Curso:
Desenvolvendo Processos de ETL com Pentaho Data Integration
(cód. ETL1000)

Iniciando o treinamento

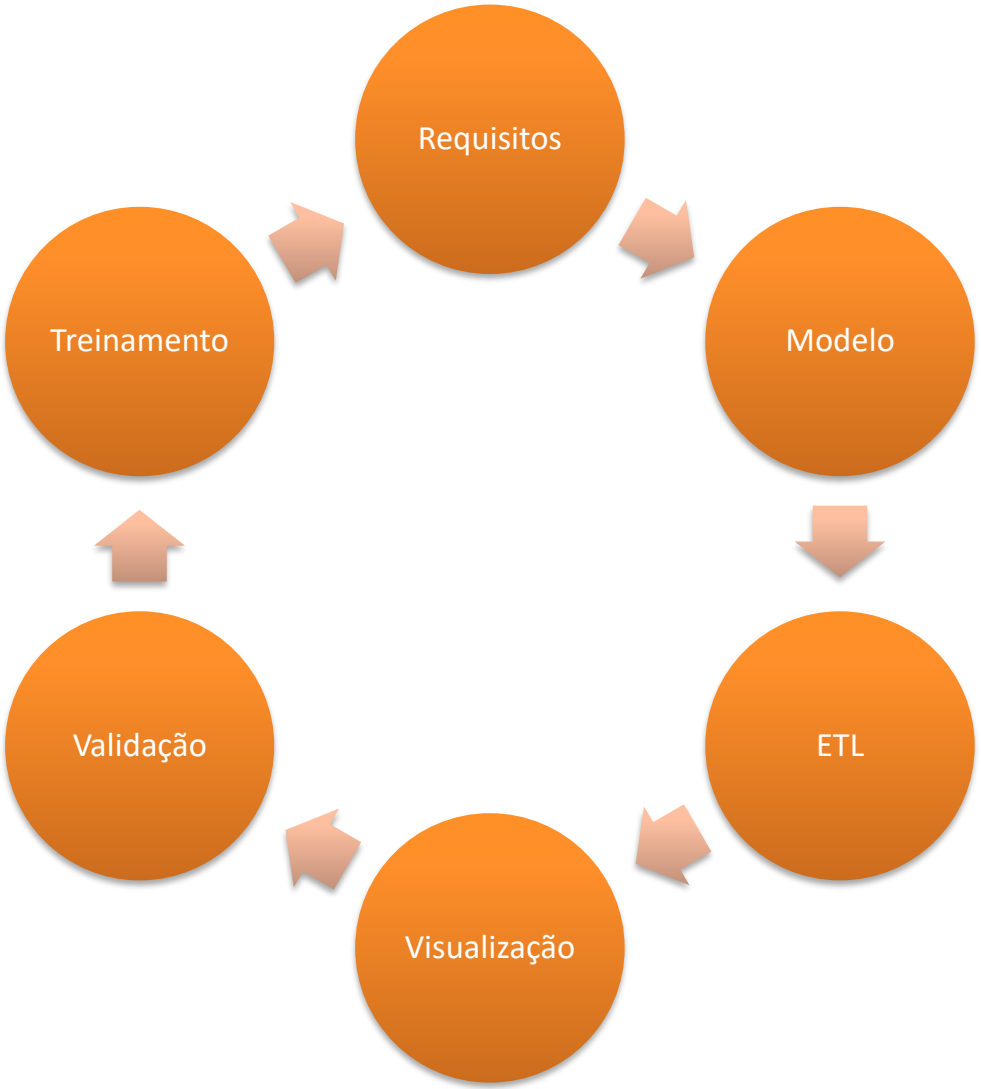
- Quem somos
- Metodologia do treinamento
- Pentaho, a solução líder em Business Analytics
- Demonstração de uso do Spoon
- Exercícios do treinamento

Quem somos

- Empresa nacional com 10 anos de mercado
- Pioneira na América Latina no uso do Pentaho há mais de 10 anos
- Estamos localizados estrategicamente na região de Jundiaí/SP
- Especialista em dados
  - Integração
  - Qualidade
  - Enriquecimento
  - ETL
  - Ingestão
  - Data Prep
  - Big Data
  - Data Science

Quem somos

- Desenvolvimento
  - In-house
  - Fábrica
- Treinamento
  - EaD
  - Online Ao Vivo
  - Presencial
  - Turmas Abertas/In-company
- Suporte Especializado
  - Ambiente Dev/QA/Prod
  - Time de desenvolvimento
- BIaaS (BI Como Serviço)



Alguns de nossos Clientes

OPENIN
YOUR PARTNER IN DATA

OPENiN

YOUR PARTNER IN DATA

Metodologia do treinamento

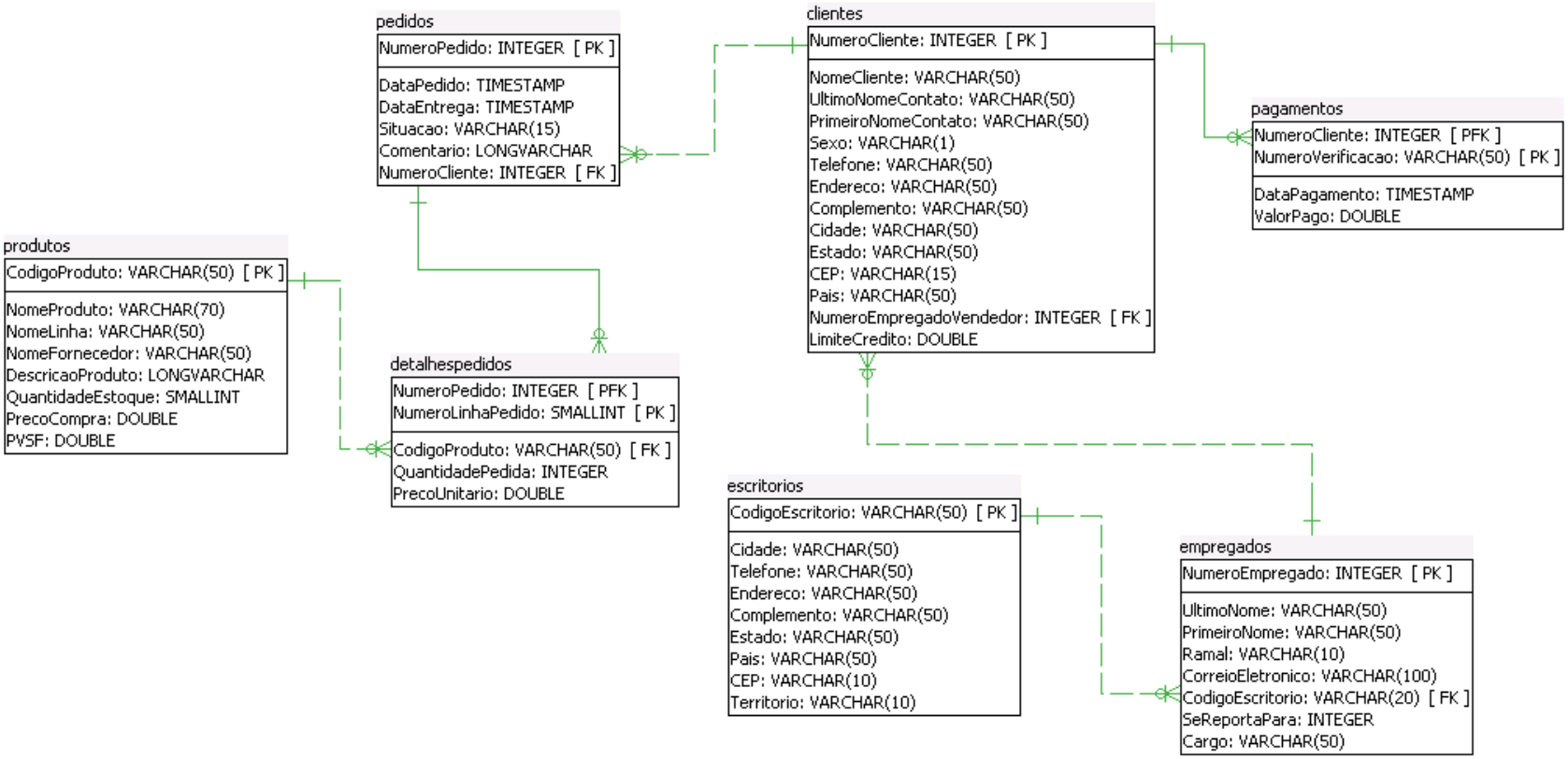
Estudo de caso

Tendo a estratégia como referência, após análise e ponderações entre os membros seniores da nossa equipe, decidimos que o processo de vendas seria o nosso objeto inicial de desenvolvimento e na sequencia viriam outros tão importante quanto

- **Vendas**
- Compras
- Receitas/Despesas
- Folha de pagamento
- ...

motor inc.

Estrutura origem do processo Vendas



OPENiN

YOUR PARTNER IN DATA

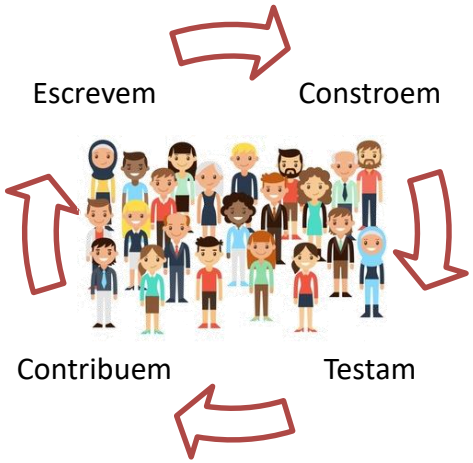
Os entregáveis do projeto

Entregáveis do projeto

- Requisitos negociais (feito)
- Modelo de dados dimensional (feito)
- Especificação do mapa de dados (feito)
- Especificação unitária do ETL (feito)
- Desenvolvimento dos comandos SQL (feito)
- Especificação do mapa de dependência (feito)
- **Desenvolvimento do ETL (a fazer)**
- Desenvolvimento do OLAP (a fazer)
- Desenvolvimento do Report (a fazer)
- Desenvolvimento do Dashboard (a fazer)

