INE 5643
Data Warehouse
Aula 10b - Back Room - Extração

Prof. Mateus Grellert
Prof. Renato Fileto
Créditos: Prof. Tite Todesco **(slides originais, adaptados pelos professores atuais)**

Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)

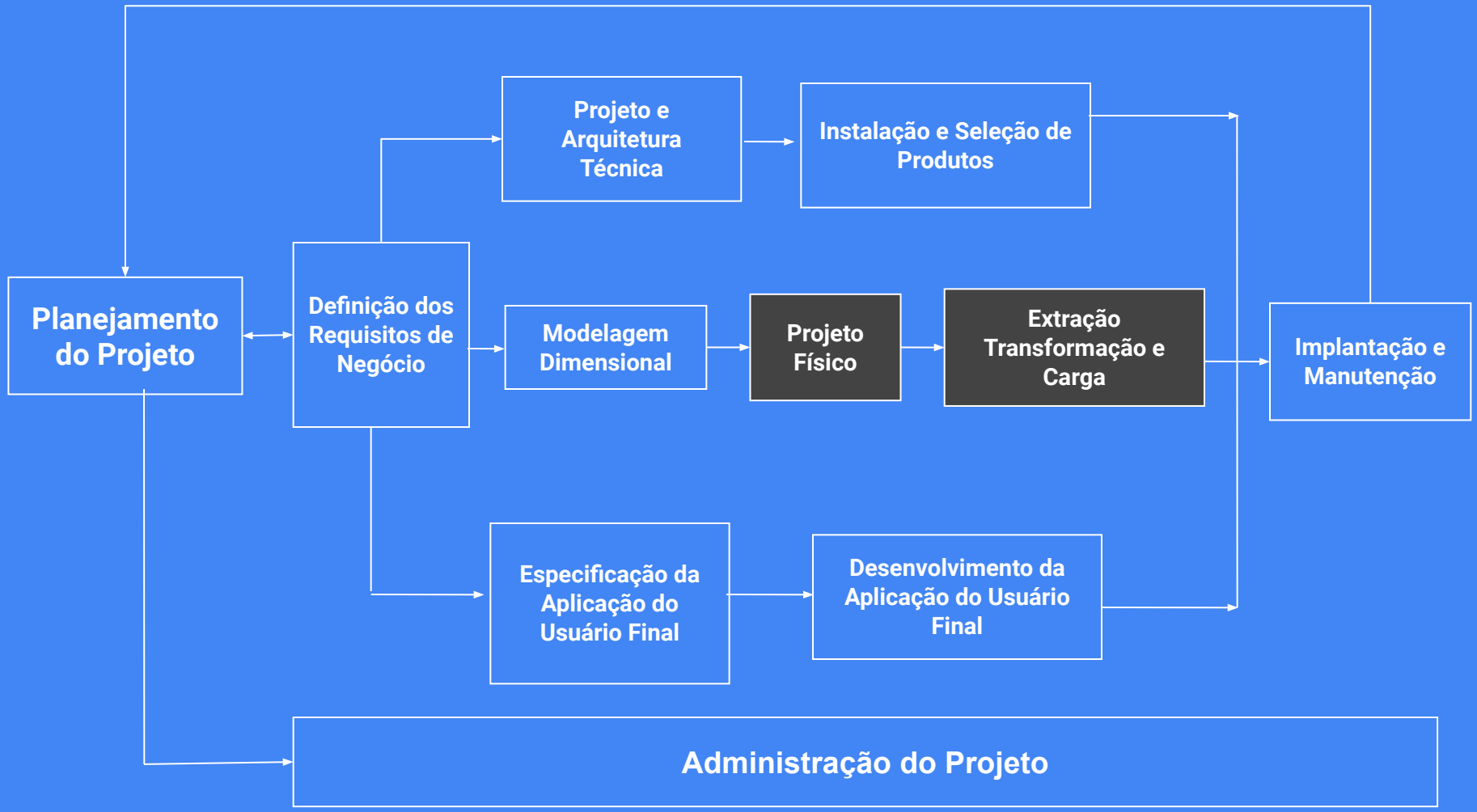
TÓPICOS

- Projeto Físico
- Área de Transição

TÓPICOS

- **Projeto Físico**
- **Área de Transição**

Próxima Aula - Ciclo de Projeto DW



O Projeto Físico

- Seleção do SGDB;
- Modelo físico de dados;
- Plano de Indexação e Particionamento;
- Projeto Inicial de Agregados.

Considerações sobre o Modelo Físico

DEFINIÇÃO DE PADRÕES

RELACIONE OS OBJETOS DO BANCO DE DADOS UTILIZADO E DEFINA OS PADRÕES DE NOMENCLATURA A SEREM ADOTADOS PARA: TABELAS, COLUNAS, ÍNDICES, ÁREA ETC

PROJETO FÍSICO DAS TABELAS E COLUNAS

APLIQUE OS PADRÕES DE NOMENCLATURA DO BANCO DE DADOS SOBRE OS OBJETOS DO MODELO LÓGICO

DEFINA OS *DATA TYPES*

DEFINA *NULL* OU *NOT NULL*

DEFINA A *PRIMARY KEY* E *FOREIGN KEYS*

ESTIMATIVA DO TAMANHO DO BANCO DE DADOS

APURE A QTDE DE TABELAS, TAMANHO DAS LINHAS, QTDE DE LINHAS, QTDE DE ÍNDICES, ÁREA DE ÍNDICE, ÁREA PARA METADADOS E A TAXA DE CRESCIMENTO

