PÓS-GRADUAÇÃO

# Modelagem e arquitetura do DW (data warehouse)

PÓS-GRADUAÇÃO

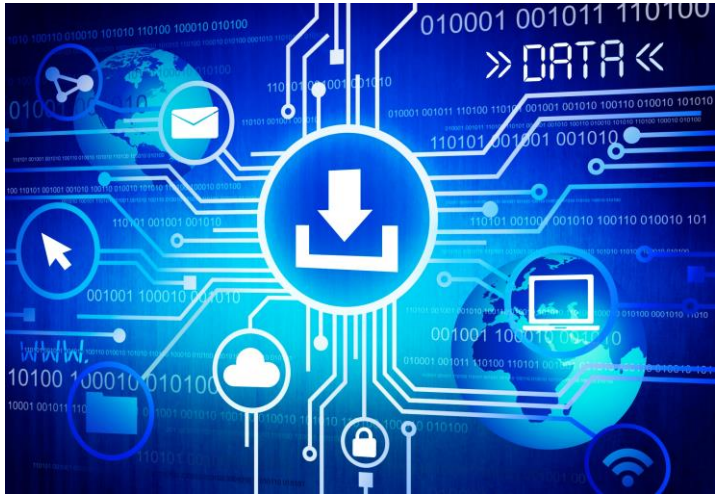
# Banco de Dados Transacionais x Banco de Analíticos

**Bloco 1**

Marise Miranda

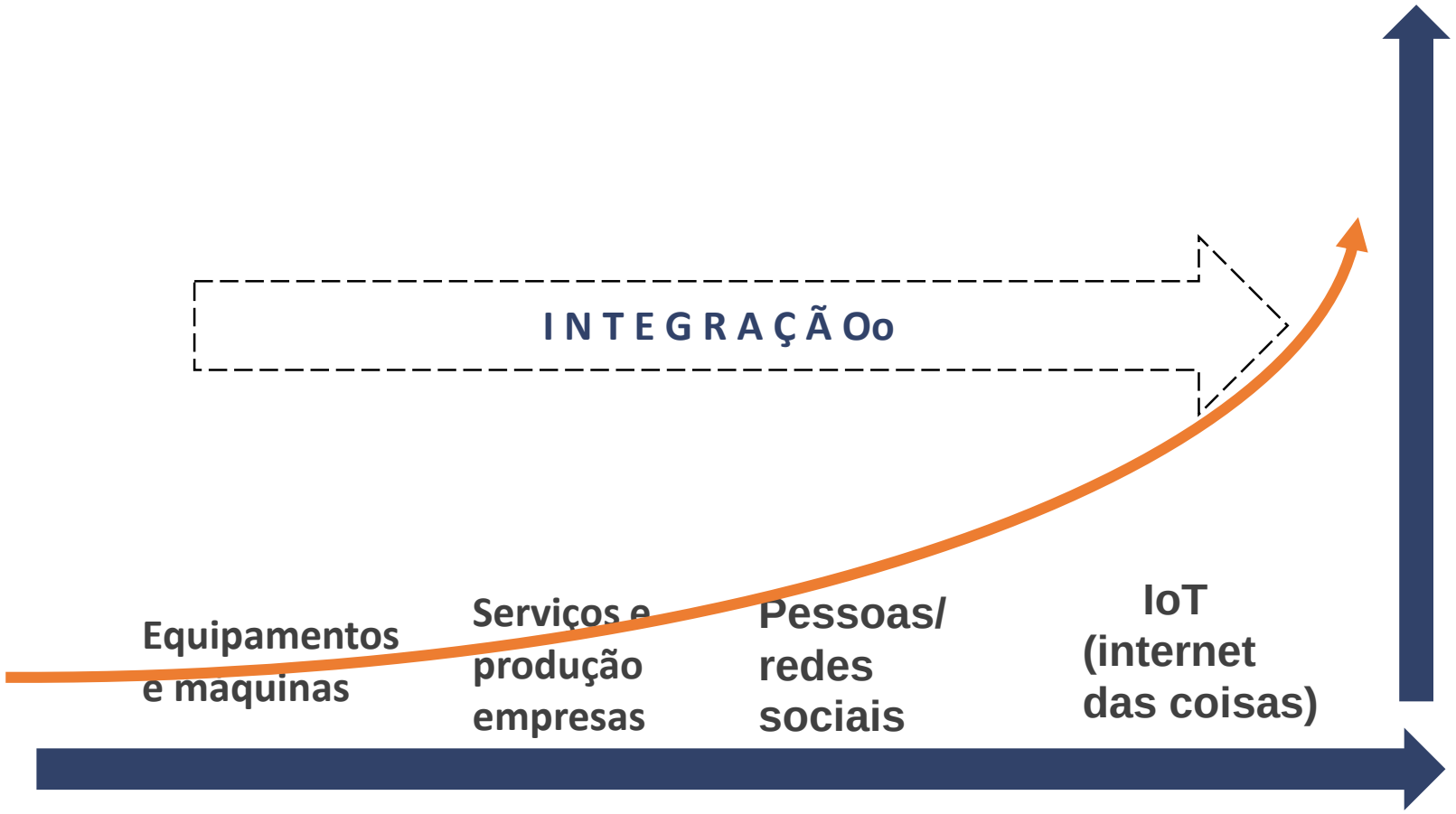
# Consolidação de diferentes fontes de dados

