



Pentaho,
garantindo o
sucesso do BI

Pentaho Data Integration

- Objetivo:
- Falar um pouco do Pentaho e suas ferramentas que vem ganhando mercado e mostrar um pouco do que elas podem oferecer tanto para projetos quanto para automatização de tarefas.
- Pentaho + PowerBI / Tableau / Qlick
- Para que você consiga aqueles gráficos e dashboards maravilhosos, sonho de todo gestor, você precisa tratar os dados e a ferramenta Pentaho vai contribuir muito com essa atividade.

O que veremos?

- Overview da Suite Pentaho e seus componentes
- Instalação do Pentaho
- Conexão com o banco de dados
- Joins com tabelas
- Criação de campos calculados e expressões
- Steps uteis para análise de dados
- Agendamento de job

O que é o Pentaho?

É uma suíte de Business Intelligence e Analytics.

É um software Open Source (gratuito, porém possui também sua versão paga, que te assegura suporte técnico por profissionais que conhecem muito da ferramenta), pertence ao grupo Hitachi Vantara.

O objetivo do Pentaho é criar um único fluxo de trabalho de dados analíticos para nossos clientes, independentemente do tamanho dos dados, carga de trabalho e caso de uso.

A suíte Pentaho é formada por um conjunto de softwares voltados para construção de soluções de BI de ponta-a-ponta, que inclui programas para extrair os dados de sistemas de origem em uma empresa, gravá-los em um data warehouse, limpá-los, prepará-los e entregá-los a outros sistemas de destino ou mesmo a outros componentes da suíte para estudar ou dar acesso aos dados ao usuário final.

Permite conexão com diversos tipos de bancos de dados e no analytics permite colocar dentro do seu fluxo Python, R, Hadoop, Spark (para processamento de big data), que são tecnologias novas e que você consegue colocar dentro do seu ETL.



Quem utiliza?

- Indústrias como assistência médica, automotiva e manufatura estão aplicando o Pentaho para integração e análise de big data para se tornarem líderes digitais em seu setor.

Para que serve?

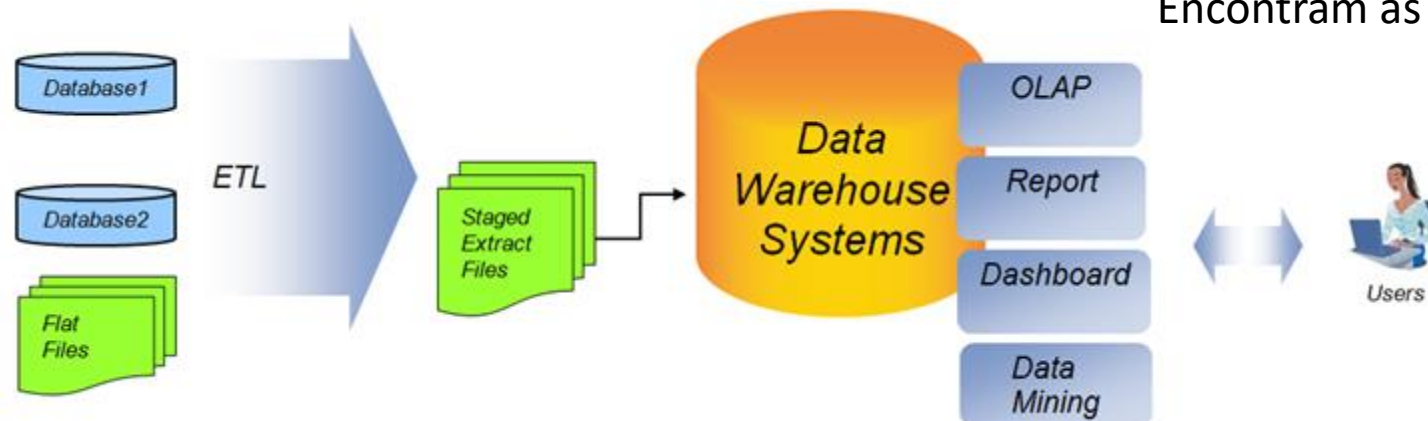
- Integração e ingestão de qualquer volume de dados.
- Automatização de análise de processos – geração de arquivos automaticamente em formato amigável para o usuário, por exemplo excel, carga de dados e subir o arquivo para o ftp ou no Google sheets
- Integração de diversas fontes de dados (Oracle, SQL, MySQL, MariaDB, noSQL), vários tipos de arquivo JSON, XML, APIs, dentre outros.
- Ferramenta fundamental para projetos de BI ou Big Data. Pois num projeto de BI, por exemplo, onde se gasta maior tempo?

É exatamente na preparação das informações e esse é o carro chefe do Pentaho.

Business Intelligence

- É o processo de transformar dados em informações relevantes para a tomada de decisão, pelos gestores.