DEFINIÇÃO DE ÍNDICES

DEFINA AS REGRAS A SEREM UTILIZADAS PARA CRIAÇÃO DE ÍNDICES

A DECISÃO ACERCA DA CRIAÇÃO DE ÍNDICES DEVE ESTAR TOTALMENTE VINCULADA AO TIPO DE ÍNDICE UTILIZADO PELO RDBMS (B-TREE, BITMAP, HASH, OUTROS)

DEFINIÇÃO DE PARTIÇÕES

DEFINA AS REGRAS PARA CRIAÇÃO DE PARTIÇÕES DE TABELAS; CONSIDERAR A QUANTIDADE DE AGREGAÇÕES UM FATOR INTERVENIENTE PARA A DEFINIÇÃO DESTAS REGRAS

ESPECIFIQUE OS CRITÉRIOS PARA DETERMINAR UNICAMENTE EM QUAL PARTIÇÃO UMA LINHA DEVE ESTAR LOCALIZADA

Projeto Físico: Modelo Físico

- Ponto de Partida: Modelo Lógico e seguir **padrões de nomenclatura;**
- Utilizar a ferramenta de modelagem de dados padrão;
- Projeto da Estrutura Física dos Dados:
  - Determinar os **tipos** de dados para as colunas;
  - Determinar opções de **NULL/NOT NULL;**
  - Chaves naturais devem ser substituídas por chaves artificiais (**surrogate keys - SKs**);
  - Especificação de chave primárias e secundárias.

Projeto Físico: Padronização da Nomenclatura

- Nome para objetos da base de dados
  - Identificador do objeto: **Conta**;
  - Tipo de objeto: **Data**;
  - Qualificadores (opcional): **Inicial**;

  Exemplo: Data_Inicial_Conta

- Similaridade dos nomes nos vários ambientes
- Definição em conjunto com a comunidade de usuários

Projeto Físico: Estimativas de Tamanho

- Qual será o tamanho do data warehouse?
- Uma estimativa é importante para:
  - dimensionamento de máquina (poder de processamento);
  - área de armazenamento necessária.

Projeto Físico: Estimativas de Tamanho

VOLUMES (exemplo) Estimativa de Ocorrências Iniciais	
Data	**1.825 (5 anos)**
Hora	**24**
Produto	**5.000**
Loja	**2.500**
Cliente	**200.000**
Fato Atômico	**1,1 bilhão**
Agregação 1	**50 milhões**
Agregação 2	**35 milhões**

ESTRATÉGIAS DE IMPLEMENTAÇÃO
ROLAP
MOLAP
HOLAP

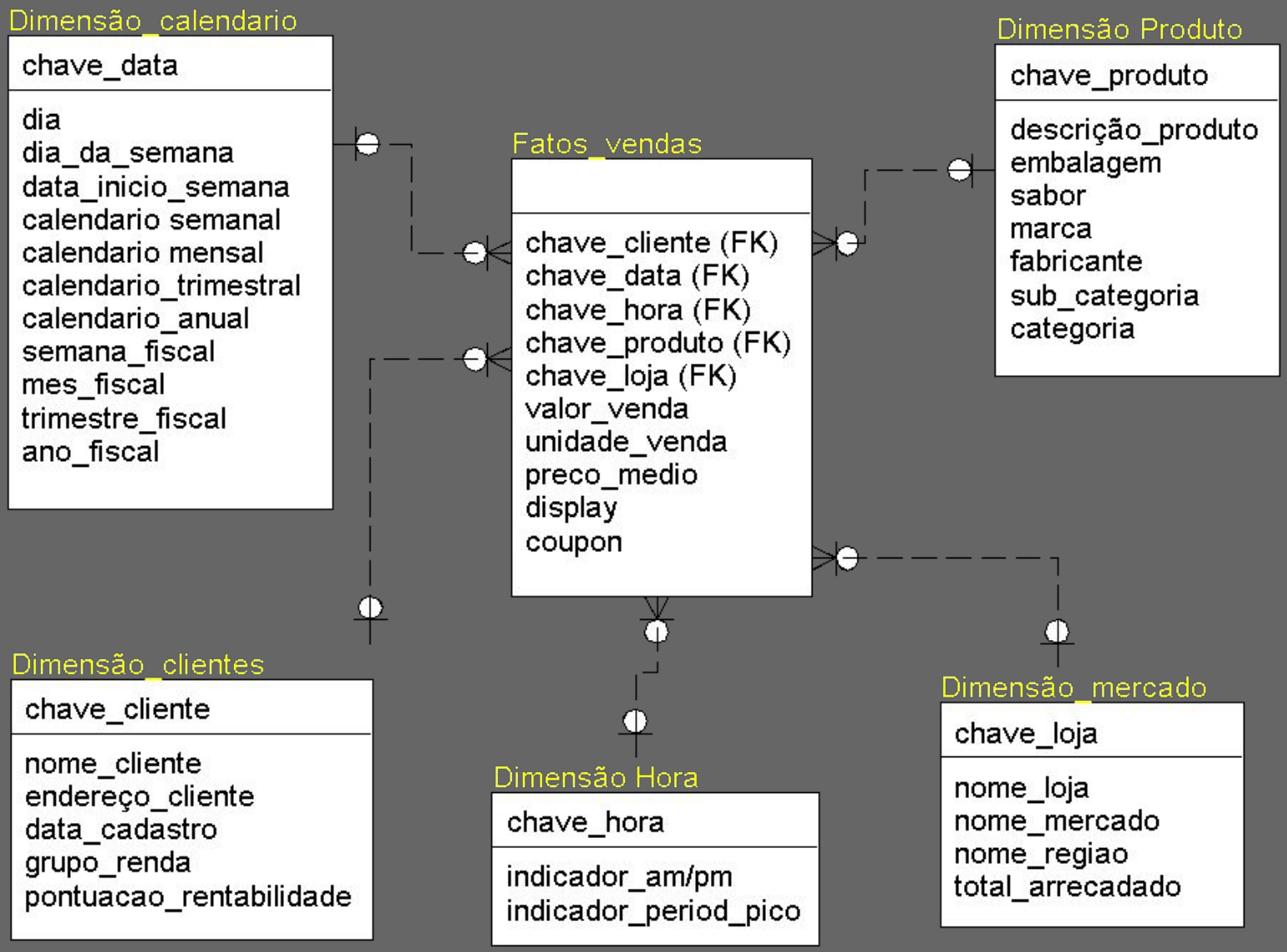
TAMANHO ESTIMADO		
	Inicial	**Final (6 meses)**
Dimensões	*33,5 Mb*	*76,5 Mb*
Fatos	*78 Gb*	*180 Gb*
Índices	*105 Gb*	*135 Gb*
Espaço Temporário	*73 Gb*	*90 Gb*
Total	***± 256 Gb***	***± 405 Gb***

