

Professor:
Fabiano Tavares



Extração, Transformação e
Carga de Dados

ANO: 2019

Sobre a disciplina

Extração, Transformação e Carga de Dados

Cronograma

Dia 23/11

- Conceitos
- Processo de ETL
- PDI
- Extração

Dia 30/11

- Extração
- Transformação
- Carga

Dia 07/12

- Carga
- Questões

Onde estamos ?

- ❖ Processo de Ciência dos Dados



Onde queremos chegar?

Extração, Transformação e Carga

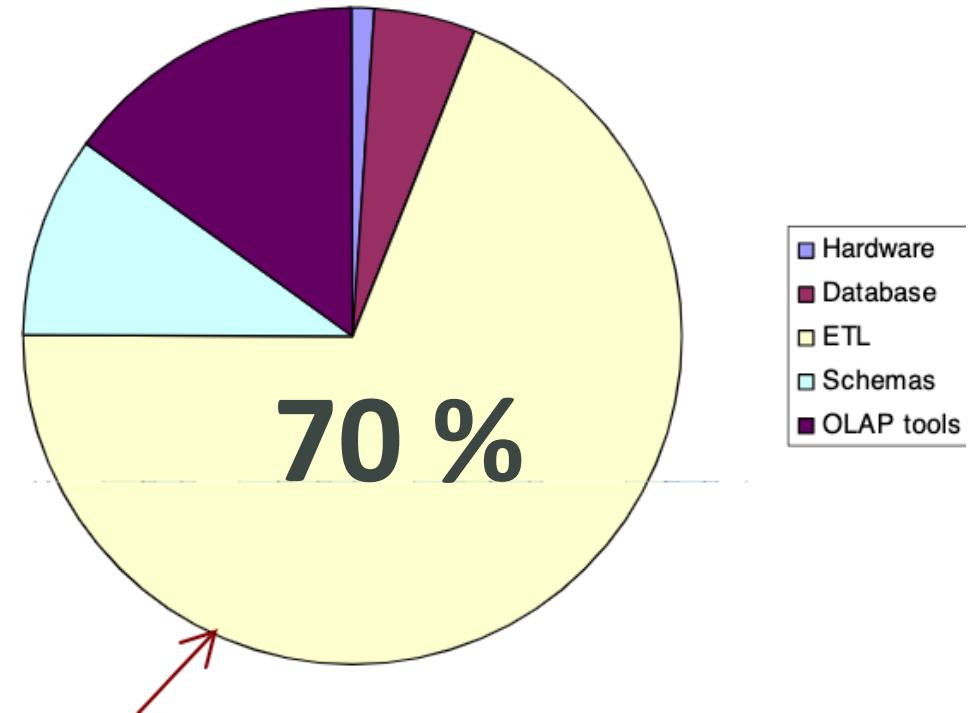


Objetivos

- ❖ Compreender os conceitos de básicos que envolvem Extração, Transformação e Carga (ETL) de dados;
- ❖ Assimilar as técnicas e estratégias de implementação do processo de ETL;
- ❖ Aplicar conhecimentos relevantes através de ferramentas.

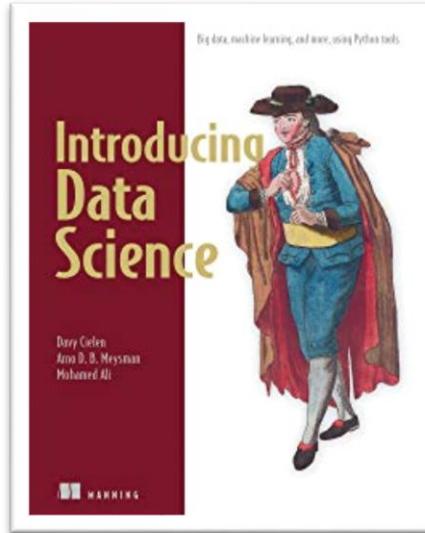
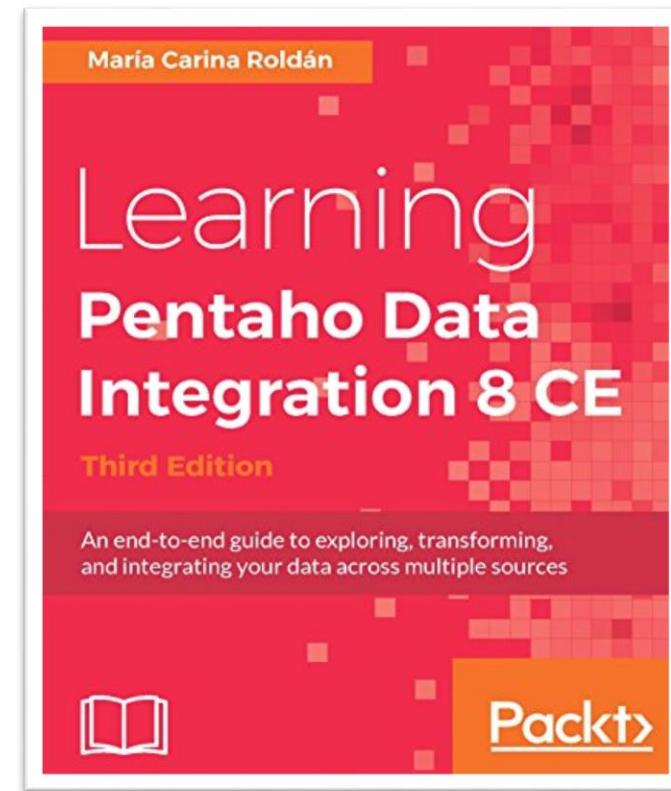
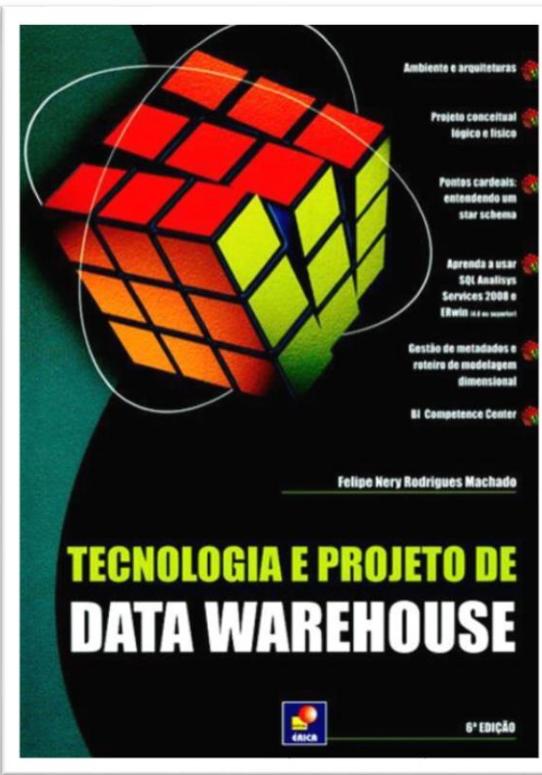
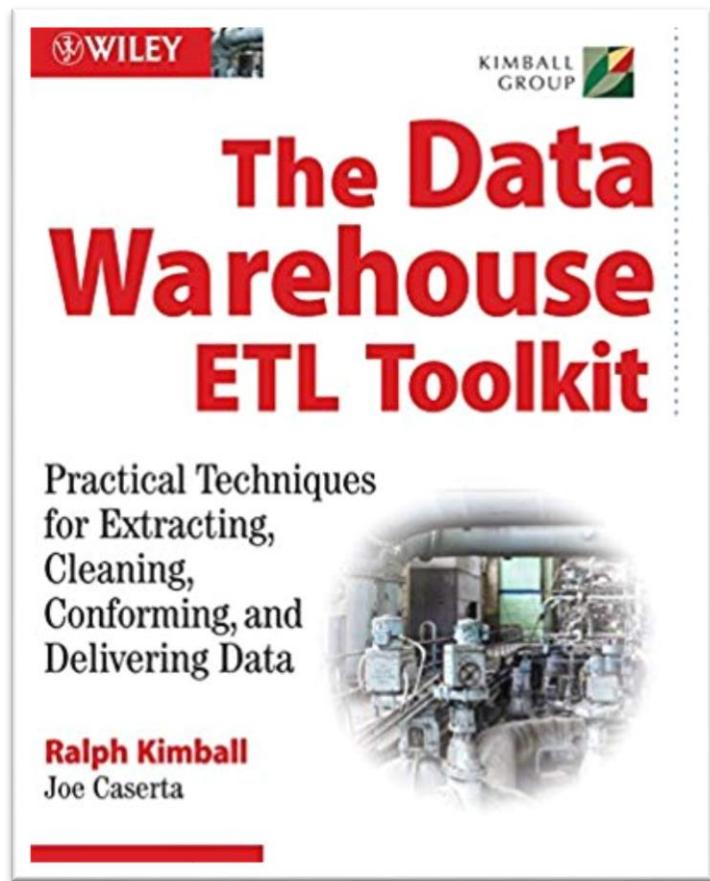
Motivação

- ❖ Dados -> Conhecimento;
- ❖ Processo base fundamental para as etapas de tratamentos, modelagem e análise do dados.



Segundo livro de Kimball & Caserta , 70% do tempo e esforço para construir um armazém de dados (data warehouse)

Bibliografia



Google

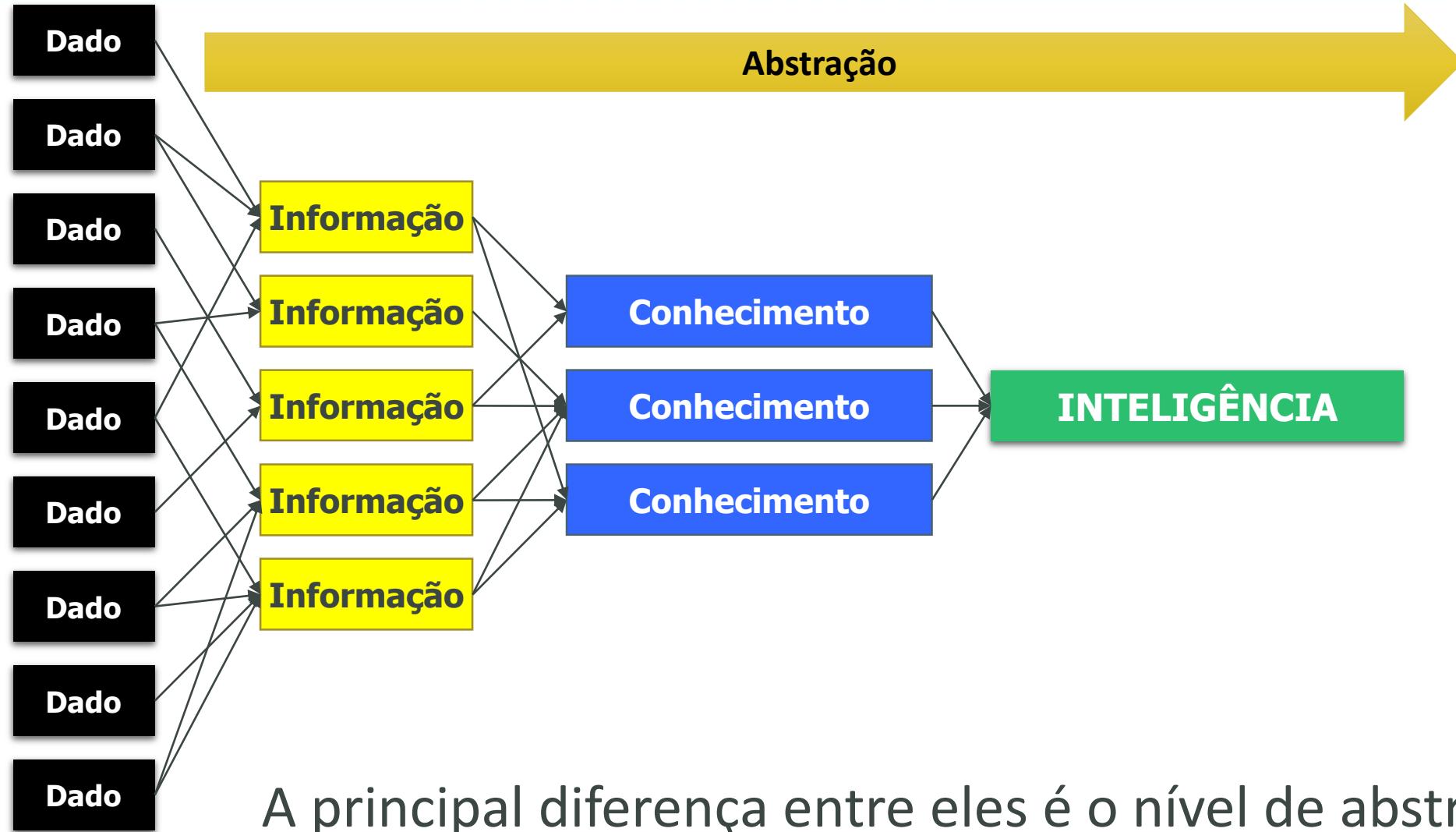
Conceitos Fundamentais

Dados, BI, DW, ID, ETL

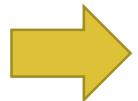
Dados

- ❖ Fatos **registrados**, e que têm um significado implícito, sobre fenômenos do mundo real;
- ❖ Tipicamente representam valores (números, caracteres) de variáveis (qualitativas ou quantitativas);
- ❖ Utilizados para transmitir, armazenar e deduzir informações.

Dados / Informação / Conhecimento



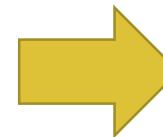
Exemplo clássico



Pingo



Vai chover !



Vamos nos molhar e talvez
não consigamos chegar a
tempo num compromisso

DADOS	INFORMAÇÃO	CONHECIMENTO
Simples observação sobre o estado do mundo.	Dados dotados de relevância e propósito.	Informação valiosa da mente humana. Inclui reflexão, síntese e contexto.
Facilmente estruturados; Facilmente obtidos por máquinas; Frequentemente quantificados; Facilmente transferíveis.	Requer unidade de análise; Exige consenso em relação ao significado; Exige necessariamente a mediação humana;	De difícil estruturação; De difícil captura em máquinas; Frequentemente tácito; De difícil transferência.
“Valor da venda” “Pressão arterial”	“Crescimento do Mercado” “Possível infarto”	“Um estudo do ciclo de vida do produto” “Análise de fatores genéticos cardiovasculares”

Dados

- ❖ Podem ter diferentes formatos:
 - Dados **estruturados** (ex. bancos de dados)
 - Dados **semiestruturados** (ex. xml, json)
 - Dados **não estruturados** (ex. documentos texto, imagens, áudio)

Dados estruturados

- ❖ Organizados e representados com uma estrutura rígida;
- ❖ Previamente planejados para armazenamento;
- ❖ Exemplos: **Banco de dados e formulários de cadastro;**
- ❖ Todos os dados de acordo com um esquema;
- ❖ Usuários precisam do esquema para formular consultas e atualizações.

modelo	cor	ano	valor	codMarca
Uno	azul	1997	14000	1
Palio	preto	1999	16000	1
Palio	azul	2008	32000	1
Palio	branco	2002	22000	1
Fiesta	branco	2000	18000	2
Ka	azul	2005	22500	2
Fiesta	azul	2007	28000	2
Monza	preto	1995	15000	3
Vectra	prata	2006	35000	3
Vectra	azul	1999	29000	3
Monza	prata	1997	12000	3
Gol	prata	1997	16000	4
Fox	preto	2002	31000	4
Gol	azul	1999	25000	4
Fox	prata	2008	39000	4
Gol	vermelho	2005	16000	4
Corola	prata	2004	60000	5
Renault 206	vermelho	2003	25000	7

Dados semiestruturados

- ❖ Organizados e representados com uma estrutura heterogênea e flexível;
- ❖ Dados podem ser incompletos e irregular;
- ❖ Tipos são apenas indicativo;
- ❖ Exemplos: XML, JSON, RDF, OWL;

```
<?xml version="1.0" encoding="utf-8"?>
<livraria>
  <livro id="L01" ano="1936">
    <autor> Jorge Amado </autor>
    <titulo>Mar Morto</titulo>
  </livro>
  <livro id="L04" ano="1930">
    <autor>
      <nome>Rachel</nome >
      <sobrenome>de Queiroz</sobrenome >
    </autor>
    <titulo>O Quinze</titulo>
    <genero> Romance </genero>
  </livro>
</livraria>
```

Dados não estruturados

- ❖ Sem estrutura (ou mínima);
- ❖ Textos, documentos, áudio, páginas de internet, emails, postagens em rede social e etc.;
- ❖ **80 a 90%** dos dados existentes no mundo.



Modelagem dos Dados

- ❖ É criar uma representação **que explique características de funcionamento e comportamento** através de estrutura de dados que **darão suporte ao processo de negócio**;
- ❖ O objetivo do **modelo de dados** é descrever os **conceitos** relevantes para um **domínio**, os **relacionamentos** entre esses conceitos e as **informações associadas** aos mesmos;
- ❖ Um modelo de dados normalmente assume a **forma de um diagrama**, apoiado por descrições textuais;
- ❖ Elementos do modelo de dados:

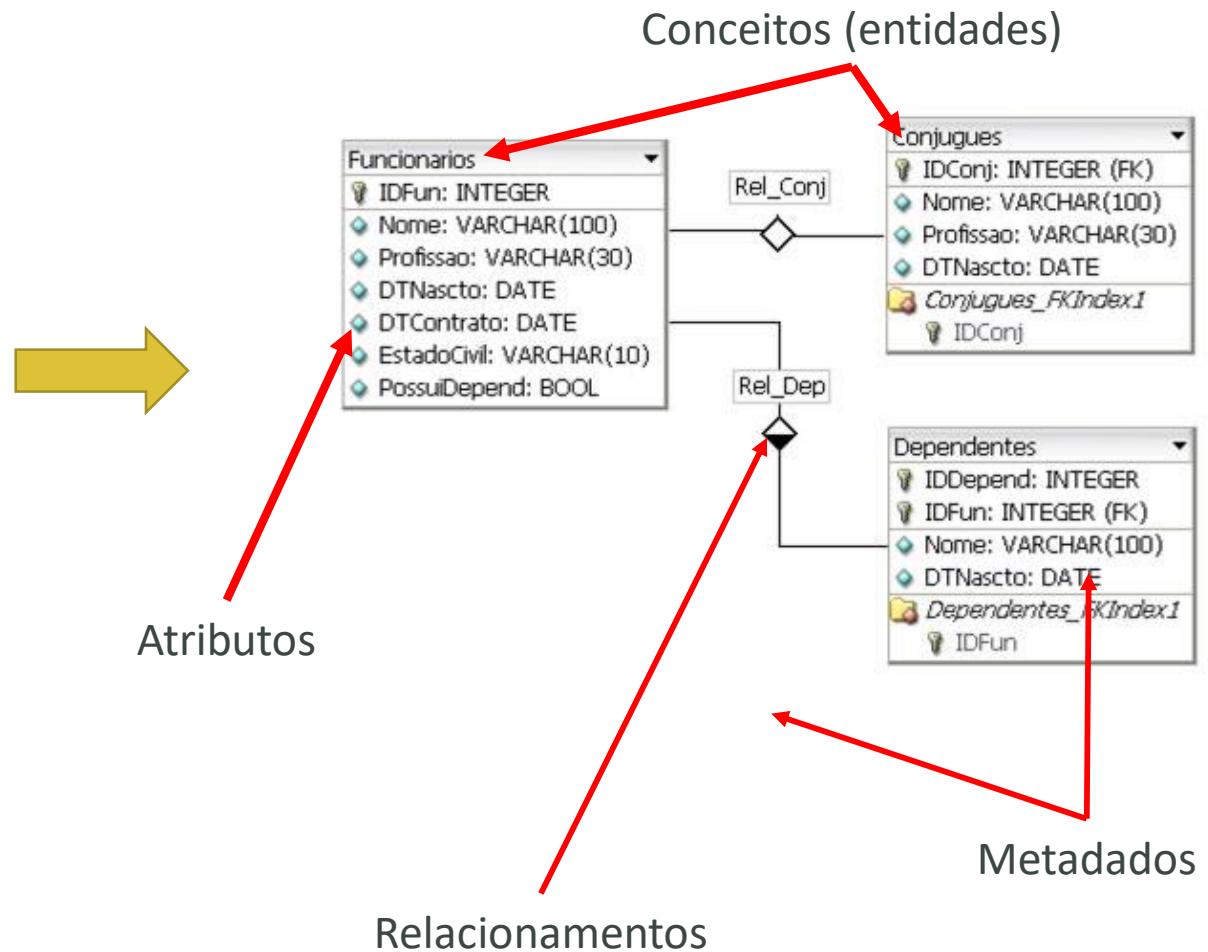


Modelo de Dados

Empresa das Nações • Rua Brasil, 2005 15º andar • Ficha Cadastral

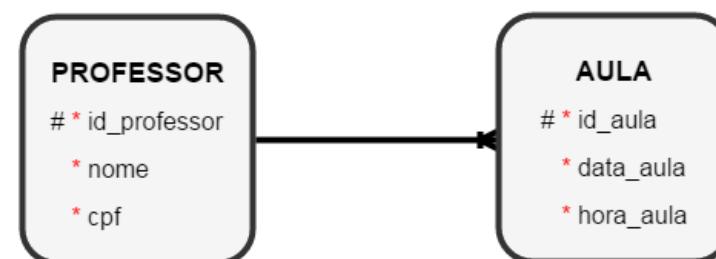
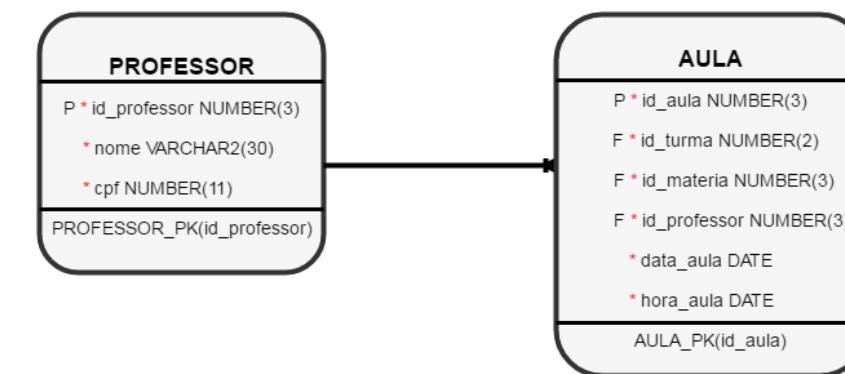
Nome: Coelho da Silva	Profissão: Assistente de Produção
Data de nascimento: 25/01/1979	Data da contratação: 25/01/2000
Estado Civil: () solteiro (X) casado () divorciado	
Nome do Cônjuge: Pascoalina Coelha da Silva	
Dependentes: () não (X) sim	
Nome: Pedro Coelho da Silva	
Nome: Pedrita Coelho da Silva	
Nome: ----	

Ficha de cadastro



Modelagem dos Dados

❖ Tipos de modelos de dados:

Modelo Conceitual	Modelo Lógico	Modelo Físico
Representação próxima do mundo real	Representação do modelo do banco de dados	Organização dos dados em disco
Ex: Modelo Entidade Relacionamento 	Ex: Diagrama Entidade Relacionamento 	Ex: Modelo Físico de ER 

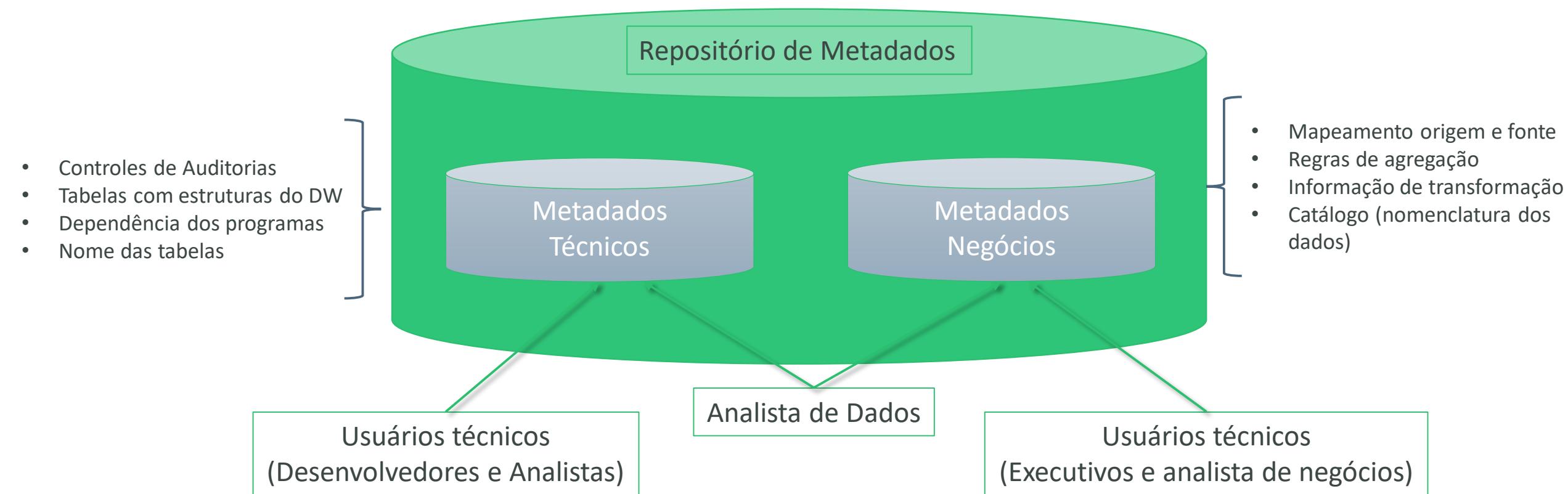
Metadados

- ❖ **Metainformação:** são dados sobre outros dados;
- ❖ Informações estruturadas que descrevem, expliquem, localizem ou tornem mais fácil recuperar, usar ou gerenciar um recurso de informações;
- ❖ Dizem do que se trata a informação (o computador comprehende);
- ❖ Facilitam entendimento dos relacionamentos e utilidade das informações;
- ❖ Organizações utilizam para conhecer e manter seus dados;
- ❖ Planilhas, documentos, diagramas, classes, tabelas sobre tabelas e Sistemas.
- ❖ Tipos: Descritivo; Estrutural; Administrativo; Referência e Estatísticos.

Metadados e ETL

- ❖ A manipulação com grandes volumes de conceitos/atributos de dados:
 - Organização do dados;
 - Comunicação efetiva entre profissionais;
- ❖ Trazem ao processo de ETL:
 - Localização;
 - Restrições para ser manipulado;
 - Interações;
 - Impactos da informação;
 - Riscos;

Metadados: Tipos para DW



Armazenamento dos dados

❖ Banco de Dados

- SGBDs – Relacional, Hierárquico, Orientados a objetos e em Rede;
- Banco orientados a documentos (NoSQL).

❖ Basicamente são dois **modelos de acesso/processamento** de dados em banco de dados:

- **OLTP** (*Online Transaction Processing* ou Processamento de Transações On-line)
 - São sistemas que se encarregam de registrar todas as transações contidas em uma determinada operação organizacional.
- **OLAP** (*Online Analytical Processing* ou Processamento Analítico On-line)
 - São sistemas capazes de manipular e analisar um grande volume de dados sob múltiplas perspectivas.

