uma solução versátil para projetos locais, em nuvem ou em ambientes distribuídos.

Comparativo: Apache Hop vs Pentaho Data Integration (PDI)

Característica	**Apache Hop**	**Pentaho Data Integration (PDI)**
Origem	Fork moderno e reescrito do PDI/Kettle	Projeto original, mantido pela Hitachi Vantara
Licença	Apache License 2.0 (open source completo)	Community Edition: LGPL Enterprise Edition: Proprietária

Governo do Projeto	Apache Software Foundation (comunidade aberta)	Hitachi Vantara (foco comercial)
Interface de Desenvolvimento	Visual (Hop GUI)	Visual (Spoon)
Modularidade	Altamente modular, orientado a plugins	Arquitetura mais monolítica
Abordagem Baseada em Metadados	Sim – pipelines, workflows, variáveis, ambientes	Parcialmente, com menos flexibilidade
Configurações de Execução	Suporte a múltiplos ambientes com "run configurations"	Limitado, dependente de configurações locais
Execução Distribuída	Suporte nativo a Spark, Flink, Beam via plugins	Requer customizações adicionais
Linha de Comando	Ferramentas modernas: `hop-run`, `hop-gui`, `hop-server`	Ferramentas legadas: `pan`, `kitchen`
Comunidade	Ativa, aberta, com crescimento contínuo	Reduzida na versão open source
Atualizações e Roadmap	Frequentes e transparentes	Lentamente atualizada na versão gratuita
Documentação	Completa e mantida pela comunidade	Limitada na versão open source

As principais vantagens do **Apache Hop** são:

Integração nativa com GIT - Não é necessário usar clientes GIT de terceiros para tornar o ambiente DevOps e DataOps mais amigável e produtivo. Existe uma interface visual no **Aoache Hop** que permite ver tudo que foi alterado, inclusive mostrando graficamente o *pipeline* ou *workflow* editado. Com certeza este é um grande avanço se comparado a todos os tipos de repositórios de artefatos (transformations e jobs) do *Pentaho Community Edition*.

Velocidade quanto a atualizações - **Pentaho** e **Apache Hop** atualmente tomam rumos diferentes, possuem objetivos diferentes e portanto têm suas atualizações seguindo por caminhos diferentes. quando há uma necessidade da comunidade sobre a atualização de uma *transform* do **Apache Hop** (equivalente ao *step* do **Pentaho**), isso acontece em uma maior velocidade.

Projeto Top Level da Apache Software Foundation - *Apache Software Foundation* é uma organização sem fins lucrativos criada para suportar os projetos de código aberto. Ser um projeto da Apache requer que o Software preencha uma séries de requisitos, o que dá grande credibilidade e robustez ao projeto.

3 Componentes do Apache Hop

Apache Hop possui três componentes principais, são eles:

Hop GUI é a interface gráfica principal do **Apache Hop**, projetada para facilitar o desenvolvimento de pipelines (antigas transformações no PDI) e *workflows* (antigos jobs). Com uma abordagem visual

e intuitiva. Elimina a necessidade de codificação ao permitir que criemos fluxos complexos de ETL (Extração, Transformação e Carga) por meio de elementos de arrastar e soltar (*drag-and-drop*). Cada *pipeline* representa uma sequência lógica de transformações de dados, enquanto *workflows* permitem orquestrar múltiplas tarefas e *pipelines* em uma ordem específica, com controle de fluxo, paralelismo e dependências. O ambiente também oferece recursos de debug, execução local, parametrização, controle de versão, validação e reutilização de metadados, o que torna o desenvolvimento altamente produtivo e sustentável.

Hop Run é uma ferramenta de linha de comando (CLI) autônoma usada para executar pipelines e *workflows* fora do ambiente gráfico. Ideal para integrações com *scripts*, automações de DevOps, servidores CI/CD ou agendamentos via **cron** (agendamento). Permite a execução headless (sem interface gráfica) com suporte completo a parâmetros, ambientes e variáveis definidos no projeto. Garante que pipelines criados visualmente possam ser facilmente executados em produção, ambientes de teste ou *containers*, promovendo flexibilidade e consistência no ciclo de vida das soluções de dados.

Hop Server é um servidor leve baseado em web, capaz de executar remotamente *pipelines* e *workflows* em ambientes distribuídos. Pode ser implantado em um ou mais nós, permite a execução paralela, balanceamento de carga e alta disponibilidade. Expõe uma API RESTful completa, possibilitando que outros sistemas ou aplicações interajam com os *pipelines* de forma programática — ideal para automações, integrações com plataformas externas e arquiteturas orientadas a eventos. Através dele, é possível orquestrar fluxos de dados complexos a partir de múltiplos servidores, promover escalabilidade horizontal e melhorar o gerenciamento das cargas de trabalho.

4 E isso tudo em contêineres do Docker

Ainda sem texto

5 Conclusão

Ainda sem texto

Sou um entusiasta do mundo **Open Source** e novas tecnologias. Qual a diferença entre Livre e Open Source? Livre significa que esta apostila é gratuita e pode ser compartilhada a vontade. Open Source além de livre todos os arquivos que permitem a geração desta (chamados de arquivos fontes) devem ser disponibilizados para que qualquer pessoa possa modificar ao seu prazer, gerar novas, complementar ou fazer o que quiser. Os fontes da apostila (que foi produzida com o LaTex) está disponibilizado no GitHub [4]. Veja ainda outros artigos que publico sobre tecnologia através do meu Blog Oficial [2].

Referências

[1] Site oficial do Apache Hop
`https://hop.apache.org/`

[2] Fernando Anselmo - Blog Oficial de Tecnologia
`http://www.fernandoanselmo.blogspot.com.br/`

[3] Encontre essa e outras publicações em
https://cetrex.academia.edu/FernandoAnselmo

[4] Repositório para os fontes da apostila
https://github.com/fernandoans/publicacoes