Curso de Data Science

Charles Adriano dos Santos
Rafael Roberto Dias

ald
eia

Manhã

Horário	Assunto
09:30	O Banco de Dados (Conceito, Criação, Linguagem SQL)
11:00	Namorando os Dados (Queries SQL)
12:30	Almoço

Tarde

Horário	Assunto
13:30	Aprendendo Linguagem R no RStudio
15:30	Analisando Qualidade dos Dados
17:00	Variáveis Relevantes / Extração de Características

Pauta

1 – O Banco de Dados (Conceito, Criação, Linguagem SQL)

2 – Namorando Dados (Queries SQL)

Nos Episódios Anteriores...

Profissão Data Science

Estatística & Ciência da Computação

Desafio Agro XP

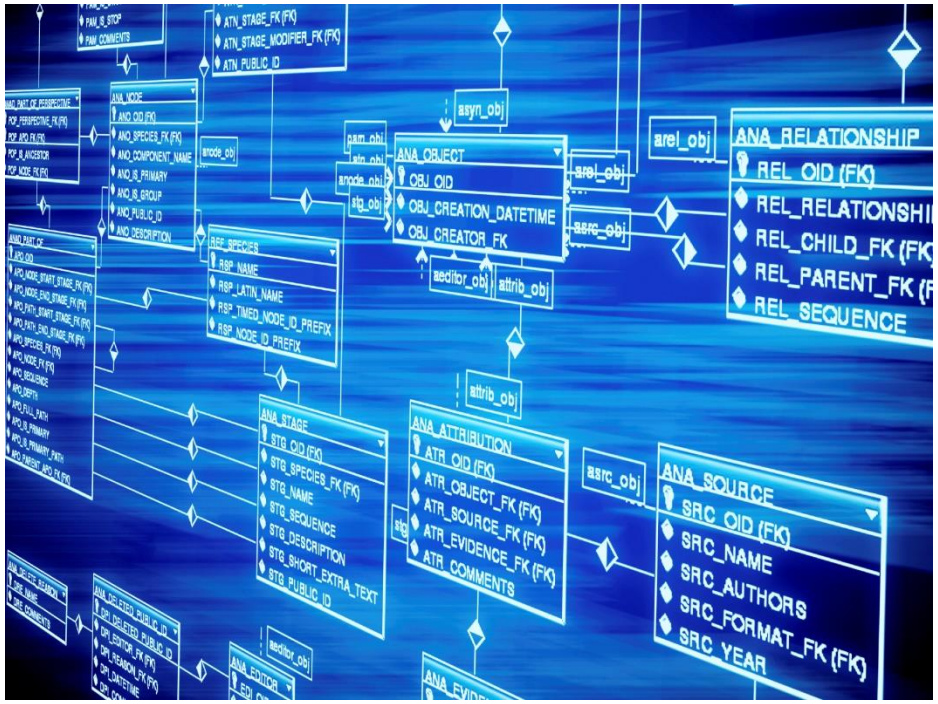
- Kanban
- Repositório
- Modelagem de Dados
- ETL

Pauta

1 – O Banco de Dados (Conceito, Criação, Linguagem SQL)

2 – Namorando Dados (Queries SQL)

Banco de Dados

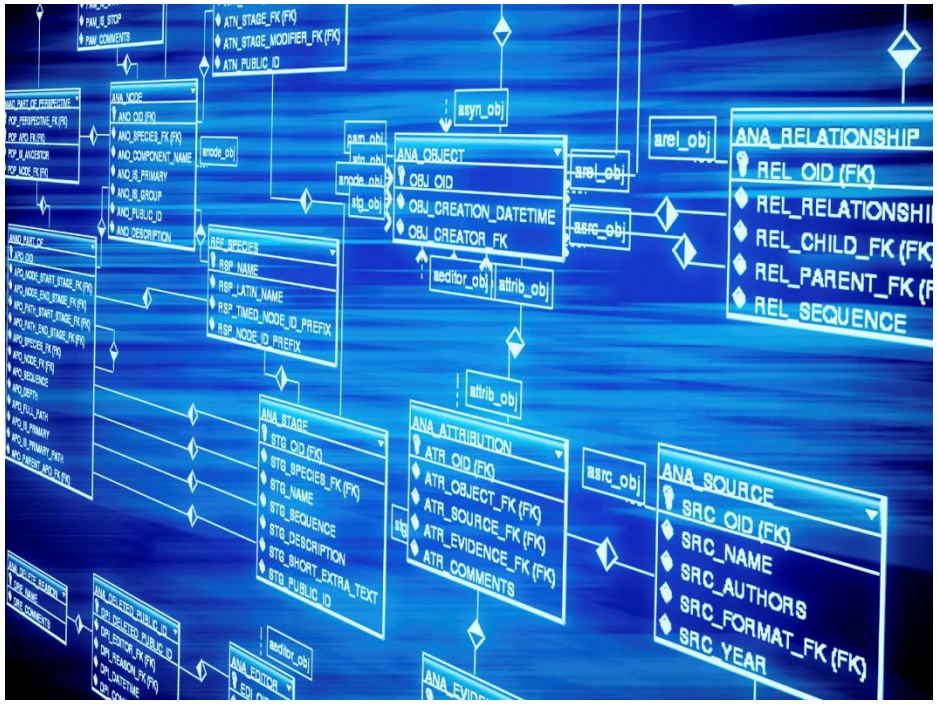


- **Bancos de dados** são conjuntos de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas

