

# Business Analytics

# About me

- Empreendedor
- Administração / Sistemas de Informação
- Mestre (UFPE/CIN)
- Data Scientist/Data Engineer
- Consultor BA a 9 anos
- Arquiteto Pentaho

Mailson Santos

# Sobre a turma!



# Agenda

## Semana 1

- O que é Business Analytics
- Extract Transform Load
- Modelagem Dimensional
- Online Analytical Processing
- Pentaho
  - Data Integration
  - Mondrian

## Semana 2

- Analytics e CRISP-DM
- Predictive Analytics
- Mining
  - R + PDI
- Pentaho
  - Pentaho Server
- Projeto

# O que é Business Analytics?



# O que é Business Analytics?

O **Business Intelligence**, ou BI, é uma técnica para **auxiliar o gestor no planejamento estratégico**. Ele é uma forma de coleta e análise de um conjunto amplo de dados de uma empresa para entender a sua performance e, a partir daí, planejar o futuro de forma mais eficiente.

Permite identificar os acertos e aquilo que não deu muito certo para corroborar as próximas decisões. Os dados no **Business Intelligence** são disponibilizados em métricas estabelecidas e planilhas relativamente complexas. É aqui que o Business Analytics, ou BA, ganha espaço.

Como uma evolução do Business Intelligence, o Business Analytics chega para facilitar esse processo para ajudar a decodificar informações e auxiliar na análise de dados de forma ainda mais eficiente para a tomada de decisões operacionais precisas.

Ambos pregam o uso da informação para melhorar o desempenho, mas o Business Analytics considera que hoje a **quantidade de informações produzida é cada vez maior e mais complexa**, trazendo à tona a necessidade de se desenvolver uma metodologia coerente com a realidade atual.

O BA, com seu enfoque mais contemporâneo, não está ganhando tanto espaço à toa. A **análise de dados** aqui é mais eficiente, vai mais fundo e permite uma compreensão dos dados que vai além dos fatos concretos.

Além de mostrar o que aconteceu, como aconteceu e quando aconteceu, o Business Analytics ajuda a responder questões relativas às razões pelas quais determinados acontecimentos ocorreram.

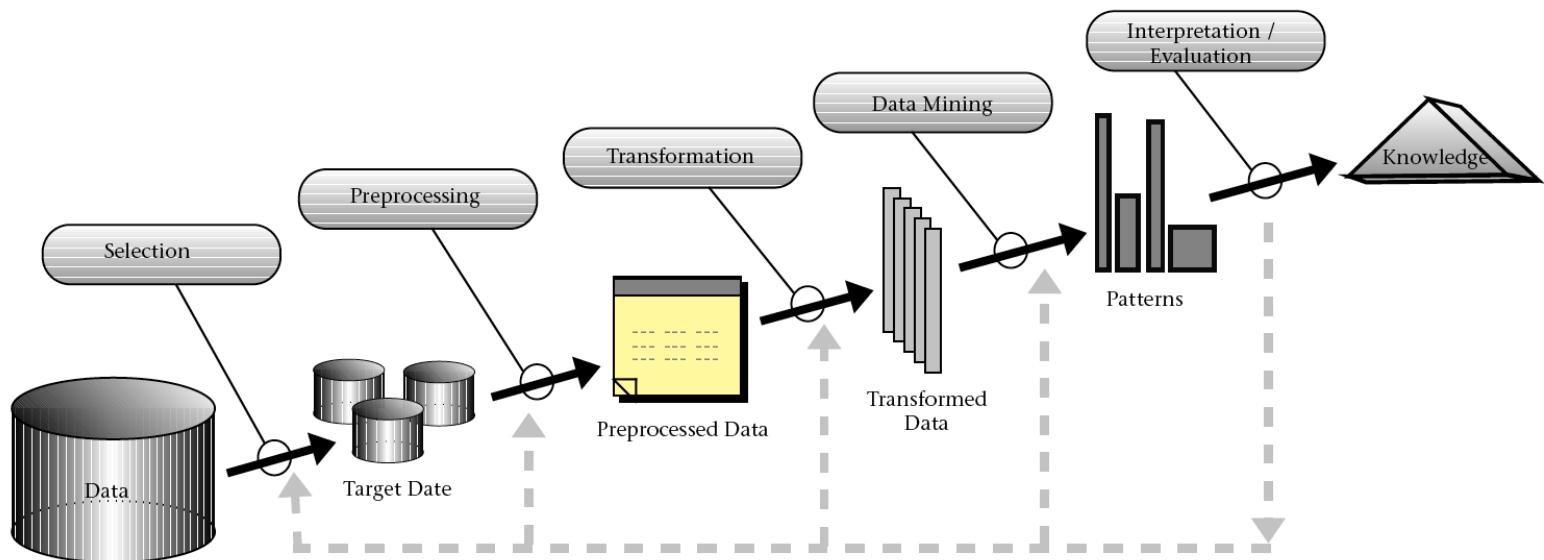
Assim como o BI, o BA faz uso de **tecnologia e da estatística** para a tradução das informações, mas permite uma investigação mais aprofundada e contínua do negócio.

# O que é Business Analytics?



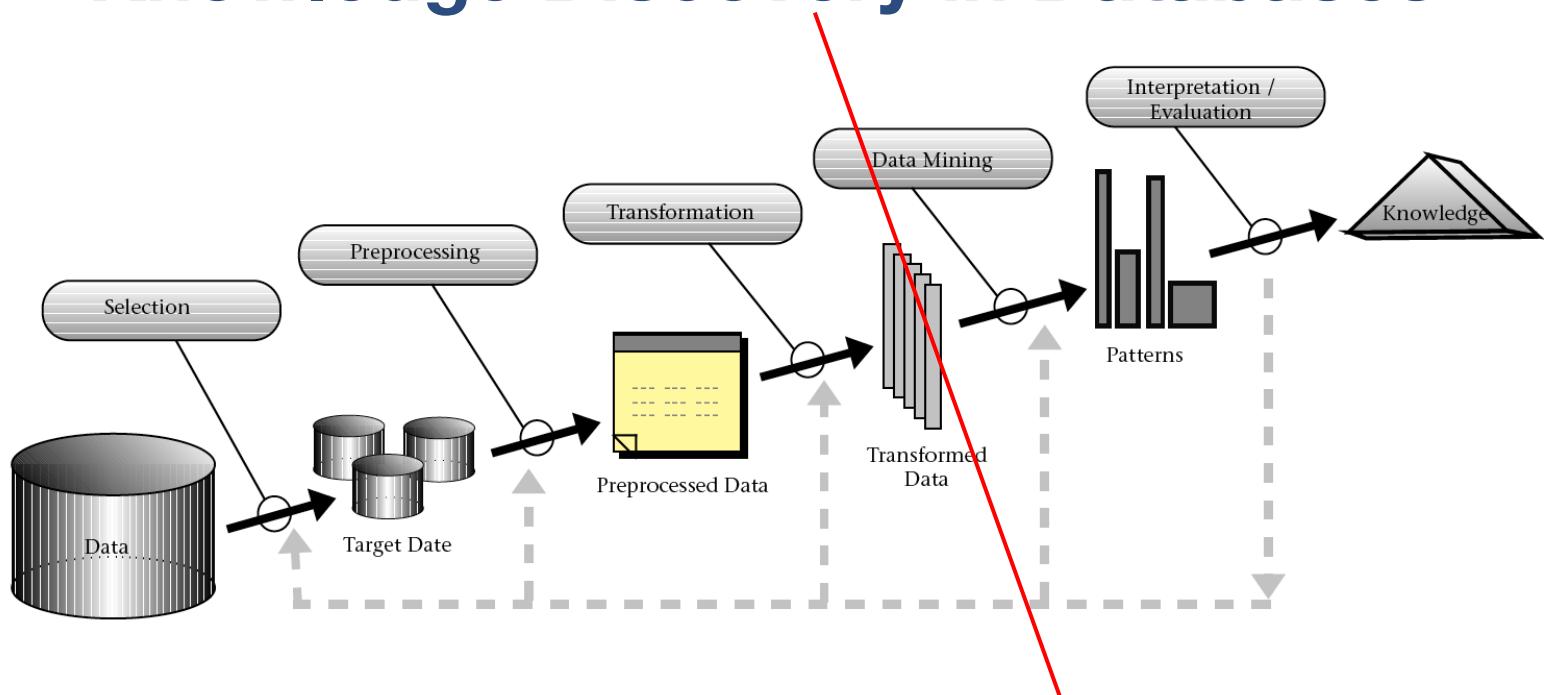
# O que é Business Analytics?

## Knowledge Discovery in Databases



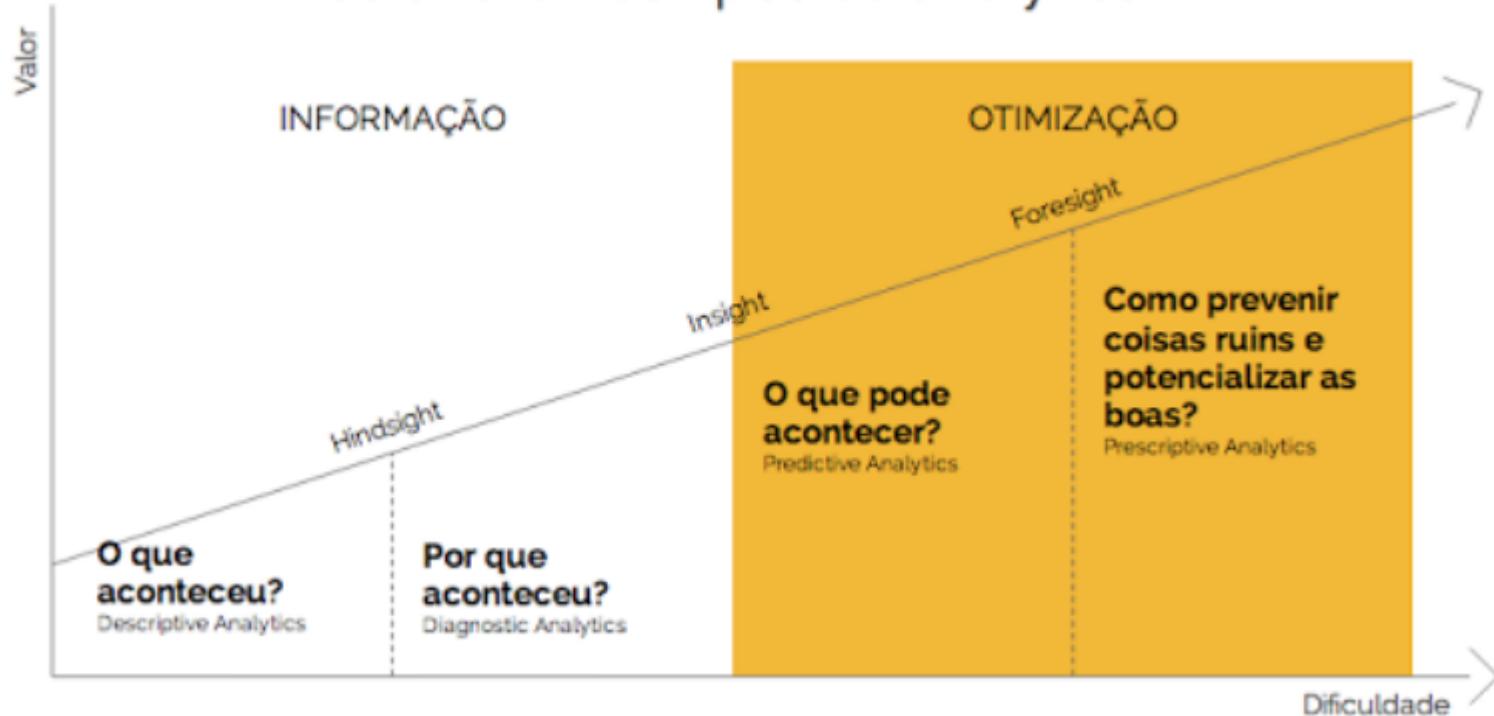
# O que é Business Analytics?

