

# Mineração e Armazenamento de Dados

---

Carlos P. Caldeira

Departamento de Informática

Universidade de Évora

[ccaldeira@di.uevora.pt](mailto:ccaldeira@di.uevora.pt)

<http://www.di.uevora.pt/~ccaldeira>

---

# Modelo dimensional

# Modelo dimensional

---

- Modelo Entidade-Relação
- Limitações dos SGBDR
- Modelo dimensional: com fácil acesso e utilização

# Modelo relacional

---

## Objectivos:

- Redundância
- Integridade

## Consequência:

- Transações
- Divisão dos dados
- Teia intrincada
- Todas as tabelas são semelhantes

# Modelo relacional

---

## Problemas:

- Queries
  - simples
  - complexas
- Diagramas demasiado complexos
- Número de possibilidades num *join*

# Modelo relacional

---

## Análise complexa?

Calcular a percentagem do total das vendas para cada produto, por categoria e cidade, comparado com o mesmo período do ano anterior.

- pergunta complicada?
- como faz o utilizador?

# Análise dos dados

---

## OLTP:

Qual foi o lucro gerado pelo novo produto?

## DSS:

Qual foi o lucro gerado pelo novo produto por mês, na Beira Litoral, por nível demográfico, em cada ponto de venda, em relação à anterior versão do produto?

# Modelo multi-dimensional

---

- Estrutura adequada aos utilizadores
- “Olhar” para os dados usando uma vista multi-dimensional
- Modelo de dados
- Diagrama



# Modelo dimensional

---

- Dois grupos: **factos** e **dimensões**
  - dados a analisar
  - atributos sobre os factos

# Tabela de factos

---

- Tabela principal
- **facto** - é uma medida da actividade

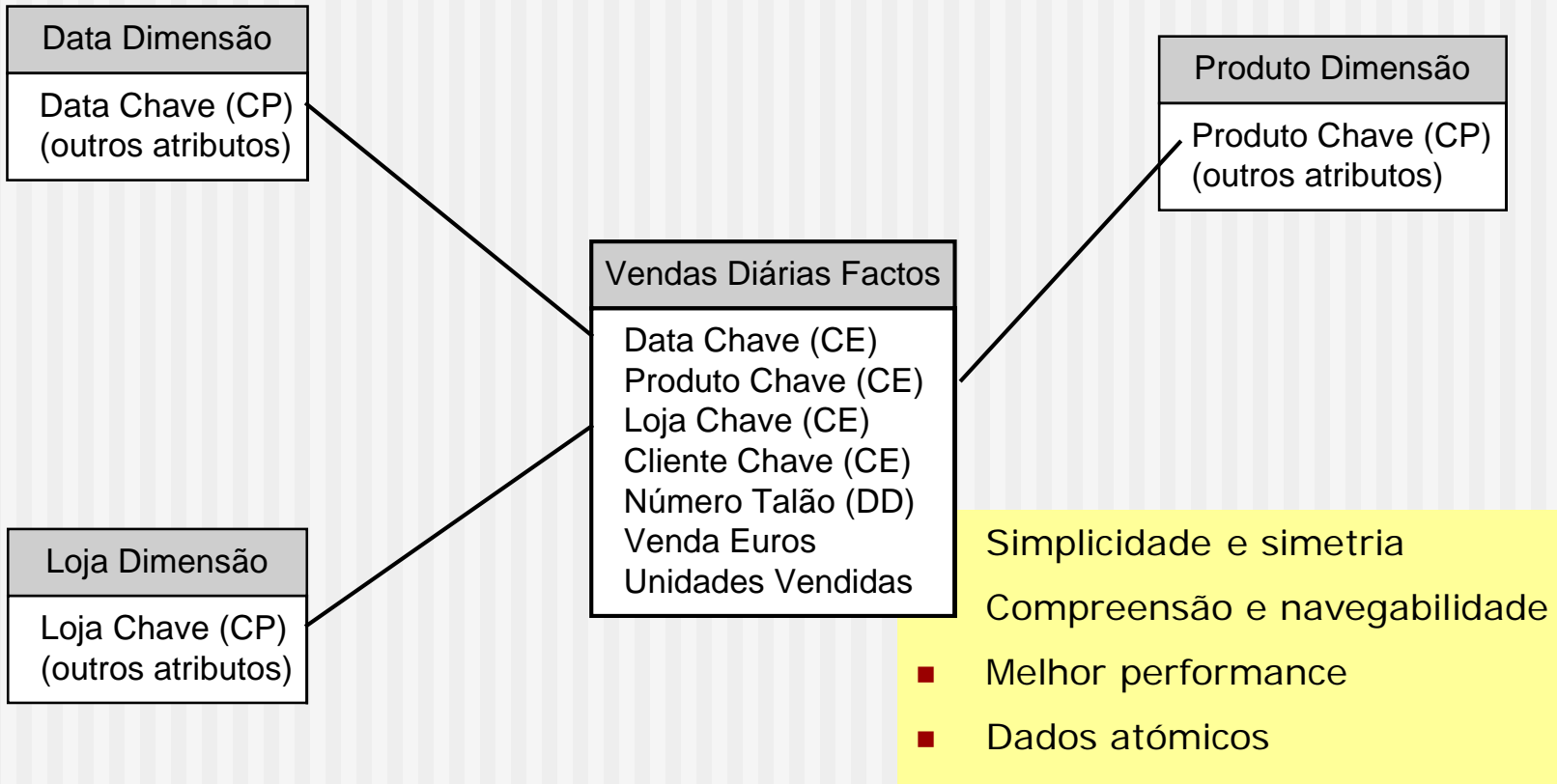
Vendas Diárias Factos
Data Chave (CE)
Produto Chave (CE)
Loja Chave (CE)
Cliente Chave (CE)
Número Talão (DD)
Venda Euros
Unidades Vendidas