- São coleções organizadas de dados que se relacionam de forma a criar algum sentido (Informação) e dá mais eficiência durante uma pesquisa ou estudo. Garantia da integridade dos dados.

- São de vital importância para empresas e há duas décadas se tornaram a principal peça dos sistemas de informação

Banco de Dados



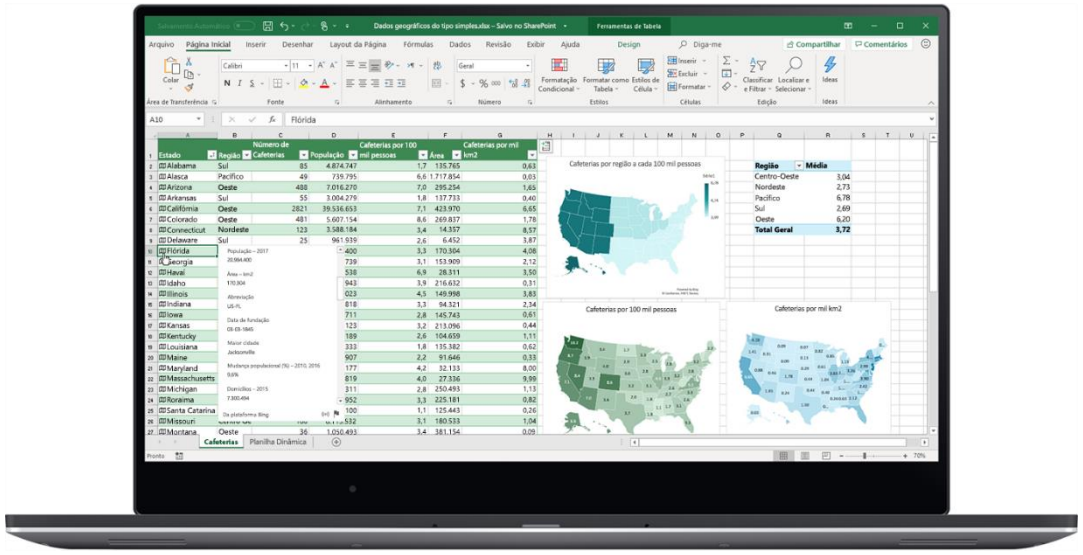
- São gerenciados por um SGDB (no nosso caso o PostgreSQL)

- Exemplos de Outros SGBDs: Relacionais → Oracle, SQL Server, MySQL, DB2, MonetDB. NoSQL → MongoDB e Cassandra e etc.

- **Banco de Dados Relacional**
  - Relações tabulares (Linha e Coluna)
  - Consistente / Íntegro
  - Relação cartesiana entre os dados
  - Custo Escalabilidade (Gerir os Dados)

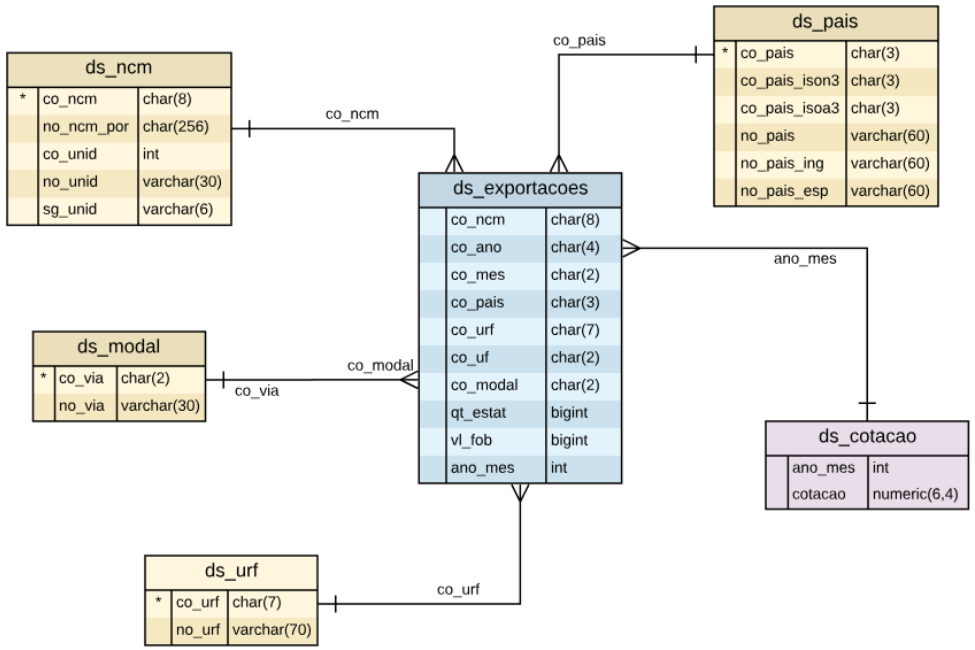
- **Banco de Dados Não Relacional (NoSQL)**
  - Orientado ao documento
  - Não garante Cosistencia/Integridade
  - Custo Menor Maior Escalabilidade (Gestão menos onerosa dos dados)

