Data Science

Charles Adriano dos Santos

Turma Março/2020


ald
eia
Sejam
bem-vindos!
Utilize a nossa redes de wi-fi:
#ALDEIA
utilizando a senha
4zamnk

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

Pauta

1 – Apresentação Alunos

2 – Professores

3 – Agenda

Apresentação Alunos

1 – Apresentação Alunos

2 – Professores

3 – Agenda

Apresentação Alunos

Galera, queremos conhecer vocês!!

Nome

Área de atuação / Empresa

O que é Data Science para vocês?

Expectativa com o curso

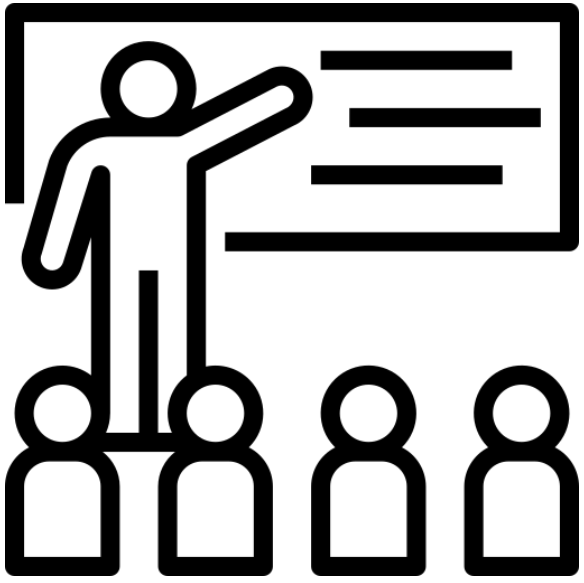
Professores

1 – Apresentação Alunos

2 – Professores

3 – Agenda

Professores



Charles Adriano dos Santos
Bacharel em Análise de Sistemas - PUCPR
Especialista em Engenharia de Software - PUCPR
Especializando em Data Science & Big Data - UFPR

Rafael Roberto Dias
Bacharel em Estatística - UFPR
Especialista em Data Science & Big Data - UFPR

Agenda

1 – Apresentação Alunos

2 – Professores

3 – Agenda

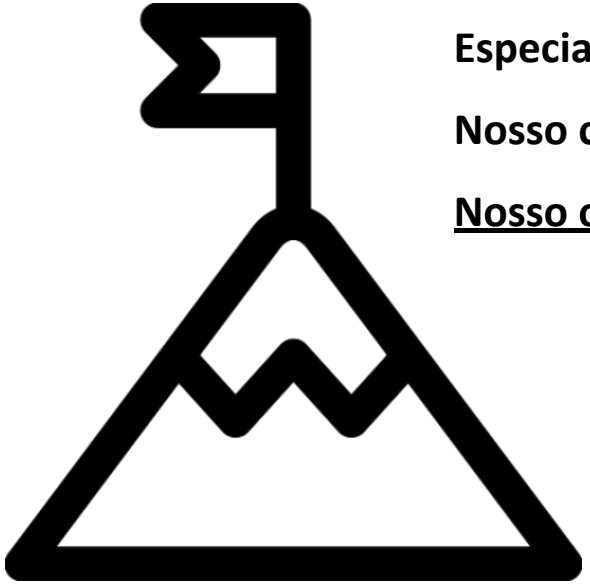
Manhã

Horário	Assunto
09:00	Apresentação e Anseios dos Futuros(as) Cientistas de Dados
09:30	Agenda, Estrutura, Objetivos e Material do Curso
09:45	O que é Data Science, Mercado Atual, Profissão, Projeções
10:30	O Trabalho do Cientista de Dados e Matriz de Habilidades
11:00	O Presente do Curso → A VM do Cientista de Dados
12:00	Almoço

Tarde

Horário	Assunto
13:00	Desafio Curso → O Desafio da AgroXP Brazil
13:30	ETL
14:30	Modelagem de Dados, Bancos de Dados, SGDB
15:30	SQL - Comando Básicos
16:30	Namorando Dados - Análise Exploratória Desafio Curso

Objetivo



Graduação área de Tecnologia da Informação de 3 a 5 anos

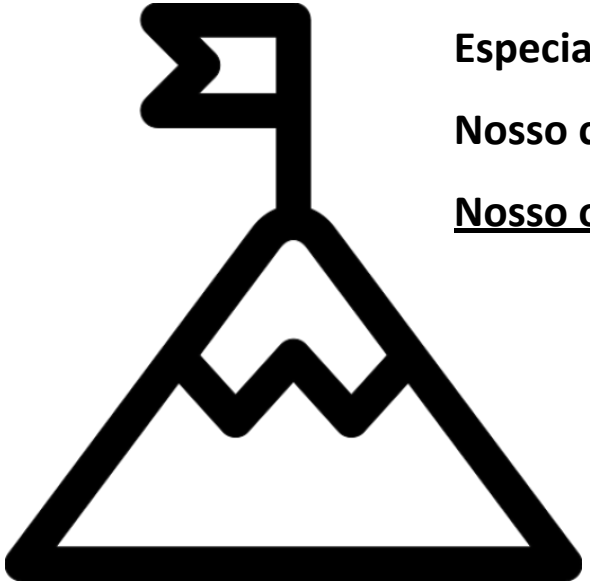
Graduação em Estatística de 4 a 6 anos (média de 5 anos)

Especializações (em TI, Estatística ou Ciência de Dados) de 1 a 2 anos

Nosso curso → 32h

<u>Nosso objetivo neste tema?</u>

Objetivo



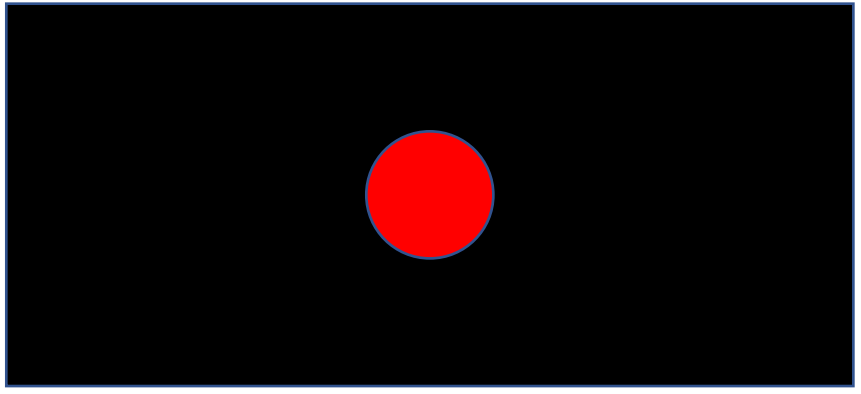
Graduação área de Tecnologia da Informação de 3 a 5 anos

Graduação em Estatística de 4 a 6 anos (média de 5 anos)

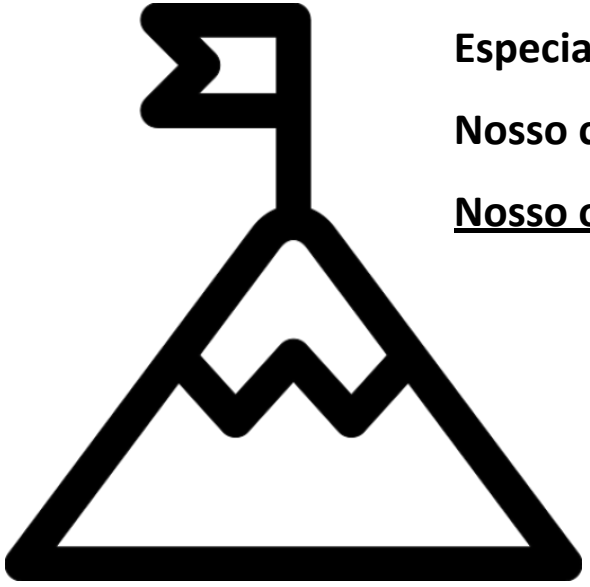
Especializações (em TI, Estatística ou Ciência de Dados) de 1 a 2 anos

Nosso curso → 32h

<u>Nosso objetivo neste tema?</u>



Objetivo



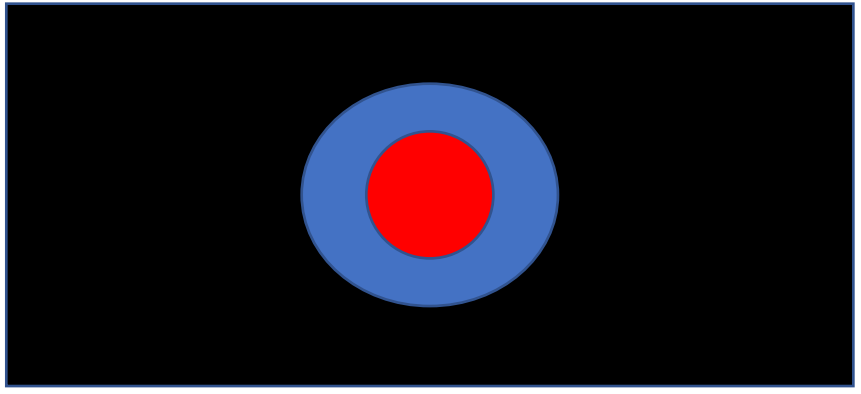
Graduação área de Tecnologia da Informação de 3 a 5 anos

Graduação em Estatística de 4 a 6 anos (média de 5 anos)

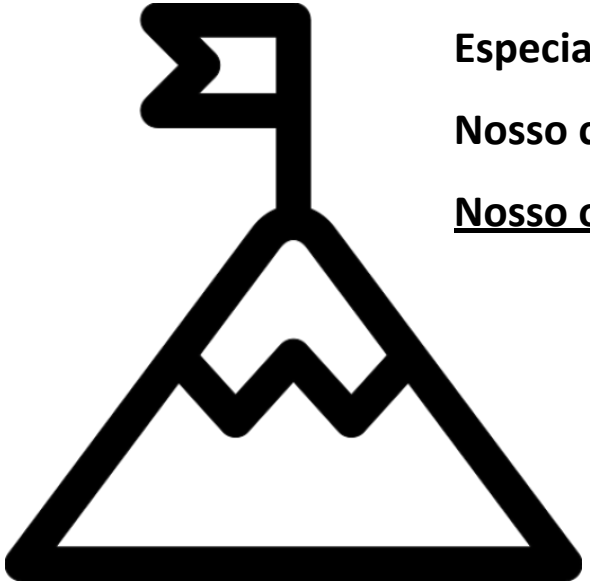
Especializações (em TI, Estatística ou Ciência de Dados) de 1 a 2 anos