Propósitos da Informação

Manter registro das operações

Suportar tomadas de decisão



Banco de Dados Operacional



Banco de Dados Analítico



- Foco em atualizar
- Alto volume de transações
- Muitas tabelas
- Dados em detalhes
- Linhas e colunas
- Baixo desempenhos das consultas
- Volátil
- Modelagem ramificada

- Em leitura e manter histórico
- Muitos dados
- Poucas tabelas
- Dados agrupados
- medidas, dimensões e fatos
- Alto desempenhos das consultas
- Não Volátil
- Modelagem simplificada
- Histórico dos dados

Arquiteturas OLAP

- ❖ **ROLAP** (Relational On Line Analytical Processing)
A consulta é enviada ao servidor de banco de dados relacional e processada no mesmo, mantendo o cubo no Servidor.
- ❖ **MOLAP** (Multidimensional On Line Analytical Processing)
processamento realizado em um servidor multidimensional.
- ❖ **HOLAP** (Hybrid On Line Analytical Processing)
é uma mistura de tecnologias onde há uma combinação entre ROLAP e MOLAP.

Propósitos da Informação

❖ Manter o registro operacional da organização

- Usuário: **Operacional**
 - Registrar novos clientes
 - Registrar pedidos



❖ Suportar tomadas de decisão

- Usuário: **Gestor**
 - Verificar estoque para decidir compras
 - Visualizar vendas para projetar metas



Problemas Relatados por Gestores

- ❖ “Nós coletamos muitos dados, mas não conseguimos *acessá-los*”;
- ❖ “Me mostre *apenas* o que é importante”;
- ❖ “Precisamos buscar dados *facilmente*”;
- ❖ “Passamos reuniões inteiras discutindo quem tem os dados corretos e não *tomamos decisão* alguma”;
- ❖ “Precisamos de decisões baseadas em *fatos*”.



Business Intelligence (BI)

❖ Inteligência de Negócios ?

- É um conceito que define o processo de **coleta, organização, análise, compartilhamento e monitoramento** de informações que permitem oferecer suporte a gestão de negócios;
- BI explica dados de eventos que já ocorreram.

❖ Objetivos:

- Interpretar dados;
- Identificar oportunidades para implementar estratégias;
- Vantagem competitiva nos negócios.

DECISÃO
ANÁLISE
INFORMAÇÃO
DADOS



Componentes de um BI

- ❖ Um sistema padrão de *business intelligence* é, portanto, composto pelos seguintes elementos:
 - **Módulo de ETL** (*extraction, transformation and loading*) – Componente dedicado à extração, transformação e carregamento de dados. É a parte responsável pela coleta das informações nas mais diversas fontes (sistemas ERP, arquivos TXT ou ficheiros Excel);
 - **Data warehouse/Data marts** – Locais onde ficam concentrados todos os dados extraídos dos sistemas operacionais. A grande vantagem de ter um repositório de dados separado consiste na possibilidade de armazenar informações históricas e agregadas, construindo assim um melhor suporte para as análises;
 - **Front-end** – Parte de um projeto de BI visível ao usuário. Pode conter dashboards, relatórios padronizados, consultas *ad hoc*, portal de intranet/Internet/Extranet, análise OLAP e funções diversas como *data mining*.

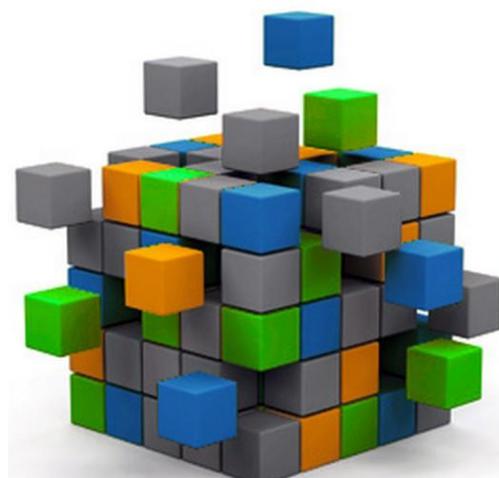
Data Warehouse

❖ Armazém de dados:

- Armazenar informações relativas às atividades de uma organização em bancos de dados de forma **integrada, variável em relação ao tempo e não volátil** para apoiar os processos de tomada de **decisões estratégicas, táticas e também operacionais** de organizações.

❖ O Data Warehouse deve:

- Prover acesso fácil à informação;
- Apresentar informação de forma consistente;
- Apresentar informações tão rápido quanto necessário;
- Ser seguro e proteger a informação;
- Ser autoridade para melhorar o processo de tomada de decisões;



DW: Orientado por Assuntos

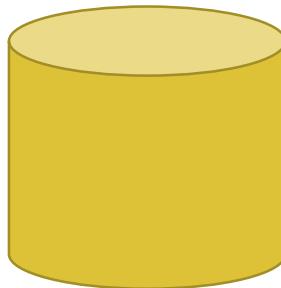
❖ Orientado por assuntos:

- Um DW sempre armazena dados importantes sobre **temas específicos** da empresa e conforme o interesse dos **processos de negócio** envolvidos.

Ambiente Transacional

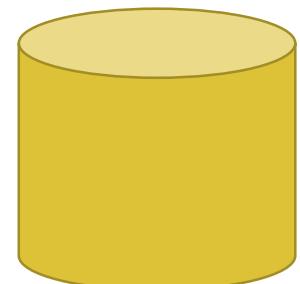


Data Warehouse



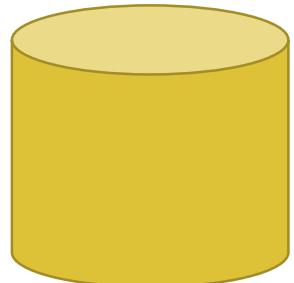
Qualidade

Data Warehouse



Vendas

Data Warehouse



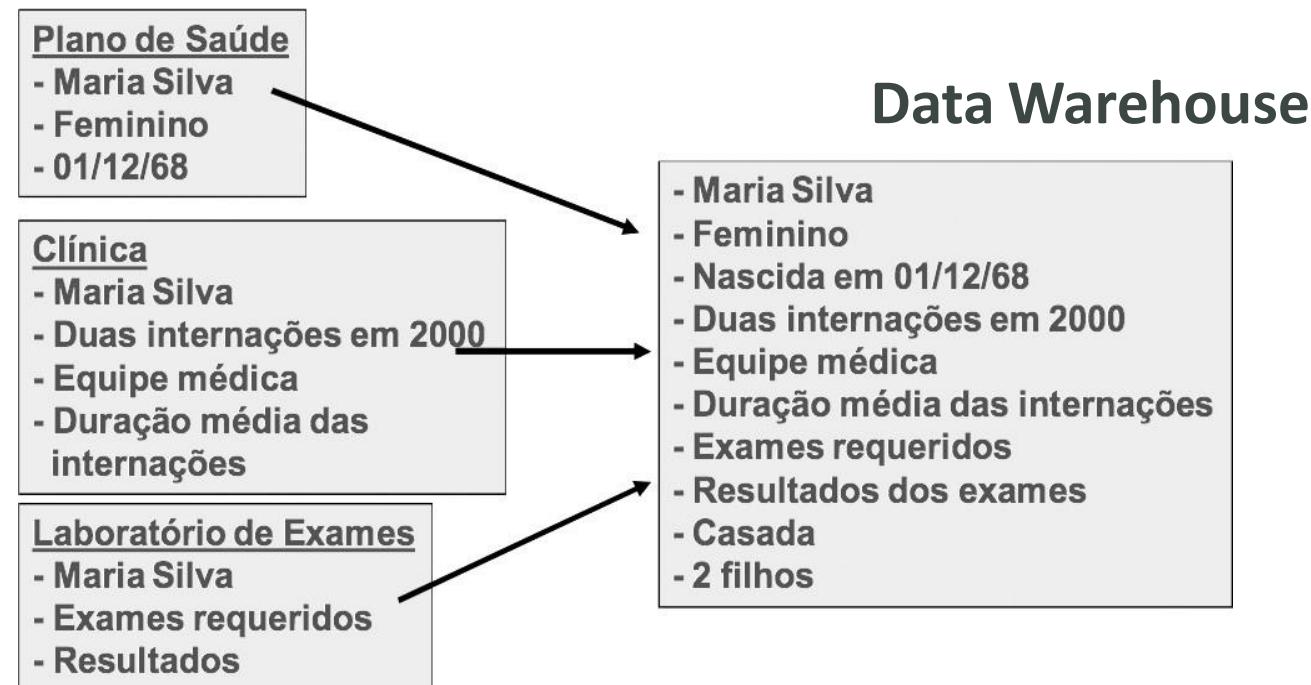
Produção

Pedido, Nota Fiscal, Produto, Clientes

DW: Integração de Dados

- ❖ O termo “integração de dados” refere-se ao processo de **combinar** dados de **diferentes fontes** para prover uma **única visão comprehensível** de todos os dados combinados.

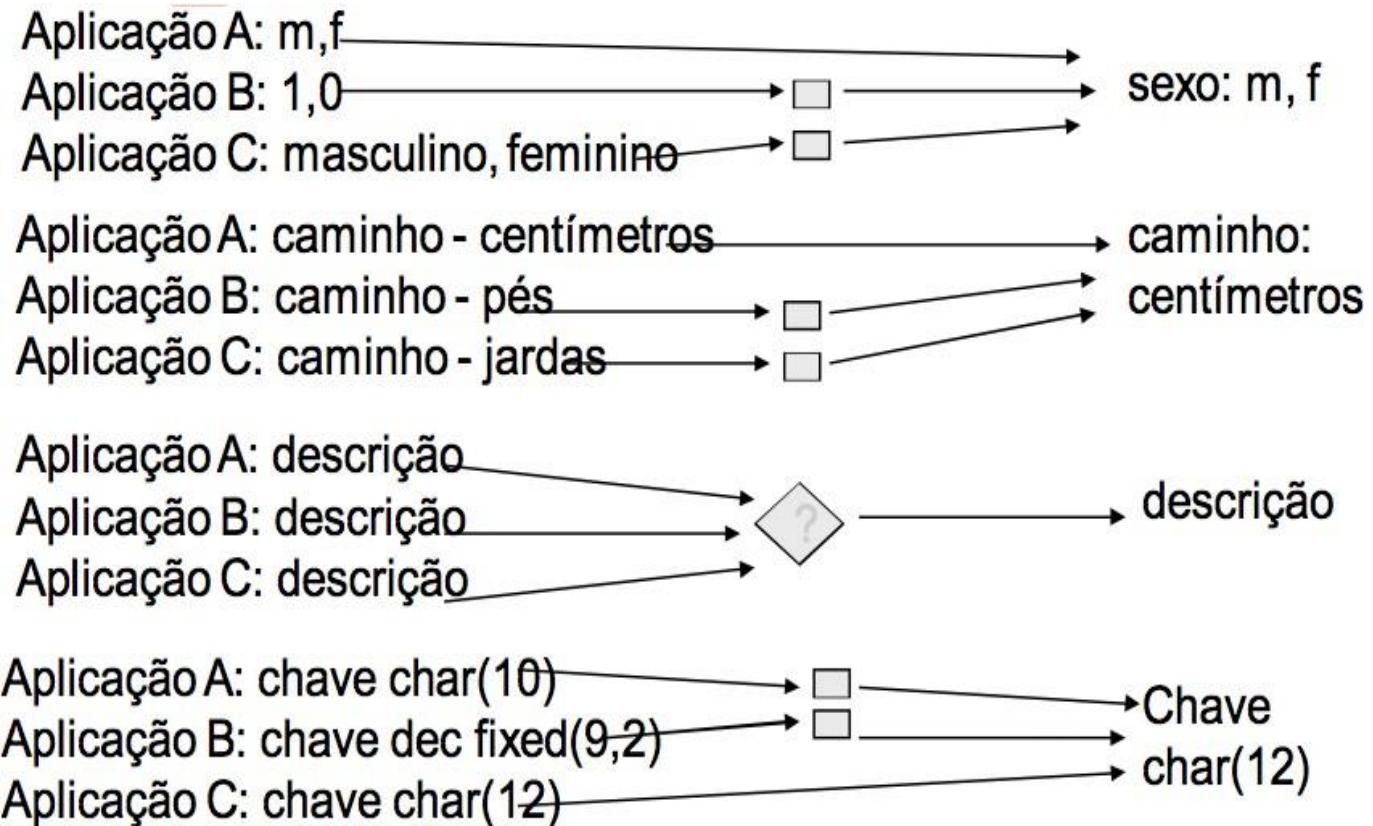
Diferentes Fontes de Dados



DW: Integração de Dados

Diferentes Fontes de Dados

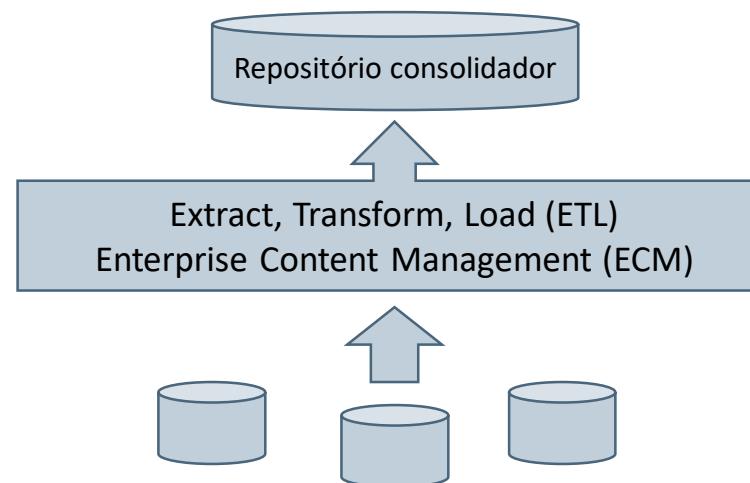
Data Warehouse



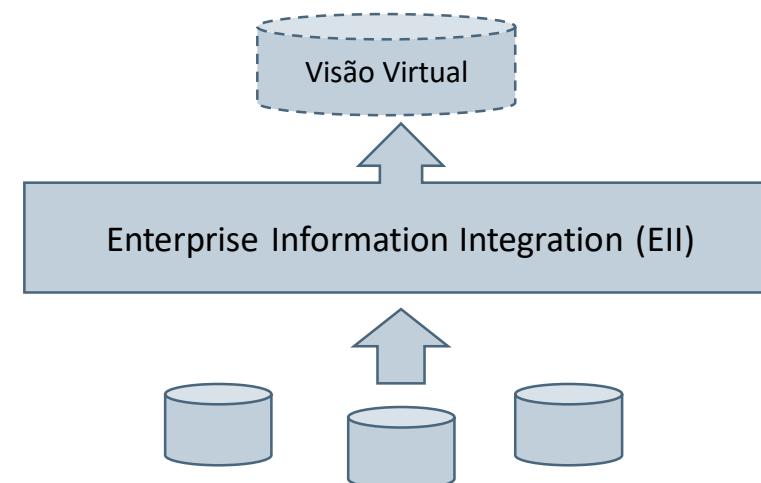
Integração de Dados

- ❖ Integração de Dados e **ETL** são termos usados indistintamente no mercado de **BI**, embora ETL seja apenas um **possível cenário** de integração de dados.

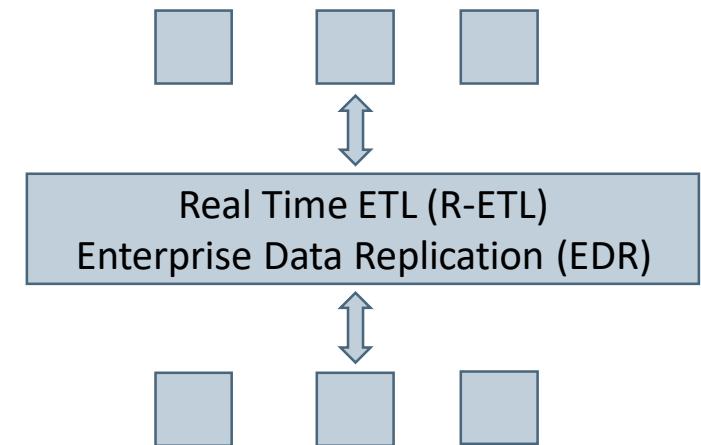
- Consolidação de Dados;



- Federação de Dados;



- Propagação de Dados.



DW: Váriavel em relação ao tempo

- ❖ Banco de dados transacional o dado se refere ao momento atual;
- ❖ Data Warehouse o dado varia em relação ao tempo.

Banco transacional

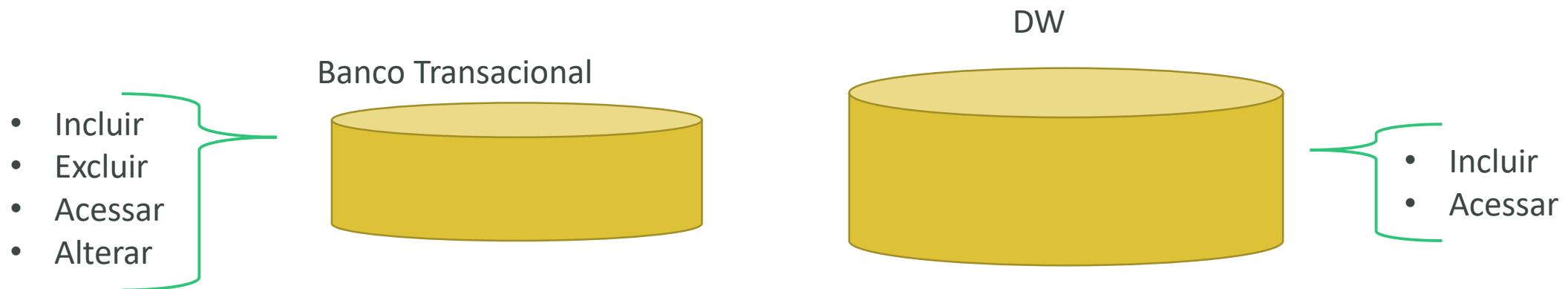
Produto	Preço
<i>Borracha</i>	R\$ 2,00
<i>Caderno</i>	R\$ 8,00
<i>Caneta</i>	R\$ 1,50
<i>Lápis</i>	R\$ 1,00

Data Warehouse

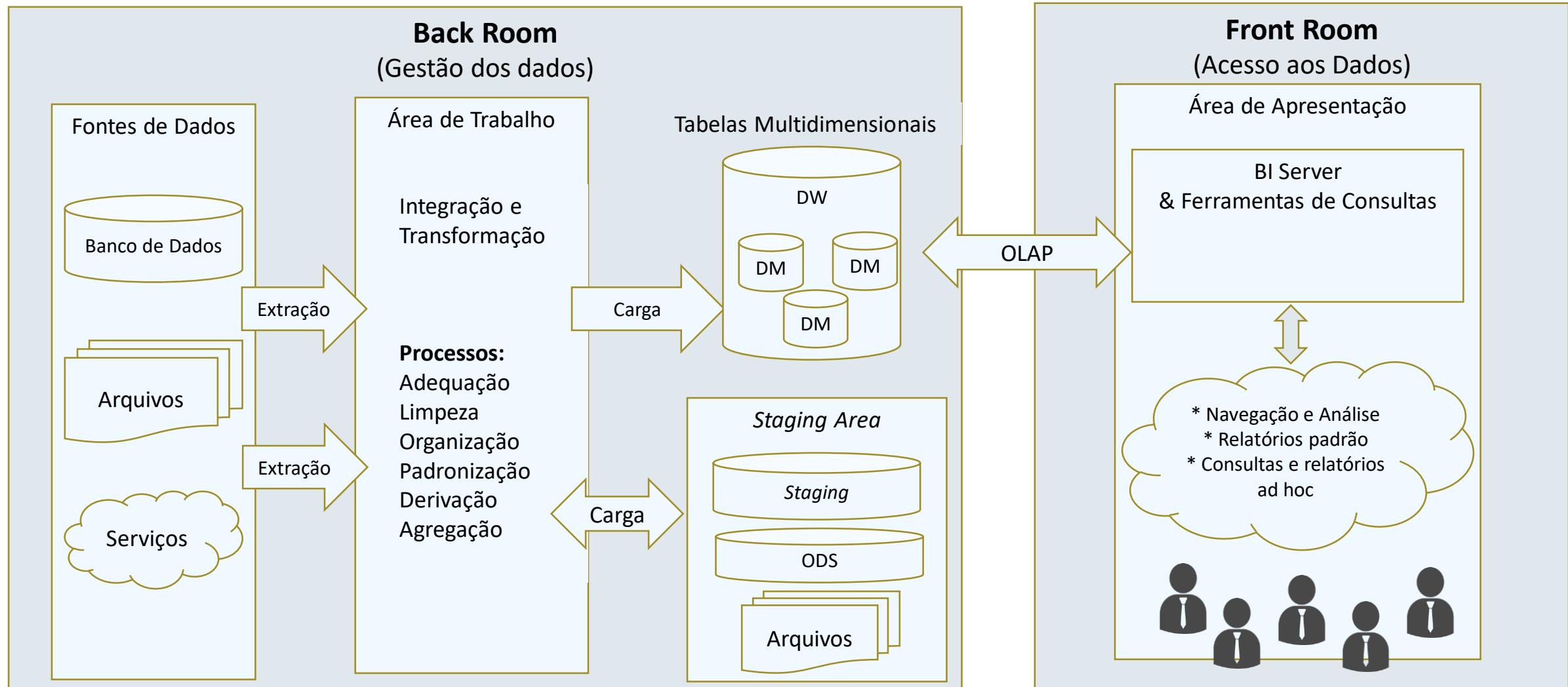
Produto	Jan/2008	Fev/2008	Mar/2008	Abr/2008
<i>Borracha</i>	R\$ 1,50	R\$ 2,00	R\$ 2,00	R\$ 2,00
<i>Caderno</i>	R\$ 8,00	R\$ 8,00	R\$ 8,00	R\$ 8,00
<i>Caneta</i>	R\$ 1,20	R\$ 1,70	R\$ 1,50	R\$ 1,50
<i>Lápis</i>	R\$ 0,85	R\$ 0,85	R\$ 0,75	R\$ 1,00

DW: Não Volátil

- ❖ Nenhum dado pode ser alterado depois de incluído no DW;
- ❖ Duas operações **carga e acesso**;



Arquitetura DW/BI Kimball

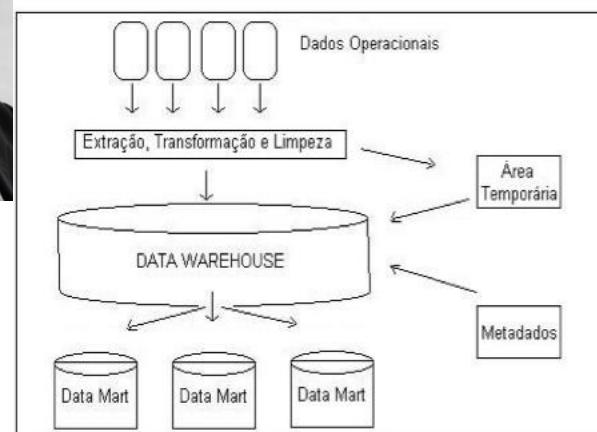


Data Marts (DM)

- ❖ São sub-conjuntos de dados de um Data warehouse;
- ❖ Agrupados por assunto em especial (ex: Vendas, Estoque, Controladoria) ou diferentes níveis de sumarização (ex: Vendas Anual, Vendas Mensal, Vendas 5 anos);
- ❖ Focam uma ou mais áreas específicas;
- ❖ Arquiteturas:

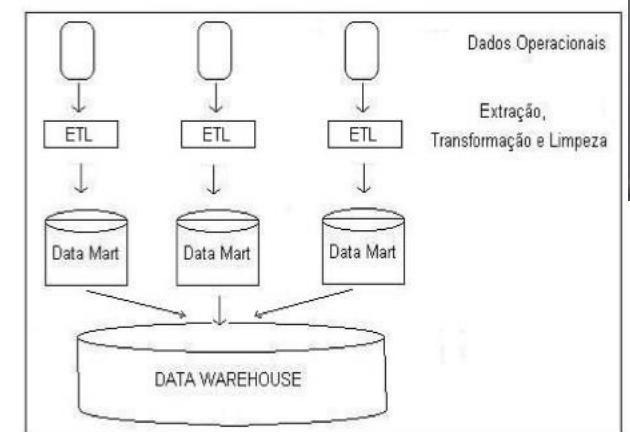


Bill Inmon



x

Ralph Kimball



❖ *Staging Area:*

- Representa uma área de trabalho que recebe as informações do ambiente OLTP e provê informações para o DW.

❖ Operational Data Store (ODS)

- Armazenamento de dados atuais ou quase atuais para suporte à decisões operacionais do dia-a-dia;
- É volátil;

❖ **Extração, Transformação e Carga (ETL)**

❖ Tabelas Multidimensional;

Modelagem Multidimensional

- ❖ Dados são organizados e estruturados em diversas **dimensões** e eventos de negócio (**fatos**) composto de **medidas** atreladas ao contexto.

- ❖ Atende a dois requisitos simultâneos:

- Entregar os dados de forma **compreensível** para os gestores;
- Entregar os dados com **rapidez**;

- ❖ Um dos tipos de modelagem multidimensional mais utilizado, é o **Star Schema** ou Esquema Estrela, desenvolvido por Ralph Kimball, um dos precursores do conceito de data warehouse.