## Knowledge Discovery in Databases

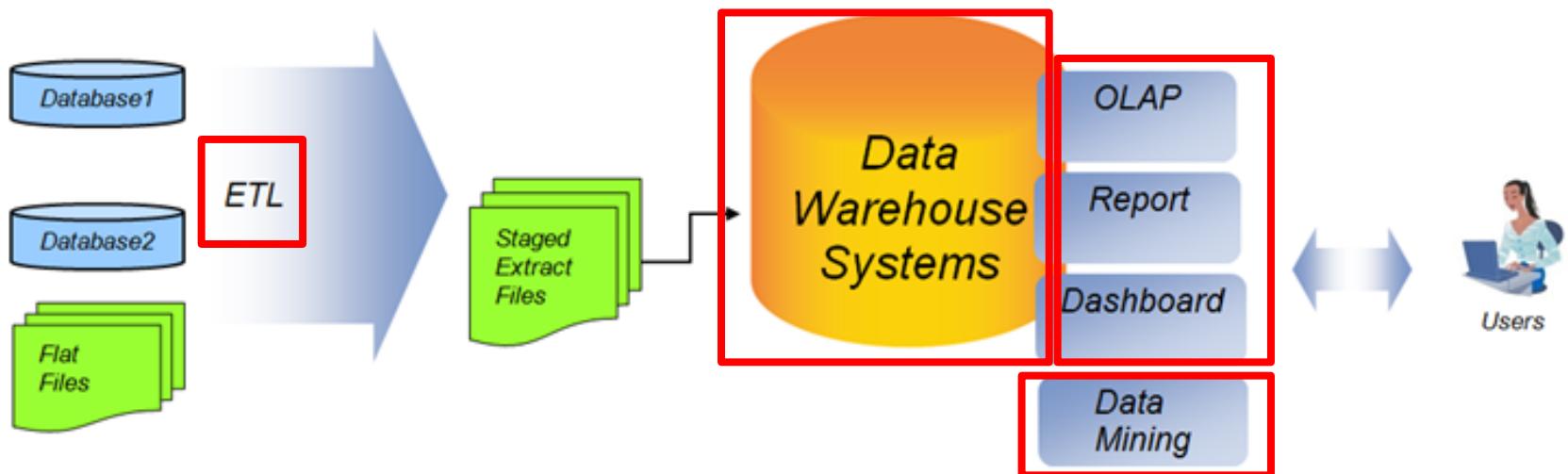


# O que é Business Analytics?

os diferentes tipos de analytics



# Camada de Informação



# Extract Transform Load

## 1. Extração

**Fontes...**

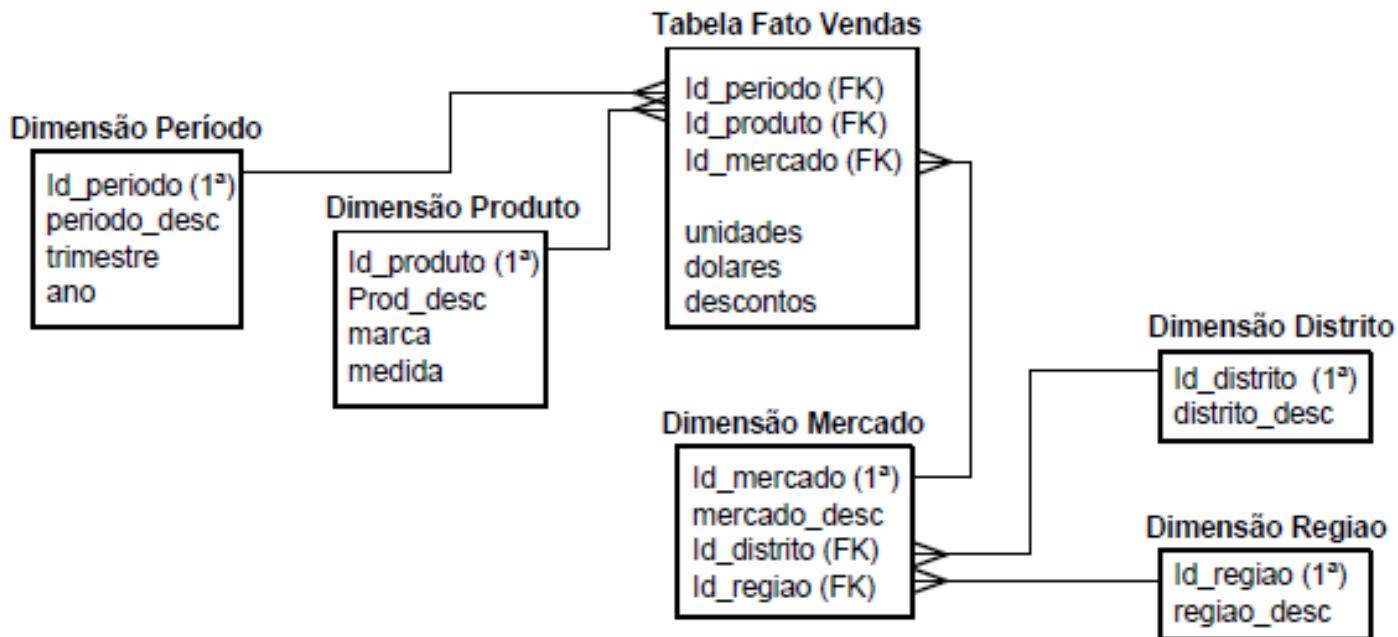
## 1. Transformação

**Quais...**

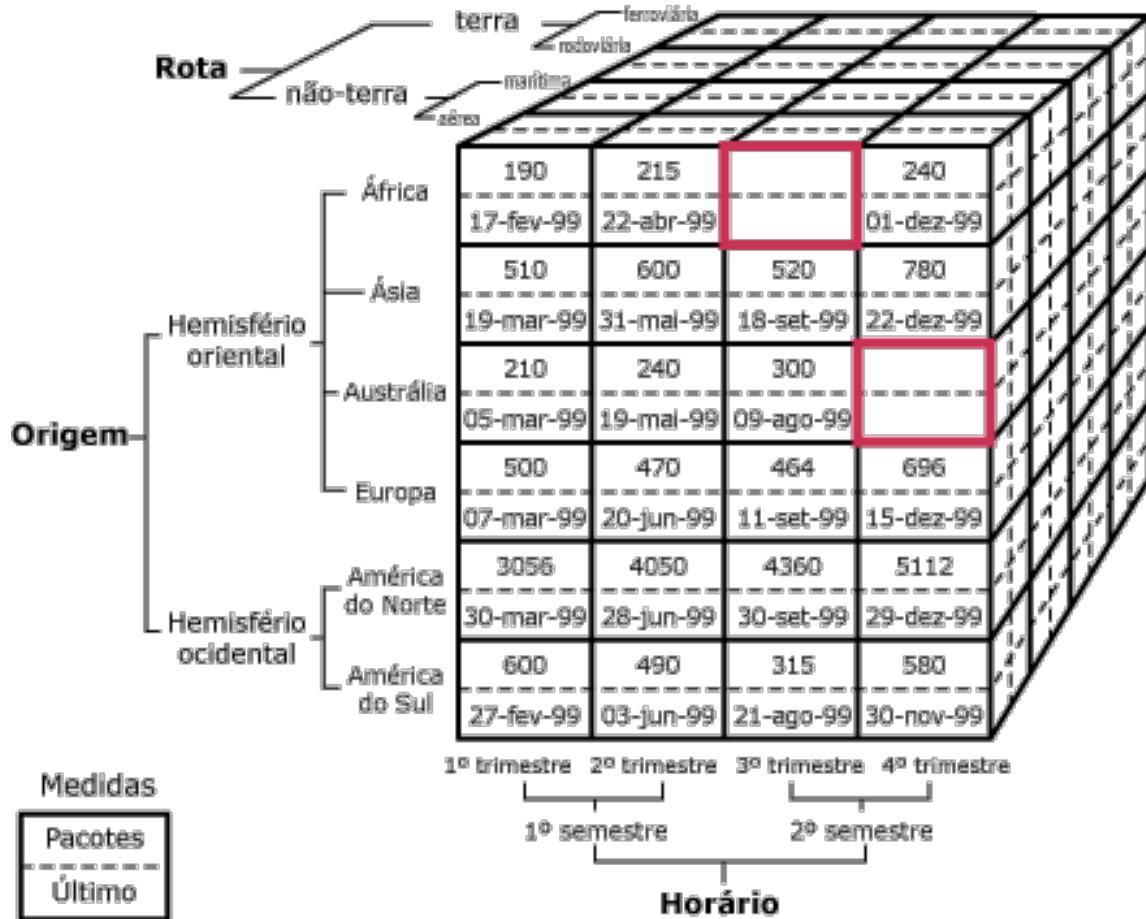
## 1. Carga

**Onde...**

# Data Warehouse



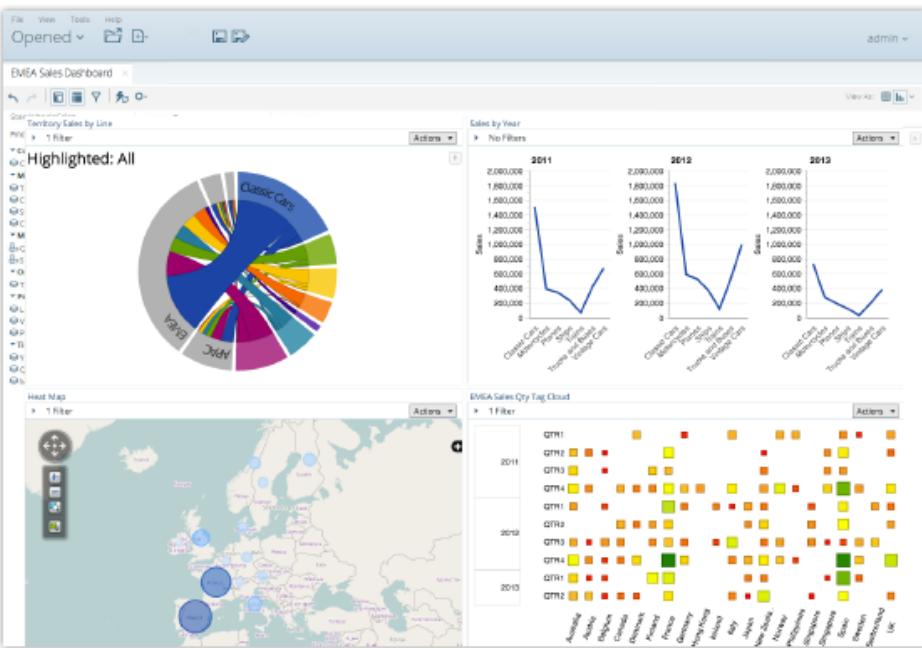
# OLAP



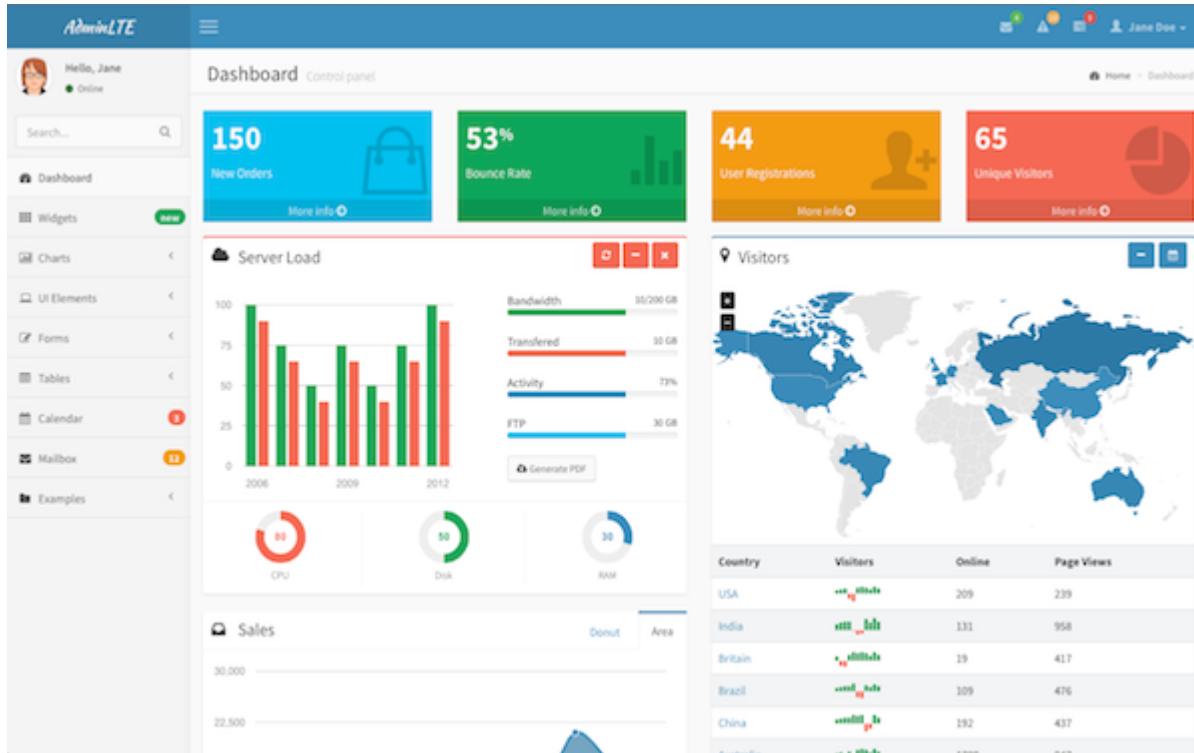
# OLAP

			Year ▾		Quarter	Month			
			1996		1997		1998		
							Quarter 1		Quarter 2
Category Name ▾	Product Name	Ship Country ▾	Ship Region	Ship City	Unit Price	Quantity	Unit Price	Quantity	Unit Price
Beverages					2033,8	1842	5058,35	3996	3622
Condiments					705,6	962	2211,85	2895	1037,4
Confections					1333,4	1357	3547,63	4137	2059,63
Dairy Products					1857,6	2086	5070	4374	1632,1
Grains/Cereals	Filo Mix				16,8	48	102,2	313	49
	Gnocchi di nonna Alice				212,8	96	1178	971	304
	Gustaf's Knackebroð				16,8	6	184,8	209	63
	Ravioli Angelo				93,6	133	148,2	124	136,5
	Singaporean Hokkien Fried Mee				54,6	37	215,6	451	56
	Tunnbröd				21,6	105	82,8	287	45
	Wimmers gute Semmelknödel				106,4	124	425,6	281	266
	Total				522,6	549	2337,2	2636	919,5
Meat/Poultry					1294,3	950	3795,48	2189	1473,09
Produce	Longlife Tofu				40	141	54	116	10
	Manjimup Dried Apples	Argentina					53	7	
							53	120	
			Brazil	RJ	84,8	37	53	28	
			SP	Campinas			53	28	
			Total		84,8	37	53	28	
			Canada				53	30	
			Denmark				53	30	
			France		42,4	20	106	35	106
			Germany		42,4	40	106	35	7
			India		42,4	40	106	35	7

# Visualizações



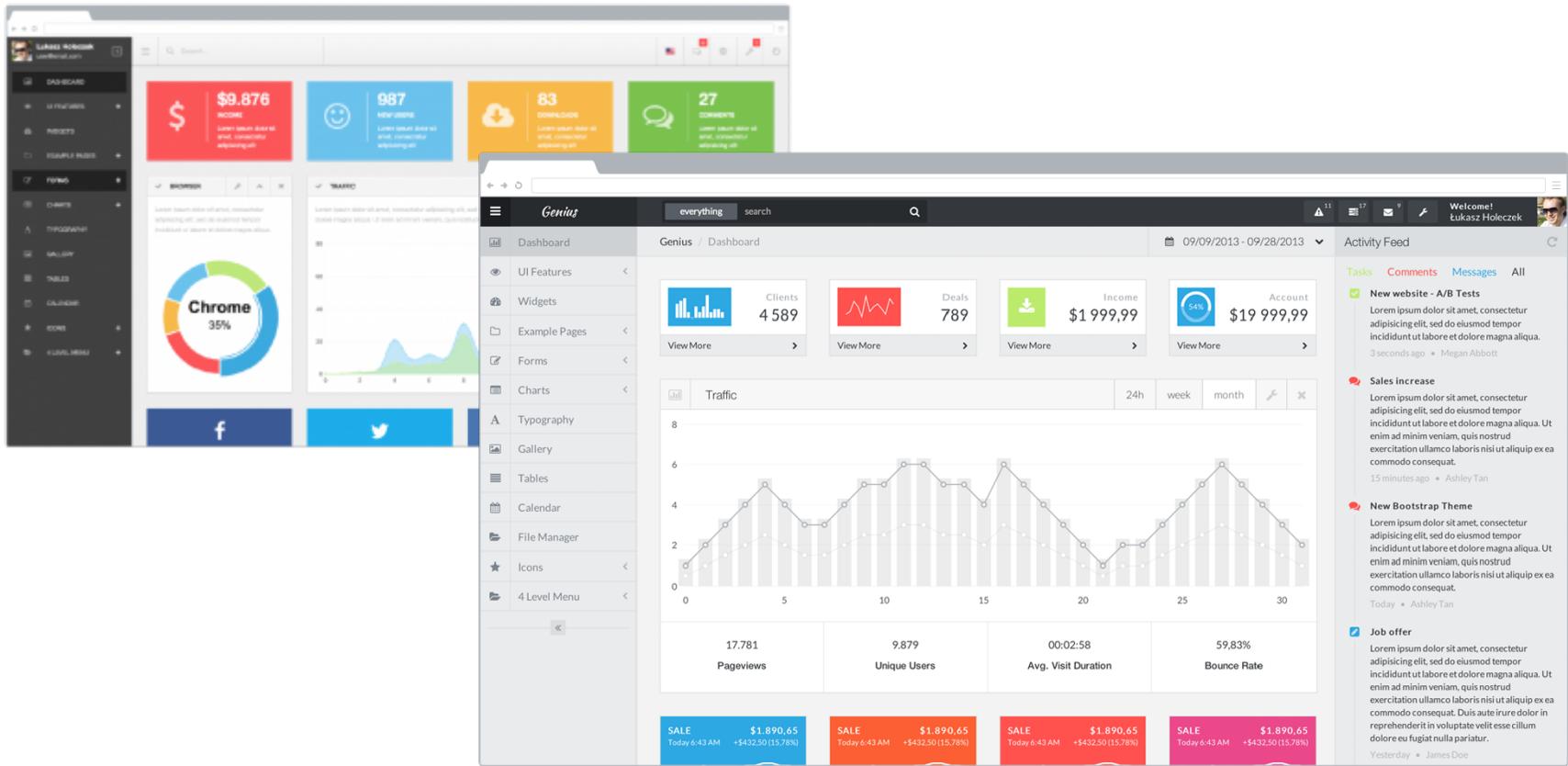
# Visualizações



# Visualizações

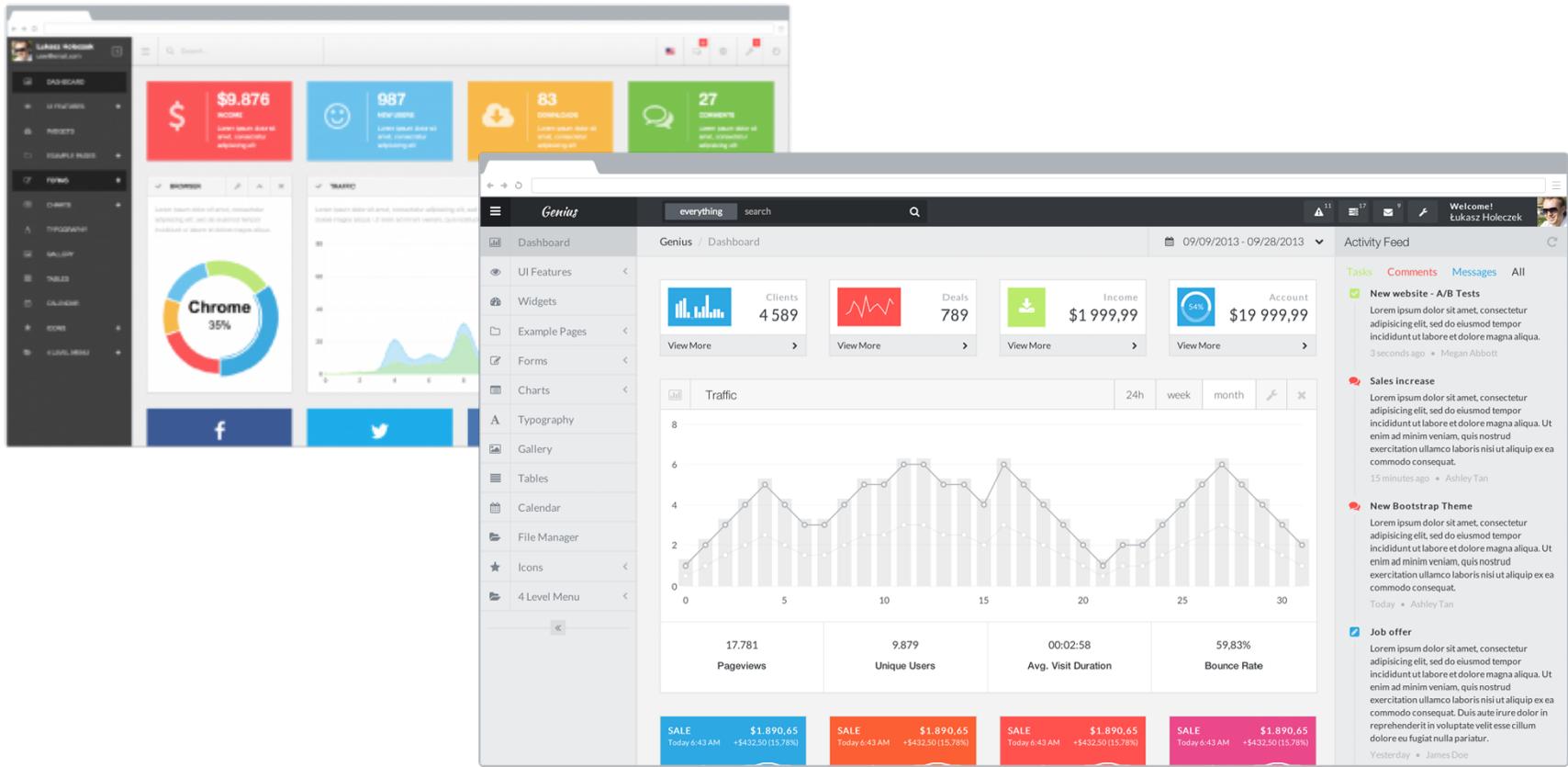


# Visualizações



The image displays two examples of dashboard visualization tools. The left side shows a dashboard builder interface with a sidebar containing navigation links such as 'Dashboard', 'UI Features', 'Widgets', 'Example Pages', 'Forms', 'Charts', 'Typography', 'Gallery', 'Tables', 'Calendar', 'Icons', and '4 Level Menu'. The main area features several cards: a red card with a dollar sign icon and '\$9.876 INCOME', a blue card with a smiley face icon and '987 NEW USERS', an orange card with a cloud icon and '83 DOWNLOADS', and a green card with a speech bubble icon and '27 COMMENTS'. Below these are social sharing buttons for Facebook and Twitter. The right side shows a live dashboard titled 'Genius' with a 'Dashboard' section. It includes four cards: 'Clients' (4 589), 'Deals' (789), 'Income' (\$1 999,99), and 'Account' (\$19 999,99). Below these are sections for 'Traffic' (a line chart showing traffic over 30 days) and summary statistics: 'Pageviews' (17.781), 'Unique Users' (9.879), 'Avg. Visit Duration' (00:02:58), and 'Bounce Rate' (59.83%). At the bottom, there are four 'SALE' cards with details: 'Today 6:43 AM +\$43250 (15.78%)', 'Today 6:43 AM +\$43250 (15.78%)', 'Today 6:43 AM +\$43250 (15.78%)', and 'Today 6:43 AM +\$43250 (15.78%)'. The right side also features an 'Activity Feed' section with posts from users like Lukasz Holeczek, Megan Abbott, Ashley Tan, and James Doe.

# Visualizações



The image displays two examples of dashboard visualization tools:

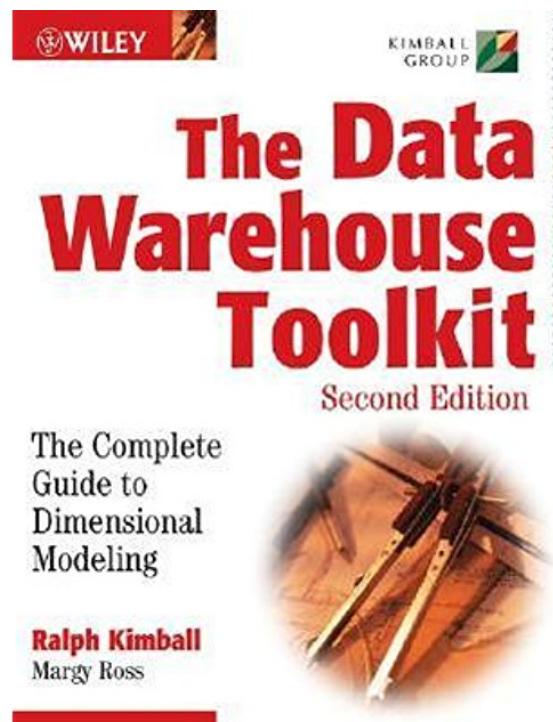
**Left Dashboard:**

- UI Features:** Shows a 3D pie chart with segments for Chrome (35%), Firefox (30%), and others.
- Widgets:** Shows a bar chart with a blue gradient.
- Example Pages:** Shows a line graph with a blue gradient.
- Forms:** Shows a form with fields for Name, Email, and Message.
- Charts:** Shows a line graph with a blue gradient.
- Typography:** Shows a chart with a blue gradient.
- Gallery:** Shows a grid of images.
- Tables:** Shows a table with columns for Product, Sales, and Profit.
- Calendar:** Shows a calendar with events.
- Icons:** Shows a collection of icons.
- 4 Level Menu:** Shows a multi-level navigation menu.