Projeto da Agregação

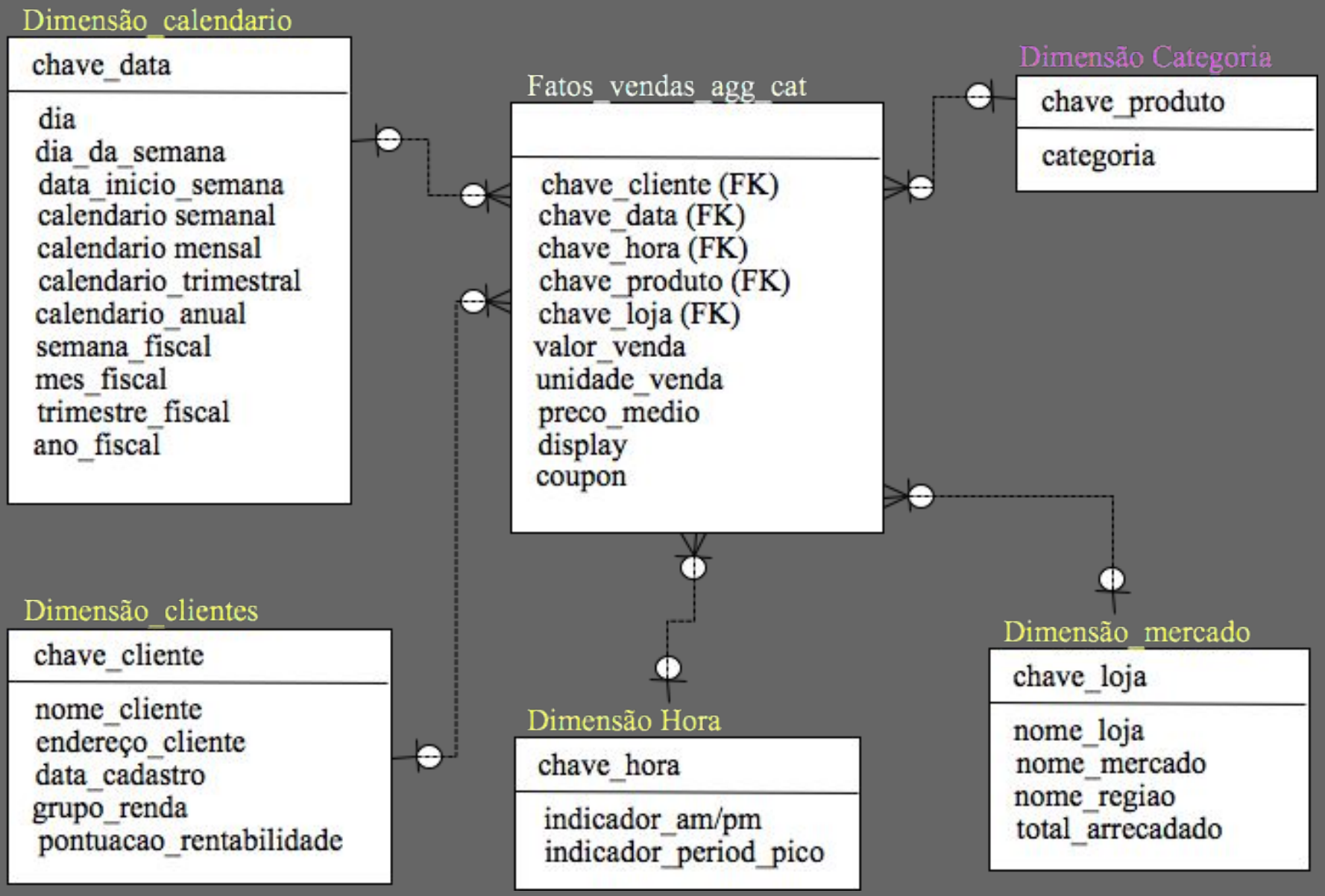
PROJETO FÍSICO: AGREGADOS

- Definir o que deve ser agregado.
- Pontos de Avaliação:
  - Requisitos das consultas **mais freqüentes**;
  - Considerar a distribuição estatística dos dados;
- Definir tabelas de fato agregados e dimensões;