Pentaho, a solução líder em Business Analytics

Modelo open source comercial



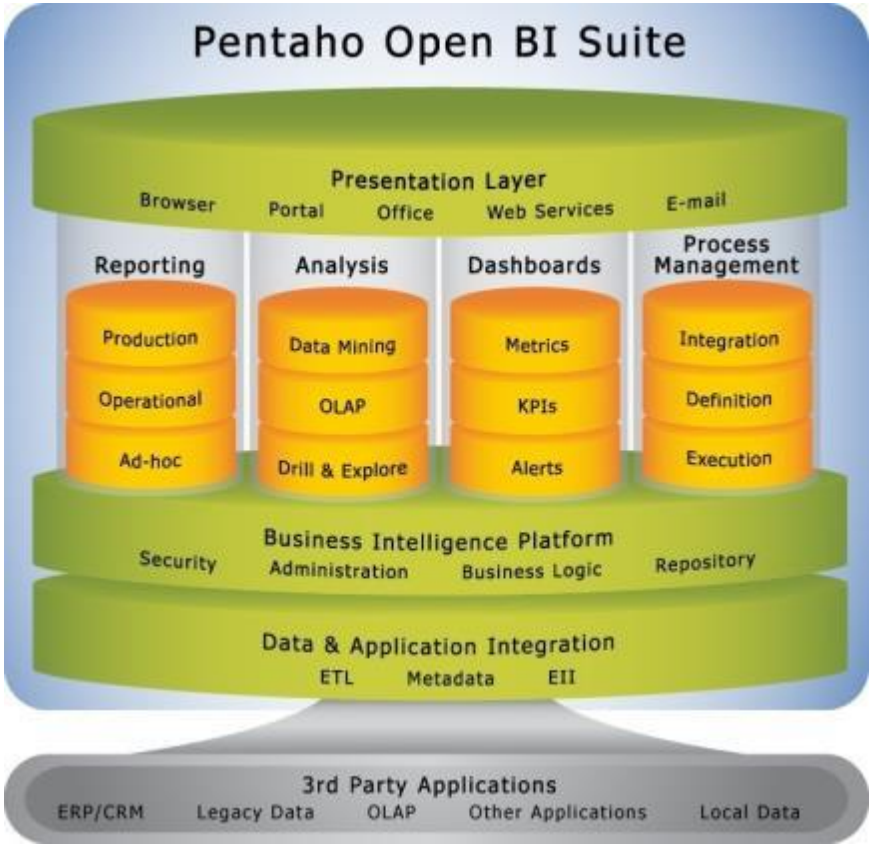
Community Edition

- Hitachi Vantara define um roadmap e comanda o desenvolvimento
- A comunidade contribui com novas funcionalidades
- Exerce um papel fundamental na evolução, melhorando e inovando os produtos

Enterprise Edition

- Produtos certificados - pronto para a produção
- Funcionalidades adicionais e aprimoramentos para facilitar o uso
- Suporte Técnico com níveis de serviços
- Comercialização baseado em assinatura, diminuindo drasticamente o investimento

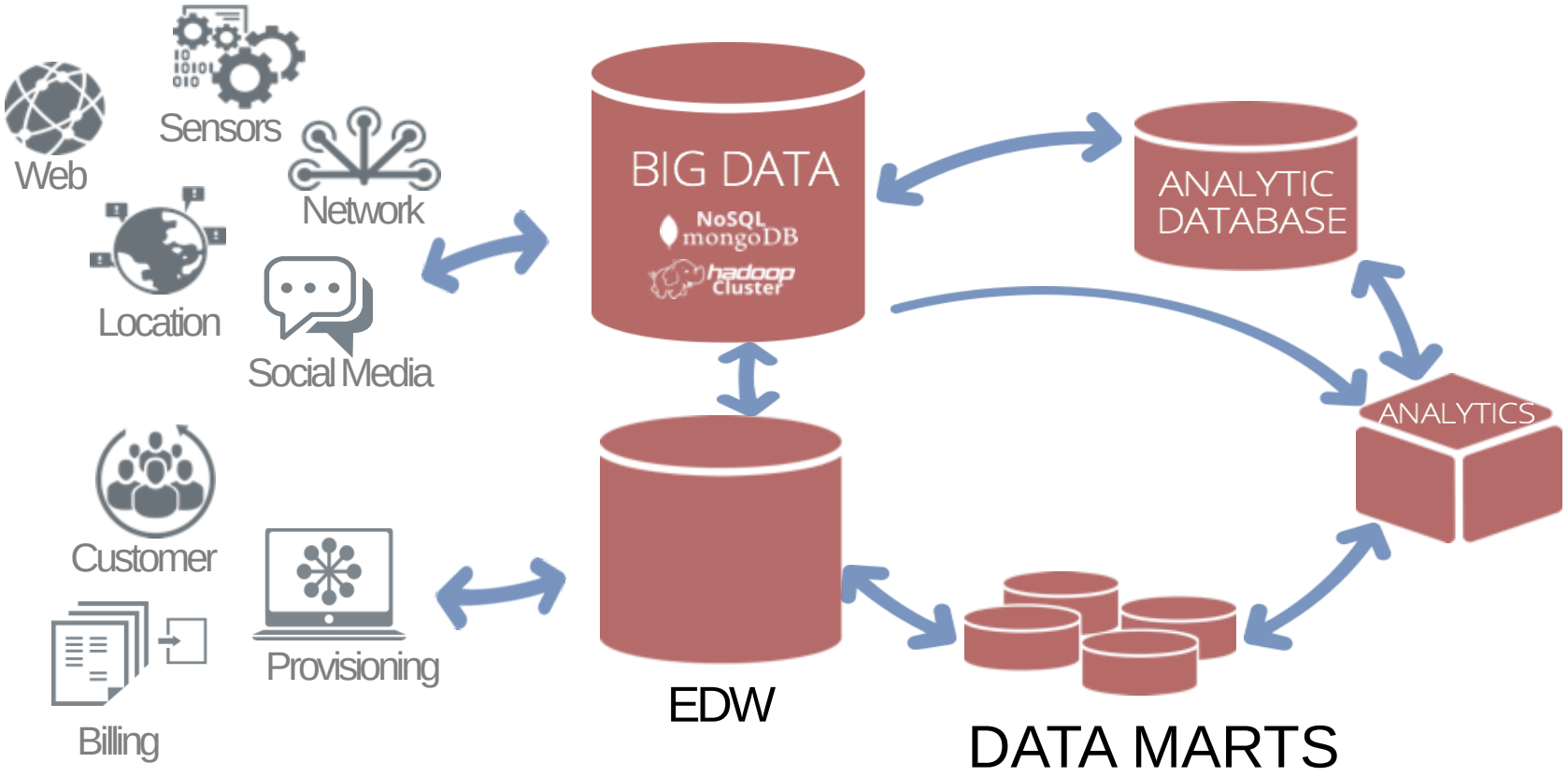
Fundamentos tecnológicos



- **Solução End-to-End**
  - ETL, OLAP, Reporting, Data Minig e Dashboard
- **Ambiente 100% J2EE**
  - Escalável, baseada em padrões
  - Interfaces Web e Thin-client
- **Multiplataforma**
  - Windows, Linux, Unix e Mac

O processo de ETL

Arquitetura de um Business Intelligence



Data Warehouse, **ETL**, Olap, Reporting e Dashboard

ETL

- O que é?
- Para que serve?

ETL, o que é?

"ETL, do inglês Extract Transform Load (Extração Transformação Carga), são ferramentas de software cuja função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e por fim a carga dos dados em um data mart ou um data warehouse. É considerada uma das fases mais críticas do Data Warehouse e/ou Data Mart."

WIKIPEDIA

ETL, para que serve?

"Os projetos de data warehouse consolidam dados de diferentes fontes. A maioria dessas fontes tendem a ser bancos de dados relacionais ou flat files (texto plano), mas podem existir outras fontes. Um sistema ETL tem que ser capaz de se comunicar com as bases de dados e ler diversos formatos de arquivos utilizados por toda a organização. Essa pode ser uma tarefa não trivial, e muitas fontes de dados podem não ser acessadas muito facilmente."

WIKIPEDIA

Características do ETL

Extração, Transformação e Carga

- Ambientes heterogêneos
- Várias fontes de dados
- Um grande tempo é aplicado nesta etapa
- Roda em batch
- Uma das fases mais crítica dentro de um projeto

Ferramentas

- Workflow (gráficas)
- Scheduler (agendador)

Procedimentos

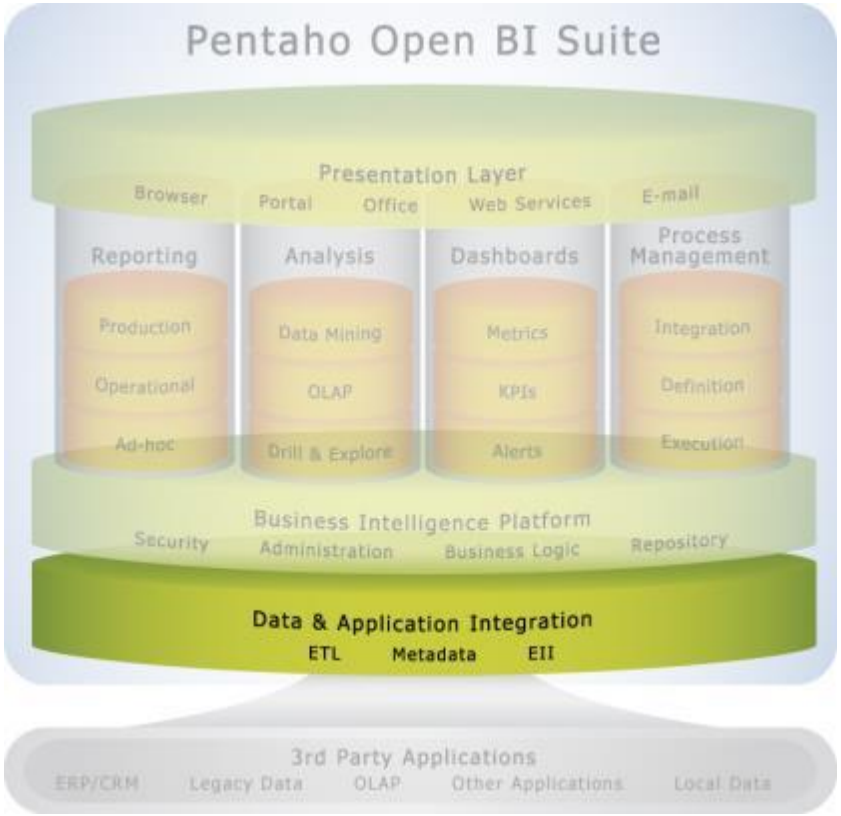
- Limpeza, Integração, Carga, Atualização, Qualidade dos dados, Enriquecimento

Perguntas e Respostas

Apresentando o Pentaho Data Integration

Apresentando o Pentaho Data Integration

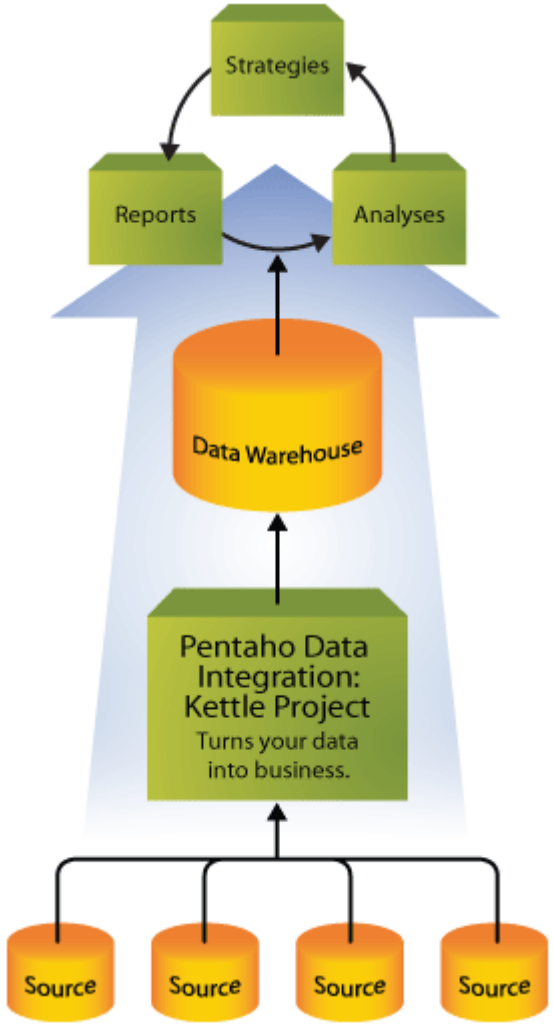
- **O mais popular projeto open source de integração de dados (Kettle)**
  - Conduzido e patrocinado pela Pentaho
  - Possui uma larga contribuição da comunidade
  - Tecnologia em franca evolução