**Right Dashboard (Genius):**

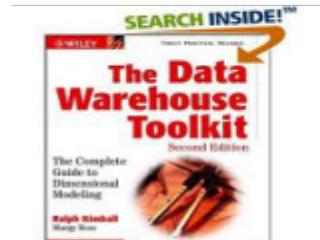
- Dashboard:** Shows four cards: Clients (4 589), Deals (789), Income (\$1 999,99), and Account (\$19 999,99).
- Traffic:** A line chart showing traffic over 30 days.
- Metrics:** Shows Pageviews (17.781), Unique Users (9.879), Avg. Visit Duration (00:02:58), and Bounce Rate (59.83%).
- Sales:** Four cards for Sales: Today 6:43 AM (\$1.890,65), Today 6:43 AM (\$1.890,65), Today 6:43 AM (\$1.890,65), and Today 6:43 AM (\$1.890,65).
- Activity Feed:** Shows posts from users like Megan Abbott and Ashley Tan.

# Data Warehouse



# Data Warehouse - Histórico

- Em 1996 Ralph Kimball lançou o livro The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling
- Melhor prática de modelagem de banco para construção OLAP;
- Técnica de modelagem **NÃO** e não uma implementação de banco;
- Pode existir com tabelas padrões de banco de dados;
- Otimizadas para agregações dinâmicas e massivas.



# Objetivos de um Data Warehouse

- "Dispomos de um enorme volume de dados na empresa, mas não conseguimos acessá-los".
- "Temos que combinar os dados a partir de fontes diversas".
- "Apenas mostra-me o que é realmente importante".
- "Fico louco quando duas pessoas apresentam exatamente os mesmos indicadores em uma reunião usando apenas números diferentes".
- "Queremos que as pessoas usem informações para dar um suporte a tomadas de decisões baseando-se em fatos".

# Objetivos de um Data Warehouse

O data warehouse deve fazer com que informações de uma empresa possam ser facilmente acessadas.

- Compreensível.
- Intuitiva e óbvia para usuários da área de negócio.
- Com separação e combinação dos dados de forma “infinita”.
- Ferramentas para acessar o data warehouse devem ser simples e fáceis de serem usadas.

Retornar os resultados no menor tempo de resposta possível.

# Objetivos de um Data Warehouse

O data warehouse deve apresentar as informações da empresa de modo consistente.

- Confiáveis.
- Obtidos cuidadosamente a partir de várias fontes na empresa, filtrados, submetidos a um controle de qualidade e liberados apenas quando estiverem prontos para serem usados.
- As informações de um processo de negócio deve coincidir com as informações de outro processo de negócio.

# Objetivos de um Data Warehouse

O data warehouse deve ser adaptável e flexível a mudanças.

- “As três certezas da vida são: Impostos, a Morte e que o usuário irá mudar os requisitos assim que tiver a primeira tela em mãos”.
- As necessidades dos usuários, as condições comerciais, os dados e a tecnologia, todos estão sujeitos às mudanças decorrentes do tempo.
- O Data Warehouse deve ser construído para que mudanças não tenham um grande impacto.

# Objetivos de um Data Warehouse

O data warehouse deve funcionar como a base para uma melhor tomada de decisões.

- O data warehouse deve conter os dados apropriados para dar suporte à tomada de decisões.
- O conceito original era usado para definir o data warehouse continua sendo: um sistema de suporte à tomada de decisões.

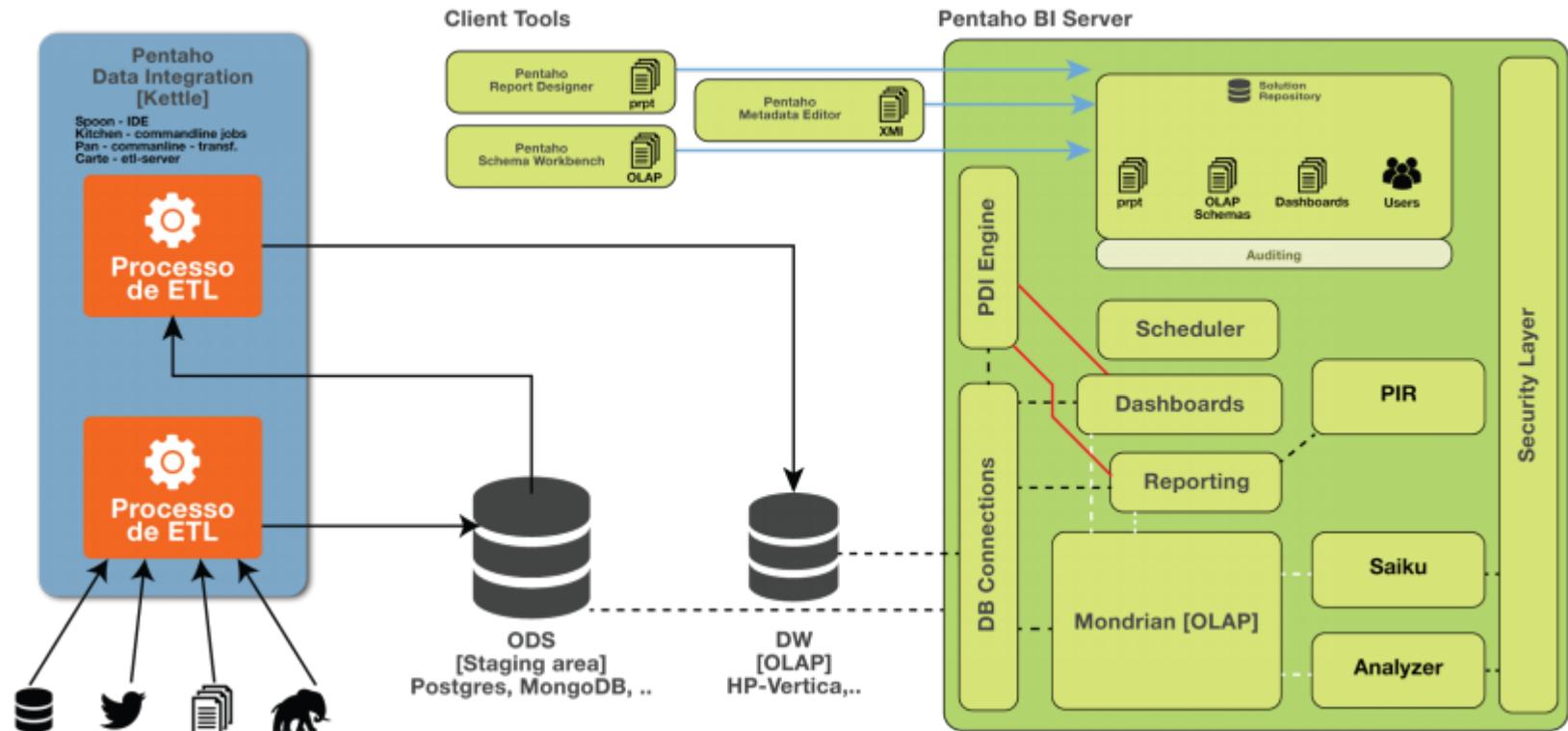
# Objetivos de um Data Warehouse

O data warehouse deve ser uma fortaleza que protege as informações.

- O data warehouse guarda informações críticas e confidenciais da empresa.
- O vazamento dessas informações podem causar verdadeiros desastres.

# Componentes de um Data Warehouse

# Componentes de um Data Warehouse



# Staging Área

- Também chamado de ODS (Operational Data Store).
- É uma base de dados com o objetivo de integrar várias fontes de dados para futuras operações adicionais sobre os dados.
- Dados Brutos.
- Liberação de Recursos das Fontes de Origem.

# Terminologia da Modelagem Dimensional

**Medidas:** Valores que se deseja medir no negócio.

- Incluem uma variedade de indicadores chaves de performance.
- Podem ser apresentadas em vários níveis de sumarização e perfurações, dependendo como as dimensões de análises são apresentadas.
- Também são chamadas de **fatos**.

# Terminologia da Modelagem Dimensional

**Dimensão:** Representa as variáveis pela qual a medição é realizada; tais como data, local, produto, cliente.

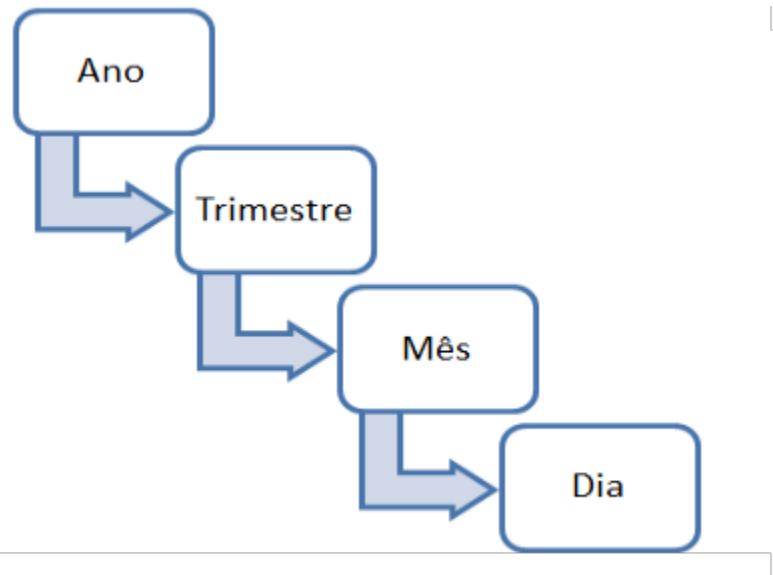
- Dimensões podem ser organizadas em hierarquias, permitindo ao usuário realizar perfuração através dos dados;

Por exemplo: As dimensões ‘Geografia’ podem conter hierarquias que iniciam com País e são perfuradas para Estado e dessa para Cidade;

- Um cuidadoso desenho de hierarquias numa dimensão facilitam para o usuário o detalhamento da informação projetando as hierarquias para serem intuitivas e acompanhando o processo de pensamento do analista.

# Terminologia da Modelagem Dimensional

**Hierarquias:** São o caminho que será percorrido na dimensão entre os níveis. São “escadas” e os níveis os “degraus” de uma dimensão.



# O que é OLAP?

Definição:

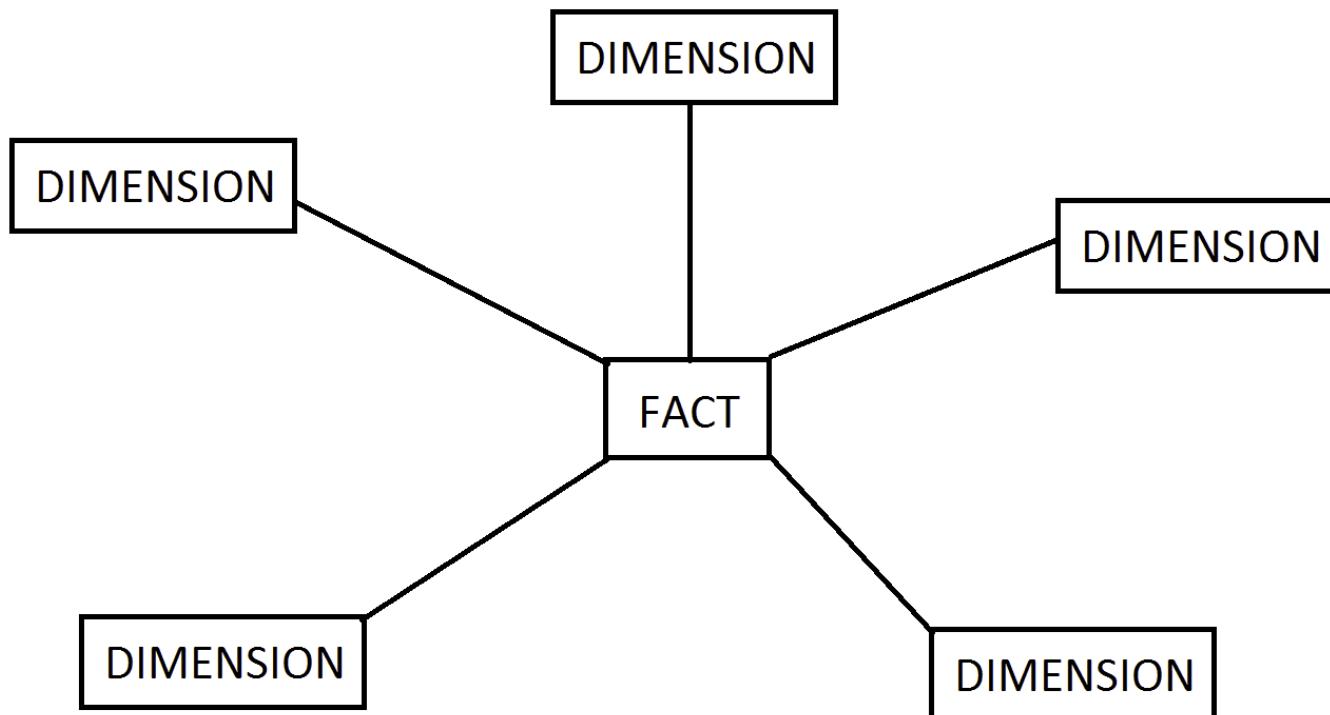
“OLAP é uma abordagem para prover respostas a perguntas de natureza multidimensional”.

- OLAP permite aos usuários realizar cortes e perfurações dos dados da forma que quiser.
- Também pode ser definida como a capacidade de manipular e analisar dados de múltiplas perspectivas.
- Um cubo OLAP (online analytical processing) é uma estrutura de dados que permite rápida análise de dados.

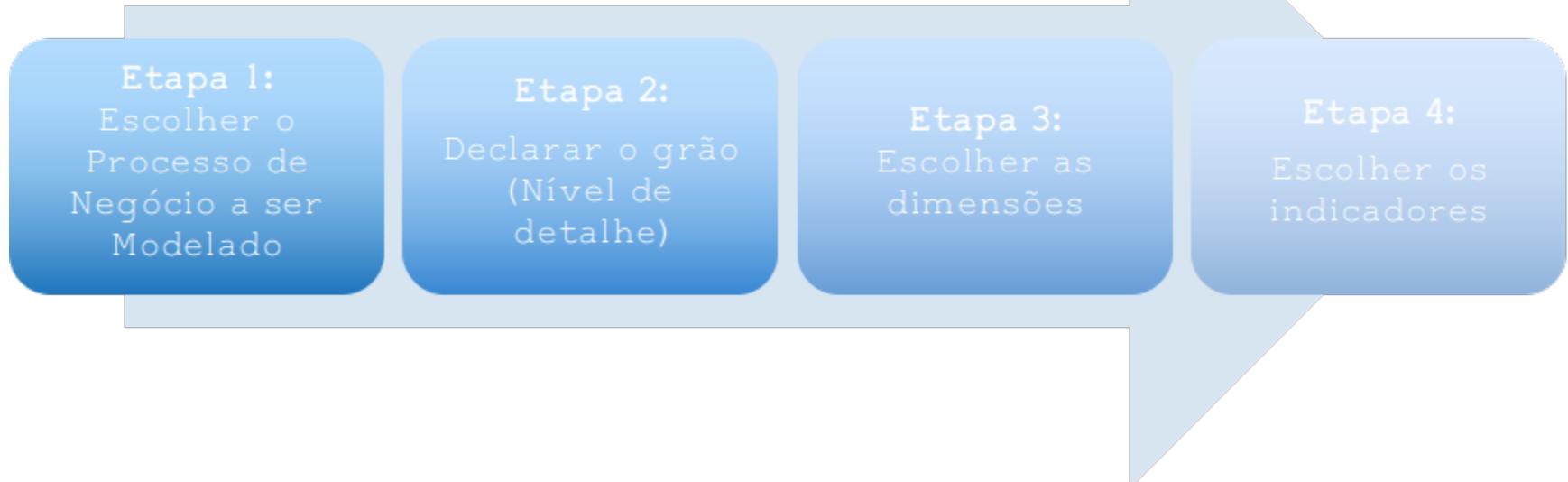
# Recursos de Soluções Analíticas

- **Foco em informações:** Projetado para investigação usuário final e exploração de dados, e não transacional.
- **Interativo:** Capaz de aceitar e agir de acordo com as perguntas do usuário de forma Ad-hoc.
- **Agregação dinâmica:** Resumo de dados detalhados.
- **Perfuração (Drilling):** Habilidade de mover-se entre níveis de granularidade de dados.
- **Slice e Dice:** Capacidade de combinar e re-combinar várias dimensões para evoluir diferentes facetas de informações.
- **Pivot:** Capacidade de oferecer comparações, revelar padrões e analisar tendências.
- **Performance:** Acesso a dados e manipulações devem ser executadas em um tempo mínimo, “Tempo de Pensamento”.

# Esquema Estrela



# Processo de Negócio

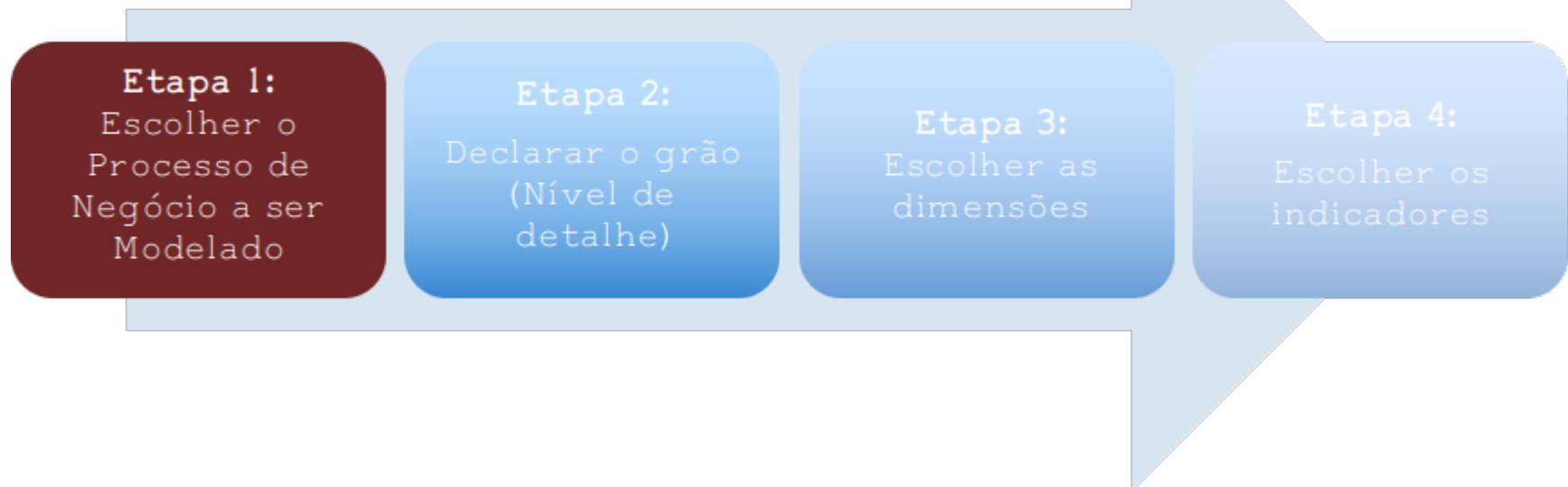


# Exemplo I - Vendas a Varejo

Um empresa que é uma rede de comercialização de carros em miniaturas com:

- 200 filiais.
- Atuando em 15 países
- 80.000 produtos

# Exemplo I - Vendas a Varejo - Etapa 1



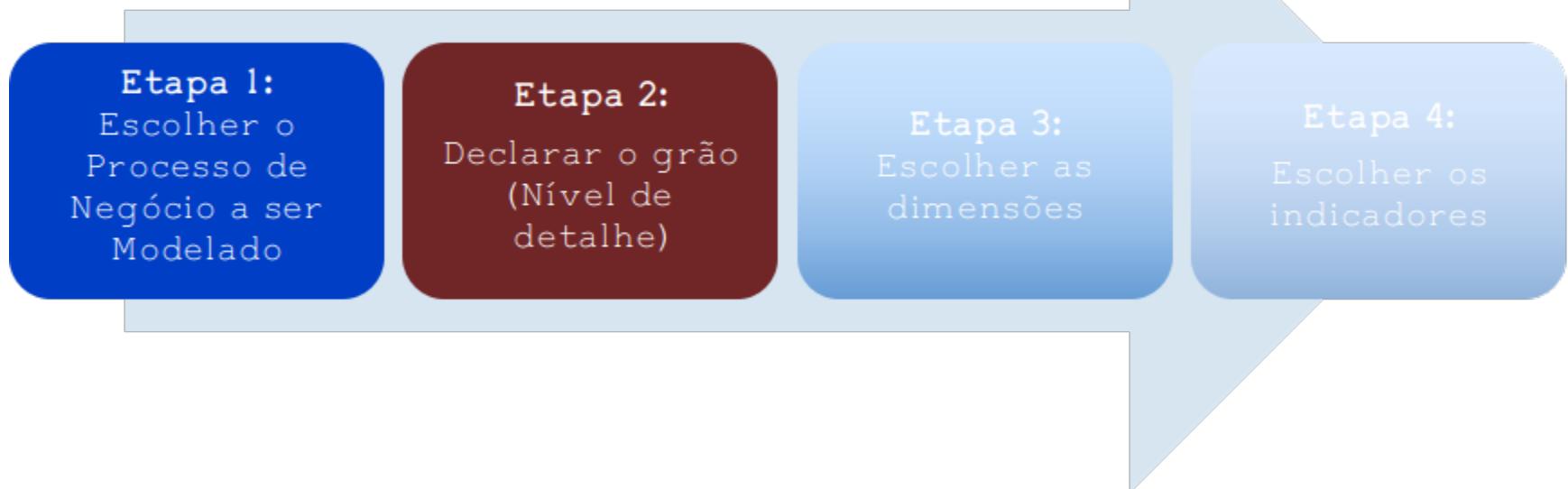
# Exemplo I - Vendas a Varejo - Etapa 1

Deve ser verificado o processo que terá mais impacto para o usuário final.