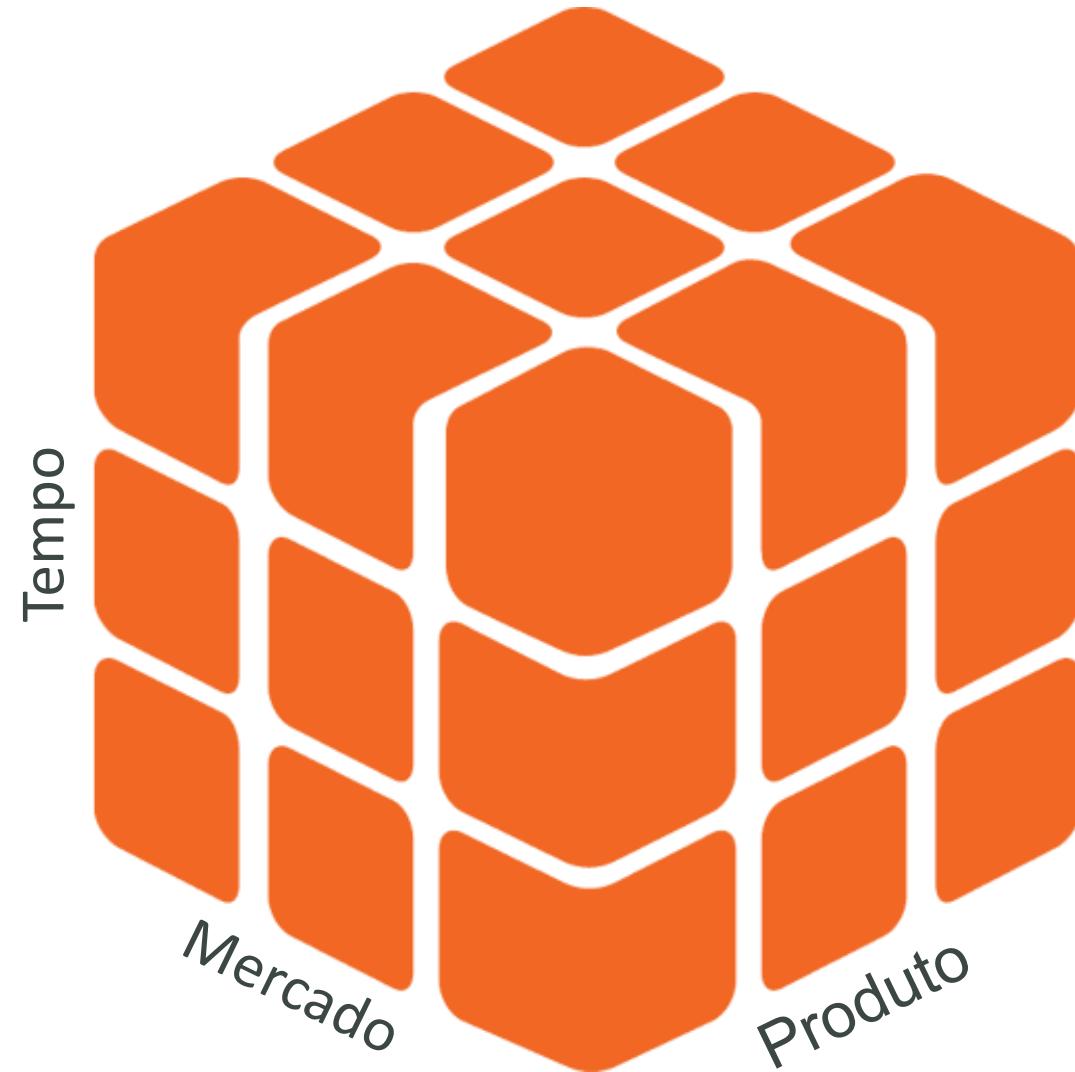
Tabela Fato de Vendas		Dimensão Tempo	
vendas_total	time_id	time_id	timestamp
2008.10	1234	1234	20191123 9:35:43

Rápido Exemplo

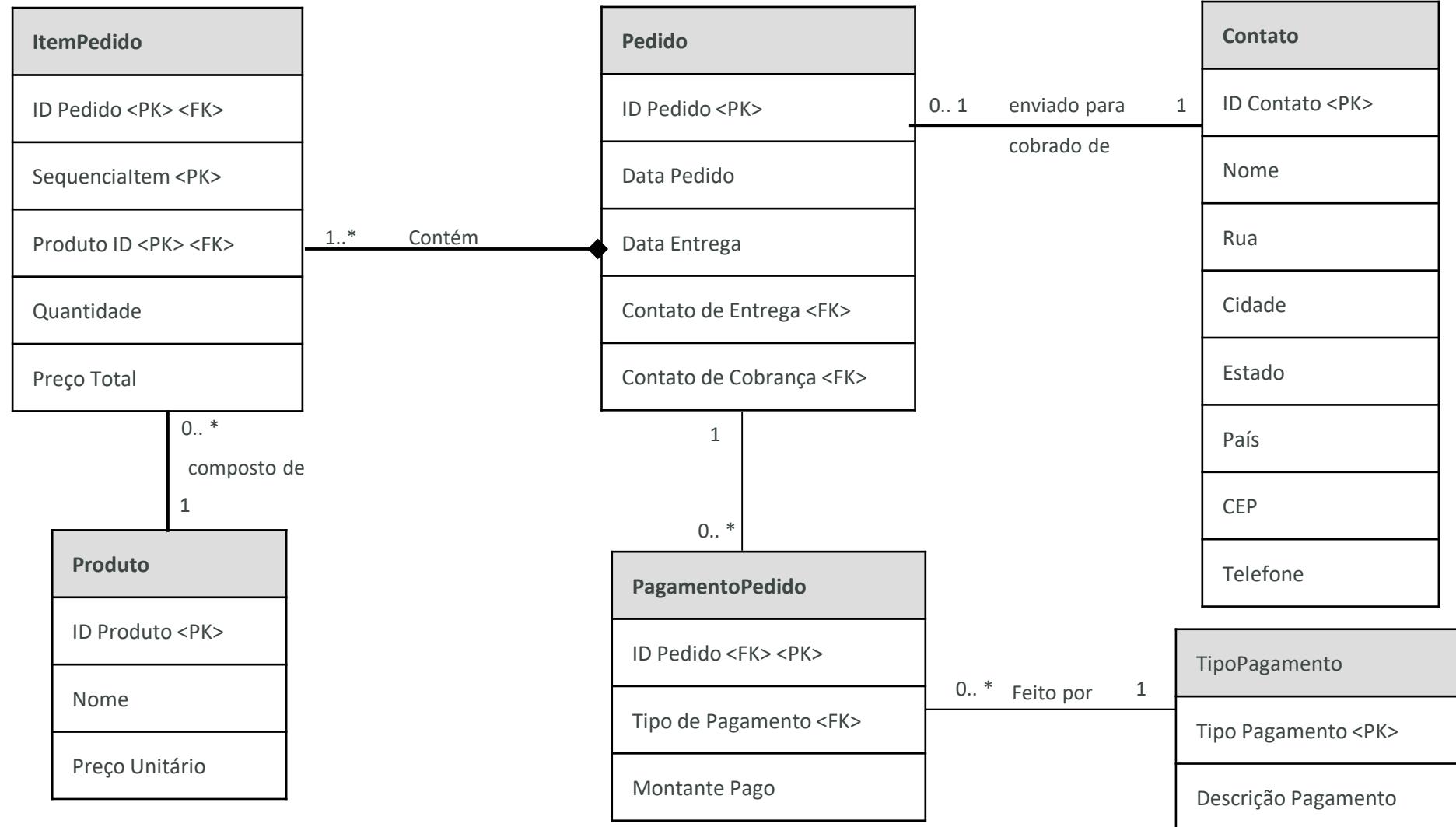
Vendemos os nossos **produtos** em
vários **mercados** e medimos nossa
performance ao longo do **tempo**



Rápido Exemplo

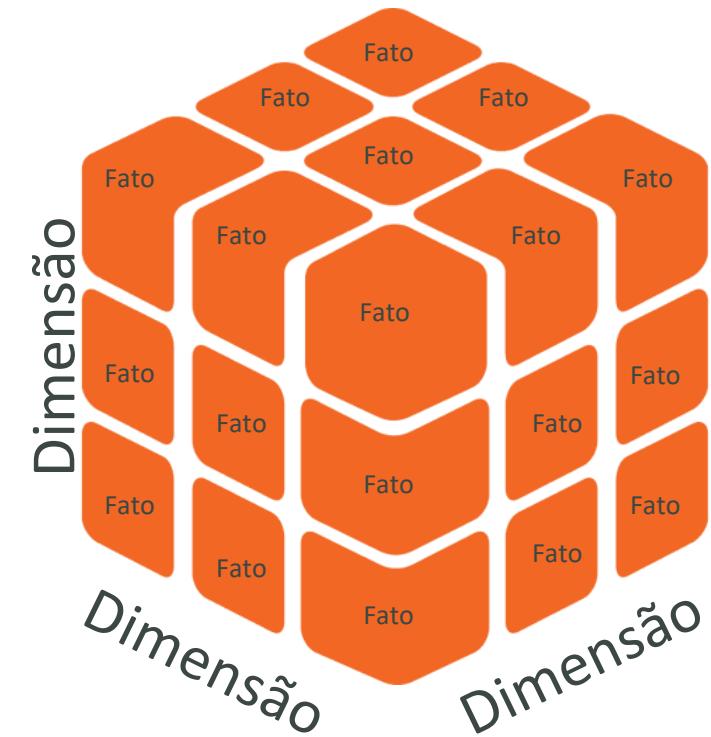
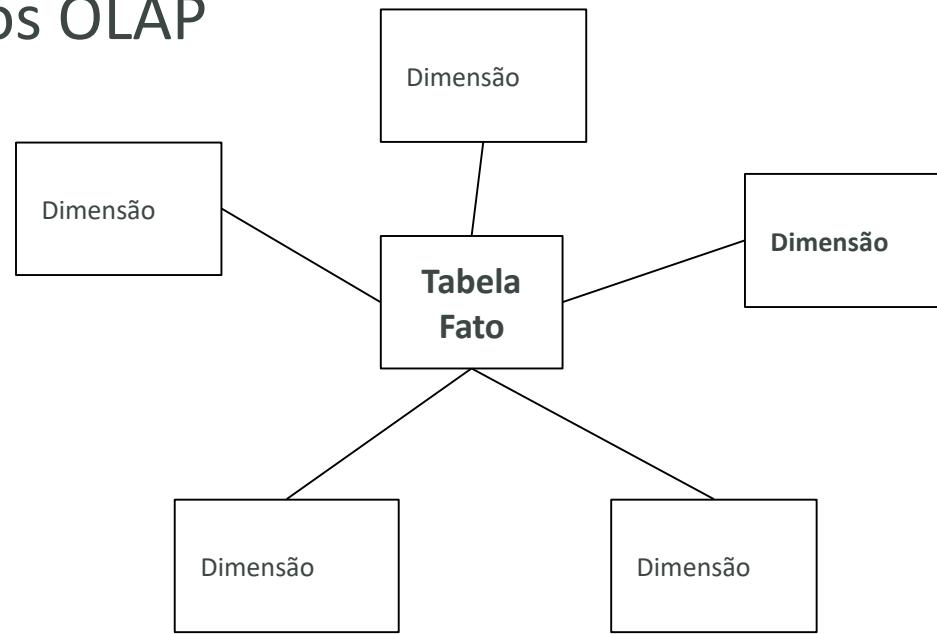


Como fazemos em bancos operacionais



Modelagem Dimensional

- ❖ Para modelagem dimensional em banco de dados relacionais utilizamos o *Star Schema* ou Esquema Estrela.
- ❖ Cubos OLAP



Componentes do Esquema Estrela

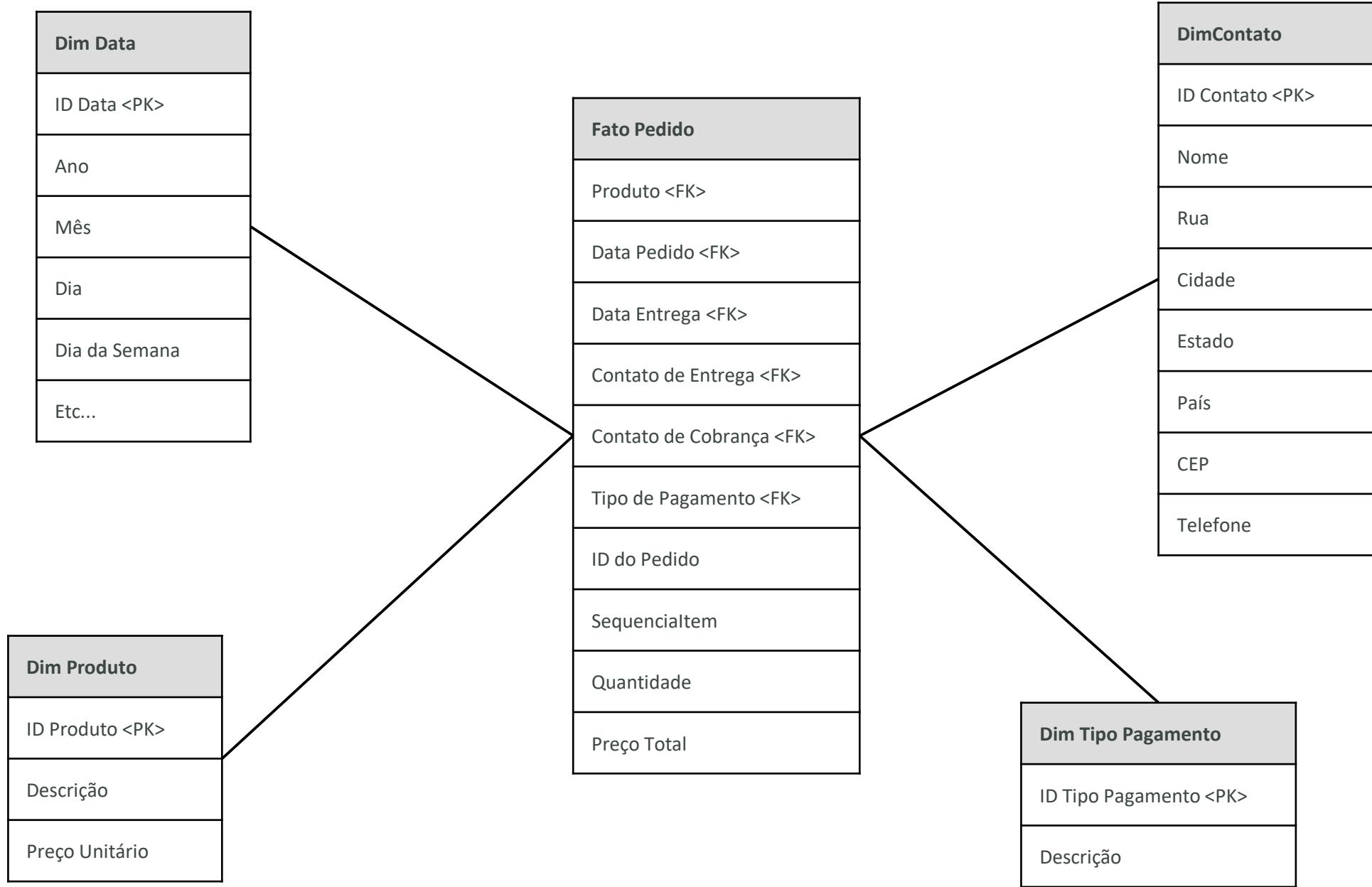
1. Tabela de Fato

- Fato é um **evento** de negócio;
- 1 linha na tabela = 1 evento registrado;
- **Medidas**: algo que pode ser sumarizado/agregado/contado;
- **Granularidade**: nível de um item, uma transação, evento ou agrupado destes;

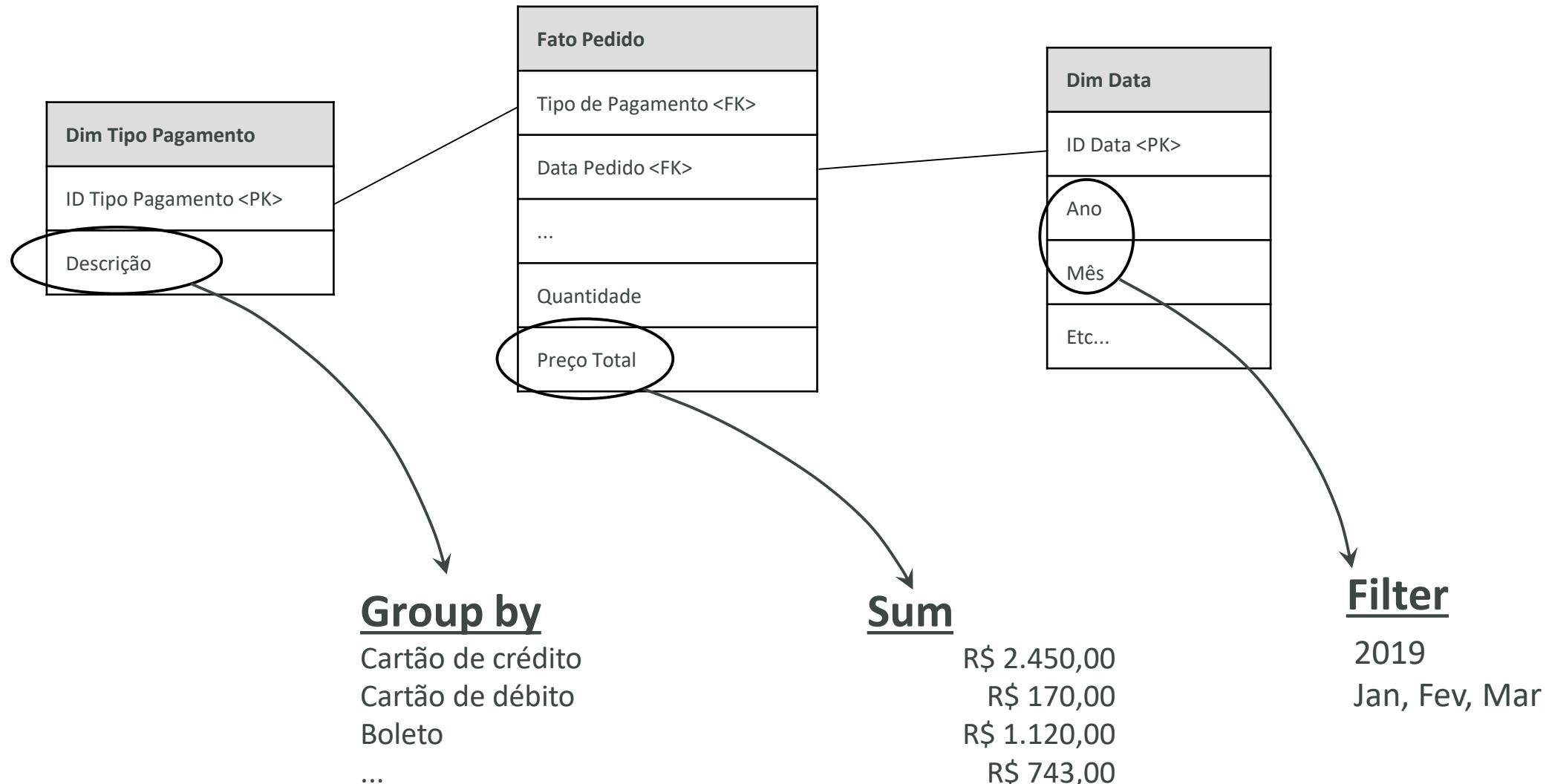
2. Tabela de Dimensão

- O **contexto** associado ao evento de negócio ocorrido;
- Quem, o que, onde, quando, como e porque (os “por”);
- Membros de uma dimensão: Hierarquia
 - Classificação de dados dentro de uma mesma dimensão;

Como fazemos no DW



Acesso aos Dados (OLTP)



Acesso aos Dados (OLAP)

- ❖ **Online Analytical Processing (OLAP)**
- ❖ As **aplicações OLAP** são usadas pelos gestores em qualquer nível da organização para lhes permitir análises comparativas que facilitem a sua tomada de decisões diárias.
- ❖ As possíveis **operações** que podem ser realizadas em um modelo OLAP são:
 - **SLICE**: selecionar dados de uma única dimensão;
 - **DICE**: extrai um subcubo da seleção de duas ou mais dimensões;
 - **DRILL DOWN**: examina dados com maior detalhe;
 - **ROLL UP**: combinação de células de uma ou mais dimensões para atingir um nível maior de generalização;
 - **PIVOT** ou **ROTATION**: visualiza dados por uma nova perspectiva.

Na prática com BI...

- ❖ Descobrir as necessidades de informação: os indicadores de negócio da empresa;
- ❖ Projetar banco de dados para armazenar essas informações;
- ❖ Disponibilizar dados históricos;
- ❖ **ETL** para DW;
- ❖ Disponibilizar dados em ferramentas;

Extração, Transformação e Carga

Histórico, Requisitos e Processo

Processo de ETL

❖ Definição original:

- O conjunto de processos para trazer dados de sistemas **OLTP** para um **data warehouse**.

❖ *Extract, Transform, Load* (Extrair, Transformar, Carregar)

- São soluções em software cuja função é a **extração de dados de diversos sistemas**, **transformação** desses dados conforme regras de negócios e por fim o **carregamento** dos dados geralmente para um **Data Warehouse**.

❖ Aplicações atuais:

- Remover erros e corrigir dados ausentes;
- Prover documentação de medidas para confiança nos dados;
- Captura o fluxo de dados transacionais;
- Ajusta dados de várias fontes para serem usados juntos;
- Estrutura os dados para serem utilizáveis pelas ferramentas do usuário final;



Breve histórico

- ❖ 1ª Geração (década de 90) – Origem do ETL e geradores de códigos:
 - Geravam códigos nativos e então estes eram customizados;
 - Única *thread* (não paralelismo);
 - Desempenho bom, mas única plataforma;
- ❖ 2ª Geração (anos 2000) – Mecanismos de ETL:
 - Engine de execução (programação em ETL);
 - *Pipelining* - processamento é feito linha por linha;
 - Workflows gráfico;
 - Sofre sobrecarga.
- ❖ 3ª Geração (atual) – Arquitetura de ETL:
 - Geram códigos nativos e possuem engines ETL;
 - Processamento paralelo nativo;
 - Ricos em tecnologias e funções de transformação.

Por que usar ETL ?

1. O principal benefício do ETL é que é muito mais **fácil e rápido** de usar do que os métodos tradicionais que movem dados escrevendo **códigos manualmente**.
2. As ferramentas ETL contêm **interfaces gráficas** que aceleram o processo de mapeamento de tabelas e colunas entre os armazenamentos de origem e de destino.

Resiliência operacional

Performance

Perfil e limpeza de dados

Fluxo visual

Reduz complexidade do gerenciamento de dados

Aprimoramento do BI ROI

Desafios

- ❖ Os processos de ETL podem ser bastante complexos e problemas operacionais significativos podem ocorrer com sistemas de ETL desenvolvidos inapropriadamente;
- ❖ Conhecer o perfil dos dados da fonte durante a análise dos dados;
- ❖ A escalabilidade de um sistema de ETL durante o seu ciclo de vida:
 - O tempo disponível para extrair dados dos sistemas de origem pode variar;
 - Processamento em lote;
- ❖ Latência dos dados;

Processamento em Paralelo

Dados

- Pela divisão de um único arquivo sequencial em arquivos de dados menores para permitir acesso em paralelo.

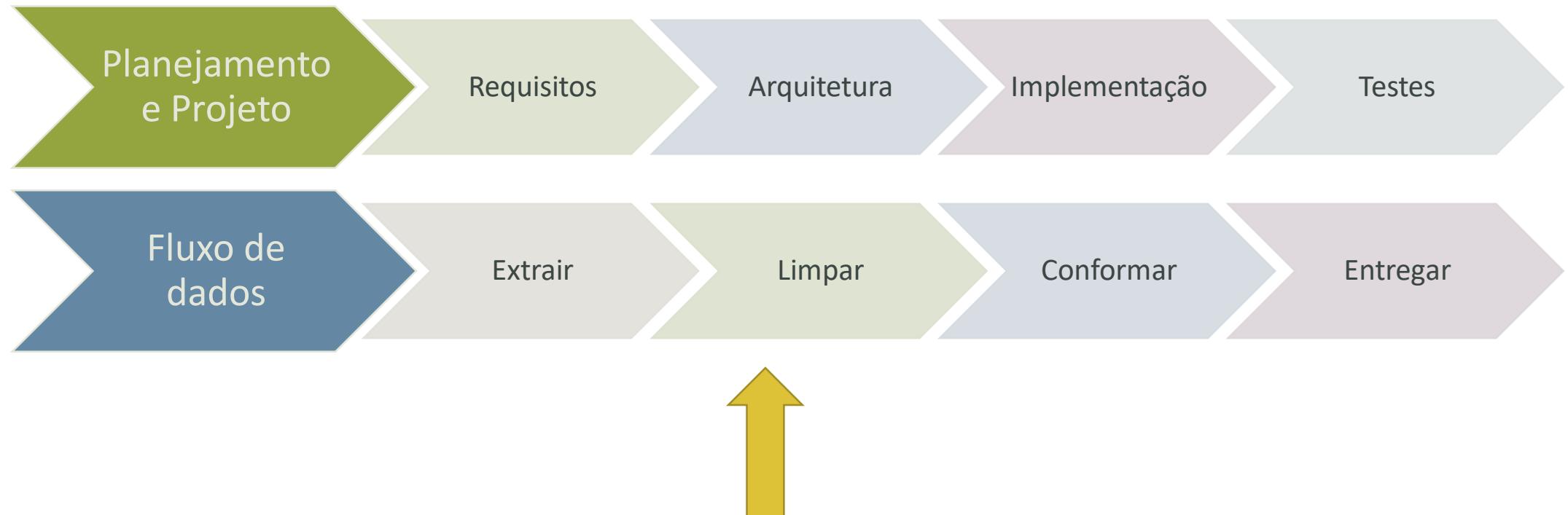
Pipeline

- Permitindo a execução simultânea de diversos componentes no mesmo fluxo de dados. Um exemplo seria a leitura de um valor no registro 1 e ao mesmo tempo juntar dois campos no registro 2.

Componente

- A execução simultânea de múltiplos processos em diferentes fluxos de dados no mesmo job.

Desenvolvimento de ETL



Embora seja conhecido pela sigla ETL, são quatro os macroprocessos, com 34 subsistemas, segundo Kimball
(<https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/>)

Projeto de ETL

- ❖ Qual a melhor maneira de projetar e construir um sistema de ETL? “**Depende**”
 - – das fontes de dados;
 - – das limitações dos dados;
 - – das linguagens de script;
 - – das ferramentas de ETL disponíveis;
 - – das habilidades do pessoal envolvido (TI e negócio);
 - – da plataforma de BI;
 - – Etc.

Requisitos para ETL

Requisitos: elementos os quais deve conviver e adaptar o seu sistema.

- ❖ Certifique-se de saber suas **necessidades antes de começar** o ETL;
- ❖ Reunir num só lugar **todos requisitos conhecidos, realidades e limitações** que podem afetar o sistema de ETL;
- ❖ Seguindo os requisitos, identificamos uma série de **decisões de arquitetura** que você precisa fazer no início do seu projeto de ETL;

Requisitos para ETL

1. Necessidades de Negócios

- Etapa mais importante e fundamental;
- Ter bem claro e documentado quais são os requisitos de negócio;
- Direcionar as escolhas das fontes de dados;
- Entrevistas com usuário final para perceber a complexidade e viabilidade dos dados.

2. Requisitos de Conformidade

- Quais são as políticas de conformidade e segurança adotadas pela empresa?
- Transparência nos dados;
- Provar não mudança dos dados (em casos financeiros);
- Provar segurança com os dados;

3. Perfil de Dados

- Um exame sistemático da qualidade, escopo e contexto de uma fonte de dados para permitir a construção de um sistema ETL.

4. Requisitos de Segurança

5. Integração de Dados

6. Latência dos Dados

- Qual é o tempo máximo permitido para disponibilização dos dados através do sistema de BI?

7. Arquivamento e Linhagem

- Ter cópias dos *stagings* subsequentes;
- Provar a descendência dos dados.

8. Interfaces do usuário final

9. Habilidades disponíveis

10. Licenças legadas

- Uso de um sistema legado para o desenvolvimento de ETL é pode ser um erro.

- ❖ **Ferramenta ETL x Codificação manual;**
 - Comprar ou desenvolver.
- ❖ **Processamento em Batch x Fluxo de dados de Streaming;**
 - O tradicional é o ETL processar periodicamente;
 - Quando o tempo de resposta é urgente em um DW/BI, a abordagem tradicional não funciona;
 - Streaming: os dados em um nível de registro fluem continuamente do sistema de origem para os bancos de dados e telas dos usuários.
- ❖ **Tarefas com dependência Horizontal x Vertical;**
 - **Horizontalmente:** permite que cada **carga** final do banco de dados seja executado de forma independente;
 - **Verticalmente:** sincroniza dois ou mais fluxos de trabalho separados para que os carregamentos finais do banco de dados ocorram simultaneamente.