- Contexto histórico
  - Produção fabril.
  - Dados ad hoc.
  - Dados centrados na produção.
- Contexto social
  - Diversidade das fontes de dados.
  - Automação de serviços.
  - Redes sociais.
  - Conectividade e internet.
- Contexto computacional
  - Conectividade e internet.
  - Processamento e armazenamento.




Fonte: Rawpixel/iStock.com

# Dados – volume e variabilidade

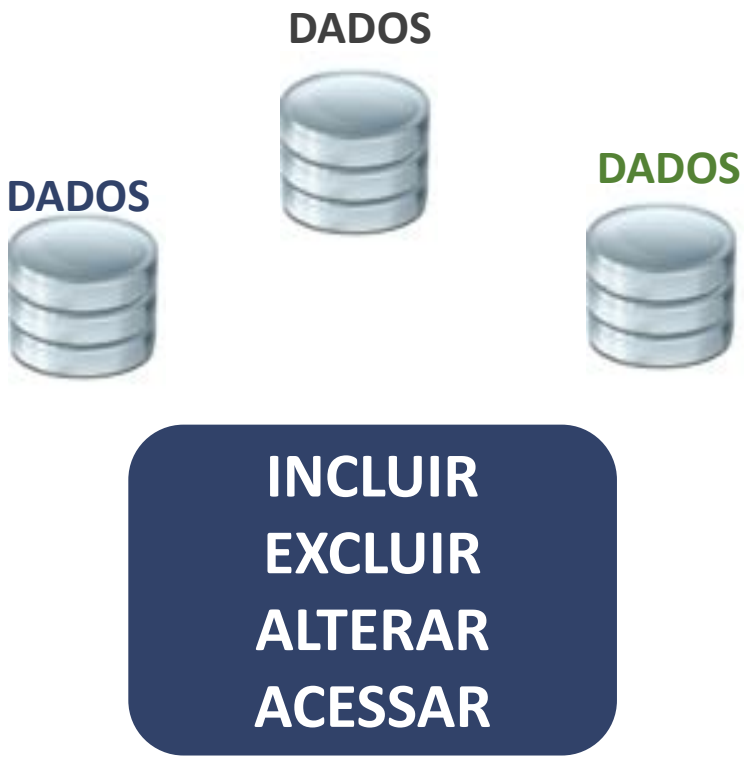



AUMENTO DO VOLUME