Exemplo:

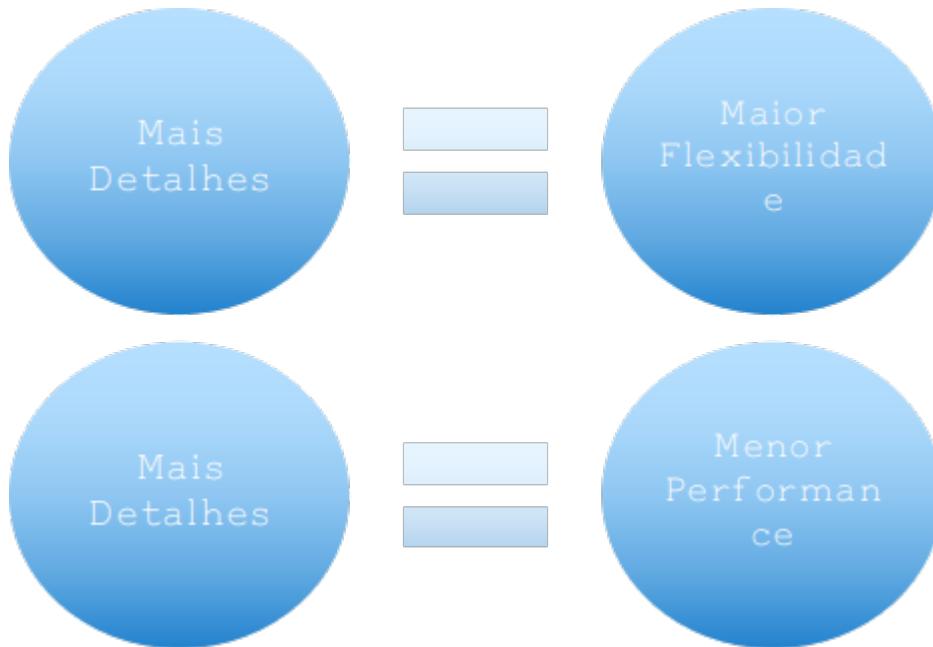
- Usuários-chaves: Gerentes;
- Cada gerente das filiais desejam analisar quais produtos estão vendendo, em quais lojas, em quais dias e em que condições promocionais.

# Exemplo I - Vendas a Varejo - Etapa 2

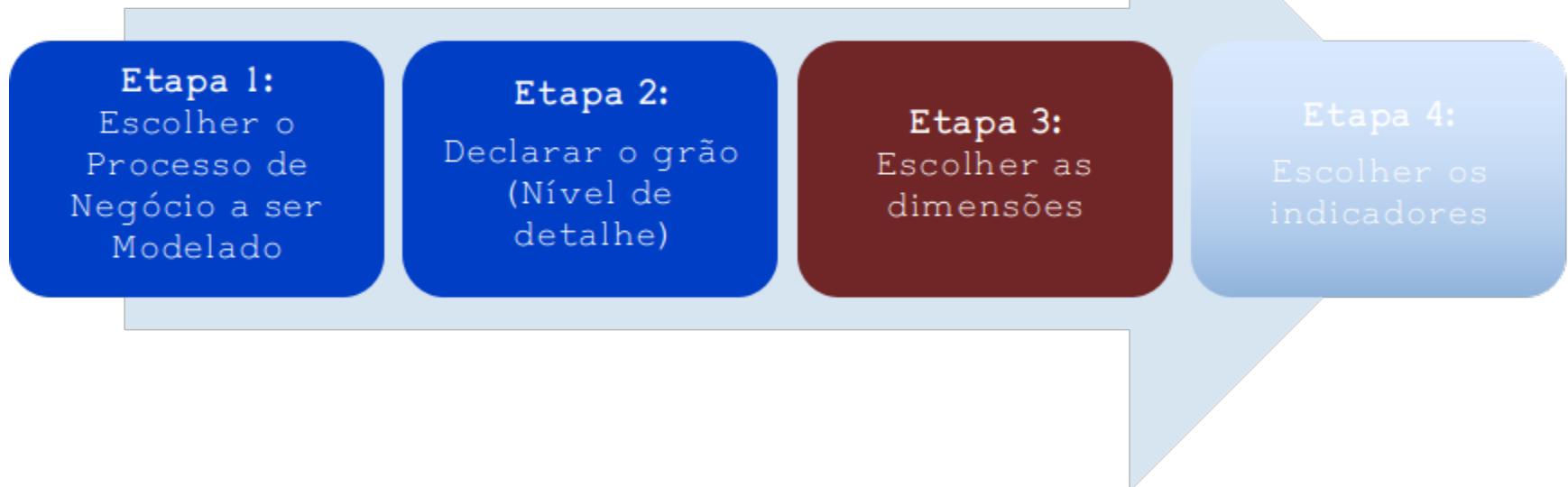


# Exemplo I - Vendas a Varejo - Etapa 2

Qual o nível de detalhe deve ficar disponível no modelo dimensional?



# Exemplo I - Vendas a Varejo - Etapa 3



## Exemplo I - Vendas a Varejo - Etapa 3

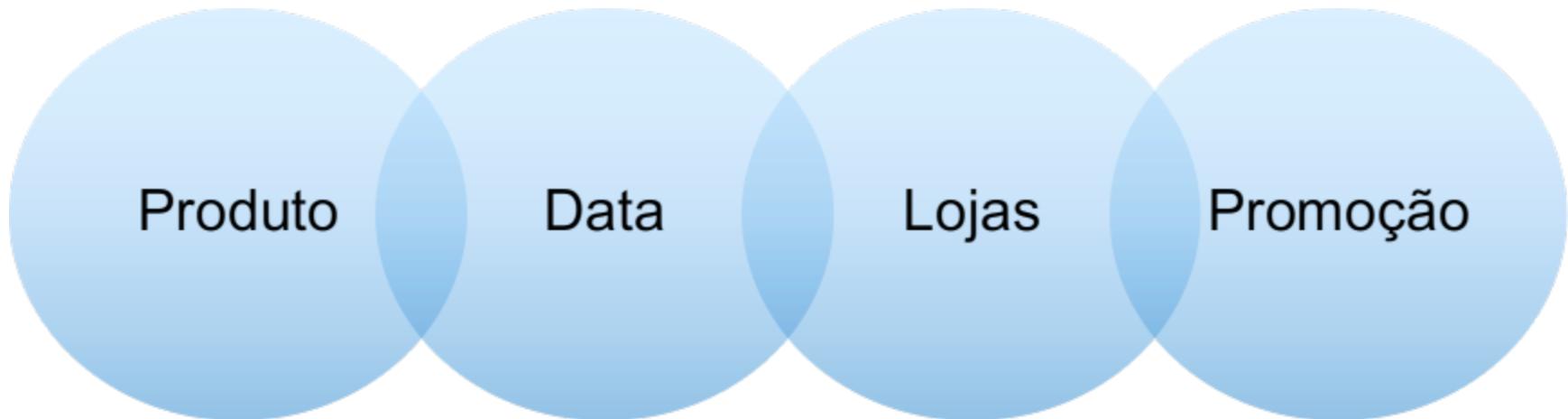
Qual o meu **produto** mais vendido?

Qual **dia da semana** tenho mais vendas?

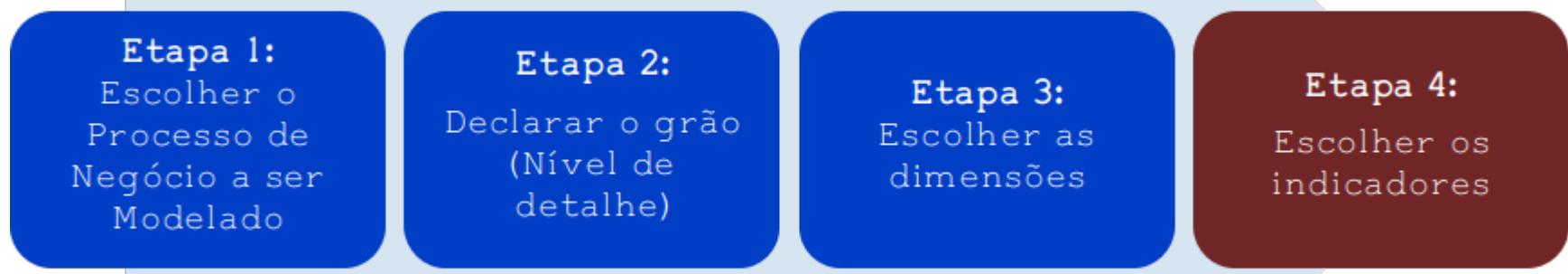
Qual **loja** tem vendido mais?

Qual **promoção** tem aumentado as vendas?

# Exemplo I - Vendas a Varejo - Etapa 3



# Exemplo I - Vendas a Varejo - Etapa 4



# Exemplo I - Vendas a Varejo - Etapa 2

“Os fatos devem ser verdadeiros para o grão” – Ralph Kimball

## Fatos

- Quantidade Vendida
- Valor Total Vendido
- Custo da Promoção
- ...

# Tipos de Dimensões e Fatos

Dimensões:

- Degenerada
- Combinada
- Slowly Changing Dimension
- Conformadas

Fatos:

- Aditivos
- Semi-aditivos
- Não-aditivos

# Dimensões Degeneradas

- Uma dimensão degenerada pode ser representado por um único atributo
- A menos que o tipo de dados seja grande, estas dimensões são armazenados como uma coluna na tabela de fatos.
- Dimensões degeneradas comumente ocorrem quando são do grão da tabela fato.
- Exemplo: Número da nota fiscal, código de venda, ticket aéreo, ordem de serviço.

# Dimensões Degeneradas

- Dimensões degeneradas também podem ser formações de dados que em vez de serem incluídos em uma tabela separada são armazenadas na tabela fato.
- Útil para dimensões de baixa cardinalidade.
- Reduz junções entre tabelas.

# Dimensões Combinadas

- Também chamadas de “Junk Dimensions”.
- Junção de dimensões de baixa cardinalidade.
- Reduz a quantidade de junções.

# Dimensões Combinadas

- Gênero
  - M / F / I
- Tipo de Promoção
  - Desconto
  - Compre um leve outro
  - Degustação



Dimensões Combinadas  
ID Dimensão / Gênero / Tipo  
1 / M / Desconto  
2 / M / Compre um leve outro  
3 / M / Degustação  
4 / F / Desconto  
5 / F / Compre um leve outro  
6 / F / Degustação  
7 / I / Desconto  
8 / I / Compre um leve outro  
9 / I / Degustação

# Slowly Changing Dimensions

- Atributos da dimensão mudam, mesmo que lentamente.
- Para cada atributo de dimensão deve ser especificado uma estratégia.

Dimensão Cliente

Atributos
Nome do Cliente
CPF do Cliente
Data de nascimento
Cidade
Estado Civil
Telefone

# Slowly Changing Dimensions

- Dimensão Cliente

Nome do Cliente, Data Nascimento

- Em caso de mudança, pouco importa o que tinha anteriormente, podendo ser sobreescrito.

Exemplos:

Vagner Silva → Wagner Silva

João Gilbert → João Gilberto

# Slowly Changing Dimensions

- Dimensão Cliente

Cidade, Estado Civil

- Em caso de mudança, todos os eventos relacionados à esse.

Exemplos:

Maceió → São Paulo  
São Paulo → Recife

Casado → Solteiro  
Solteiro → Casado

# Slowly Changing Dimensions

- Dimensão Cliente

Telefone

- Não preciso saber todo o histórico da pessoa, mas uma referência anterior pode ser útil para o departamento de cobrança

Exemplos: Antigo  
82 9931.6799

Atual  
83 9922.5566

# Slowly Changing Dimensions

- Tipo 1 – Sobrescrever

Ignora o histórico e sobrescrever as colunas, utilizado quando não importa os valores anteriores.

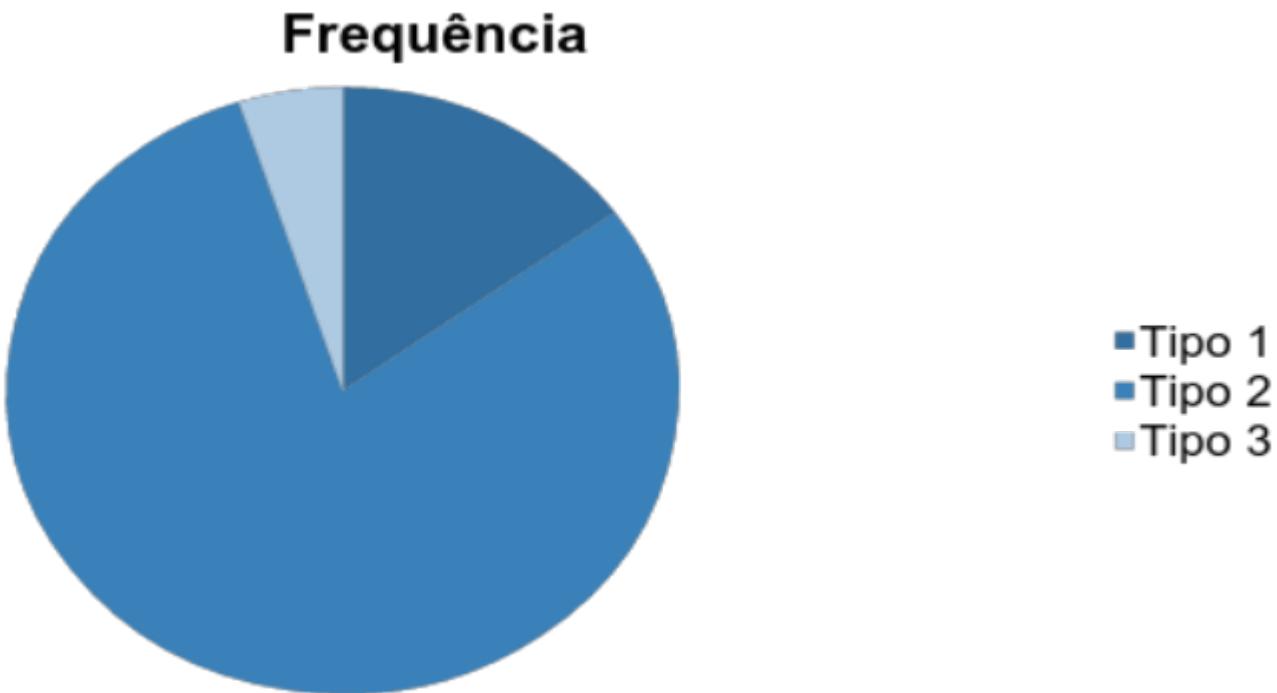
- Tipo 2 – Inserção

Criar um histórico de todas as modificações de status.

- Tipo 3 – Realidade Alternada

Mantém o histórico apenas da última mudança.

# Slowly Changing Dimensions



# Exemplo: Estoque

- Quantos produtos temos disponível até o final do mês?
- Qual deve ser o fluxo de vendas para zerar o estoque?
- Em quantos dias precisamos pedir aos fornecedores mais produtos?



# Exemplo: Estoque

- Instantâneo da posição de estoque no dia.
- Fatos semi-aditivos.
- Qual o valor que representa o estoque no mês passado?

