

aldeia

Sejam bem-vindos!



Utilize a nossa redes de wi-fi:

#ALDEIA

utilizando a senha

b8xygf

A Aldeia é muito mais que espaço

Somos um movimento de desenvolvimento de realizadores.

Temos tudo que realizadores precisam para fazer uma ideia dar certo.

<http://aldeia.cc>

Cursos
Confrarias
Coworking
Offices
Networking
Eventos
Acelerações



Não passe perrengue

Tem água e café à vontade, e um doce e um salgado para você pegar na hora que quiser.

Temos banheiros nos dois andares da **Cândido**:

- Primeiro andar: atrás da recepção
- Segundo andar: ao lado da escada

E atrás da recepção na unidade **Estação**.

Se algo não estiver certo, fale com a nossa equipe

Faça parte da nossa Tribo

Receba os **materiais do curso** e seu **certificado** de participação por meio da nossa comunidade virtual.

Acesse <https://aldeia.cc/chamado> e faça sua solicitação para fazer parte da plataforma, utilizando o e-mail da compra do curso para se identificar.



Tire uma foto deste QR code e vá direto para a página da Tribo

Data Science

Charles Adriano dos Santos
Rafael Roberto Dias



Pauta

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – Namorando os Dados

7 – Welcome to R

8 – Homework



Agenda

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – Namorando os Dados

7 – Welcome to R

8 – Homework



Manhã

Horário Assunto

09:30 Framework para Modelagem

10:30 Extract, Transform and Load (ETL)

12:00 Estratégia para os dados (Modelo de Dados)

12:30 Almoço



Tarde

Horário Assunto

- 13:30 O Banco de Dados (Conceito, Criação, Linguagem SQL)
- 15:00 Namorando os Dados (Queries SQL)
- 16:30 Welcome to R – Parte I
- 18:00 Homework (ETL, Namorando Dados SQL, R)



Framework para Modelagem

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

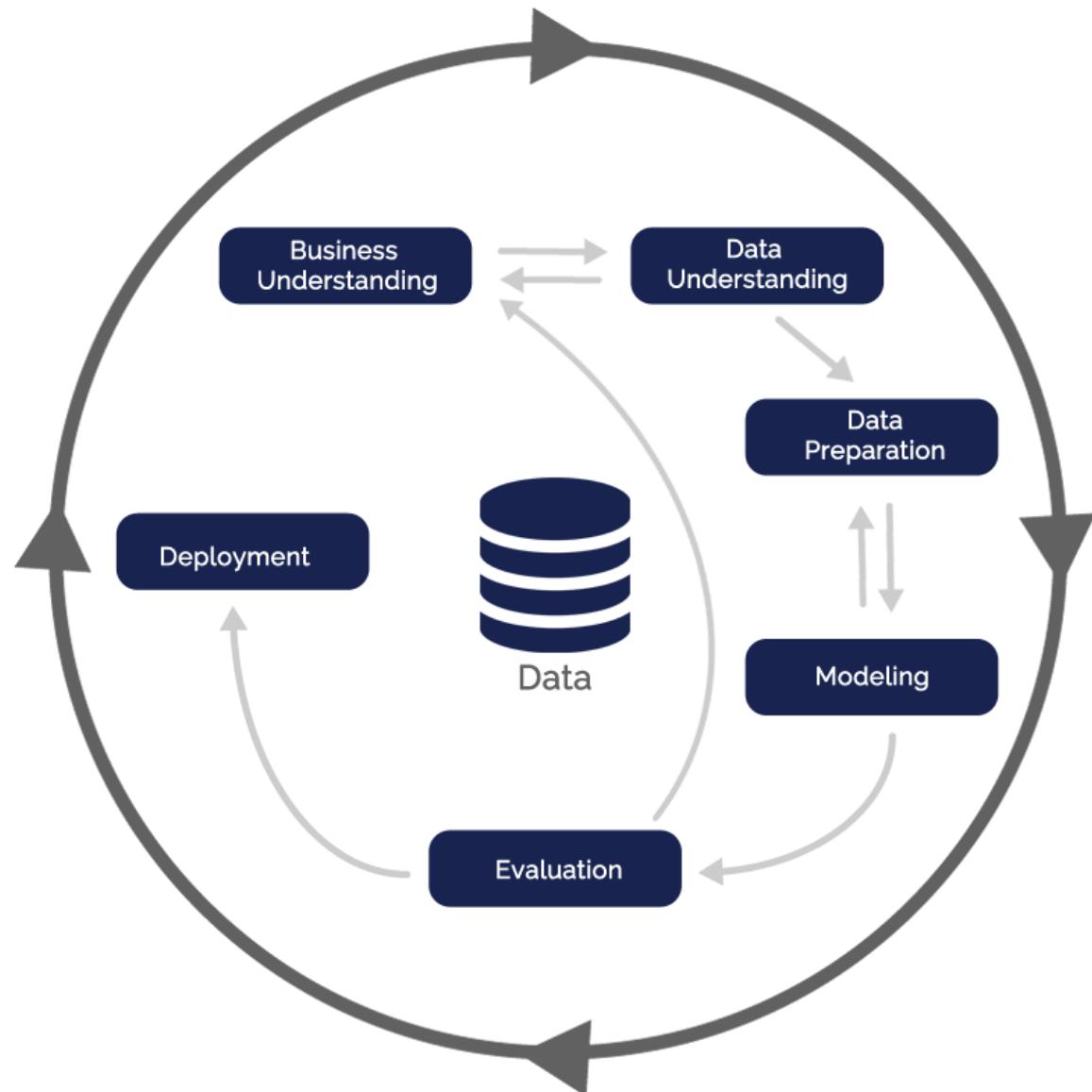
6 – Namorando os Dados

7 – Welcome to R

8 – Homework



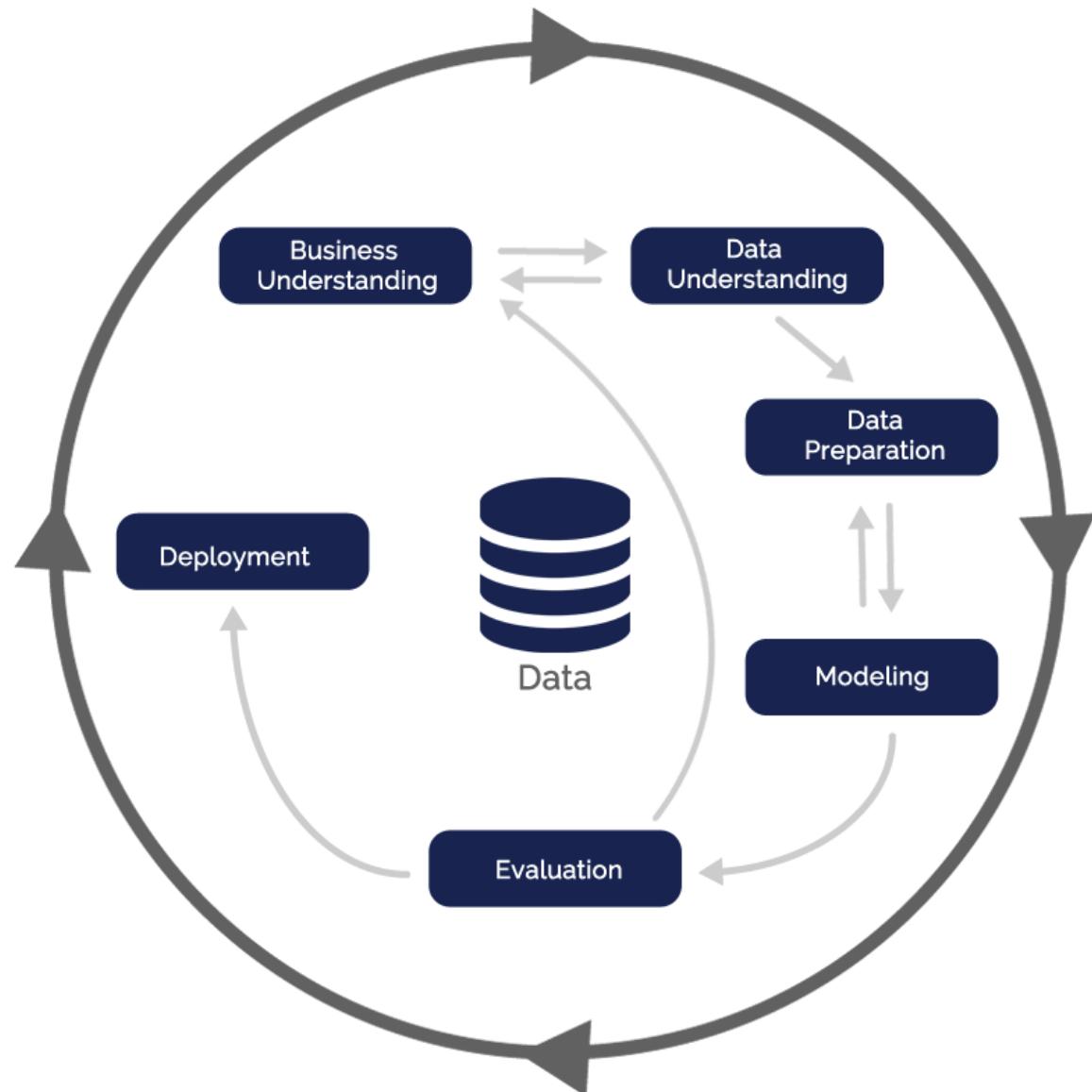
Framework para Modelagem



CRISP-DM: Cross Industry Standard Process for Data Mining

- É a técnica mais utilizada em Mineração de Dados
- Principal vantagem é que pode ser aplicada a qualquer tipo de negócio
- Pode ser utilizada para Data Science pela sua simplicidade
- Consegue-se juntar esta técnica com Scrum