	Vantagens	Desvantagens
Codificação Manual	Técnicas de programação orientadas a objeto.	Em alguns casos possuem uma menor velocidade de execução.
	O teste do código ETL possuem muitas tecnologias.	Manutenção de metadados seja feita separadamente.
	Codificação manual é mais flexível.	Manutenção desses Sistemas de ETL.
	Os metadados criados podem ser controlados de uma forma mais direta.	
Ferramenta ETL	Fluxo visual e Auto documentação.	Custo da licença.
	Técnicos que não sejam programadores podem utilizar.	Flexibilidade reduzida.
	Conectores pré-programados	Incerteza.
	Repositório integrado de metadados.	
	Monitoramento.	
	Limpeza avançada de dados.	

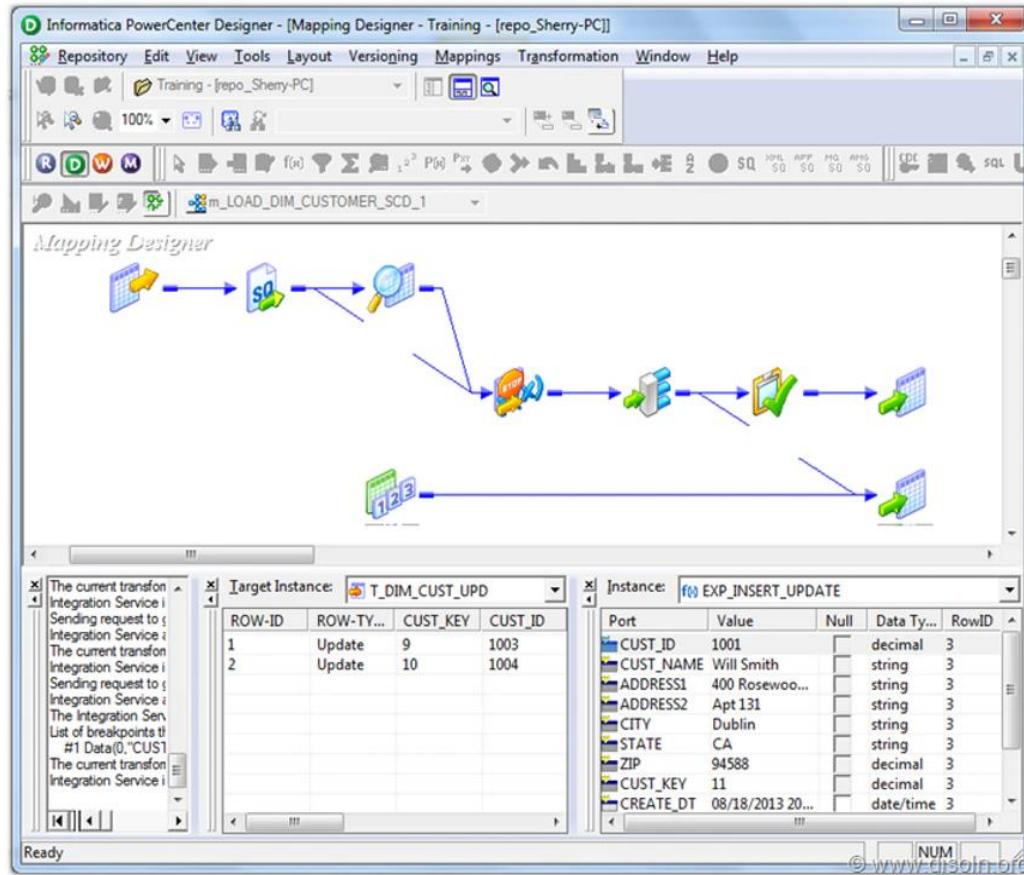
❖ Ferramenta ETL x Codificação manual, qual utilizar ?

1. A complexidade da transformação de dados;
 - Quanto mais complexo melhor uma ferramenta.
2. Necessidade de limpeza de dados;
 - Haverá limpeza profunda antes de serem armazenados no DW?
3. Volume de dados.
 - Qual será o volume de dados a serem extraídos?
 - Todos esses dados vão passar por uma transformação e limpeza?

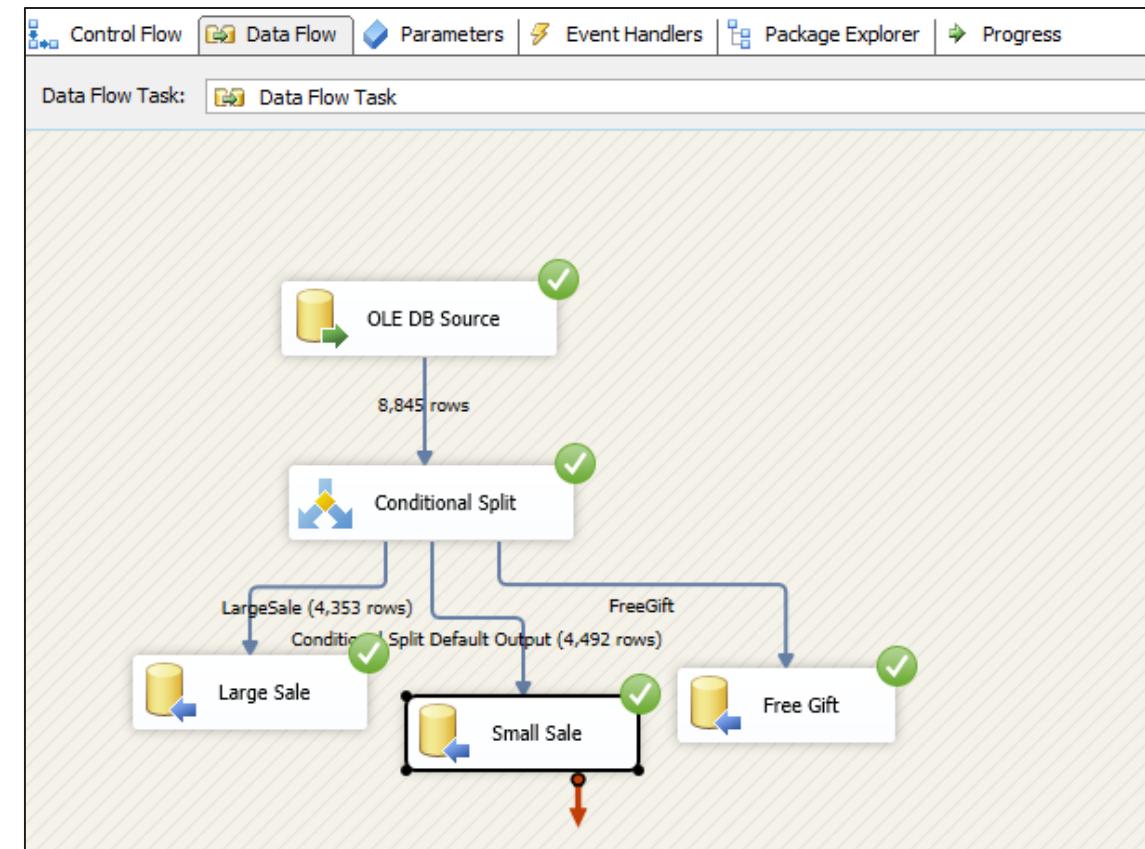
- ❖ Automatização
 - Agendar, executar, pausar e parar.
- ❖ Monitoramento
 - Gerenciar a exceção;
 - Gerenciar a qualidade;
 - Recuperação e Restart;
 - Metadados de execução;
 - Segurança:
 - Controle de usuários;
 - Controle nos log e acesso aos dados de *staging*.

Ferramentas de ETL

Informática Power Center

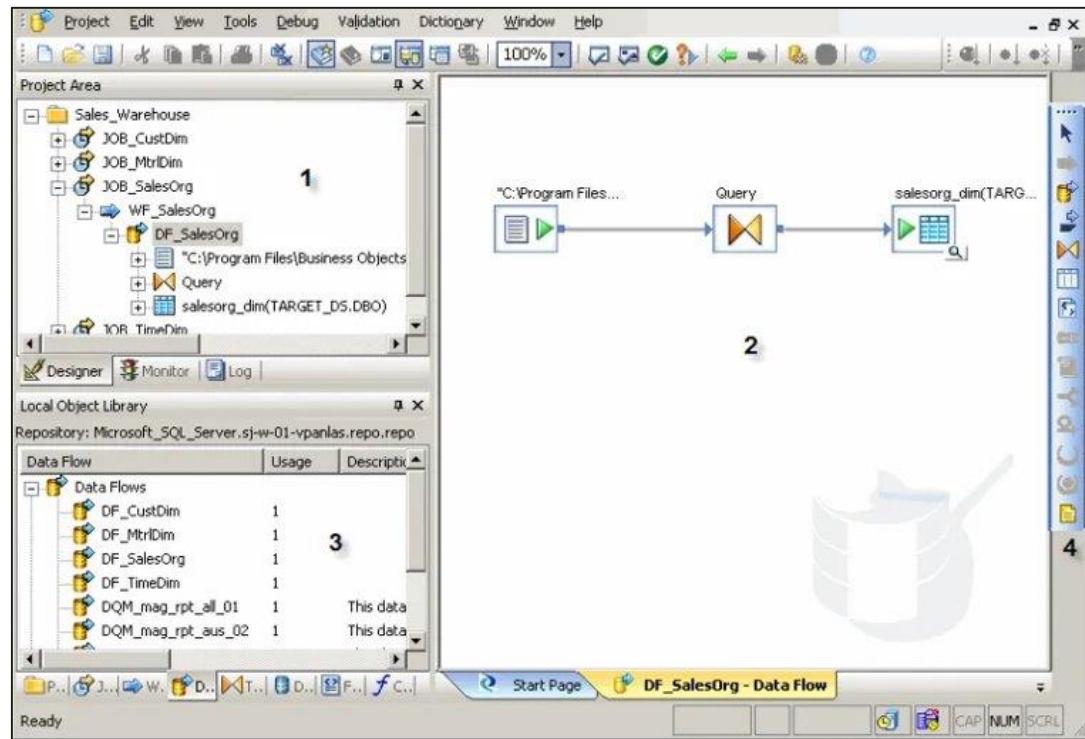


Microsoft SQL Server Integration Services (SSIS)

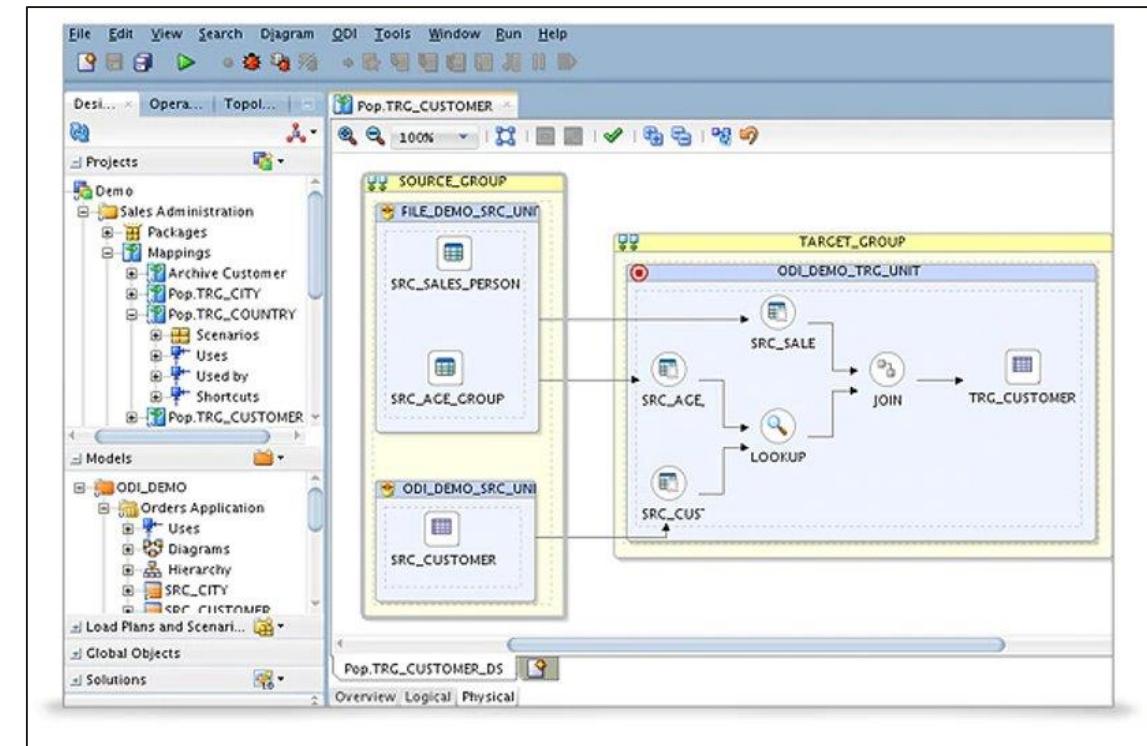


Ferramentas de ETL

SAP Business Objects Data Services

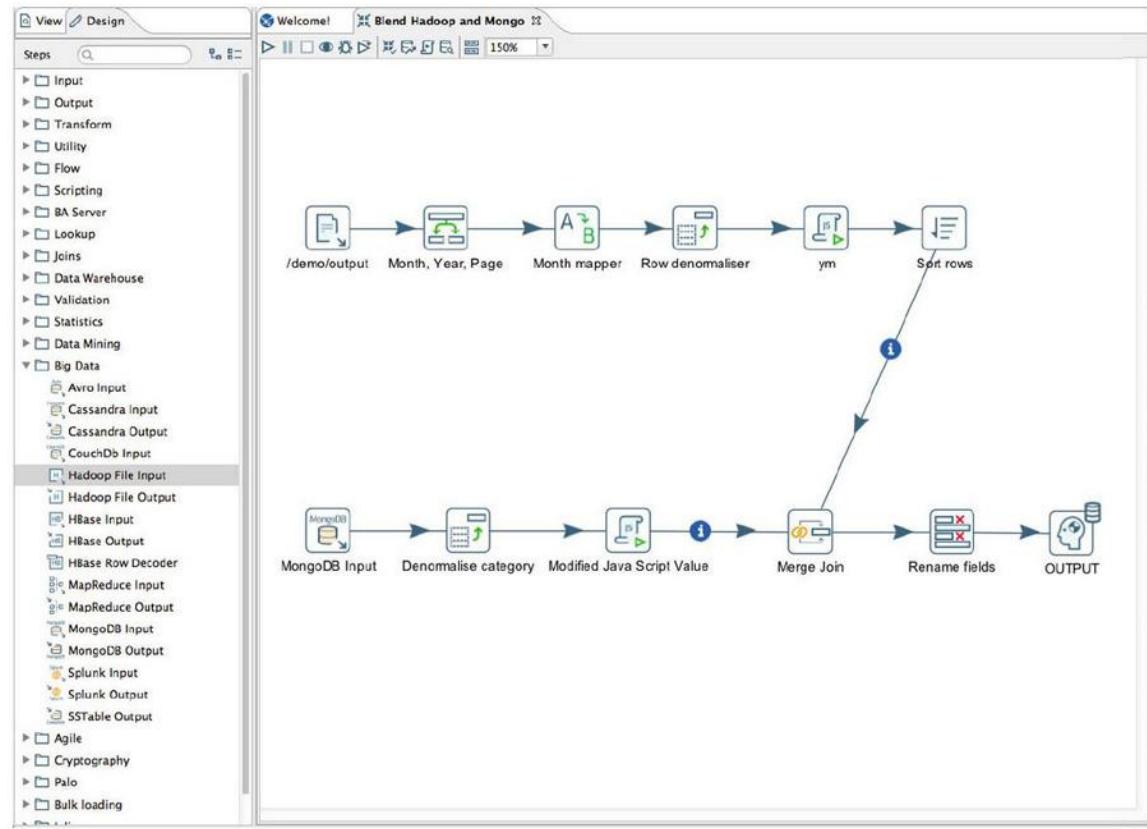


Oracle Data Integrator (ODI)

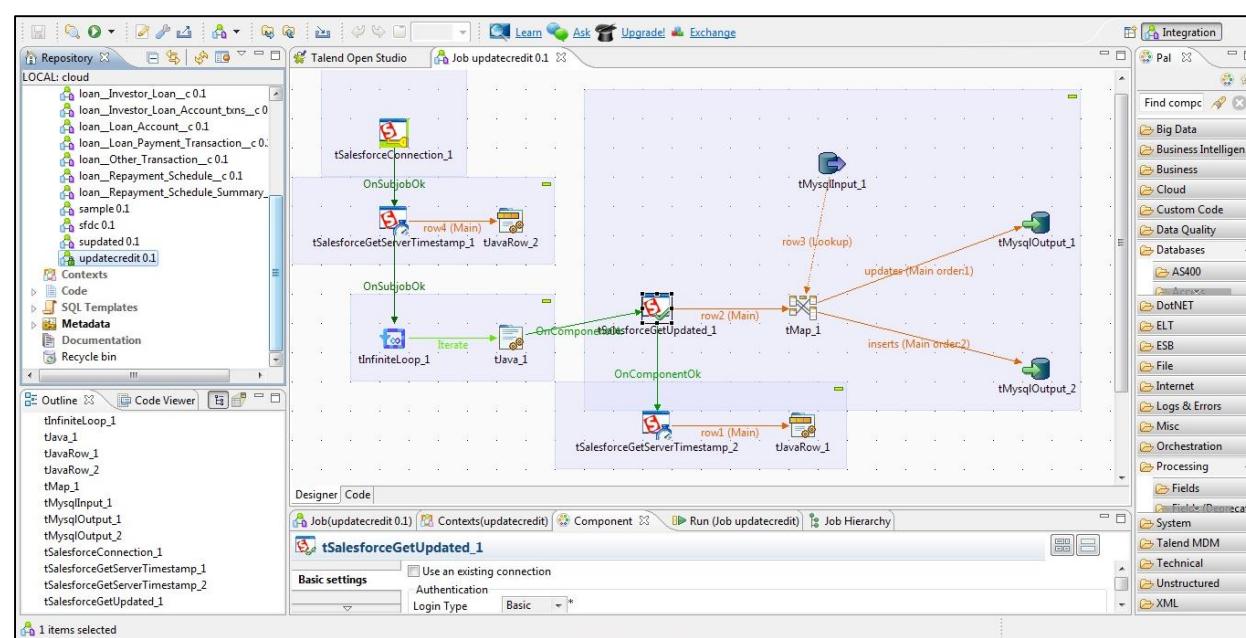


Ferramentas de ETL

Pentaho Data Integration



Talend Open Studio Data Integration



Ferramentas ETL

Soluções proprietárias

- ❖ Informática Power Center
- ❖ Microsoft SQL Server Integration Services (SSIS)
- ❖ Oracle Data Integrator (ODI)
- ❖ SAP Business Objects Data Services

Open Source

- ❖ Pentaho PDI - Kettle
- ❖ Talend Open Studio
- ❖ CloverETL
- ❖ KETL

Na prática um exemplo ciclo...

- ❖ Planejamento e Projeto
- ❖ Criar base de referência dos dados
- ❖ Extrair dos fontes
- ❖ Validar, Limpar, Agregar, Regra de negócio, Qualidade
- ❖ Carregar em *Staging*
- ❖ Relatório de auditória e diagnóstico
- ❖ Publicar

Pentaho Data Integration

PDI, Spoon, Kettle, Job...

- ❖ Solução de BI Open Source;
- ❖ Pentaho é uma suíte de softwares para inteligência empresarial;
- ❖ Desenvolvido em Java;
- ❖ Possui versões CE (Community Edition) e EE (Enterprise Edition).



Pentaho - Componentes

- ❖ Solução completa para BI (ETL, Reporting, Dashboards, OLAP, Mineração de dados, etc)



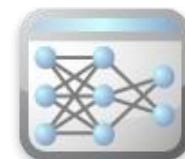
Pentaho Data Integration: Kettle;



Pentaho Analysis Services: Mondrian
OLAP server, Jpivot;



Pentaho Reporting: JFreeReport,
Jasper Report, Birt;



Data Mining

Pentaho Data Mining: Weka;



Pentaho Dashboard: Ctools, CDE,
CDA;



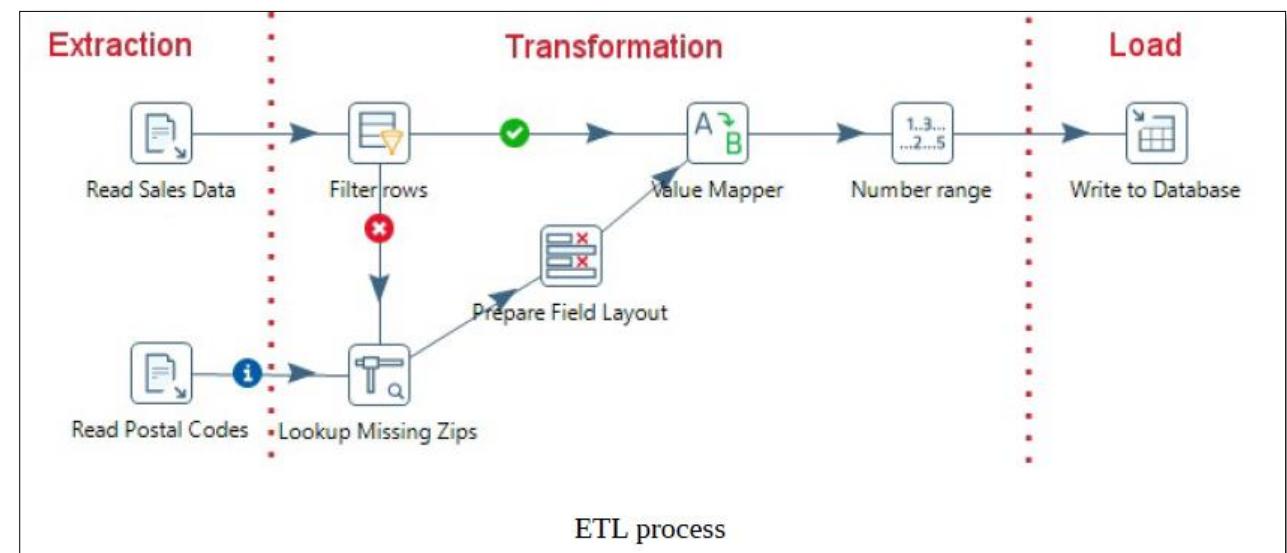
Pentaho BI Server;

Pentaho Data Integration ou Kettle

- ❖ *Pentaho Data Integration* (PDI, também chamado Kettle) é um componente da suíte do Pentaho responsável pelos processos de Extração, Transformação e Carga (ETL).



Kettle
Extraction
Transportation
Transformation
Loading
Environment



Instalando o PDI

❖ Download:

- \\172.17.21.3\publica\etl
- <https://sourceforge.net/projects/pentaho/files/Pentaho%208.3/client-tools/pdi-ce-8.3.0.0-371.zip/download>

❖ Descompactar:

- pdi-ce-8.3.0.0-371.zip



./data-integration

❖ Executar:

• Se Linux:

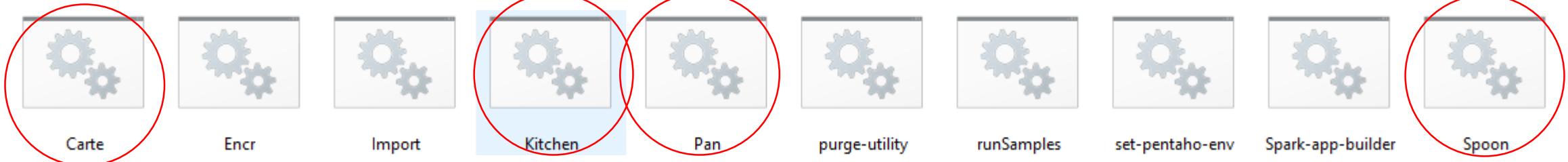
- #chmod 755 spoon.sh
- Executar # ./spoon.sh

• Se Windows:

- Executar spoon.bat

Componentes do PDI

- ❖ **Spoon** – ferramenta gráfica com que se desenha e testa todo processo do PDI
- ❖ **Pan** – Execução de **Transformações** em linha de comando.
- ❖ **Kitchen** – Execução de **Jobs** em linha de comando.
- ❖ **Carter** – Servidor para execução remota.
- ❖ **Repositório** – Os metadados das transformações e Jobs. Podem ser persistidos em um banco de dados, em arquivos ou em um servidor.



