



ETL com Pentaho Data Integration PDI

Ecila Alves de Oliveira Migliori

M.a Engenharia de Produção (UNIP)

Especialista em Ciência de Dados e Big Data (PUC-MG)

Especialista em Análise de Sistemas (Mackenzie)

Administração (UNIP)

Processamento de Dados (Mackenzie)

Bolsista Técnico na UNIFESP – SoUCiência - <https://souciencia.unifesp.br/>

17 anos de docência (UNIP e UNICID)

Currículo Lattes: <https://lattes.cnpq.br/1235717715073144>

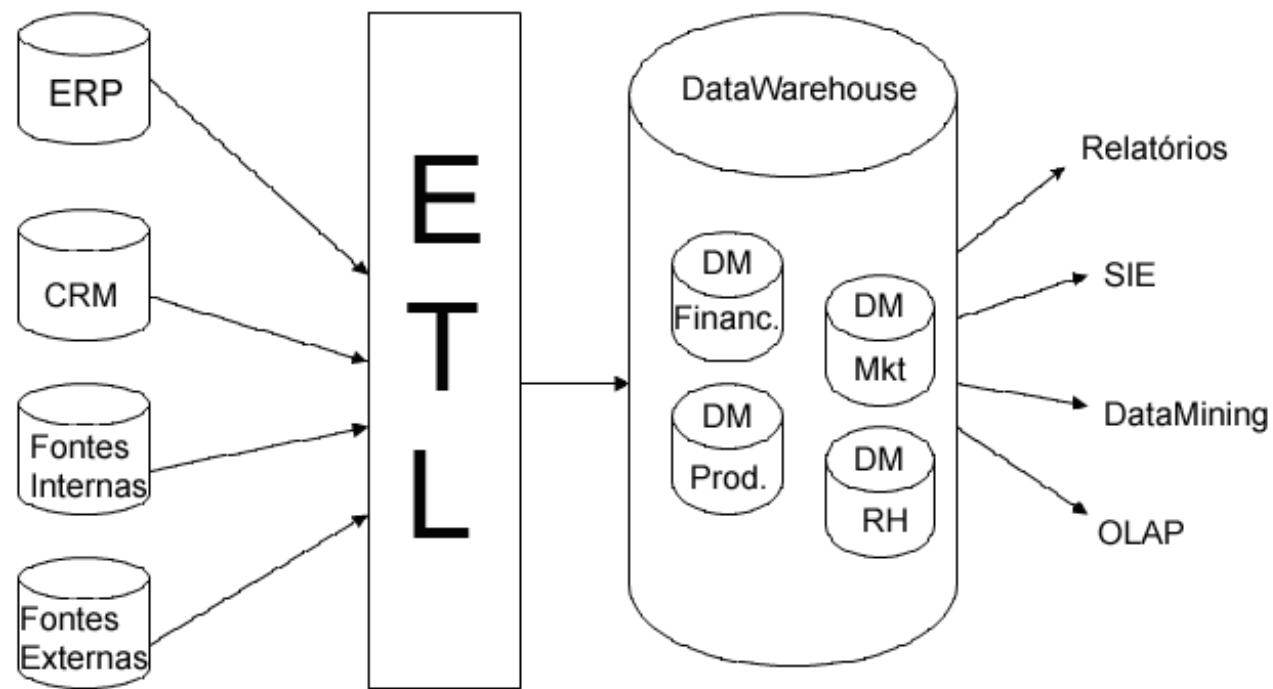
Contato: ecilaoliveira@uol.com.br

WhatsApp: (11)9.9806-7958

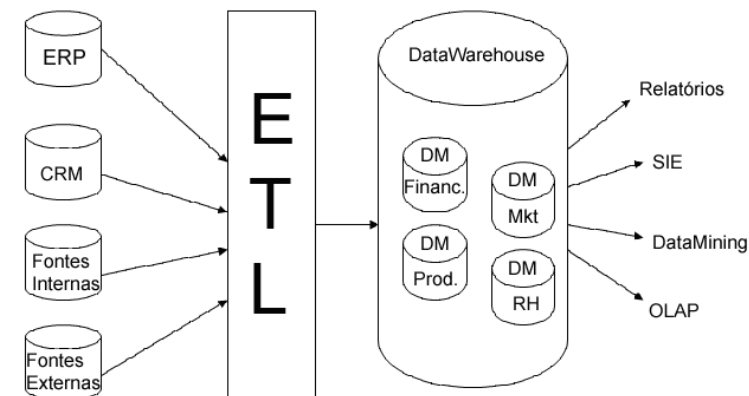
Arquitetura de um BI

Uma solução de BI, qualquer que seja a ferramenta, sempre terá os mesmos elementos.