# Exemplo: Estoque

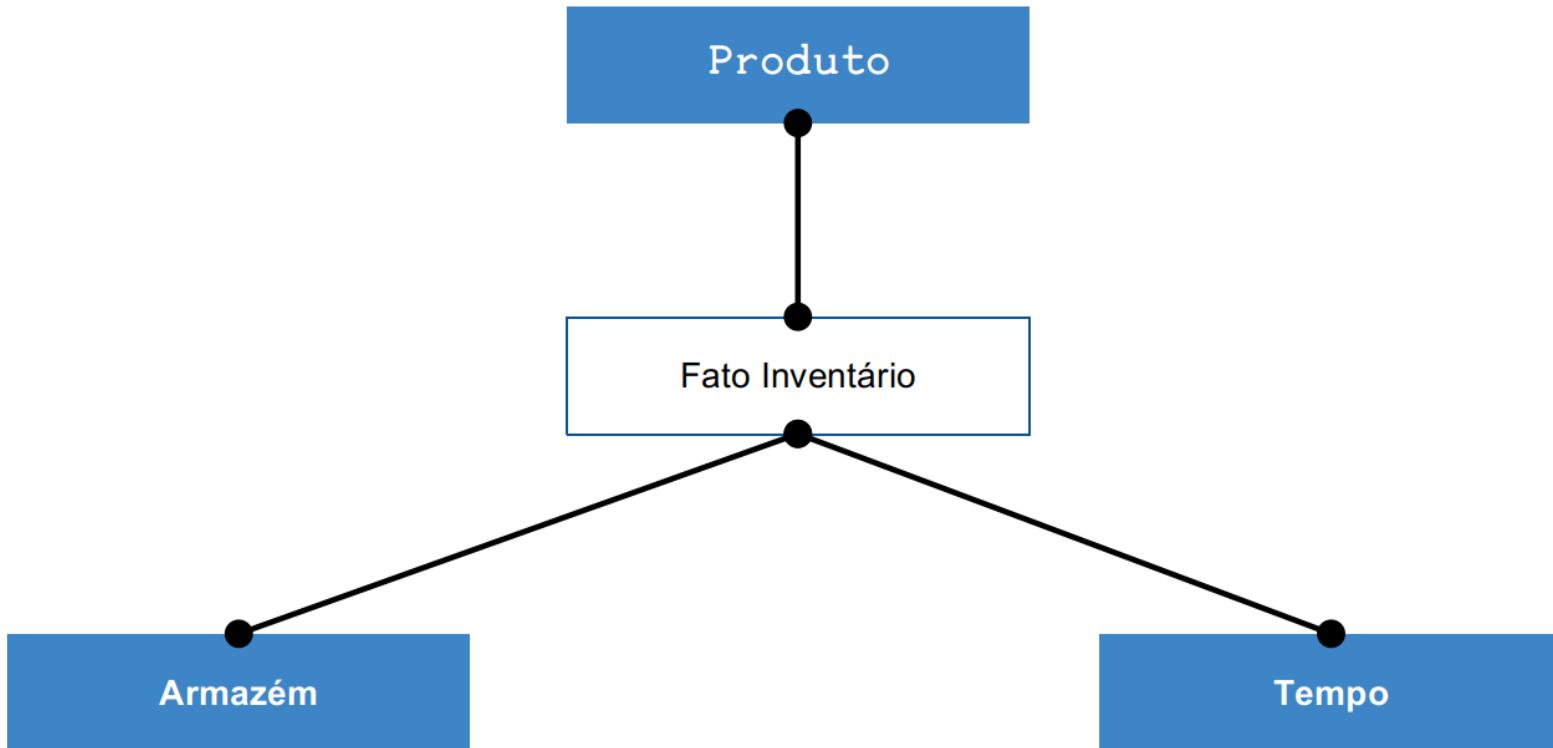
Produto	Data	Qtd
Refrigerante	10/01/2013	1000
Refrigerante	12/01/2013	930
Refrigerante	13/01/2013	890
Refrigerante	15/01/2013	813
Refrigerante	20/01/2013	700
Refrigerante	23/01/2013	600



Produto	Data	Qtd
Refrigerante	10/01/2013	1000
Refrigerante	11/01/2013	1000
Refrigerante	12/01/2013	930
Refrigerante	13/01/2013	890
Refrigerante	14/01/2013	890
Refrigerante	15/01/2013	813
Refrigerante	16/01/2013	813
Refrigerante	17/01/2013	813
Refrigerante	18/01/2013	813
Refrigerante	19/01/2013	813
Refrigerante	20/01/2013	700
...	...	...
Refrigerante	23/01/2013	600
Refrigerante	31/01/2013	600

Qual é a quantidade de estoque de Refrigerante no fim do mês de janeiro?

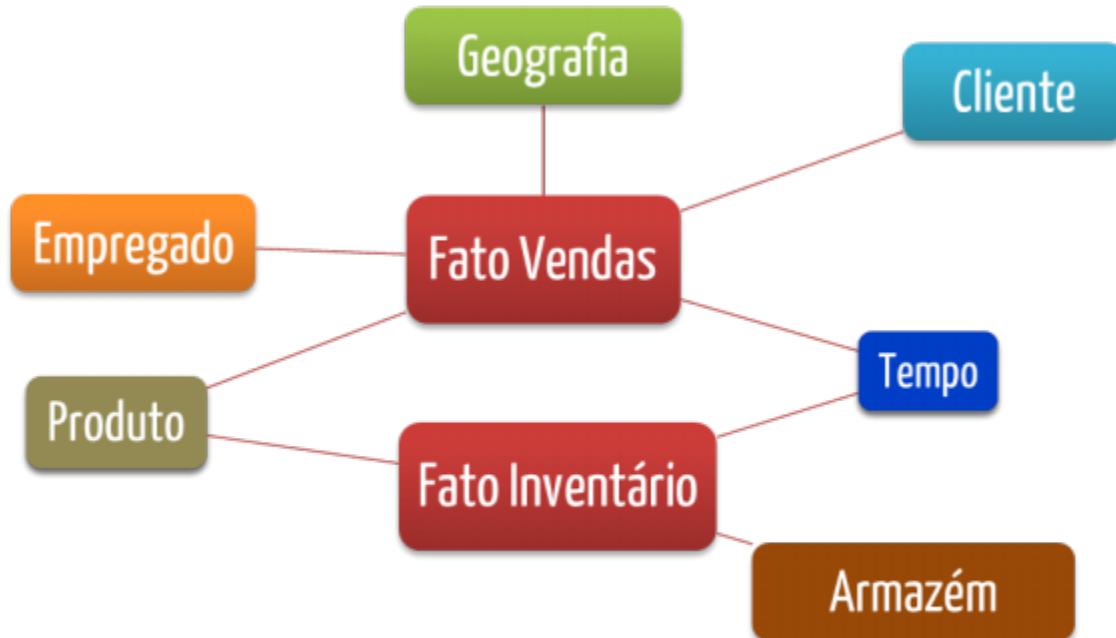
# Exemplo: Estoque



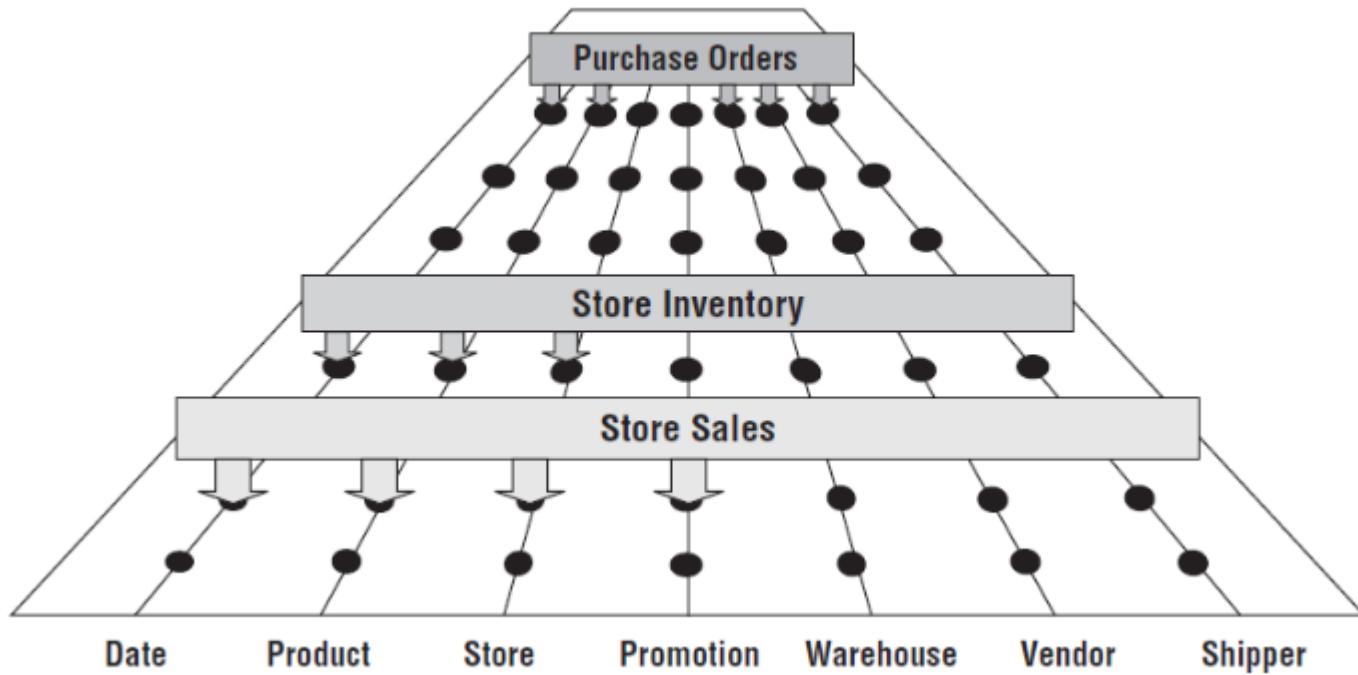
# Dimensões Conformadas

- Dimensões conformadas são dimensões que podem ser compartilhadas entre esquemas estrelas.
- Permite projeto de banco de dados analítico escalável.
- Permite análise e agregação em diferente áreas.

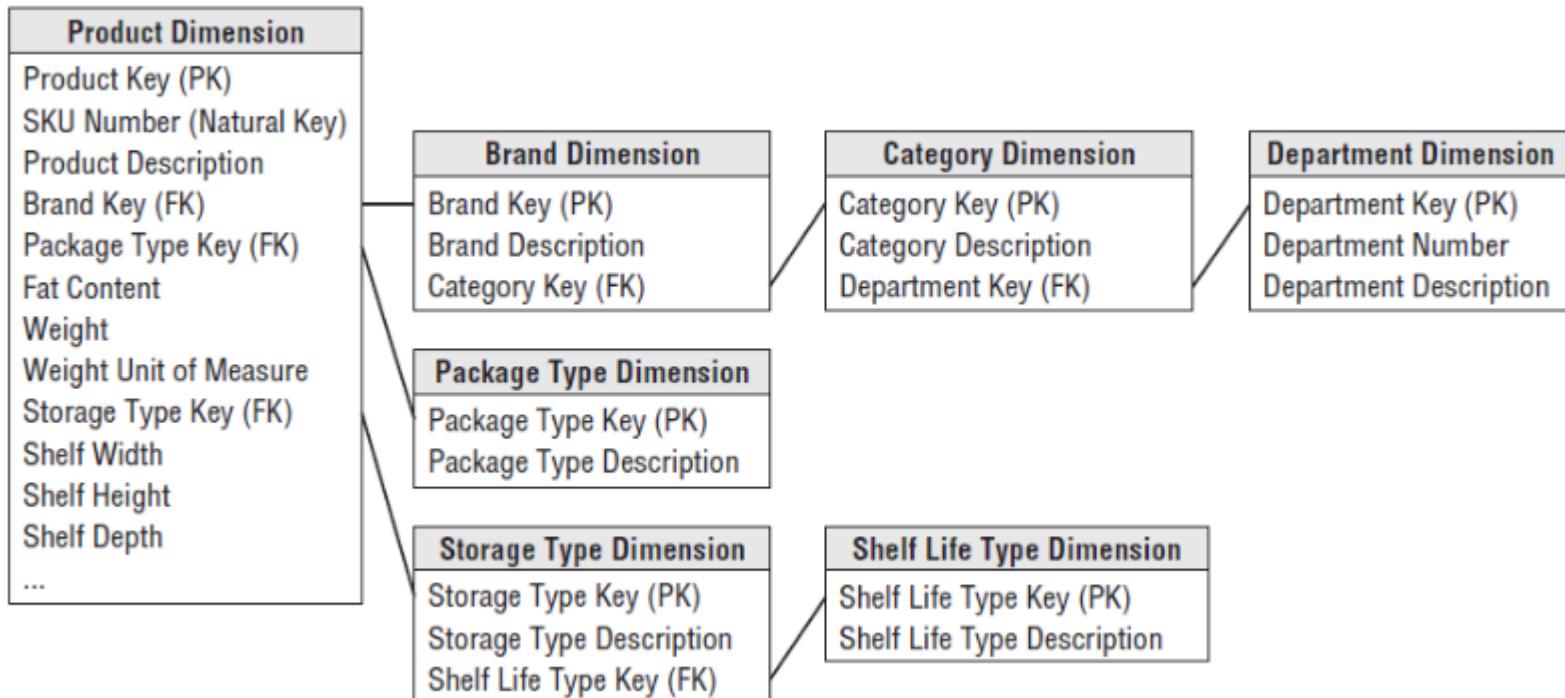
# Dimensões Conformadas



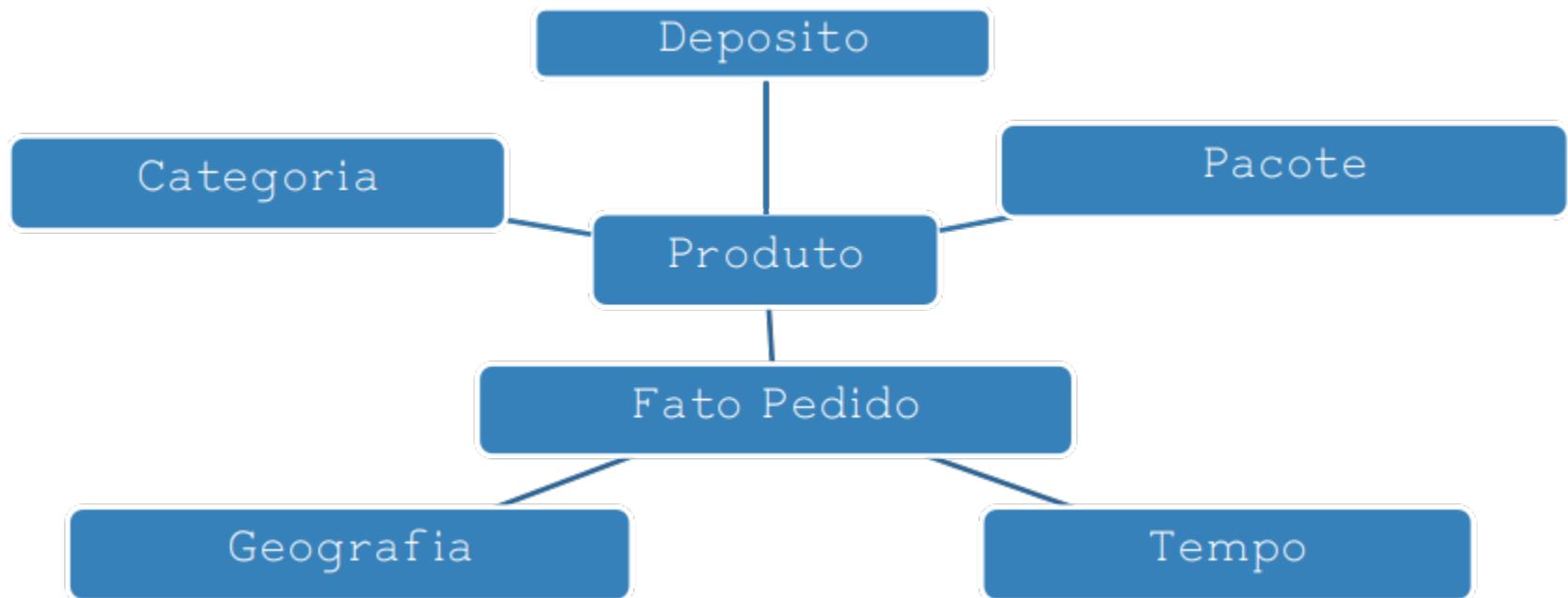
# Dimensões Conformadas - Bus Matriz



# Snow Flake



# Snow Flake



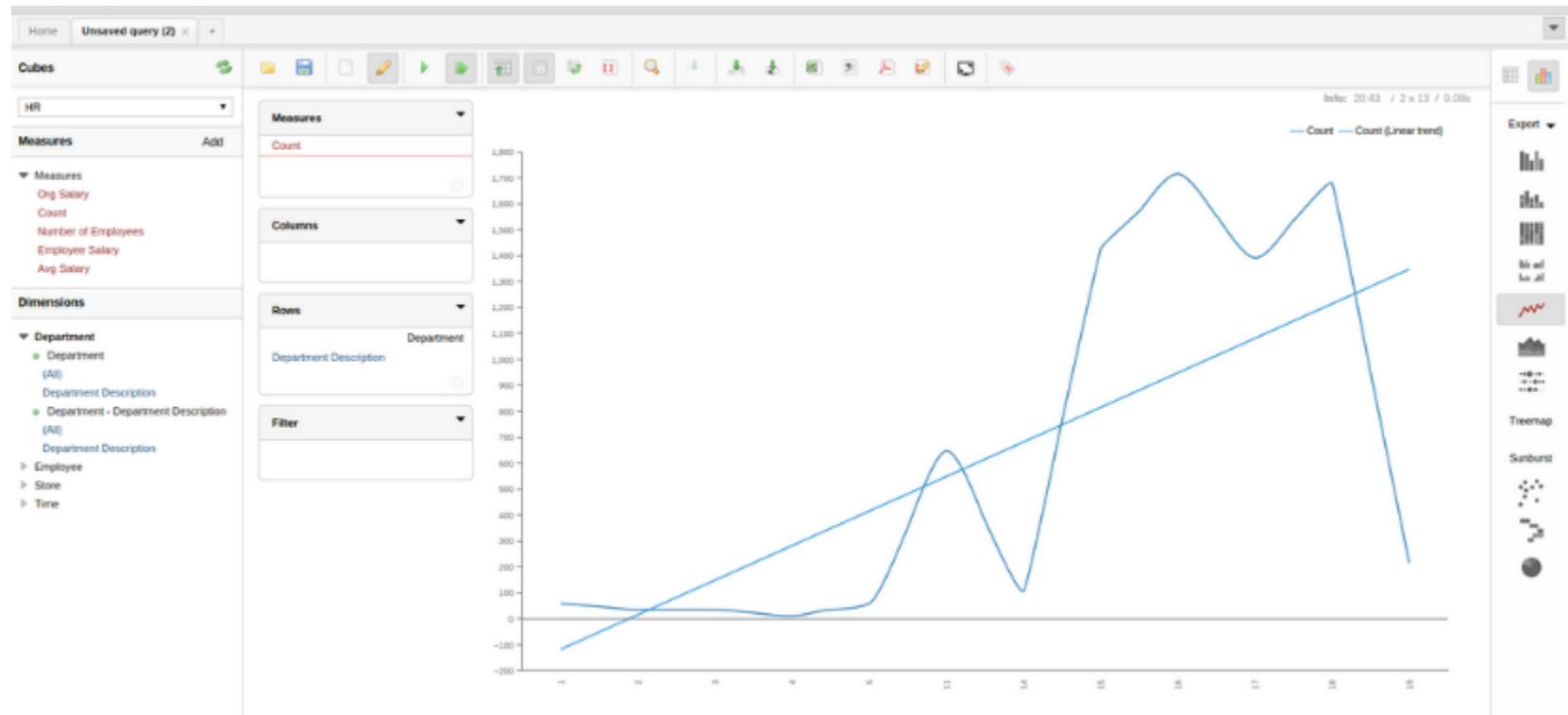
# OLAP - Online Analytical Processing

# OLAP

É uma categoria de software que permite aos analistas, gerentes e executivos obter *insights* sobre dados através de um acesso interativo, rápido e consistente para uma ampla variedade de pontos de vista da informação, que foi transformada a partir de dados brutos para refletir a dimensionalidade real da empresa

O poder da ferramenta permite aos analistas a escolha de medidas e dimensões que precisam analisar, assim como efetuar Drill, Slice and Dice, Up, Filter e Pivot para descobrir novos relacionamentos, oportunidade e problemas.