Outras instalações

❖ JAVA

- https://www.java.com/pt_BR/download/

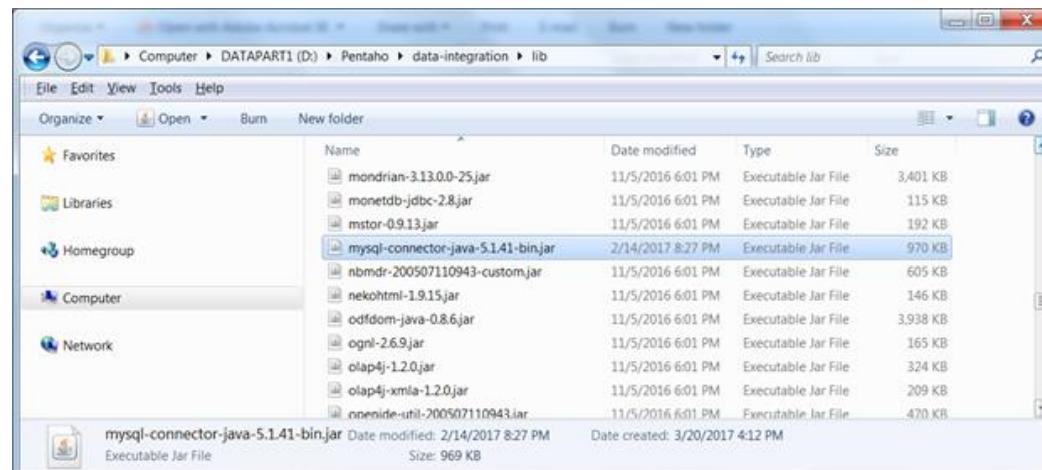
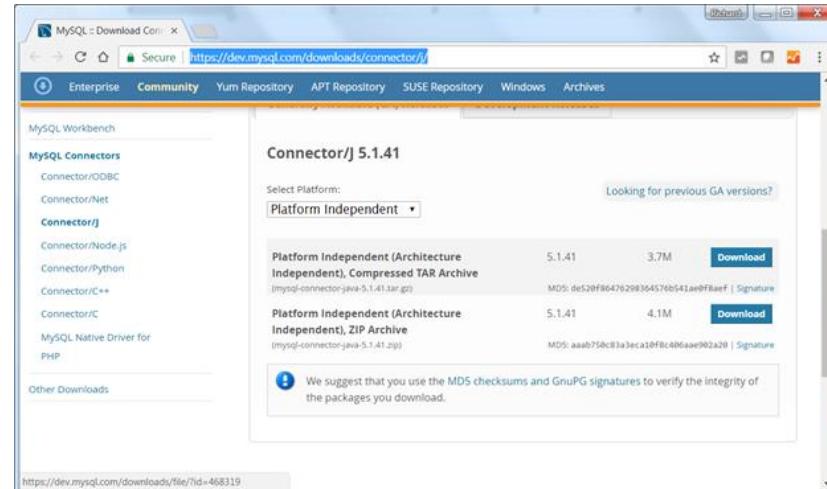
❖ Configurar JAVA_HOME

❖ Instalar banco de dados MySQL:

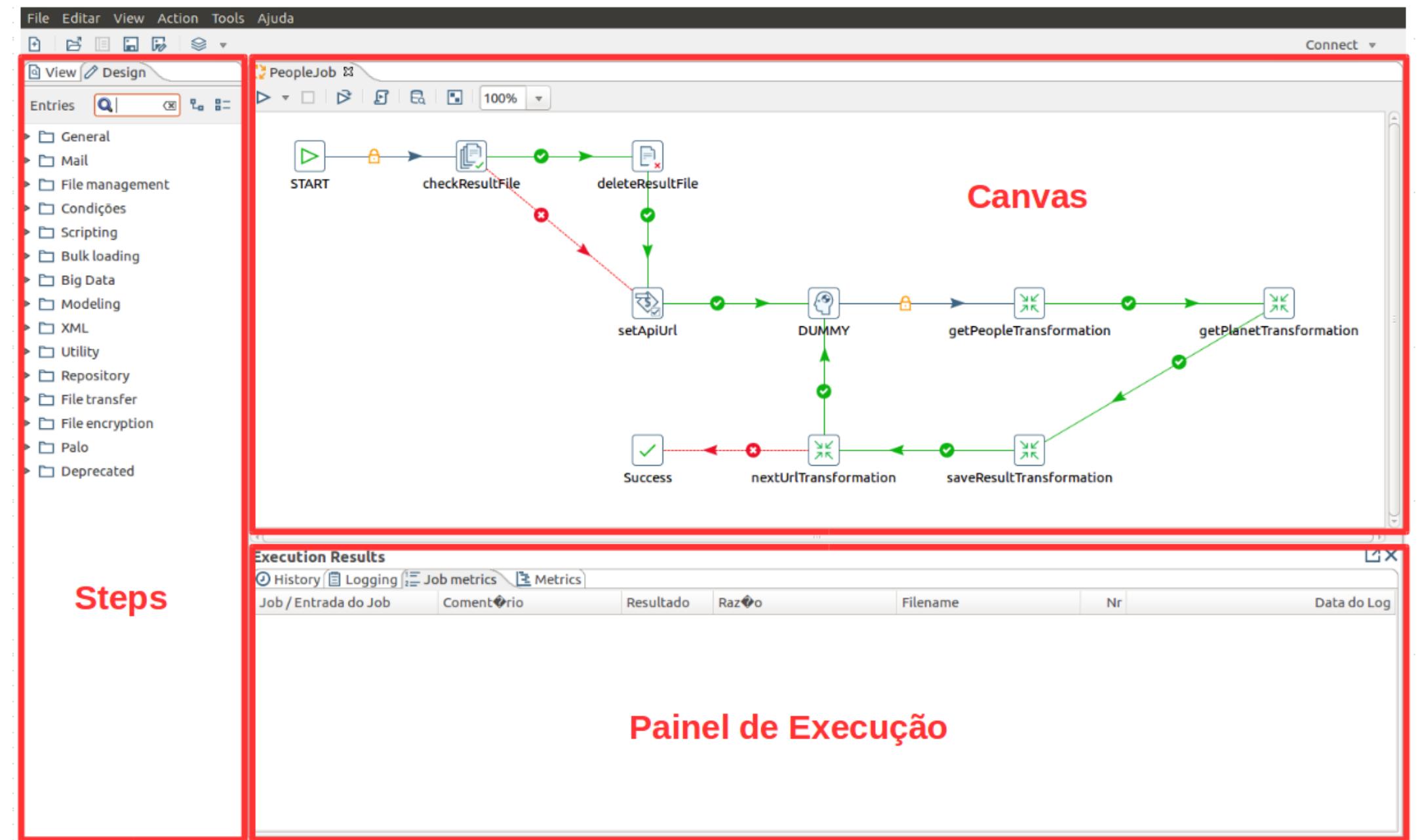
- <https://dev.mysql.com/downloads/installer/>

❖ Download do Drive JDBC MySQL:

- <https://dev.mysql.com/downloads/connector/j/>
- Copiar mysql-connector-java-5.1.41-bin.jar para pasta .\data-integration\lib



Interface Gráfica - PDI



Elementos Básicos

❖ MAPA ETL:

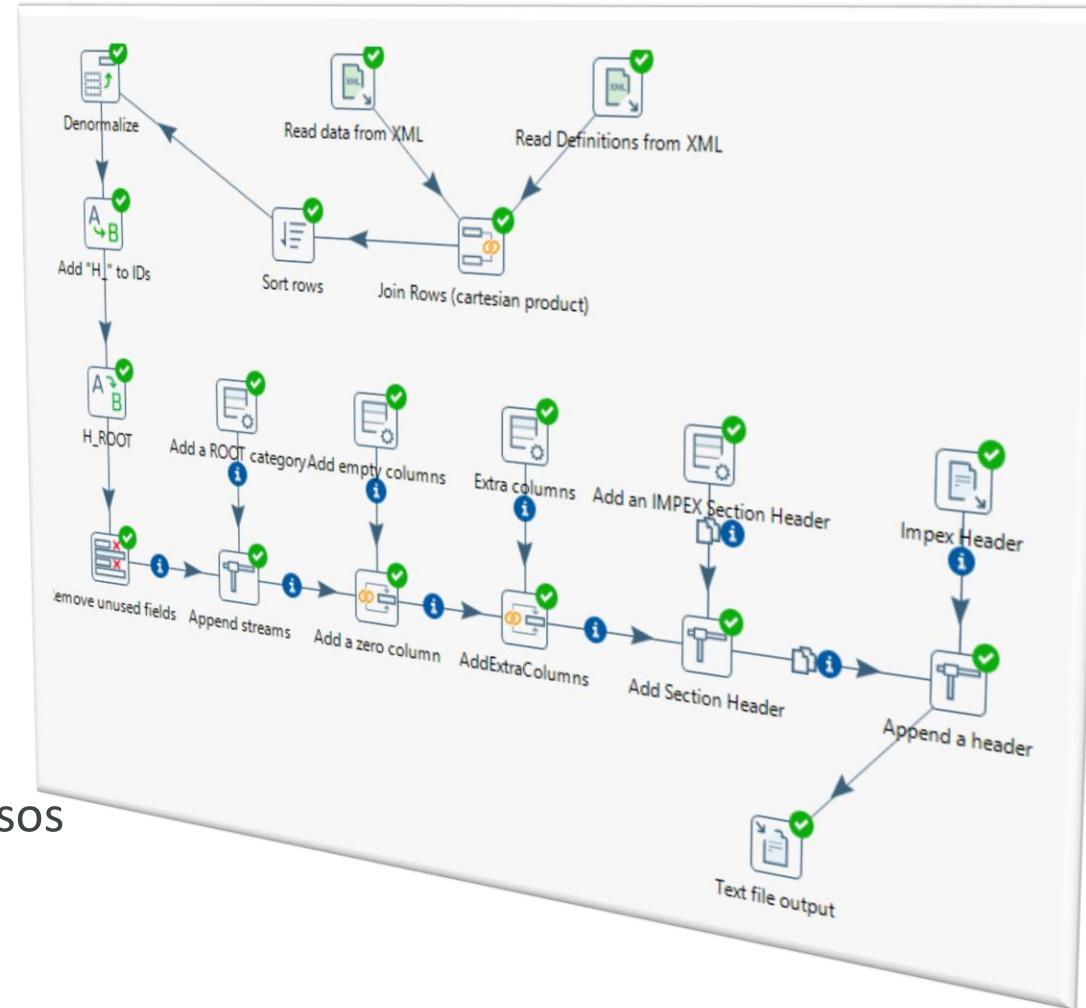
- Transformation - *.KTR
- Job - *.KJB

❖ STEP:

- Um passo é uma unidade mínima dentro de uma transformação;
- Grande variedade de passos;
- Agrupada em categorias (input , Output, etc);
- Os tipos básicos são: entrada , transformação , saída.

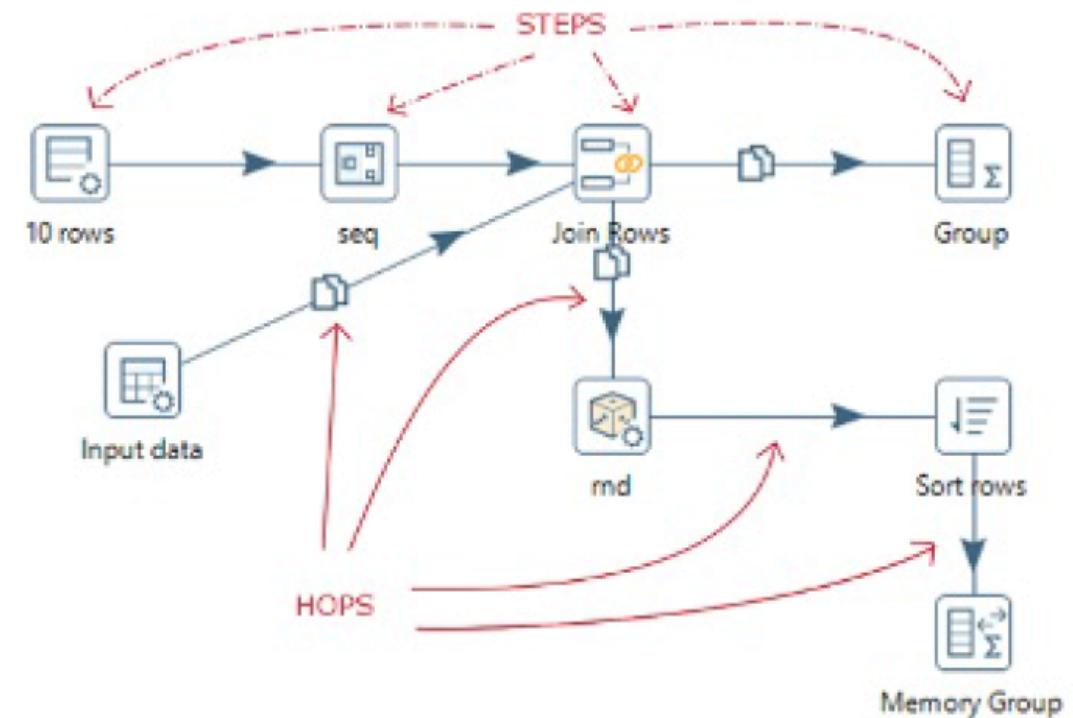
❖ HOP:

- Representação gráfica do fluxo de dados entre dois passos (conexão);
- Um deles Origem e outro Destino.

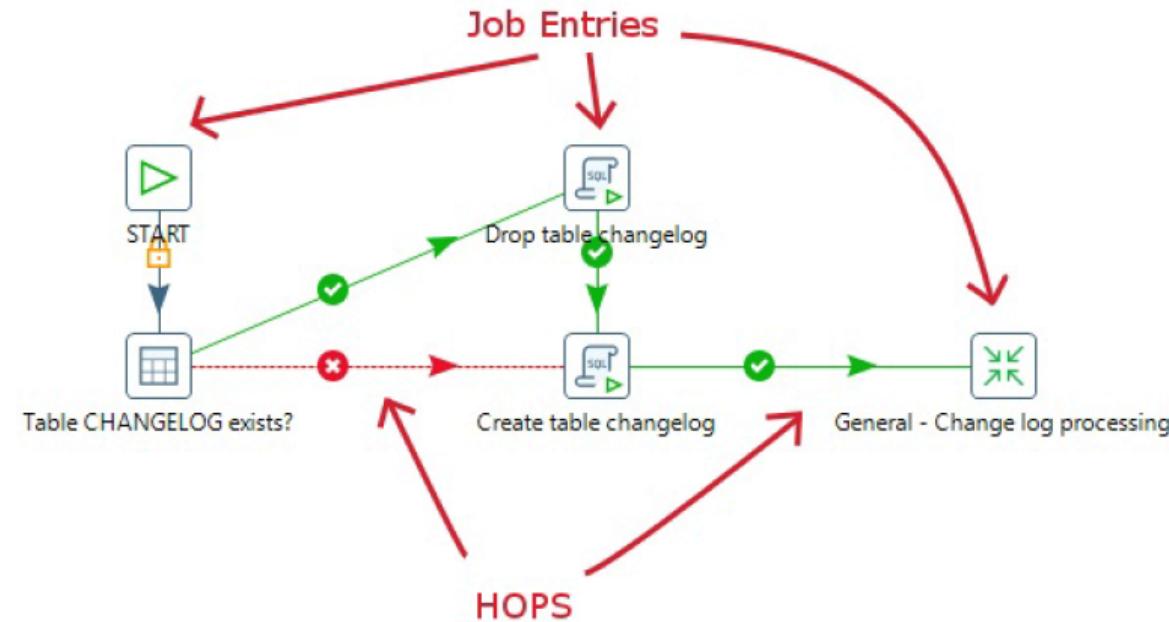


Transformação

- ❖ Consiste de uma coleção de steps de transformação;
- ❖ Cada *step* denota uma **operação** do processo de ETL;
- ❖ A saída de um *step* produz um **conjunto de registros**;
- ❖ É recomendado 1 transformação para cada dimensão ou tabela fato;
- ❖ É **executado em paralelo e de forma síncrona**.

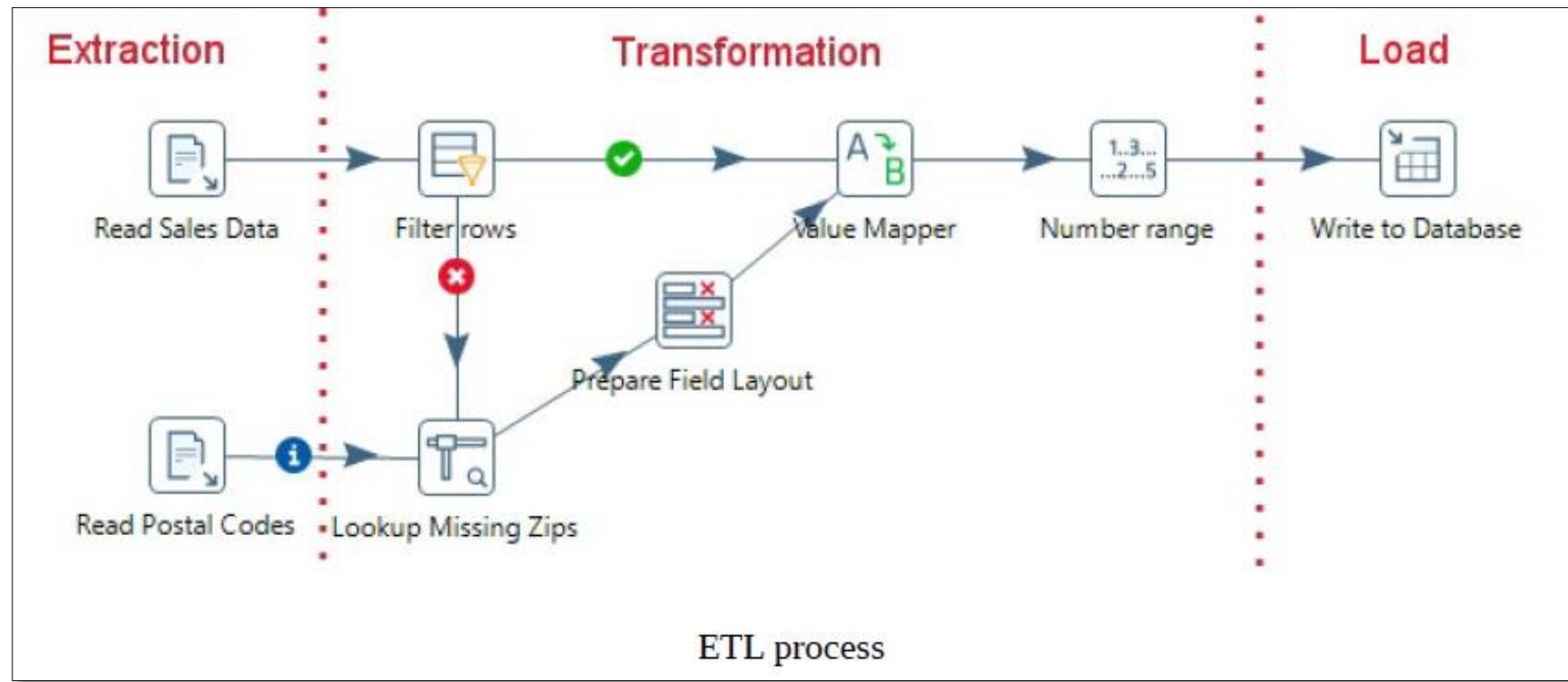


- ❖ Consiste de uma coleção de transformações ou de *steps* de *Jobs*;
- ❖ Cada entrada do *job* denota uma **tarefa** do processo de ETL;
- ❖ A saída de cada entrada do *job* produz um **status** de execução;
- ❖ Carga da tabela fato;
- ❖ É executado de forma **sequencial** e caminhos **assíncrono**.



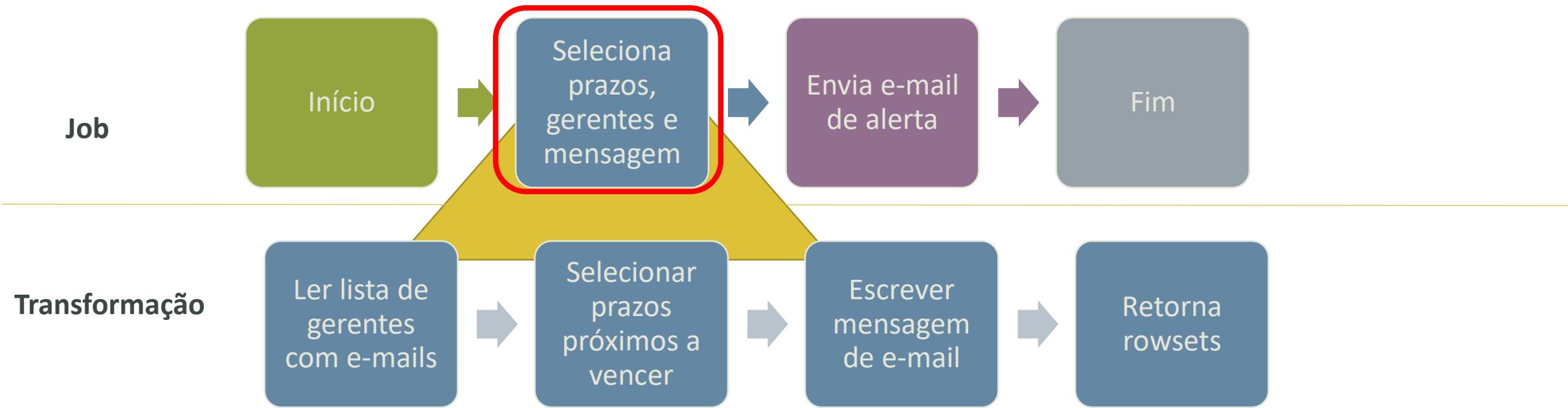
Mapa ETL

- ❖ Representação de um fluxo de dados, de tarefas ou algoritmo.

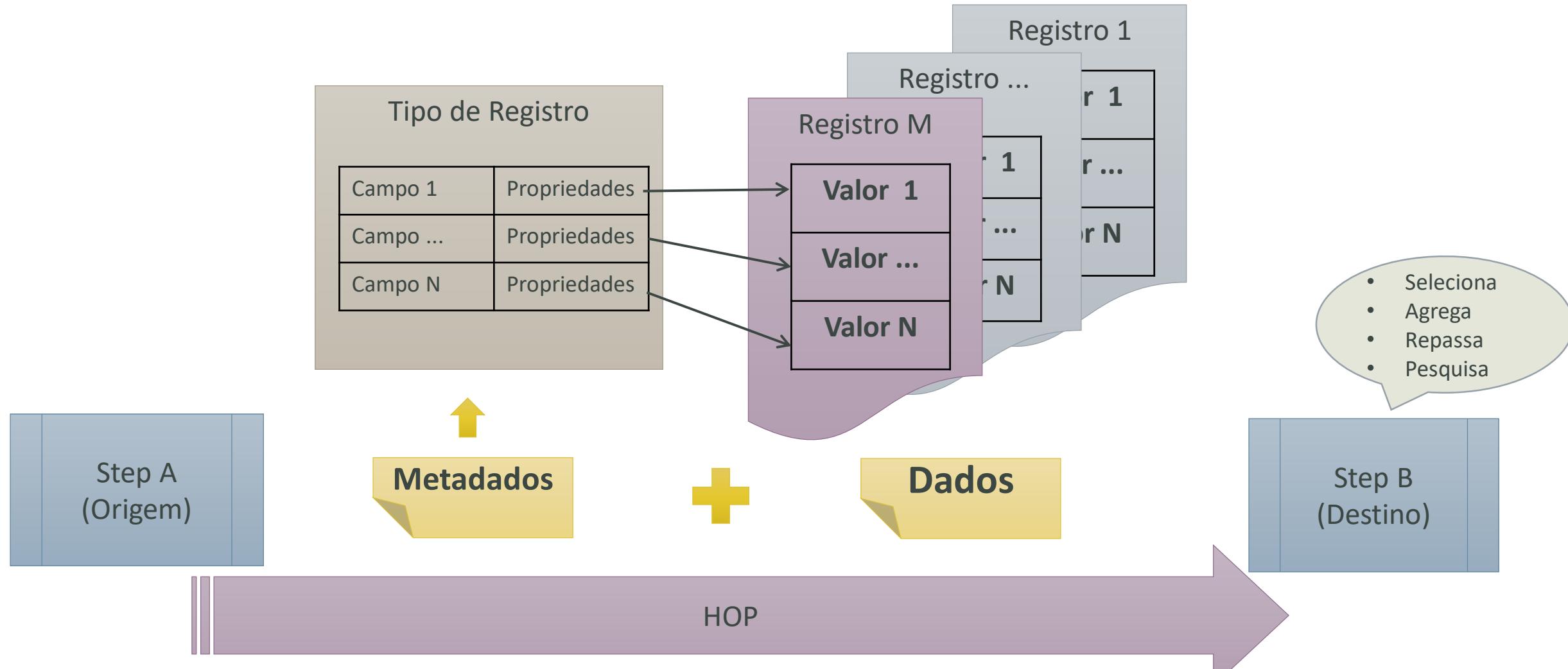


Resolvendo problemas

- ❖ **Problema:** Gerentes perdem prazos para solicitar compras.
- ❖ **Solução:** Notificar coordenadores do encerramento do prazo.
- ❖ **Restrição:** O sistema de compras é externo.



Fluxo de Dados



Fluxo de dados

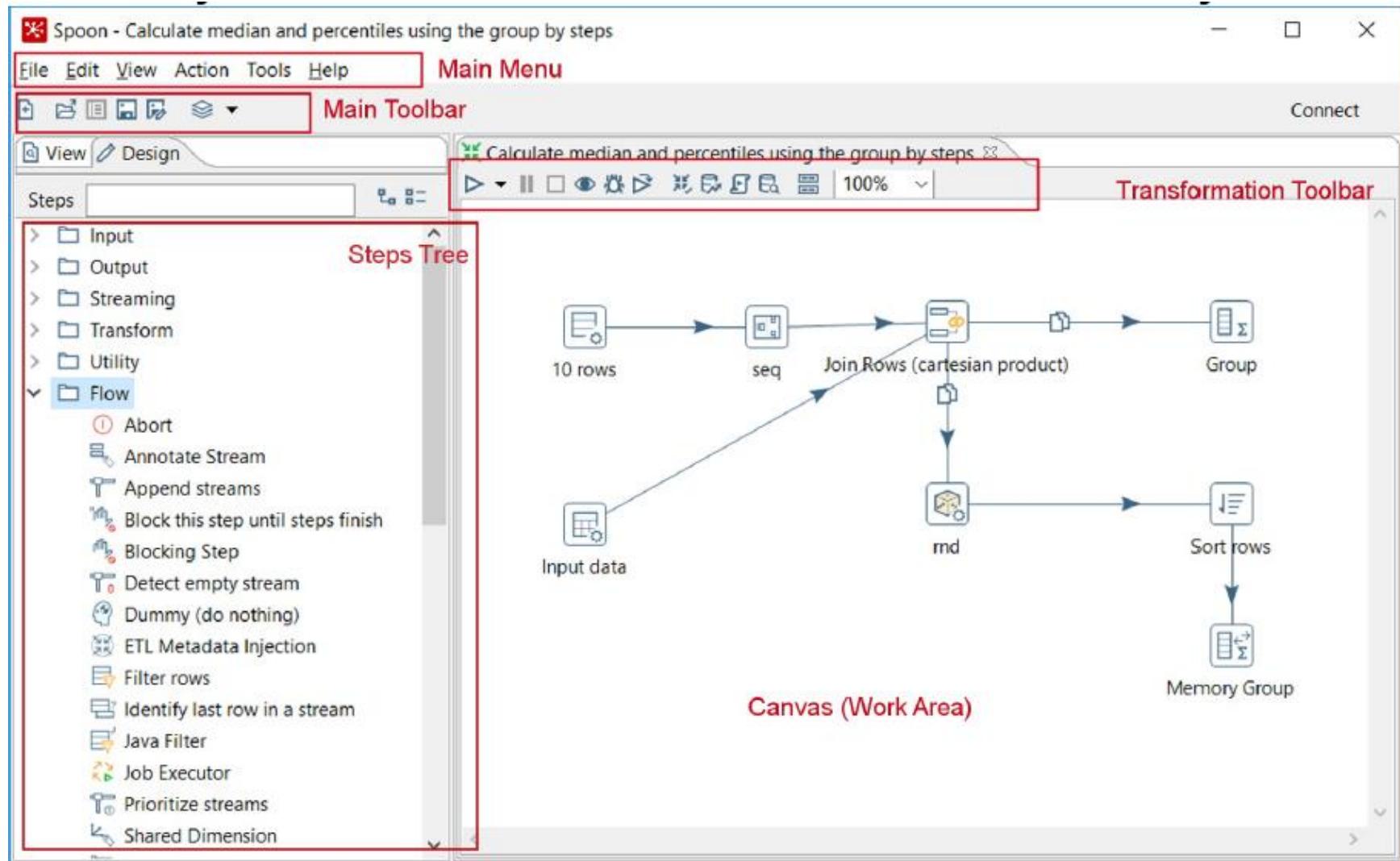
❖ Rowset

- Dados representados de forma tabular (datasets)
- Cada coluna representa um **campo**
 - Nome (obrigatório)
 - Tipo: Number (float), String, Date, Boolean, Integer e Big Number
- Cada **linha** corresponde a um membro do dataset