AUMENTO DA VARIABILIDADE

Fonte: elaborada pela autora

# Banco de dados transacionais




BD TRANSACIONAL

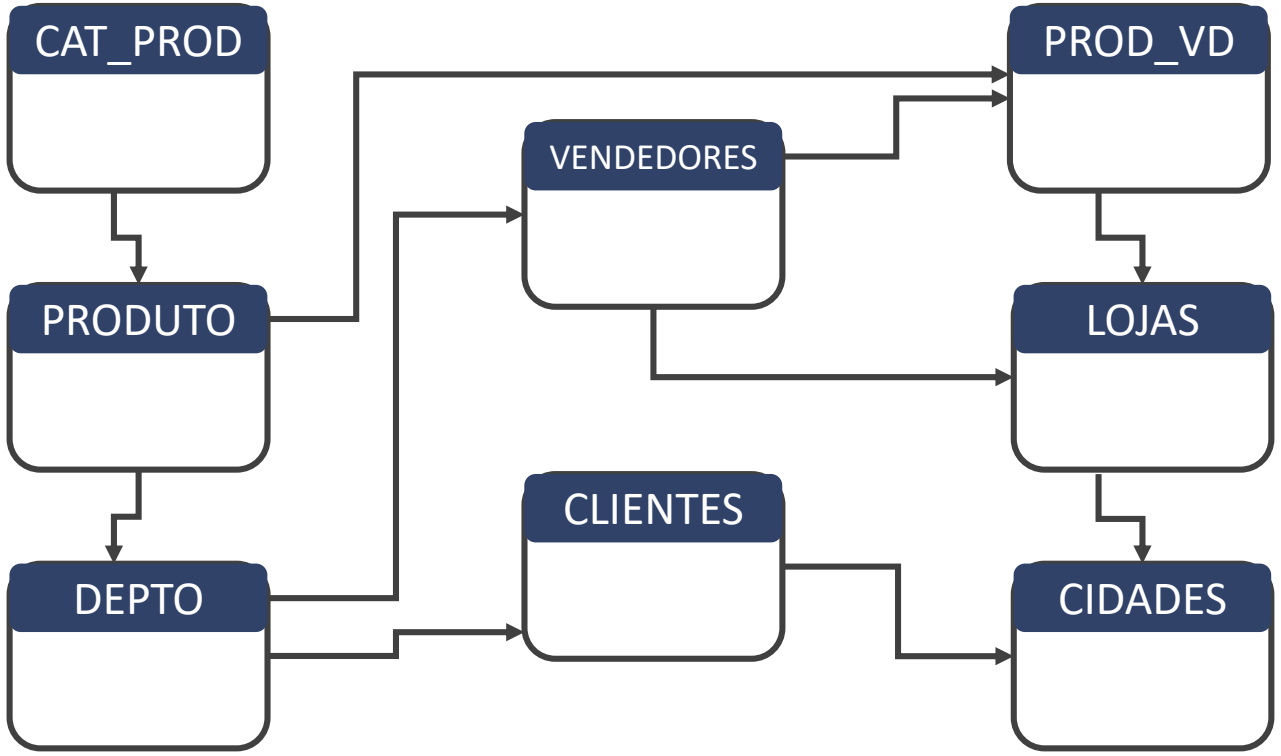



Diagrama ER – base de dados transacional

# Banco de dados Transacionais → Modelo DW

Dimensão 1

Dimensão 2

Dimensão 3

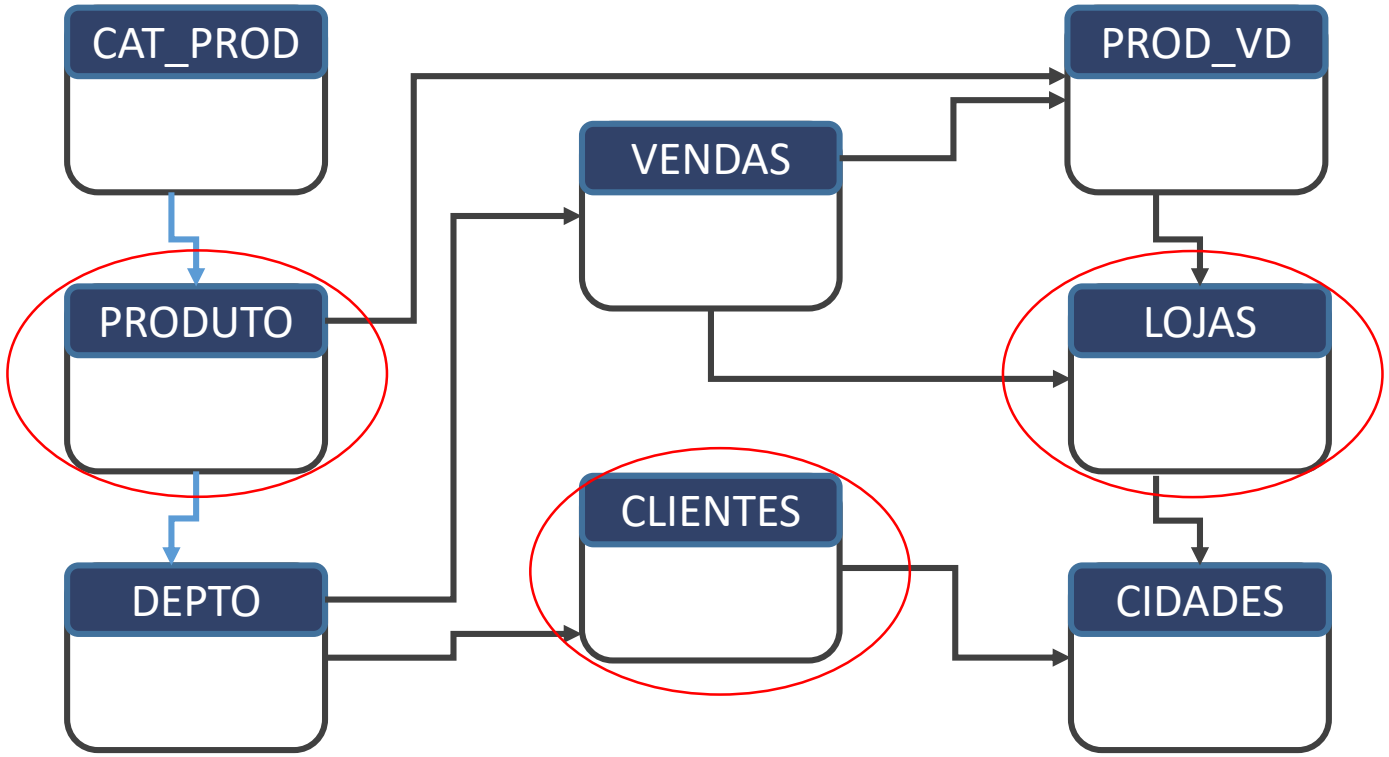


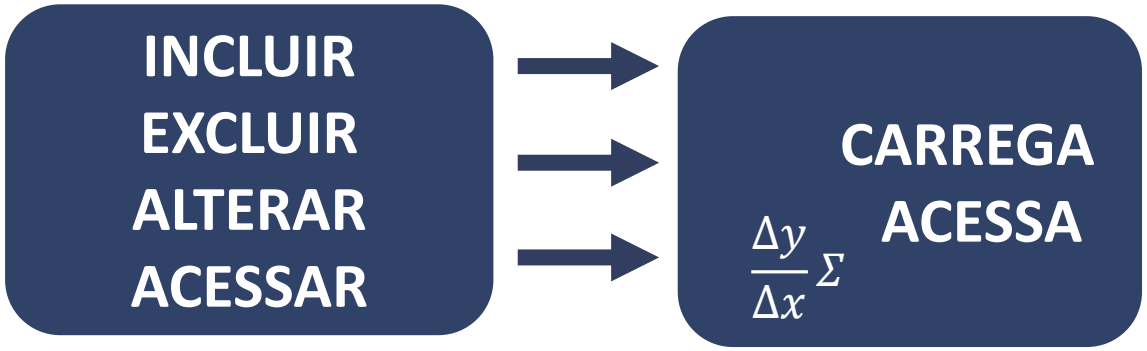

Diagrama ER – base de dados transacional → Diagrama ER do *data warehouse*