Arquitetura de um Projeto de BI:



Data Warehouse: Banco de dados onde se Encontram as tabelas Dimensões e Fatos

Visualização de Dados

The diagram features a central iceberg floating in a dark blue sea. The iceberg is divided into two main sections: a small, light blue tip above the water and a large, dark blue base below the water. The top section is labeled 'Visualização de Dados' and includes 'Data Science & Machine Learning', 'Dashboards', and 'Relatórios Automatizados'. The bottom section is labeled '80%' and includes 'Extração de Dados', 'Entender Requisições De Dados', 'Transformação de Dados', 'Modelagem De Dados', 'Agendamento de Jobs', and 'Filtros e Customização De Processos de Dados'. The background is a light blue sky with a few birds flying.

Data Science &
Machine Learning

Dashboards

20%

Relatórios
Automatizados

Relatórios

Extração de Dados

Agendamento de Jobs

Entender Requisições
De Dados

Filtros e Customização
De Processos de Dados

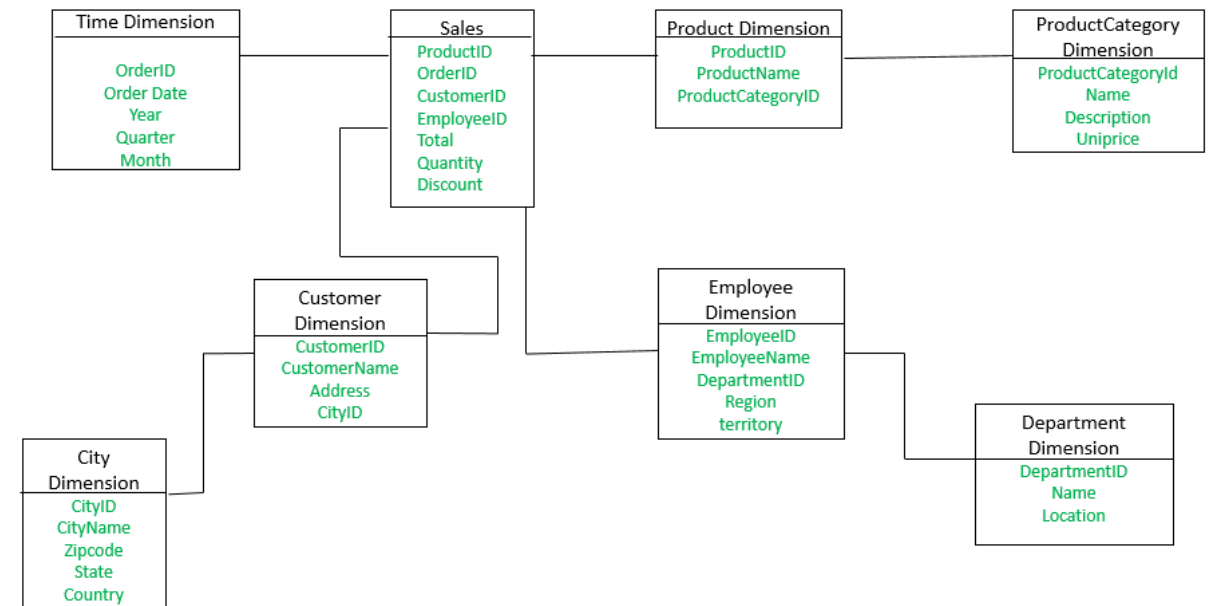
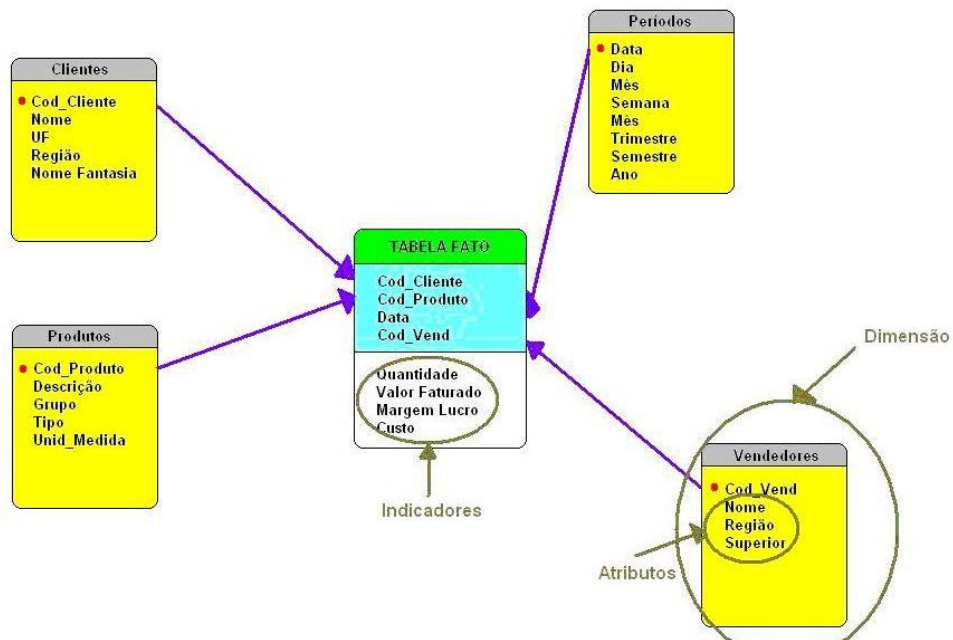
80%