# OLAP



# OLAP

**Data Analytics**

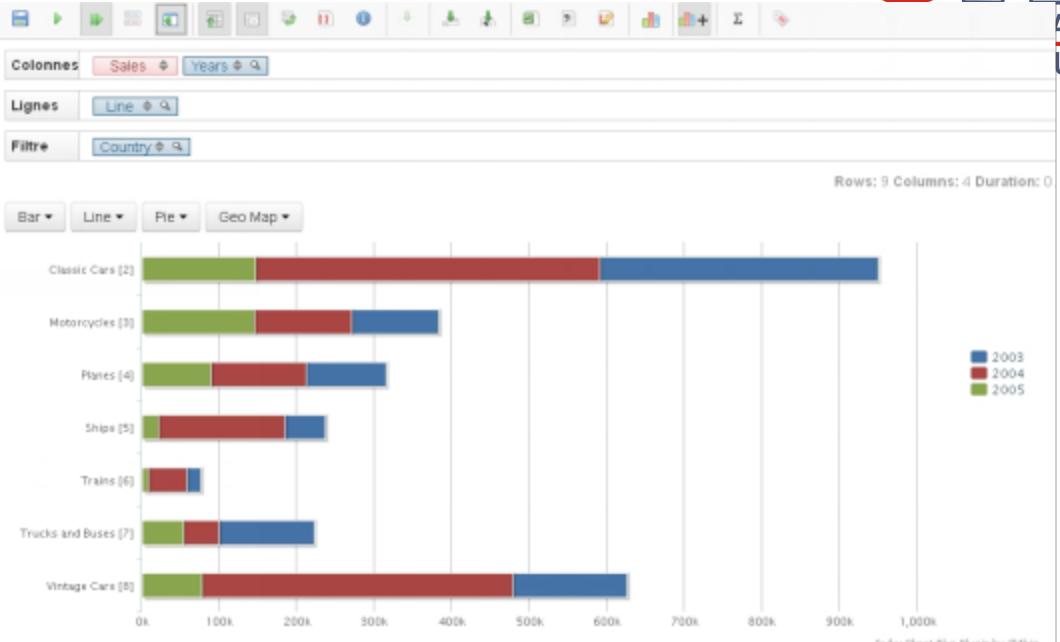
**Cubes**  
SteelWheelsSales

**Dimensions**

- Customers
- Markets
  - (All)
  - Territory
  - Country
  - State Province
  - City
- Order Status
  - (All)
  - Type
- Product
  - (All)
  - Line
  - Vendor
  - Protect
- Time
  - (All)
  - Years
  - Quarters
  - Months

**Measures**

- Measures
  - Quantity
  - Sales



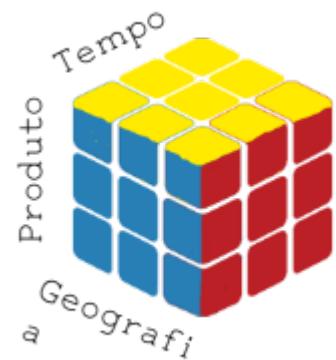
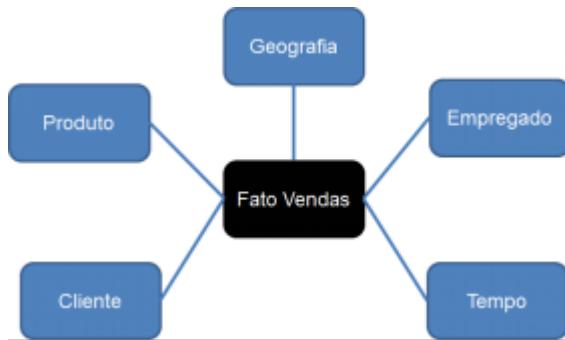
# Benefícios dos Cubos OLAP

- O modelo de dados relacional é impulsionado pela necessidade de juntar várias tabelas. A complexidade e o número de join depende da complexidade do esquema.
- O esquema em estrela simplifica o modelo de dados para ferramentas de consulta.
- Em contraste, o cubo multi-dimensional isola o usuário do esquema subjacente para que a ferramenta de consulta gere o processo de associação para o usuário.
- O usuário pode então se concentrar em analisar os dados sem se preocupar com as estruturas de esquema de apoio.

# Lógico x Físico

## Cubo

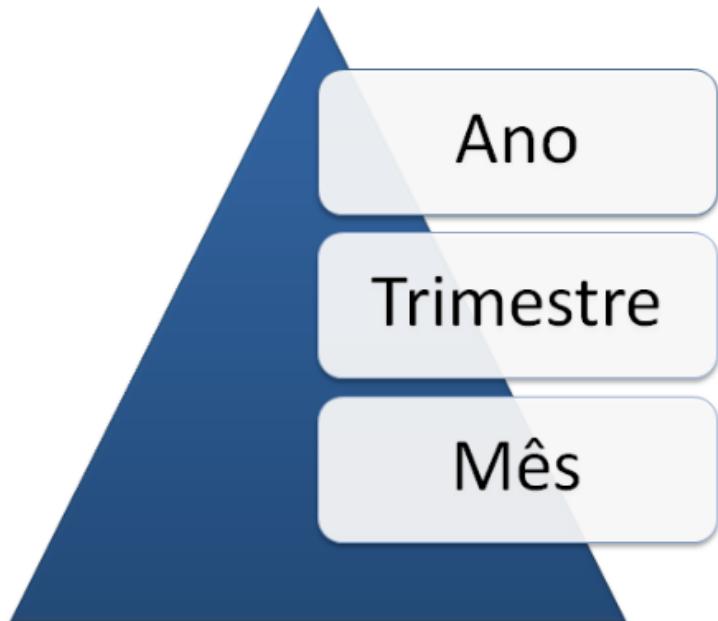
- Construção lógica que contém dimensões e medidas
  - Dimensões
    - Hierarquias
    - Níveis
    - Atributos
  - Medidas



Cube é Lógico e o Star Schema é o Físico

# Hierarquia

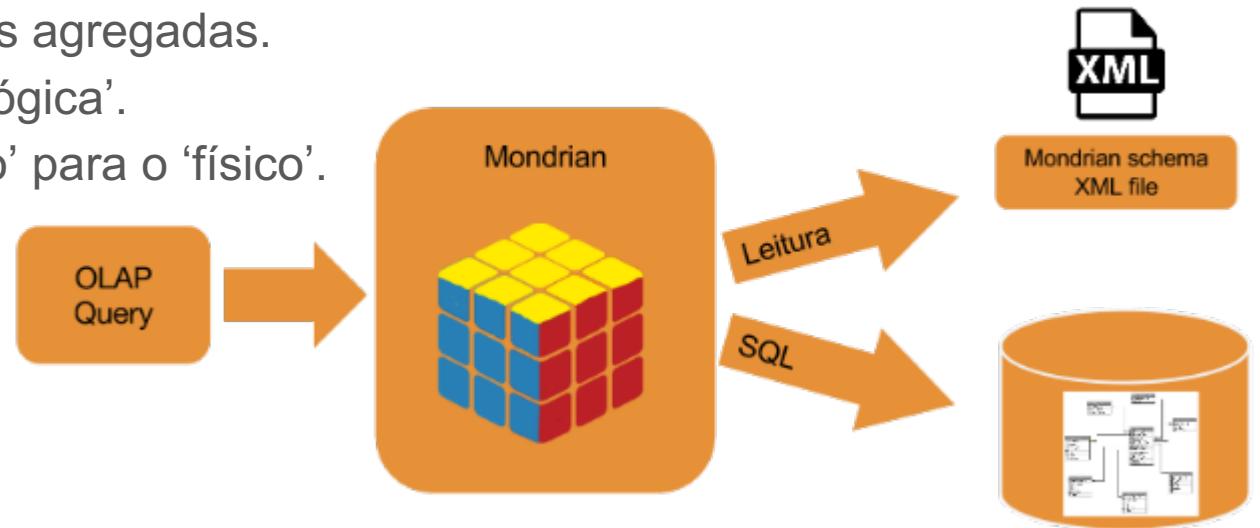
O Dado é sumarizado para cada nível da hierarquia



# Pentaho Mondrian

Um arquivo XML define:

- Cubos (Medidas + Dimensões).
- Segurança e tabelas agregadas.
- Ele define a parte ‘lógica’.
- Ele mapeia o ‘lógico’ para o ‘físico’.



# Pentaho Mondrian

É um Engine ROLAP

- Java = Runs on everything

Engine, cria visões multidimensionais do dado num banco de dados.

# Pentaho Mondrian

## Recursos:

- Compilador de MDX
- XML para Analysis e compilador JOLAP
- Suporte a agregação, calculos e categorização dos dados para bancos de dados relacionais, incluindo Oracle, MySQL, Microsoft SQL Server, IBM DB2 e muitos outros.
- Otimização automática:
- Expressões pré-compiladas de MDX permitem ao servidor do Mondrian otimizar as consultas, escolhendo entre processar a expressão na memória ou enviar a consulta para o RDBMS.

A ferramenta visual para auxiliar a criação de cubos é o **Schema Workbench**.

ETL

# ETL

ETL é um processo de integração de dados que envolve:

- **Extrair** dados de uma fonte de dados;
- **Transformar** de acordo com o banco de dados destino;
- **Carregar** (Load) dados em um destino.

ETL é necessário para:

- **Integrar** dados de fontes heterogêneas;
- **Limpeza** de dados, remoção de problemas transacionais e introduzir consistência;
- **Reestruturar** dados para otimizar o processamento analítico.

# Extração

- Acesso a fontes de dados.
- Captura eficiente de cada fonte de dados:
  - Presença mínima, importante para os administradores de banco de dados.
  - Apenas obtém os dados que são necessários.
  - Só executa com a frequência necessária para cumprir a exigência do processamento analítico.
- Geração de formatos de dados que podem ser consumidos de maneira consistente por transformações e processos de carga.

# Extração

- Potencialmente pode armazenar os dados em uma área comum (ODS) para cruzamentos de transformações e integração.
- Pode envolver um passo no transporte para mover dados extraídos do banco de dados do servidor de origem para o servidor de destino.

# Transformação

- Transforma o formato de dados extraídos da origem para a estrutura de banco de dados de destino.
- Pode envolver “passos” que realizam o seguinte:
  - Sumarização de valores de múltiplas linhas de dados;
  - Impor regras de qualidade de dados e processamento de exceção;
  - Segregação de específicas colunas para carga;
  - Decodificação de valores;
  - Codificação de valores de forma livre para corresponder ao formato de dados de destino;
  - Processamento de cálculo de dados;
  - Geração de chaves técnicas;
  - Troca de eixos dos dados;
  - População de tabelas agregadas.

# Carga

- Movendo dados transformados para a fonte de destino de forma correta e eficiente.
- Carregamento em massa (**Bulk loading**) para grandes volumes de dados.
- Pode exigir re-indexação ou tarefas de ajustes de dados físicos antes e/ou depois do processo de carga.

# Pentaho Data Integration

# Pentaho Data Integration

Data Integration (ou Kettle) é uma ferramenta poderosa que possui capacidades de Extração, Transformação e Carga (ETL), usando uma inovadora abordagem orientada a metadados.



# Pentaho Data Integration

Fácil de Usar:

- 100% orientada a metadados (define O QUE fazer e não COMO);
- Sem geração de código extra, menor complexidade;
- Configuração simples, interface intuitiva e de fácil manutenção.

Flexibilidade:

- Nunca força um certo caminho para o usuário;
- Arquitetura “plugável” que permite a extensão de funcionalidades.
- Arquitetura baseada em padrões modernos;
- 100% construído em Java, permitindo suporte multi-plataforma;
- Mais de 200 (e crescendo) “out-of-the-box” objetos de mapeamento (Steps e Jobs Entries).

# Usos Comuns

- Corrigindo e modificando dados para melhorar análise em Agile BI.
- Enriquecimento de informações através de “looking up” de dados em várias fontes de informação (arquivo CSV, arquivos text, planilhas e etc).
- Exportar um banco(s) de dados para um arquivo(s) CSV ou outro banco de dados.
- Importar dados para um banco de dados, oriundos de um arquivo de texto ou planilhas.
- Migração de dados.
- Exploração dos dados existentes em uma base de dados (tabelas, visões e etc).

# Usos Comuns

- Popular um Data Warehouse
- Construção de suporte para slowly changing dimensions, junk dimensions e outros conceitos de data warehouse.

# Conceitos: Transformações e Jobs

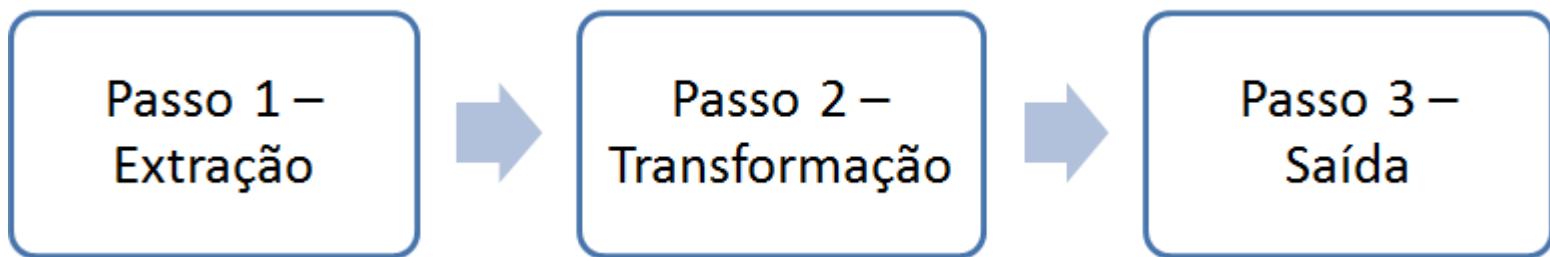
- Transformações
  - Processo atômico que buscam de um ou mais fontes de entrada, executam transformações e carregam um ou mais saídas.
  - Entradas podem ser tabelas, arquivos, XMLs e outros.
  - Saídas podem ser tabelas, arquivos, XMLs e outros.
  - Pode ter uma lógica de exceção para tratar arquivos “ruins”, usando remediação inteligente ou falhas programáticas, dependendo do projeto.
  - Geração de estatísticas de auditoria (registros lidos, escritos e etc...) e logs.

# Conceitos: Transformações e Jobs

- Job
  - Um processo coordenado através de uma ou mais transformações.
  - É geralmente utilizado para agendar a execução.
  - Introduz processos lógicos como ramos, paralelismos e outros.
  - Usualmente permite automação de notificações de resultado via e-mail ou alertas.

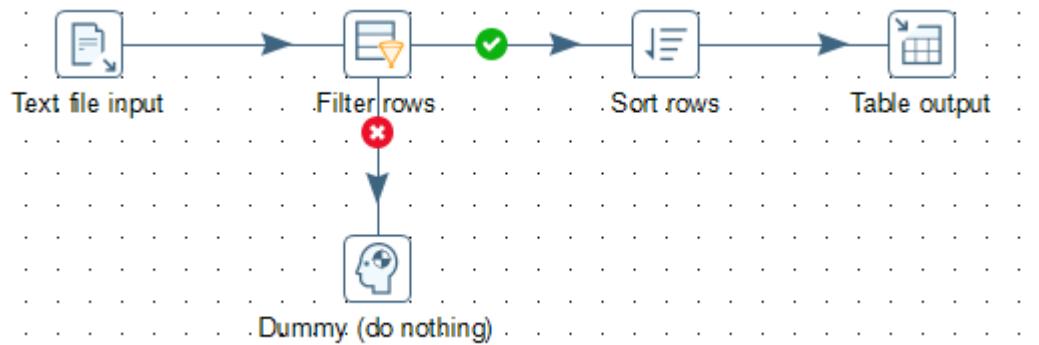
# O que é uma transformação no PDI

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



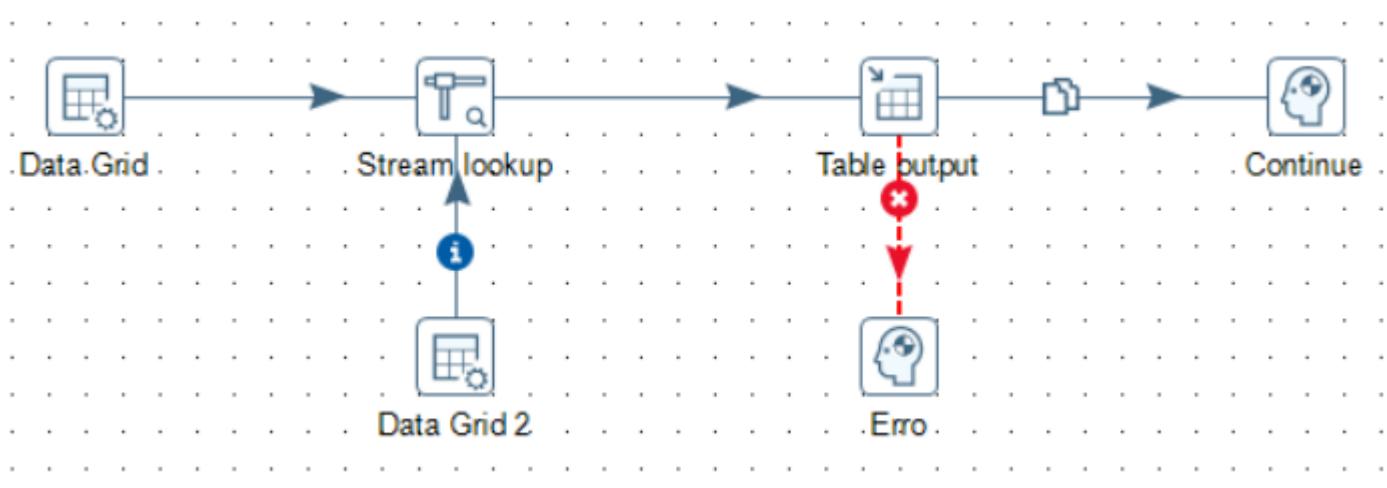
# Transformações

- Transformações são uma rede de tarefas lógicas (steps).
  - Ler um arquivo.
  - Filtrá-lo.
  - Ordená-lo
  - Carregá-lo em uma tabela MySQL.



