
Data Science
Charles Adriano dos Santos
Rafael Roberto Dias
ald
eia

Pauta

1 – Agenda

2 – Nos Episódios Anteriores...

3 – Virtual Machine

4 – Desafio - Kanban

5 – Desafio – Modelo de Dados

6 – Best Practice - Versionamento

7 – Desafio - ETL

8 – O Banco de Dados

9 – Desafio – Namorando Dados (SQL)

Pauta

1 – Agenda

2 – Nos Episódios Anteriores...

3 – Virtual Machine

4 – Desafio - Kanban

5 – Desafio – Modelo de Dados

6 – Best Practice - Versionamento

7 – Desafio - ETL

8 – O Banco de Dados

9 – Desafio – Namorando Dados (SQL)

Manhã

Horário	Assunto
09:30	Nos Episódios Anteriores... (revisão)
10:00	Virtual Machine
10:30	Desafio – Montando o Kanban
11:00	Desafio – Modelo de Dados
12:00	Best Practice – Versionamento
12:30	Almoço

Tarde

Horário	Assunto
13:30	Desafio - ETL
15:30	O Banco de Dados
17:30	Desafio – Namorando os Dados (SQL)
18:30	Encerramento

Nos Episódios Anteriores...

Ciência de Dados

- O conceito
- Os desafios e etapas
- A profissão
- Carreira
- Habilidades

Estatística

- Conceitos
- Média, Moda, Mediana, Variância, Desvio Padrão
- Modelos
- Séries Temporais, Regressão Logística

Ciência da Computação

- Conceitos
- Dado, Sistemas Binários, Lógica Booleana
- Conjuntos e Matrizes
- Banco de Dados, Algoritmo, Cluster/Cloud, Machine Learning

Nos Episódios Anteriores...

Ciência de Dados

- O conceito
- Os desafios e etapas
- A profissão
- Carreira
- Habilidades

Estatística

- Conceitos
- Média, Moda, Mediana, Variância, Desvio Padrão
- Modelos
- Séries Temporais, Regressão Logística

Ciência da Computação

- Conceitos
- Dado, Sistemas Binários, Lógica Booleana
- Conjuntos e Matrizes
- Banco de Dados, Algoritmo, Cluster/Cloud, Machine Learning

Nos Episódios Anteriores...

Ciência de Dados

- O conceito
- Os desafios e etapas
- A profissão
- Carreira
- Habilidades

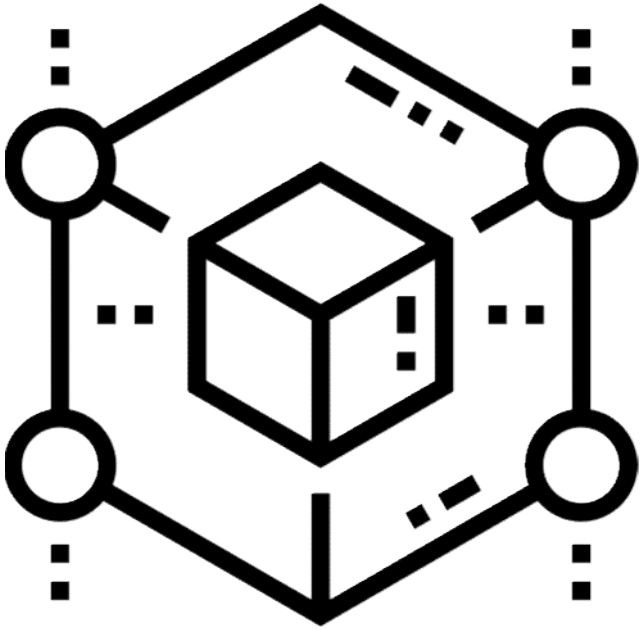
Estatística

- Conceitos
- Média, Moda, Mediana, Variância, Desvio Padrão
- Modelos
- Séries Temporais, Regressão Logística

Ciência da Computação

- Conceitos
- Dado, Sistemas Binários, Lógica Booleana
- Conjuntos e Matrizes
- Banco de Dados, Algoritmo, Cluster/Cloud, Machine Learning