- Servidor de ETL
- Data Warehouse
- Servidor de Exploração de Dados

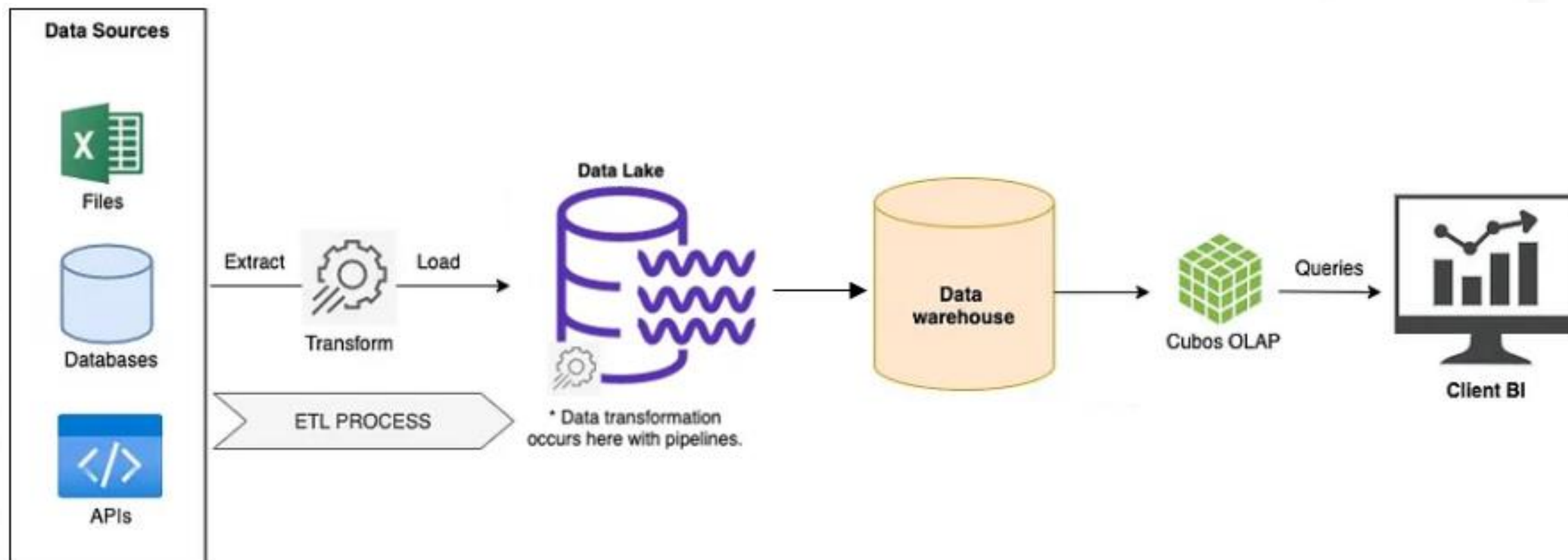


Arquitetura de um BI



- **Servidor de ETL:** ETL nesse contexto significa a máquina que vai executar o processo de extração, transformação e carga das fontes de dados para dentro do DW.
- **Data Warehouse:** DW refere-se ao servidor de banco de dados - hardware e software - que vai cumprir a função de armazém de dados para a solução de BI da empresa. Para definir esse componente é importante conhecer o volume de dados que será carregado inicialmente, a que velocidade (em bytes ou registros por mês) ele vai crescer, quantos usuários poderão consultá-lo e quantas faces do cubo multidimensional (dimensões) ele vai ter.
- **Servidor de Exploração de Dados:** uma vez que os dados estejam disponíveis no DW, os usuários começam a acessá-los e a explorá-los para resolver suas diversas necessidades: medir o desempenho da empresa, responder as perguntas estratégicas, táticas e até mesmo operacionais, planejar e avaliar o resultado das ações e um inimaginável sem números de usos.

Arquitetura de um BI



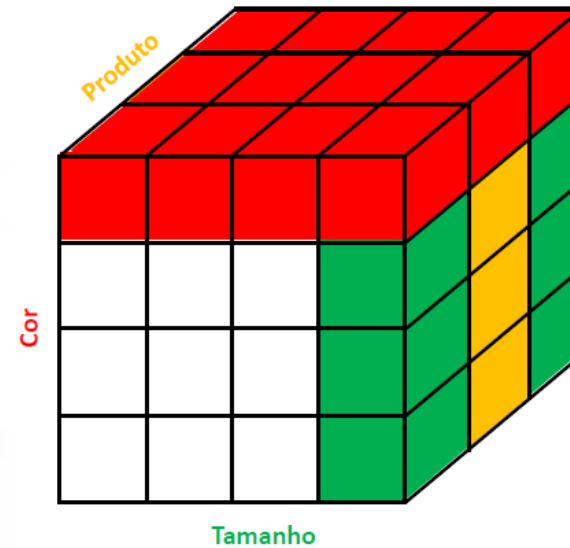
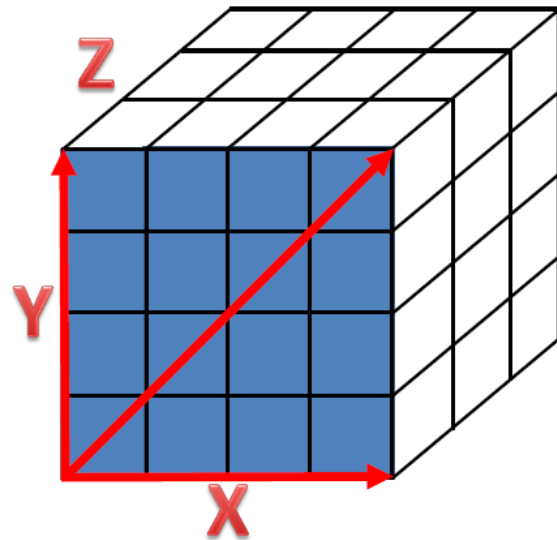
Processamento OLTP vs OLAP