# Hops

- Info Steps: Quando o dado é recolhido de para outro step
- Error handling steps: Quando a manipulação de erro é utilizada



# Motor de Workflow

- Todos os steps são inicializados e executados em paralelo:
  - Não é possível prever sequência de inicialização.
- O Kettle é capaz de processar um número infinito de linhas:
  - Desempenho varia em velocidade e consumo de memória.
  - É possível definir quantidade máxima de linhas executadas por steps
  - Transformation Properties > Miscellaneous > Number of rows in rowset

# Componentes do PDI

## Spoon

- Ambiente gráfico para modelagem de transformações e jobs

## Kitchen

- Aplicativo para execução de *jobs* via linha de comando

## Kettle

## Pan

- Aplicativo execução de transformações via linha de comando

# Inicialização

Pan

Windows

Pan.bat

Linux ou MacOS

pan.sh

Kitchen

Windows

Kitchen.bat

Linux ou MacOS

kitchen.sh

Spoon

Windows

Spoon.bat

Linux ou MacOS

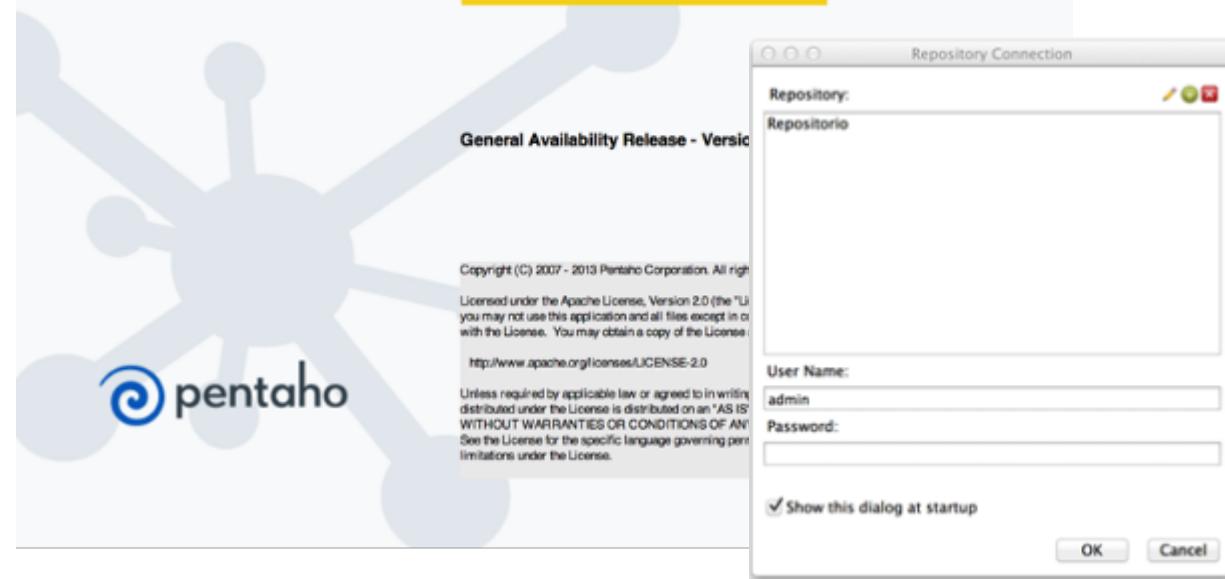
spoon.sh

Todos executáveis encontram-se dentro de:

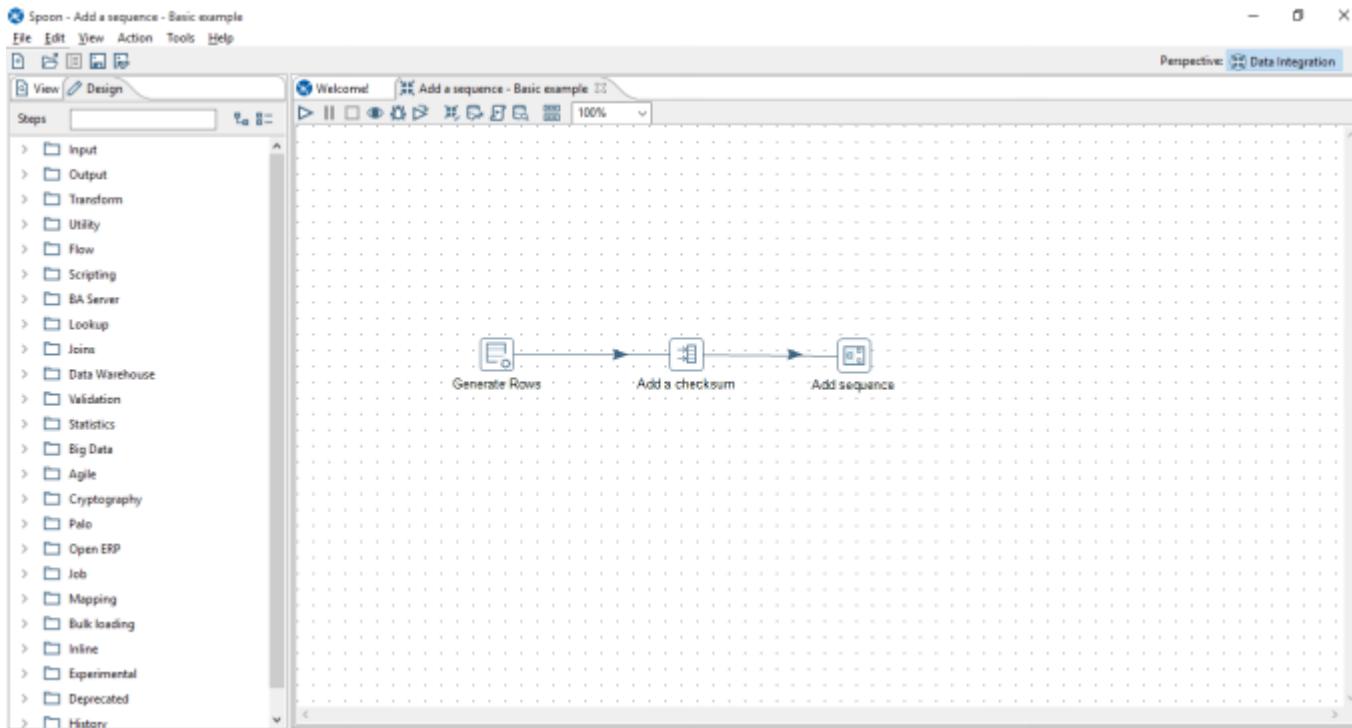
{diretório de instalação} data-integration/

# Spoon

## Pentaho Data Integration

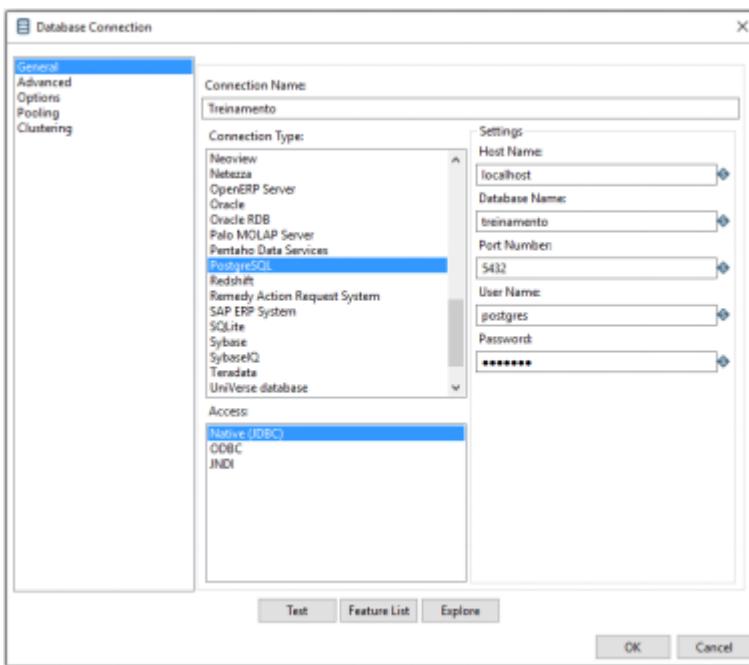
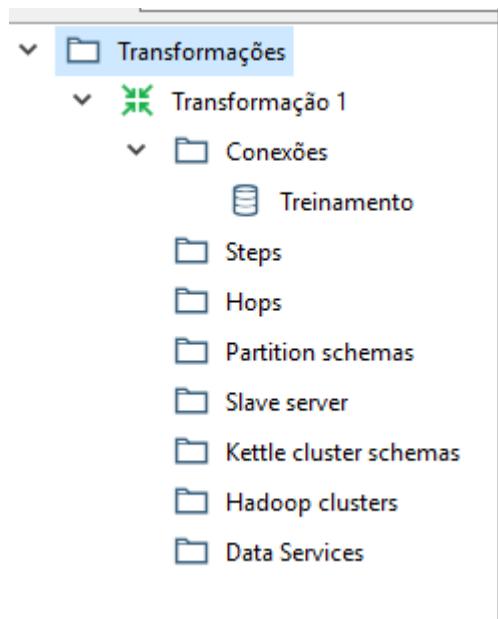


# Spoon - Elementos



# Principais Steps

# Conexões

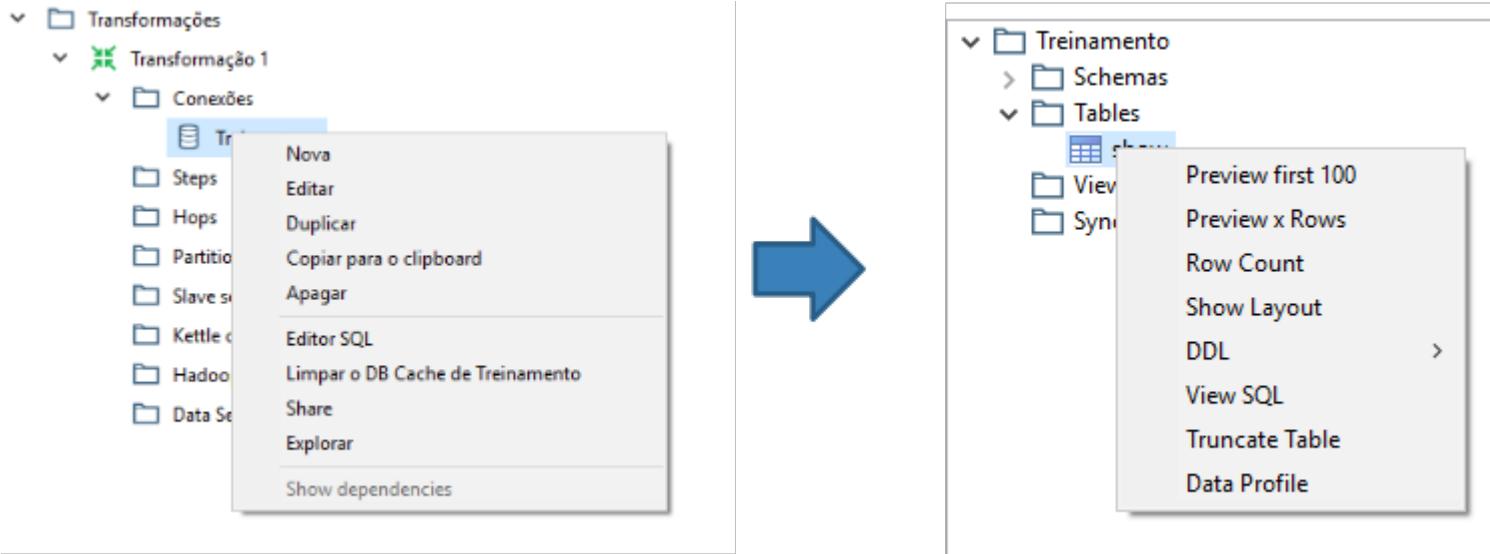


# Conexões

- Nativamente o PDI vem com diversos drivers JDBC.
- Drivers adicionais podem ser acrescentados em data-integration/lib/
- Drivers que já estão na pasta lib podem ser substituídos caso a versão do banco seja alterada.

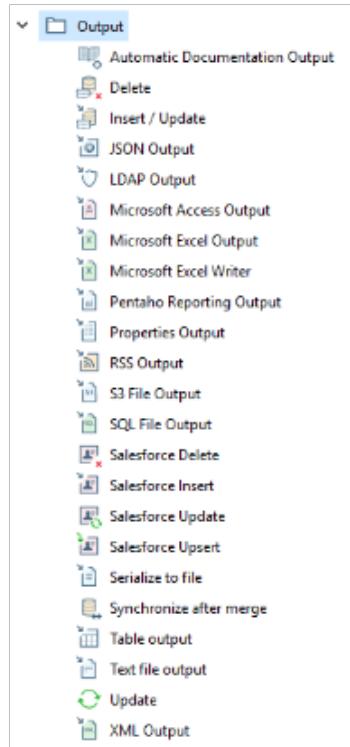
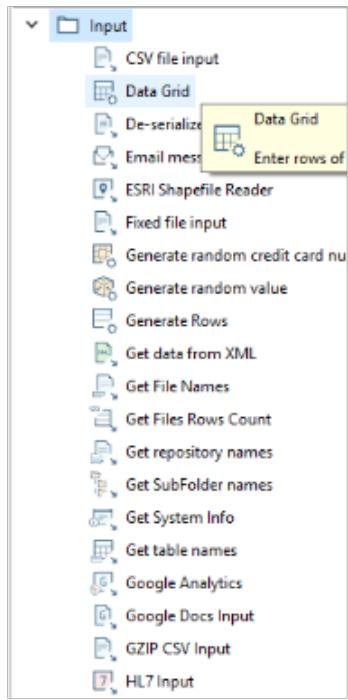
# Database Explore

- Permite com que os usuários explorem o banco de dados sem a necessidade de sair da ferramenta.
- Possui tela com opções de contexto.



# Inputs e Outputs

## Inputs

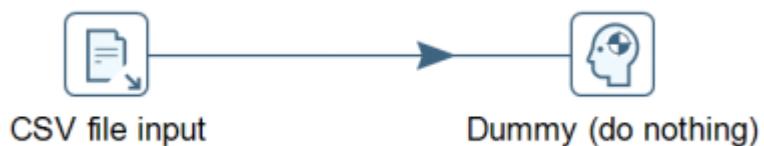


# CSV Input

- Lê arquivos de texto separado por caractere.
- Apesar de se chamar CSV (comma separated value – valor separado por vírgula), qualquer caractere pode ser usado.
- Mais rápido quando comparado ao Text File Input.



CSV file input



CSV Input

Step name: **CSV Input**

Filename: \${Internal.Transformation.Filename.Directory}/files/customers-100.txt

Delimiter: ,

Enclosure: "

Max buffer size: 50000

Lazy conversion:

Header row present:

Add filename to result:

The row number field name (optional):

Running in parallel:

New line possible in fields:

File encoding: UTF-8

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	id	Integer		3	0	€	,	-	left
2	name	String		10					none
3	firstname	String		13					none
4	zip	Integer		5	0	€	,	-	left
5	city	String		8					none
6	birthdate	Date	yyyy/MM/dd	10		€	,	-	none
7	street	String		11					none
8	houseNr	Integer		3	0	€	,	-	left
9	stateCode	String		9					none
10	state	String		30					none

Help OK Get Fields Preview Cancel

# Fixed File Input



Fixed file input

- Lê arquivos de texto delimitado exclusivamente por tamanho.
- Para campo é necessário definir atributos de tamanho, padding e alinhamento.
- Tamanho é definido pela quantidade de caracteres de cada campo.



Fixed file input



Dummy (do nothing)

# Microsoft Excel Input



Microsoft Excel Input

- Este step importa planilhas a partir de arquivos em formato:
- Microsoft Excel 2003.
- Microsoft Excel 2007.
- OpenOffice Calc.
- O preenchimento das abas Files, Sheets e Fields é obrigatório

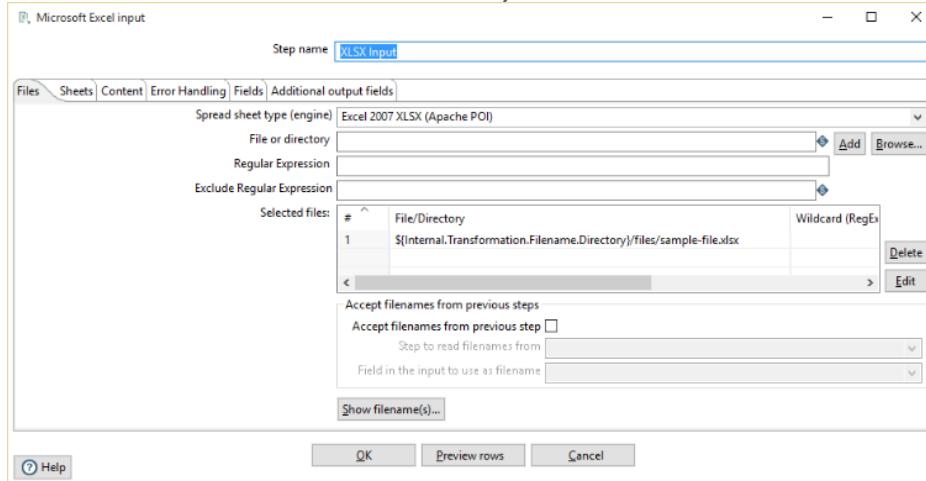
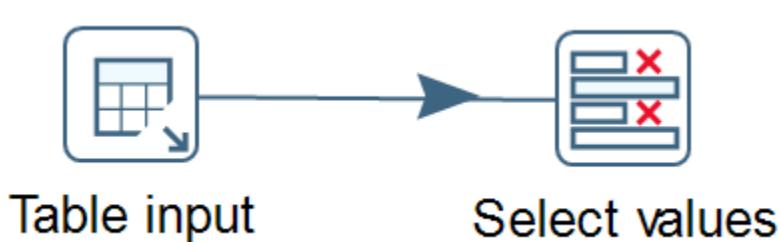




Table input

# Table Input

- Step utilizado para extrair informações a partir de um banco de dados.
- Necessita de uma conexão válida para o SGBD.
- Gerador de SQL.
- Consultas parametrizadas.
- Não permite SQL injection (usar step Execute SQL).





Get System Info

# Get System Info

- Recupera informações do ambiente PDI.
- Pode trazer os tipos de informação:
  - Informações de data e hora
  - Variáveis de Ambiente JAVA
  - Metadados transformação de tempo de execução
  - Argumentos de linha de comando



Text file output

# Text File Output

- Exporta dados em diferentes formatos de arquivos, incluindo CSV.
- As opções fornecidas pelo PDI para gerar saídas de arquivo de texto são:
  - Extensão.
  - Apende.
  - Separador.
  - Delimitador (Enclosure).
  - Cabeçalho/Rodapé.
  - Compressão.
  - Incluir número do step/Data/tempo no nome do arquivo.
  - Codificação.