Nosso curso → 32h

<u>Nosso objetivo neste tema?</u>



Objetivo



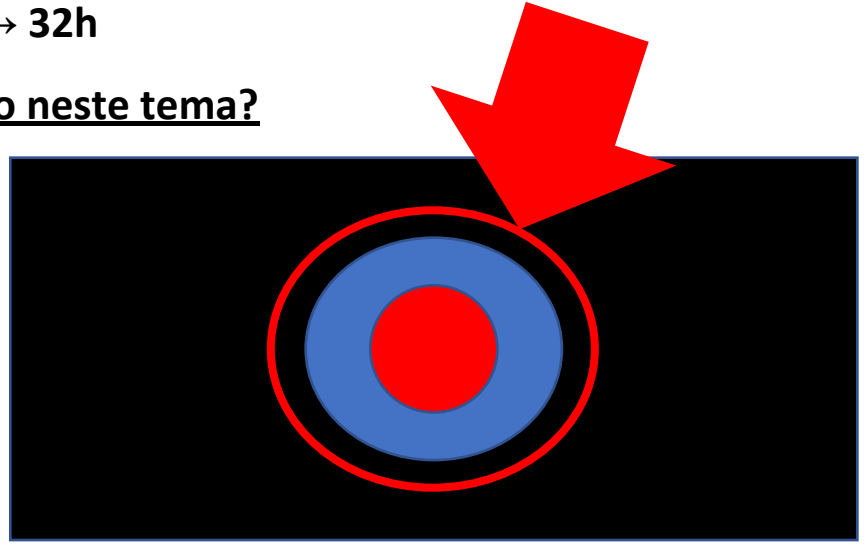
Graduação área de Tecnologia da Informação de 3 a 5 anos

Graduação em Estatística de 4 a 6 anos (média de 5 anos)

Especializações (em TI, Estatística ou Ciência de Dados) de 1 a 2 anos

Nosso curso → 32h

<u>Nosso objetivo neste tema?</u>

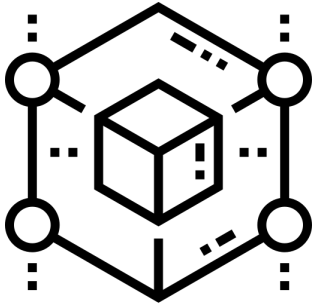


Material do Curso

Nosso repositório → https://github.com/datasciencealdeia/202003.git

Apresentações, Scripts, Dados Utilizados e etc, sempre atualizados

Virtual Machine



1) Download do Virtual Box (versão 6.0.4)

https://www.virtualbox.org/wiki/Downloads

2) Download da nossa VM (~7 Gb)

google drive

3) Executar o Virtual Box

Opção Arquivo > Importar Appliance > Escolher o arquivo baixado anteriormente para importar (DataScience.ova)

user: ds

passwd: ds2019Xpto

IMPORTANTE!!! A VM será a base para os exercícios após o almoço e IMPRESCINDÍVEL estar instalada e funcionando para o melhor aproveitamento

Pauta

1 – O que é Data Science

2 – Material Curso

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – SQL Básico

7 – Namorando Dados

O que são dados?

The Economist | Topics | Current edition | More | Subscribe

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Print edition | Leaders ›
May 6th 2017

A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A

FORTUNE

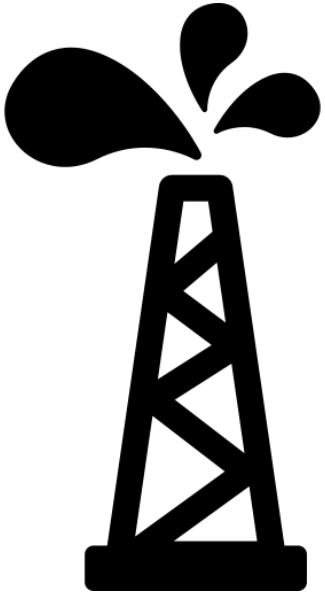
LEADERSHIP • ON LEADING

Intel CEO Says Data is the New Oil

By SUSIE GHARIB June 7, 2018

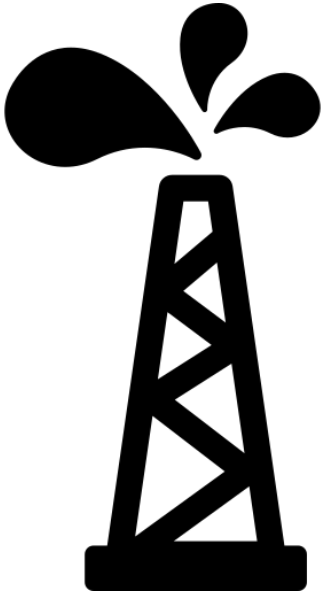
Brian Krzanich believes big data will dramatically change the world.

O que são dados?



O dado (no singular) é mesmo o novo petróleo?

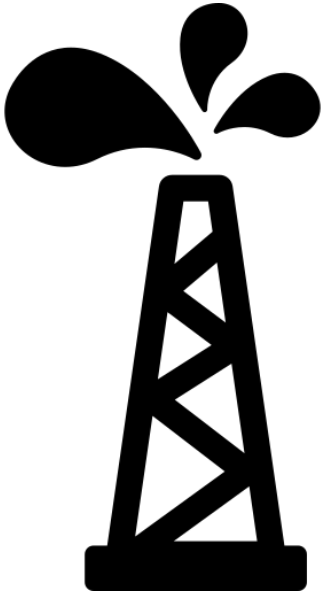
O que são dados?



O dado (no singular) é mesmo o novo petróleo?

O dado é a matéria-prima. Sem ser “refinado” não gera valor.

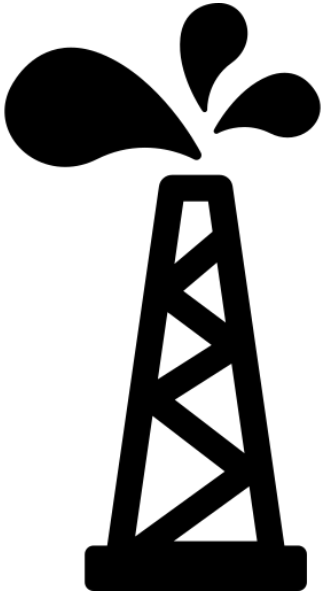
O que são dados?



Dados refinados, trabalhados geram **INFORMAÇÃO**

Em computação temos estudos referente a **teoria da informação**… ela vem estudando relatos de 30mil anos atrás, do homem primitivo, buscando se comunicar, expressar de forma a ser compreendido seus pensamentos internos. Todos estes “dados” utilizados por nós buscam, de forma encadeada, transmitir uma informação a respeito de um tema.

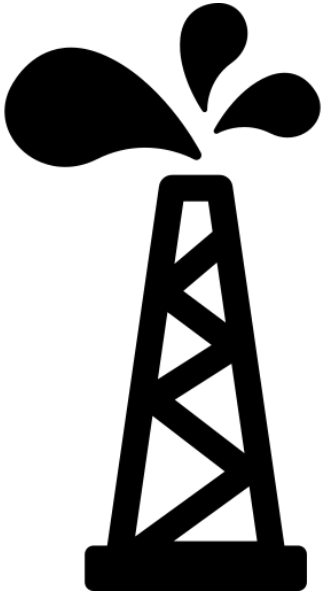
O que são dados?



Porém a parte nobre dos Dados não é apenas se tornar informação, mas sim um conjunto de informações que gere **CONHECIMENTO**

Dados → Informação → Conhecimento

O que são dados?



CONHECIMENTO É O NOVO ATIVO DE FATO

28/03/2019 - 15H51 - ATUALIZADA ÀS 15H51 - POR ÉPOCA NEGÓCIOS ONLINE

McDonald's investe US$ 300 milhões para adquirir startup de IA e big data

Primeira mudança após a compra da Dynamic Yield deverá ser vista nos drive thrus da rede de fast food

Guerra comercial custou bilhões de dólares aos EUA e à China em 2018, diz economista

Disputa atingiu setores como automobilístico, tecnologia e, acima de tudo, agricultura

Reuters

Indústria 4.0 pode economizar R$ 73 bilhões ao ano para o Brasil

Os ganhos de eficiência produtiva correspondem a uma economia de R$ 31 bilhões

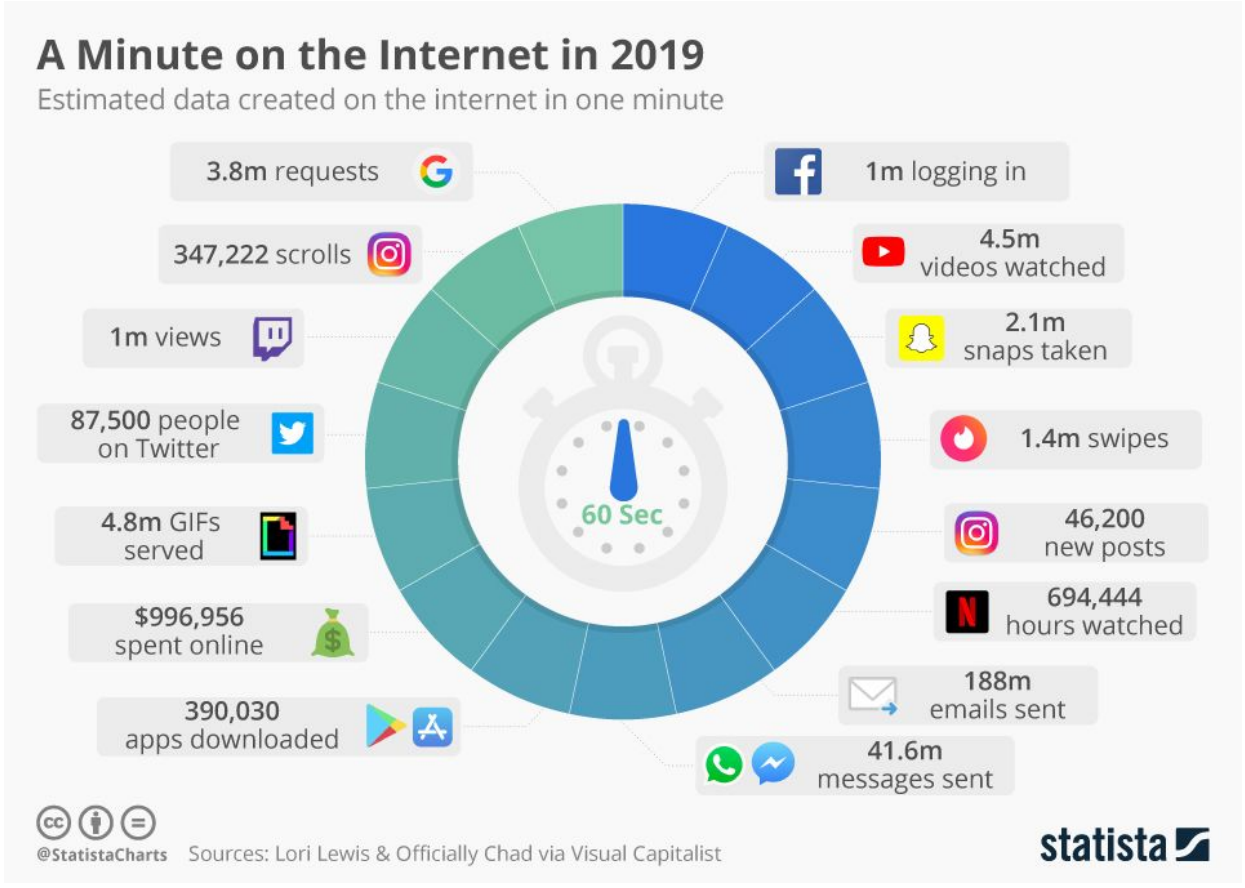
Fernando Rotta | 20/12/2017

Está chegando a era dos super-humanos. E eles são chineses

Tiago Cordeiro, especial para a Gazeta do Povo [22/03/2019] [11:09]

Qual a quantidade de dados são gerados por dia?

Qual a quantidade de dados são gerados por dia?