Framework para Modelagem

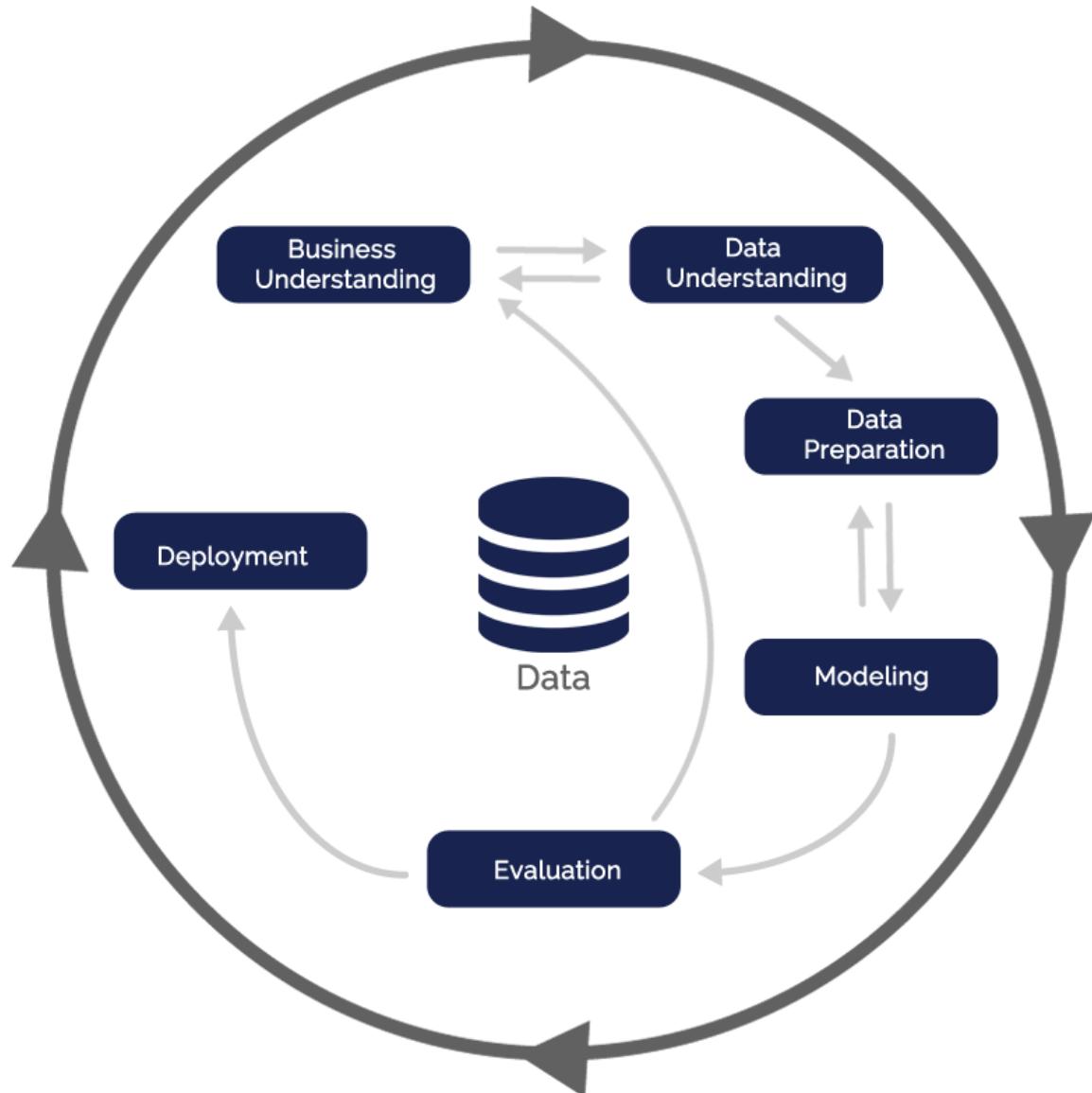


Business Understanding

Nesta fase será necessário identificar a questão a ser resolvida e/ou modelada, sendo necessário formalizar:

- **Background:** Explicar a situação da empresa/entidade e como os produtos serão direcionados para entrega da modelagem
- **Objetivo Modelagem:** Descrever com clareza quais serão as entregas
- **Critério de sucesso:** Definir os KPIs que serão avaliados para comprovar eficácia do modelo, assim como acompanhamento da sua acurácia

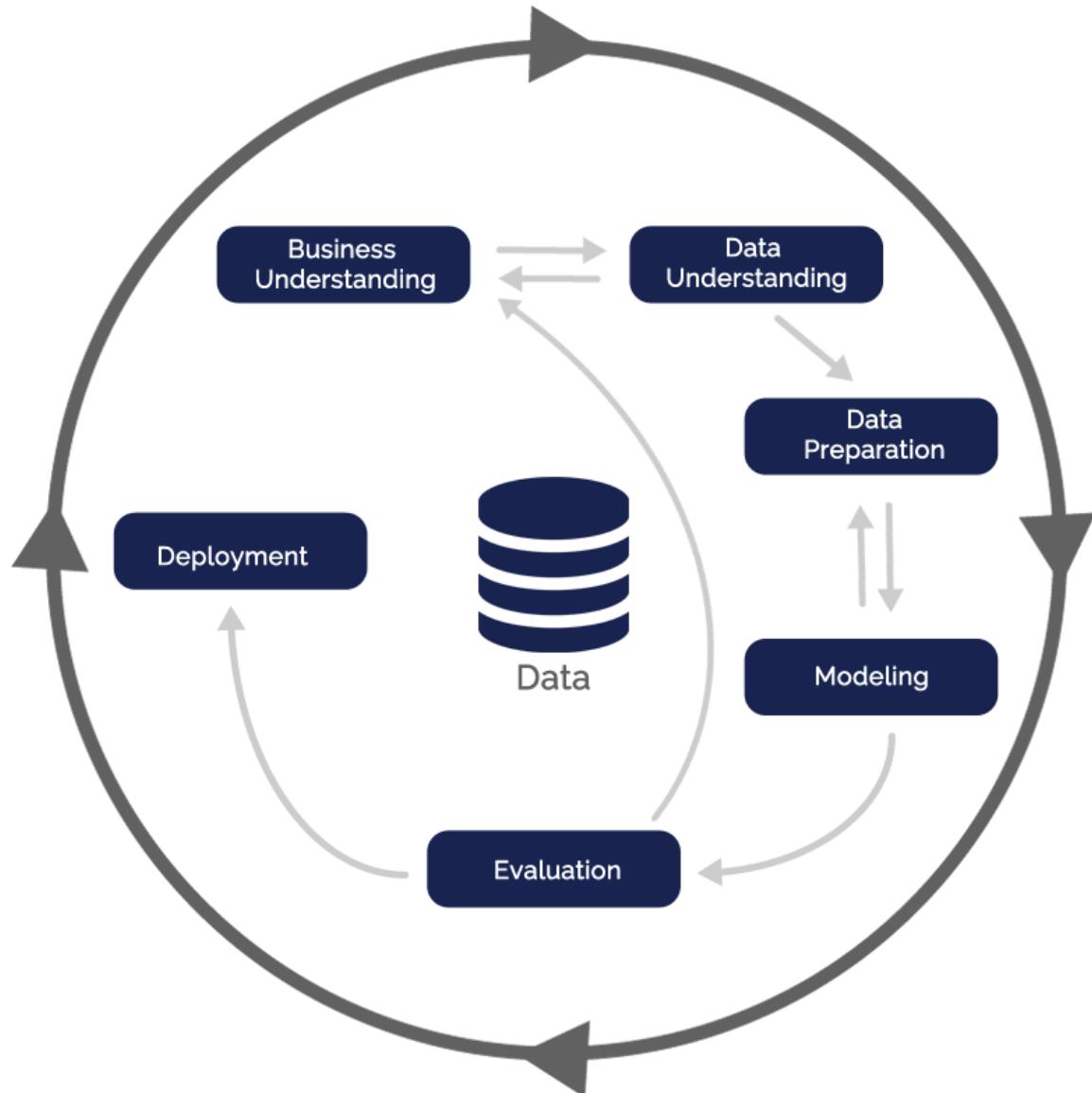
Framework para Modelagem



Data Understanding

- É de conhecimento que coletar e tratar o dado é uma tarefa responsável por mais de 70% do tempo gasto por um Data Scientists
- Exatamente sobre isso que essa fase e a próxima dizem respeito. Aqui, com uso de estatísticas, será necessário: **Coletar, Descrever, Explorar e verificar a qualidade dos dados**

Framework para Modelagem



Data Preparation

Consiste na preparação dos dados para modelagem:

Data Selection: Selecionar os dados que serão usados no modelo. Importante documentar o motivo de tê-los escolhidos

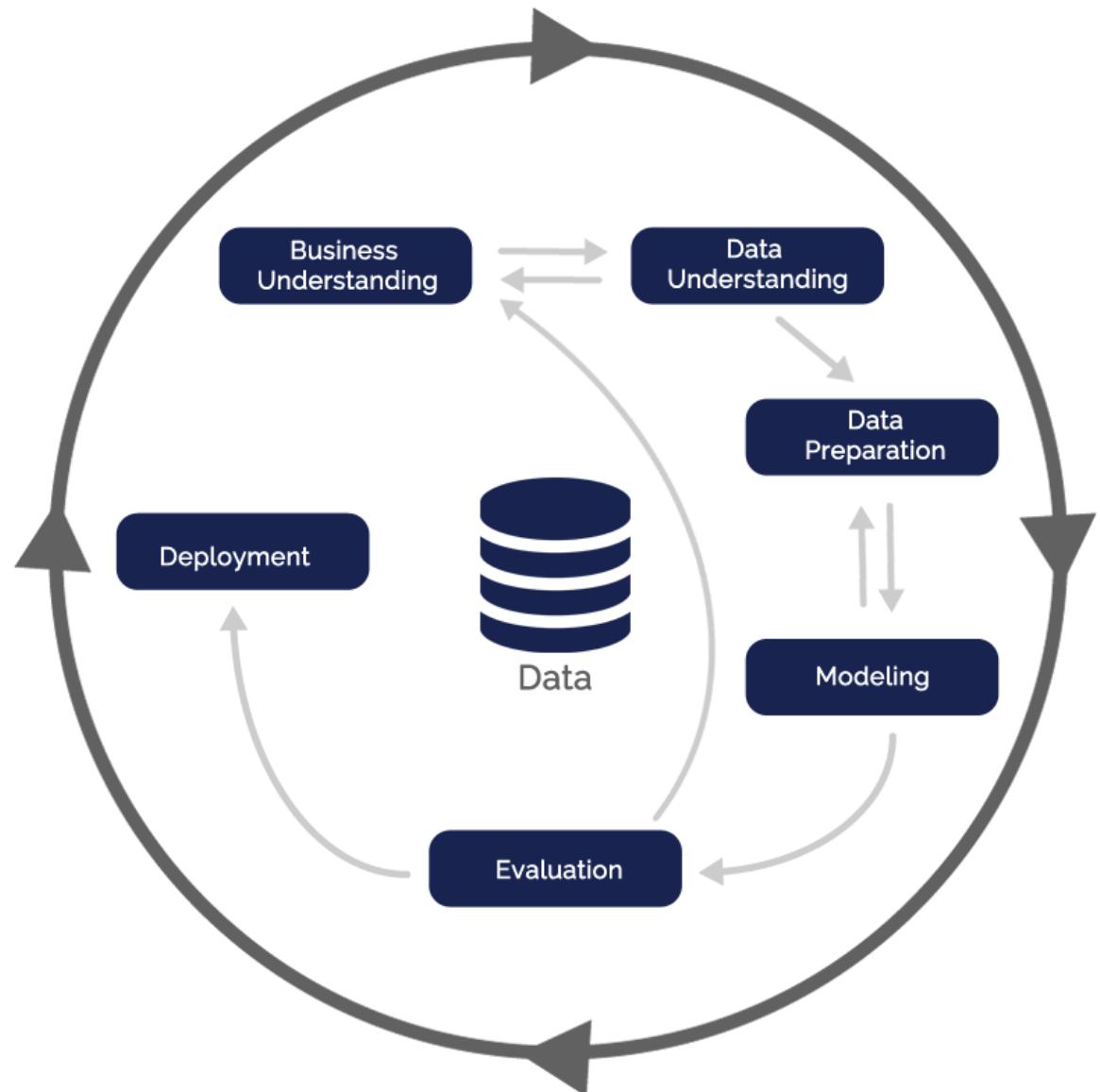
Data Cleaning: Datas em formato incorreto e números inteiros sendo interpretados como string são alguns dos exemplos de questões a serem tratadas

Construct Data: Algumas vezes é necessário criar novos dados para modelagem. Por exemplo, um novo campo ou coluna indicando que a data é feriado ou qual dia da semana representa