Apresentando o Pentaho Data Integration

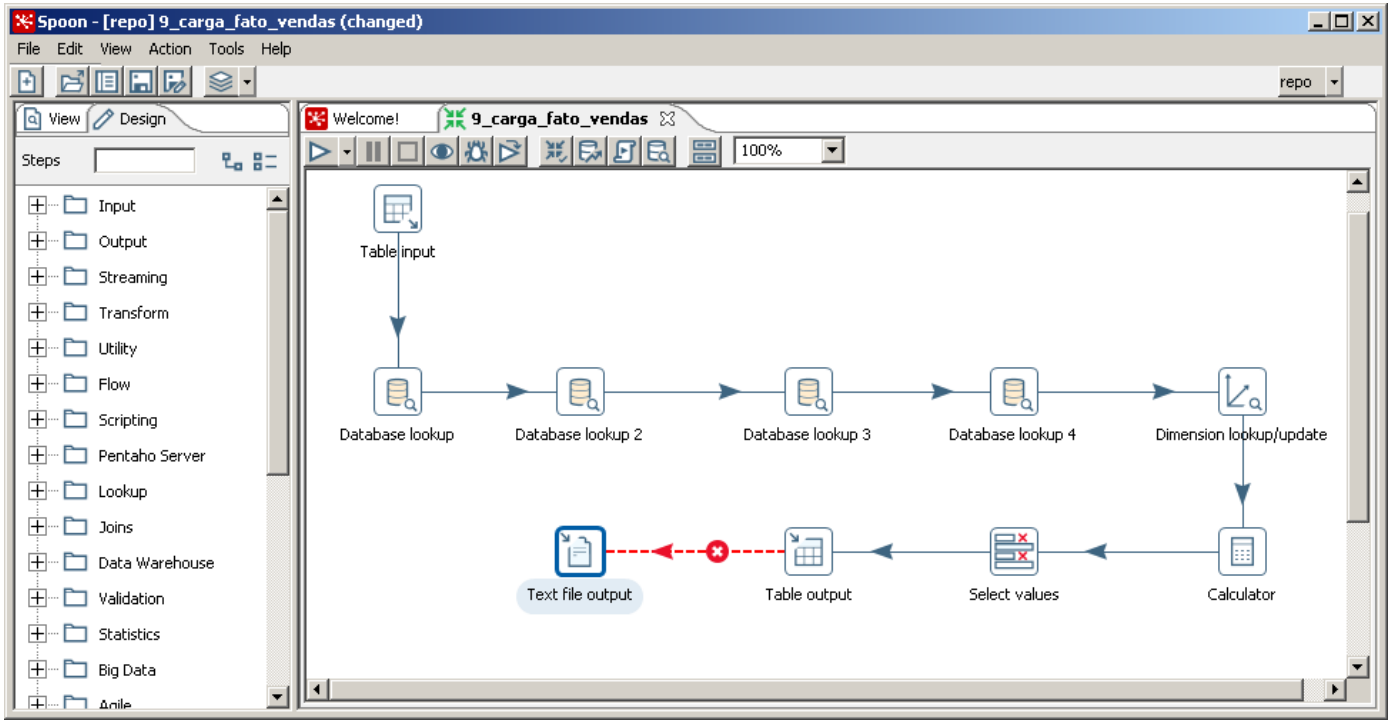
Oportunidades de uso:

- Popular Data warehouse
- Exportar dados para vários formatos
- Importar dados oriundos de diversas fontes
- Migração de dados entre aplicações
- Integração de dados entre aplicações
- Apoio a metodologias ágeis
- Suporte ao Big Data



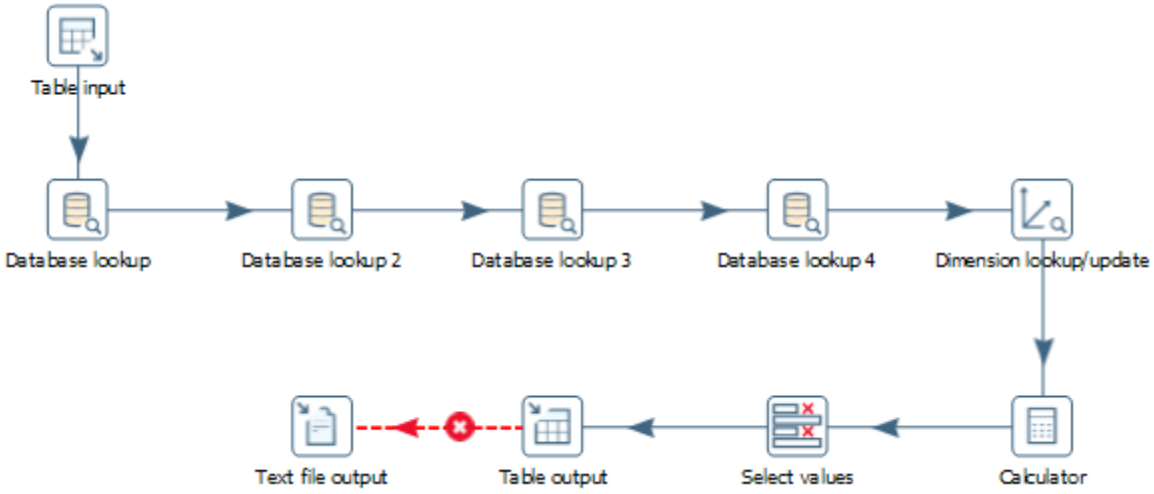
Componentes do PDI - Spoon

- **Spoon**
- Pan
- Kitchen
- Carte



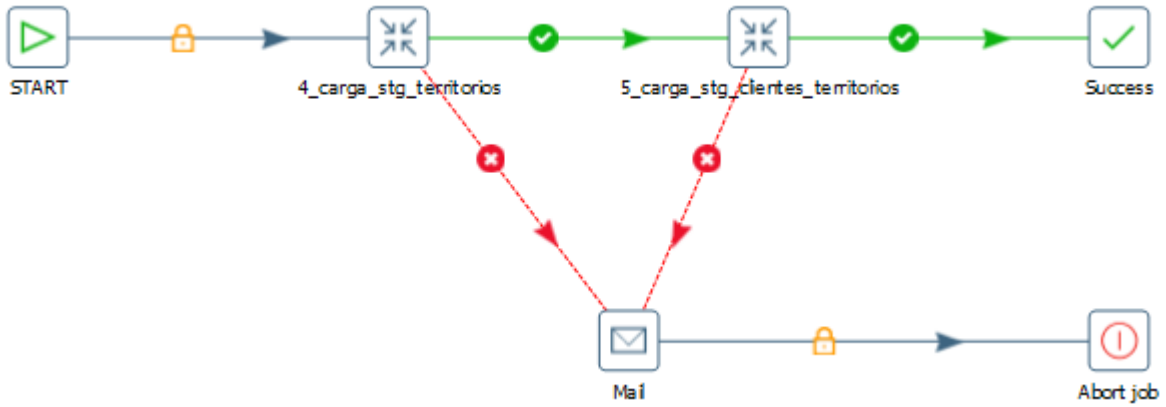
Componentes do PDI – Pan

- Spoon
- **Pan**
- Kitchen
- Carte



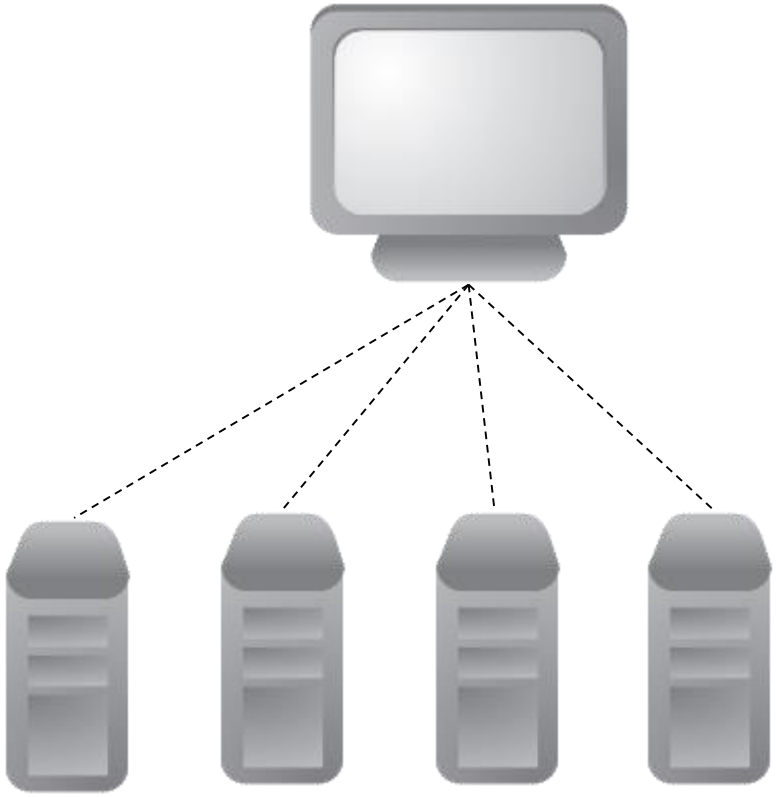
Componentes do PDI – Kitchen

- Spoon
- Pan
- **Kitchen**
- Carte



Componentes do PDI - Carte

- Spoon
- Pan
- Kitchen
- **Carte**



Pré-requisitos

Para instalar o Pentaho Data Integration você deve possuir familiaridade em administração de sistemas e execução de comando via linha de comando.