Gerindo os Dados → Cluster e Cloud

Computação em Nuvem é a distribuição de serviços de computação – servidores, armazenamento, bancos de dados, redes, software, análises, inteligência e muito mais pela Internet (“a nuvem”), proporcionando inovações mais rápidas, recursos flexíveis e economia na escala

O que é Data Science?

Ramo da Ciência especializada em:

- **Coleta**
- **Armazenamento**
- **Visualização**
- **Transformação**
- **Análise**
- **Modelagem de Dados**

Com foco principal na obtenção de subsídios para tomada de decisões!

Profissão, Carreira, Mercado Atual e Projeção

- ✔ **A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais**

Profissão, Carreira, Mercado Atual e Projeção

- ✔ **A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais**

- ✔ **Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO**

Profissão, Carreira, Mercado Atual e Projeção

- ✔ **A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais**

- ✔ **Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO**

- ✔ **Este profissional é o responsável por gerar conhecimento para tomada de decisões rápidas e precisas**

Profissão, Carreira, Mercado Atual e Projeção

- ✔ **A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais**
- ✔ **Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO**
- ✔ **Este profissional é o responsável por gerar conhecimento para tomada de decisões rápidas e precisas**
- ✔ **Inclusive, é responsável por automatizar as tomadas de decisões em tempo real (Aprendizado de Máquina & Inteligência Artificial)**

Profissão, Carreira, Mercado Atual e Projeção

- ✔ **A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais**

- ✔ **Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO**

- ✔ **Este profissional é o responsável por gerar conhecimento para tomada de decisões rápidas e precisas**

- ✔ **Inclusive, é responsável por automatizar as tomadas de decisões em tempo real (Aprendizado de Máquina & Inteligência Artificial)**

- ✔ **Logo, são bem remunerados:**
  https://www.glassdoor.com.br/Sal%C3%A1rios/data-scientist-sal%C3%A1rio-SRCH_KO0,14.htm

Profissão, Carreira, Mercado Atual e Projeção

- **O mercado brasileiro acordou para o valor desta profissão graças às <u>startups</u>**
- São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados
- Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real
- E as demais? Estão correndo atrás do **prejuízo!**

EXAME

Imposto de Renda Venezuela Previdência Concur

Por que o Nubank sempre busca cientistas de dados e paga até R$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por **Udacity**
30 jun 2018, 09h00

Profissão, Carreira, Mercado Atual e Projeção

- **O mercado brasileiro acordou para o valor desta profissão graças às <u>startups</u>**
- **São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados**
- **Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real**
- **E as demais? Estão correndo atrás do prejuízo!**

EXAME

Imposto de Renda Venezuela Previdência Concur

Por que o Nubank sempre busca cientistas de dados e paga até R$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por **Udacity**
30 jun 2018, 09h00

Profissão, Carreira, Mercado Atual e Projeção

- **O mercado brasileiro acordou para o valor desta profissão graças às <u>startups</u>**
- **São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados**
- **Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real**
- **E as demais? Estão correndo atrás do prejuízo!**

EXAME

Imposto de Renda Venezuela Previdência Concur

Por que o Nubank sempre busca cientistas de dados e paga até R$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por **Udacity**
30 jun 2018, 09h00

Profissão, Carreira, Mercado Atual e Projeção

- **O mercado brasileiro acordou para o valor desta profissão graças às startups**
- **São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados**
- **Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real**
- **E as demais? Estão correndo atrás do prejuízo!**

EXAME

Imposto de Renda Venezuela Previdência Concur

Por que o Nubank sempre busca cientistas de dados e paga até R$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por **Udacity**
30 jun 2018, 09h00

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

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

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

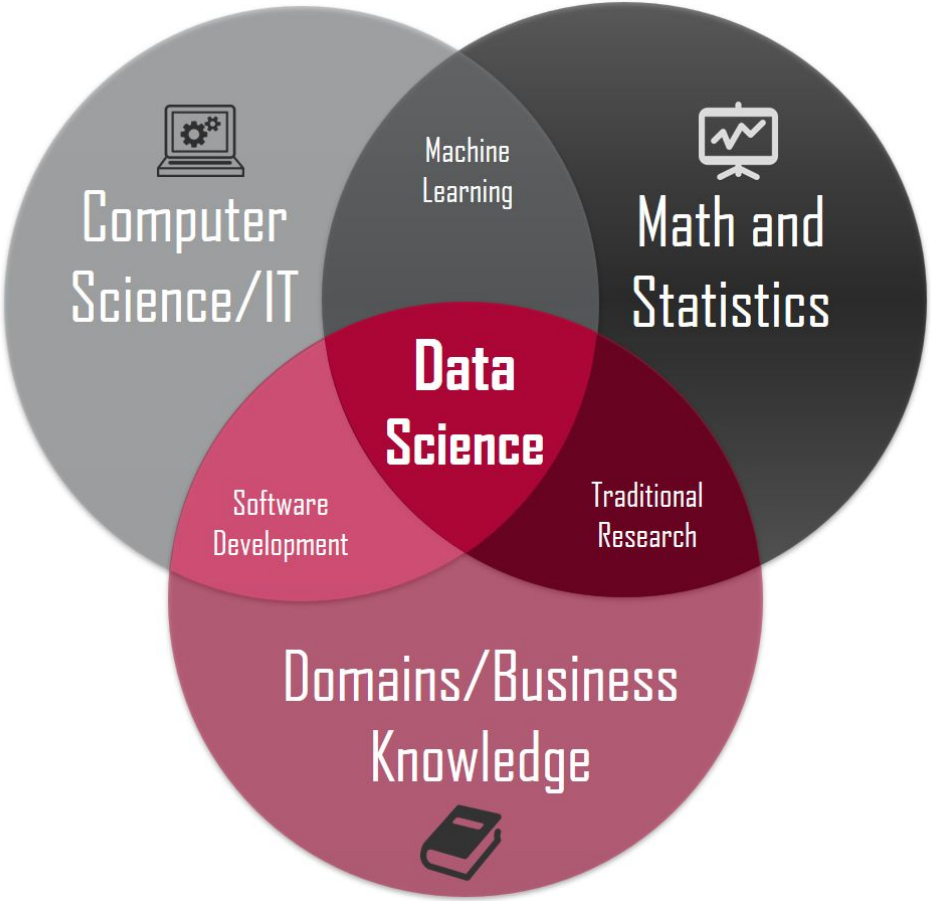
O Trabalho do Cientista de Dados

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

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**

Habilidades



Habilidades

Habilidades	Analista de Dados	Engenheiro de Machine Learning	Engenheiro de Dados	Cientista de Dados
Ferramentas de programação				
Visualização de Dados				
Conhecimento do Negócio				
Estatística				
Data Preparation				
Machine Learning				
Engenharia de Software				
Análise Multivariada				

 **Nice to have**

 **Importante**

 **Imprescindível**

Pauta

1 – O que é Data Science

2 – Material Curso

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – SQL Básico

7 – Namorando Dados

Material do Curso

Best Practice → Versionamento

Utilização de repositório (geralmente na cloud) para sincronizar e criar versões de arquivo.

É uma boa prática quando estamos gerando scripts / códigos fontes.

Soluções mais difundidas utilizadas:

SVN → https://subversion.apache.org/

GiT → https://git-scm.com/

Nosso repositório! → https://github.com/datasciencealdeia/202003.git

Apresentações, Scripts, Dados Utilizados e etc, sempre atualizados

Material do Curso

Principais Comandos GiT

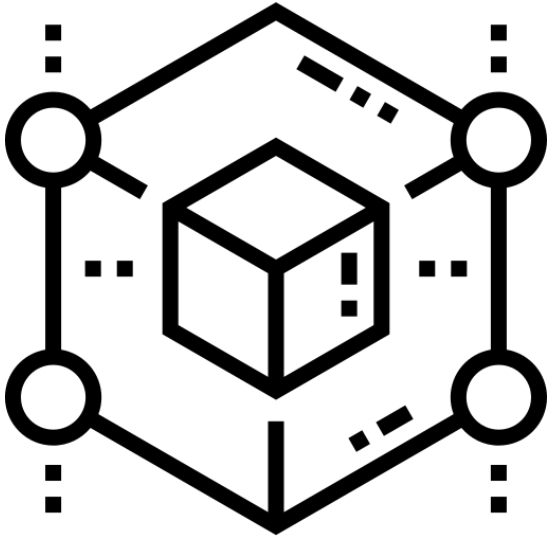
https://woliveiras.com.br/posts/comandos-mais-utilizados-no-git/

O que usaremos no curso:

git pull (baixar versão atualizada do repositório para vm)

*git status (*status do repositório local da vm)

Virtual Machine



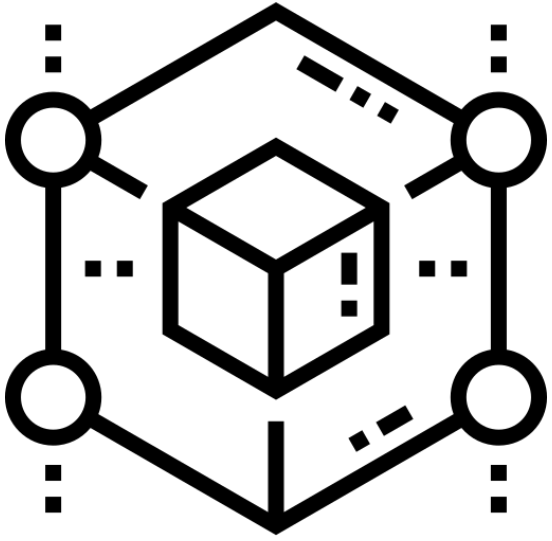
Conceito da Virtualização de Ciência da Computação

Software que simula um computador (máquina virtual) utilizando recursos do computador que está instalado (máquina host)

Pode ter configuração dimensionada com simples configuração: Memória RAM, HD, Placa de Rede e etc. (Sempre limitada a configuração física do computador host)

Utilizado na prática nas máquinas cluster/cloud

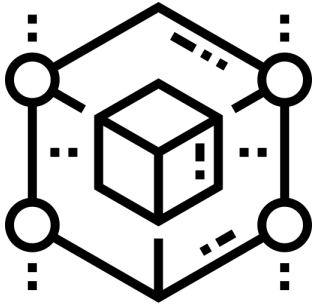
Virtual Machine



Configuração da Nossa Virtual Machine (VM)

- Sistema Operacional: Ubuntu 16.04.02 LTS 32 Bits
- 4 Gb de Memória RAM
- 50 Gb de HDD
- Softwares Embarcados:
  - ETL → Pentaho 5.0.1
  - Banco de Dados → PostgreSQL 10.7
  - Linguagens → R e Python 3

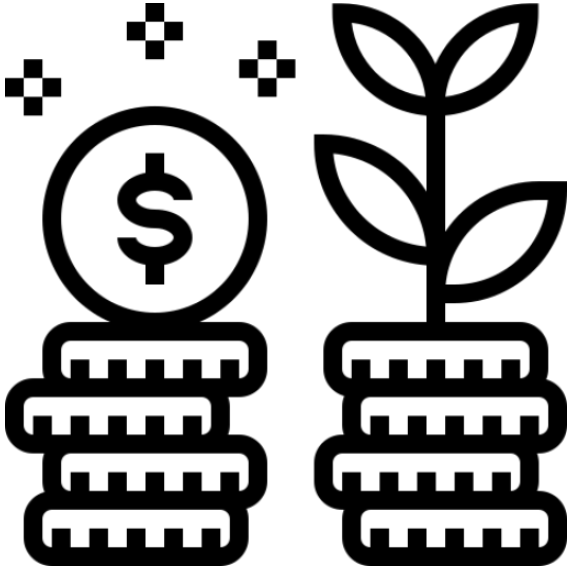
Virtual Machine



Dúvidas?