# Modelo DW

Dimensão 1

Dimensão 2

Dimensão 3




BD TRANSACIONAL

MODELO DW

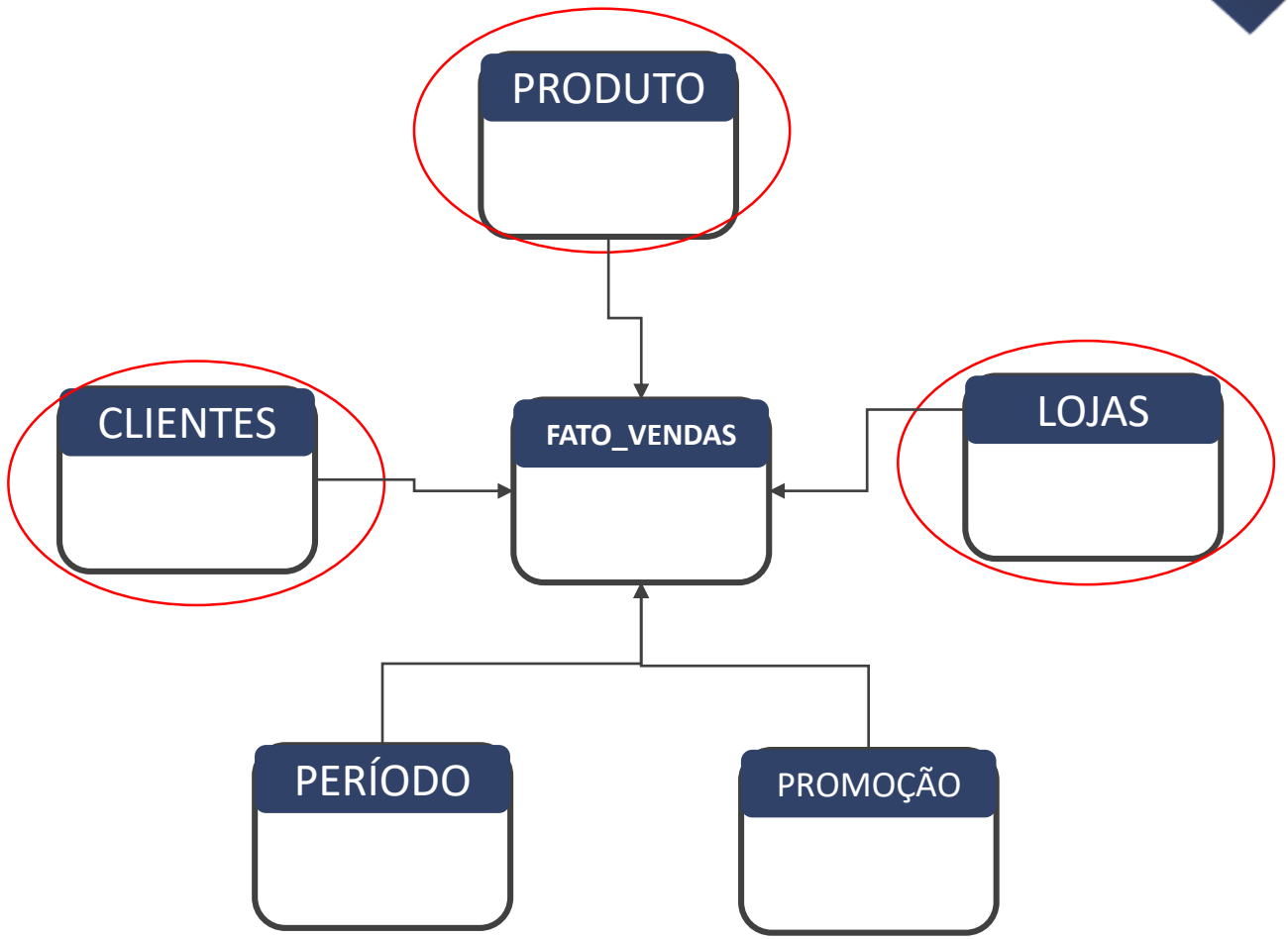



Diagrama ER do *data warehouse*

# Modelo DW → Banco de dados analíticos

**INCLUIR**
**EXCLUIR**
**ALTERAR**
**ACESSAR**

BD TRANSACIONAL

**OLTP**

BD ANALÍTICO

**CARREGA**
**ACESSA**

MODELO DW

$$\frac{\Delta y}{\Delta x} \quad \Sigma \quad x_{y^2} \quad \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

**OLAP**

Fonte: ojogabonitoo/iStock.com

Fonte: zmicierkavabata/iStock.com

# Conclusão

Um banco de dados transacional integra as bases de dados da organização que, em geral, advém dos processos.

Um banco de dados analítico integra um conjunto de dimensões construídas dentro da modelagem do DW. Em geral, são consumidas como apoio a tomada de decisão.

Um DW é o caminho intermediário entre os dados transacionais e os dados analíticos.