Virtual Machine



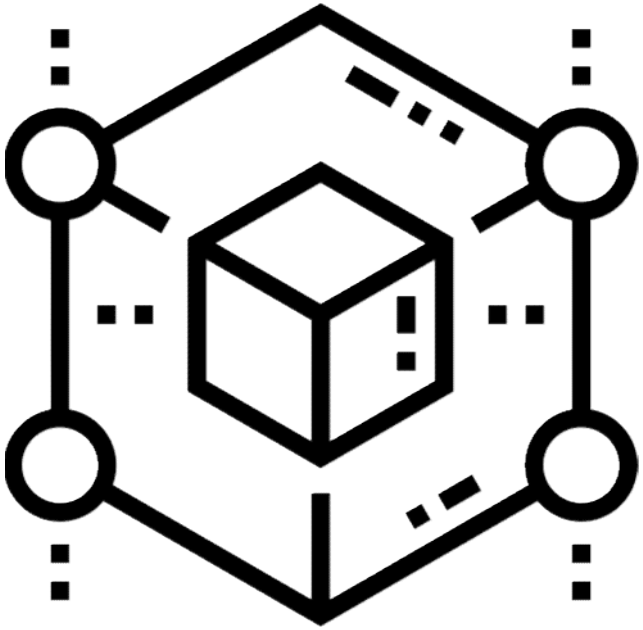
Conceito da Virtualização de Ciência da Computação

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

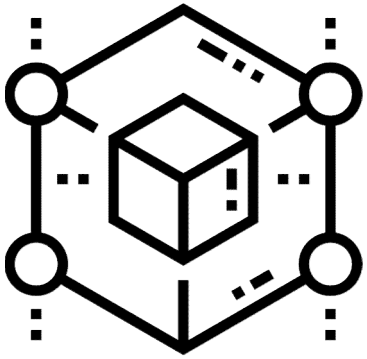
Virtual Machine



Configuração da Nossa Virtual Machine (VM)

- Sistema Operacional: Ubuntu 16.04.02 LTS 32 Bits
- 1 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
  - Versionamento → GitEye 2.2

Virtual Machine



http://dontpad.com/datasciencealdeia/toolkit

1) Download do Virtual Box (versão 6.0.4)

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

2) Download da nossa VM

https://drive.google.com/open?id=1OVr0BVaok5cfAKJ8XinzFY0dF-W_RpQK

3) Executar o Virtual Box

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

IMPORTANTE → Em Mac Address Policy escolher Generate new MAC... > Importar

user: ds

passwd: ds2019Xpto

Desafio – Montando o Kanban

Relembram do Desafio?

*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.*

Desafio – Montando o Kanban

Relembram do Desafio?

*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.*

Vamos ao planejamento!

Que tal usarmos um Kanban?

Desafio – Montando o Kanban

Kanban

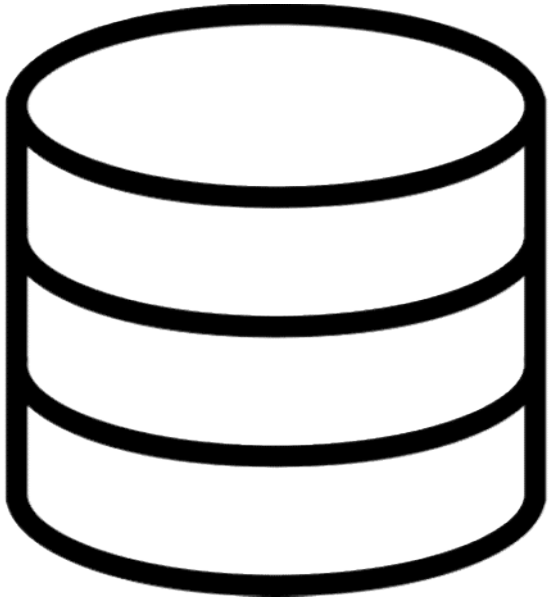
Ferramenta para sinalização de fluxos de produção, utilizada inicialmente no âmbito fabril porém hoje difundida em várias áreas.

Muito utilizado em gerenciamento de projeto em metodologia ágil (Agile!)