Momento Help Desk VM

Desafio Curso → AgroXP Brazil



Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodites*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.

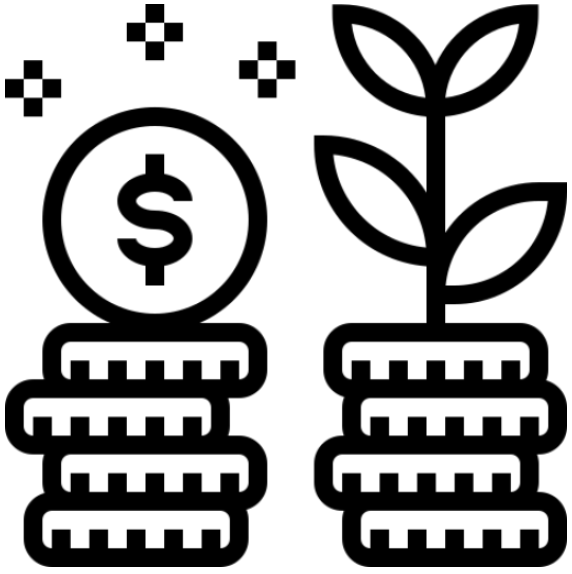
Você possui os seguintes dados:

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)

Desafio Curso → AgroXP Brazil



Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodites*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.

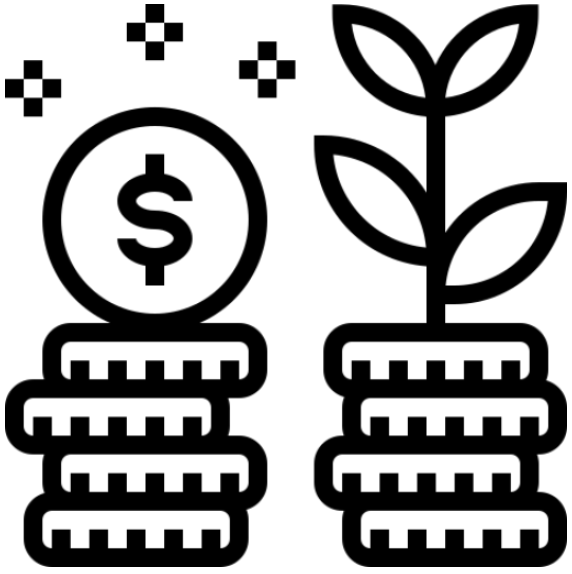
Você possui os seguintes dados:

1) Ministério de Desenvolvimento Indústria e Comércio --> apresenta os dados de TODOS commodities exportados no País desde 1997 até abril 2019(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)

Desafio Curso → AgroXP Brazil



Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodites*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.

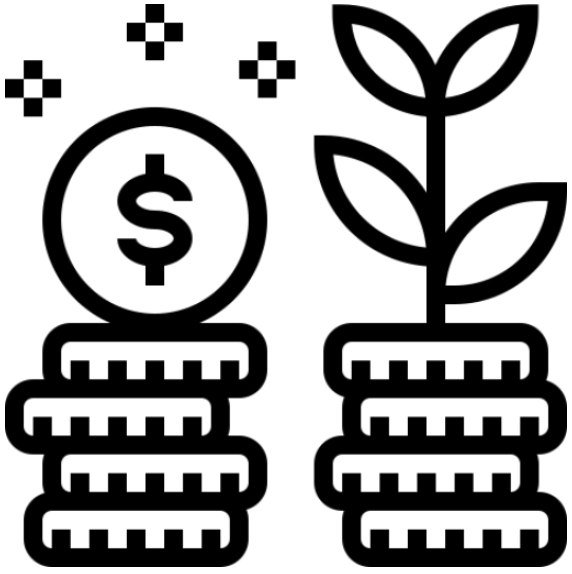
Você possui os seguintes dados:

1) Ministério de Desenvolvimento Indústria e Comércio --> apresenta os dados de TODOS commodities exportados no País desde 1997 até abril 2019 (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)

Desafio Curso → AgroXP Brazil



Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodites*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.

Você possui os seguintes dados:

1) Ministério de Desenvolvimento Indústria e Comércio --> apresenta os dados de TODOS commodities exportados no País desde 1997 até abril 2019 (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)

Pauta

1 – O que é Data Science

2 – Material Curso

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – SQL Básico

7 – Namorando Dados

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

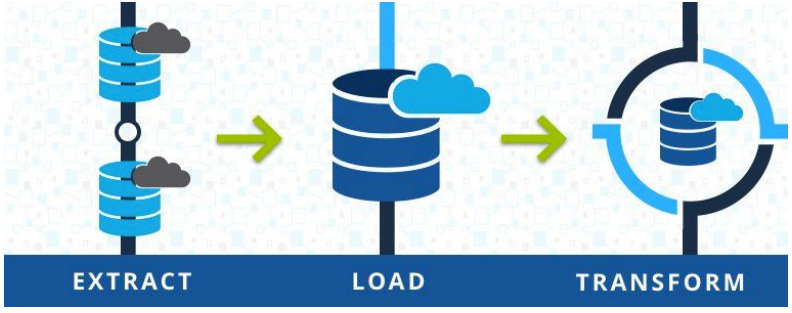
ETL – Extract, Transform,

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, 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 ferrramentas 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,

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,

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/YYY

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 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.

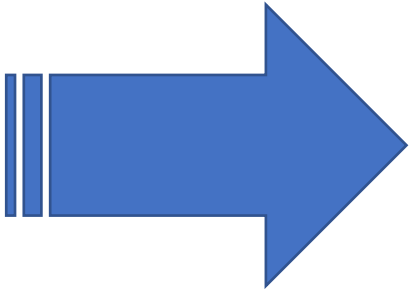
Pentaho Data Integration (Kettle) → Framework com soluções para fluxo de automação de forma produtiva, profissional e didática.

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

ETL – Extract, Transform,

Bora lá Praticar Galera???

ETL na Prática – Exercício



XML ⟶ Dados de Funcionários

CSV ⟶ Dados de Relógio Ponto

EXCEL ⟶ Cargos e Salários

Planilha Excel com:

Matrícula

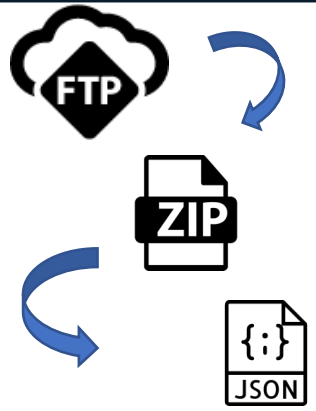
Nome Funcionário

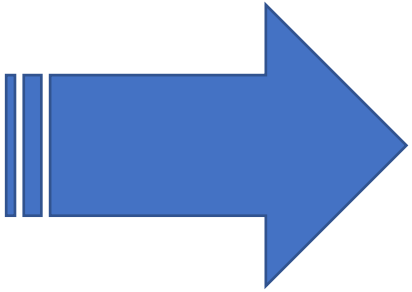
Cargo

Valor Hora

Dia e Hora Marcação Ponto

ETL na Prática – Homework





JSON Com remuneração variável por funcionário

ftp server: ftp.drivehq.com

User: datascienceandbigdata@gmail.com

Password: ds2019FTP

Diretório: GroupWrite

Arquivo: remunera.zip

Planilha Excel com:

Matrícula

Nome Funcionário

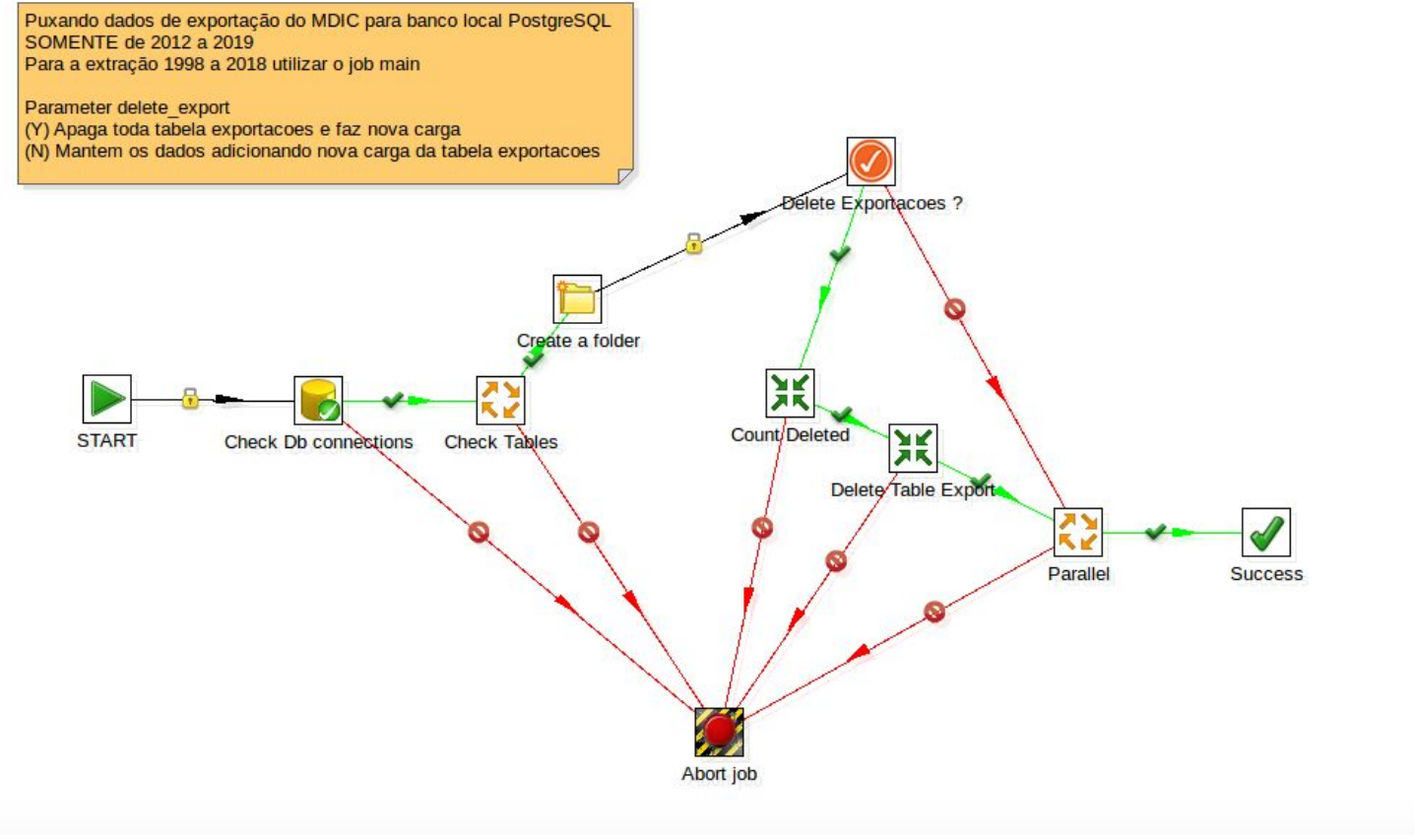
Cargo

Valor Hora

Último Dia e Hora Marcação Ponto

Total Remuneração Variável

ETL na Prática – Desafio AgroXP - Pentaho



Pauta

1 – O que é Data Science

2 – Material Curso

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – SQL Básico

7 – Namorando Dados

Estratégia – Modelo de Dados

O que é um Modelo de Dados?