- **OLTP** (On-Line Transaction Processing - Processamento On-Line de Transações) e está diretamente relacionada aos sistemas **TRANSACIONAIS** da empresa, tais como: Contas a Pagar, Contas a receber, Faturamento, Livros Fiscais, Contas-Correntes e etc...
- **OLAP** (On-Line Analytical Processing - Processamento Analítico On-Line) está relacionado à metodologia de acesso e busca às informações em ambientes de Data Warehouse através de ferramentas específicas para este fim e tem como um dos objetivos principais a geração de informações **GERENCIAIS**.

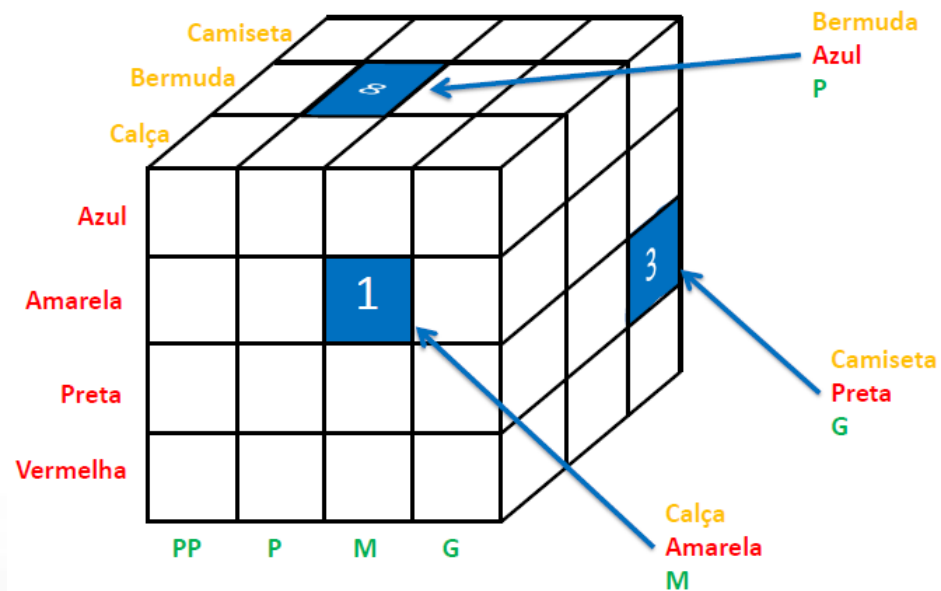


Funcionalidade de OLAP permite agregar dados consultados em cubos multidimensionais

Cubos Multidimensionais



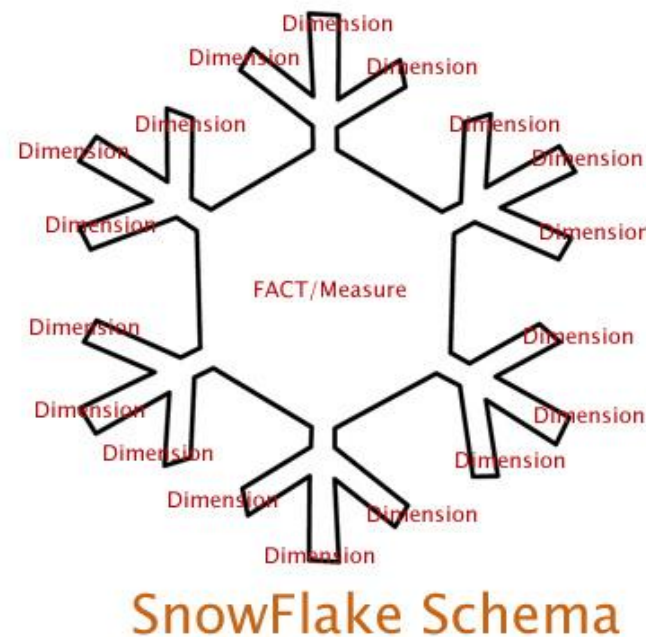
- Dimensões
- Fato



Cubos Multidimensionais - Topologias

No BI existem esquemas lógicos para a modelagem dos dados:

- Star Schema
- Snowflake Schema



Cubos Multidimensionais

✓ Visão dimensional

- Vendas por Linha
- Vendas por Linha e Vendedor
- Vendas por Linha, Vendedor e Produto
- Vendas por Linha, Vendedor, Produto e ano

✓ Agregação

- Vendas = SOMA dos itens vendidos

File View Tools Help

Opened ▾

admin ▾

Saiku Analytics x

Cubos

SteelWheelsSales ▾

Medidas Adicionar

Quantity
Sales

Dimensões

Customers
Markets
Order Status
Product
(All)
Line
Vendor
Product
Time
(All)
Years
Quarters
Months