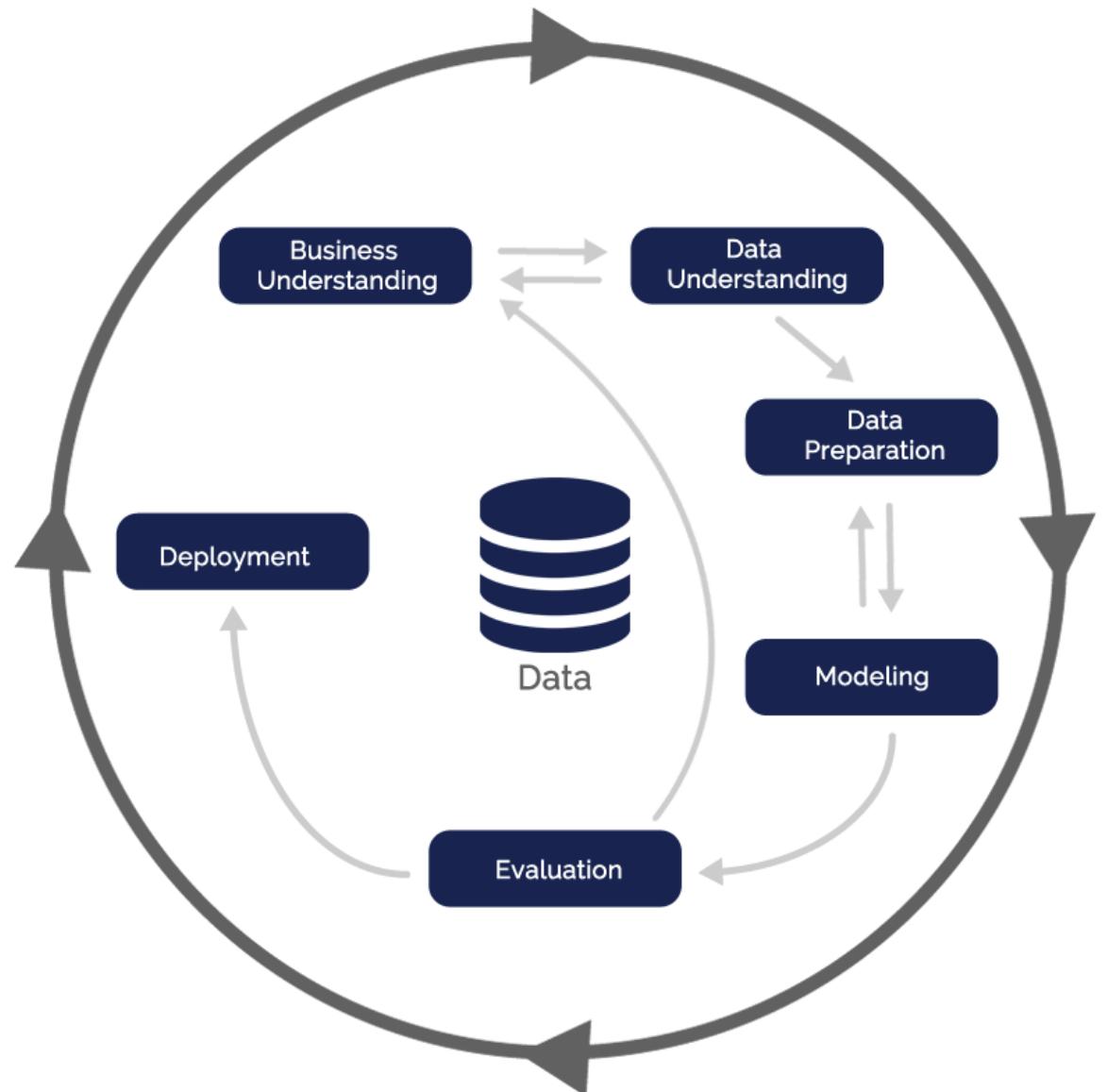
Integrating Data: Necessário para juntar fontes de dados diferentes

Framework para Modelagem

Modeling



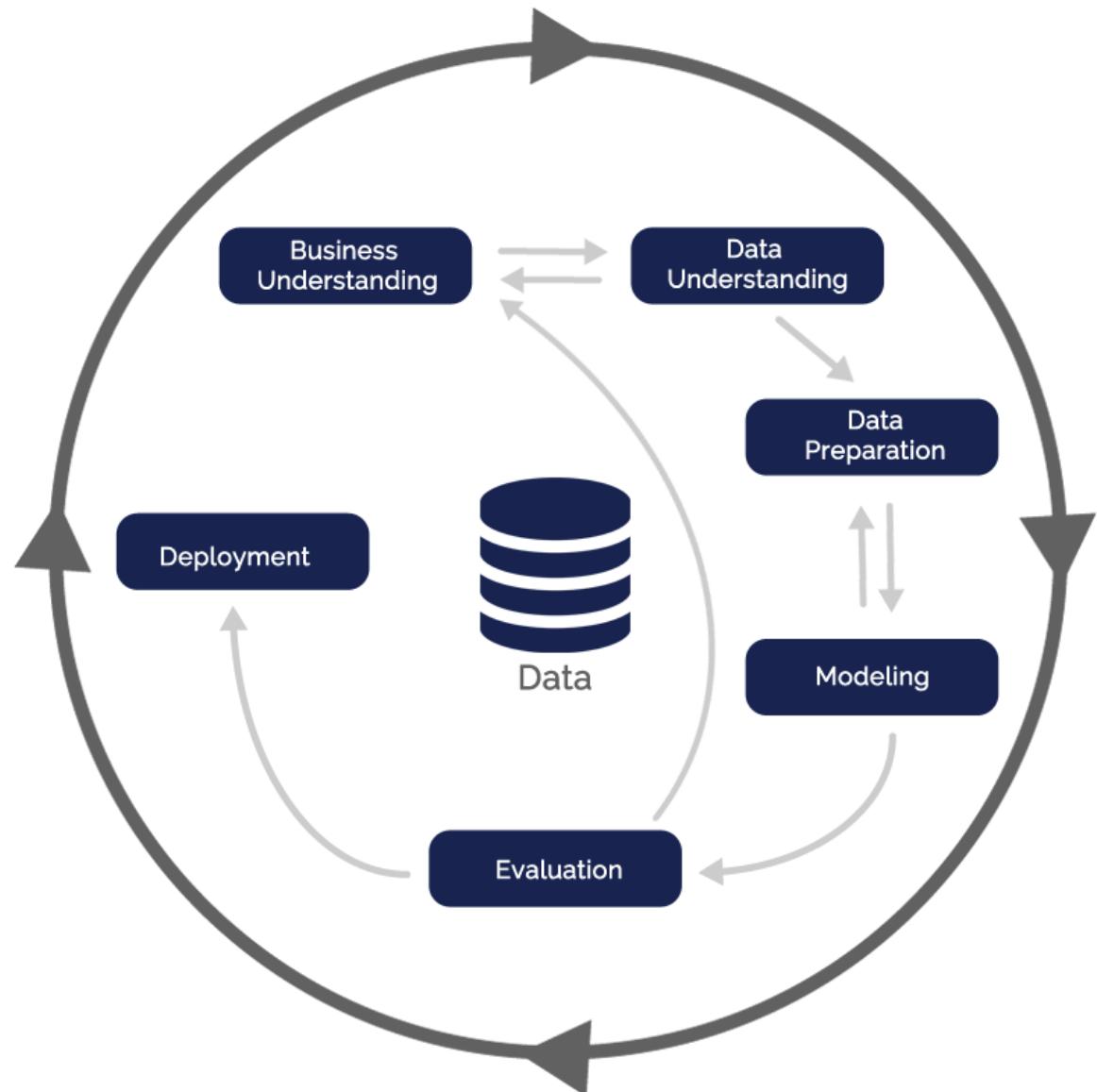
Framework para Modelagem



Modeling



Framework para Modelagem

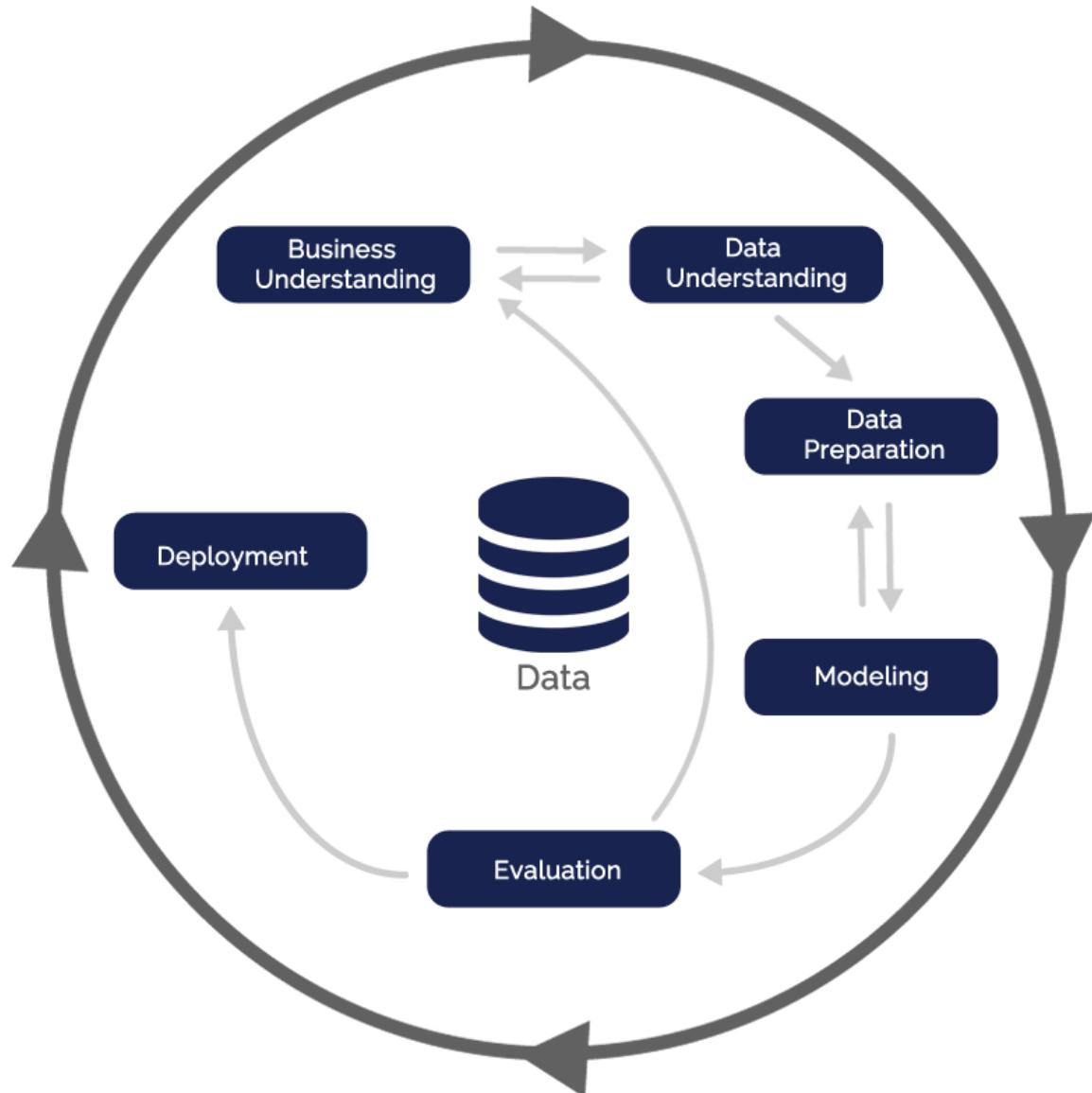


Evaluation

Validação dos resultados da modelagem, utilizando os critérios de sucesso definidos na primeira fase

- No caso de não atingimento dos KPIs, será necessário voltar a primeira fase para entendimento da falha de planejamento
- No caso de atingimento adequado das metas, pode-se prosseguir para fase de implementação da modelagem

Framework para Modelagem



Deployment

- Nesta fase será hora de colocar o modelo em produção, para que possa ser usado
- O deployment encerra a entrega do produto final
- Após será necessário acompanhar e monitorar os resultados
- Sempre que necessário a modelagem deverá ser **readaptada para manter a sua acurácia**

Extract, Transform and Load

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – Namorando os Dados

7 – Welcome to R

8 – Homework



O Trabalho do Cientista de Dados

- 1. Definição do problema e levantamento de perguntas a serem respondidas**
- 2. Planejamento do processo de Data Science**
- 3. Coleta de dados**
- 4. Processamento e limpeza dos dados**
- 5. Armazenamento dos dados**
- 6. Análise de dados**
- 7. Construção e validação de algoritmos e modelos**
- 8. Data Visualization**
- 9. Disseminação da informação**
- 10. Colocar modelo em produção**

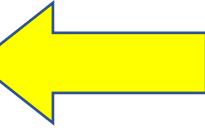


O Trabalho do Cientista de Dados

- 1. Definição do problema e levantamento de perguntas a serem respondidas**
- 2. Planejamento do processo de Data Science**
- 3. Coleta de dados**
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
- 10. Colocar modelo em produção**