Estratégia – Modelo de Dados

O que é um Modelo de Dados?

VAMOS MONTAR UM MODELO DE DADOS

Estratégia – Modelo de Dados

O que é um Modelo de Dados?

Estratégia – Modelo de Dados

O que é um Modelo de Dados?

Representação conceitual, lógica, que expressa os relacionamentos entre os dados em um banco de dados

Estratégia – Modelo de Dados

O que é um Modelo de Dados?

Representação conceitual, lógica, que expressa os relacionamentos entre os dados em um banco de dados

Banco de Dados pode representar qualquer organização para melhor relacionar os dados que precisamos armazenar a fim de descrever uma informação.

Estratégia – Modelo de Dados

IMPORTANTE: Para entender o modelo ou desenhar um modelo mais adequado aos seus dados deve-se perguntar: O que quero entender/representar com estes dados?

P.ex:

→Quais as versões de veículos "peruas"?

→Quais as versões da linha pesada que possuem GPS com piloto automático?

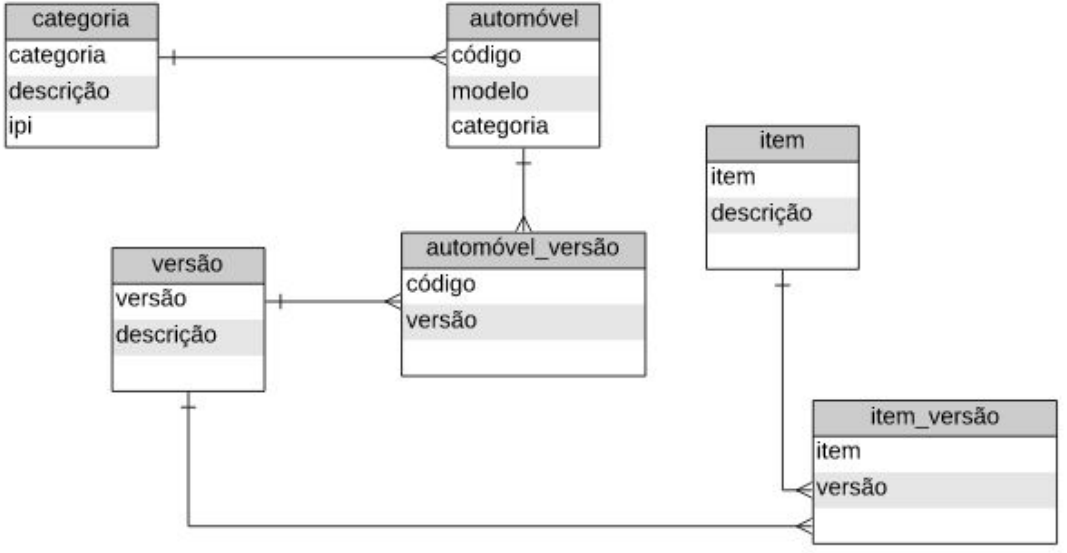
→Quantos automóveis da categoria Utilitário tem versões com ar condicionado de série?

Estratégia – Modelo de Dados

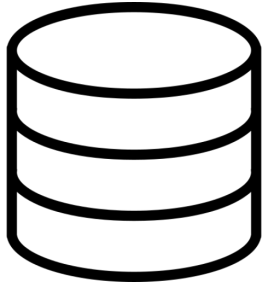
A representação gráfica do modelo se dá através do MER (Modelo de Entidades e Relacionamentos) ou DER (Diagrama de Entidades e Relacionamentos)

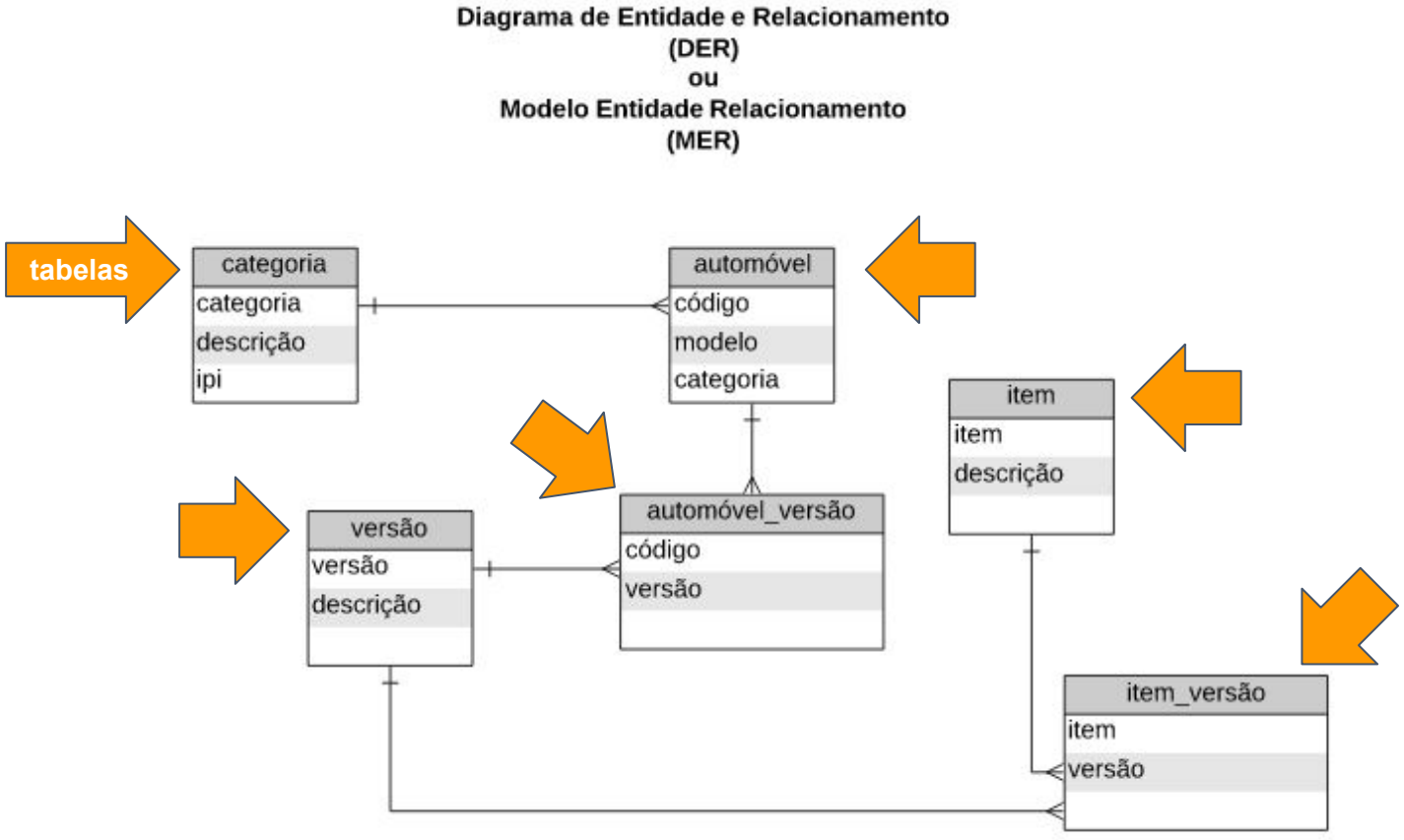
Estratégia – Modelo de Dados

Diagrama de Entidade e Relacionamento
(DER)
ou
Modelo Entidade Relacionamento
(MER)