Transformação de
Dados

Modelagem
De Dados

Tipologias

- Star Schema – Uma Fato para várias Dimensões
- Snowflake Schema – Uma Fato e várias dimensões com suas dimensões



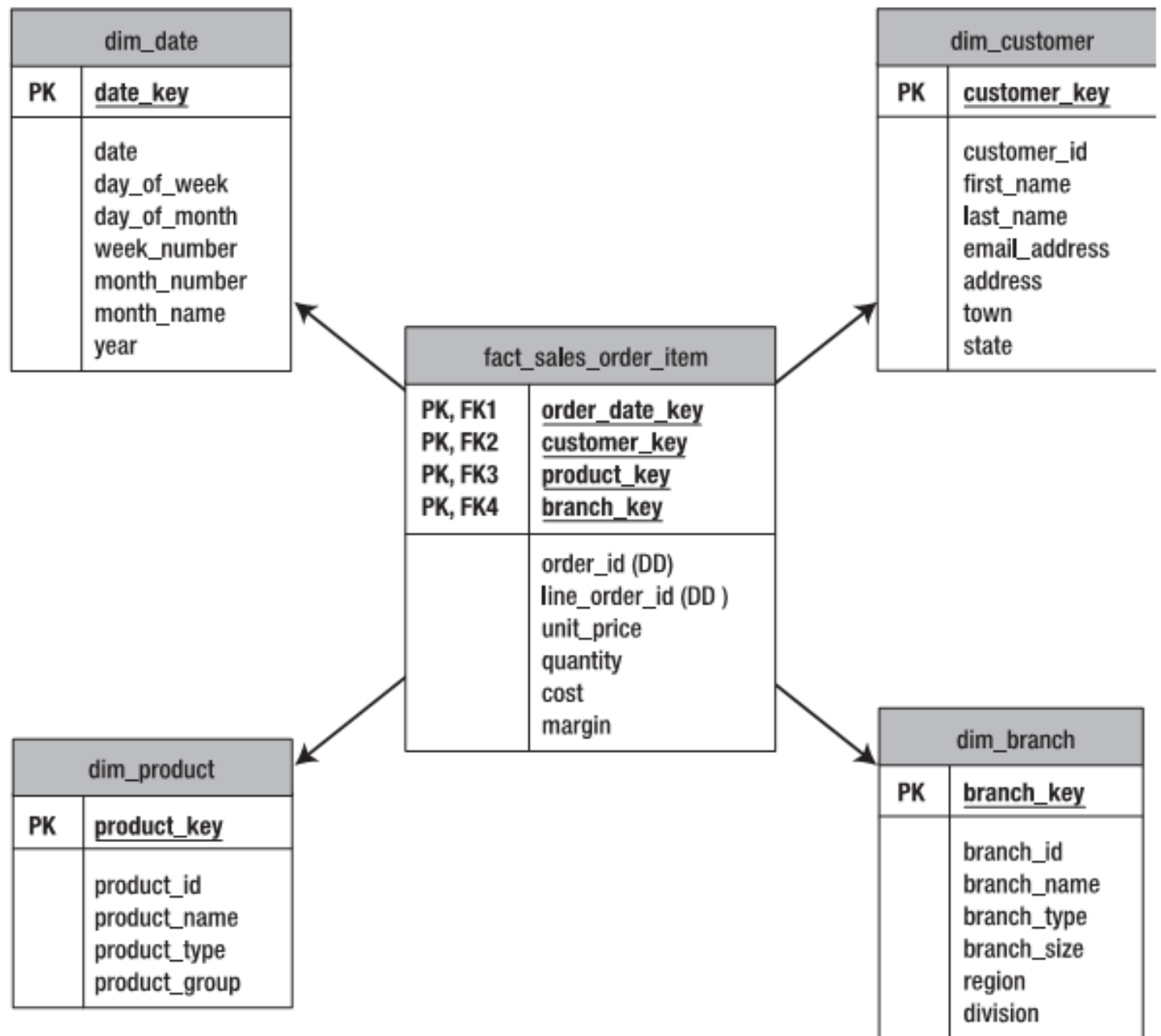
Tipos de Tabelas - DW

- Dimensões: Armazenam todos os registros descritivos, que farão referência à tabela Fato.