Banco de Dados - Relacionais



- **Relações Matriciais / Tabulares (Tabelas)**

Banco de Dados - Relacionais



- **Relações Matriciais / Tabulares (Tabelas)**
- Todos os dados de um banco de dados relacional são armazenados em **tabelas**
- Uma tabela é uma simples estrutura de linhas e colunas
- Linha → Registro / Coluna → Atributo
- As tabelas associam-se entre si por meio de regras de relacionamentos, que consistem em associar um ou vários atributos de uma tabela com um ou vários atributos de outra tabela

Banco de Dados

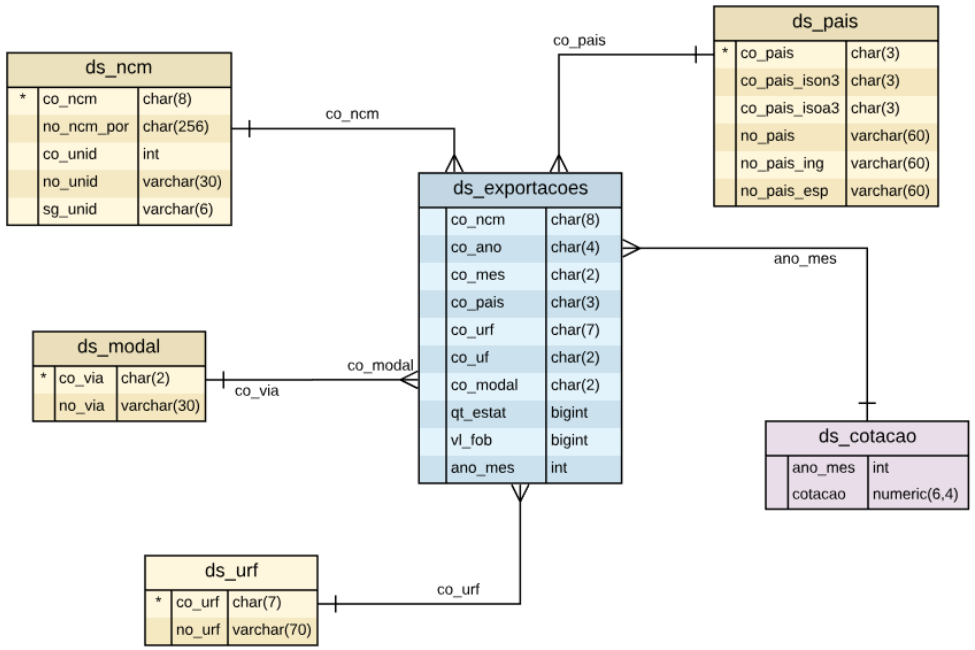
- **Registros (ou tuplas)**

- Tupla = Registro = Linha = Conjunto de Colunas

- Tabela = Entidade = Conjunto de Tuplas

	CO_ANO	CO_MES	CO_NCM	CO_UNID	CO_PAIS	SG_UF_NCM	CO_VIA	CO_URF	QT_ESTAT	KG_LIQUIDO	VL_FOB
1	1997	3	41043911	15	149	RS	1	1010500	3987	4150	16725
2	1997	5	63019000	10	97	MG	7	145200	0	1002	8420
3	1997	6	87168000	11	586	RS	7	145300	48	153	915