Exemplo: Sem Agregação



Exemplo: Agregado para Categoria



Exemplo de substituição

```
select p.categoria, sum(f.qtde_venda)
from fatos_vendas f, dimensao_produtos p,
        dimensao_tempo t, dimensao_mercado l
where f.chave_produto = p.chave_produto and
           f.chave_loja = l.chave_loja and
           f.chave_cal = t.chave_cal and
           t.dia_da_semana = ‘Sábado’ and
           l.regiao = ‘SUL’
group by p.categoria
```

Exemplo de substituição

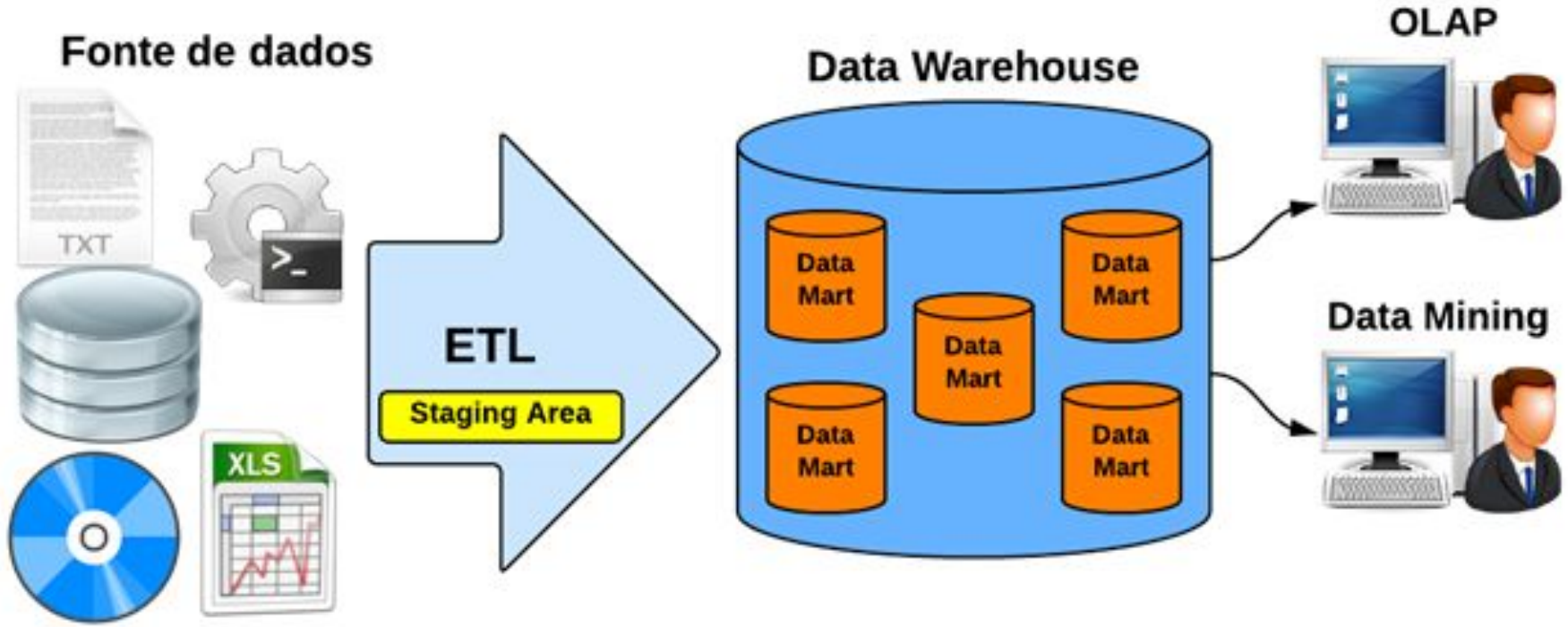
```
select p.categoria, sum(f.qtde_venda)
from fatos_vendas_agg_cat f, dimensao_categoria p,
        dimensao_tempo t, dimensao_mercado l
where f.chave_produto = p.chave_produto and
           f.chave_loja = l.chave_loja and
           f.chave_cal = t.chave_cal and
           t.dia_da_semana = ‘Sábado’ and
           l.regiao = ‘SUL’
group by p.categoria
```

TÓPICOS

- Projeto Físico
- Área de Transição

Como implementar eficientemente **extração, transformação e carga** de dados íntegros a partir de **uma ou mais** base de dados de **atualizadas** continuamente por sistemas corporativos *online* ?

Prática Comum



Fonte

Área de Transição - Staging Area

- Responsável pela correta extração de dados do ponto A para o ponto B na formatação e tempo apropriado;
- É onde o processo de **ETL** de dados ocorre;
- Exige definição da arquitetura necessária.

Área de Transição - Staging Area

- Tipo de Armazenamento
  - Flat Files;
  - Tabelas Relacionais;
  - Estruturas proprietárias usadas pelas ferramentas de estagiamento de dados.

- A escolha depende da qualidade e cronograma do projeto envolvido.