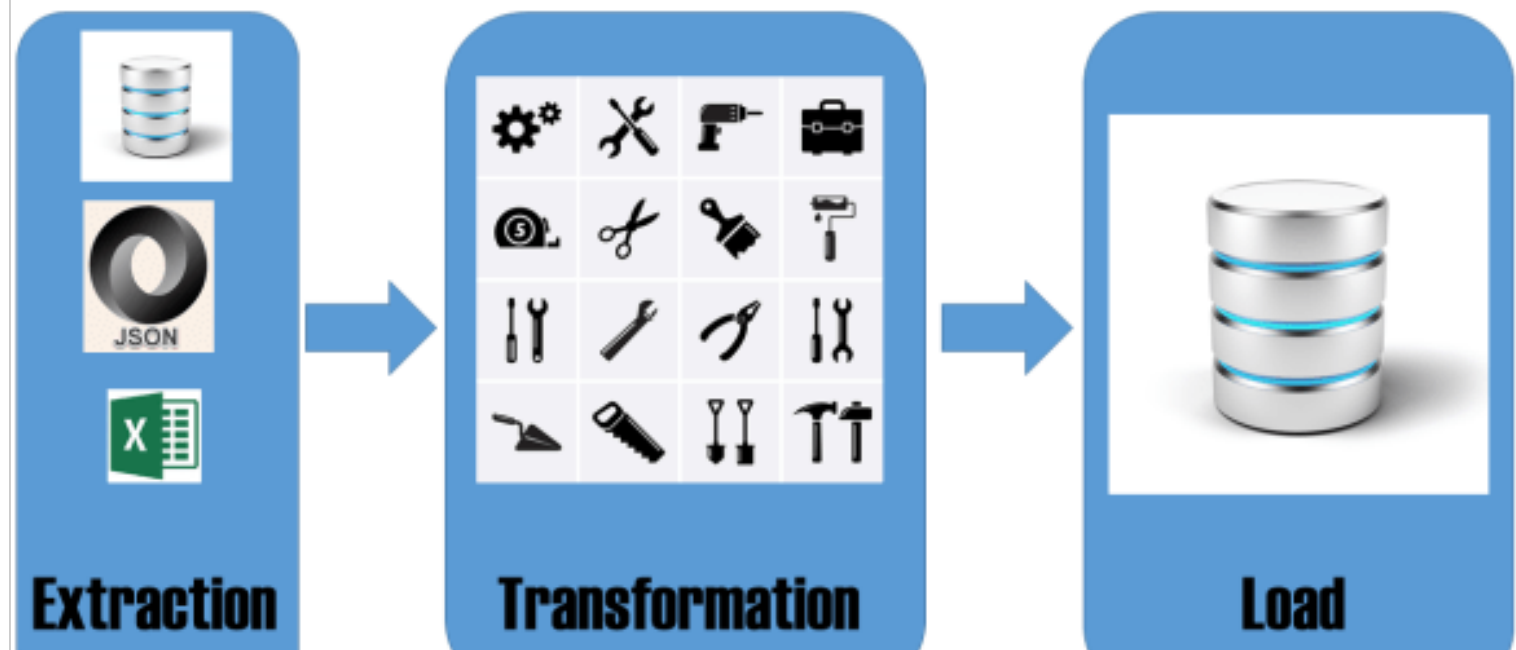
Ex.: Tenho o código do cliente na tabela fato e para saber mais informações teremos a dimensão Cliente com o código do cliente, nome, endereço, etc.

Fatos: A tabela fato é a principal tabela do Data Warehouse, ela se conectará nas dimensões. Nesta tabela serão armazenadas duas coisas: as **Métricas**, que são os fatos propriamente dito, e as **Foreign Keys** que são as chaves para relacionar os dados das Dimensões com a Fato.

Modelagem Data WareHouse



ETL



- **Extract** – Extrair dados
- **Transform** – Transformar, criar cálculos, processar registros nulos, duplicados, etc.
- **Load** – Carregar os dados no data warehouse

Cada vez que surgir a necessidade de novas tabelas e novas informações, o processo de ETL ocorrerá.

Pentaho Data Integration (PDI)

- Migração de dados de um servidor para outro
- Tratamento de dados, através de Vários componentes
- Limpeza de Dados. Ex.: registros nulos num campo Descrição, pode ser verificado e pode ser tratado com o texto “Não Informado”.
- Criação de métricas e indicadores, Levar o cálculo de métrica já pronto para a ferramenta de frontend que vai ler o calculo pronto e vai ganhar em processamento.
- Exportar banco de dados para outros formatos
- Consumo de web services e API. Ex.: Pega o API de outra empresa, cruza os dados com o seu ERP e carrega no Data Warehouse, podendo demonstrar no dashboard um comparativo de sua empresa com o cenário externo.
- Integração com Diversos Bancos de Dados