Banco de Dados



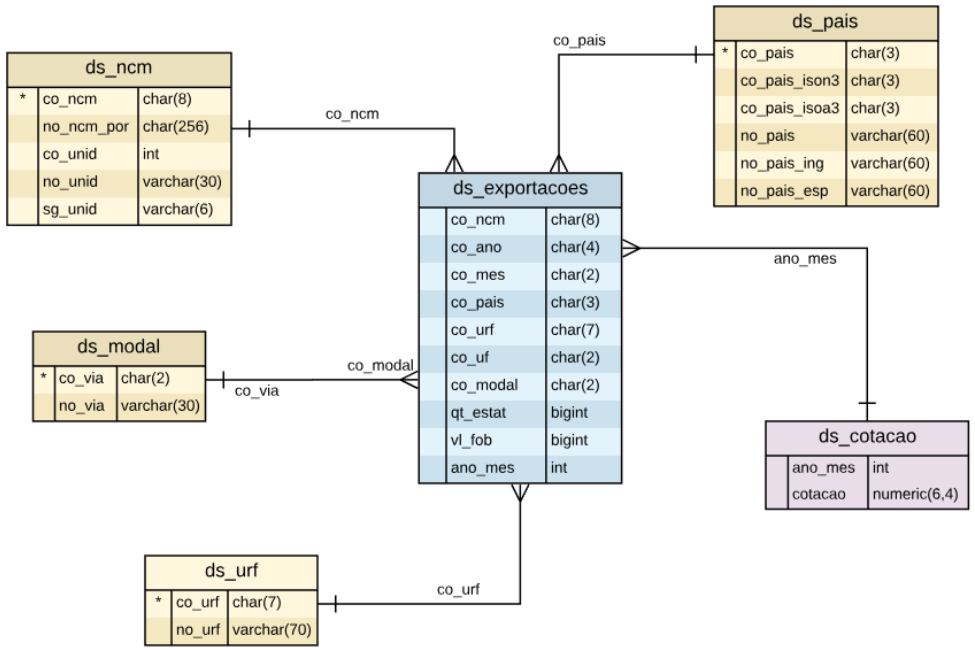
- **Chave**

- Integridade → Tupla/Registro/Linha única

- Chave primária: (PK - Primary Key)
  - ➢ A chave primária nunca se repetirá

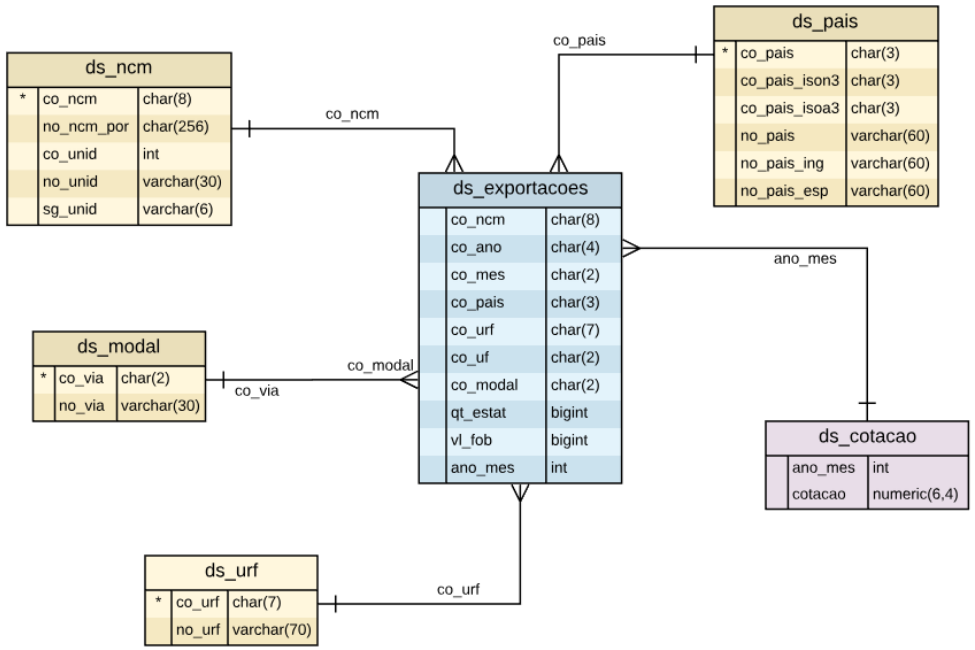
- Chave Estrangeira: (FK - Foreign Key) é a chave formada através de um relacionamento com a chave primária de outra tabela.
  - ➢ Define um relacionamento entre as tabelas e pode ocorrer repetidas vezes
  - ➢ Caso a chave primária seja composta na origem, a chave estrangeira também o será

Banco de Dados



- **Índices:**
- Coluna/Atributos utilizados para performance na recuperação da informação
- O SGDB define o plano de acesso e qual índice utilizar
- Possui um custo ótimo para recuperar o registro porém um custo alto no armazenamento do registro

Banco de Dados - SQL



- SQL – Structure Query Language

- Linguagem declarativa implementada pelos SGBDs para consulta aos dados armazenados no banco

- ANSI padroniza a linguagem porém cada SGBD implementa alguma modificação na versão. Ex:

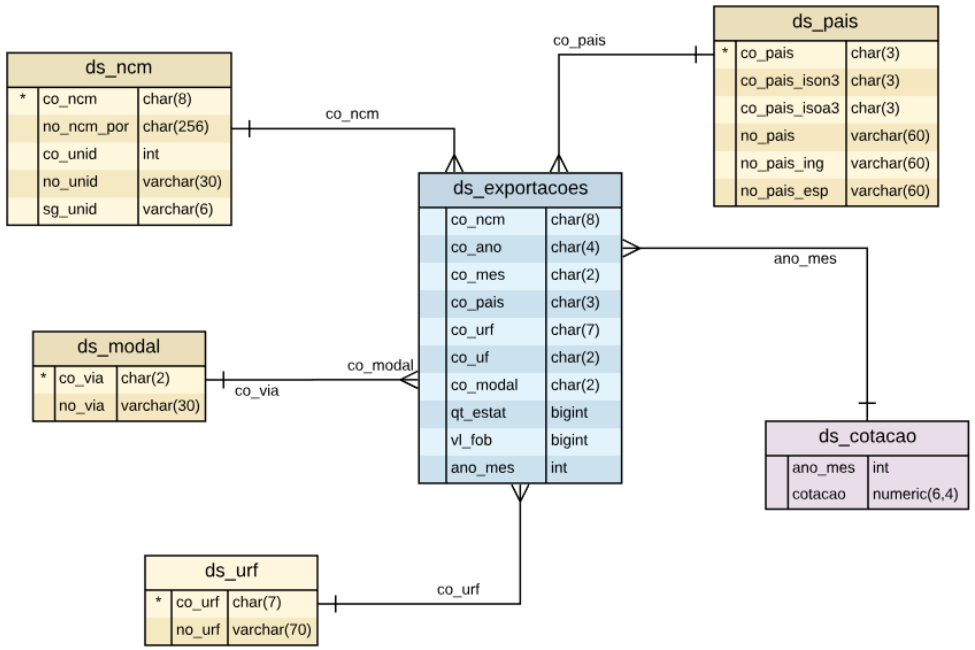
- Oracle →

  SELECT sysdate FROM dual; --Data e hora atual do SGBD

- PostgreSQL →

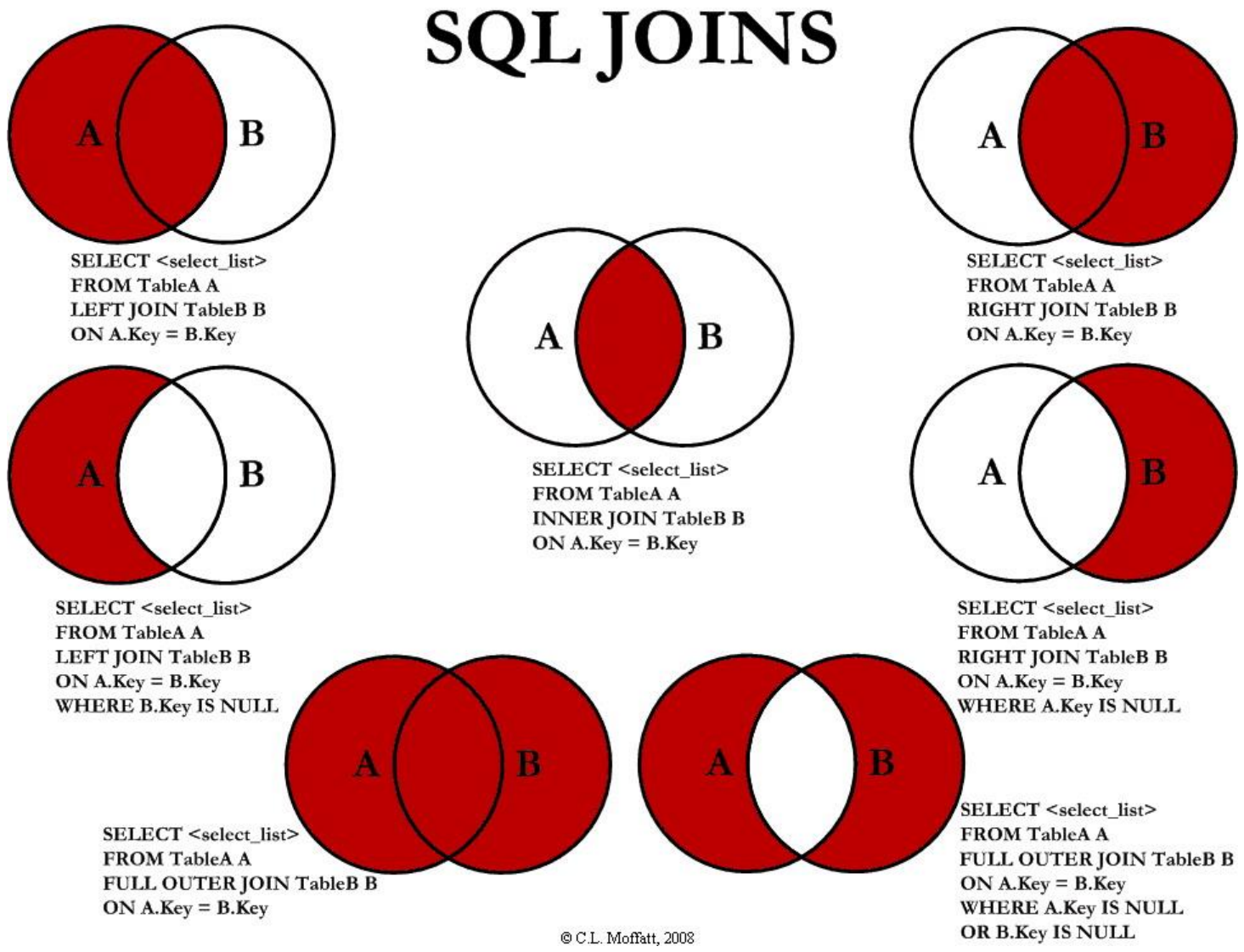
  SELECT CURRENT_TIME; --Somente hora
  SELECT CURRENT_DATE; --Somente a Data