Medidas

Sales

Colunas

Years Time

Linhas

Line Product
Vendor
Product

Filtros

Notice: You have 18 free logins remaining. [To find out more click here.](#)

Info: 11:31 / 6 x 111 / 0.12s

Line	Vendor	Product	Years	2003	2004	2005
Classic Cars	Autoart Studio Design	1968 Ford Mustang	Sales	62.140	67.155	23.974
		1958 Chevy Corvette Limited Edition		11.553	20.145	15.537
		1966 Shelby Cobra 427 S/C		19.093	19.272	10.243
	Carousel DieCast Legends	1982 Camaro Z28		35.028	53.640	15.612
		1949 Jaguar XK 120		30.263	36.802	16.573
		1952 Alpine Renault 1300		72.913	86.832	31.328
	Classic Metal Creations	1956 Porsche 356A Coupe		53.675	54.519	32.433
		1957 Corvette Convertible		43.517	71.055	22.543
		1961 Chevrolet Impala		21.293	44.201	17.895
	Exoto Designs	1965 Aston Martin DB5		47.005	43.386	16.460
		1952 Citroen-15CV		37.567	38.385	16.937
		1969 Chevrolet Camaro Z28		28.475	30.688	11.584
Gearbox Collectibles	1992 Porsche Cayenne Turbo Silver			34.793	54.520	12.842
				30.169	32.048	17.451
				54.683	50.502	23.133
	1976 Ford Gran Torino			40.163	64.615	23.879
				59.305	78.485	18.229

Saiku

Cubos Multidimensionais

Análise, não simplesmente relatório !!!!

Selecionar

Filtrar

Detalhar

Line	Classic Cars	Motorcycles	Planes	Ships	Trains	Trucks and Buses	Vintage Cars
Years	Sales	Sales	Sales	Sales	Sales	Sales	Sales
2003	1.514.407	397.220	347.755	244.821	72.802	420.430	679.949
2004	1.838.275	590.580	528.928	375.672	124.750	531.976	997.560
2005	738.738	286.325	200.074	128.178	36.917	201.875	388.718

Line		Classic Cars	Motorcycles	Planes	Ships	Trains	Trucks and Buses	Vintage Cars
Years	Months	Sales	Sales	Sales	Sales	Sales	Sales	Sales
2003	Jan	41.192	-	-	-	4.934	36.801	46.827
	Feb	20.464	25.784	39.205	27.050	4.330	-	24.002
	Mar	105.027	12.639	-	-	-	9.908	46.931
	Apr	59.874	23.476	36.563	27.399	4.756	28.791	20.750
	May	98.179	22.097	-	-	-	25.363	47.034
	Jun	50.257	2.642	34.592	29.045	10.071	17.370	26.583
	Jul	113.426	37.924	-	-	-	26.100	48.036
	Aug	48.407	44.165	33.938	25.607	7.375	15.032	23.285
	Sep	137.667	3.156	-	20.564	6.035	43.506	53.046
	Oct	241.145	68.277	84.713	35.980	10.234	46.912	102.703
	Nov	452.924	114.740	98.965	79.175	22.523	127.063	191.330
	Dec	145.846	42.320	19.778	-	2.545	43.584	49.421
2004	Jan	122.792	41.201	35.454	29.480	7.583	5.152	74.917
	Feb	137.641	49.067	37.660	30.612	4.968	33.710	25.042
	Mar	88.390	-	14.419	27.715	9.642	37.390	64.587

Saiku

ETL – Extract / Transform / Load

Para buscar dados e transformá-los em algo valioso que responda a uma pergunta concreta do negócio, é necessário, sobretudo, lidar com dois grandes problemas: a **coleta** e o **armazenamento**.

A **coleta** é um desafio, pois esses dados são gerados em formatos distintos, em tamanhos distintos e sem nenhuma estrutura organizada.

O **armazenamento** é outro desafio, pois é preciso colocar os dados em uma disposição que permita a inserção em alguma tecnologia similar a algum tipo de base de dados.

ETL – Extract / Transform / Load

✓ EXTRAÇÃO

- O processo de ETL se inicia obtendo os dados necessários para atender a um determinado problema, sendo que os dados podem estar localizados em uma ou mais fontes externas de dados ou, até mesmo, em diferentes formatos de dados.
- Como exemplo das fontes de dados, temos os bancos de dados relacionais, bancos de dados NoSQL, Planilhas Excel, arquivos de texto, entre outros.
- Essa é uma das etapas mais importantes já que é ela que vai definir se é possível ou não extrair todas as informações relevantes para serem estudados. Além disso, os dados precisam ser extraídos diversas vezes e de forma periódica para que o arquivo esteja sempre atualizado.