Software

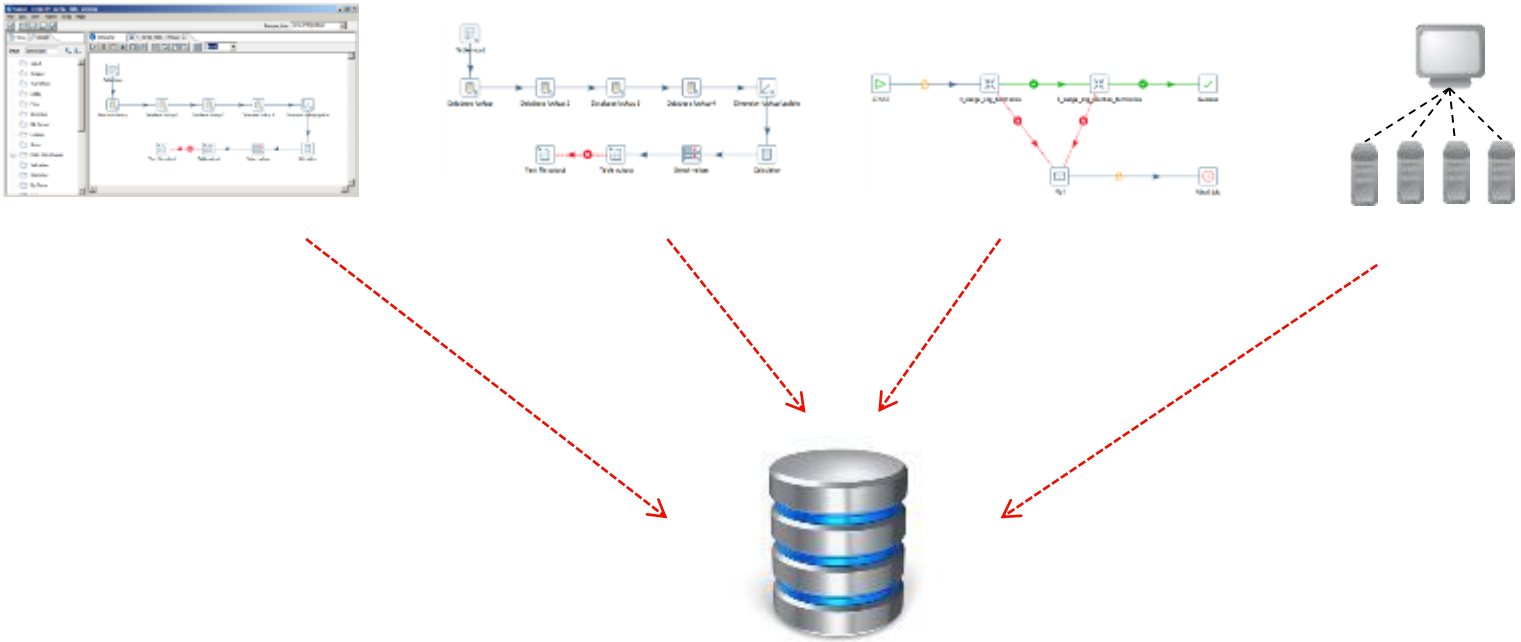
- Windows, Mac, Unix e Linux (32/64 Bits)
- Java Runtime Environment 1.8 (JRK8 - 32/64 Bits)

Hardware

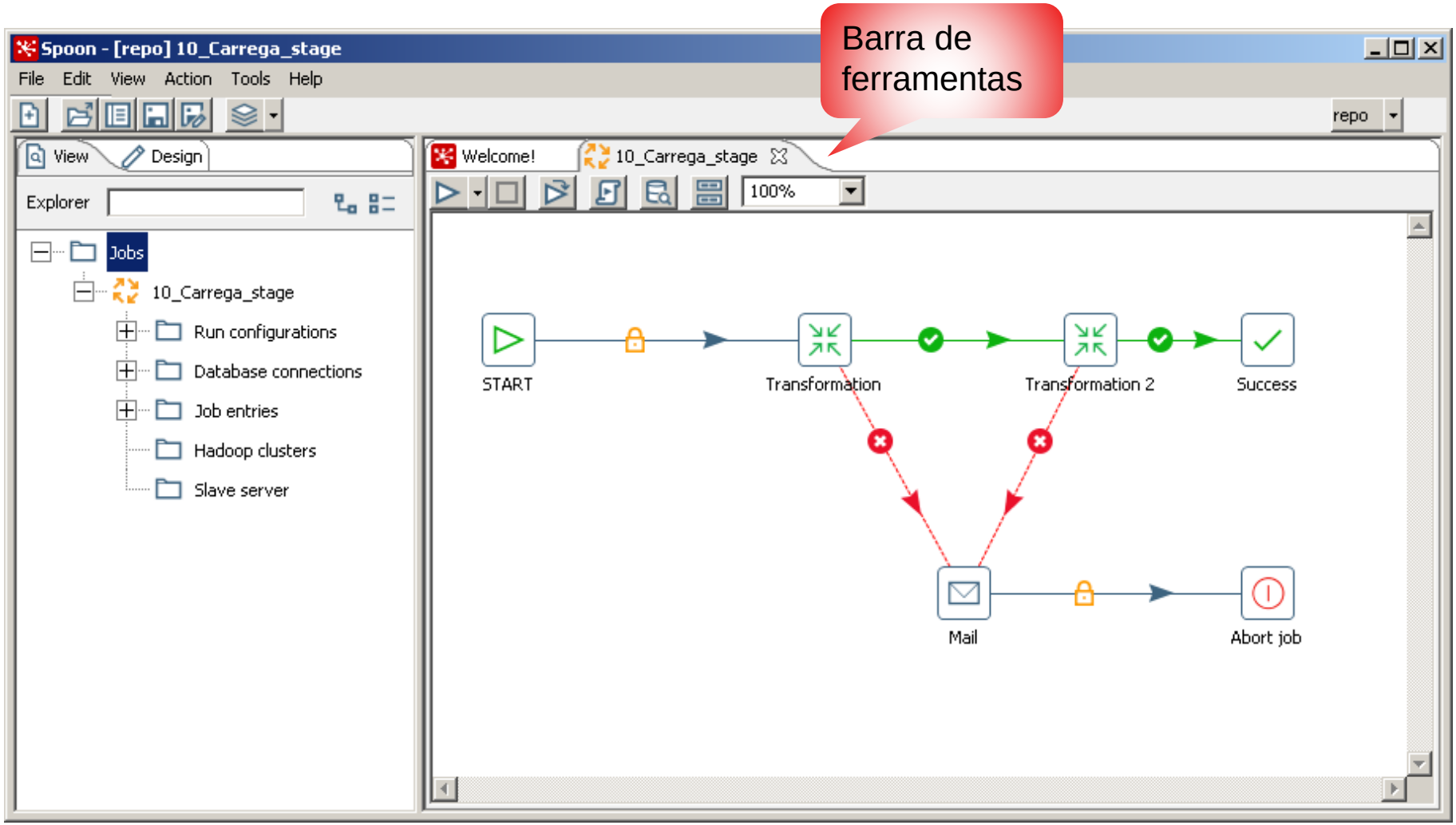
- 4 GB RAM (recomendado)
- 5 GB Espaço em disco (recomendado)
- Processador 32/64 Bits Dual-Core ou Core-2-Dual
- CPU 1.8GHz ou superior

Usando e Iniciando o Spoon

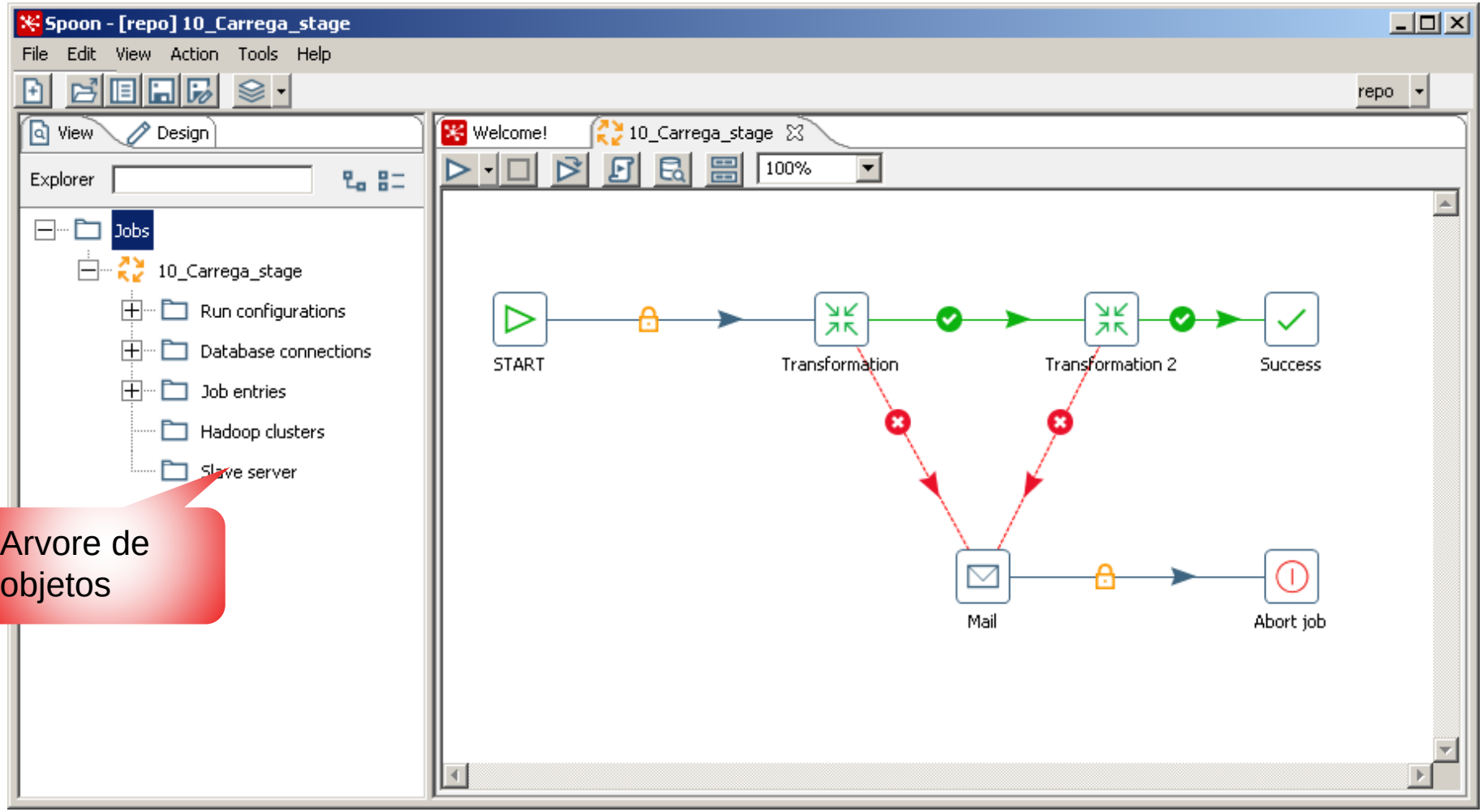
- Em nosso treinamento faremos uso de repositório baseado em arquivo, todos o metadados estarão armazenados em arquivos de sistema com extensão KTR (Transformations) ou KJB (Jobs), em formato XML
- Iniciar **Spoon.bat** (Windows) ou **Spoon.sh** (Linux, MacOS) na pasta do Pentaho data Integration