Banco de Dados - SQL

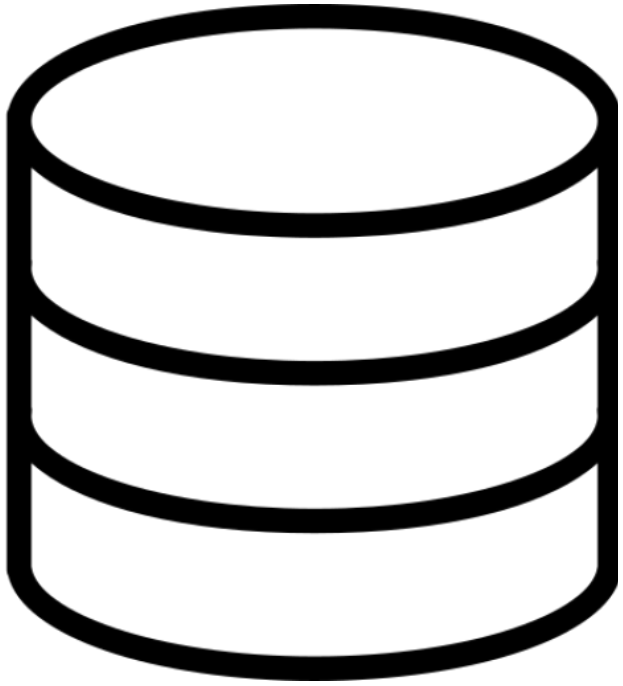


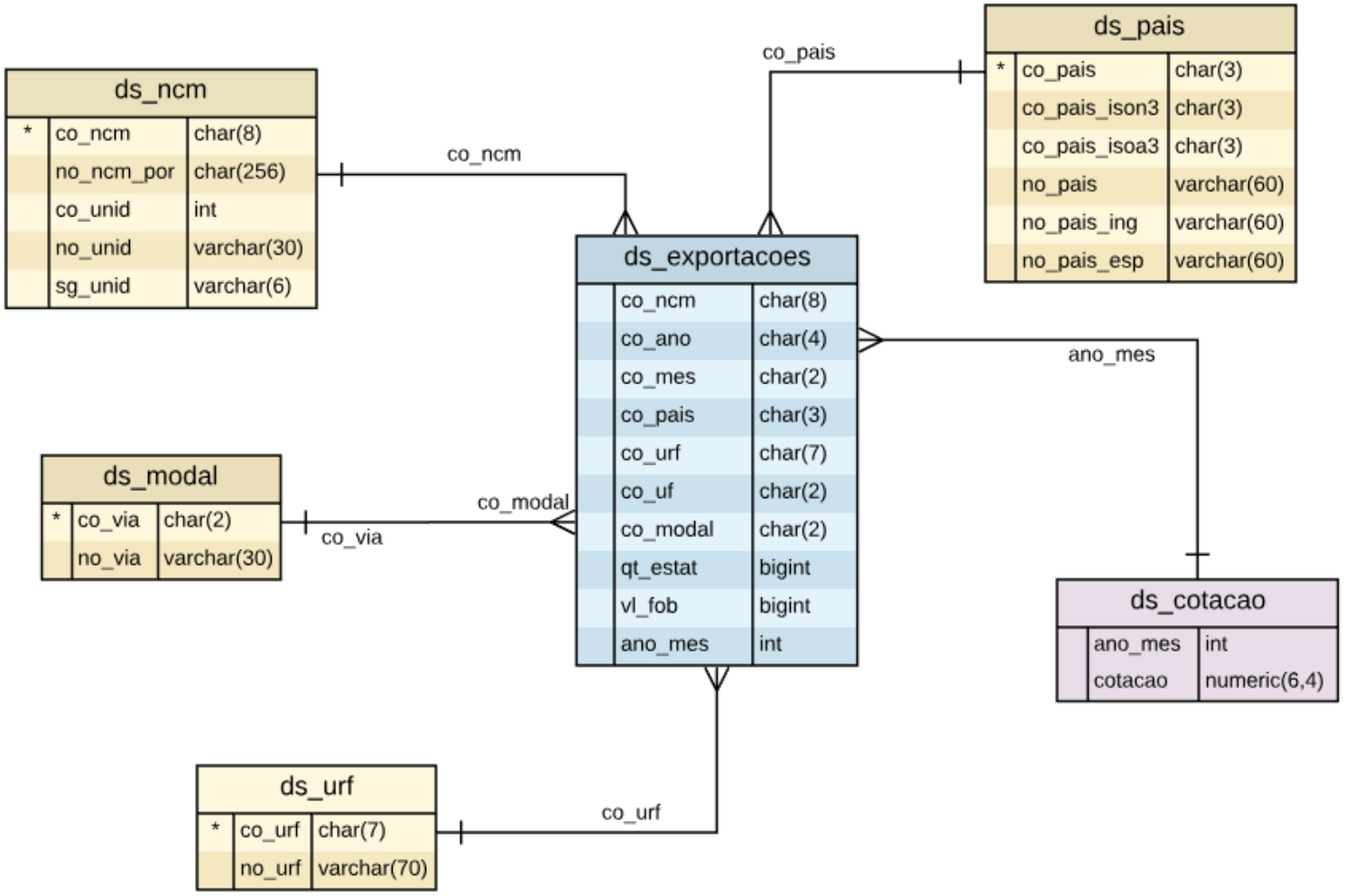
- SQL – Structure Query Language
- Subtipos da linguagem SQL (mais utilizados):
  - **DDL** → Definição de Dados / Altera estrutura da tabela/entidade (Ex: CREATE TABLE)
  - **DML** → Manipulação de Dados / Altera o conteúdo das colunas/atributos de tupla(s) (Ex: UPDATE)
  - **DTL** → Transação de Dados (Ex: Commit / Rollback)
- **DQL** → Consulta de Dados (SELECT)