Componentes PDI



Spoon – Criação de Transformação e Jobs



Pan – Executa as Transformações de forma agendada e automática

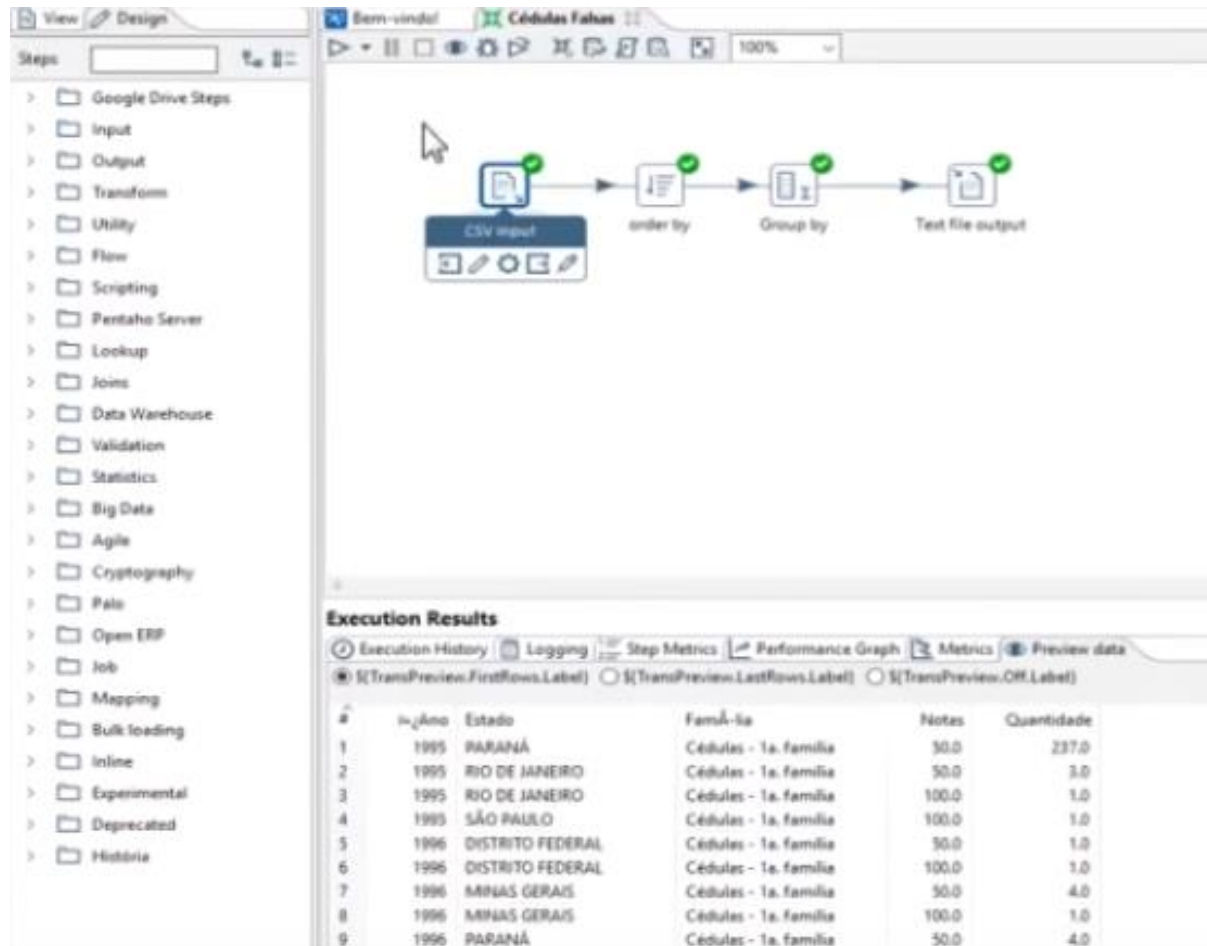
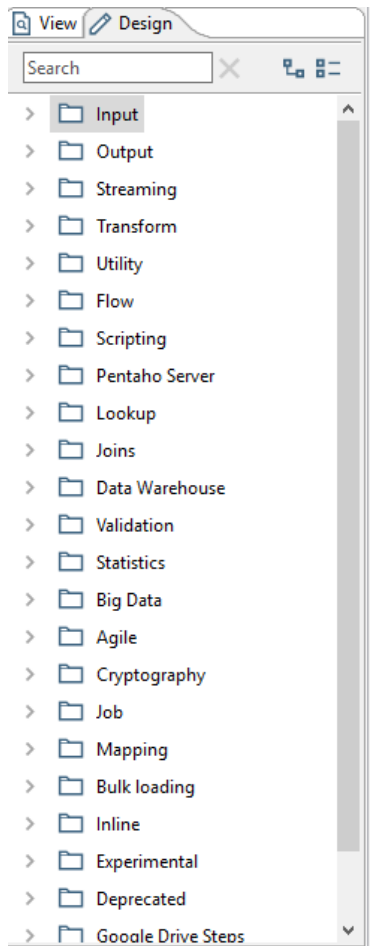