ETL – Extract / Transform / Load

✓ TRANSFORMAÇÃO

- Assim que a etapa de extração é concluída inicia a fase da transformação, preparação e adaptação dos dados extraídos. É nessa etapa que um conjunto de regras são inferidas nos dados com o objetivo de transformá-los no formato adequado para poderem ser carregados no banco de dados de destino.
- A transformação não consiste apenas no mapeamento de colunas e tabelas para o destino correto, mas também são feitas as modificações necessárias conforme as restrições de integridade propostas.
- Lógica de execução para tratar dados “ruins”.
 - padronizar nomes;
 - combinar e eliminar dados duplicados;
 - normalizar cálculos;
 - correções dos dados;
 - remover colunas ou linhas desnecessárias;
 - corrigir erros de digitação.

ETL – Extract / Transform / Load

✓ CARGA

- A etapa de carregamento dos dados se inicia após a finalização da transformação dos dados.
- Nesse momento, os dados são carregados em uma estrutura, como o DW, para que sirvam ao propósito de análise.
- É importante que a frequência com que os dados serão extraídos e armazenados esteja bem definida, permitindo assim dimensionar corretamente o repositório, garantindo o desempenho adequado para realizar sua função.
- Com os dados armazenados no banco de destino, pode-se dar início a leitura e análise conforme a necessidade dos cientistas de dados que poderão manipular e gerar relatórios precisos.



Suíte Pentaho

Sobre a Pentaho









- Pentaho Corporation é uma empresa fundada em 2004 sendo uma iniciativa pioneira da comunidade de desenvolvimento Open Source para proporcionar ferramentas de Business Intelligence (BI) para que as organizações melhorem sua performance, eficiência e efetividade na gestão da informação.
- O Pentaho surgiu com o desejo de alcançar uma mudança positiva no mercado de análise de negócio dominada por grandes vendedores que ofereciam produtos baseados em plataformas com custo elevado. A partir daí, cinco experientes desenvolvedores de sistemas (daí a origem do prefixo Penta) fundaram o Pentaho.

Sobre a Pentaho

- Reconhecida como líder da classe open source em BI e Integração de Dados
- Média de um download a cada 30 segundos
- Dos downloads: 75% Windows, 11% Linux e 9% MAC
- Mais de 1.500 clientes em 65 países
- Mais de 12.000 ambientes de produção
- Prêmios:



Principaux Clients



Clientes Brasil



CAIXA



Telefonica

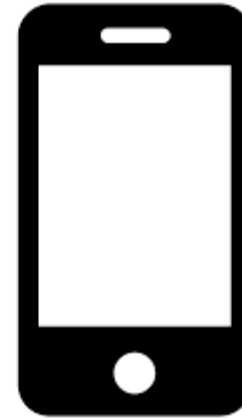


EXÉRCITO BRASILEIRO



Pentaho Business Analytics

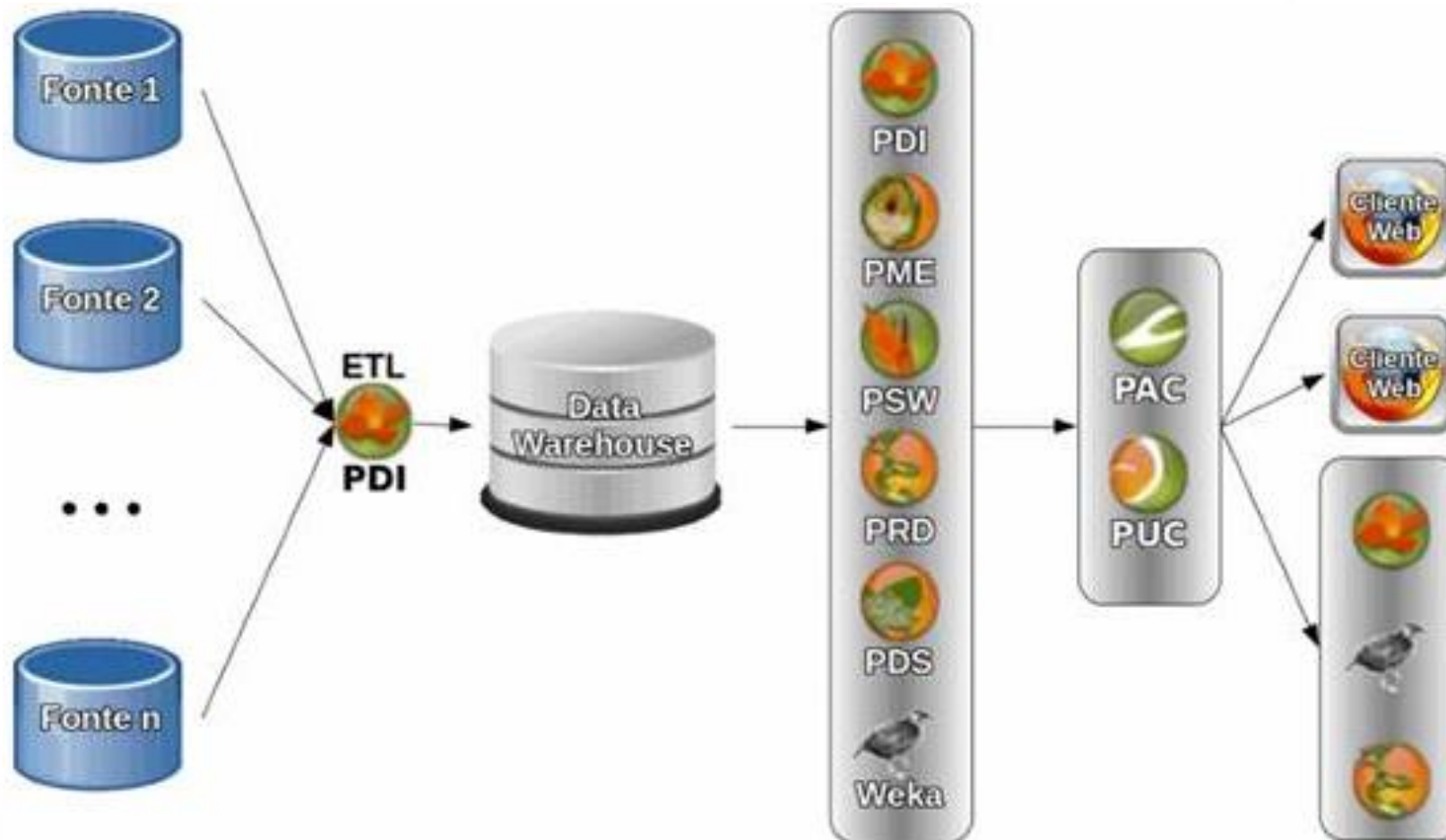
Suporte Multiplataforma:



Plataforma Pentaho

- ✓ Pentaho – BI Server (Servidor de Exploração de Dados)
- ✓ **Pentaho Data Integration (PDI)** (Servidor de ETL)
- ✓ Pentaho Schema Workbench (PWS) (Modelo lógico multidimensional)
- ✓ Pentaho Report Designer (PRD)
- ✓ Painéis (Dashboards) (CTools)
- ✓ Mineração de Dados (Weka)

Arquitetura Pentaho



Processo de criação de Solução de BI Padrão com Pentaho



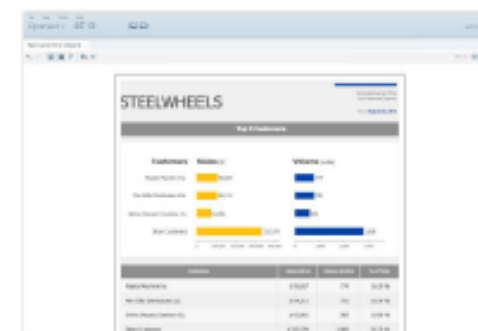
Dashboards

Visualizações, Indicadores, Métricas e Self-services



Reporting

Relatórios Operacionais e Interativos



Pentaho BI Suíte



Analysis

Análises Self-services e Interativas



Data Integration

Integração e limpeza de dados com alto desempenho



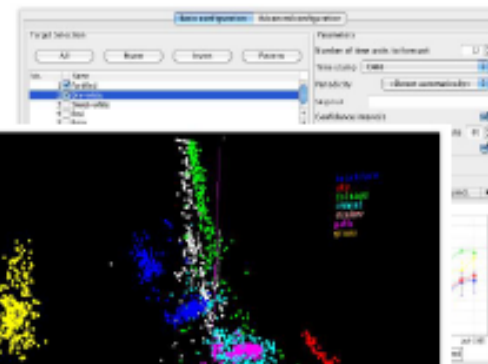
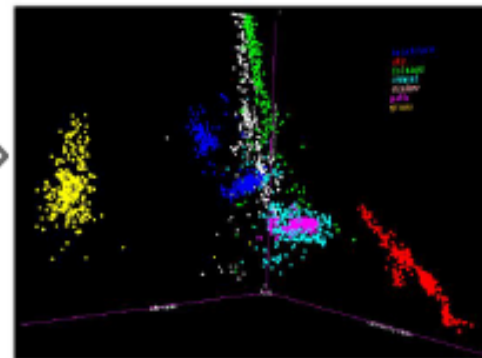


Data mining pode identificar relacionamento entre seus dados



Data Mining

Análises preditivas
avancadas



Data Mining é usado para encontrar correlações e padrões, de dados, dentro da base