Namorando os Dados (Queries SQL)



Desafio – Modelo de Dados





Pauta

1 – O Banco de Dados (Conceito, Criação, Linguagem SQL)

2 – Namorando Dados (Queries SQL)

Pauta

1 – Aprendendo Linguagem R no RStudio

2 – Analisando Qualidade dos Dados

3 – Variáveis Relevantes

Pauta

1 – Aprendendo Linguagem R no RStudio

2 – Analisando Qualidade dos Dados

3 – Variáveis Relevantes

Quais são os principais softwares Estatísticos?

- MiniTab - Software Matemático e Estatístico
- SAS - Statistical Analysis System
- SPSS - Statistical Package for the Social Sciences
- S-PLUS - Versão paga do R
- Python - Linguagem Interpretada
- R - (Ross e Robert)

Detalhes Software R

- **Linguagem Alto Nível -** Longe do código de máquina e mais próximo à linguagem humana
- **Interpretada -** O programa resultante não é executado diretamente pelo sistema operacional ou processador
- **Script -** Programas escritos para um sistema de tempo que automatiza a execução de tarefas
- **Orientada a objetos -** Abstração, Encapsulamento, Herança e Polimorfismo

Detalhes Software R

O R disponibiliza uma ampla variedade de

- Técnicas estatísticas
- Gráficos
- Modelos Lineares
- Modelos não Lineares
- Testes estatísticos clássicos
- Análises de Séries Temporais
- Classificação
- Agrupamento
- Machine Learning
- Artificial Inteligence

Detalhes Software R

- O R é utilizado através de um Interpretador de comandos
- Ao escrever 4 + 4 na linha de comando, obtém-se o seguinte resultado:

```
> 4 + 4
[1] 8
>
```

- A linguagem R suporta matrizes aritméticas, escalares, vetores, matrizes, quadros de dados (similares a tabelas numa base de dados relacional) e listas

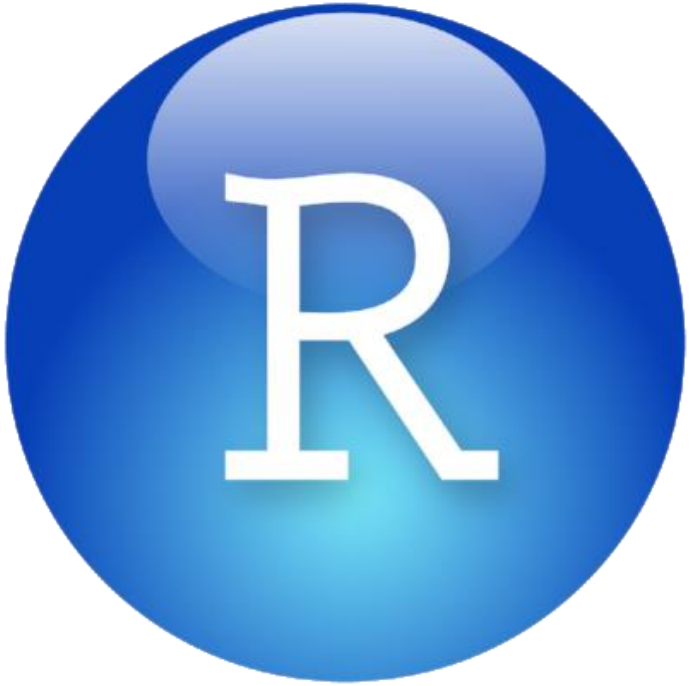
Detalhes Software RStudio

- RStudio é um software livre de ambiente de desenvolvimento, e que possui uma interface gráfica amigável