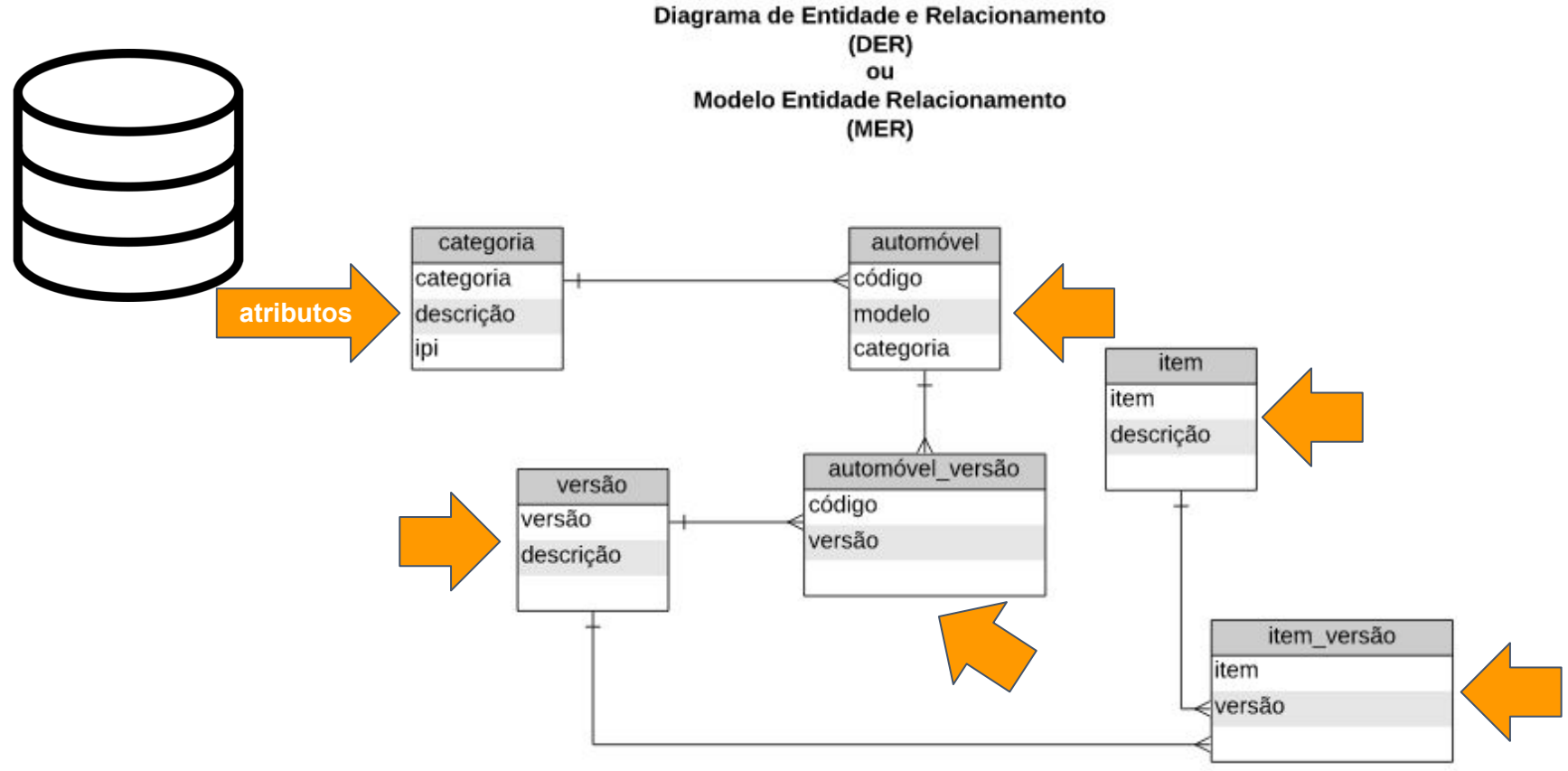


Estratégia – Modelo de Dados

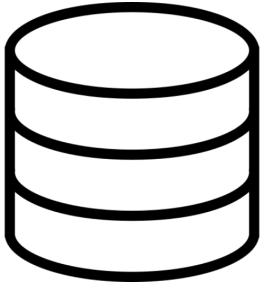


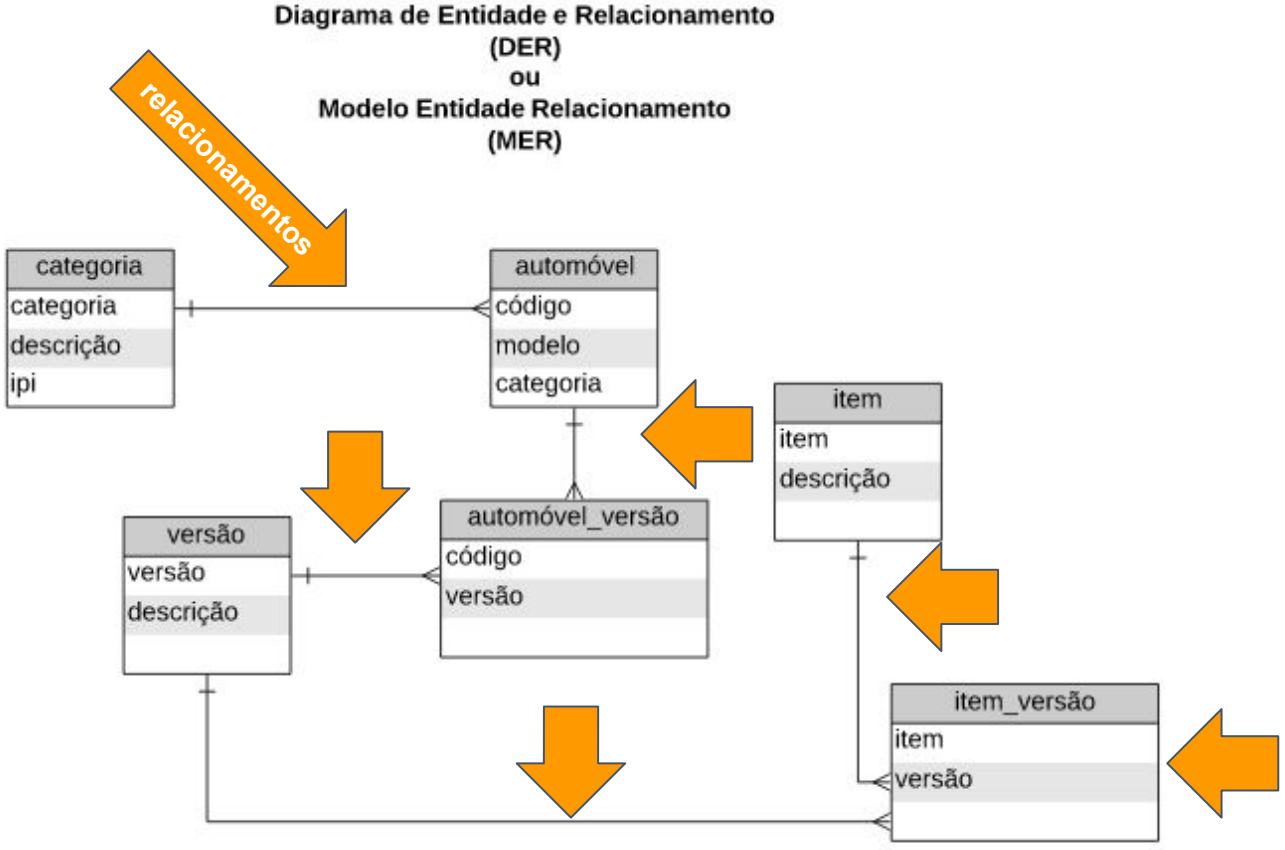


Estratégia – Modelo de Dados



Estratégia – Modelo de 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)

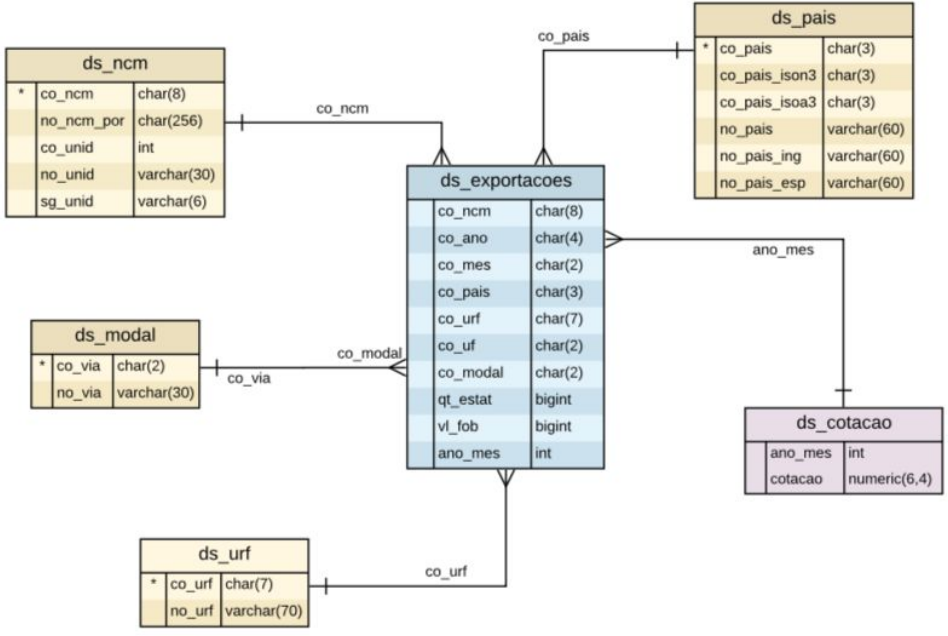
2) Tabelas auxiliares de nomenclatura de produtos com NCM – Nomenclatura Comum do Mercosul (formato .xls)

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:

Pauta

1 – O que é Data Science

2 – Material Curso

3 – Extract, Transform and Load

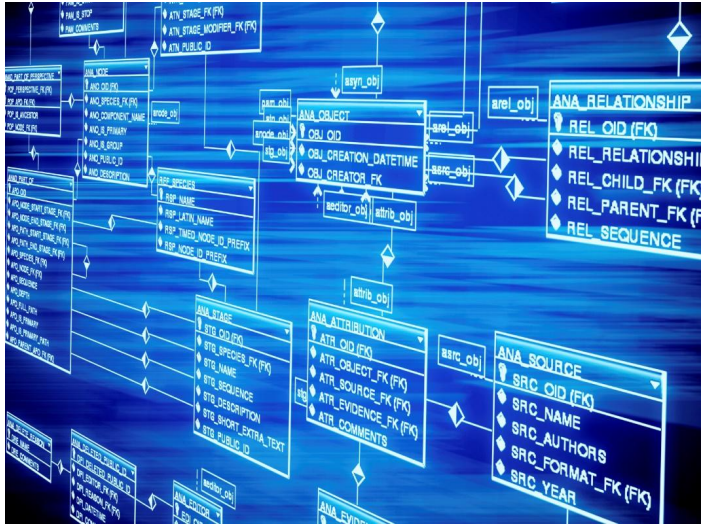
4 – Modelo de Dados

5 – Banco de Dados

6 – SQL Básico

7 – Namorando Dados

Banco de Dados

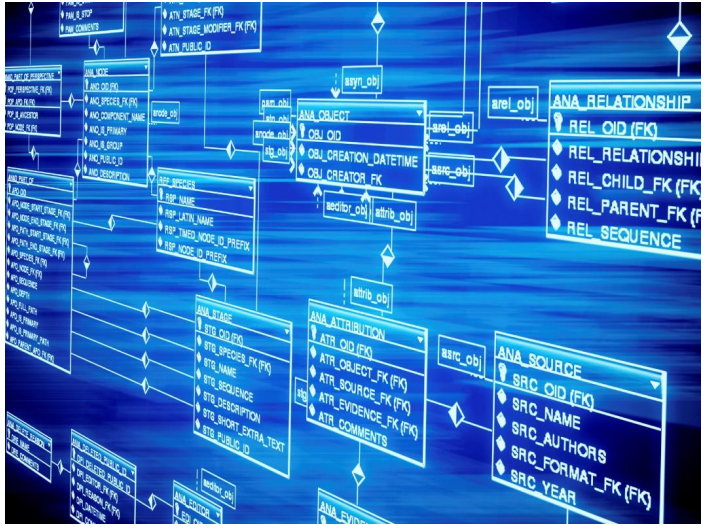


- **Bancos de dados** são conjuntos de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas

- 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

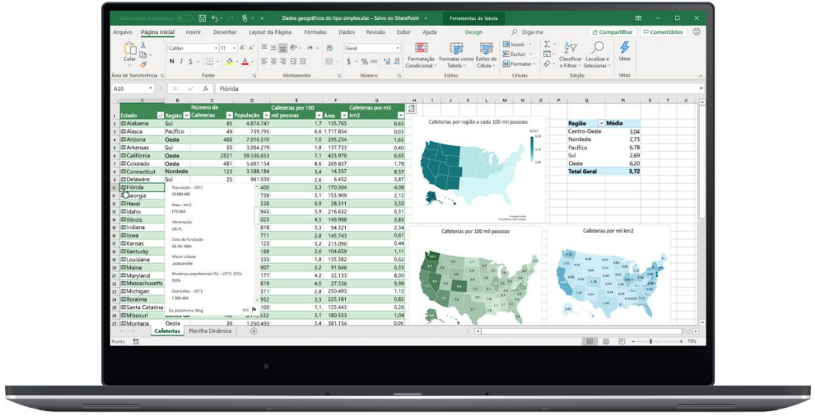


- São gerenciados por um SGDB (no nosso caso o PostgreSQL). Exemplos de Outros SGBDs:
  Relacionais (meados 1970) □ Oracle, SQL Server, MySQL, DB2, MonetDB.
  NoSQL (1998) □ MongoDB e Cassandra e etc.