O Trabalho do Cientista de Dados

- 1. Definição do problema e levantamento de perguntas a serem respondidas**
- 2. Planejamento do processo de Data Science**
- 3. Coleta de dados** 
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



ETL – Extract, Transform, Load

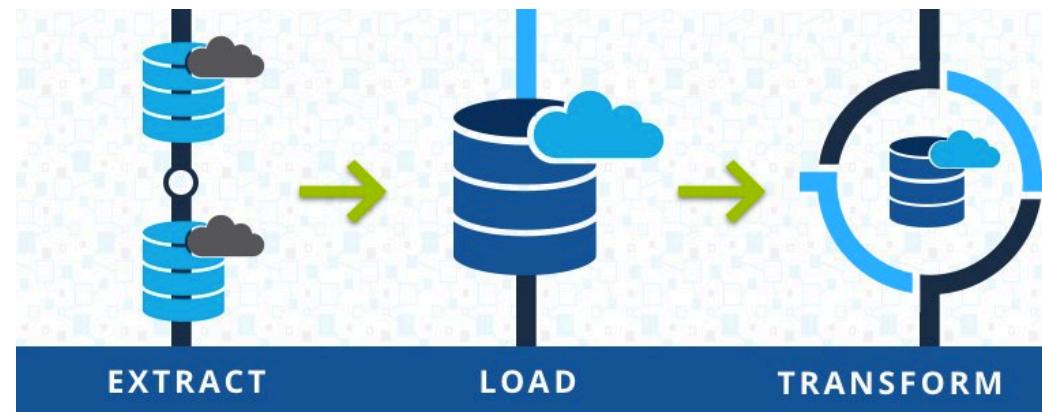
ETL, do inglês **Extract Transform Load** (*Extrair Transformar Carregar*), é uma técnica de processamento de dados extração destes dados de diversos fontes, **transformação** (conforme regras do negócio) e **carregamento** dos dados depurados (Data Mart e/ou Data Warehouse):



- Extração de dados de fontes externas
- Transformação dos dados para atender necessidades de negócios
- Carregamento dos dados

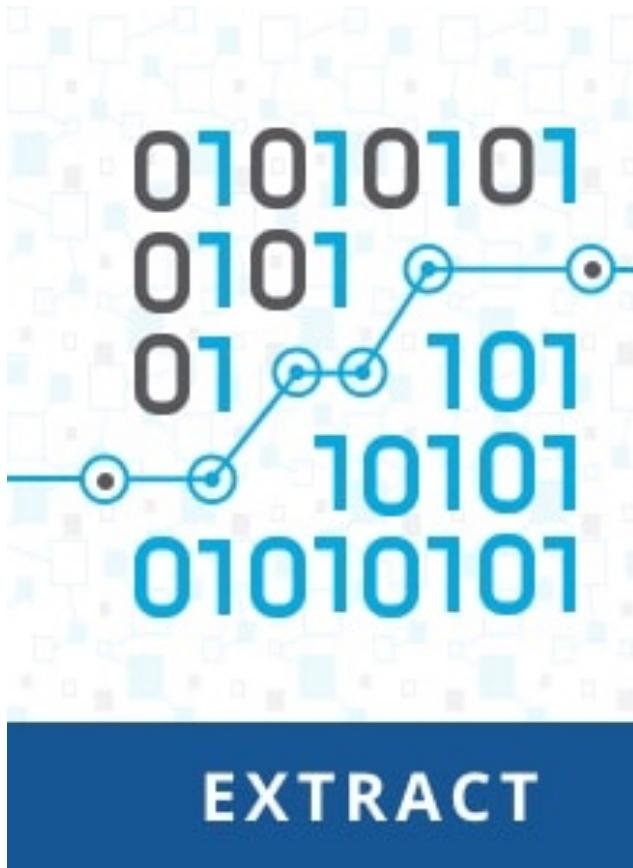
ETL vs ELT

ELT, do inglês **Extract Load Transform**(*Extrair Carregar Transformar*), similar a técnica ETL, porém fazendo o Extração e Carga primeiro para depois aplicar a Transformação



- Conceito que surgiu para melhor tratar Big Data (Algoritmo Map Reduce e ferramentas que o implementam como Hadoop, Spark, utilizam esta técnica)
- Mais ágil e rápido para processo de extração e carga de grandes volumes de dados
- Os dados precisam passar pela regra de transformação para serem utilizados

ETL – Extract, Transform, Load



Extração é a primeira parte do processo de ETL é a extração de dados dos sistemas de origem.

Definição das fontes e dados a serem utilizados.

Ex:

Banco de Dados (Dados de Funcionários)

Arquivo .csv (Dados de Relógio Ponto)

Planilha .xlsx (Cargos e Salários)

ETL – Extract, Transform, Load



Transformação definir e aplicar regras sobre os dados extraídos para melhor utiliza-los.

Podem ser regras de agrupamento de distintas fontes de dados, regras de discretização, transformações de data, hora, escalas e etc.

Exemplo:

Converter salário de valor inteiro para valor decimal

Converter MM/DD/YYYY de data para DD/MM/YYYY

Calcular o valor hora de um funcionário conforme sua remuneração e quantidade de horas trabalhadas em contrato

ETL – Extract, Transform, Load



Carregamento consiste em publicar estes dados. Esta publicação pode ser em um DW (Data Warehouse), em um Data Mart, em uma tabela para ser consumida por um BI ou mesmo uma aplicação e etc.

A temporização e o alcance de reposição ou acréscimo constituem opções de projeto estratégicas que dependem do tempo disponível e das necessidades de negócios

ETL na Prática - Pentaho



O ETL ou ELT por ter uma técnica pode ser implementado em qualquer linguagem.

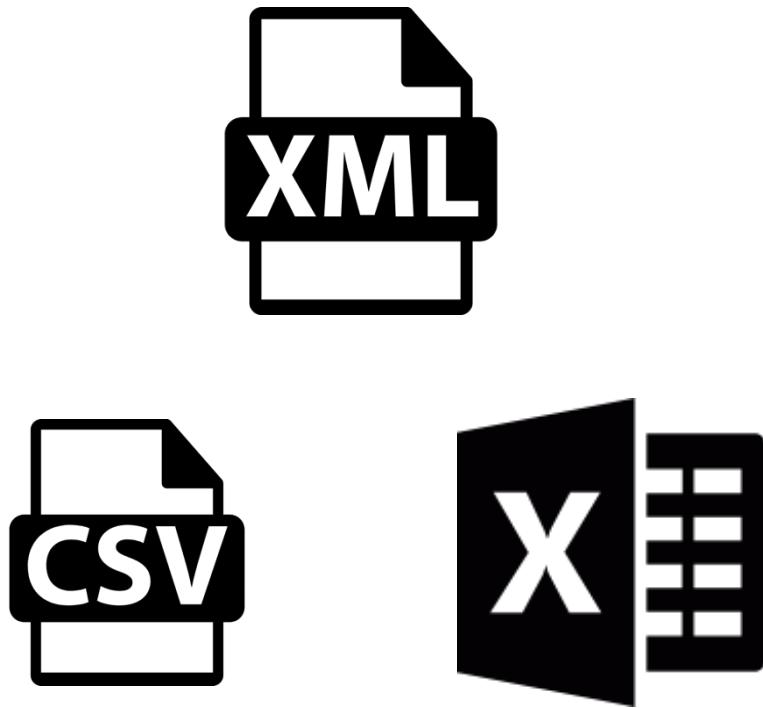
Pentaho Data Integration (Kettle) → Framework com soluções para fluxo de automação de forma produtiva, profissional e didática.

<https://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>

ETL – Extract, Transform, Load



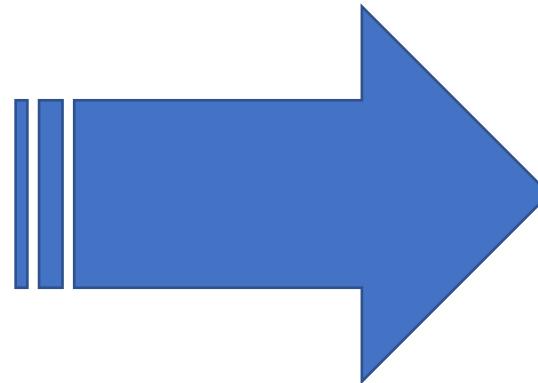
ETL na Prática – Exercício



XML → Dados de Funcionários

CSV → Dados de Relógio Ponto

EXCEL → Cargos e Salários



Planilha Excel com:

Matrícula

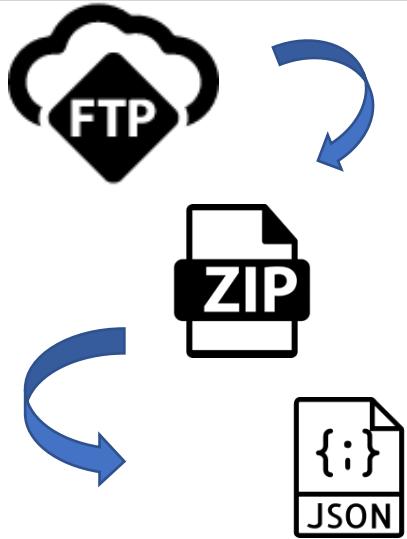
Nome Funcionário

Cargo

Valor Hora

Dia e Hora Marcação Ponto

ETL na Prática – Homework



JSON → Com remuneração variável por funcionário

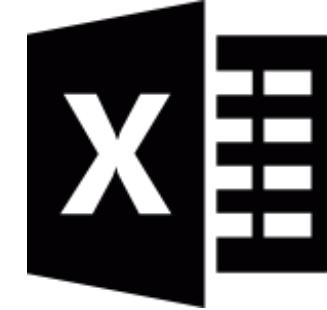
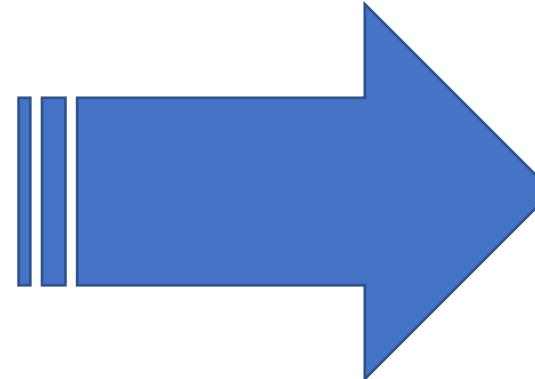
ftp server: [ftp.drivehq.com](ftp://ftp.drivehq.com)

User: datascienceandbigdata@gmail.com

Password: ds2019FTP

Diretório: GroupWrite

Arquivo: remunera.zip



Planilha Excel com:

Matrícula

Nome Funcionário

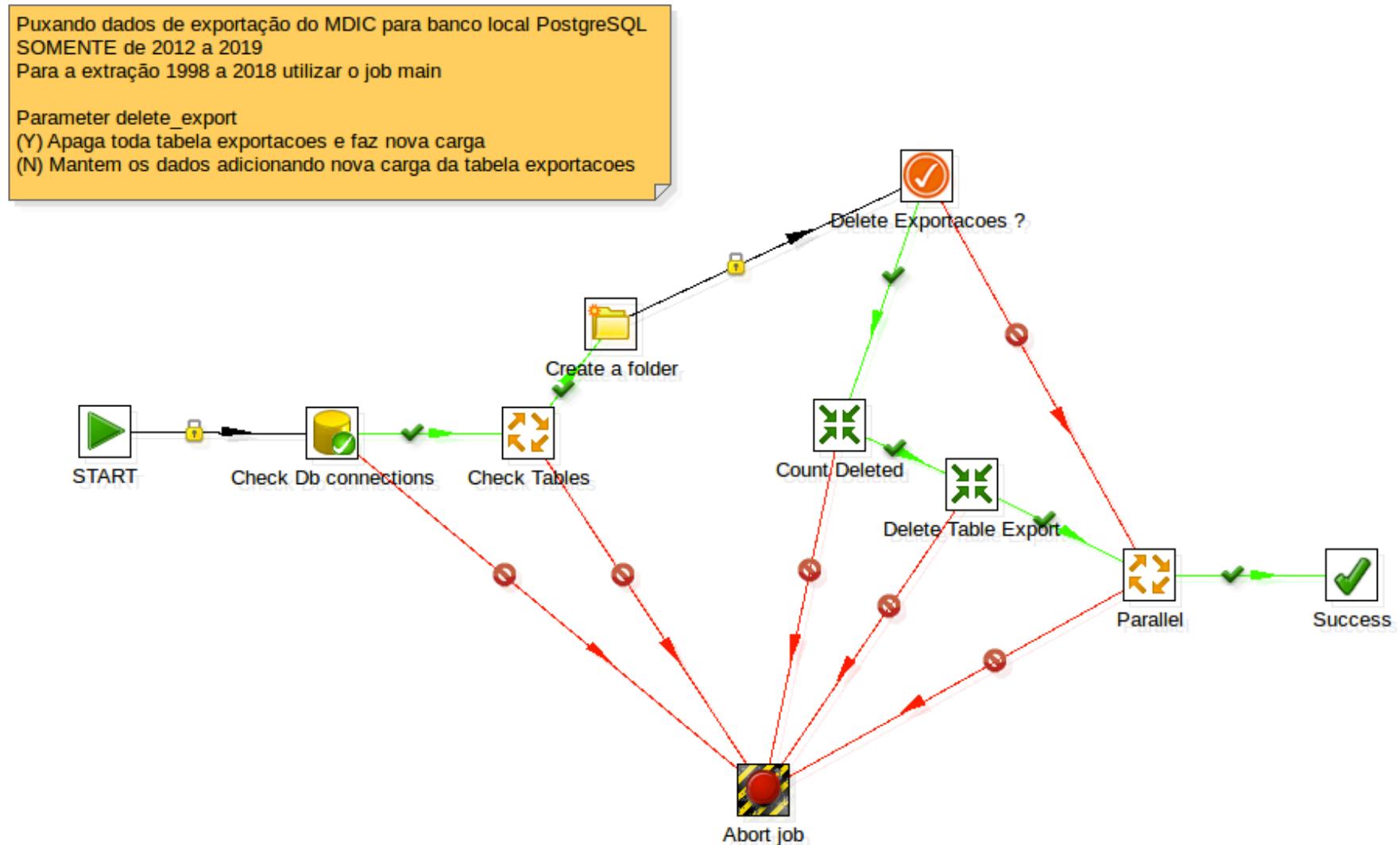
Cargo

Valor Hora

Último Dia e Hora Marcação Ponto

Total Remuneração Variável

ETL na Prática – Desafio AgroXP - Pentaho



Modelo de Dados

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – Namorando os Dados

7 – Welcome to R

8 – Homework

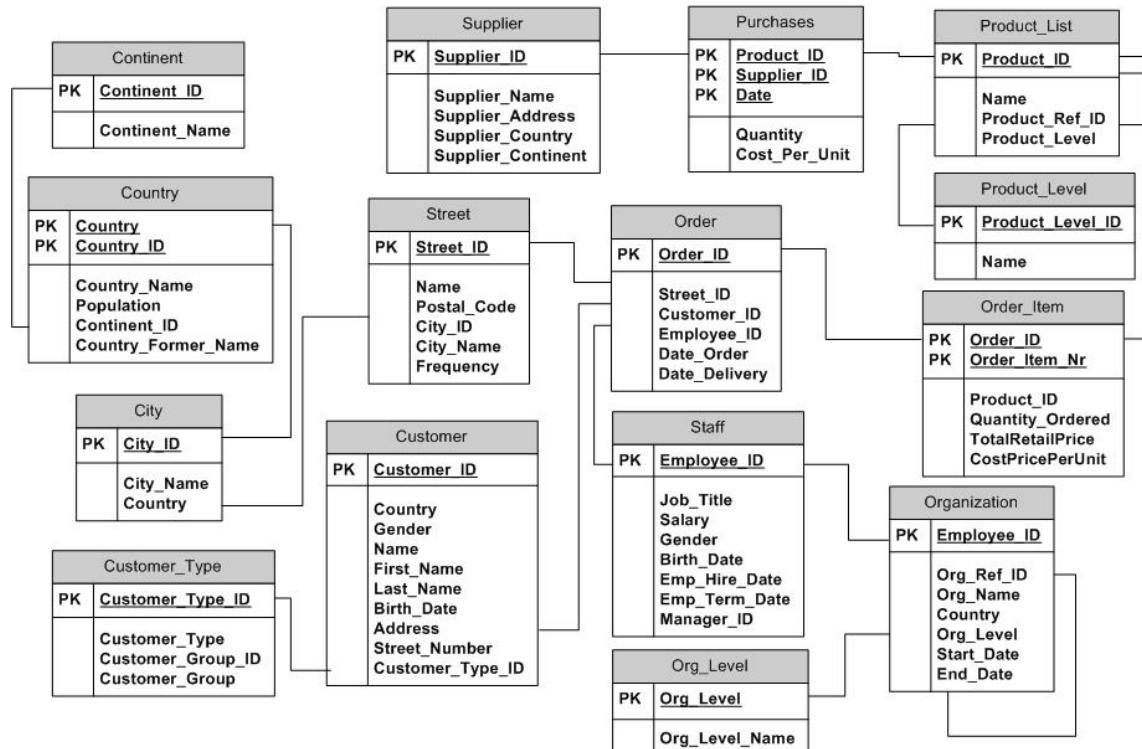


Estratégia – Modelo de Dados

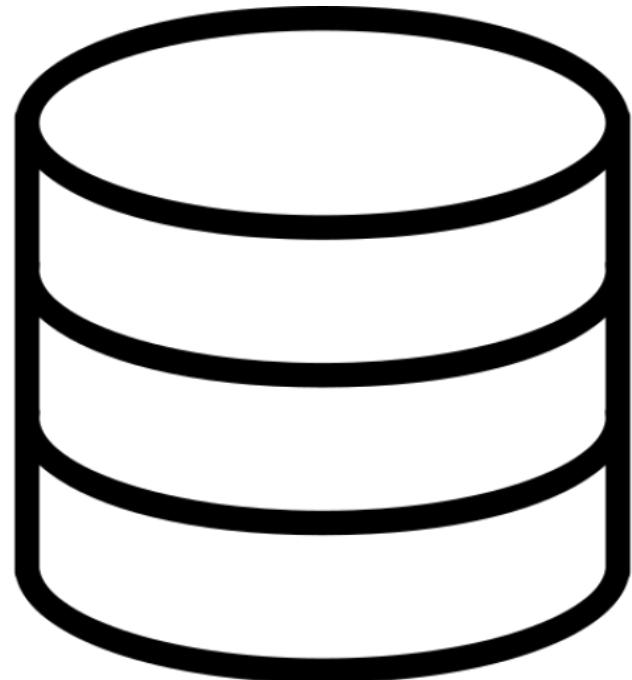


O que é um Modelo de Dados?

É a representação de todos os dados de maneira lógica. Cada dados estará em uma tabela, com atributos, que possuem relação com outros dados.

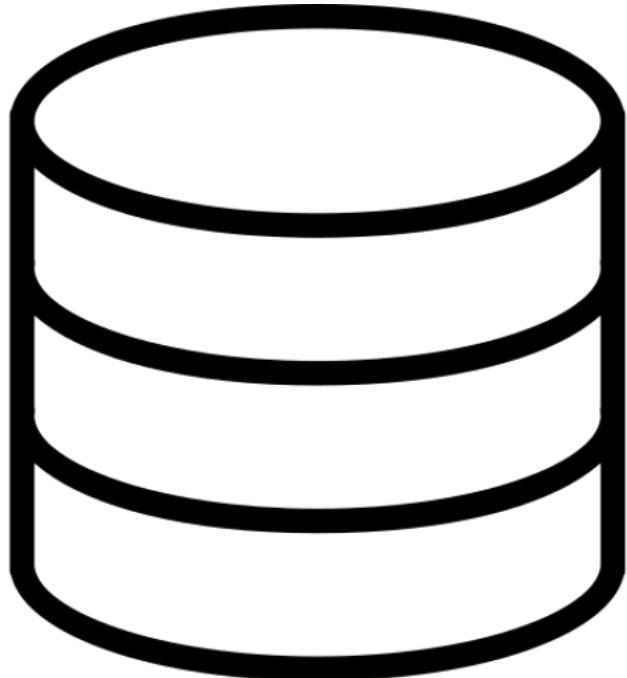


Estratégia – Modelo de Dados



E o nosso modelo do desafio Agro XP Brazil?

Estratégia – Modelo de Dados



Desafio AgroXP Brazil

Qual o dados que estou utilizando?

Você possui os seguintes dados:

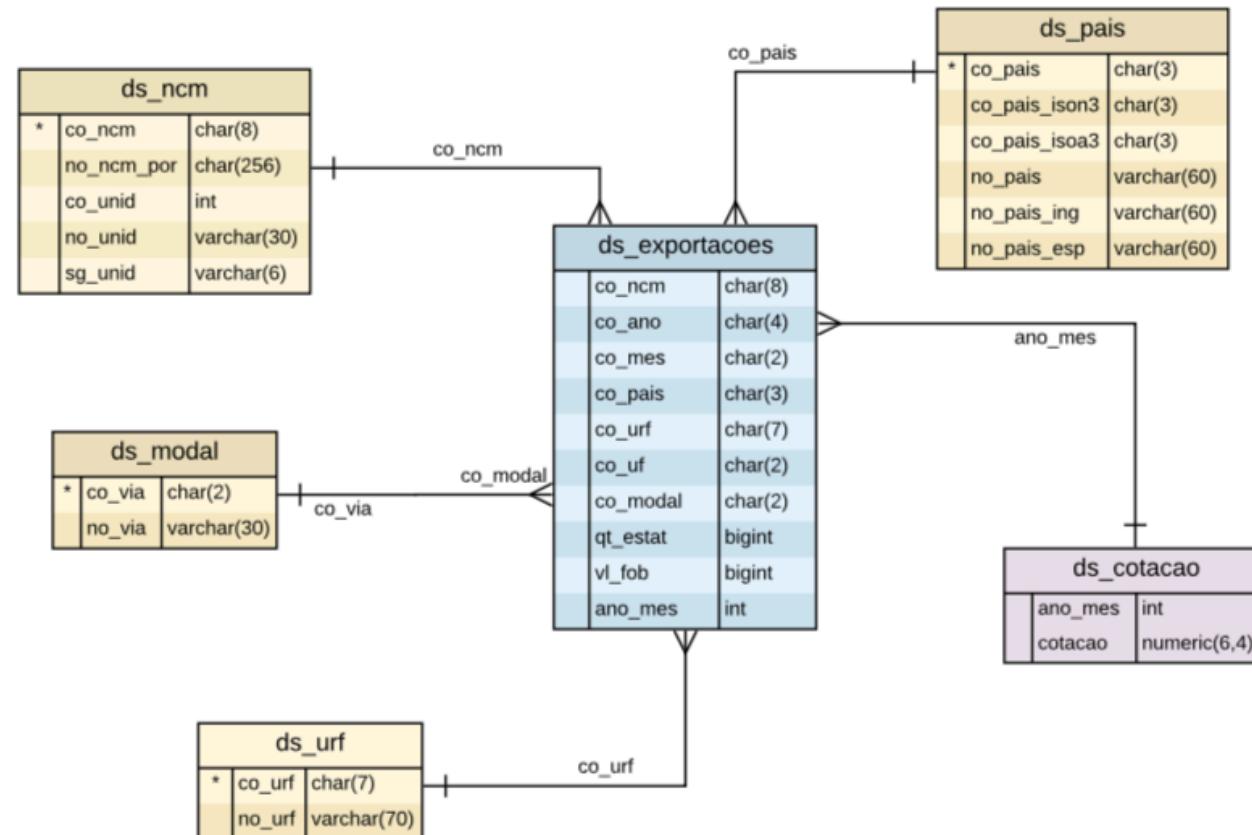
- 1) [Ministério de Desenvolvimento Indústria e Comércio](#) --> apresenta os dados de *TODOS commodities exportados no País desde 1997 até 1 mês atrás (formato .csv)*
- 2) *Tabelas auxiliares de nomenclatura de produtos com NCM – Nomenclatura Comum do Mercosul (formato .xls)*
- 3) *Taxa cambial mensal desde 1997 (formato .csv)*

Vamos entender e montar um modelo!