# Tabelas de dimensões

---

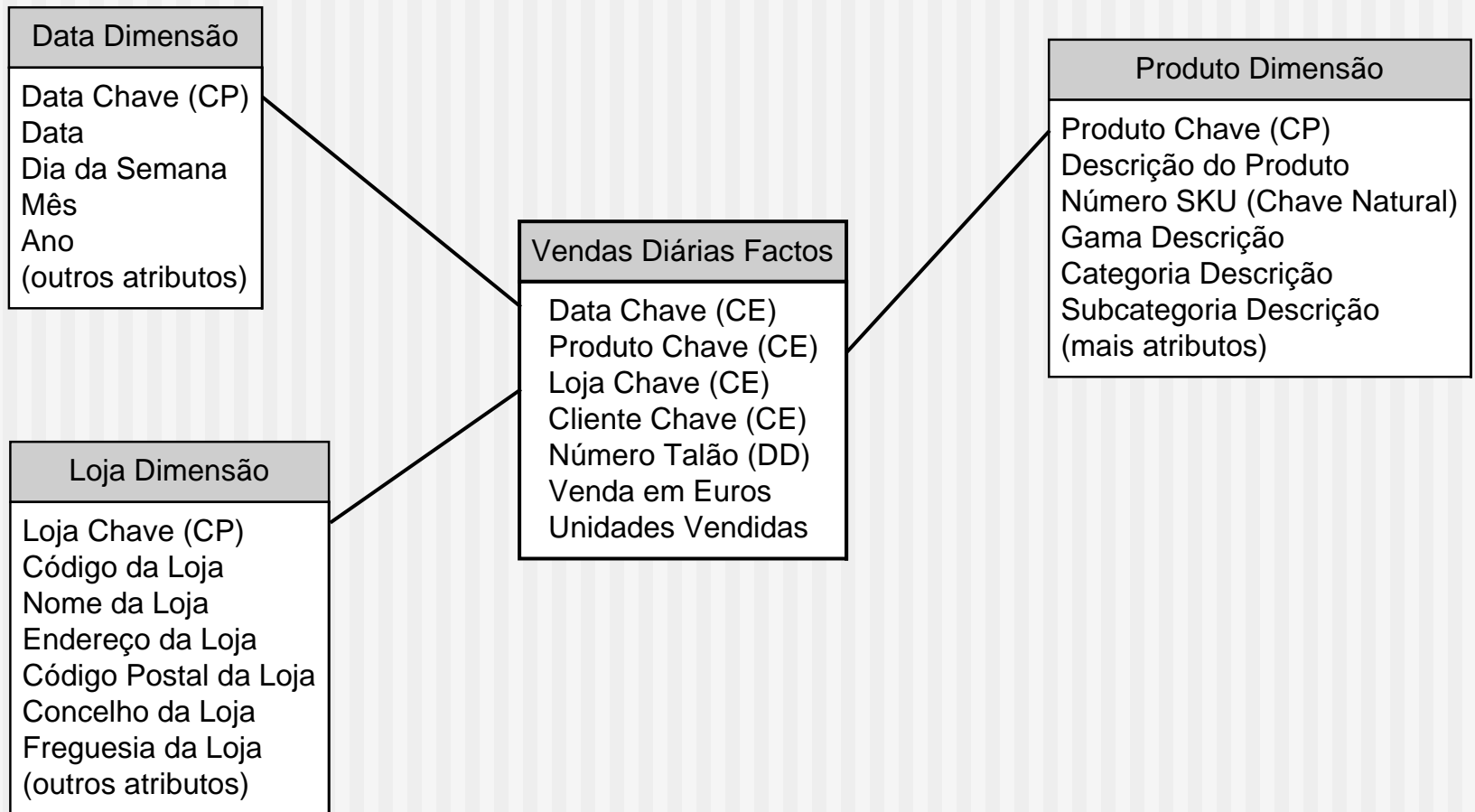
- Contêm a descrição textual do negócio
- Muitos atributos
- Cada dimensão é definida pela PK
- Principal fonte de informação
- Manter o DW utilizável
- Não-normalizadas