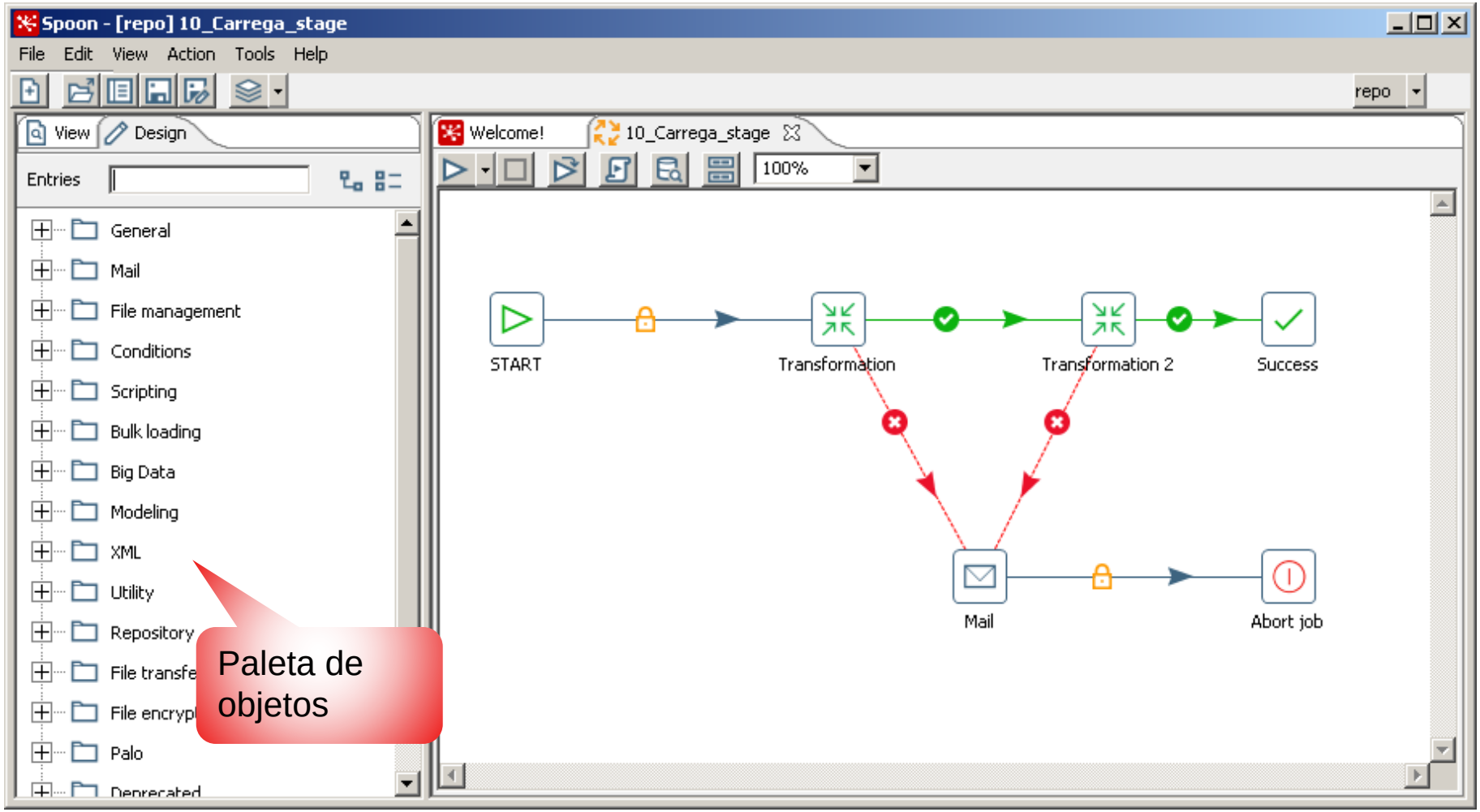
A interface Spoon



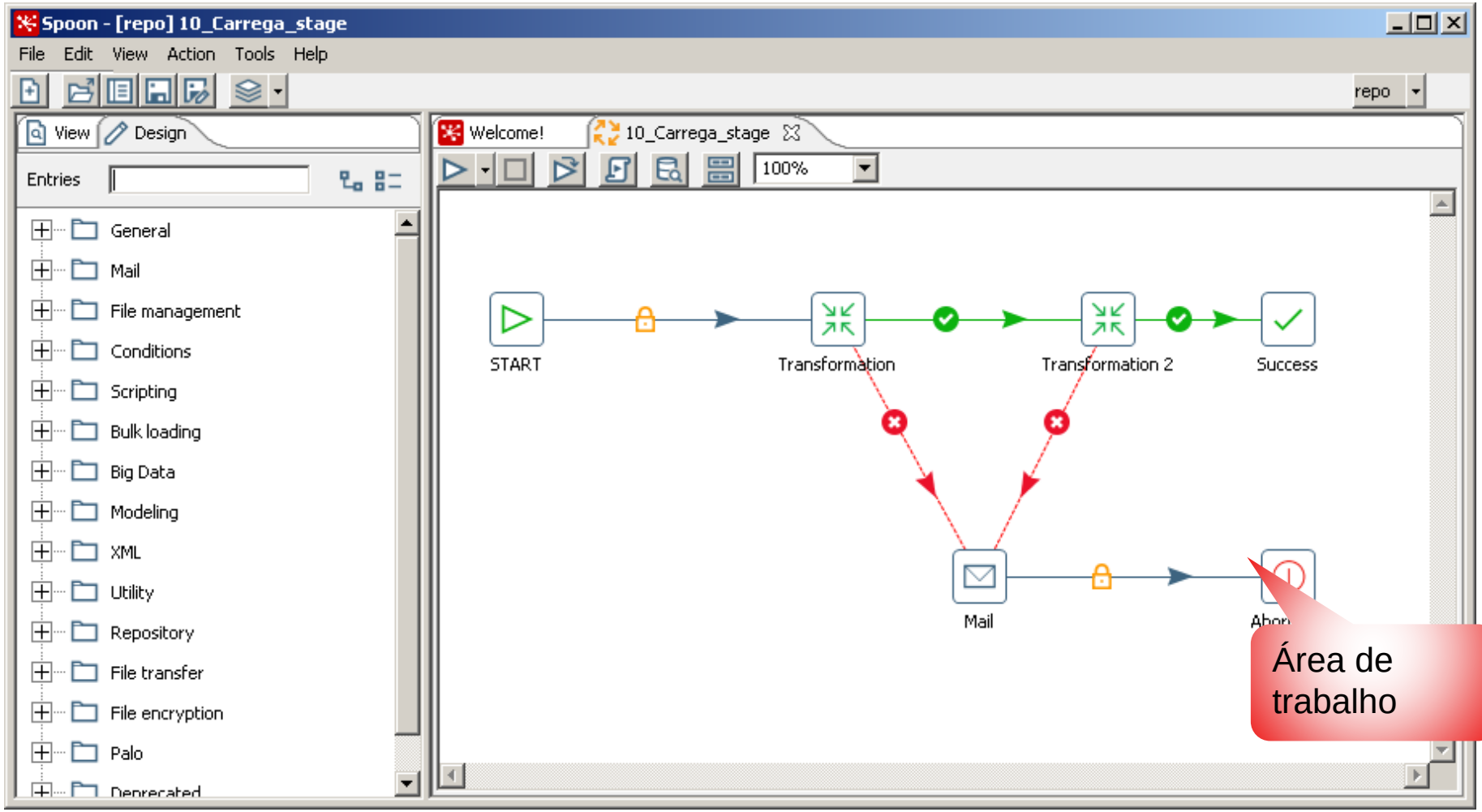
A interface Spoon



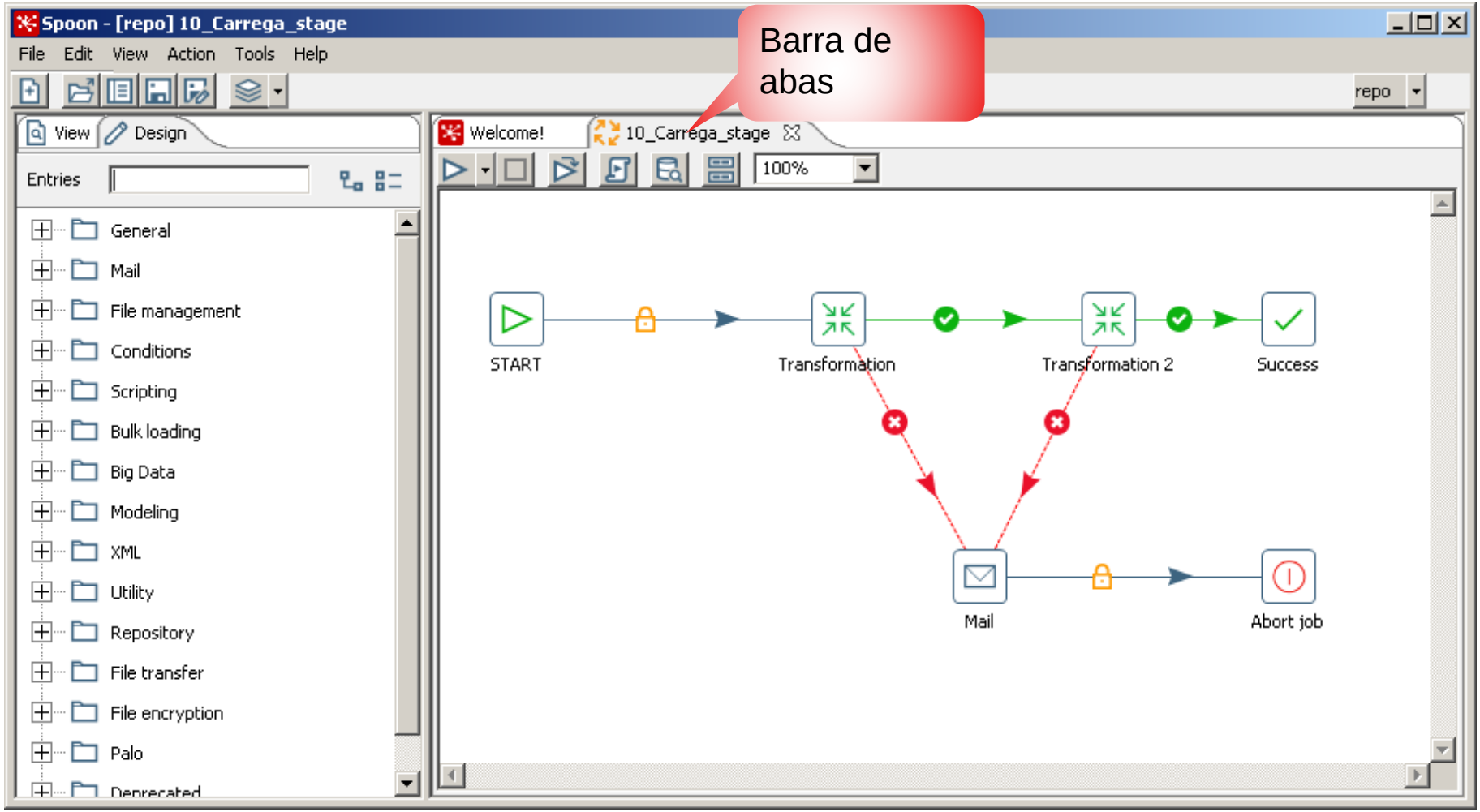
A interface Spoon



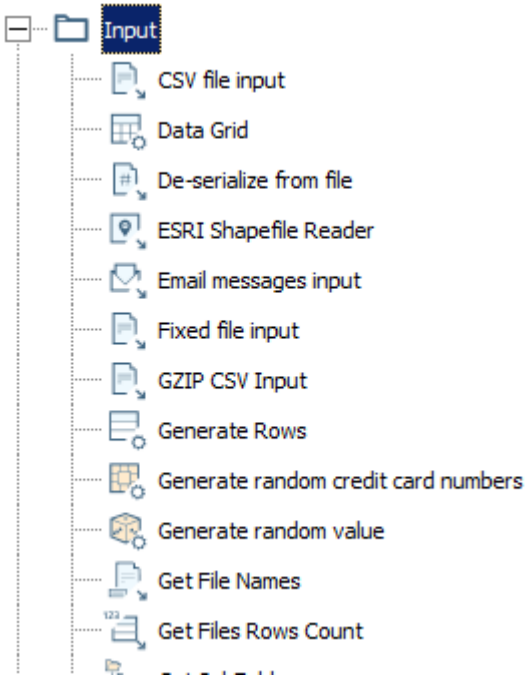
A interface Spoon

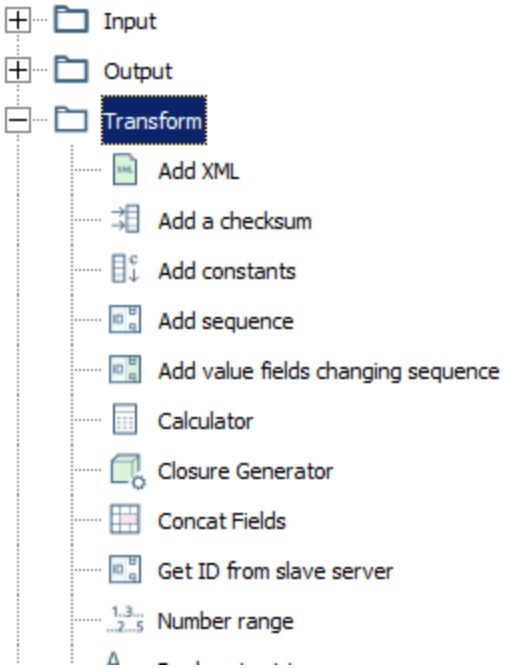


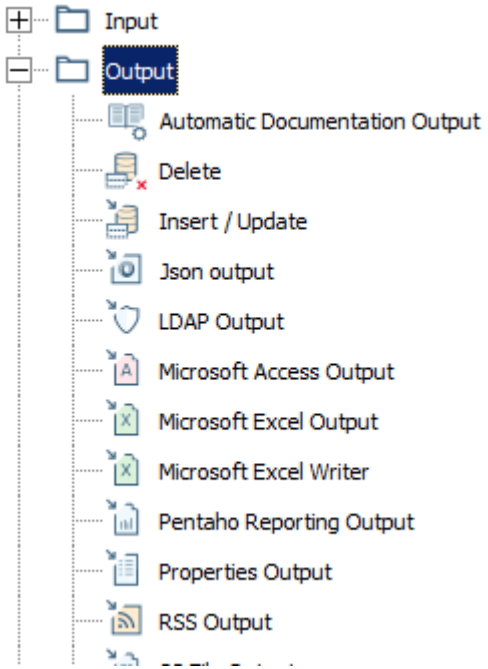
A interface Spoon



Steps







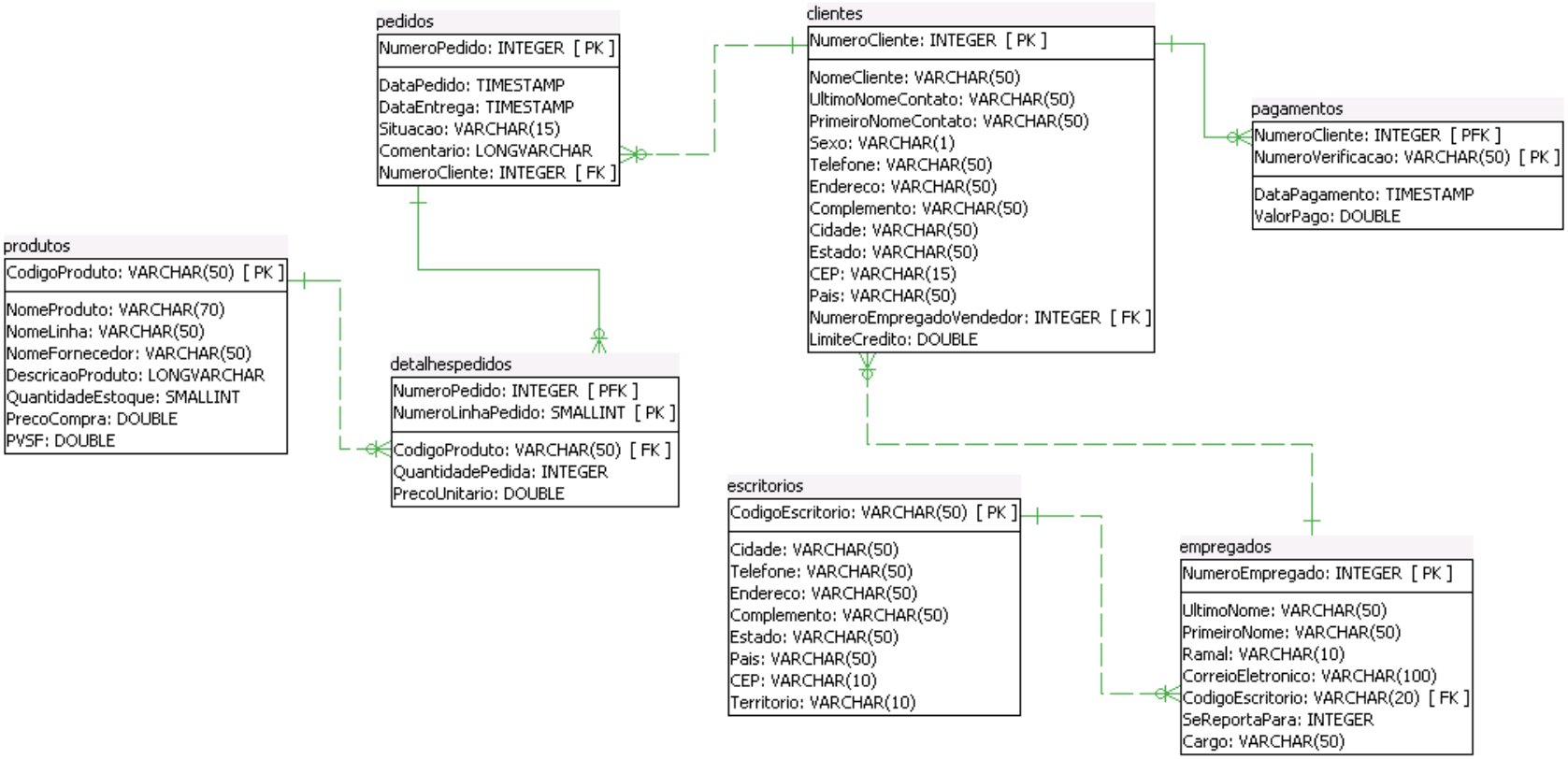
Perguntas e Respostas

OPENiN

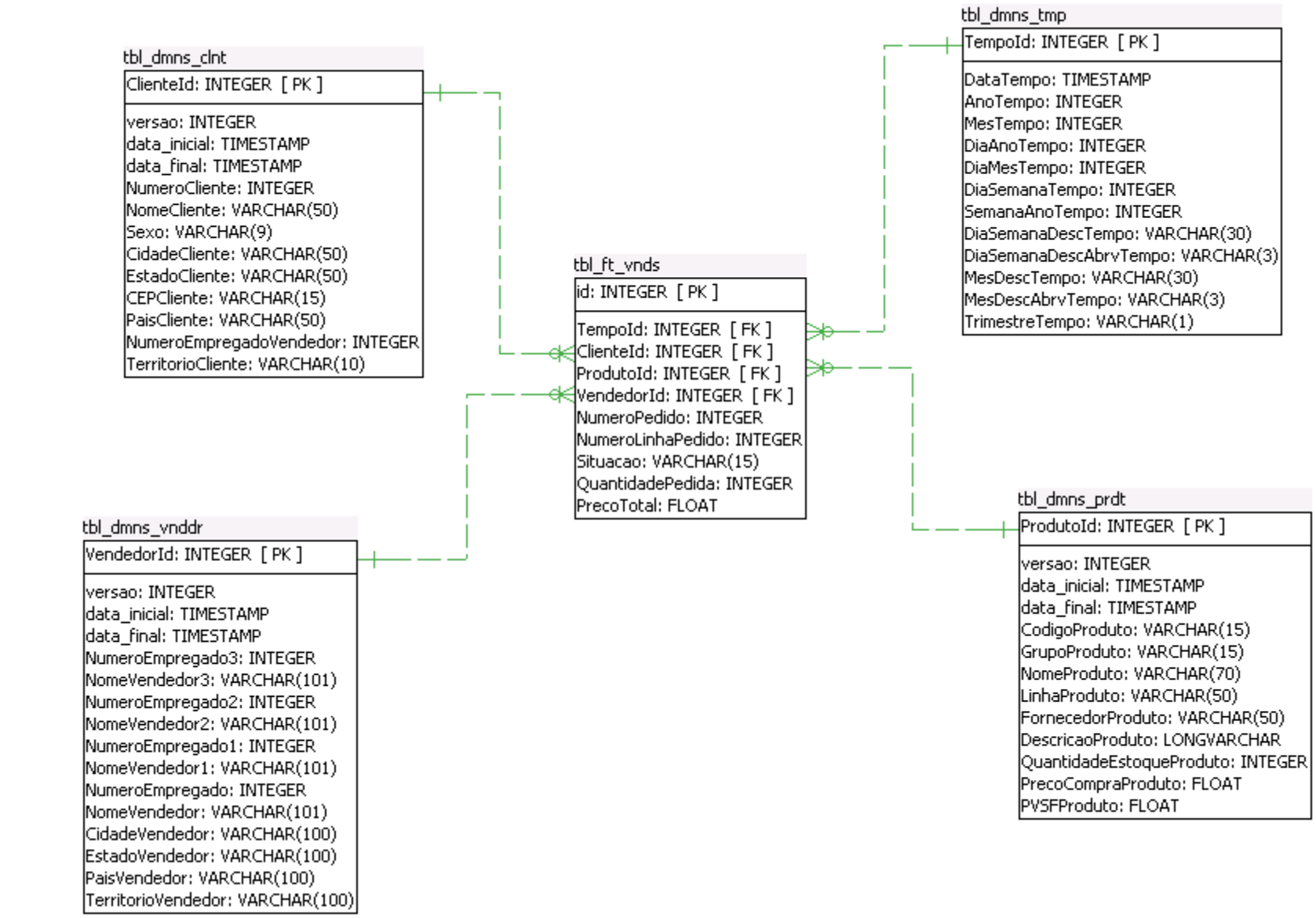
YOUR PARTNER IN DATA

Os modelos: Origem e Destino

Modelo relacional



Modelo dimensional



Perguntas e Respostas

Demonstrando o uso do Spoon

Exercícios

Check List

- Configuração da memoria do spoon
- Configuração do driver jdbc
- Restaurar os bancos

Exercícios do treinamento

- O processo de ETL (do inglês Extract, Transform and Load)
- Conceituação sobre estratégias de cargas
- Apresentando o Pentaho Data Integration (Kettle)
  - Passos para popular um Staging Area com transformações
  - Passos para popular um Data Warehouse (Data Marts) com transformações
  - Trabalhando com repositório de metadados
  - Criando e Compartilhando conexões JDBC e JNDI
  - Usando variáveis para conexões e caminhos de arquivos/pastas
  - Dimensão Tempo com Role Playing Dimension
  - Dimensão Degenerada
  - Padrões de projeto (Design Patterns)
  - Inserção e atualização de registros (Slowly Change Dimension, type 1 e 2)
  - Gerenciamento de erro em tempo de execução (Error handling)
  - Identificação visual de gargalos de processamento
  - Orquestrando processos de ETL com jobs
  - Executando processos de ETL por linha de comando
  - Agendando processos de ETL
  - Enviando e-mail de sucesso e/ou falha
  - Visualizando e gerando logs de execução

Exercício 1

- Criando e editando parâmetros de conexões

  O objetivo deste exercício é apresentar ao aluno a possibilidade de trabalhar com parâmetros de conexões externos a ferramenta, permitindo uma grande facilidade nas mudanças de configurações relacionadas aos ambientes de bancos de dados, como Desenvolvimento, Homologação e Produção.