Estratégia – Modelo de Dados



Desafio AgroXP Brazil – Modelo de Dados



Fontes dos Dados:



Banco de Dados

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

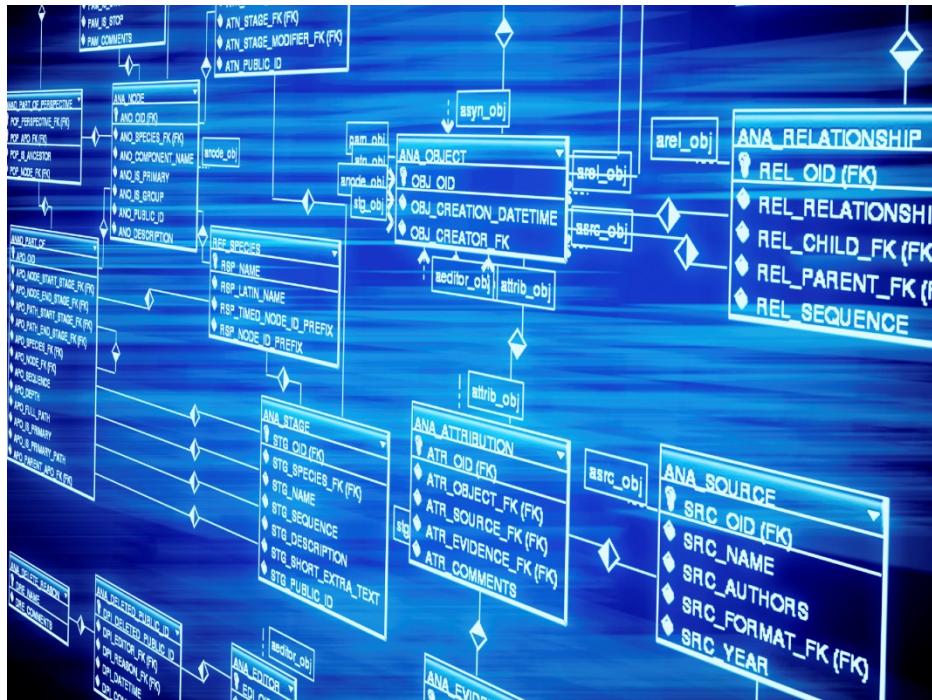
6 – Namorando os Dados

7 – Welcome to R

8 – Homework

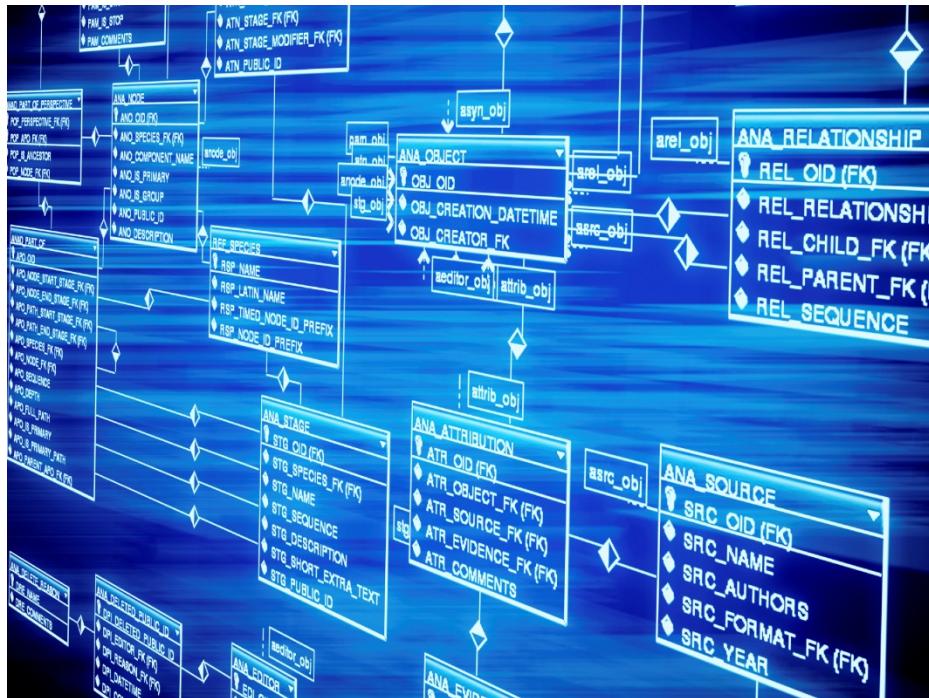


Banco de Dados



- **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

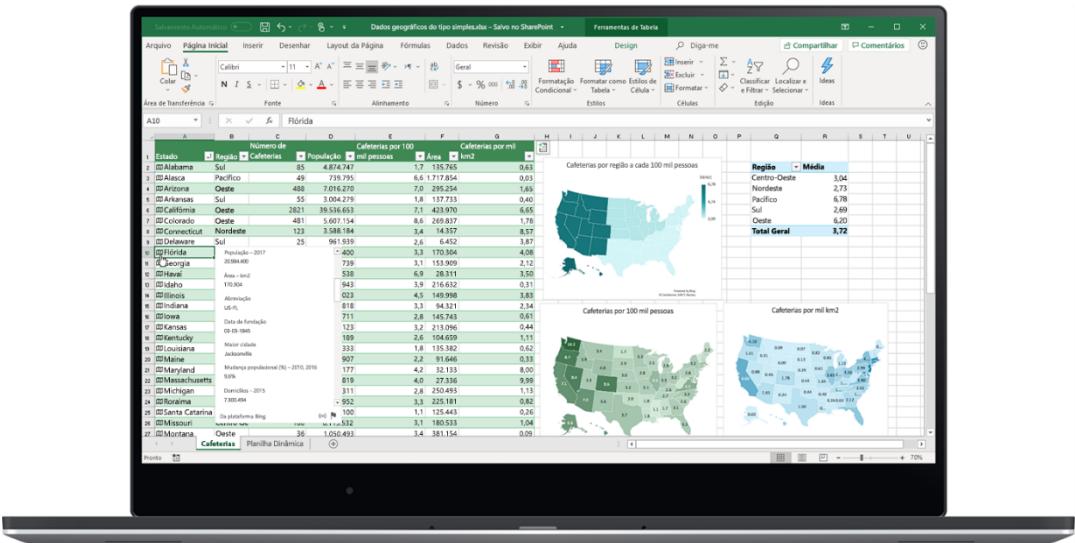
Banco de Dados



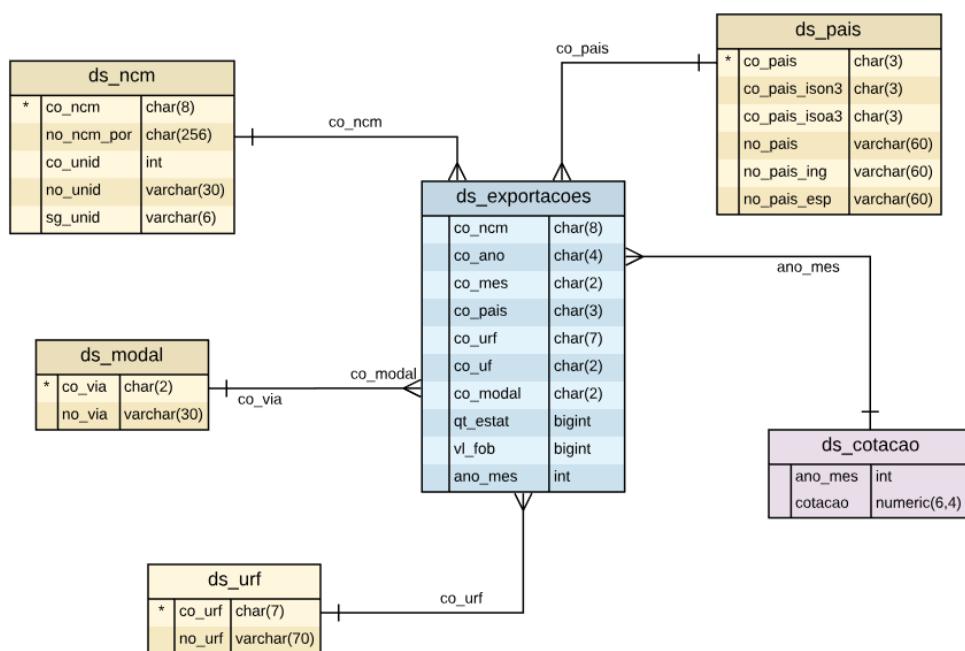
- São gerenciados por um SGDB (no nosso caso o PostgreSQL). Exemplos de Outros SGBDs:
Relacionais (meados 1970) → Oracle, SQL Server, MySQL, DB2, MonetDB.
NoSQL (1998) → MongoDB e Cassandra e etc.
- **Banco de Dados Relacional**
 - Relações tabulares (Linha e Coluna)
 - Consistente / Íntegro
 - Relação cartesiana entre os dados
 - Custo Escalabilidade (Gerir os Dados)
- **Banco de Dados Não Relacional (NoSQL)**
 - Orientado ao documento
 - Não garante Consistência/Integridade
 - Custo Menor Maior Escalabilidade (Gestão menos onerosa dos dados)

Banco de Dados - Relacionais

- Relações Matriciais / Tabulares (Tabelas)



Banco de Dados - Relacionais



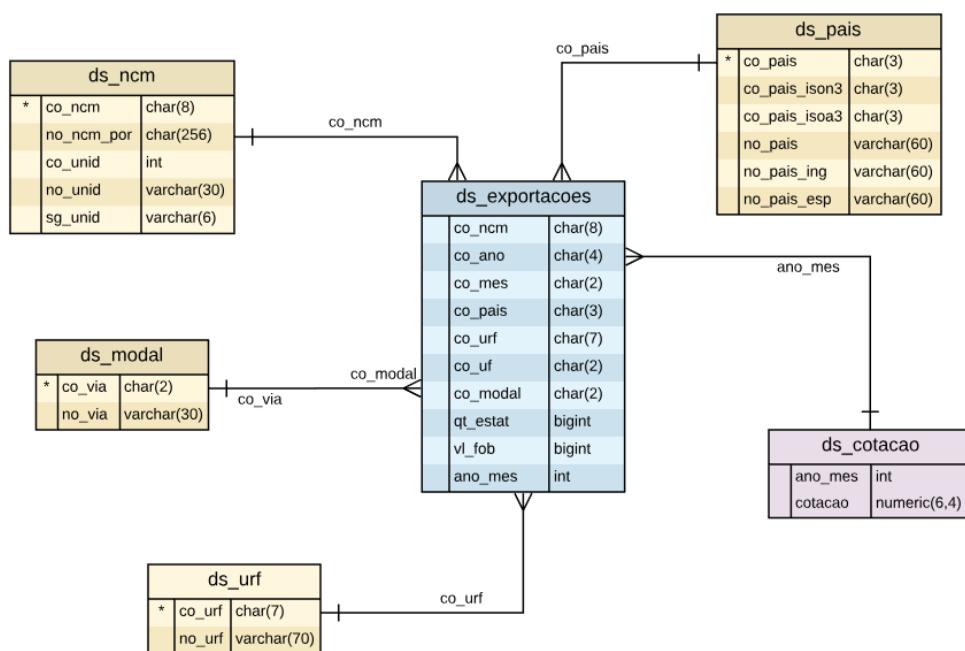
- **Relações Matriciais / Tabulares (Tabelas)**
- Todos os dados de um banco de dados relacional são armazenados em **tabelas**
- Uma tabela é uma simples estrutura de **linhas** e **colunas**
- **Linha** → Registro / **Coluna** → Atributo
- As tabelas associam-se entre si por meio de regras de relacionamentos,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