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

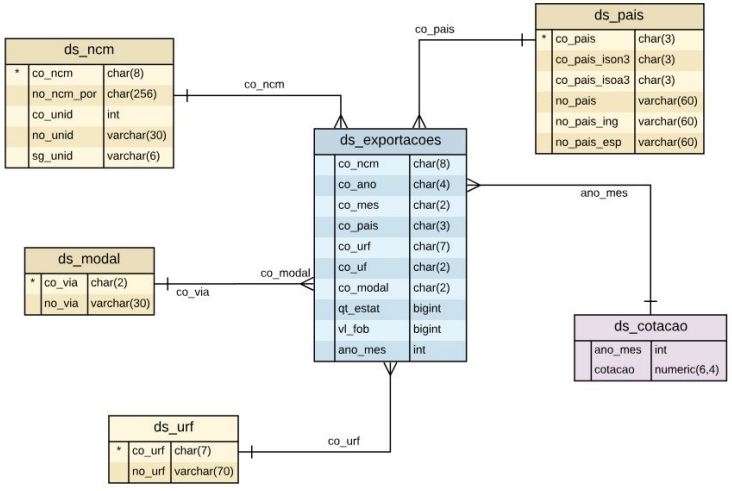
- **Banco de Dados Não Relacional (NoSQL)**
  - Orientado ao documento
  - Não garante Consistência/Integridade
  - Custo Menor Maior Escalabilidade (Gestão menos onerosa dos dados)

Banco de Dados - Relacionais



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

Banco de Dados - Relacionais



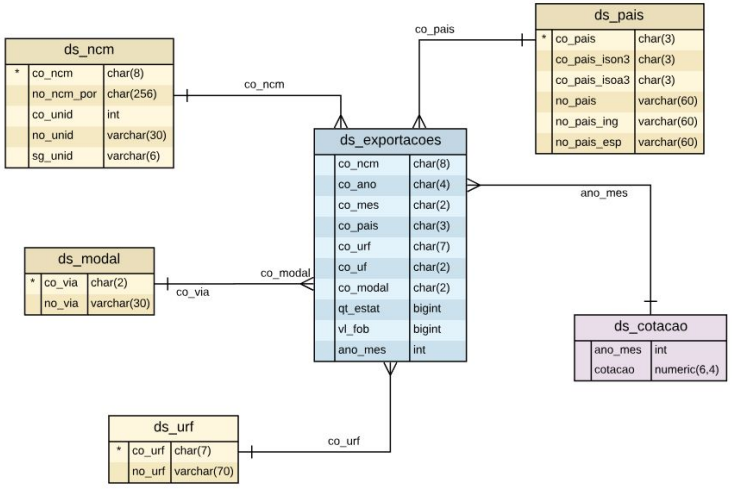
- **Relações Matriciais / Tabulares (Tabelas)**
- Todos os dados de um banco de dados relacional são armazenados em **tabelas**
- Uma tabela é uma simples estrutura de linhas e colunas
- Linha ⟶ Registro / Coluna ⟶ Atributo
- As tabelas associam-se entre si por meio de regras de relacionamentos, que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



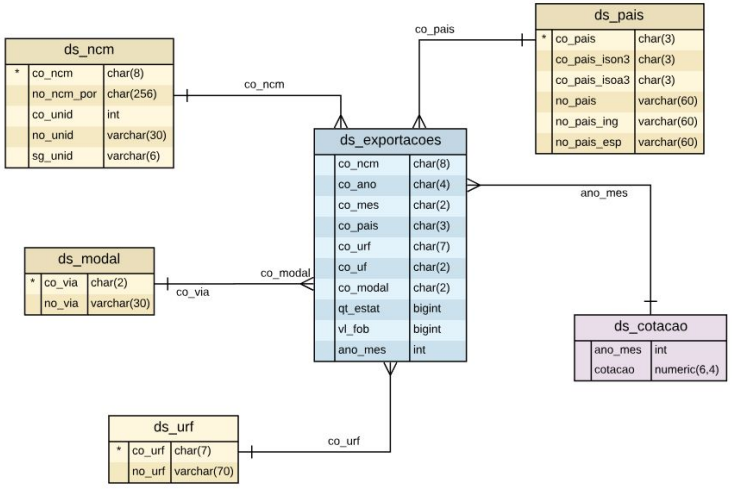
- **Chave**

- Integridade → Tupla/Registro/Linha única

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

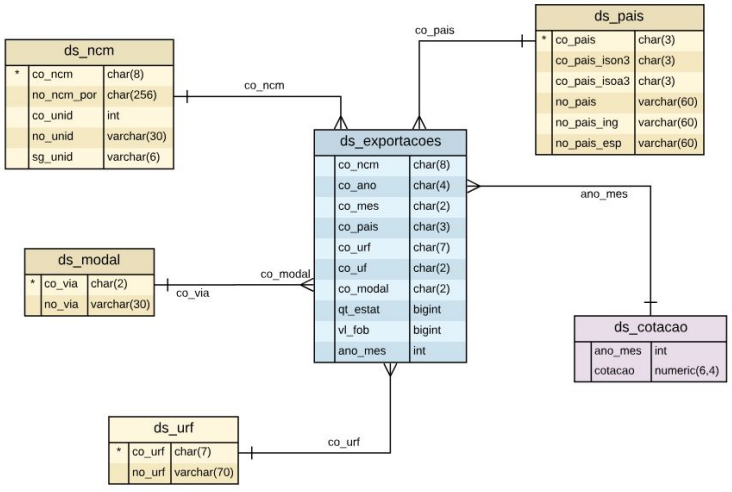
- Chave Estrangeira: (FK - Foreign Key) é a chave formada através de um relacionamento com a chave primária de outra tabela.
  → Define um relacionamento entre as tabelas e pode ocorrer repetidas vezes
  → Caso a chave primária seja composta na origem, a chave estrangeira também o será

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



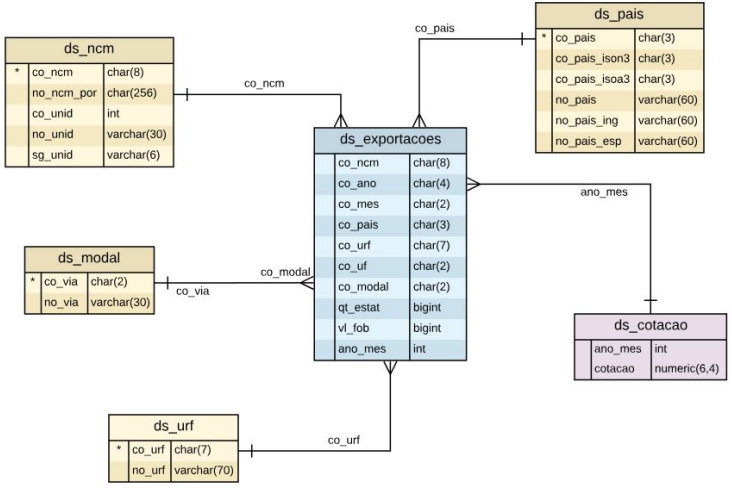
- SQL – Structure Query Language
- Linguagem declarativa implementada pelos SGBDs para consulta aos dados armazenados no banco
- ANSI padroniza a linguagem porém cada SGBD implementa alguma modificação na versão. Ex:
- Oracle:

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

- PostgreSQL:

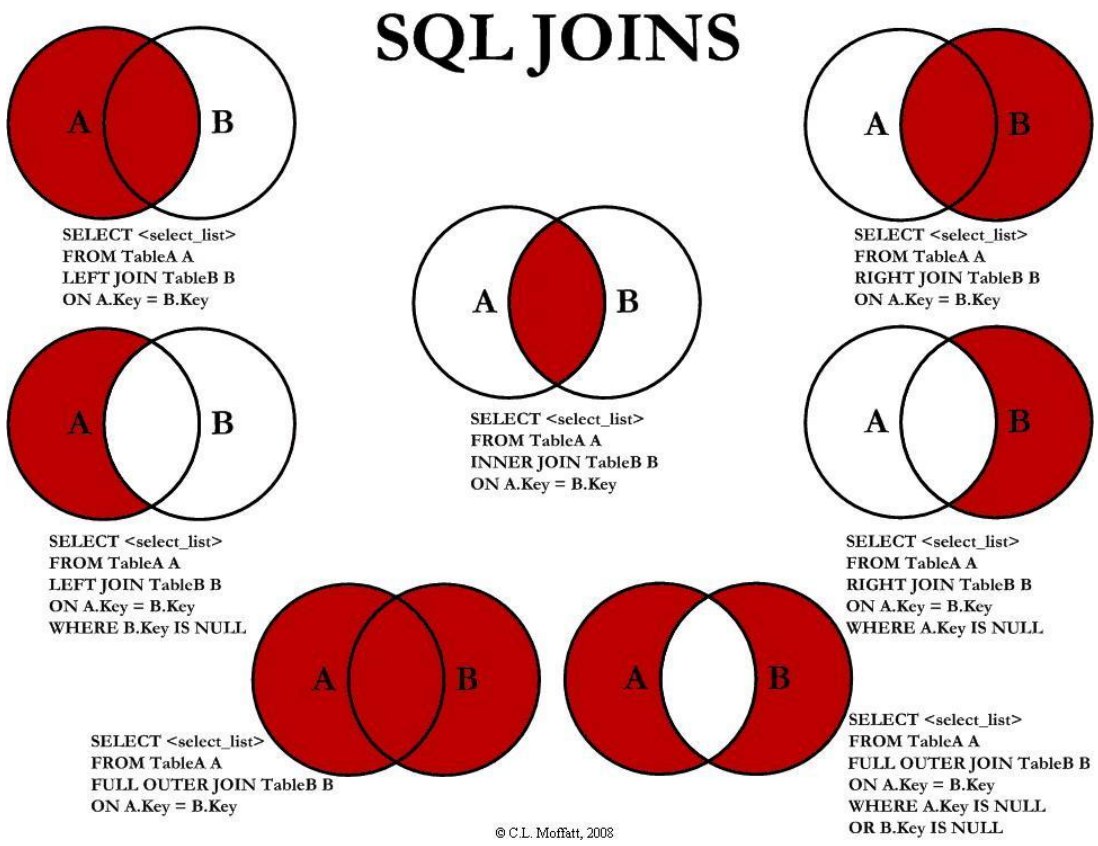
  SELECT CURRENT_TIME; --Somente hora
  SELECT CURRENT_DATE; --Somente a Data