- O R Studio é uma interface para o R, com diversas utilidades diferentes que a tornam uma ferramenta mais simples em comparação ao R

- Ele possui duas versões: RStudio Desktop, que roda localmente em desktop e RStudio Server, que permite acessá-lo usando um navegador web enquanto ele roda em um servidor GNU/Linux remoto

Bora Praticar?



Pauta

1 – Aprendendo Linguagem R no RStudio

2 – Analisando Qualidade dos Dados

3 – Variáveis Relevantes

Analisando a Qualidade dos Dados

- Objetivo nesta etapa do estudo é verificar a qualidade dos dados para entender quais tem potencial de fazer parte do estudo

- Foco maior em verificar se existem dados faltantes ou nulos que podem interferir no estudo

- Também aqui começa o entendimento de como cada variável ajuda a explicar o evento em estudo

- Aqui começam as **descobertas** do Cientista de Dados

Pauta

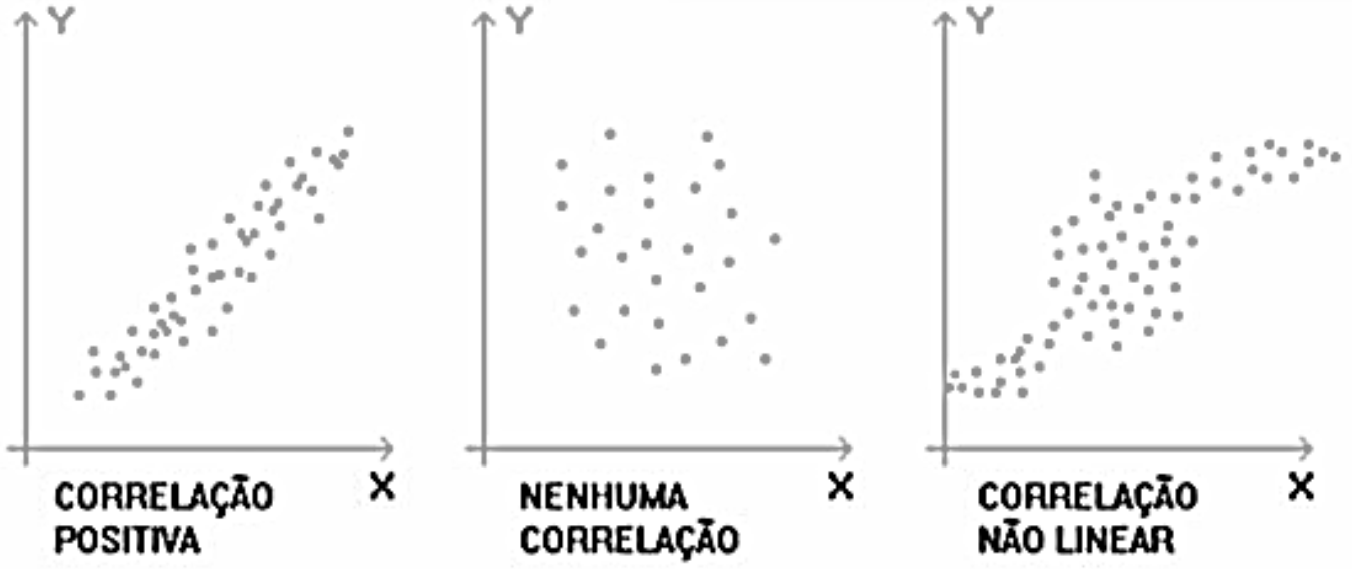
1 – Aprendendo Linguagem R no RStudio

2 – Analisando Qualidade dos Dados

3 – Variáveis Relevantes

Variáveis Relevantes

- Objetivo nesta etapa do estudo é verificar a como as variáveis se relacionam entre si
  - **Foco maior aqui é entender a correlação entre as variáveis**
- O modelo ou a metodologia que será utilizada para responder as perguntas do estudo dependem dos achados desta etapa



Obrigado !

Charles Adriano dos Santos
charles.a.santos@caelis.it
chadri
41 99144 6663

Rafael Roberto Dias
rafael.dias@madeiramadeira.com.br
rafael-roberto-dias-00b39123
41 99672 7170