Nosso Kanban

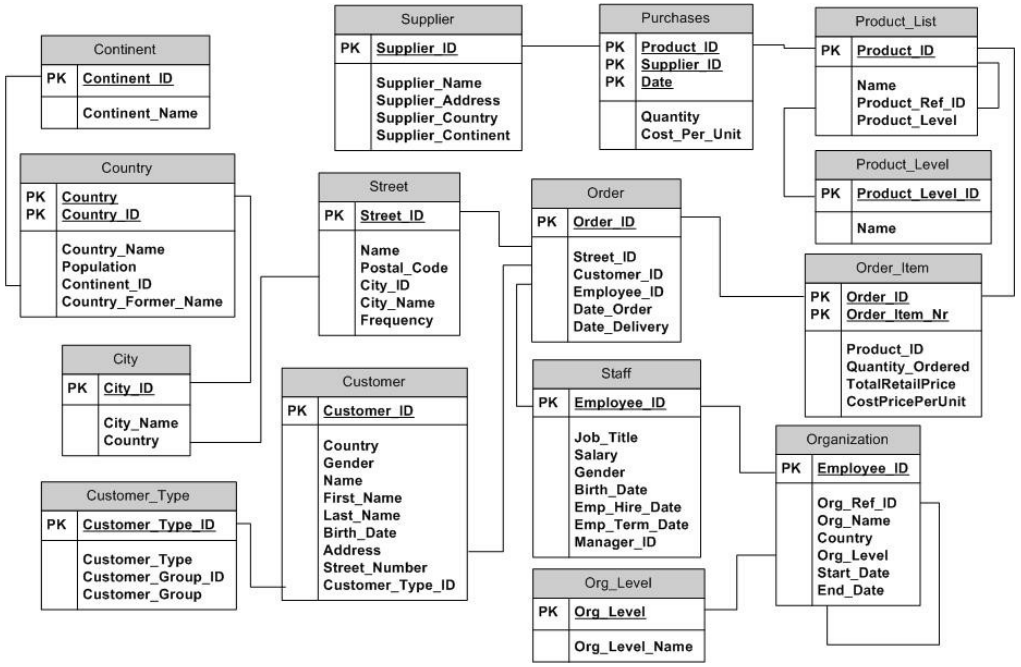
https://scrumy.com/culprits85mythic

Desafio – Modelo de Dados

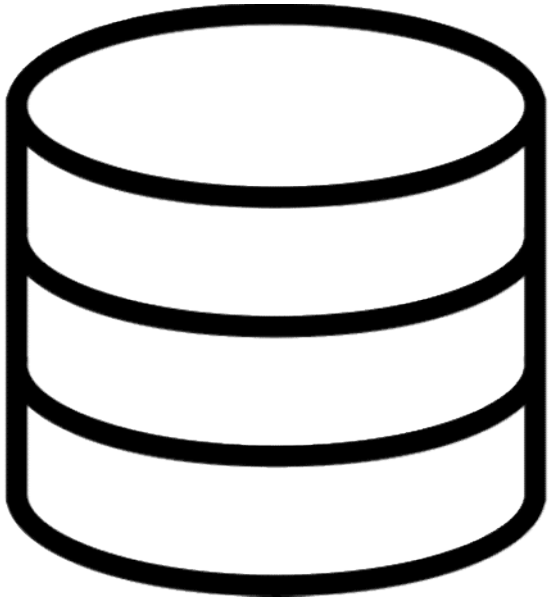


O que é um Modelo de Dados?

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

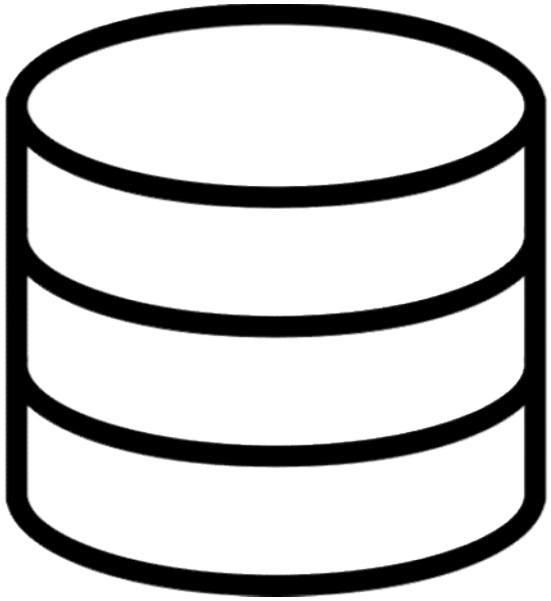


Desafio – Modelo de Dados



E o nosso modelo do desafio Agro XP Brazil?