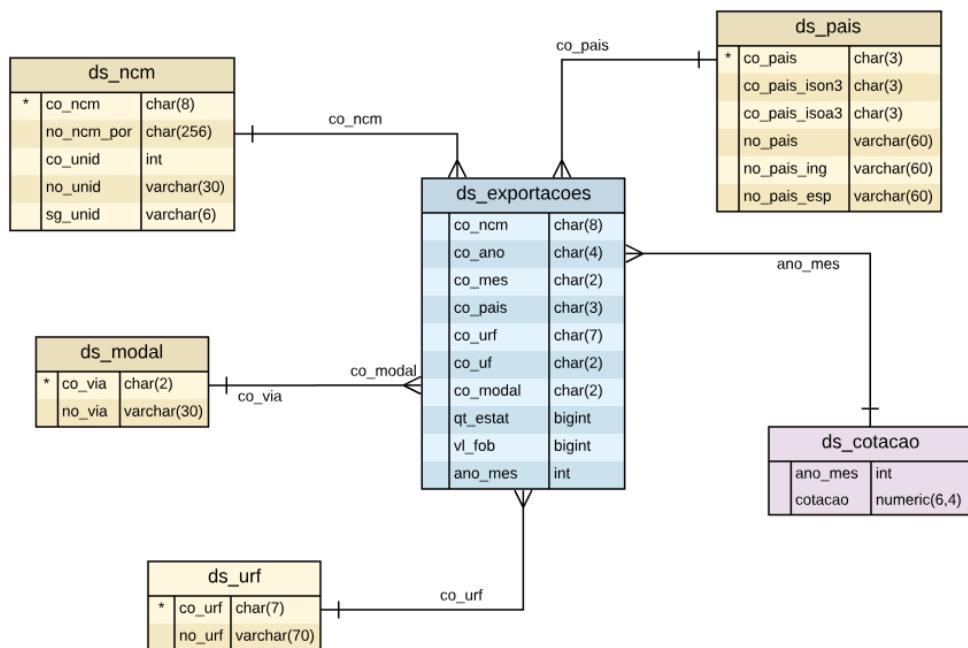
▲	CO_ANO	CO_MES	CO_NCM	CO_UNID	CO_PAIS	SG_UF_NCM	CO_VIA	CO_URF	QT_ESTAT	KG_LIQUIDO	VL_FOB
1	1997	3	41043911	15	149	RS	1	1010500	3987	4150	16725
2	1997	5	63019000	10	97	MG	7	145200	0	1002	8420
3	1997	6	87168000	11	586	RS	7	145300	48	153	915

Banco de Dados



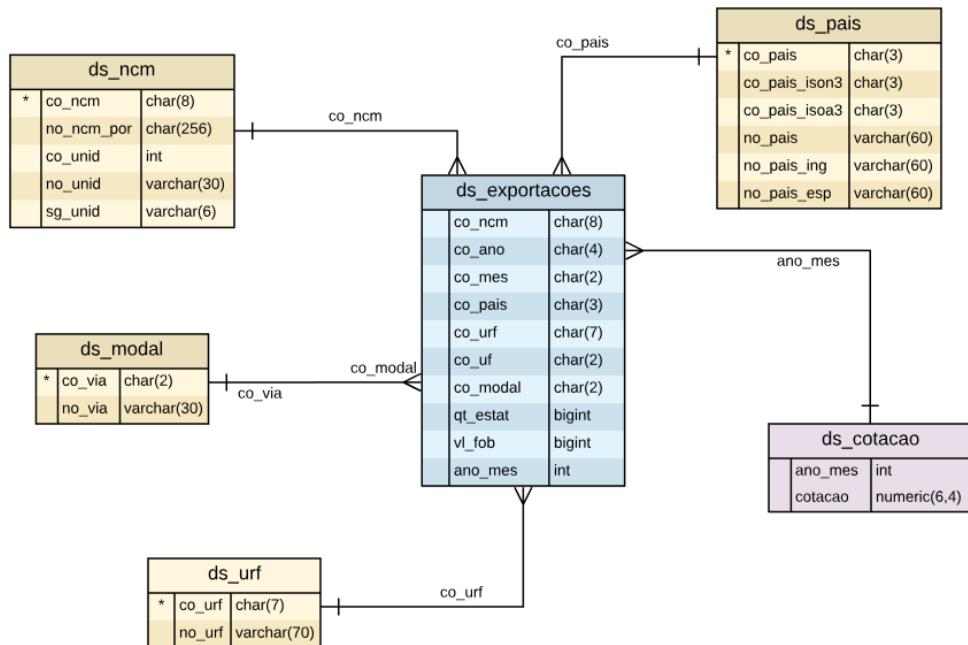
- **Chave**
- **Integridade → Tupla/Registro/Linha única**
- **Chave primária: (PK - Primary Key)**
 - A chave primária nunca se repetirá
- **Chave Estrangeira: (FK - Foreign Key) é a chave formada através de um relacionamento com a chave primária de outra tabela.**
 - Define um relacionamento entre 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

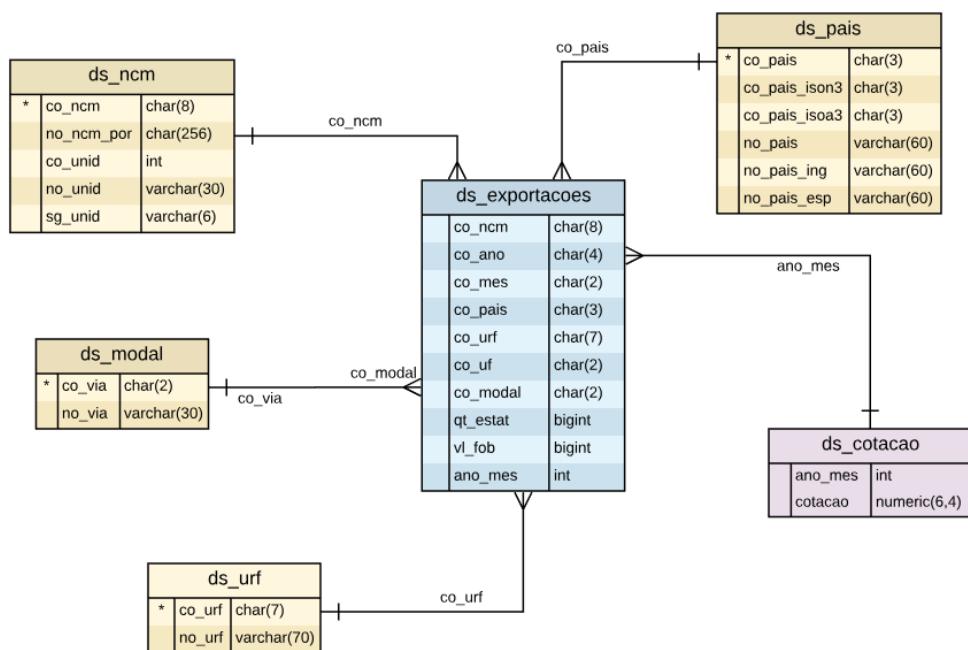
Banco de Dados - SQL



- 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

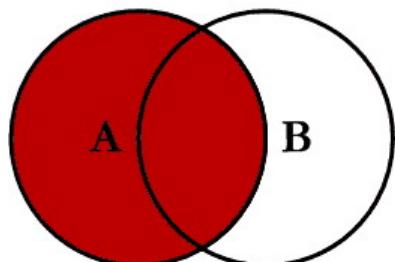
Banco de Dados - SQL

- 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)

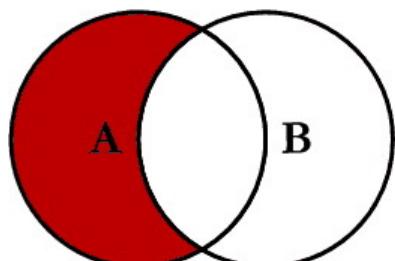


Queries Seleção

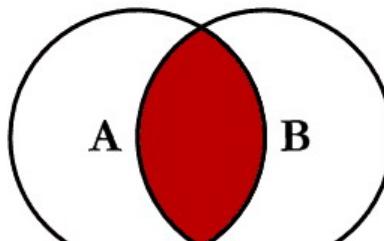
SQL JOINS



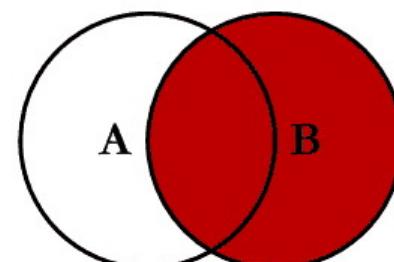
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



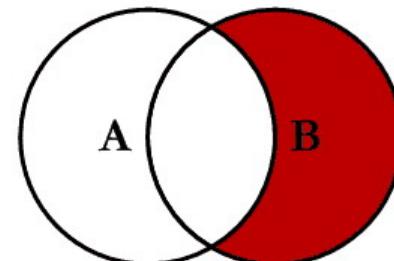
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```



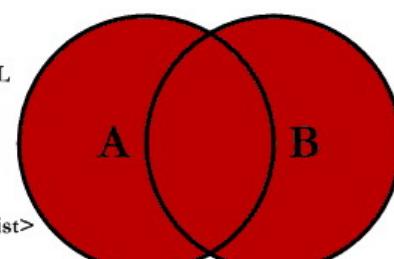
```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```

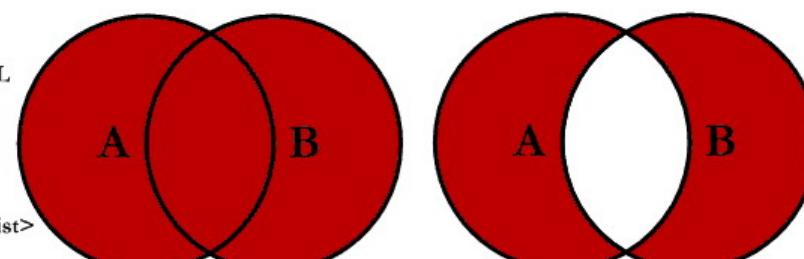


```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```

© C.L. Moffatt, 2008

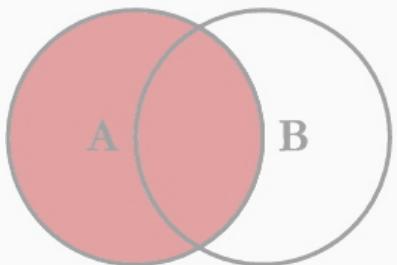


```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```



Queries Seleção

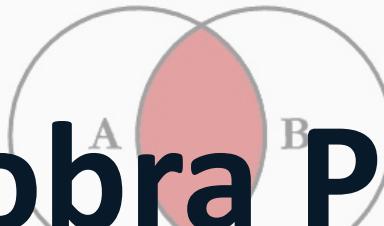
SQL JOINS



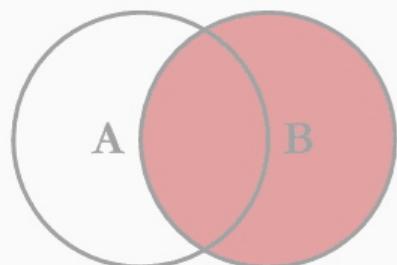
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```



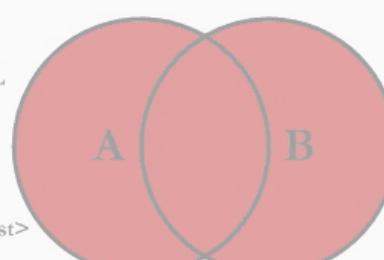
```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