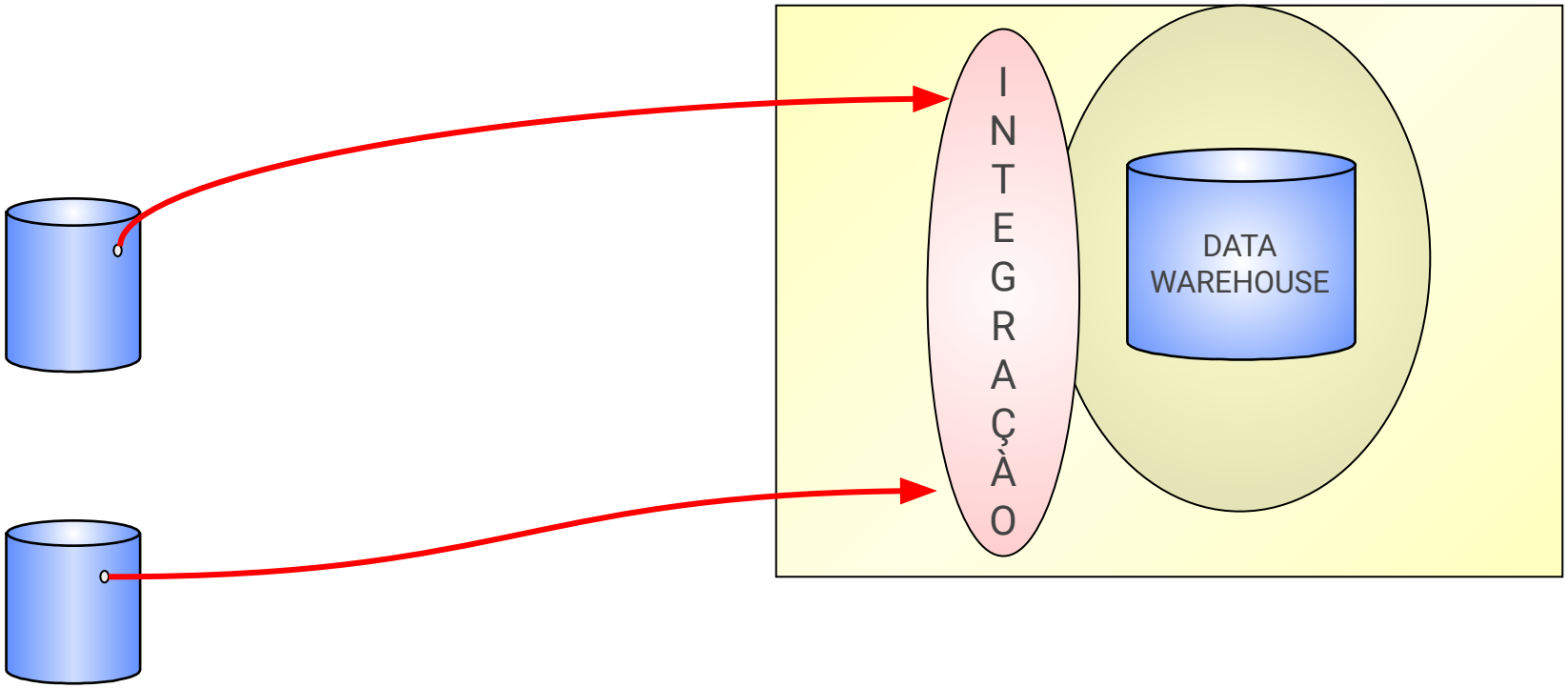
Visão do Projeto da Área de Transição

Planejamento
1 - Plano básico (visão geral do processo);
2 - Seleção de ferramenta de extração;
3 - Projeto detalhado;

Carga das Dimensões
4 - Construir a carga de uma dimensão estática;
5 - Construir a carga de uma dimensão de modificação lenta;
6 - Construir a carga das dimensões remanescentes;

Carga das Tabelas de Fatos e Automação
7 - Construir a carga das tabelas de fatos;
8 - Construir a carga incremental;
9 - Construir a carga de tabela agregação/MOLAP;
10 - Desenvolver a automação da transição.

Área de Transição



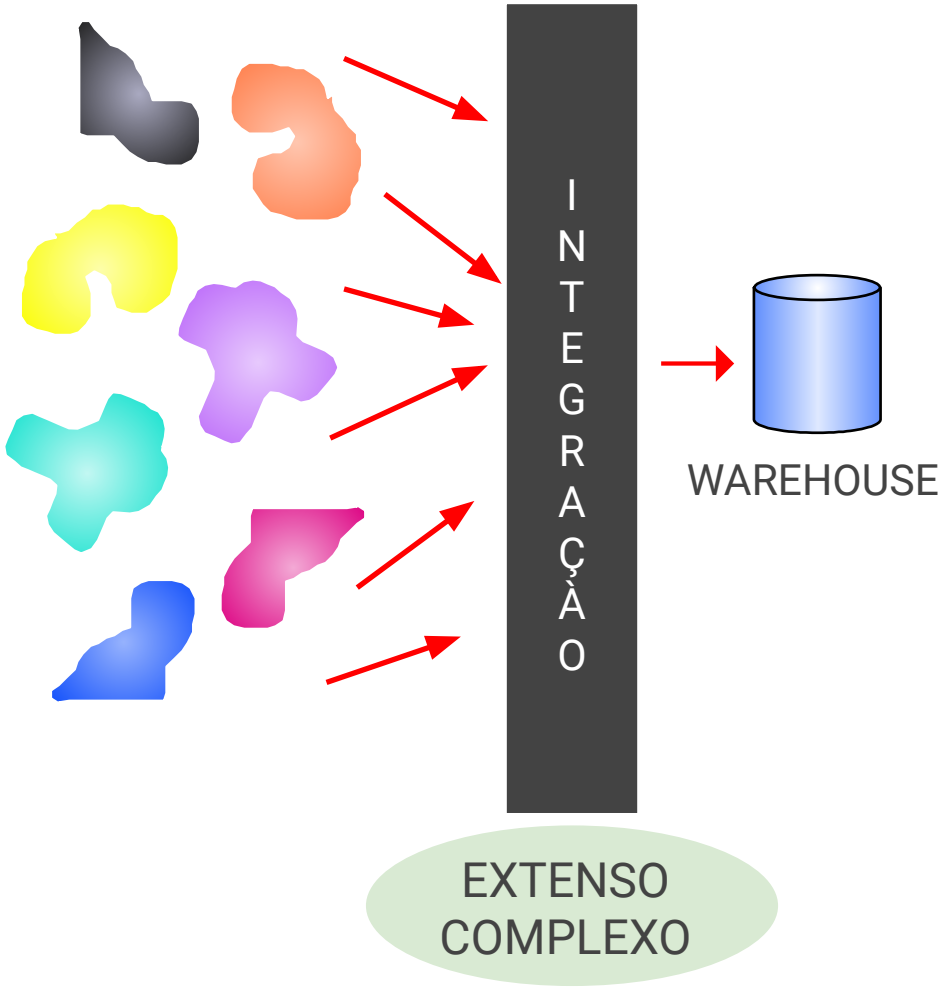
CAMADA ONDE OS DADOS NÃO INTEGRADOS DO AMBIENTE TRANSACIONAL SÃO COMBINADOS E TRANSFORMADOS EM DADOS CORPORATIVOS.