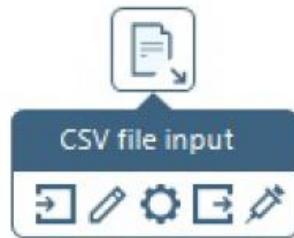
❖ Streams

- Dados enviados de um step para outro
 - Os hops apenas repassam o fluxo de dados
- Cada step pode ter um rowset de **entrada** e outro de **saída**
- Botão direito -> Mostra campos de entrada/saída

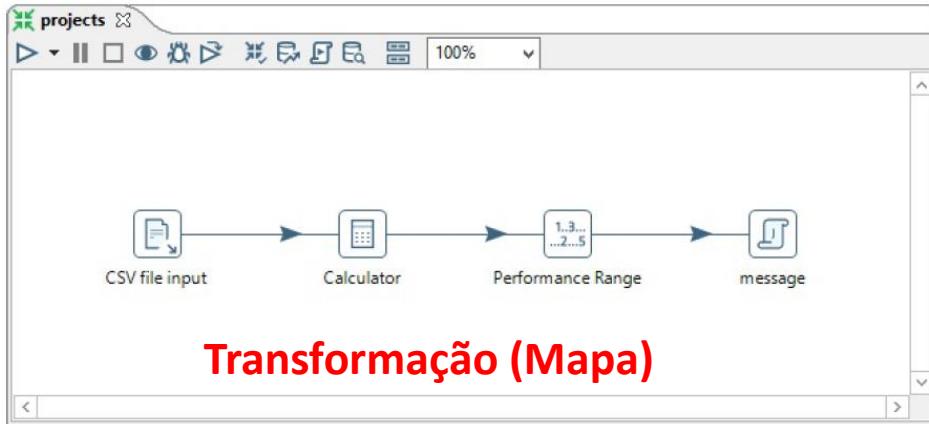
Kettle



Kettle



Step

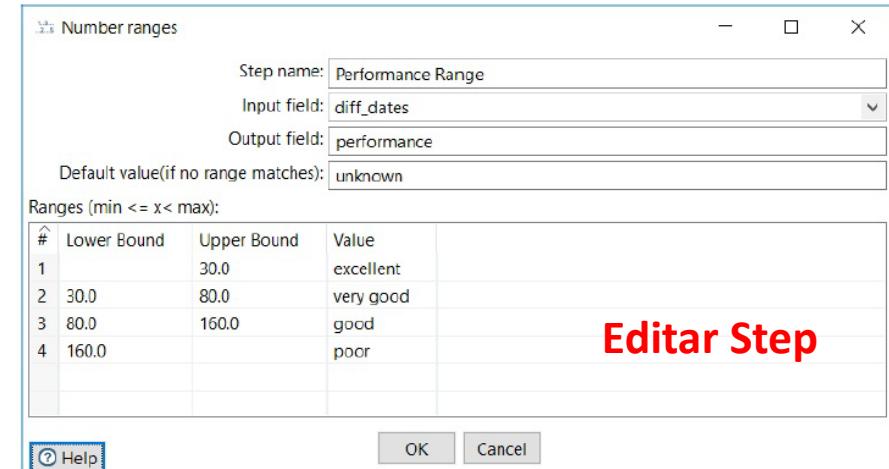


Examine preview data

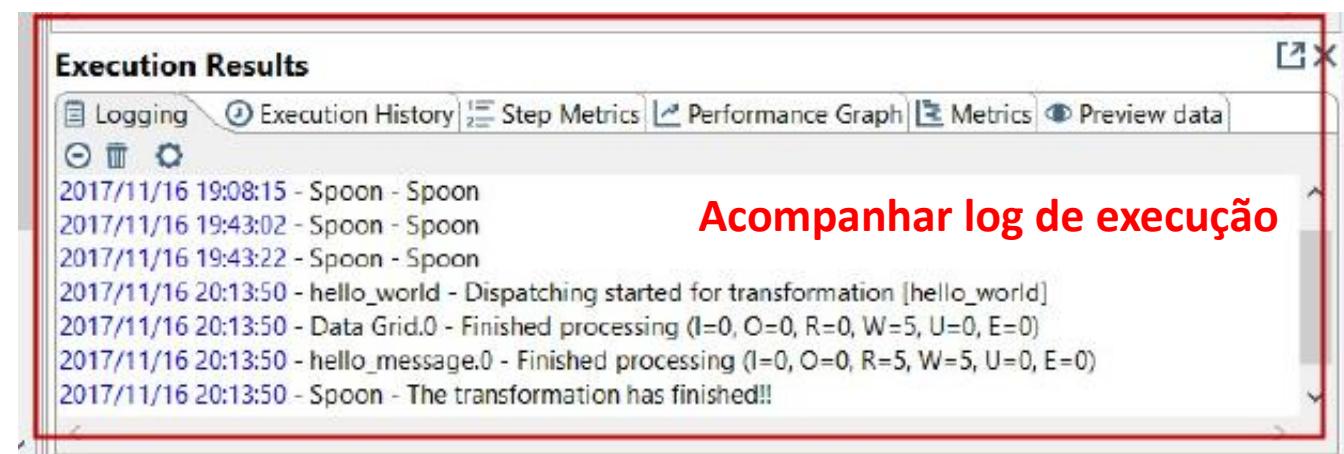
Rows of step: hello_message (5 rows)

#	people	hello_message
1	John	Hello, John!
2	Mary	Hello, Mary!
3	Ammy	Hello, Ammy!
4	Erik	Hello, Erik!
5	Laura	Hello, Laura!

Acompanhar os dados



Editar Step



Executando ETLs

❖ Pan (*.sh *.bat)

- file = Arquivo (Transformação in XML) para executar.

❖ Exemplo:

- **pan.sh -file minha-transformacao.ktr**

❖ Exemplo com Log:

- **pan.sh -file transformacao.ktr -logfile log.txt**

❖ Kitchen (*.sh *.bat)

- file = Arquivo (Transformação in XML) para executar.

❖ Exemplo:

- **kitchen.sh -file meu-job.kjb**

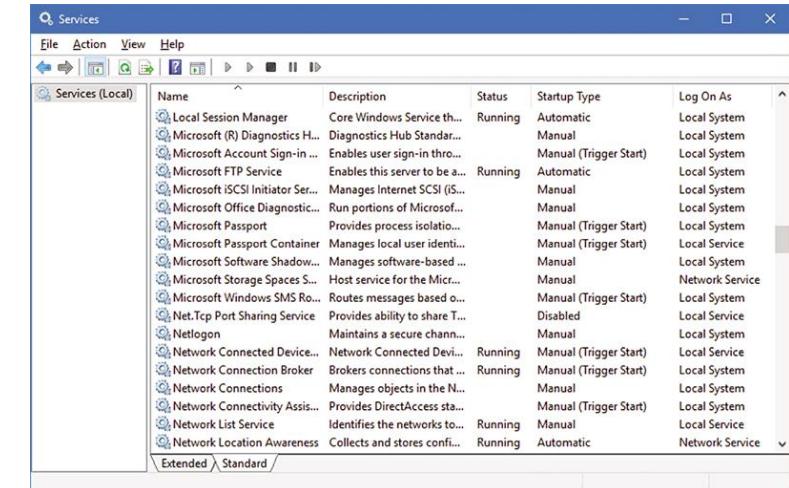
❖ Exemplo com Log:

- **kitchen.sh -file job.kjb -logfile logjob.txt**

Automatizar

- ❖ Serviços do Windows
- ❖ Crontab do Linux
- ❖ Chamar execução remotamente via
Carte (SOA)

```
root@PO20XIDCMF # crontab -l
#ident  "#0 (#)root      1.21    04/03/23 SMI"
#
# The root crontab should be used to perform accounting data collection.
#
#
10 3 * * * /usr/sbin/logadm
15 3 * * 0 /usr/lib/fs/nfsfind
30 3 * * * [ -x /usr/lib/gss/gsscred_clean ] && /usr/lib/gss/gsscred_clean
#10 3 * * * /usr/lib/krb5/kprop_script __slave_kdcs
30 23 * * * [ -x /usr/local/bin/syschk.sh ] && /usr/local/bin/syschk.sh
30 23 * * * [ -x /usr/local/bin/inventory.sh ] && /usr/local/bin/inventory.sh
38 23 * * * /src/oracledump.sh >> /src/oracledump.log
30 00 * * * /src/datatransfer/dataload_new.sh > /tmp/nfc_dataload.log
00 10 * * * /netflow1/nflowbck/crtbl.sh
00 09,13,17 * * * /src/loadBackupData.sh
00 08 * * * /src/bkp_policy_mismatch.sh
15 00 * * * /src/datatransfer/create_folder_date_DT.sh > /tmp/create_folder.log
#1,6,11,16,21,26,31,36,41,46,51,56 * * * * /netflow2/AsLogsFtp/flukeFtp.sh
```



PDI Status

Transformation name	Carte Object ID	Status	Last log date	Remove from list
credit-carte	62fab675-2a90-4320-a09a-35070f449244	Finished	2018/03/27 16:09:27.194	Remove

Job name | Carte Object ID | Status | Last log date | Remove from list |

Configuration details:

Parameter	Value
The maximum size of the central log buffer	10000 lines
The maximum age of a log line	1440 minutes
The maximum age of a stale object	1440 minutes
Repository name	

These parameters can be set in the slave server configuration XML file: configuration.xml

Exemplos – Modelos PDI

- ❖ Base de conhecimento:
- ❖ Exemplos:
 - Na pasta do ./pentaho-integration/samples/transformations
 - Documentação de Step
 - <http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps>

LAB 1 - PDI

Repositório, Criação de Jobs e Transformação

* Arquivo lab1_pdi.pdf

Extração

Coleta de dados, Diferentes fontes de dados

Extração de Dados

- ❖ Primeira tarefa que deve ser realizada (Input Data);
- ❖ Coleta de dados relevantes propagados para o *Data Warehouse*;
- ❖ Atividades de Extração:
 - 1. Captura dos Dados**
 - A seleção de dados de diferentes fontes e formatos;
 - Identificação de mudanças desde a última extração;
 - Validação e descarte de dados de acordo com regras de negócio e padrões.
 - 2. Staging**
 - Armazenamento temporário dos dados

Extração de Dados

❖ Desafios

- Que dados extrair ?
- Quais filtros devem ser aplicados ?
- Processo simples porém demorado.

❖ Começar os estudos por:

- Conhecer os dados e *metadados*;
- Análise da contribuição das **fontes** para composição do sistema final;
- Definir **estratégia** de captura dos dados;
- Checar a **política** de disponibilização dos dados.

Diversidade de fontes

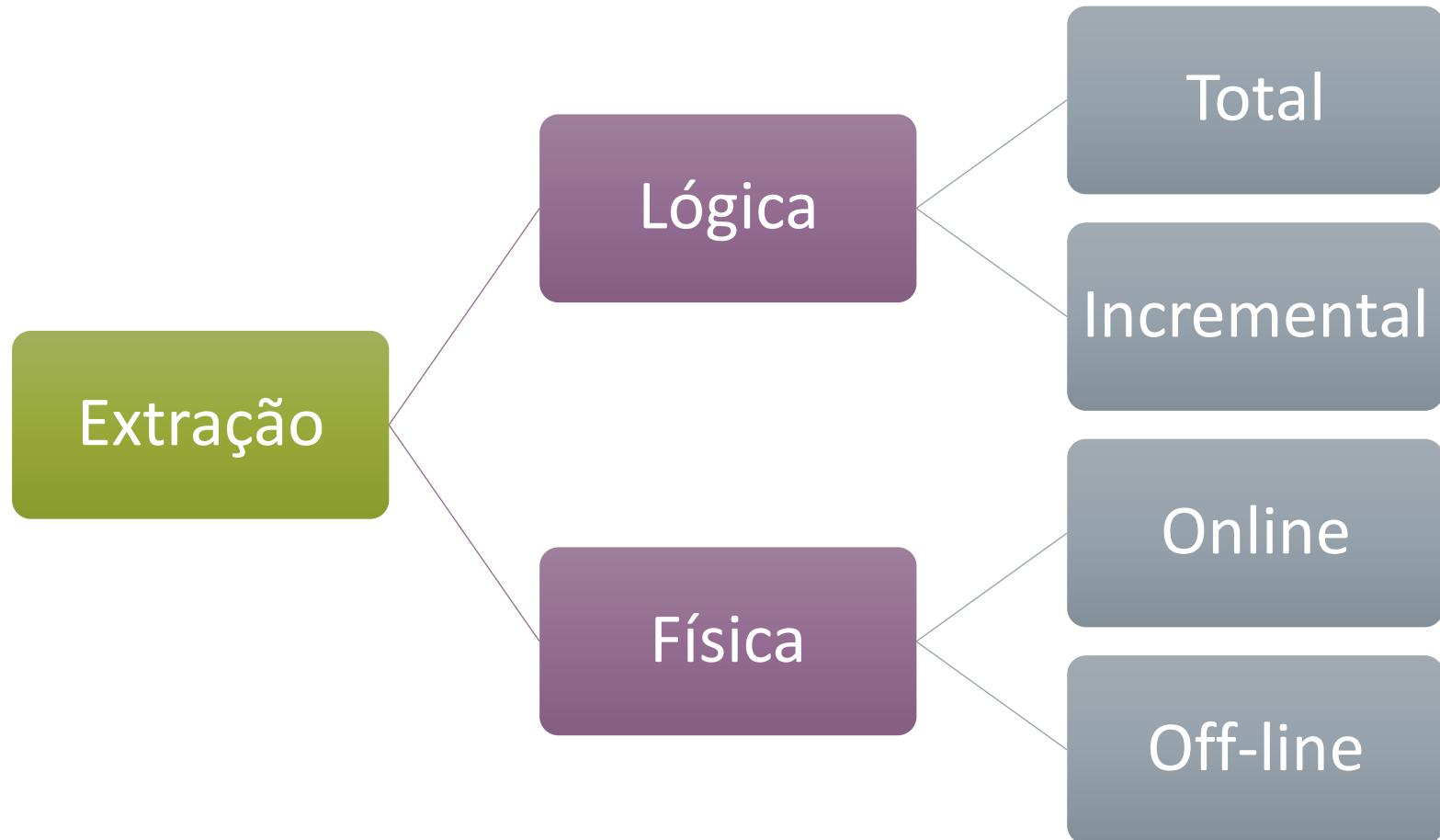
❖ Desafios

- Diferentes tecnologias.
- Diferentes tempos.
- Diferentes modelos.

❖ Restrições

- A fonte de dados deve sofrer o **mínimo de sobrecarga** durante a extração, já que outras atividades administrativas também acontecem durante esse período.
- Deve haver um **mínimo de interferência** com a configuração do software no lado da fonte de dados por razões técnicas e políticas.

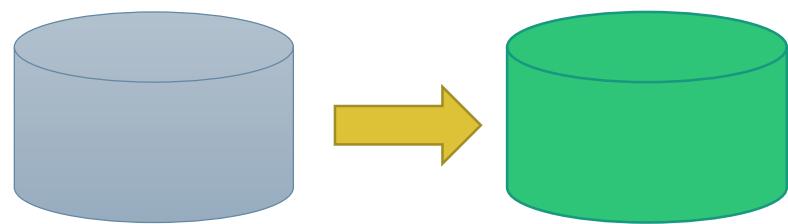
Classificação da Extração



Extração Lógica

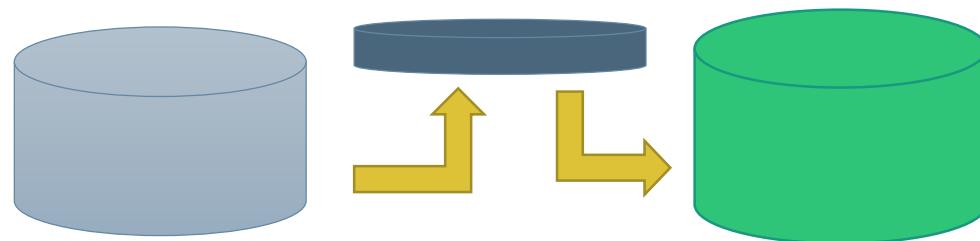
❖ Total

- Não há controle das alterações que houveram com os dados nas fontes;
- Pode levar muito tempo;
- Carga inicial;
- Atualização global.



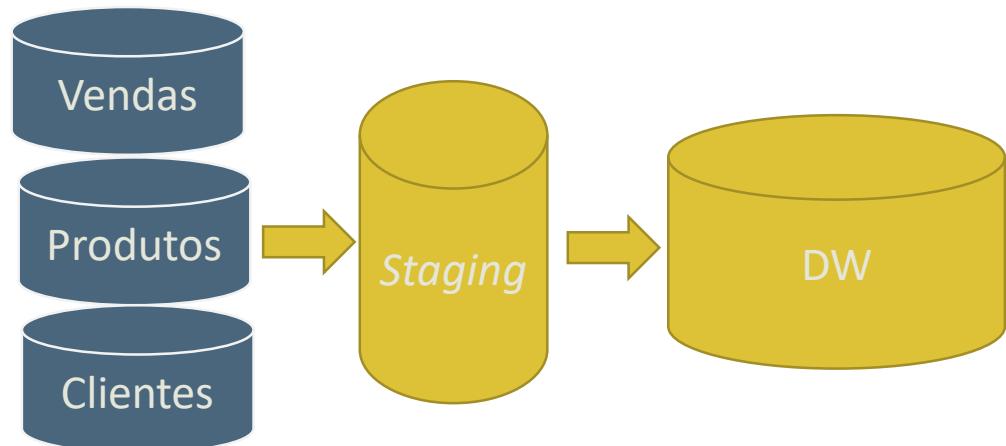
❖ Incremental

- Possui ponto lógico de corte;
- Usa algum mecanismo para perceber a mudança;
- Tempo curto e atividade periódica.

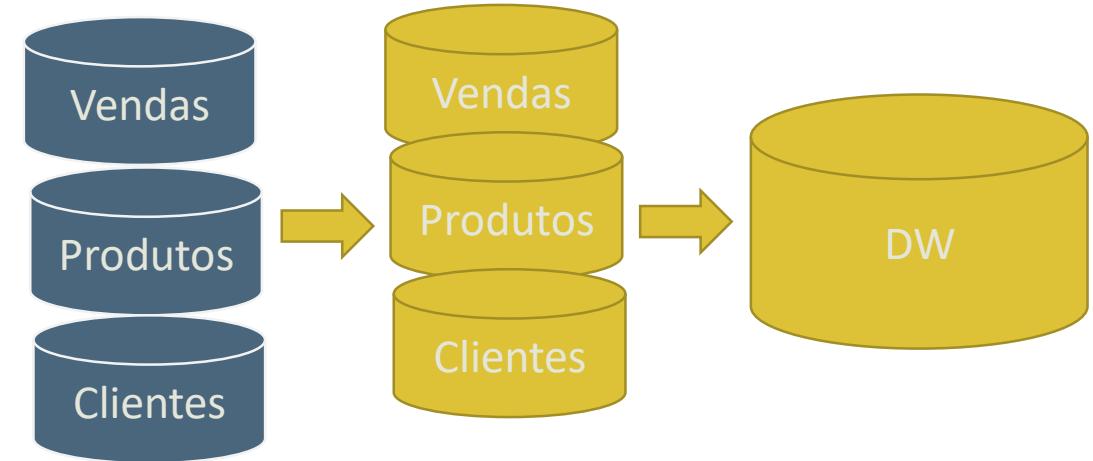


Extração Física

❖ Online



❖ Off-line



Processo de Extração (3 subsistemas - Kimball)

1. Perfil dos dados:

- É uma análise técnica dos dados para descrever seu conteúdo, consistência e estrutura. Ela desempenha duas funções: a estratégica e a tática.
- Mitigar riscos com a qualidade do dados.
- Mapa lógico dos dados.

2. Captura de Dados Alterados (CDC):

- Detecta mudanças nos dados e executa uma extração seletiva e incremental. Utiliza algoritmos de CRC.

3. Sistema de Extração:

- Leitura dos dados de diversas fontes para *Staging Area*, integração, migração e dispor para DW/BI.

Mapa lógico dos dados

- ❖ É um documento que mostra a relação entre a fonte primária e o ponto final e geralmente é apresentado em forma de tabelas ou em formato de planilha.
- ❖ Componentes: **Destino**, **Origem**, **Transformação**.

Mapa lógico dos dados

	A	B	C	D	E	F	G	H	I	J	K
1	TARGET						SOURCE				
2	Tabela	Campo	Tipo	Descrição	Tipo Tabela	SCD		Tabela	Campo	Tipo	Descrição
3	dim_usuario	id_usuario	int	chave primaria	dimensao	0	<->	-	-	-	-
4	dim_usuario	cod_usuario	int	chave natural	dimensao	0	<->	users	id	int	chave primária
5	dim_usuario	nome_usuario	varchar	nome do usuario	dimensao	0	<->	users	firstname.lastname	varchar	nome do usuário
6	dim_usuario	email_usuario	varchar	email do usuario	dimensao	0	<->	users	mail	varchar	e-mail do usuário
7	dim_situacao	id_situacao	int	chave primaria	dimensao	0	<->	-	-	-	-
8	dim_situacao	cod_situacao	int	chave natural	dimensao	0	<->	issue_statuses	id	int	chave primária
9	dim_situacao	situacao	varchar	nome da situacao	dimensao	0	<->	issue_statuses	name	varchar	nome da issue
10	dim_situacao	tipo_situacao	varchar	tipo da situacao	dimensao	0	<->	issue_statuses	is_closed	boolean	tipo da situação aberta/fechada
11	dim_projeto	id_projeto	int	chave primaria	dimensao	0	<->	-	-	-	-
12	dim_projeto	id_projeto_pai	int	id do projeto pai	dimensao	0	<->	-	-	-	-
13	dim_projeto	cod_projeto	int	chave natural do projeto	dimensao	0	<->	project	id	int	chave primaria
14	dim_projeto	cod_projeto_pai	int	chave natural do projeto pai	dimensao	0	<->	project	parent_id	int	chave do projeto pai
15	dim_projeto	nome_projeto	varchar	nome do projeto	dimensao	0	<->	project	name	varchar	nome do projeto
16	dim_sprint	id_sprint	int	chave primaria	dimensao	0	<->	-	-	-	-
17	dim_sprint	cod_sprint	int	chave natural	dimensao	0	<->	versions	id	int	chave primária
18	dim_sprint	ano	varchar	ano da sprint	dimensao	0	<->	versions	created_on	timestamp	ano da sprint
19	dim_sprint	projeto	varchar	projeto mes	dimensao	0	<->	custom_fields	possible_values	varchar	projeto mês
20	dim_sprint	sprint	varchar	nome da sprint	dimensao	0	<->	versions	descripton	varchar	nome da sprint
21	dim_sprint	sprint_num	varchar	numero da sprint	dimensao	0	<->	versions	name	varchar	numero da sprint
22	dim_sprint	projeto_num	varchar	numero do projeto	dimensao	0	<->	custom_fields	created_on	timestamp	numero do projeto

Mapa lógico dos dados

- ❖ Tarefas durante desenvolvimento do mapa lógico:
 1. Identificar as fontes de dados;
 2. Coletar documentação dos sistemas de origem;
 3. Criar um relatório de rastreamento dos sistemas de origem:
 1. Responsável por cada fonte; Assunto; Área; Negócio envolvido; Plataformas; Servidor de produção, etc;
 4. Identificar e analisar os modelos e diagramas;
 5. Analisar amostras dos dados;
 6. Regras de negócio para o processo ETL;
 7. Formas de integração dos dados;

Captura de Dados Alterados - CDC

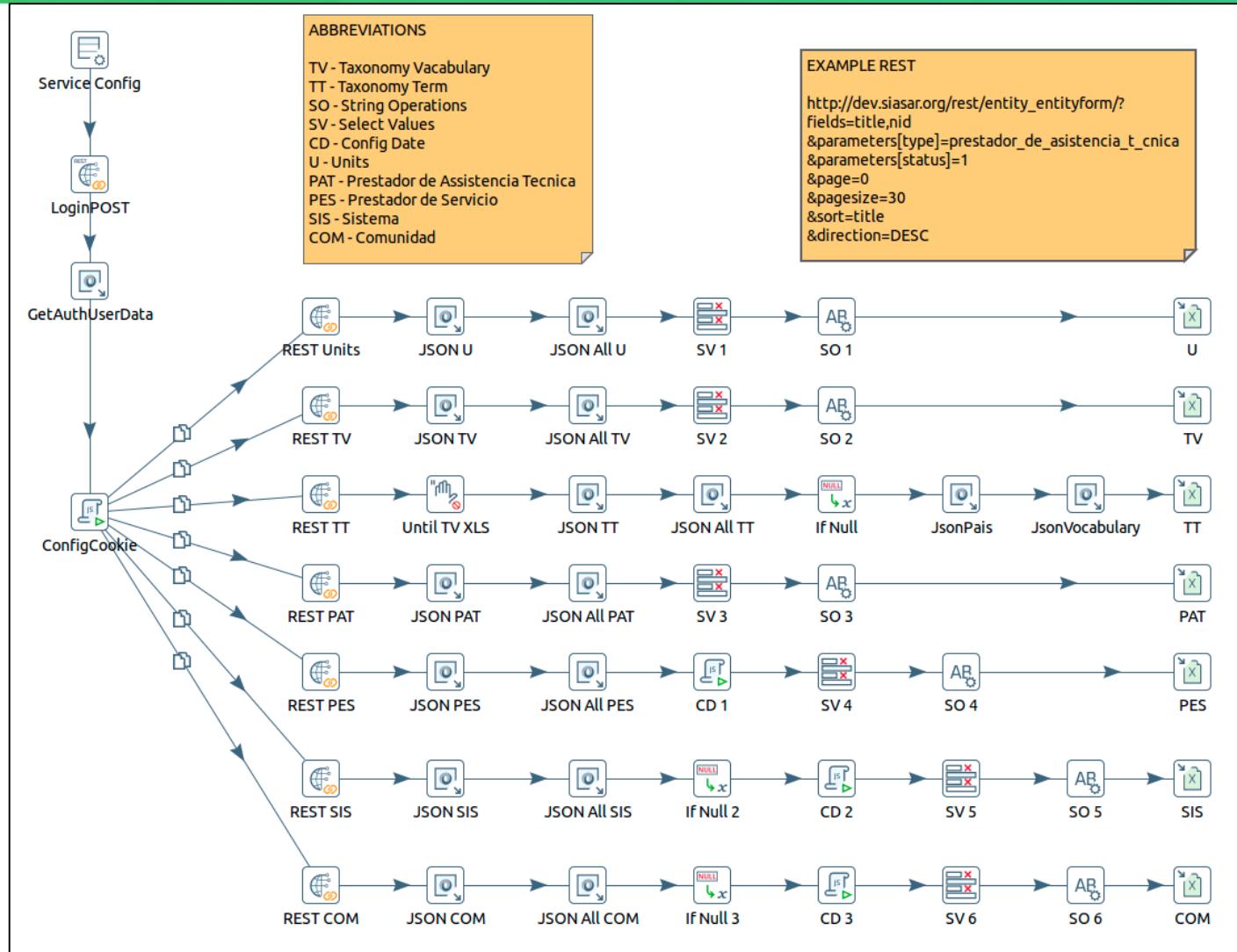
- ❖ Usando informações de auditoria
 - capturar alterações. Mostrar a hora em que o registro foi adicionado ou modificado.
- ❖ Minerar ou detectar logs de banco de dados
 - Tira uma foto instantânea da detecção de alterações.
- ❖ Extrações cronometradas
 - seleção de dados com base no tempo.
- ❖ Processo de eliminação
 - Armazena uma cópia de cada extração anterior na Staging Area e compara linha por linha. Apenas dados diferentes são enviados.
- ❖ Sistema de notificação
 - Se o sistema de origem puder fornecer uma notificação de que um registro foi alterado e descrever a alteração.