PÓS-GRADUAÇÃO

# Os datasets para o BD Analítico

**Bloco 2**

Marise Miranda

# *Datasets* não podem ser chamados de conjunto de dados

- **Datasets são a matéria prima das análises.**
- Datasets  são coleções de dados tabulares em formato tabela, linha e coluna, as linhas são os registros/campos e as colunas são características.
- **Dataset é uma coleção de registros de dados logicamente relacionados e armazenados.**
- Dataset é ad hoc, personalizado, autocontido, sem formatação.
- **Datasets podem ser enriquecidos por meio de cruzamento com outros dados.**
- Relatórios não são datsets.
- **Conjuntos de dados tem maior abrangência.**


Fonte: matejmo/iStock.com

# Consolidação de diferentes fontes de dados

dados.gov.br
PORTAL BRASILEIRO DE DADOS ABERTOS


75,0,190,80,91,193,371,174,121,-16,13,64,-2,?,63,0,52,44,0,0,32,
,32,0,0,0,56,0,0,0,0,0,0,80,0,0,0,0,0,0,0,0,0,0,0,0,40,52,0,0,28
-0.2,0.0,6.1,-1.0,0.0,0.0,0.0,0.6,2.1,13.6,30.8,0.0,0.0,1.7,-1.0,0.6
0,5.7,-1.0,0.0,0.0,-0.1,1.2,14.1,22.5,0.0,-2.5,0.8,0.0,0.0,0.0,0.0,1
-10.0,0.0,0.0,0.6,5.9,-3.9,52.7,-0.3,0.0,15.2,-8.4,0.0,0.0,0.0,0.9,5
56,1,165,64,81,174,401,149,39,25,37,-17,31,?,53,0,48,0,0,0,24,0,
,0,0,20,0,0,0,0,0,0,0,24,52,0,0,16,0,0,0,0,0,0,0,32,52,0,0,20,0,
,0.0,7.2,0.0,0.0,0.0,0.0,0.4,1.5,17.2,26.5,0.0,0.0,5.5,0.0,0.0,0.0,0.0,0
0.9,0.0,0.0,0.4,0.7,8.3,12.3,0.2,0.0,2.2,0.0,0.0,0.0,0.0,-0.2,0.8,6.
0.9,3.8,-5.7,27.7,-0.2,0.0,9.5,-5.0,0.0,0.0,0.0,0.5,2.6,11.8,34.6,-0
54,0,172,95,138,163,386,185,102,96,34,70,66,23,75,0,40,80,0,0,24
0,0,128,0,0,0,24,0,1,0,0,0,0,0,24,36,76,0,100,0,0,0,0,0,0,0,40,2
,0,0,0,0,0,1.0,0.0,4.5,-2.8,0.0,0.0,0.0,0.3,2.5,-2.2,19.8,0.8,-0.4,6
8.5,0.5,0.0,1.7,-2.7,0.0,0.0,-0.2,1.0,-9.4,-1.2,0.4,0.0,4.9,0.0,
0.0,5.8,-4.1,4.0,-0.5,0.4,0.3,20.4,23.3,0.7,0.0,10.0,-5.7,0.0,0.
55,0,175,94,100,202,380,179,143,28,11,-5,20,?,71,0,72,20,0,0,48,
0,60,44,0,0,32,0,0,0,0,0,0,56,0,0,0,0,0,0,0,0,0,0,0,0,40,44,0,0,
0,0.9,0.0,7.8,-0.7,0.0,0.0,1.1,1.9,27.3,45.1,0.1,0.0,9.1,-2.6,0.
9,0.0,3.2,-0.4,0.0,0.0,0.7,1.2,9.4,18.0,-0.1,0.0,5.1,-2.5,0.0,0.
0,-7.9,0.0,0.0,0.1,4.1,7.6,51.0,0.4,0.0,15.0,-5.5,0.0,0.0,0.1,3.


**Detalhes da Produção Intelectual Técnica de Programas de Pós-Graduação Strict...**

Os dados da seção contêm informações sobre o detalhamento da produção intelectual técnica dos programas dos programas de pós-graduação stricto sensu do ano de 2017.

CSV XLSX PDF HTML

**Detalhes da Produção Intelectual Bibliográfica de Programas de Pós-Graduação ...**

Os dados da seção contêm informações sobre o detalhamento da produção intelectual bibliográfica dos programas de pós-graduação stricto sensu no ano de 2017.