Pentaho BI Suíte

DASHBOARDS

RELATÓRIOS

ANALYSIS

DATA INTEGRATION

DATA MINING



Usuários de Negócio

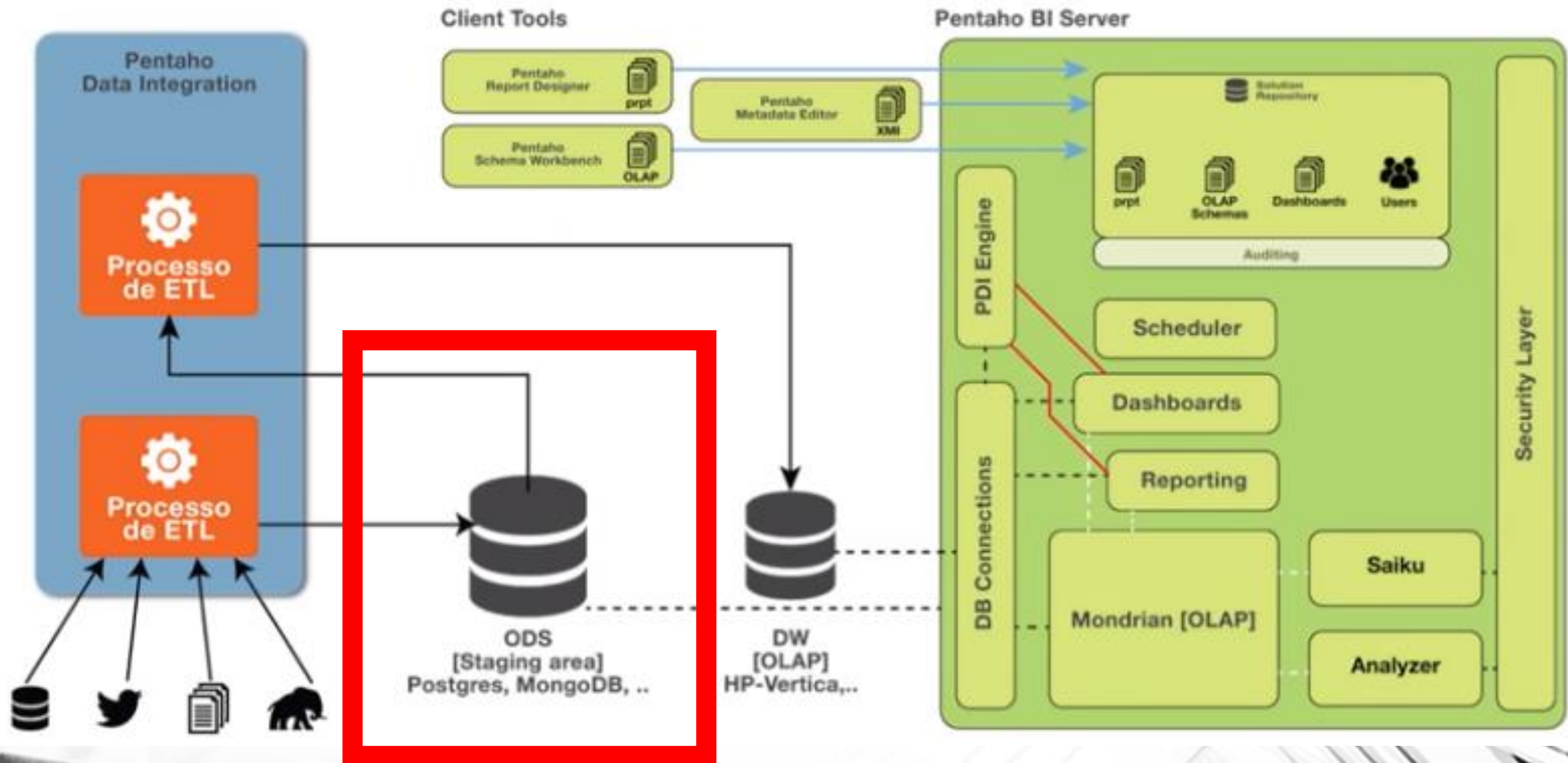


Desenvolvedores e DBAs



Usuários Avançados

Arquitetura Pentaho - PDI



Pentaho Data Integration (PDI)

O PDI é bastante fácil e intuitivo de usar. Todos os processos são criados através de ferramenta gráfica sem escrever “nenhuma” linha de código. A lógica do processo é realizada através da composição dos itens da ferramenta.

Características do PDI:

- Pode ser usado de forma independente ou como parte da suíte do Pentaho.
- É a ferramenta open source disponível mais conhecida.
- Suporta um grande conjunto de formatos de entrada e saída de dados, incluindo arquivos texto, arquivos .xls (Excel), arquivos .mdb (access), além de vários SGBD's.
- Orientado a fluxo de dados.

Pentaho Data Integration (PDI)

- Migração de dados de um servidor para outro
- Tratamento dos dados
- Limpeza dos dados
- Criação de métricas e indicadores
- Exportar banco de dados para outros formatos
- Consumo de Web Services e API (por exemplo, API Trello)
- Integração com diversos banco de dados (mais de 20)
- Integração com várias plataformas de Big Data (Hadoop, Spark, Python, R,...)

Pentaho Data Integration (PDI)

✓ TRANSFORMAÇÃO

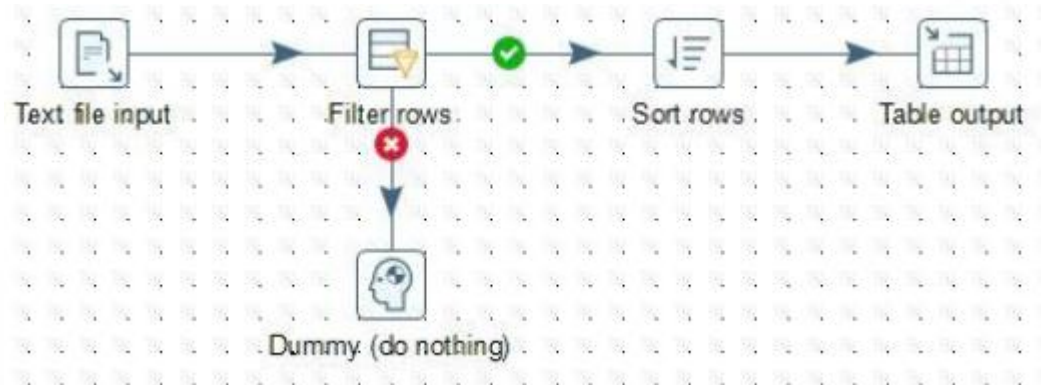
- Uma transformação representa o movimento dos dados que estão sendo trabalhados.
- Uma transformação representa uma coleção de passos (*steps*)
 - Cada passo executa uma tarefa ou operação específica em uma coleção de dados ou em um simples registro.
- Os *steps* se ligam entre si através de *hops* que guiam os dados passo a passo.