Kitchen – Executa os Jobs de forma agendada e automática



Carte – Web Server para execução remota das transformações e Jobs

Exemplo: Trabalhando com Arquivos



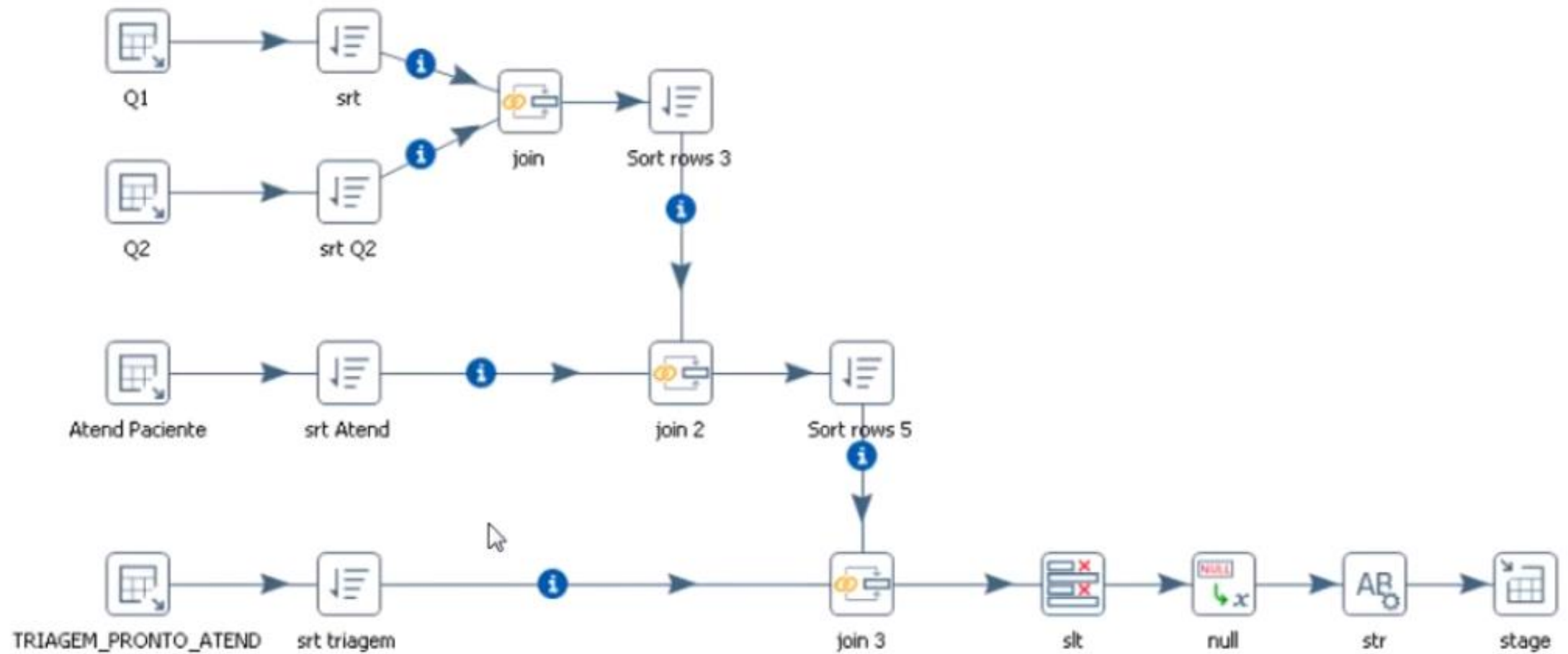
Entrada: Arquivo .csv

Ordenação de Registros

Agrupamento de Dados Ex.:
Agrupar por Estado somando
a quantidade

Saída em outro arquivo excel, txt

STAGE



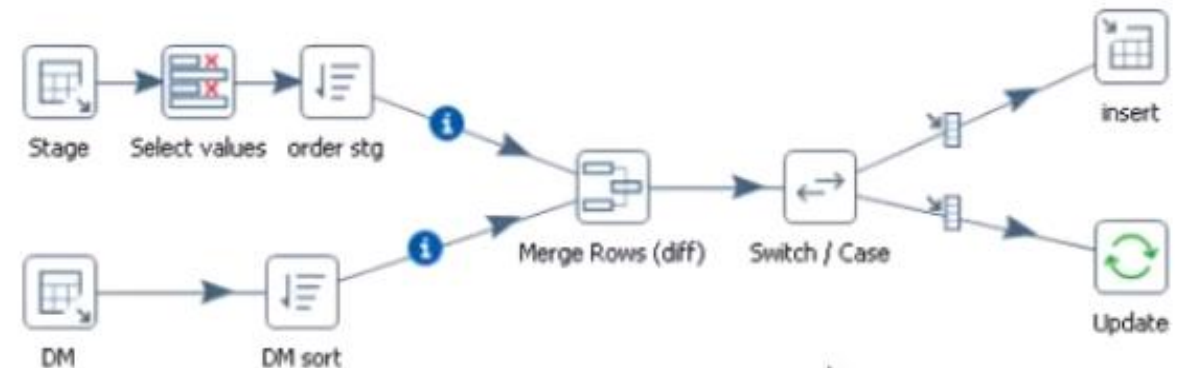
Stage e Data Mart

Stage