CSV XLSX PDF HTML

**Discentes dos Programas de Pós-Graduação stricto sensu no Brasil 2004 a 2012**

O arquivo apresenta informações sobre as atuações de pessoas declaradas como discentes pelos Programas de Pós-Graduação (PPG) stricto sensu no período de 2004 a 2012. Reúne as...

CSV XLSX PDF HTML

## Index of /ml/machine-learning-databases/arrhythmia

| Name | Last modified | Size | Description |
|---|---|---|---|
| Parent Directory | | - | |
| arrhythmia.data | 01-Apr-1998 12:14 | 393K | |
| arrhythmia.names | 19-Mar-1998 16:17 | 6.0K | |

*Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 443*

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

# Conclusão

Quando se organiza um *dataset,* este deve partir do princípio do que ser quer representar.  O que  cada  atributo significa.

Analogamente, um *dataset* deve preservar  as mesmas características quando exposto à amostragem.

A abordagem por amostra dos dados não deve implicar na modelagem das análises.

PÓS-GRADUAÇÃO

# Teoria em prática

**Bloco 3**

Marise Miranda

# Criação de um *dataset* – Estudo de Caso – Cafeteria

Criar *dataset* para análises, é uma tarefa complexa. Suponha que você tenha que criar um *dataset* em um Excel, relativo ao desempenho de uma cafeteria. São vendidos em média 200 cafés diariamente. A cafeteria abre de domingo a domingo, das 7 horas da manhã até às 20 horas.

Além do tradicional café, água, chá e leite também são vendidos. E ainda podem ser consumidos pão de queijo, pão com manteiga e brigadeiro.

A média de vendas de todos as bebidas, com exceção do café, é de R$ 77,00 diariamente.

A média diária de vendas dos produtos alimentícios fica em R$ 65,00 por dia.

# Criação de um *dataset* – Estudo de Caso – Cafeteria

Você poderá elaborar um *dataset* simulando as vendas dos itens descritos. Leve em consideração o desempenho de um mês dessa cafeteria.

Dica: os dados podem ser preenchidos na tabela como data, hora, produto, tipo de produto, quantidade, número do pedido, ID do pedido.

Em 10 anos de vendas de café, houve uma variação de 16%, variando de R$ 1,00

para R$ 4,40 reais.

720.000 cafés vendidos em 10 anos, aproximadamente.

Tabule os dados e crie conceito de perspectiva analítica, siga o modelo.

# Verificando as regras de negócio da cafeteria

Tabule os dados e crie conceito de perspectiva analítica. Siga o modelo.

| ID pedido | data | hora | produto | tipo de produto | qte | nº pedido | preço uni | preço tot |
|---|---|---|---|---|---|---|---|---|
| 1 | 14/01/2019 | 07:23 | bebida | café expresso | 2 | 321 | R$4,40 | R$8,80 |
| 2 | 14/01/2019 | 07:25 | bebida | média | 1 | 322 | R$5,40 | R$9,40 |
| | | | comida | pão de queijo | 1 | 322 | R$4,00 | |
| 3 | 14/01/2019 | 07:27 | comida | pão manteiga | 1 | 323 | R$2,50 | R$2,50 |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| Etc. | Etc. | Etc. | Etc. | Etc. | Etc. | Etc. | Etc. | Etc. |

Dica: 200 * 30 dias = 6000/ mês, 72.000/ano, em 10 anos = 720.000 cafés

PÓS-GRADUAÇÃO

# Dica do professor

**Bloco 4**

Marise Miranda

# Ferramentas SGBD de mercado e projeto Anvisa

**Pesquise na internet  quais são as principais ferramentas SGBDs, e compare suas vantagens e desafios. Fica como dica o Mysql e o SQL Server.**


# Ferramentas SGBD de mercado e projeto Anvisa

Acesse o portal da Anvisa, que apresenta o controle de medicamentos do governo brasileiro, e verifique como os dados dos produtos para saúde homologados estão disponibilizados para acesso público. Há informações sobre o projeto e quais tipos de dados estão disponibilizados, bem como o modo de como acessá-los.

ANVISA

# Referências

INMON, W. H. **Como Construir o Data Warehouse.** Rio de Janeiro: Campus, 1997.

KIMBALL, R. ***The Data Warehouse* Toolkit:** guia completo para modelagem dimensional. Rio de Janeiro: Campus, 2002.