Pentaho Data Integration (PDI)

✓ TRANSFORMAÇÕES

- Transformações são uma rede de tarefas lógicas (*steps*):
 - Ler um arquivos,
 - Filtrá-lo,
 - Ordená-lo,
 - Carregá-lo em uma tabela PostgreSQL, por exemplo.



✓ JOBS

- Processo que determina o fluxo de execução de transformações ou de outros jobs.
 - Definir o fluxo de dependências e em qual ordem as transformações devem ser executadas.
- Agendados pelo Sistema Operacional para executar os fluxos de informações periódicos.
- Preparar para execução, avaliando as condições como: o meu arquivo fontes está disponível? ou a tabela X existe?
- Gerenciar arquivos, como enviar ou baixar arquivos usando FTP e copiar ou excluir arquivos.
- Permite automação de notificações de resultados (de sucesso ou de falha) por e-mail ou alertas.

Pentaho Data Integration (PDI)

✓ Componentes



Agendada e automática

Agendada e automática

Web Service para execução remota das transformações e Jobs

Pentaho Data Integration (PDI)

MÃOS À OBRA



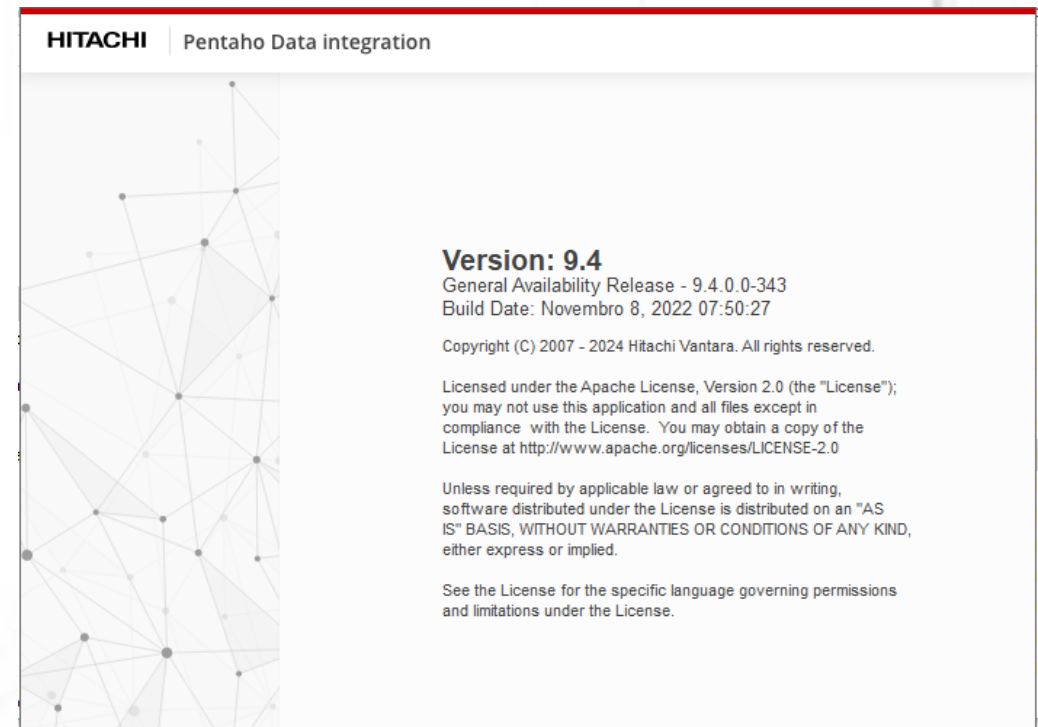
Pentaho Data Integration (PDI)

✓ Preparando o ambiente

- PDI – download
 - <https://www.hitachivantara.com/en-us/products/pentaho-plus-platform/data-integration-analytics/pentaho-community-edition.html>
 - O PDI não requer instalação. Basta descompactar o arquivo .zip dentro de uma pasta qualquer.

Sugestão: drive:\ **pentahoce** \data-integration

drive:\ **pentahoce** data-integration \ **spoon.bat**



Pentaho Data Integration (PDI)

- ✓ **Preparando o ambiente**
 - Java
 - <https://www.oracle.com/br/java/technologies/downloads/#jdk21-windows>
 - PostgreSQL
 - <https://www.postgresql.org/download/>