Desafio – 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)

4) Base de contratos com rentabilidade obtida pela empresa em cada produto negociado nos últimos 6 anos

Vamos entender e montar um modelo!

Best Practice - Versionamento

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/

Vamos criar nosso repositório! → https://github.com/

ELT – Extract, Transform, Load

ETL, do inglês **Extract Transform Load** (*Extrair Transformar Carregar*), são ferramentas de software cuja função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e por fim o carregamento dos dados geralmente para um Data Mart e/ou Data Warehouse:

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

ELT – Extract, Transform, Load

Extração é a primeira parte do processo de ETL é a extração de dados dos sistemas de origem. A maioria dos projetos de data warehouse consolidam dados extraídos de diferentes sistemas de origem. Cada sistema pode também utilizar um formato ou organização de dados diferente.

ELT – Extract, Transform, Load

O estágio de **transformação** aplica uma série de regras ou funções aos dados extraídos para derivar os dados a serem carregados. Algumas fontes de dados necessitarão de muito pouca manipulação de dados. Em outros casos, podem ser necessários um ou mais de um dos seguintes tipos de transformação:

ELT – Extract, Transform, Load

- Seleção de colunas para carregar
- Tradução de valores codificados (se o sistema de origem armazena 1 para sexo masculino e 2 para feminino, mas o data warehouse armazena M para masculino e F para feminino)
- Codificação de valores de forma livre (mapeando “Masculino”,“1” e “Sr.” para M)
- Derivação de um novo valor calculado (flag 0,1)

ELT – Extract, Transform, Load

- Junção de dados provenientes de diversas fontes
- Resumo de várias linhas de dados (total de vendas para cada loja e para cada região, por exemplo)
- Geração de valores de chaves substitutas ou Transposição ou rotação
- Quebra de uma coluna em diversas colunas

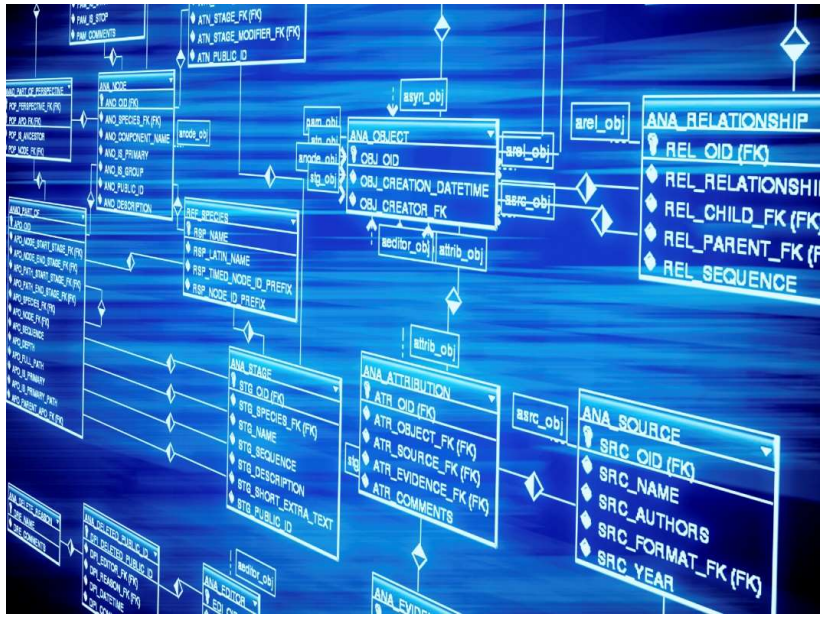
ELT – Extract, Transform, Load

A fase de **carregamento** consiste na colocação dos dados no Data Warehouse (DW)

