Pentaho / Kettle

Instalação

Com relação à instalação o Pentaho se mostra de fácil execução. Não contém um instalador específico para cada sistema operacional, mas executáveis em geral, e têm como dependência apenas o java sdk 8. O arquivo de download é grande, na versão 9.2.0.0, lançada em 02 de agosto de 2021, contém 1,9 GB compactado. Dentro dele contém os arquivos necessários para execução em ambiente Linux, Windows e Mac.

Para a utilização do programa é necessário apenas descompactar, ter certeza que o java está configurado nas variáveis de ambiente e executar o arquivo relacionado a cada sistema operacional.

Documentação Oficial

Em questão de documentação, o Pentaho mostra ter os principais pontos necessários em geral: instalação, primeiros passos e de consulta. Porém, o que me saltou aos olhos, foi o quanto há apoio por outros grupos. Blogs diretamente apontados no site oficial do serviço, vídeos, fóruns.

Destaque para os vídeos da comunidade hospedados no You Tube. Enquanto on-line encontramos a documentação apenas em Inglês, há uma farta gama de facilitadores que dispõe seus vídeos em português.

Aprendizagem

Nesse quesito ele se mostra bastante prático. Utilizando muito a ideia de arrastar, soltar e ligar. O engenheiro deve conhecer os dados e sua estrutura, mas para a utilização dos componentes não é necessário uma grande bagagem técnica. Em geral SQL pode ser bastante útil para algumas atividades com dados.

Se houver necessidade de utilizar outras linguagens, é possível realizar com instalando um plugin, sendo então disponível para Python, Php entre outras.

A comunidade é o grande diferencial. Contribuindo com inúmeras dicas, tutoriais e artigos que ajudam o usuário iniciante a conseguir montar seus primeiros jobs e cadeias de transformação.

Ação

A integração de dados do Pentaho permite que o usuário realize diversas atividades rotineiras de engenharia de dados. A ferramenta possui conexão com inúmeros bancos de dados. Embora ela se destaque com bancos relacionais, não relacionais também podem ser trabalhados com a ferramenta. Há, inclusive, integração com serviços distribuídos. Hadoop é suportado.

Em caso de falha em algum ponto do pipeline ou que fontes não estejam disponíveis, uma mensagem de e-mail pode ser configurada para gerar um alerta. Assim uma rápida reação ao problema ocorrido pode ser posta em prática. Por outro lado, possibilita também que informes de sucesso sejam enviados a um destinatário.

Outro ponto que vale pontuar, é a possibilidade de agendar uma atividade para ser realizada. Deixando a automação da execução aplicável para quando se desejar. Um arquivo que é disponibilizado mensalmente poderá então ser recuperado sem chamadas desnecessárias.

Levando-se em conta os diferentes tipos de arquivos, criamos três laboratórios que simulam casos de uso reais do Pentaho Data Integration. São eles:

Arquivos com base em documentos: csv, xls, etc;

Bancos de dados: mysql, mariadb, postgres, etc;

Dados de internet via API: json, xml.

Descrição de caso de uso: trabalhando com arquivos documentais

Descrição de caso de uso: trabalhando com banco de dados

Descrição de caso de uso: trabalhando com dados de internet via API

Como a ferramenta trabalha a seleção dos dados?

Exige programação

Drag and drop

Esforço necessário para seleção

Como e qual é o esforço para a realização de transformação em outras etapas do pipeline

Tipos de dados suportados pela ferramenta?

Tempo necessário para implementação de um pipeline dependendo da fonte por ferramenta.

Tipo final de dado.

Implementando as mesmas etapas do pipeline, chega-se ao mesmo resultado final independente da ferramenta?

Que tipo de cenário pode ser utilizado para cada ferramenta?