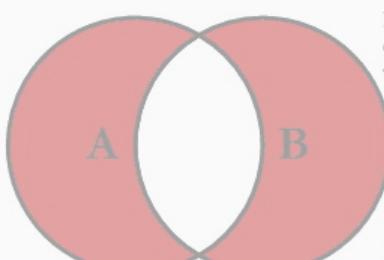
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

Namorando os Dados

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

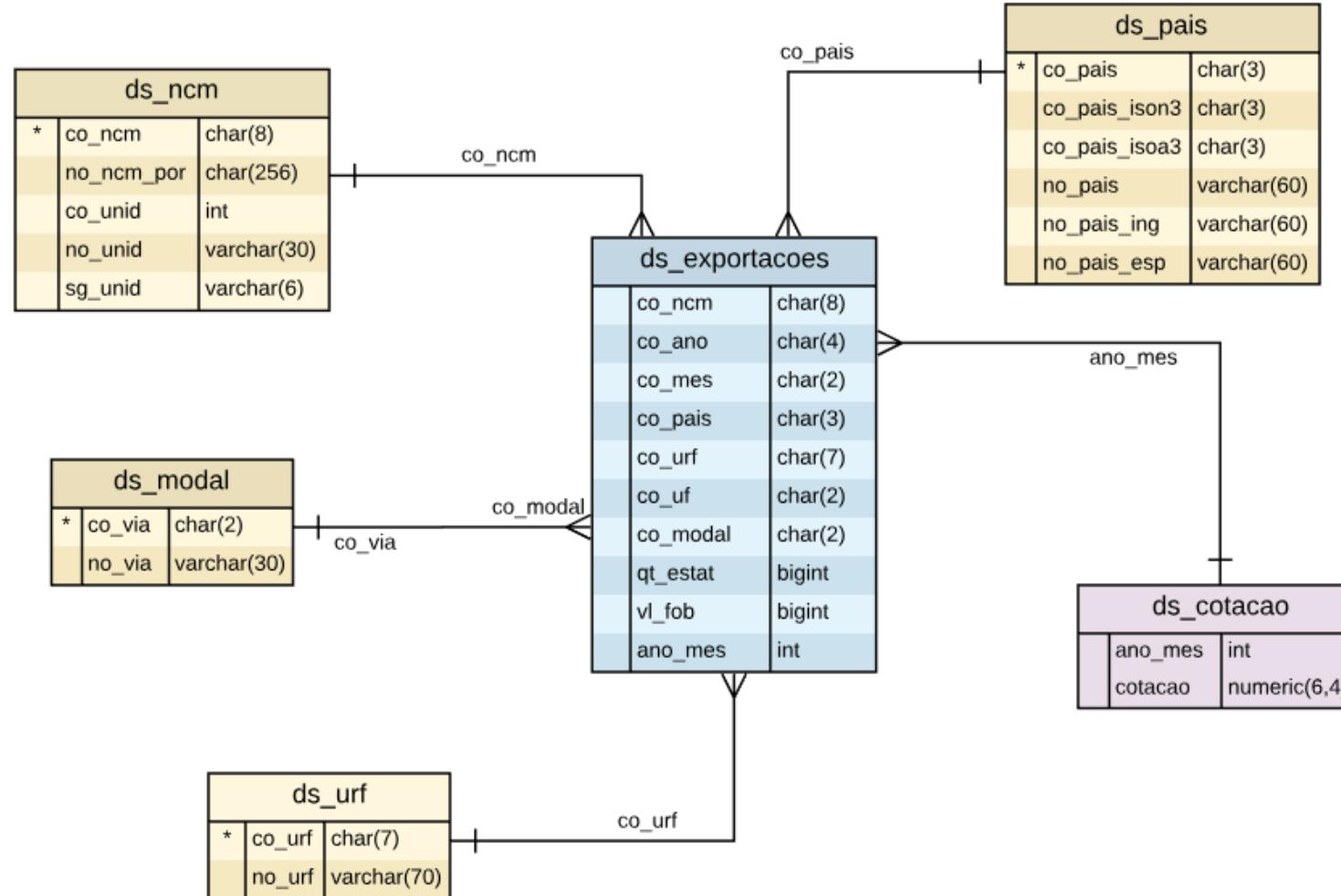
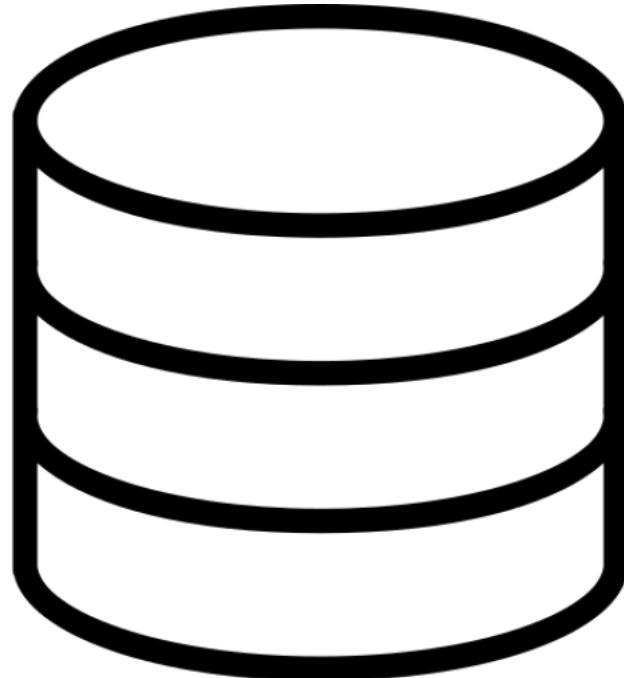
6 – Namorando os Dados

7 – Welcome to R

8 – Homework

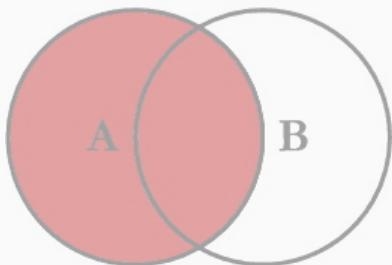


Desafio – Modelo de Dados



Namorando os Dados (Queries SQL)

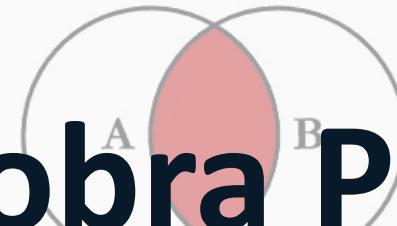
SQL JOINS



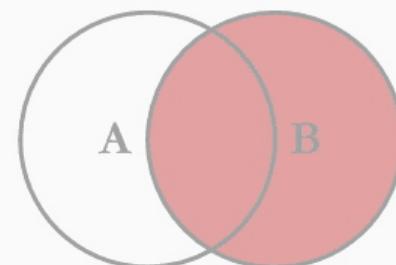
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```



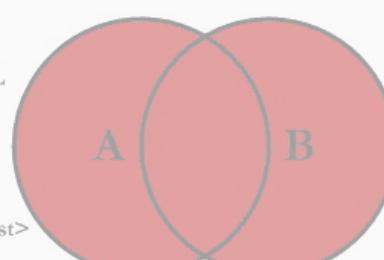
```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



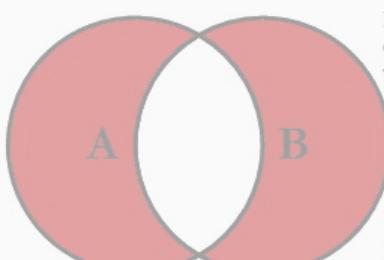
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

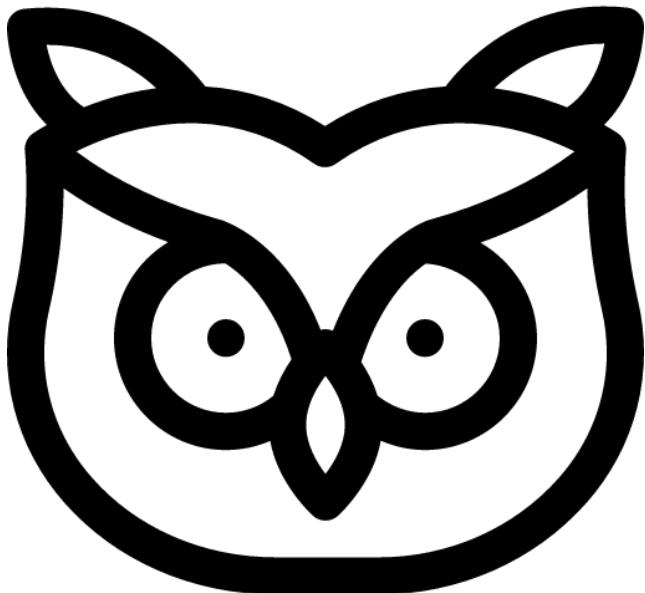
Mãos à obra Pessoal!!!

O Trabalho do Cientista de Dados > Desafio Curso

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



Quero Saber Mais...



HITACHI
Inspire the Next

Hitachi Vantara | Pentaho Documentation

Version 7.1 | How can we help you?

Home > Documentation > 7.1

Tutorials

Last updated: Nov 29, 2016

[Getting Started with PDI >](#)

If you are new to PDI, start here. These tutorials provide step-by-step instructions for creating and refining transformations and jobs using the PDI client (Spoon).

[PDI Transformation Tutorial](#)

[PDI Job Tutorial](#)

[Getting Started with PDI and Hadoop](#)

PostgreSQL Tutorial



Welcome to R

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – Namorando os Dados

7 – Welcome to R

8 – Homework



Welcome to R

1 – Aprendendo Linguagem R no RStudio

2 – Analisando Qualidade dos Dados

3 – Variáveis Relevantes



Aprendendo Linguagem R no RStudio

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 Intelligence

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

Aprendendo Linguagem R no RStudio



**Bora
Praticar?**



Homework

1 – Agenda

2 – Framework para Modelagem

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

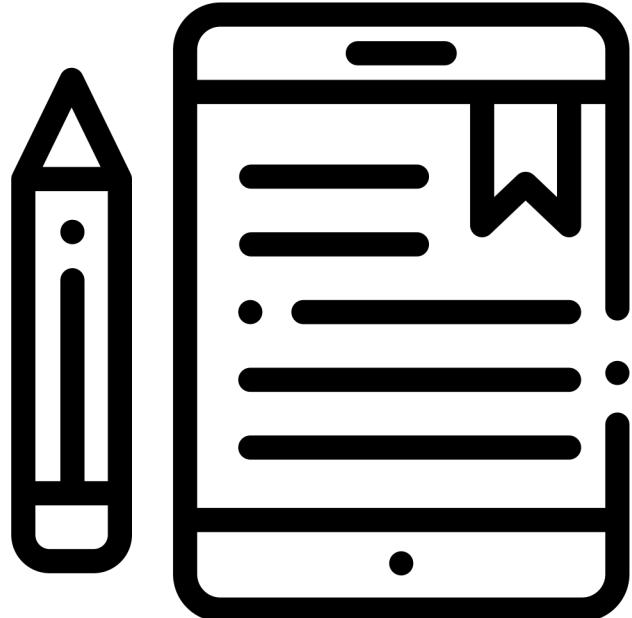
6 – Namorando os Dados

7 – Welcome to R

8 – Homework



Homework



Importante praticar durante a semana a lista de exercícios para melhor fixação da parte prática do conteúdo:

- ETL
- Query SQL – Namorando Dados
- Lista R

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