Instabilidade da Área de Transição

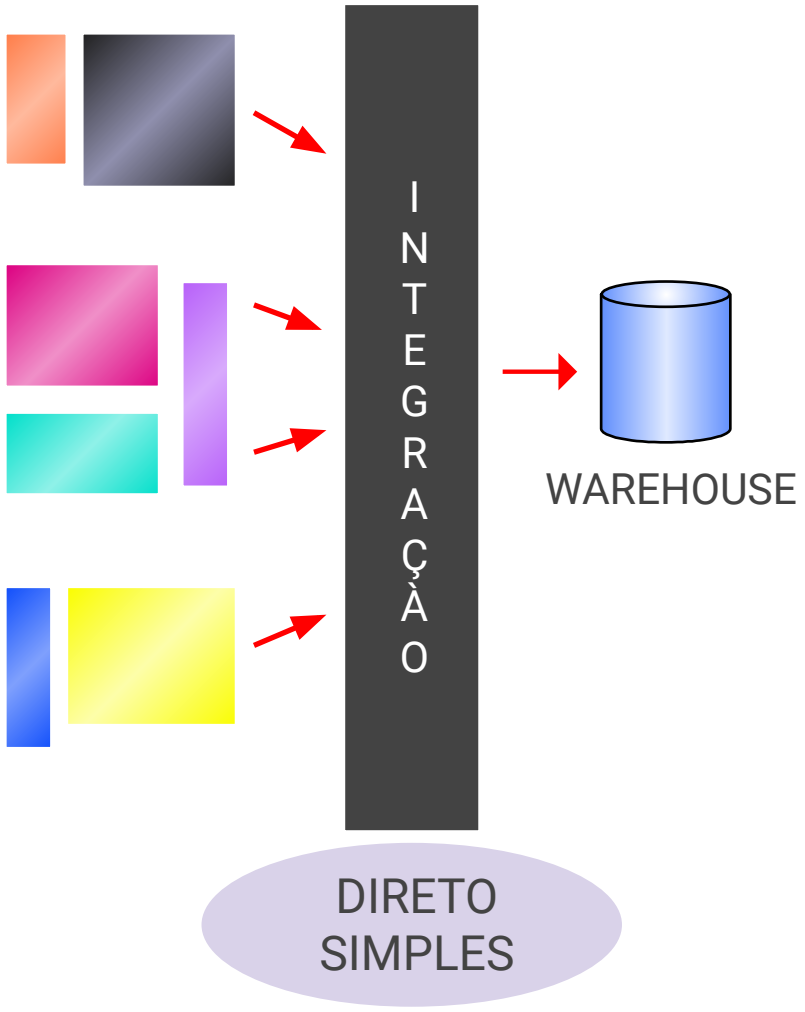
- o ambiente das aplicações é constantemente alterado
- o DWH normalmente é construído de forma incremental
- o DWH normalmente é construído de forma iterativa