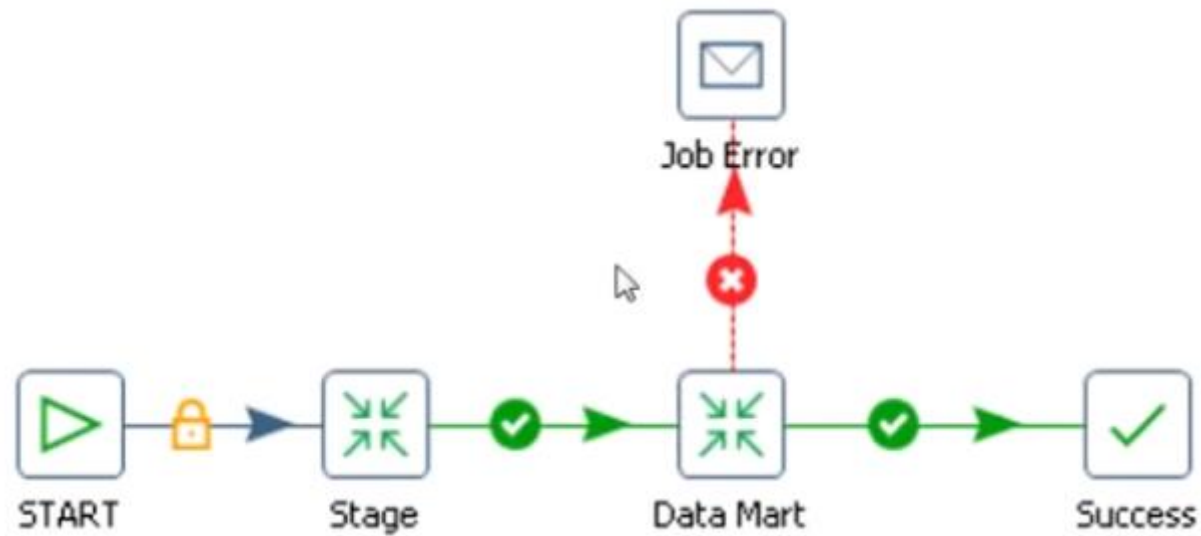
Processo simples de carga de dados.

Data Mart



Carga incremental com switch case verificando se o registro é novo ou se já existe para ser atualizado.

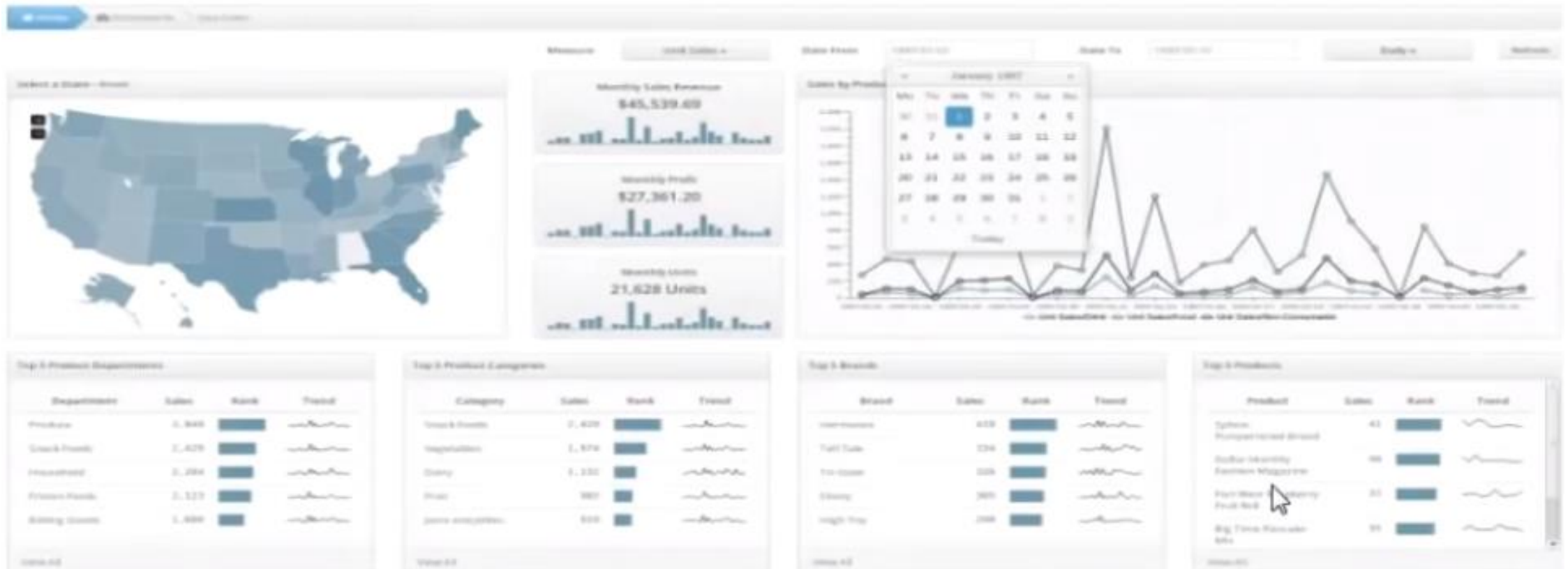
Criando Jobs:



Início o Job, carrega o Stage, carrega o Data Mart, se der erro enviar um email para toda a equipe

Exibição dos dados carregados

Dashboard construído no Pentaho Server componente CDE.