# Modelo dimensional

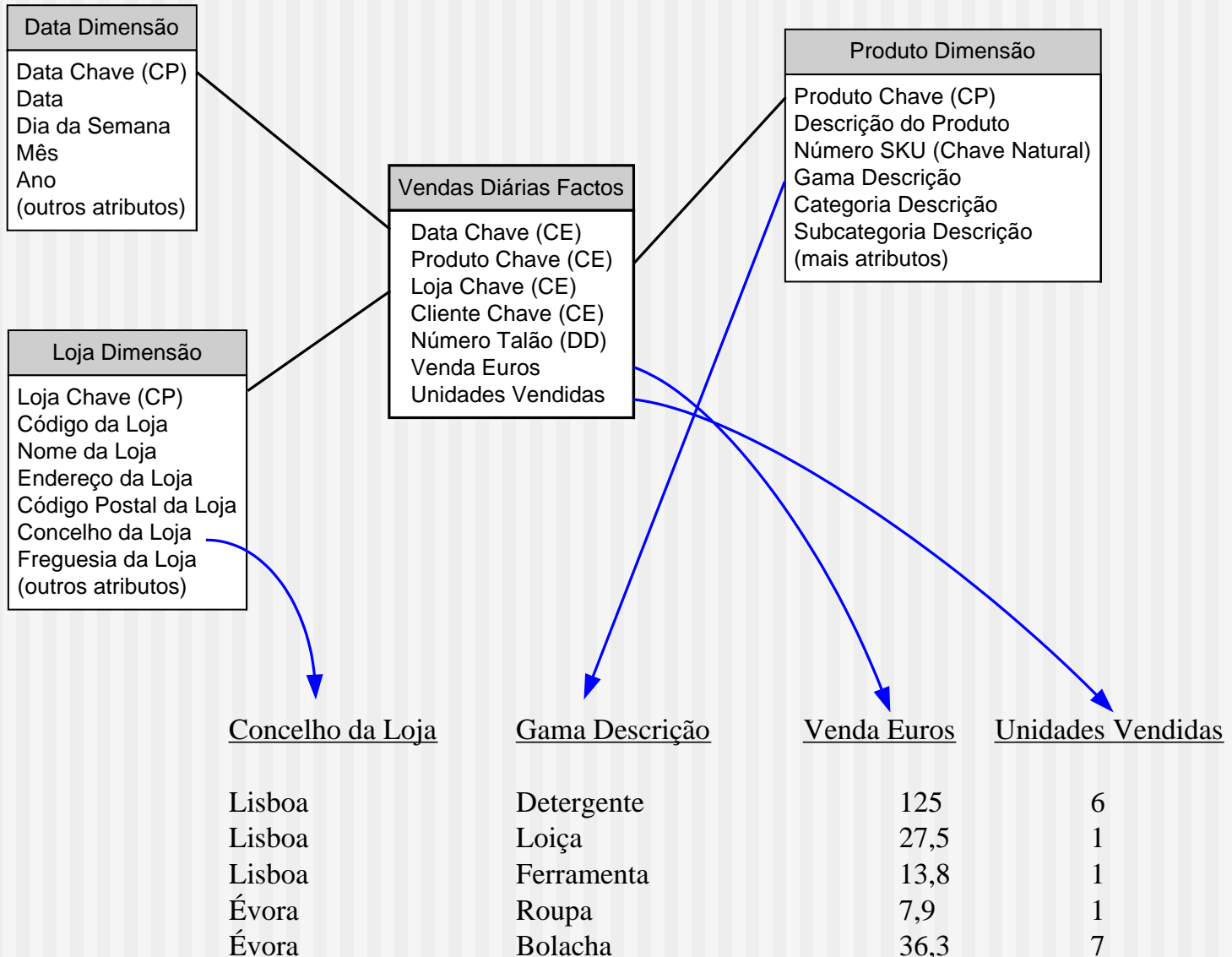


## Star Join Schema

# Complementaridade entre...



# Complementaridade entre...



# Asneiras...

---

## Utilização de nomes crípticos:

- T\_Colegio
- T\_Grau
- T\_Especialidade
- Num\_AI
- Doc\_Id
- Cd\_cr\_key
- etc. ....

# Asneiras...

## Tabela tradução de códigos

Máquinas : Tabela		
	Identificação	Designação da Máquina
	+	1 Chamua aivecas 3F14"
	+	2 Grade 18D24"
	+	3 Distrib adubo centrif 600l
	+	4 Fresa 220cm
	+	5 Fresa aradora 300cm
	+	7 Pulverizador 600l
	+	8 Reboque 7000kg
	+	9 Plantador batata 2L
	+	10 Arrancador batata 1L
	+	11 Vibrocultor ligeiro 36D10cm
	+	12 Enfardadeira
	+	13 Carregador frontal
	+	14 Rolo 300cm
	+	15 Plantador tomate 2L



---

Modelo dimensional

**Casos de Estudo**

# As quatro fases do desenho

---

## 1. Escolha do processo a modelar

- Um processo não é...
- Publicar os dados apenas uma vez

# As quatro fases do desenho

---

## 2. Declaração do grão do processo

- Nível de detalhe, o que significa...
- Exemplos:
  - Cada um dos medicamentos...
  - A classificação a cada disciplina...
  - O balanço mensal de uma conta...
  - O bilhete de embarque...

# As quatro fases do desenho

---

## 3. Escolha das dimensões

- Quais são as propriedades dos dados?
- Se a granularidade for bem escolhida...

# As quatro fases do desenho

---

## 4. Identificação dos factos numéricos

- O que se está a medir?
- Os factos têm que estar de acordo com o nível de detalhe definido no ponto 2.