Tempo médio para a construção do exercício: **10 minutos**
Complexidade para a construção do exercício: **baixa**

Explicando os parâmetros

des_star/type=javax.sql.DataSource

des_star/driver=com.mysql.jdbc.Driver

des_star/url=jdbc:mysql://localhost:3306/motor-inc-star

des_star/user=root

des_star/password=root

ATENÇÃO: Depois de colado o texto ao lado no arquivo jdbc.properties, confira se não há espaços a direita e também não deixe espaço entre os blocos de texto.

Exercício 2

- Criando o repositório de metadados

  O objetivo deste exercício é apresentar ao aluno como fazer para criar um repositório de metadados baseado em arquivos para ser utilizado pelas ferramentas da solução Pentaho Data Integration.

Tempo médio para a construção do exercício: **10 minutos**
Complexidade para a construção do exercício: **média**

Exercício 3

- Criando e editando variáveis

  O objetivo deste exercício é apresentar ao aluno como fazer para utilizar o mecanismo de variáveis. Este recurso é muito importante, pois permite grande facilidade no uso de configurações como caminhos de diretórios em ambientes diferentes e nome de conexões em ambientes de desenvolvimento, homologação e produção.

Tempo médio para a construção do exercício: **10 minutos**
Complexidade para a construção do exercício: **baixa**

Explicação das variáveis

```
#Variável que define um lugar para armazenar arquivos
caminho=C:\\treinamento\\design-tools\\data-integration\\export\\

#Variável que define qual JNDI será utilizado
star=des_star
```

ATENÇÃO: O conteúdo da variável caminho deve condizer com o local na sua estrutura de diretórios e cuidado com os espaços em branco a direita de cada valor

Exercício 4

- Criando um processo de carga na Staging 1

  O objetivo desta *transformation* é apresentar ao aluno um caso de uso da área de estagiamento. Nesta *transformation* iremos juntar informações de empregado com território, mas extraindo esses dados em momentos distintos. Os dados carregados por esta *transformation* servirá de apoio ao próximo exercício.

Tempo médio para a construção do exercício: **15 minutos**
Complexidade para a construção do exercício: **baixa-média**

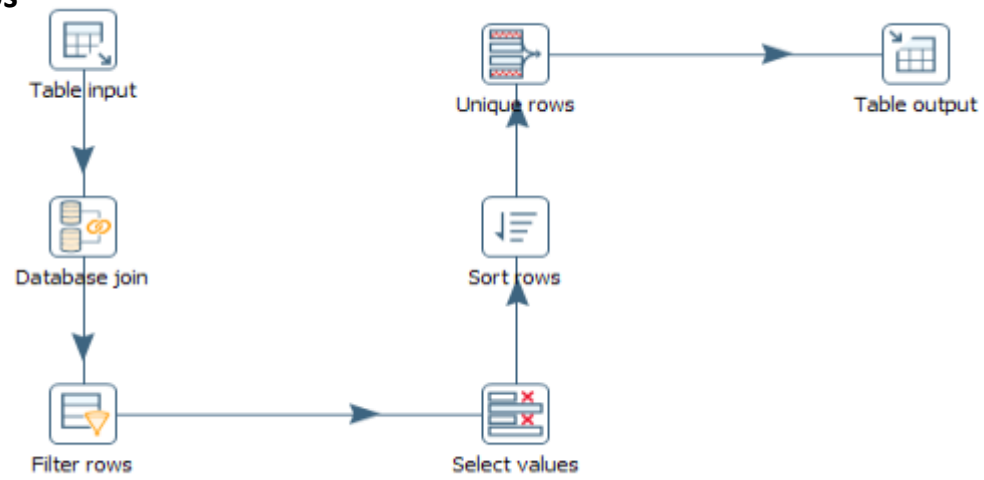
Exercício 5

- Criando um processo de carga na Staging 2

  O objetivo desta *transformation* é apresentar ao aluno um outro caso de uso da área de estagiamento. Nesta *transformation* iremos juntar informações de clientes como País e complementar com informações vindas do vendedor, como Território, pois a tabela de clientes não possui a informação de Território. É uma espécie de *Data Quality*. O resultado desta *transformation* será utilizado para compor com os outros dados de clientes no processo de carga da dimensão cliente.

Acompanhe a explicação do instrutor

Tempo médio para a construção do exercício: **35 minutos**
Complexidade para a construção do exercício: **média**

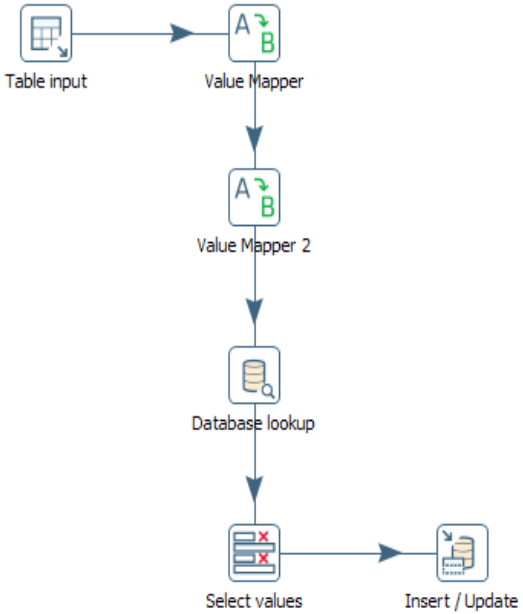


Exercício 6

- Criando o processo de carga da Dimensão Cliente

  O objetivo desta *transformation* é apresentar ao aluno como utilizar uma informação gerada em processo anterior e que está residida na área de estagiamento. Nesta *transformation* iremos complementar as informações de clientes com o território, pois grande parte dos clientes não possui a informação de território em sua tabela.

Tempo médio para a construção do exercício: **30 minutos**
Complexidade para a construção do exercício: **média**

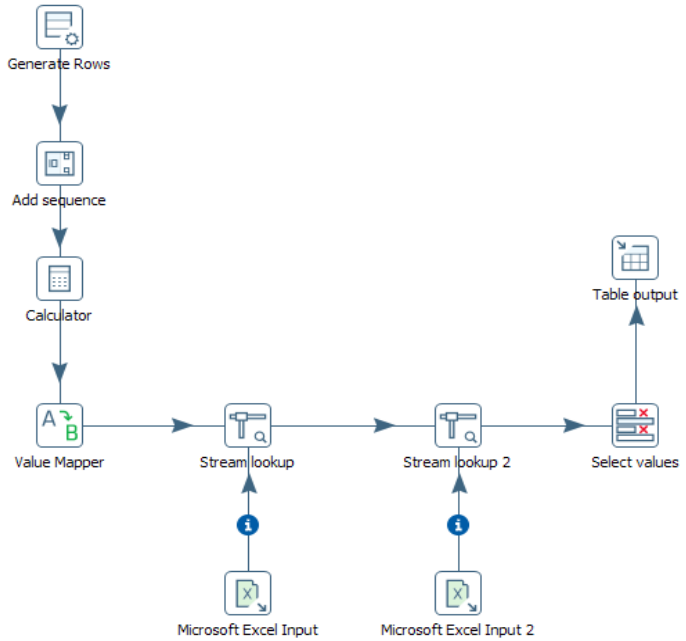


Exercício 7

- Criando o processo de carga da dimensão Tempo

  O objetivo desta *transformation* é apresentar ao aluno a construção de um processo que irá popular a dimensão tempo. Nesta transformação utilizaremos informações manuais e inseridas nos próprios *steps*, dando uma excelente ideia do potencial da ferramenta em termos de inserir informações não sistematizadas.

Acompanhe a explicação do instrutor

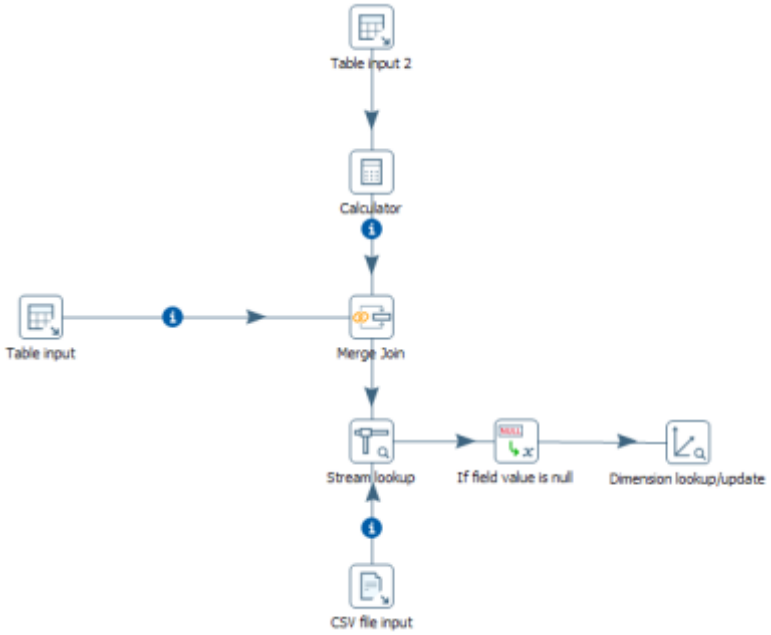


Exercício 8

- Criando o processo de carga da dimensão Vendedor

  O objetivo desta transformação é apresentar ao aluno a possibilidade em se desnormalizar uma estrutura de dados e ainda obter informações de uma fonte em arquivo texto. Nesta transformation estamos desnormalizando a estrutura comercial da motor-inc, a fim de, criar uma estrutura dimensional dinamica, permitindo a inserção novos vendedores de forma automática. Aqui teremos também o controle de versionamento, SCD do tipo 2