Avaliação da Complexidade do Processo

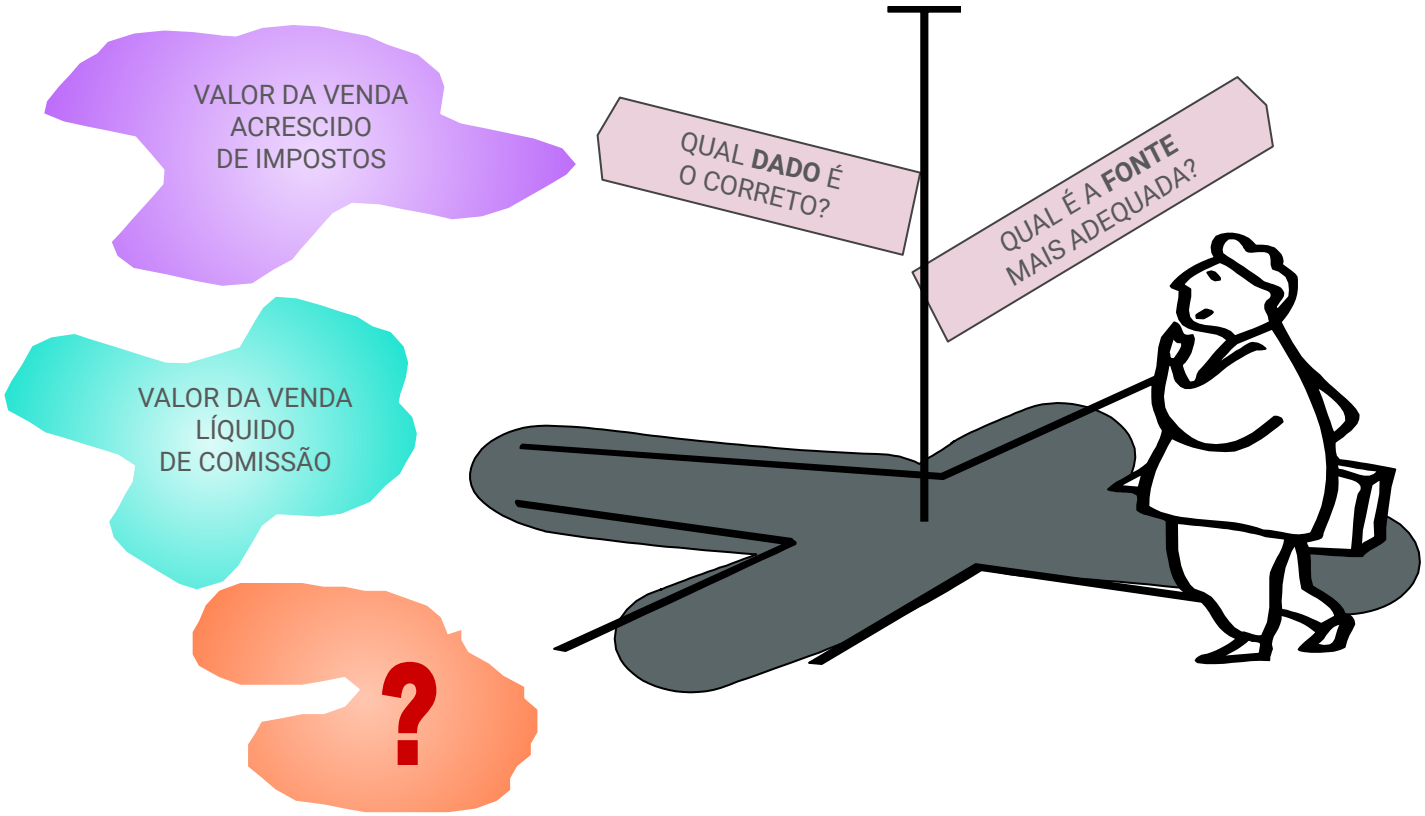
Cenário 1



Cenário 2



Identificação das Fontes de Dados

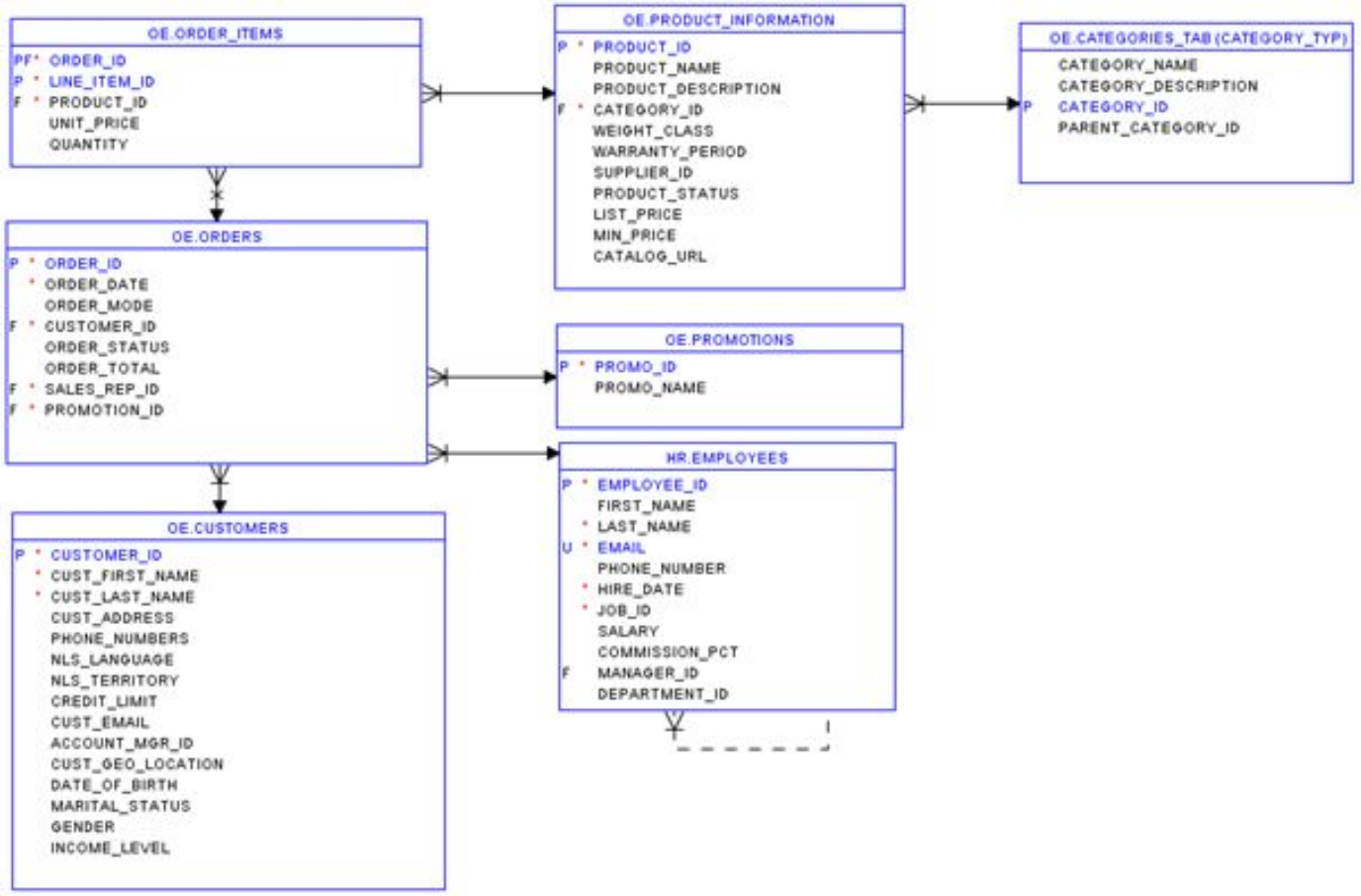


INTEGRAÇÃO

A PARTIR DO SIGNIFICADO DO DADO NO DATA WAREHOUSE DEVE SER IDENTIFICADO, ENTRE OS SISTEMAS OPERACIONAIS, O DADO A SER UTILIZADO COMO FONTE DE DADO

Um Exemplo de Extração com PDI

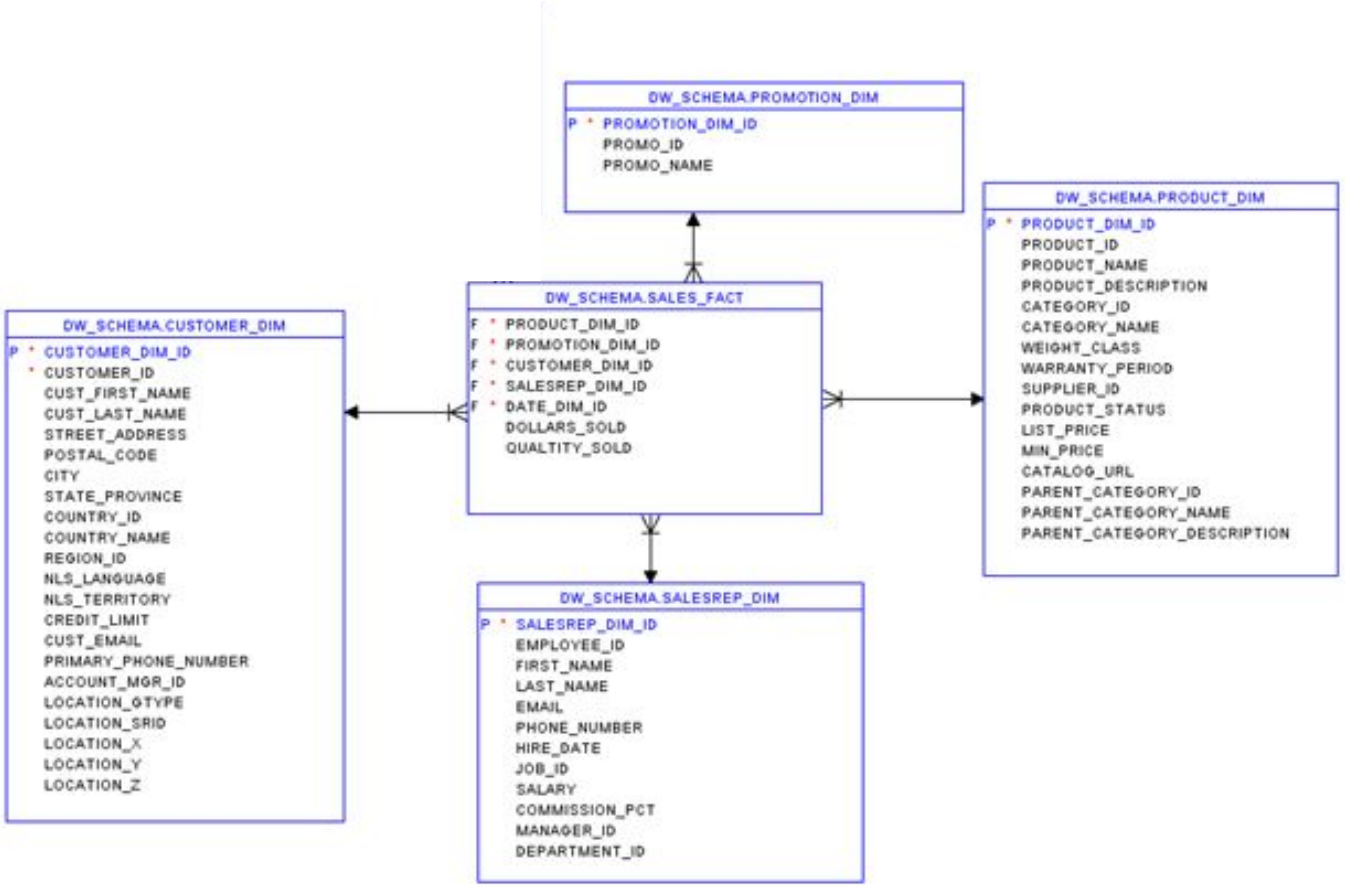
- Modelo transacional



Fonte: https://holowczak.com/building-etl-transformations-in-pentaho-data-integration-kettle/

Um Exemplo de Extração com PDI

- Vamos ver como extrair esses dados e criar as devidas dimensões com o Pentaho



Fonte: https://holowczak.com/building-etl-transformations-in-pentaho-data-integration-kettle/

Por hoje é só! 💪

Próxima aula:

- Transformação de Dados
- Prática no PDI