# Empresa distribuição (retail case)

---

- Cada loja: mercearia, drogaria, lácteos, padaria, ...
- Cada loja tem  $\approx 60\ 000$  produtos:
  - código de inventário
  - 5 000 não têm UPC

1. Processo
2. Grão
3. Dimensões
4. Factos

# Empresa distribuição

---

## 1. Estudar as vendas nos POS

- Analisar o comportamento dos clientes
- Tipo de vendas por loja
- Promoções

# Empresa distribuição

---

## 2. Cada linha numa venda no POS

- Dados atômicos são *naturalmente dimensionais*
  - alto valor analítico
  - agregado de múltiplas formas
  - Qualquer estudo-resumo pode ter base nos dados atômicos mas...



# Empresa distribuição

---

## 3. Derivação das dimensões

- Loja, produto e data
  - **dimensões primárias**
- Promoção
  - **dimensão adicional**

# Empresa distribuição

---

## 4. Escolha dos factos

- Os factos têm que alinhar com o grão escolhido
- Que factos são recolhidos num POS?
  - a quantidade vendida
  - o valor da venda

# Data - dimensão

## Data Dimensão

Data Key (PK)

Data

Descrição

Dia da Semana

Dia Juliano

Semana Juliana

Mês Juliano

Número do Dia no Mês

Número do Dia no Ano

Número do Dia no Mês Fiscal

Número do Dia no Ano Fiscal

Nome do Mês

Número do Mês no Ano

Mês-Ano (MM-YYYY)

Trimestre

Semestre

Ano

Semana Fiscal

Mês Fiscal

Número do Mês Fiscal no Ano

Número da Semana Fiscal no Ano

Mês-Ano Fiscal (MM-YYYY)

Indicação de Feriado

Tipo de Dia da Semana

Estação de Vendas

Acontecimento Importante

(outros atributos)