Sistema de Extração: Fontes de Dados

- ❖ Arquivos
- ❖ Banco de dados
- ❖ Web Service
- ❖ *Big Data*
- ❖ *Streaming*
- ❖ Outros



Sistema de Extração: Fontes de Dados



Extraindo dados de arquivos

- ❖ Arquivos planos
- ❖ Dados semiestruturados e não estruturados
- ❖ Arquivos:
 - TXT
 - CSV
 - EXCEL
 - XML
 - JSON
 - FIXED
- ❖ Principais propriedades dos steps de extração:
 - Nome e localização do arquivo
 - Descrição do conteúdo
 - Separador, codificação, cabeçalho, etc.
 - Depende do tipo do step
 - Campos
 - Filtros:
 - Pular linhas em branco, ler as primeiras n linhas, etc.
 - Uso de expressões regulares
 - Ler vários arquivos

Seleção de dados

❖ Operações básicas:

- Selecionar e Alterar campos
- Remover Campos
- Alterar metadados dos campos

❖ Uso de parâmetros:

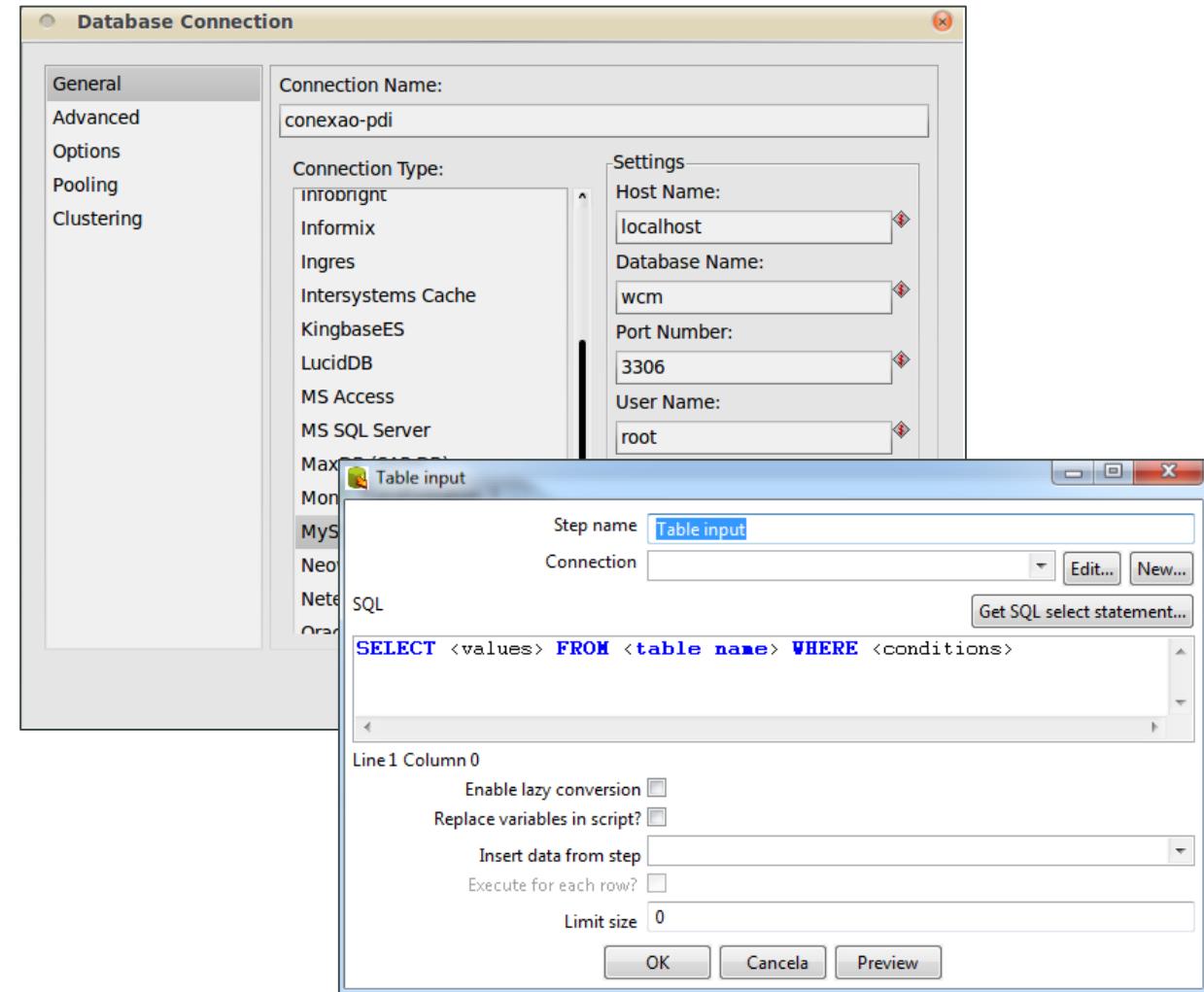
- Controle
- Filtro
- Variáveis
- Constantes
- Internacionalização

❖ Tipos de dados:

- *Number*
 - Tamanho e precisão
- *Date*
 - Especificar máscara. Ex: yyyy/MM/dd
- *String*
- *Boolean*
- *Integer*
- *Big Number*
- *Binary*

Extraindo dados de Banco de Dados

- ❖ Adicionar adaptadores do Banco.
- ❖ Configurar uma conexão.
- ❖ Linguagem de Consulta de Dados
 - (DQL – Data Query Language)
- ❖ Principais steps:
 - Table input
 - Lookup
 - Scripting



Extraindo dados de Web Service

- ❖ Permite a interoperabilidade de dados entre aplicações;
- ❖ Padrões:
 - **SOAP** (*Simple Object Access Protocol*)
 - Usa o arquivo **WSDL** para descrever as estruturas da comunicação.
 - Formato **XML**
 - **REST** (*Representation State Transfer*)
 - Formato **JSON** (JavaScript Object Notation)
- ❖ Podem ser realizados usando um step HTTP/HTTPS
 - Se necessário alterar o cabeçalho de comunicação.



LAB 2 – Extração de Dados

Diversas fontes de dados

* Arquivo lab2_extrair.pdf

Transformação de Dados

Validar, Limpar, Agregar, Regra de negócio e Qualidade dos dados

Transformação de Dados

Processo de aplicação de uma série de regras de negócio ou funções sobre os dados extraídos para prepará-los para o carregamento no destino final;

- ❖ Segunda etapa do processo ETL;
- ❖ É a etapa mais importante e opcional;
- ❖ Agrega valor aos dados;
- ❖ Produzir dados limpos, condensados, novos, completos e padronizados, respectivamente;

Transformação de Dados



❖ Limpar

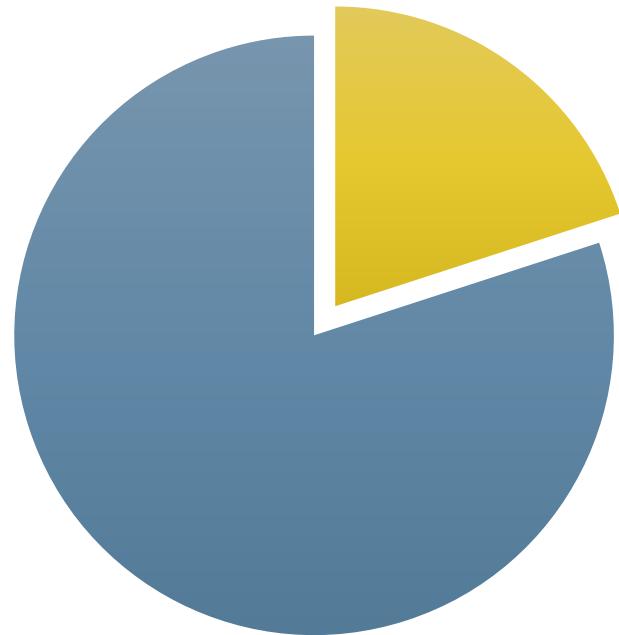
- significa **identificar e corrigir** os erros e omissões nos dados.

❖ Conformar

- significa resolver os **conflictos** para que dados possam ser usados juntos em um Data Warehouse.

Transformação de Dados

- ❖ 80% aplicando regras de negócio e de integridade de dados.



- ❖ 20% das regras de conversão dos dados.

Desafios da transformação

- ❖ Diversidades de fontes:
 - Baixa qualidade em legados;
 - Reconciliar/padronizar;
 - Boa fonte = pouca transformação.
- ❖ Dados conforme a necessidade do negócio;
- ❖ Definem a forma de **transporte** dos dados:
 - Orquestrar o Fluxo de dados;
 - Dados devem passar **áreas intermediárias**.
- ❖ Aplicam **funções** que agregam valor aos dados.



Problemas (conflitos)



❖ Nível de Schema

a) Conflitos de nomenclatura

- Onde o mesmo nome é usado para objetos diferentes (homónimos) ou nomes diferentes são usados para o mesmo objeto (sinónimos)

b) Conflitos estruturais

- Lidar com diferentes representações do mesmo objeto em diferentes fontes, ou converter tipos de dados entre fontes e o DW.

Problemas (conflitos)



- ❖ Nível de Registro
 - a) Registros **duplicados**
 - b) Registros **contraditórios**
 - c) Registros **inconsistentes**
 - a) Desacordo com regra de negócio e domínio

Problemas (conflitos)



❖ Nível de Valores

- a) Diferentes **representações** de valores
 - Para o sexo: “Homem”, “M” ou “1”

- a) Diferentes **interpretações** de valores
 - Americano “mm/dd/yy” vs. Europeu “dd/mm/yy”

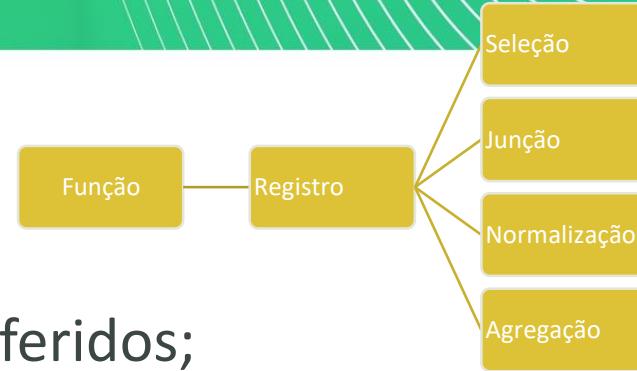
Funções de transformação



Funções de transformação

❖ Funções ao nível de Registro

- Seleção
 - Os dados são particionados de acordo com os critérios preferidos;
- Junção
 - Junção de dados de várias fontes;
- Normalização
 - Decomposição de relações com anomalias para produzir relações menores, bem estruturados;
- Agregação
 - Transformar dados de um nível detalhado para um nível de resumo.



Funções de transformação

❖ Funções ao nível de Campo

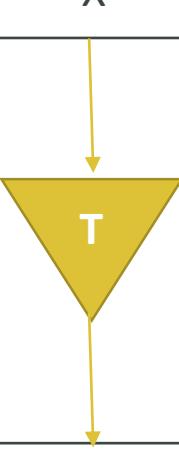
- Função de **campo único**

- Converte dados de um campo de fonte único para um campo de destino único.



Registro de Origem

ID			X					
----	--	--	---	--	--	--	--	--



Registro de Destino

ID			F(x)					
----	--	--	------	--	--	--	--	--

Funções de transformação

❖ Funções ao nível de Campo

- Função de **diversos campos**

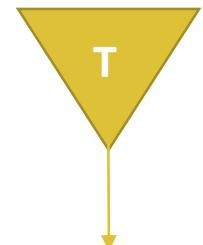
- converte dados de um ou mais campos de origem para um ou mais campos de destino.



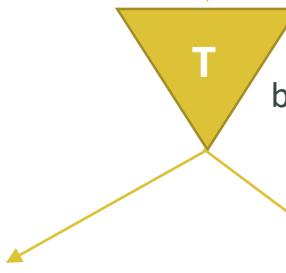
Registro de Origem

ProdID	Nome		Data		Codigo	
--------	------	--	------	--	--------	--

A) Muitas origens para um destino



b) Uma origem para vários destinos



Registro de Destino

ProdID		Codigo		Marca		Loja	
--------	--	--------	--	-------	--	------	--

Transformação na prática...

❖ Selecionar

- Apenas determinadas colunas para carregar ou remover.
- Filtrar apenas registros não nulo.

❖ Tradução de valores codificados:

- Ex: 1 para “Masculino” e 2 para “Feminino”

❖ Codificando valores de forma livre:

- Ex: mapeando “Masculino”, “1” e “Sr.” para M

❖ Derivação de um novo valor calculado

- Ex: montante_vendas = qtde * preço_unitário

Transformação na prática...

- ❖ **Classificando ou ordenando** os dados com base em uma lista de colunas para melhorar o desempenho da pesquisa;
- ❖ **Junção** de dados de várias fontes (**Lookup e Merge**) e de dados duplicados;
- ❖ **Agregação**
 - Ex: resumindo várias linhas de dados: vendas totais para cada loja e para cada região etc. Contabilizar quantidade de filmes em uma categoria.
- ❖ **Geração de valores** de chaves substitutas (*surrogate keys*)

Transformação na prática...

❖ Transposição ou rotação

- Ex: transformando várias colunas em várias linhas ou vice-versa;

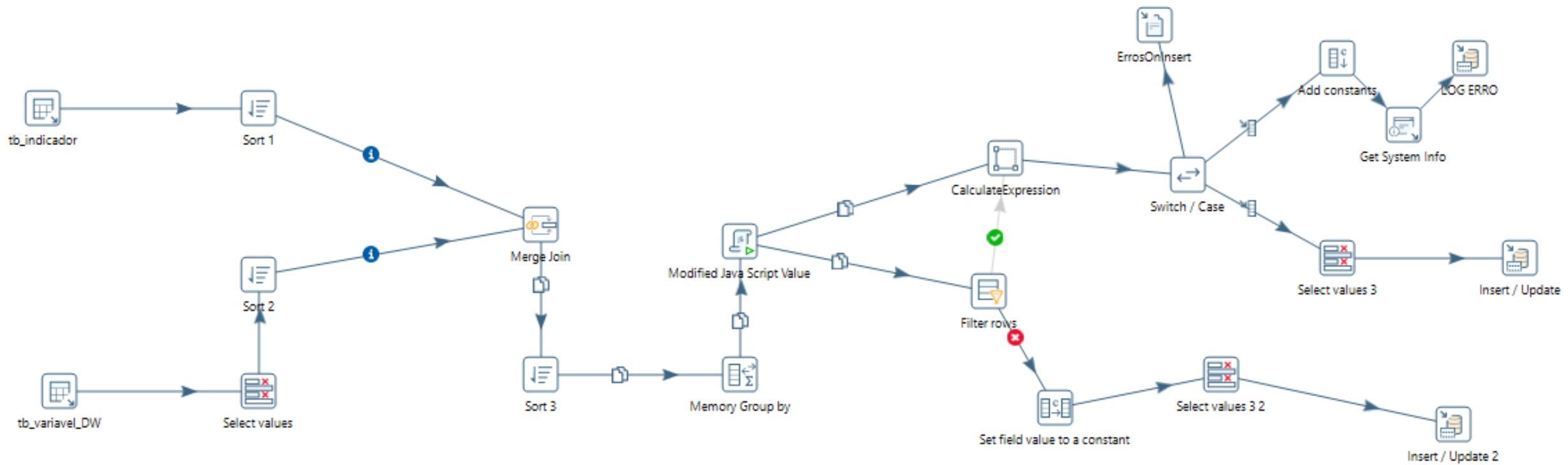
❖ Dividir uma coluna em várias colunas

❖ Desagregação de colunas repetidas em uma tabela separada com detalhes separada

- Ex: movendo uma série de telefone1, telefone2, telefone3 para uma tabela separada.

❖ Aplicar qualquer forma de validação de dados;

Exemplo de Transformação



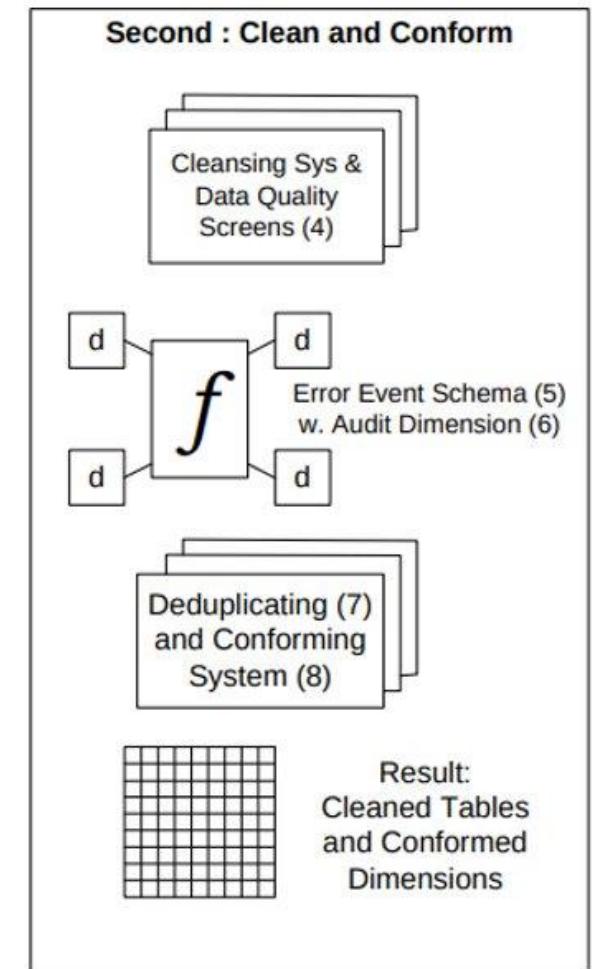
Subsistemas para DW (Kimball)

- ❖ Limpeza dos Dados
- ❖ Qualidade dos Dados
- ❖ Acompanhamento de Erro
- ❖ Criação de Dimensão de Auditoria
- ❖ Remover duplicidades dos dados
- ❖ Conformidade dos dados

Cleaning machinery →

Cleaning control →

Integration →



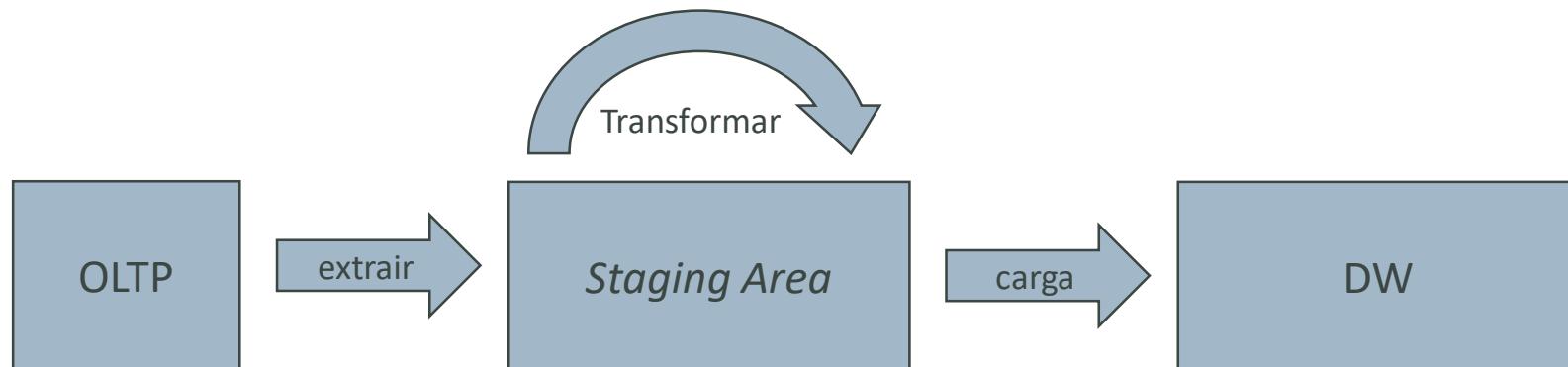
Carga de Dados

Carregar dados para DW

Carga de Dados

Carrega os dados ao **destino final**, que pode ser qualquer armazenamento de dados, incluindo um arquivo simples até um armazém de dados (DW).

- ❖ É a última etapa do processo ETL;
- ❖ A fase de colocação dos dados no Data Warehouse (DW);
- ❖ Entrega os dados nas tabelas de **dimensão** e **fato** de acordo com a **modelagem dimensional**;



Carga de Dados

- ❖ O processo de ETL termina quando o modelo dimensional for: **atualizado, indexado, provido das agregações** apropriadas e teve sua **qualidade garantida**;
- ❖ Ao final, o processo deve **notificar à comunidade** de negócios que os dados mais novos foram publicados;
- ❖ Depois dessa etapa o dados ficam disponíveis para a **Área de Apresentação**;

Classificação da Carga

❖ *Full Load*

- Os dados que se encontram no sistema de destino serão apagados (*truncate* se for necessário) e carregados os que foram extraídos para esse destino;

❖ *Incremental*

- Quando os dados do destino não estiverem atualizados, para que haja uma atualização de somente os dados em falta.

Modelagem Dimensional

- ❖ Dados são organizados e estruturados em diversas **dimensões** e eventos de negócio (**fatos**) composto de **medidas** atreladas ao contexto.
- ❖ Para modelagem dimensional utilizamos o *Star Schema* ou Esquema Estrela.



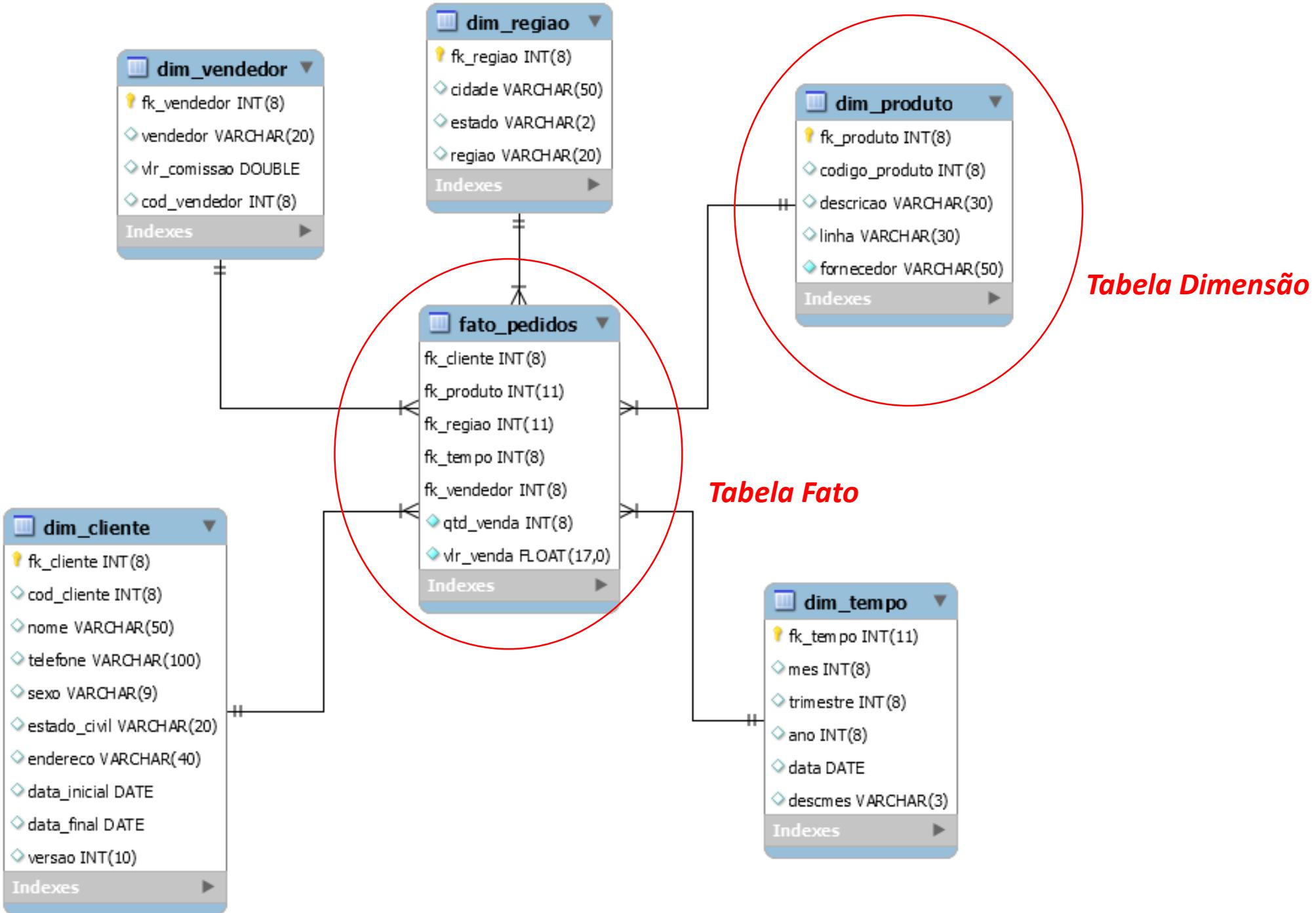
Componentes do Esquema Estrela

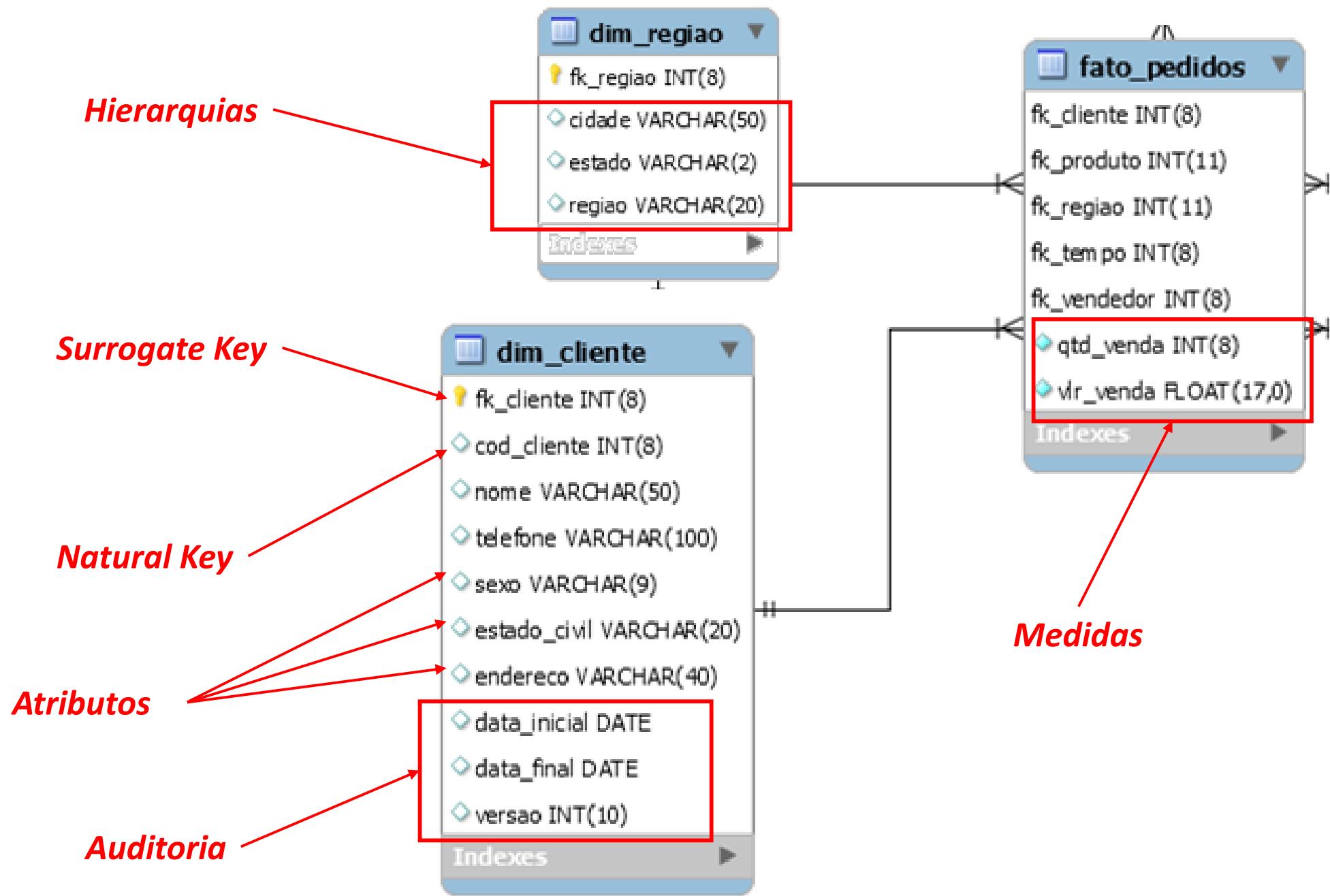
1. Tabela de Fato

- Fato é um **evento** de negócio;
- 1 linha na tabela = 1 evento registrado;
- **Medidas**: algo que pode ser sumarizado/agregado/contado;
- **Granularidade**: nível de um item, uma transação, evento ou agrupado destes;