Acompanhe a explicação do instrutor

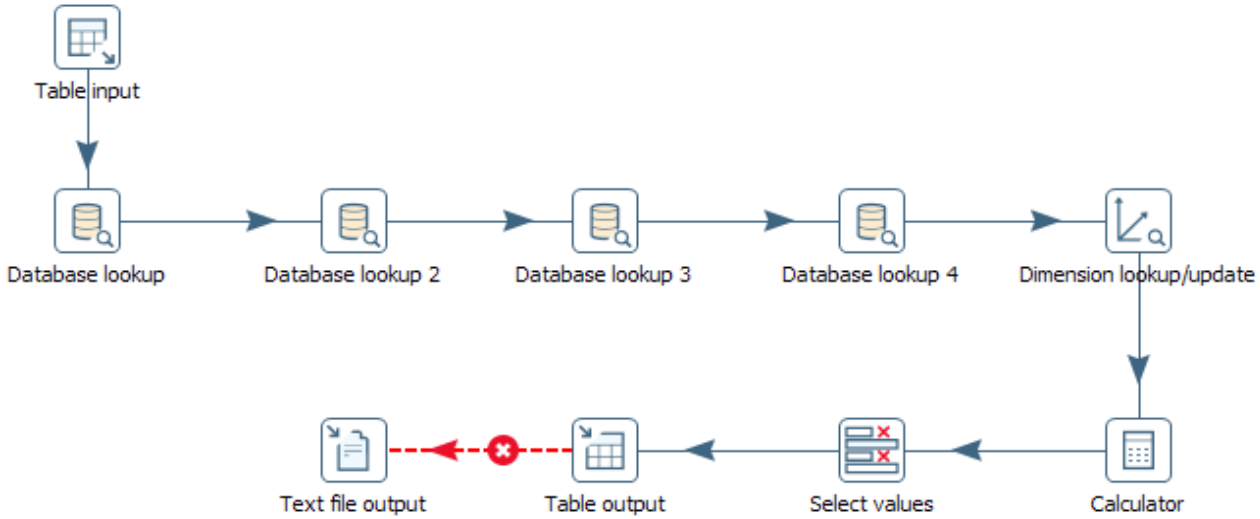


Exercício 9

- Criando o processo de carga da Fato Vendas

  O objetivo desta *transformation* é apresentar ao aluno a criação do processo de carga da tabela fato Vendas. Nesta *transformation*, além das informações extraídas da origem, existe um processo troca (lookup) de informações entre as dimensões. A carga de dados é sempre inserção e caso haja algum erro de violação de constraint o registro será expurgado para um arquivo texto. Neste momento iremos também gerar uma informação nova, um procedimento de enriquecimento de dados.

Tempo médio para a construção do exercício: **40 minutos**
Complexidade para a construção do exercício: **média-alta**

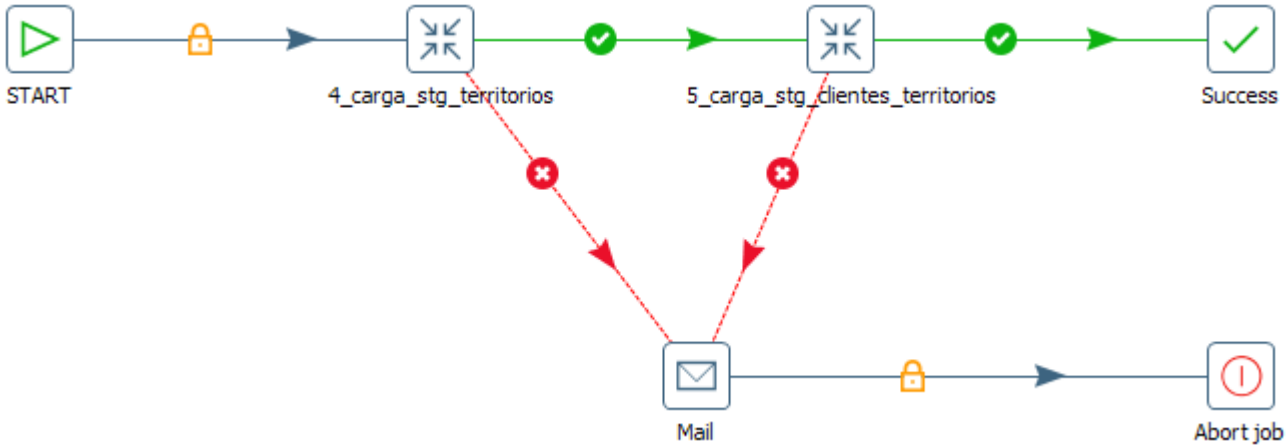


Exercício 10

- Orquestrando a execução dos processos da Staging

Neste Job, apresentaremos ao aluno, como encadear as chamadas, primeiramente pelos processos da área de estagiamento, resultando em sucesso ou em falha e neste caso, enviando e-mail de erro.

Tempo médio para a construção do exercício: **15 minutos**
Complexidade para a construção do exercício: **baixa**



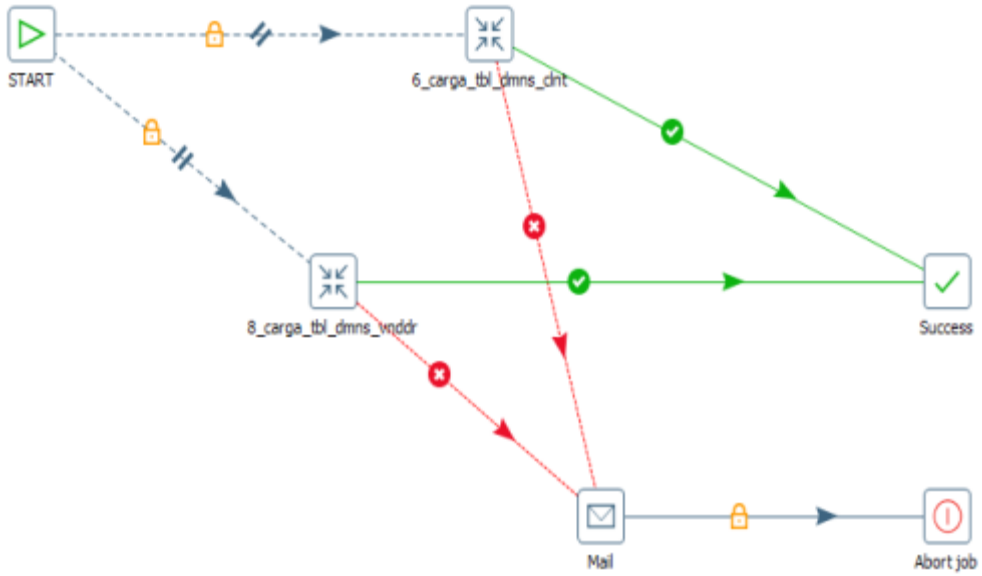
Exercício 11

- Orquestrando a execução dos processos Dimensões

  Neste Job, apresentaremos ao aluno como encadear as chamadas pelos processos que carregam as dimensões de Produto, Cliente, Tempo e Vendedor, resultando em sucesso ou em falha e neste caso, enviando e-mail de erro.

Acompanhe a explicação do instrutor

Tempo médio para a construção do exercício: **20 minutos**
Complexidade para a construção do exercício: **média**



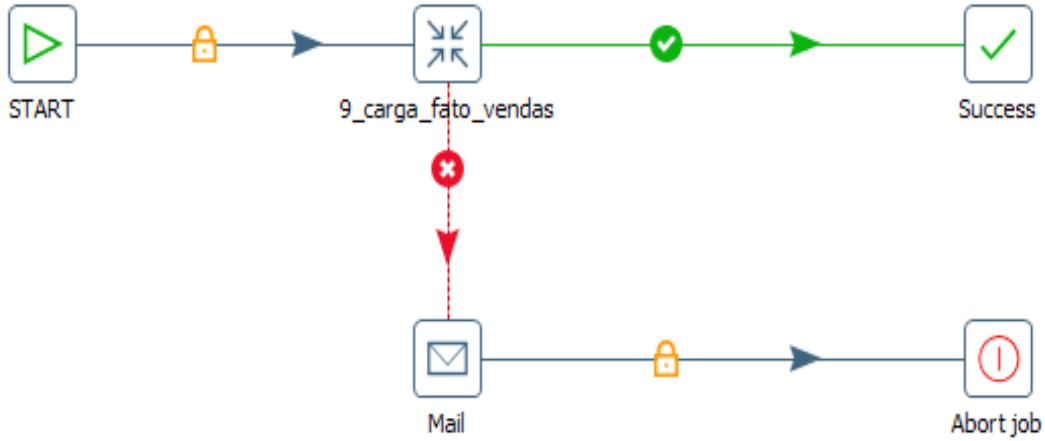
Exercício 12

- Orquestrando a execução do processo Fato

Neste Job, apresentaremos ao aluno como encadear a chamada ao processo da tabela fato vendas, resultando em sucesso ou em falha e neste caso, enviando e-mail de erro.

Tempo médio para a construção do exercício: **5 minutos**
Complexidade para a construção do exercício: **baixa**

Acompanhe a explicação do instrutor

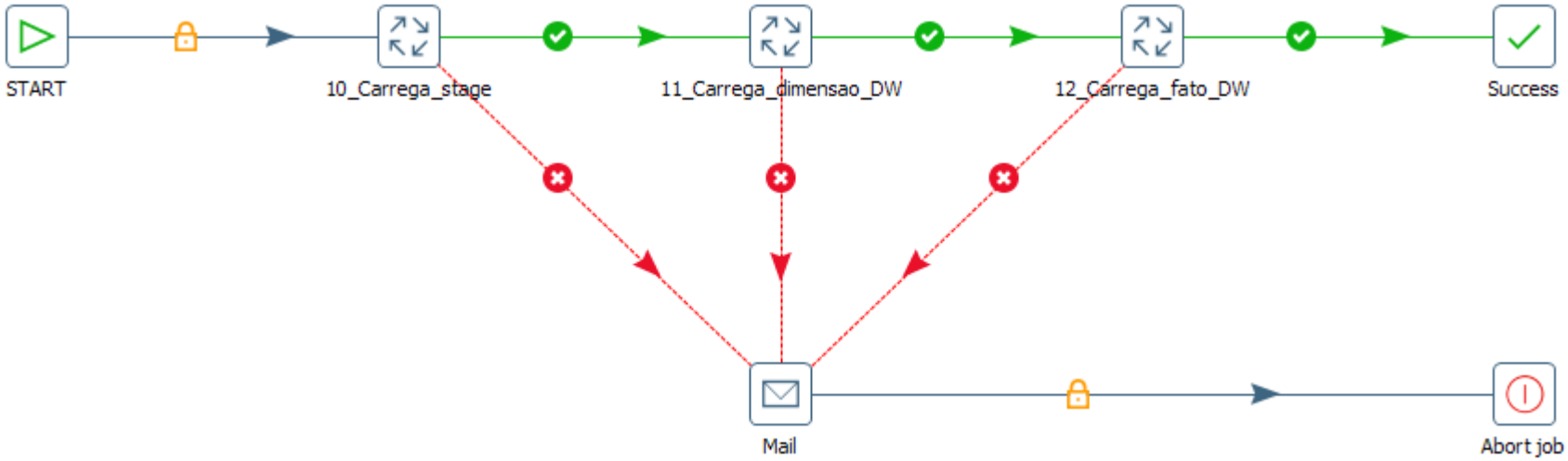


Exercício 13

- Orquestrando todos os processos do DW

  Neste Job, apresentaremos ao aluno como encadear as chamadas primeiramente da área de estagiamento, na sequencia dos processos das dimensões e por último o processo da tabela fato vendas, resultando em sucesso ou em falha e neste caso, enviando e-mail de erro.

Tempo médio para a construção do exercício: **5 minutos**
Complexidade para a construção do exercício: **média**



Exercício 14

- Executando um Job através de linha de comando

  Neste exercício apresentaremos ao aluno como executar um Job através de linha comando, usando a ferramenta Kitchen.

Tempo médio para a construção do exercício: **20 minutos**
Complexidade para a construção do exercício: **média**

Perguntas e Respostas

Final do treinamento ETL1000