Banco de Dados - SQL

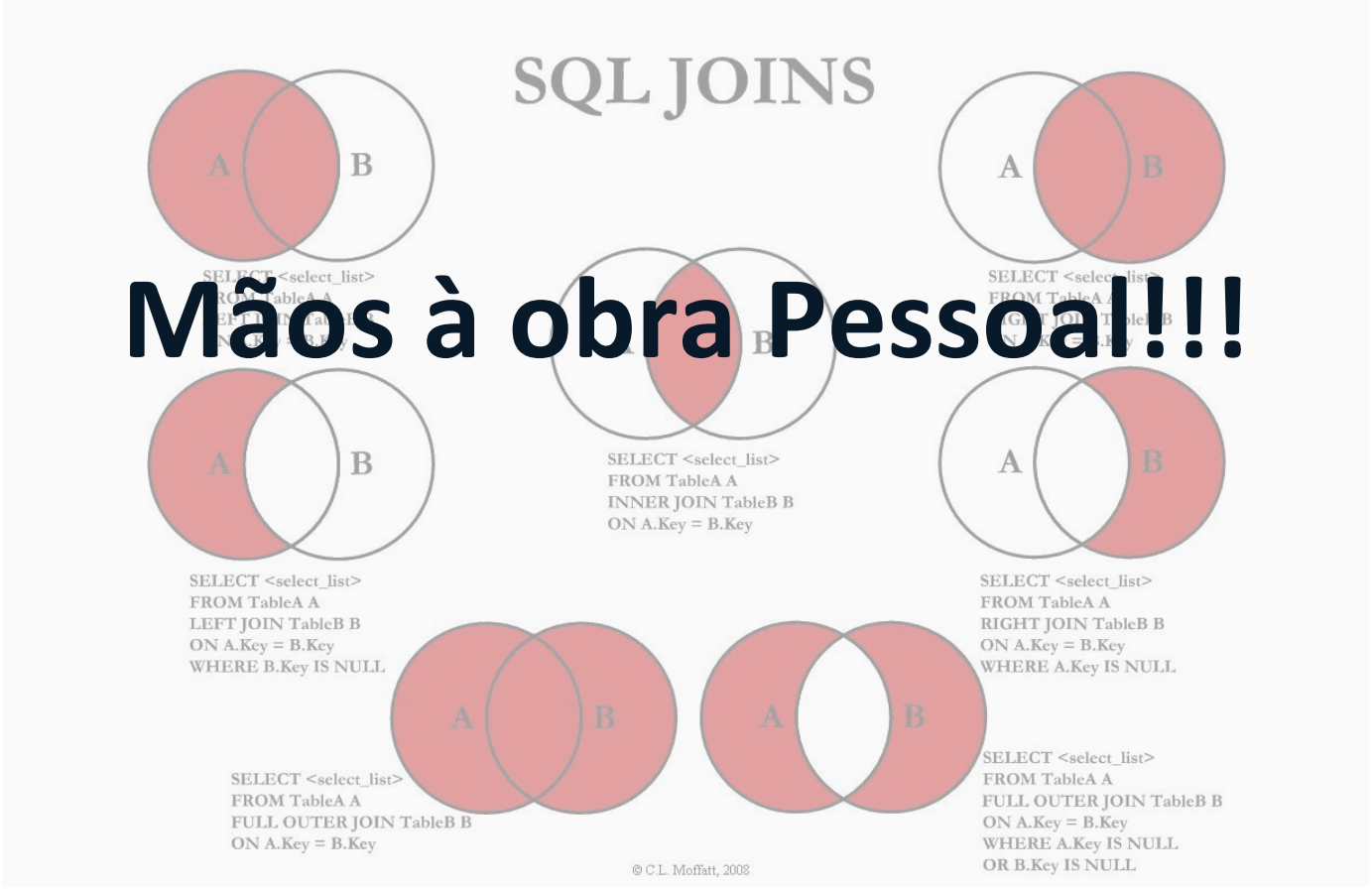


- SQL – Structure Query Language
- Subtipos da linguagem SQL (mais utilizados):
- **DDL:** Definição de Dados / Altera estrutura da tabela/entidade (Ex: CREATE TABLE)
- **DML:** Manipulação de Dados / Altera o conteúdo das colunas/atributos de tupla(s) (Ex: UPDATE)
- **DTL:** Transação de Dados (Ex: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

Queries Seleção



Queries Seleção



Mãos à obra Pessoal!!!

Pauta

1 – O que é Data Science

2 – Material Curso

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

6 – SQL Básico

7 – Namorando Dados

Pauta

1 – O que é Data Science

2 – Material Curso

3 – Extract, Transform and Load

4 – Modelo de Dados

5 – Banco de Dados

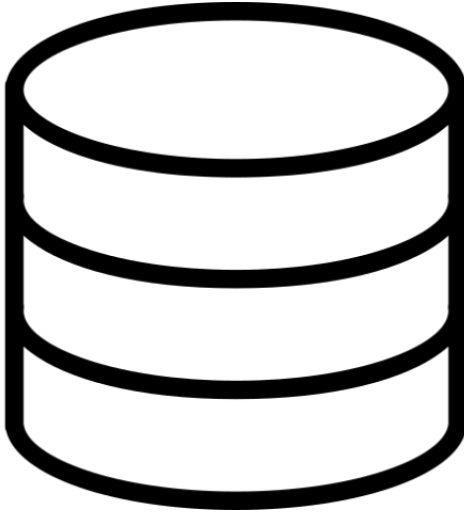
6 – SQL Básico

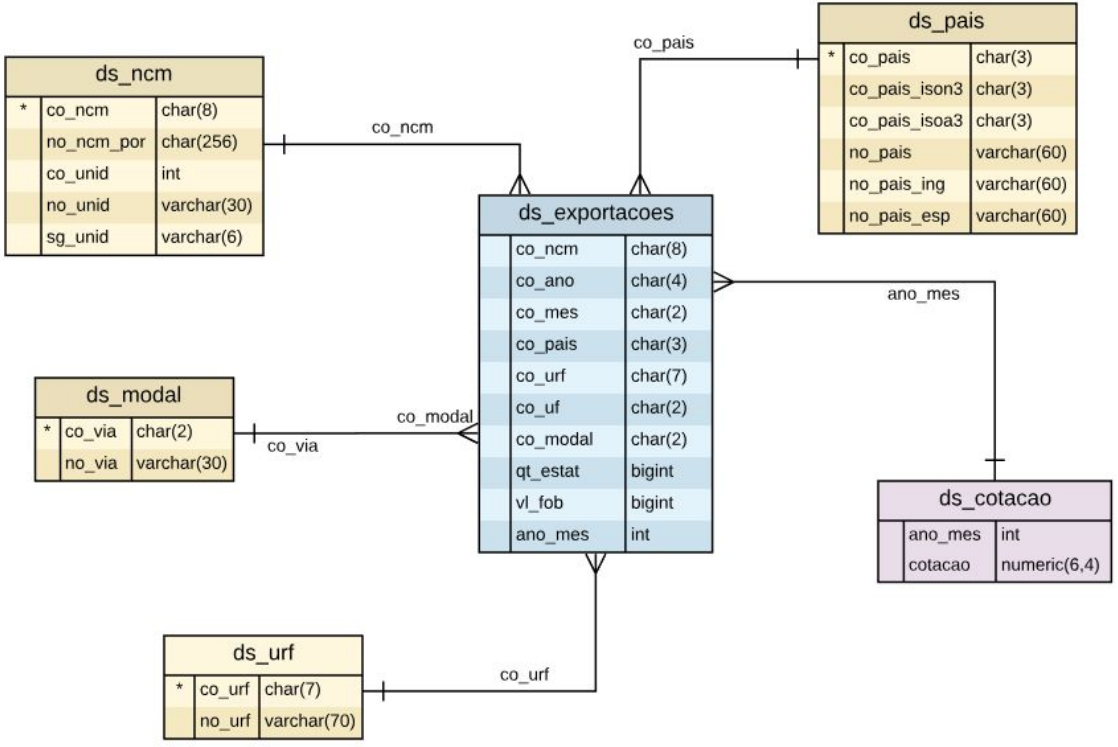
7 – Namorando Dados

O Trabalho do Cientista de Dados > Desafio Curso

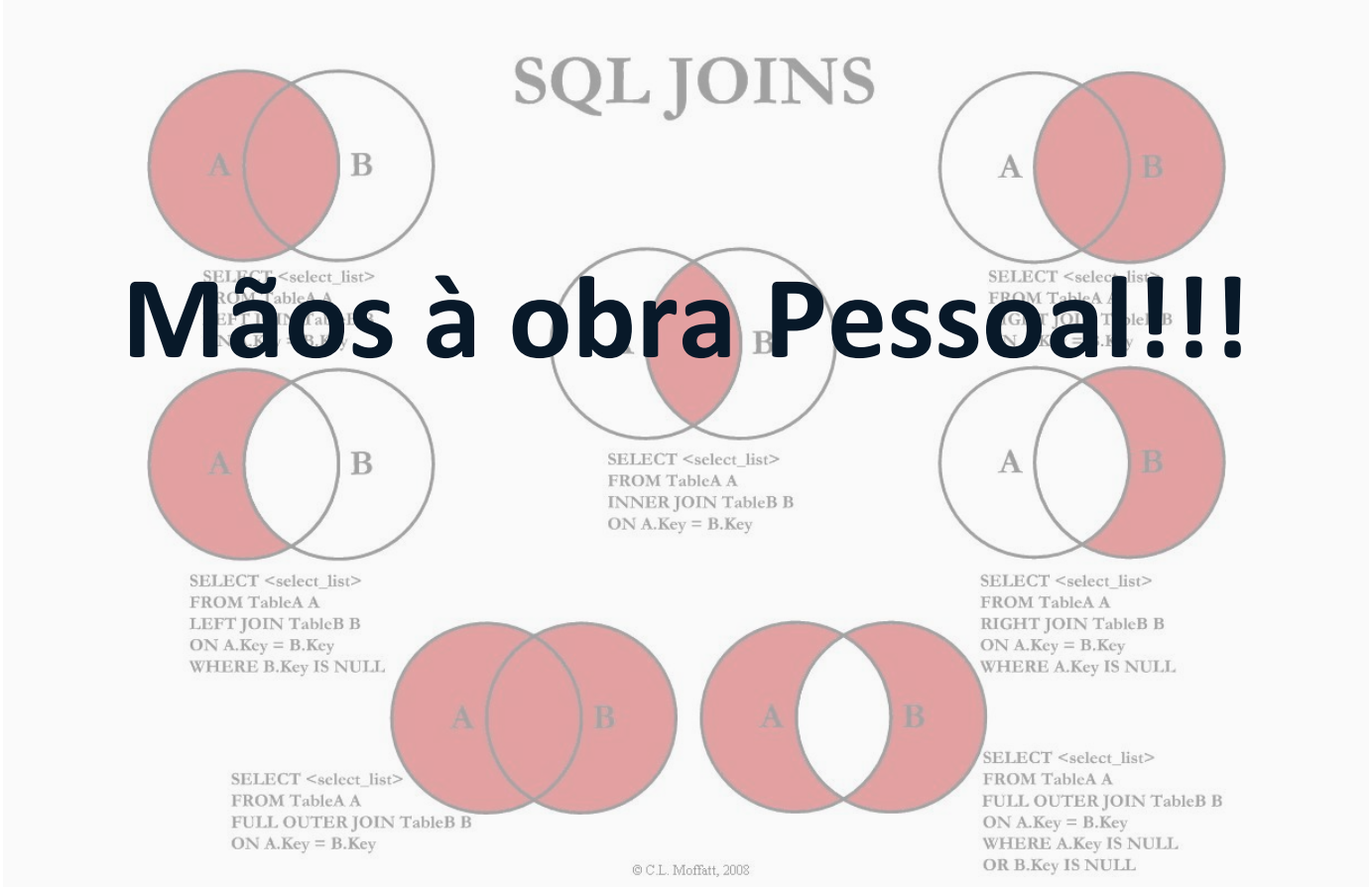
1. Definição do problema e levantamento de perguntas a serem respondidas ✓
2. Planejamento do processo de Data Science ✓
3. Coleta de dados ✓
4. Processamento e limpeza dos dados
5. Armazenamento dos dados ✓
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9.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

Desafio – Modelo de Dados



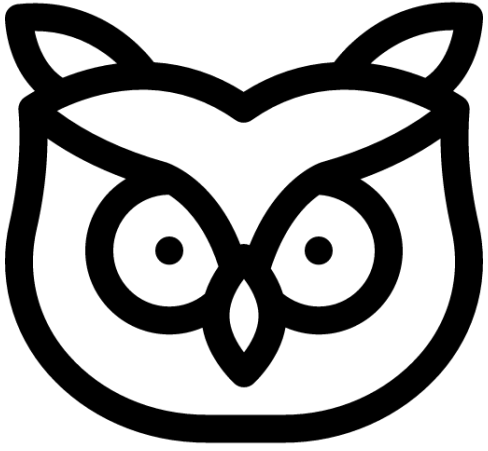


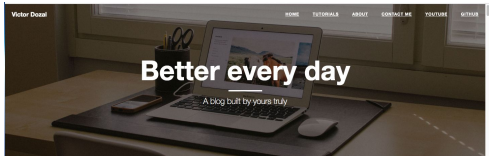
Namorando os Dados (Queries SQL)



Mãos à obra Pessoal!!!

Quero Saber Mais...





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