2. Tabela de Dimensão

- O **contexto** associado ao evento de negócio ocorrido;
- Quem, o que, onde, quando, como e porque (os “por”);
- Membros de uma dimensão: Hierarquia
 - Classificação de dados dentro de uma mesma dimensão;





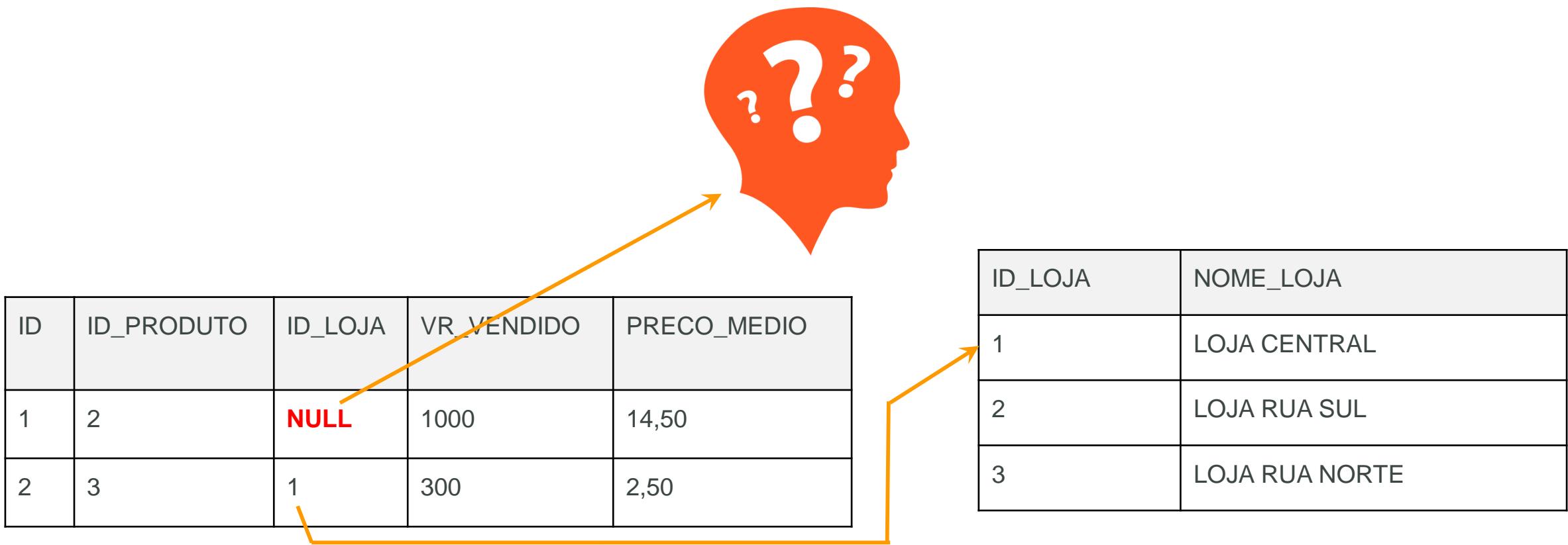
Estrutura de uma Tabela de Fatos

- ❖ Contém **medidas** numéricas vindas de **eventos** do mundo real;
- ❖ Uma **linha da fato** corresponde a uma **medição feita**;
- ❖ Para cada uma das **dimensões** associadas à tabela de fato, existe uma coluna contendo **chaves estrangeiras**;
- ❖ As tabelas são o principal alvo de **computações** e **agregações** dinâmicas requisitadas pelas consultas.

Valores Nulos na Tabela de Fato

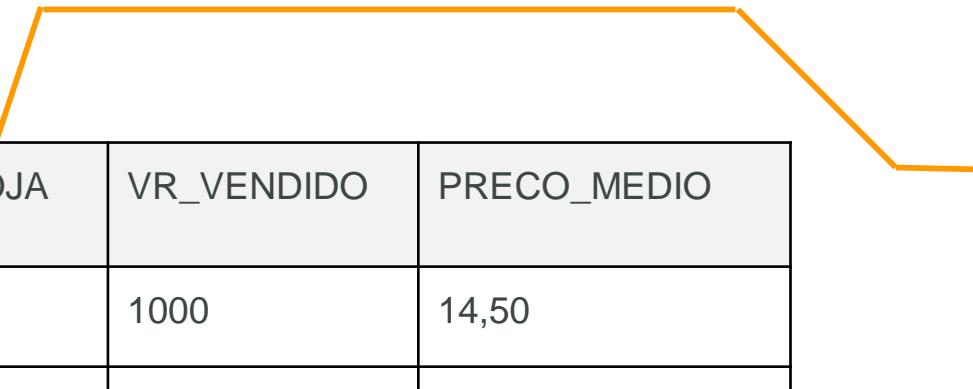
- ❖ SUM, COUNT, MIN, e MAX se **comportam como esperado** quando encontram valores nulos, ignorando esses valores.
- ❖ Mas, valores nulos nas colunas de chaves estrangeiras da fato causam, automaticamente, **violação da integridade de referências**
 - A chave estrangeira não corresponde a nenhuma chave na tabela de dimensão.
- ❖ **Crie uma linha padrão na tabela de dimensão** que representa o desconhecido ou o inaplicável.

Valores Nulos na Tabela de Fato



Valores Nulos na Tabela de Fato

ID	ID_PRODUTO	ID_LOJA	VR_VENDIDO	PRECO_MEDIO
1	2	0	1000	14,50
2	3	1	300	2,50

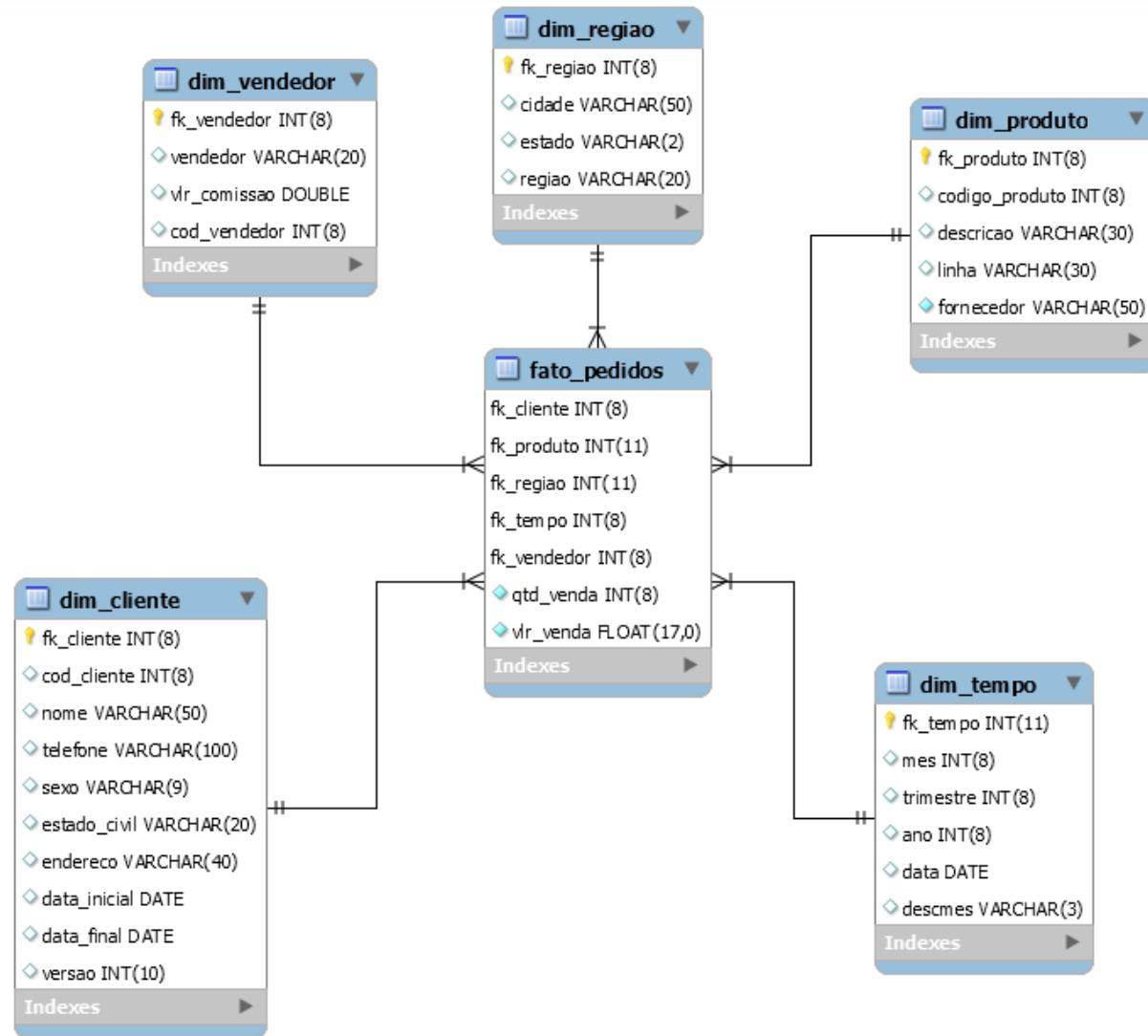


ID_LOJA	NOME_LOJA
0	NÃO SE APLICA
1	LOJA CENTRAL
2	LOJA RUA SUL
3	LOJA RUA NORTE
99999	DESCONHECIDO

Tipos de Tabela de Fatos

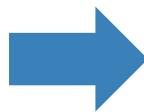
	Transacional	Snapshot Periódico	Snapshot Acumulado
<i>Peridiocidade</i>	Ponto de transação discreta no tempo	Instantâneos recorrentes em intervalos regulares e previsíveis	Intervalo de tempo indeterminado para evolução de pipeline / fluxo de trabalho
<i>Grão</i>	1 linha por transação ou linha de transação	1 linha por período de snapshot	1 linha por ocorrência de pipeline
<i>Dimensão Data</i>	Data da Transação	Data do Snapshot	Data para principais marcos do pipeline
<i>Atualizações da Tabela de Fatos</i>	Sem atualizações, só correção de erros.	Sem atualizações, só correção de erros.	Atualizado sempre que ocorre atividade de pipeline

Tabelas de Fato Transacionais



Tabelas de Fato de *Snapshot* Periódico

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

Tabelas de Fato de *Snapshot* Acumulativo

Estado Inicial de Recebimento do Pedido

Lote Recebido	Data de Recebimento	Data Inspeção	Data de Entrega	Chave Produto	Quantidade Recebida	Lag Recebimento a Inspeção	Recebido a Entregue Lag
101	20130101	0	0	1	100		

Inspeção do Pedido e Separação Para a Entrega

Lote Recebido	Data de Recebimento	Data Inspeção	Data de Entrega	Chave Produto	Quantidade Recebida	Lag Recebimento a Inspeção	Recebido a Entregue Lag
101	20130101	20130103	0	1	100		2

Produto Entregue No Cliente

Lote Recebido	Data de Recebimento	Data Inspeção	Data de Entrega	Chave Produto	Quantidade Recebida	Lag Recebimento a Inspeção	Recebido a Entregue Lag
101	20130101	20130103	20130104	1	100		2

3

Estrutura de uma Tabela de Dimensão

- ❖ Toda tabela de dimensão deve possuir apenas **uma coluna de chave primária**;
- ❖ A **chave primária** deve estar **associada** às linhas da **tabela de fatos** através de chaves estrangeiras;
- ❖ **Rótulos de relatórios**, geralmente, veem dos **atributos** das dimensões;
- ❖ Os **atributos** servem para **filtrar, agrupar e restringir** as consultas de usuários e aplicações de BI.

Estrutura de uma Tabela de Dimensão

❖ Chave Substituta (Surrogate key) das Dimensões:

- Crie uma chave substituta que consiste de um simples inteiro atribuído em sequência do 1 até onde for necessário;

❖ Hierarquias:

- Dimensões podem conter mais uma hierarquia natural

■ Exemplo:

- Dimensão tempo (dia, mês, trimestre, semestre, ano)
- Dimensões de localização (cidade, estado, país, continente)

❖ Dimensão Tempo:

- Uma dimensão quase onipresente em DWs e projetos de BI;
- Permite a navegação dos dados através dos anos, meses, dias, etc.

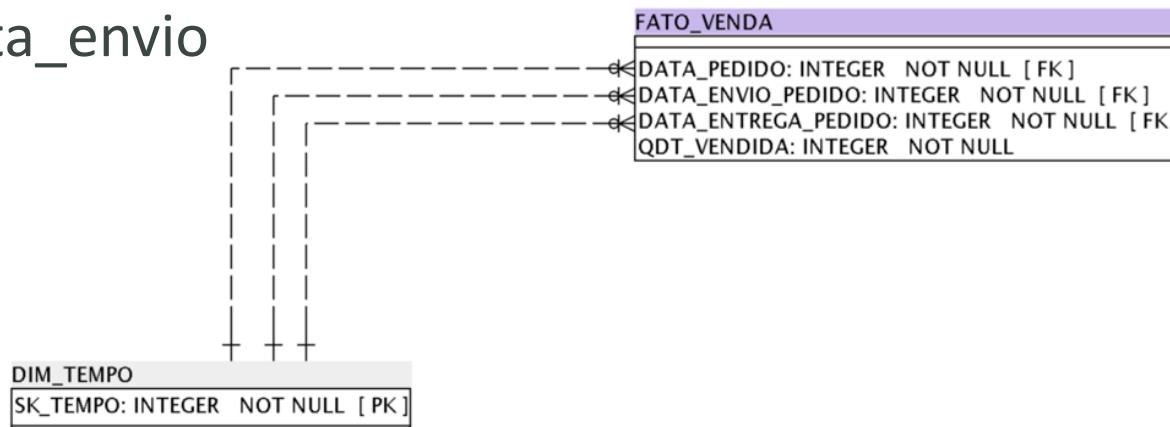
Estrutura de uma Tabela de Dimensão

❖ Dimensões Degeneradas:

- É a dimensão que não mereceu ser uma tabela dimensão e foi inserida como coluna na fato;
 - Exemplo: numero_fatura

❖ Dimensões *Role-Playing*:

- Uma dimensão física com múltiplas referências na tabela de fato.
 - Exemplo na tabela fato: data_pedido, data_envio



Estrutura de uma Tabela de Dimensão

❖ *Junk Dimensions:*

- Os processos de negócio podem conter conjuntos de dados, como flags e tipos, de domínio e de baixa cardinalidade;
- Reduz a quantidade de joins.

Gênero
M
F
I

Tipo de Promoção
Desconto
Compre um leve outro
Degustação

Combinações



ID	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

Estrutura de uma Tabela de Dimensão

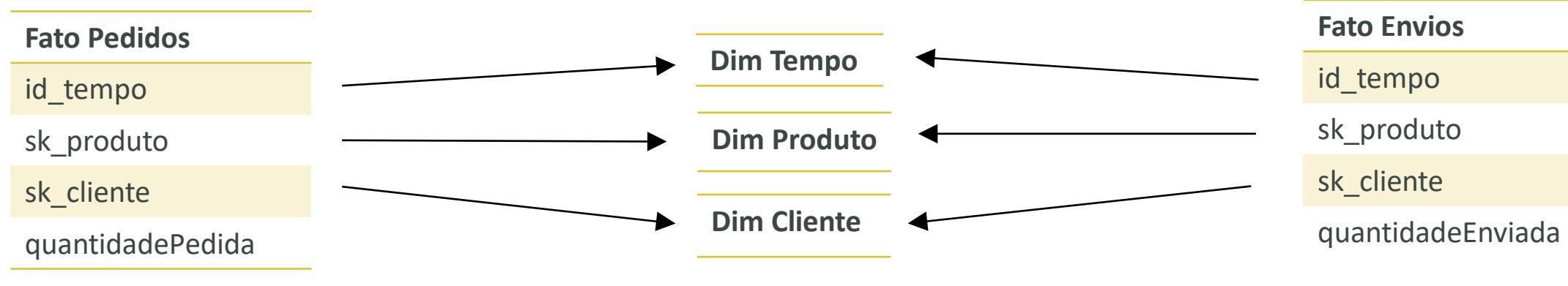
❖ Dimensões Snowflake:

- Dimensões cujas estruturas hierárquicas estão implementadas de forma normalizada;



Estrutura de uma Tabela de Dimensão

❖ Dimensões Conformadas:



Produto	Quantidade Pedida	Quantidade Enviada
Produto 1	125	125
Produto 2	250	200
Produto 3	400	0

Carga em Dimensões

❖ *Slowly Changing Dimensions (SCD)*

- Uma técnica para atualizar os atributos da tabela dimensão.
- **Tipo 0:** Retenha o original
 - Depois que os dados foram inseridos, não precisa mais atualizar.
 - Ex: dimensão de tempo
- **Tipo 1:** Sobrescreva
- **Tipo 2:** Adicione uma Nova Linha
- **Tipo 3:** Adicione um Novo Atributo

Carga em Dimensões

❖ Tipo SCD 1: Sobrescreva

Nome do Cliente, Data Nascimento

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

Exemplos:

Leonardo Luís → Leonardo Luiz
Jon Snow → John Snow

Carga em Dimensões

❖ Tipo SCD 2: Adicione uma Nova Linha

- Novas linhas são adicionadas contendo os valores atualizados
- Cada nova linha criada possui uma nova chave substituta
- Adicione três colunas na tabela de dimensão:
 - data efetiva da linha
 - data de validade da linha
 - indicador de linha atual (ou número da versão).

Exemplos:

Recife → Rio de Janeiro
Rio de Janeiro → Brasília

Solteiro → Casado
Casado → Divorciado

dim_cliente	
fk_cliente	INT(8)
cod_cliente	INT(8)
nome	VARCHAR(50)
telefone	VARCHAR(100)
sexo	VARCHAR(9)
estado_civil	VARCHAR(20)
endereco	VARCHAR(40)
data_inicial	
data_final	
versao	

Carga em Dimensões

❖ Tipo SCD 2: Adicione uma Nova Linha

- Inserir

cod_cliente	nome	estado_civil
1123	Fulando da Silva	Solteiro



fk_cliente	cod_cliente	nome	estado_civil	data_ini	data_fim	versao
1	1123	Fulando da Silva	Solteiro	03/08/2016	31/12/2099	1

- Atualizar

cod_cliente	nome	estado_civil
1123	Fulando da Silva	Casado



fk_cliente	cod_cliente	nome	estado_civil	data_ini	data_fim	versao
1	1123	Fulando da Silva	Solteiro	03/08/2016	07/12/2019	1
1	1123	Fulando da Silva	Casado	07/12/2019	31/12/2099	2

Carga em Dimensões

❖ **Tipo SCD 3: Adicione um Novo Atributo**

- É criado uma nova coluna na tabela de dimensão
- O valor antigo do atributo SCD 3 é colocado nessa nova coluna
- O novo valor do atributo é colocado na coluna original
- Também chamada de *Realidade Alternativa*

Exemplos: Tel Antigo
55 85 5555.5555

Tel Atual
55 85 4444.4444

Gerenciamento de ETL



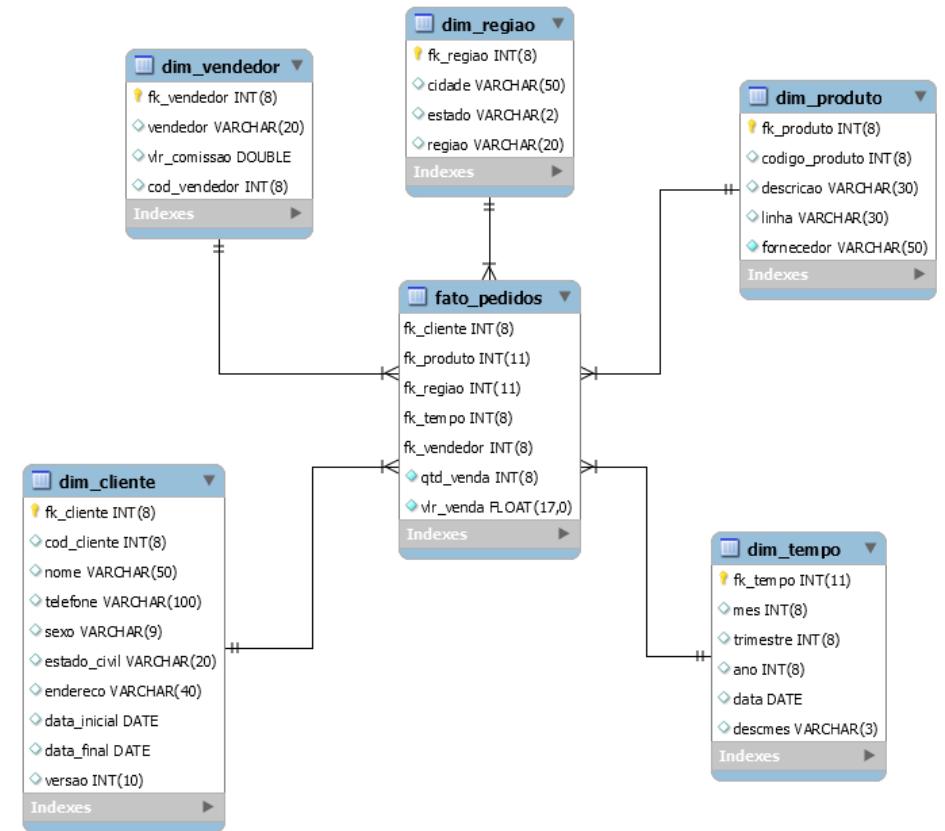
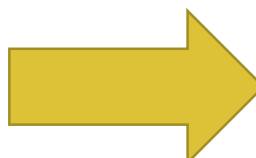
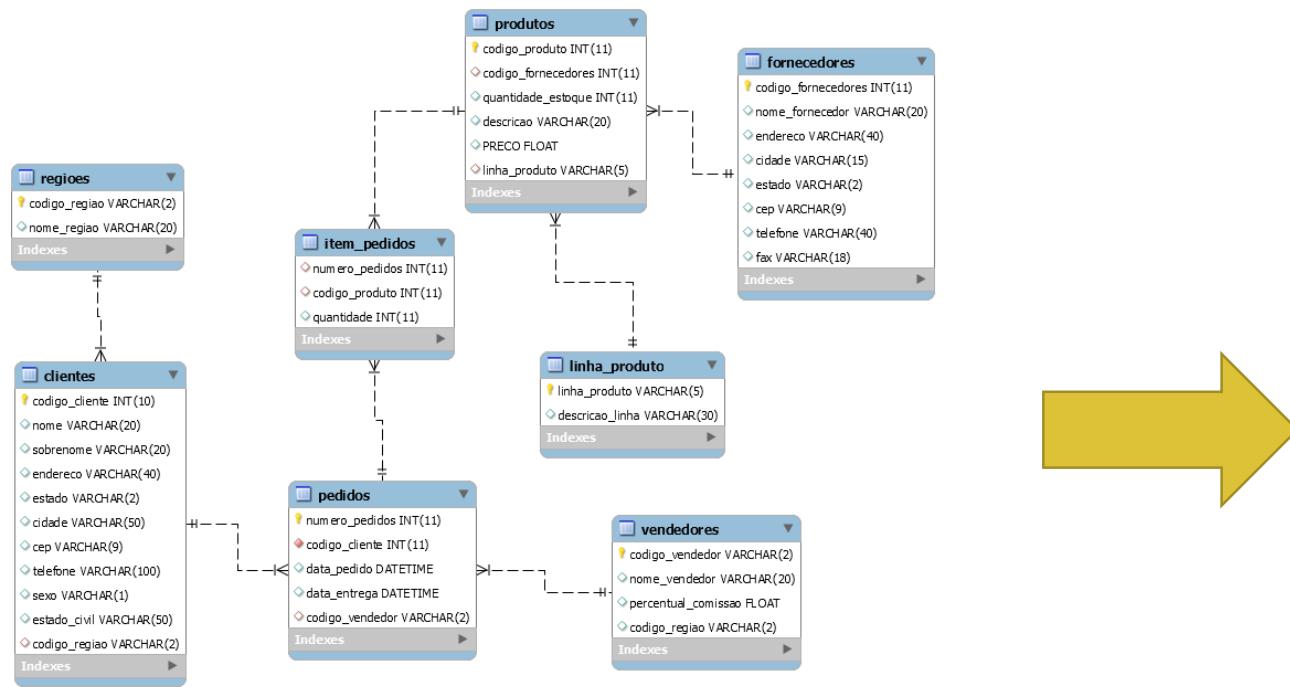
- ❖ Um DW de **sucesso** é uma fonte **confiável** para as tomas de decisões nas empresas;



- ❖ Agendamento de tarefas
- ❖ Sistema de Backup
- ❖ Recuperação e *Restarts*
- ❖ Controle de Versão
- ❖ Migração de Versão
- ❖ Monitoramento do fluxo de trabalho

LAB IV – Carga de Dados

- ❖ Migração de Dados
- ❖ Carga no DW



LAB IV – Carga de Dados