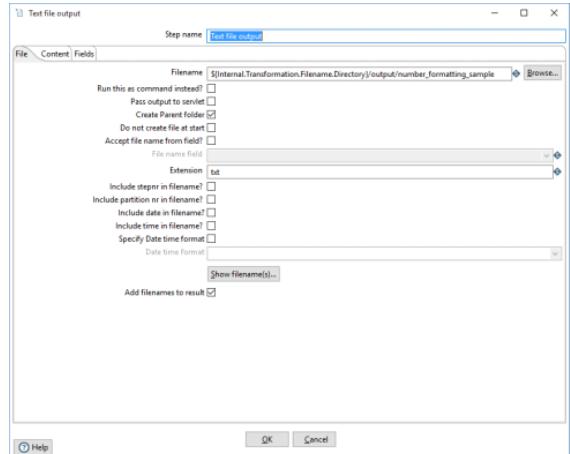
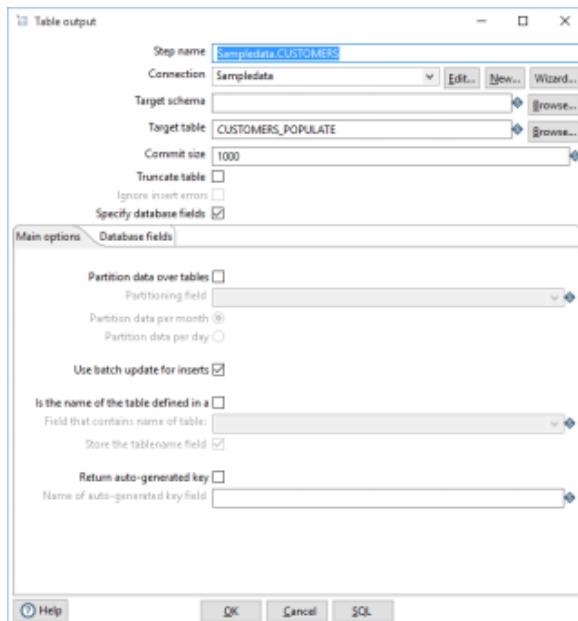




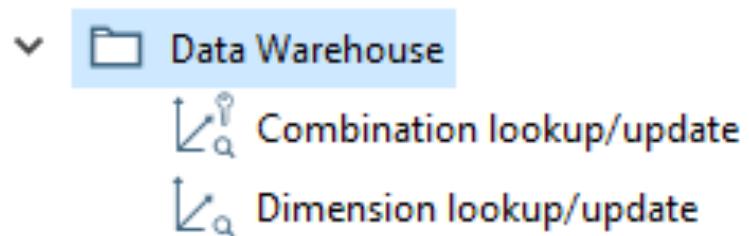
Table output

# Table Output

- Executa apenas Insert de dados para uma tabela de um banco de dados.
- Requer conexão com banco de dados válida.
  - Atributos
  - Tabela de destino.
  - Tamanho do commit.
  - “Truncar” tabela de destino.
  - Usar batch update para inserts.
  - Retornar chave artificial gerada.
  - Nome da tabela definida em um campo



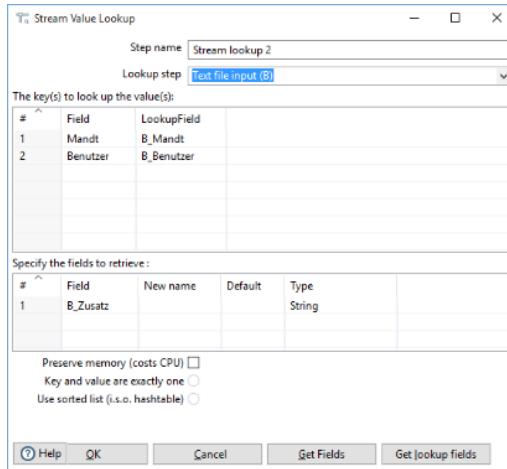
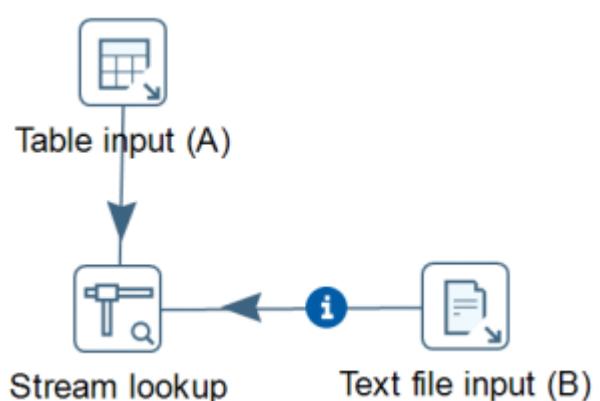
# Data Warehouse



# Stream Lookup

O tipo de etapa de pesquisa Fluxo permite consultar dados usando informações provenientes de outras etapas da transformação. Os dados provenientes da etapa Fonte é lido primeiro na memória e é, então, usada para procurar dados do fluxo principal.

No exemplo, a transformação acrescenta informações provenientes de um arquivo de texto (B) com dados vindos de uma tabela de banco de dados (A). Informação de B é usada para executar as pesquisas, tal como indicado pela opção Fonte passo mostrado ao lado.



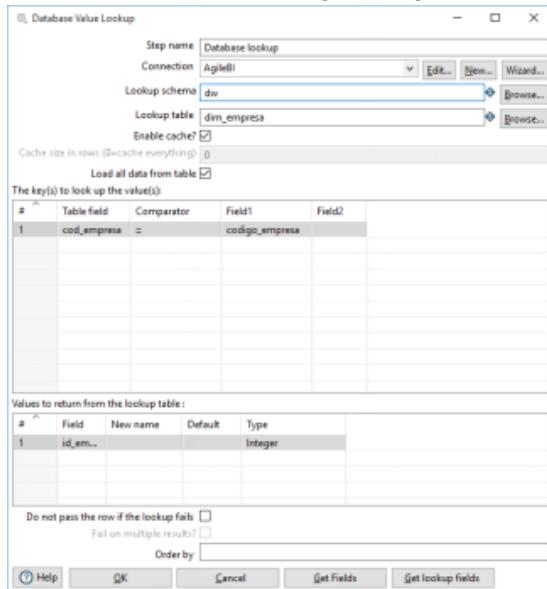
Stream lookup

# Database Lookup



Database lookup

O step Database Lookup permite procurar valores em uma tabela de banco de dados. Valores de pesquisa são adicionados como novos campos para o fluxo.



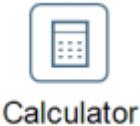


Select values

# Select Values

Este step é útil para selecionar, renomear, alterar os tipos de dados e configurar o comprimento e precisão dos campos no fluxo.

#	Fieldname	Rename to	Length	Precision
1	v_date	IN_DATE		
2	START_DAY			
3	Days_since			
4	Year			
5	Month			
6	DayOfYear			
7	DayOfMonth			
8	DayOfWeek			
9	WeekOfYear			
10	Quarter			
11	date_pk			
12	QuarterName			
13	YearQuarterName			
14	DayOfWeekDesc			
15	DayOfWeekShortDesc			
16	DayOfWeekendInd			
17	MonthDesc			
18	MonthShortDesc			
19	week_of_month_number			
20	week_of_month_name			
21	week_name			



# Calculator

- O Calculator disponibiliza uma lista de funções que podem ser executadas nos valores dos campos.
- Cálculos mais rápidos que o Step JavaScript.
- Campos temporários podem ser removidos.

The screenshot shows the configuration interface for a 'Calculator' step. At the top, there's a title bar with the window title 'Calculator'. Below it, the 'Step name' is set to 'Calculator'. The main area is titled 'Fields:' and contains a table for defining new fields:

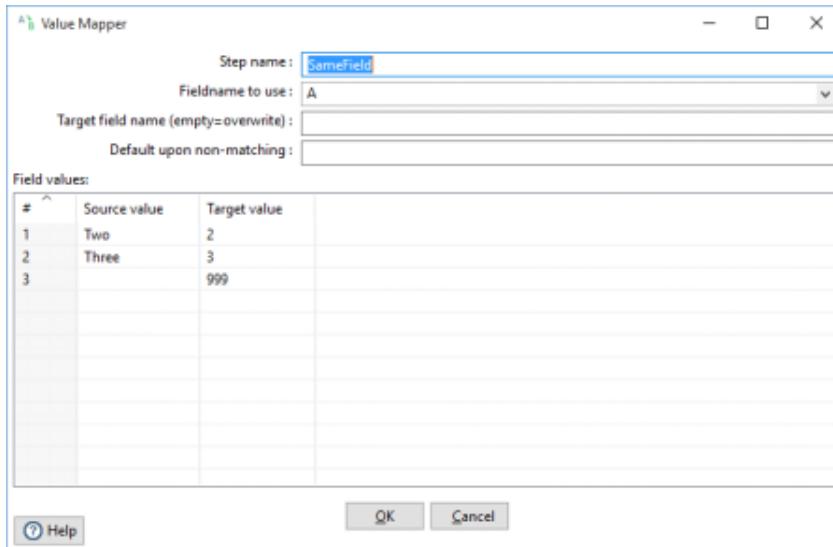
#	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove
1	valor_bruto	A * B	quantida...	preco_u...		Number			
2									



Value Mapper

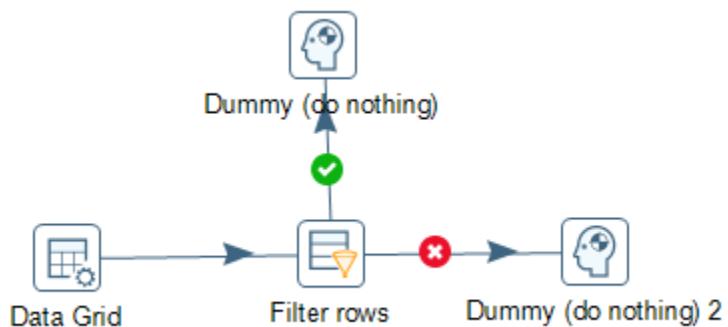
# Value Mapper

- Realiza o mapeamento entre campos



# Filter Rows

- Permite filtrar linhas com base em condições e comparações. Uma vez que este passo está ligado a uma etapa anterior (um ou mais e receber entrada), você pode clicar no "<field>", "=" e "<value>" áreas para a construção de uma condição.

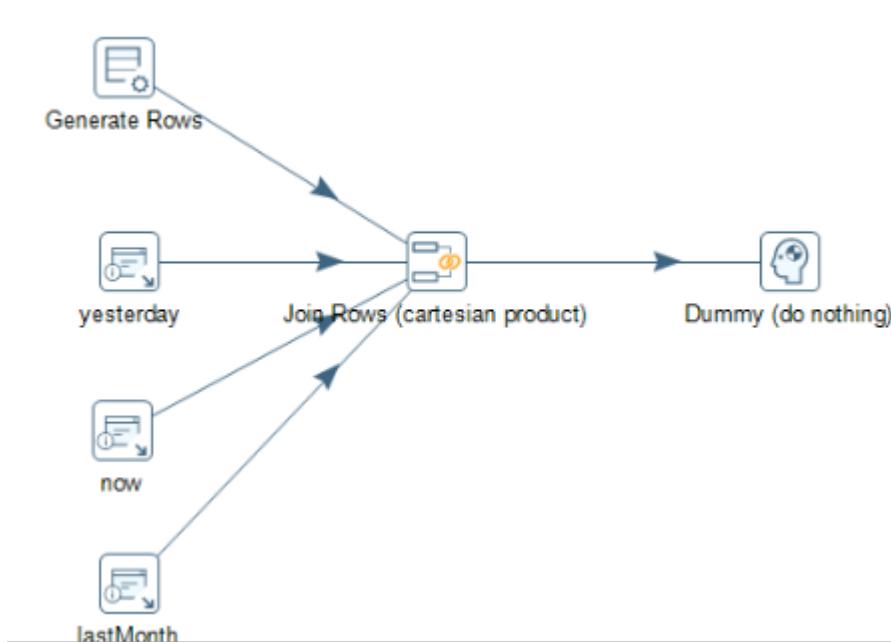




Join Rows (cartesian product)

# Join Rows

- O step Join Rows permite produzir combinações (produto cartesiano) de todas as linhas da entrada de fluxos, como mostrado abaixo:





Merge Join

# Merge Join

O step Merge Join executa uma clássica junção entre conjuntos de dados com dados provenientes de dois steps de entrada distintos.

As opções de junção incluem:

INNER;

LEFT OUTER;

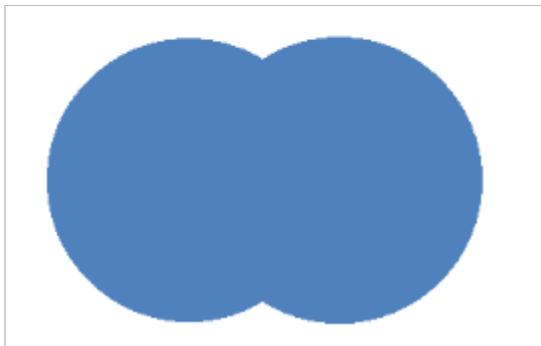
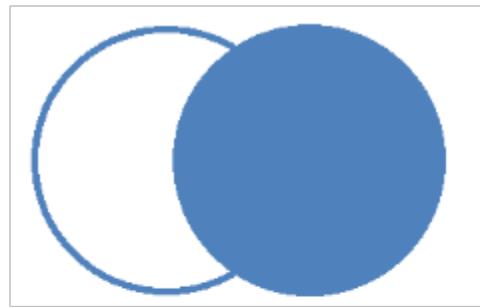
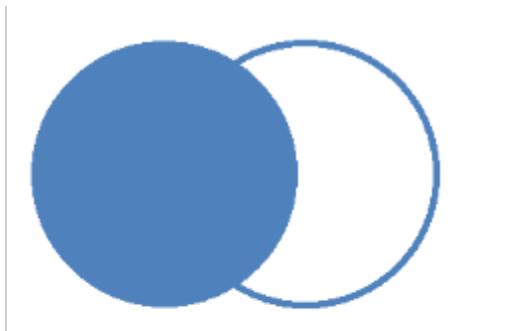
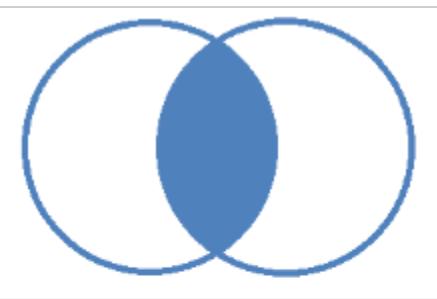
RIGHT OUTER;

FULL OUTER.



Merge Join

# Merge Join



# Criando uma Dimensão Tempo

# Jobs e Transformações

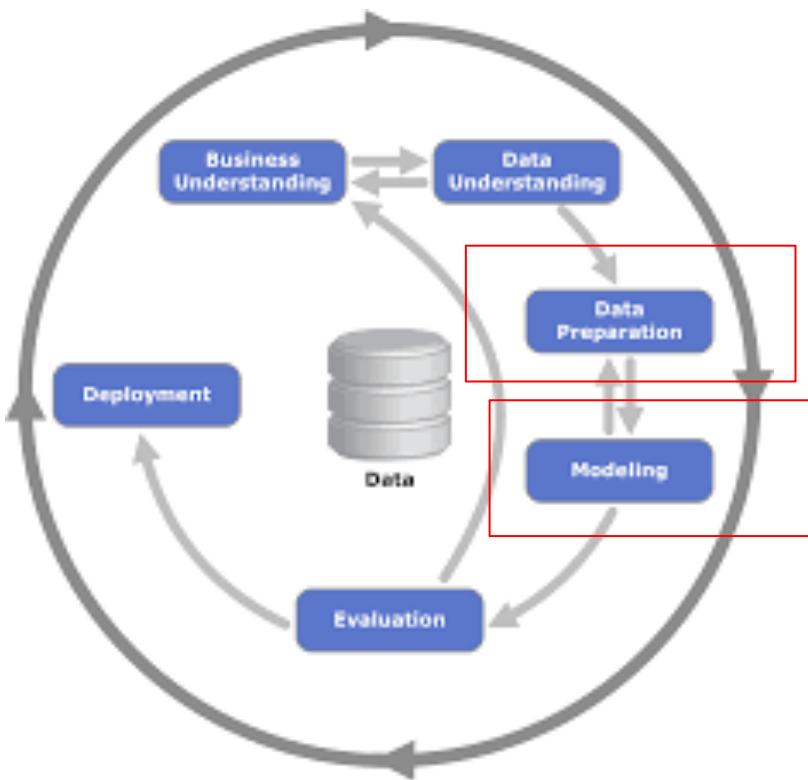
# Semana 2

# Avaliação Atividade

# CRISP-DM



# CRISP-DM + PDI



# Data Preparation

Visa a preparação dos dados, que geralmente não estão dispostos em formato adequado, para a aplicação dos algoritmos de descoberta, análise e extração do conhecimento.

1. As grandes bases de dados são altamente susceptíveis a ruídos, valores faltantes e inconsistência
2. Dados limpos e consistente são requisitos básicos para sucesso da mineração de dados.
3. Esse processo tem como objetivo assegurar a qualidade dos dados.

# Data Preparation

Conhecimento sobre o domínio auxilia em todas as etapas do processo de KDD. **D3M** (Domain Driven Data Mining)

1. Seleção de Dados
2. Limpeza
3. Transformação
4. Integração de dados
5. Formatação sintática

# Data Preparation - Seleção dos Dados

1. Simplicidade do modelo gerado
2. Relevância dos atributos
3. Redundância entre atributos
4. Aumento da Acurácia
5. Principais formas:
  - a. Segmentação dos dados
  - b. Eliminação Direta
  - c. Amostragem aleatória
  - d. Agregação

# Data Preparation - Limpeza

- Remoção de Ruídos
- Atributos incompletos ou sem informação
- Principais formas:
  - Exclusão de casos
  - Preenchimento de valores
  - Preenchimento com valores globais constantes
  - Preenchimento com medidas estatísticas
  - Preenchimento com métodos de mineração de dados

# Data Preparation - Transformação de dados

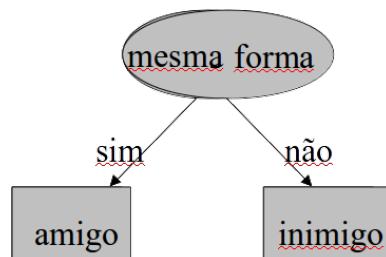
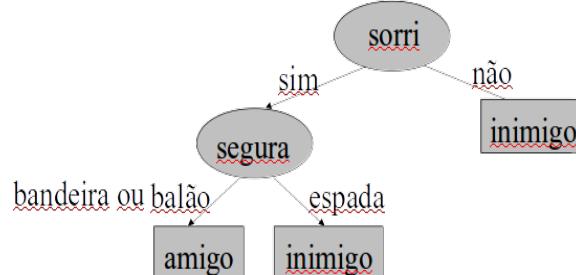
- Generalização
- Normalização
- Transformação Númerico para categórico
- Transformação Categórico para Númerico

# Data Preparation - Construção de dados

- Criação de novos atributos

<u>Cabeça</u>	<u>Corpo</u>	<u>Sorri</u>	<u>Segura</u>	<u>Classe</u>
Triangular	Triangular	Sim	Balão	Amigo
Quadrada	Quadrado	Sim	Balão	Amigo
Redonda	Redondo	Sim	Bandeira	Amigo
Quadrada	Triangular	Não	Espada	Inimigo
Triangular	Redondo	Sim	Espada	Inimigo
Redonda	Quadrado	Não	Bandeira	Inimigo

<u>Cabeça</u>	<u>Corpo</u>	<u>Sorri</u>	<u>Segura</u>	<u>Mesma forma</u>	<u>Classe</u>
Triangular	Triangular	Sim	Balão	Sim	Amigo
Quadrada	Quadrado	Sim	Balão	Sim	Amigo
Redonda	Redondo	Sim	Bandeira	Sim	Amigo
Quadrada	Triangular	Não	Espada	Não	Inimigo
Triangular	Redondo	Não	Espada	Não	Inimigo
Redonda	Quadrado	Não	Bandeira	Não	Inimigo



# Modeling

- R Script Executor
- Weka Steps
- Ferramentas de Mineração de Dados!
  - Plugins
  - Chamadas Shell
  - Integração no nível de saída.

# Evaluation

- Pós-processamento
- Criação de Data Mart com os Dados (Continuação do ETL pós modelagem)
- Desempenho avaliado diretamente no modelo OLAP, levando ao OLAM (OLAP + DataMining)
- Kolmogorov-Smirnov Curve (KS Test)
- ROC Curve

# Vantagens



# Usando o R com PDI

Plugin instalável no Marketplace: Help > Marketplace

<http://wiki.pentaho.com/display/EAI/R+script+executor>

<http://dekarlab.de/wp/?p=5>

# Usando o R com PDI

Baixar arquivos.

# Pentaho Server

# Projeto