Referências Bibliográficas

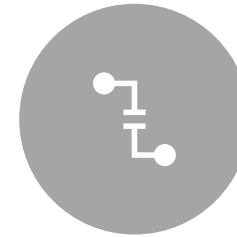
Links:

- Pentaho Fórum: <http://forums.pentaho.org>

Começando!



VOCÊ É O
RESPONSÁVEL
PELOS DADOS QUE
O SEU CLIENTE
TOMA DECISÕES.



ATUALMENTE A
TOMADA DE
DECISÕES, PELOS
GESTORES, É
ORIENTADA A
DADOS.

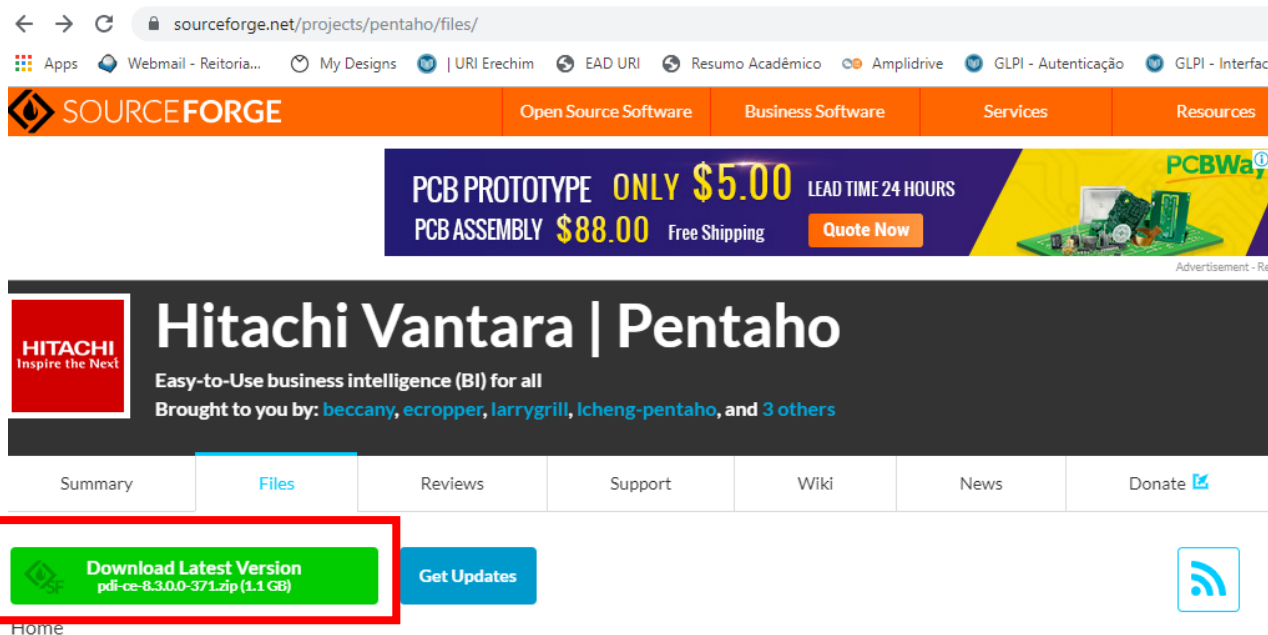


O SEU TRABALHO
GERA MILHÕES EM
FATURAMENTO.

Instalação do Pentaho:

- jdk-8u221-windows-x64
- Parametrização de variáveis de ambiente (no windows)
- Instalação do Pentaho:

<https://sourceforge.net/projects/pentaho/files/Pentaho%208.3/>



sourceforge.net/projects/pentaho/files/

Apps Webmail - Reitoria... My Designs | URI Erechim EAD URI Resumo Acadêmico Amplidrive GLPI - Autenticação GLPI - Interf

SOURCEFORGE Open Source Software Business Software Services Resources

PCB PROTOTYPE ONLY \$5.00 LEAD TIME 24 HOURS
PCB ASSEMBLY \$88.00 Free Shipping Quote Now

HITACHI Inspire the Next
Hitachi Vantara | Pentaho
Easy-to-Use business intelligence (BI) for all
Brought to you by: beccany, ecropper, larrygrill, lcheng-pentaho, and 3 others

Summary Files Reviews Support Wiki News Donate

Download Latest Version
pdi-ce-8.3.0.0-371.zip (1.1 GB)

Get Updates

Home

O Pentaho não é um executável, é necessário rodar o Server e em seguida na pasta do Data Integration executar o componente Spoon ou outro.

Spoon.bat no Windows ou Spoon.sh no Linux



Vamos a prática!

Obrigada! ;)

Linkedin: Tatiana Cavalheri

E-mail: tatiana.cavalheri@gmail.com