Dependendo das necessidades da organização, este processo varia amplamente

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

Banco de Dados

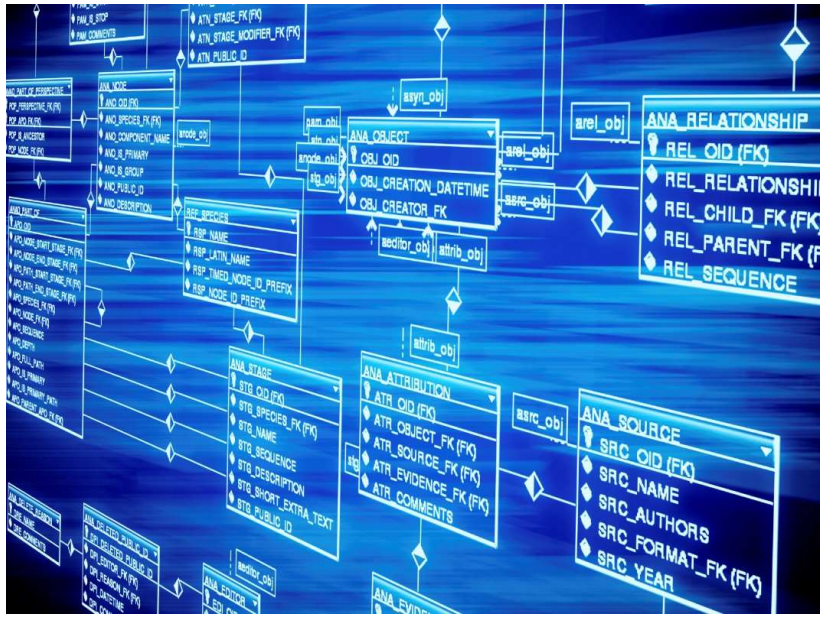


- **Bancos de dados** são conjuntos de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas

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

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

Banco de Dados



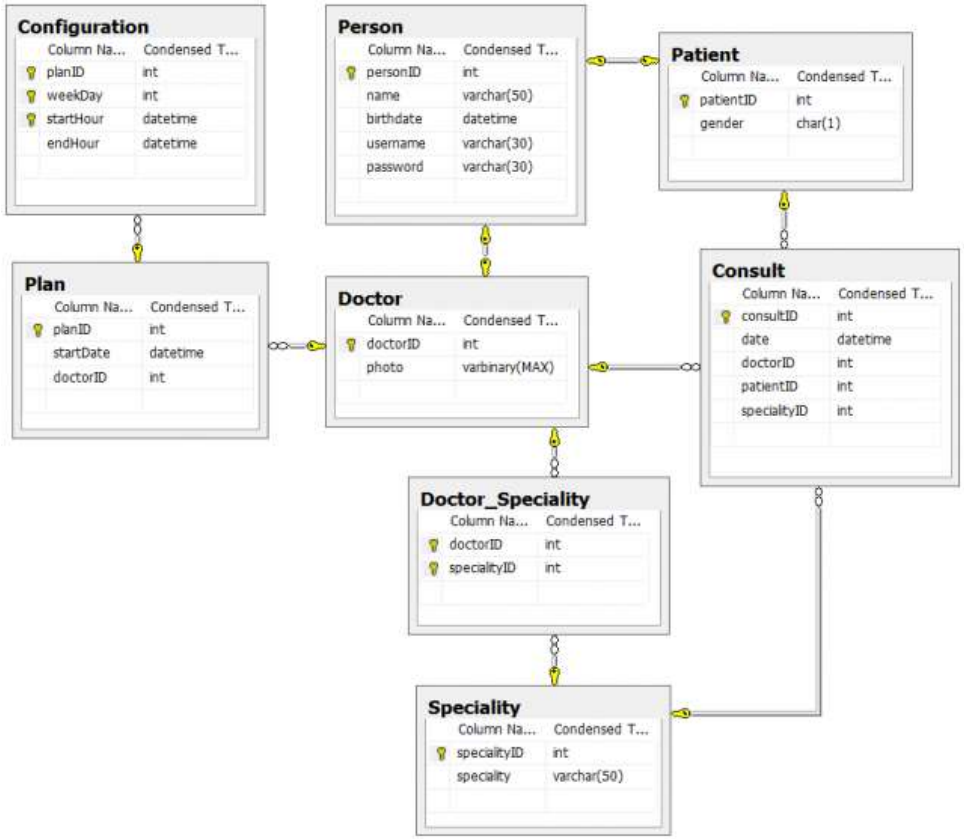
- **Banco de Dados Relacional (Modelo Relacional)**

- Foram desenvolvidos para prover acesso facilitado aos dados, possibilitando que os usuários utilizassem uma grande variedade de abordagens no tratamento das informações

- Nos Bancos de Dados Relacionais os usuários podem fazer perguntas relacionadas aos negócios por meio de vários pontos

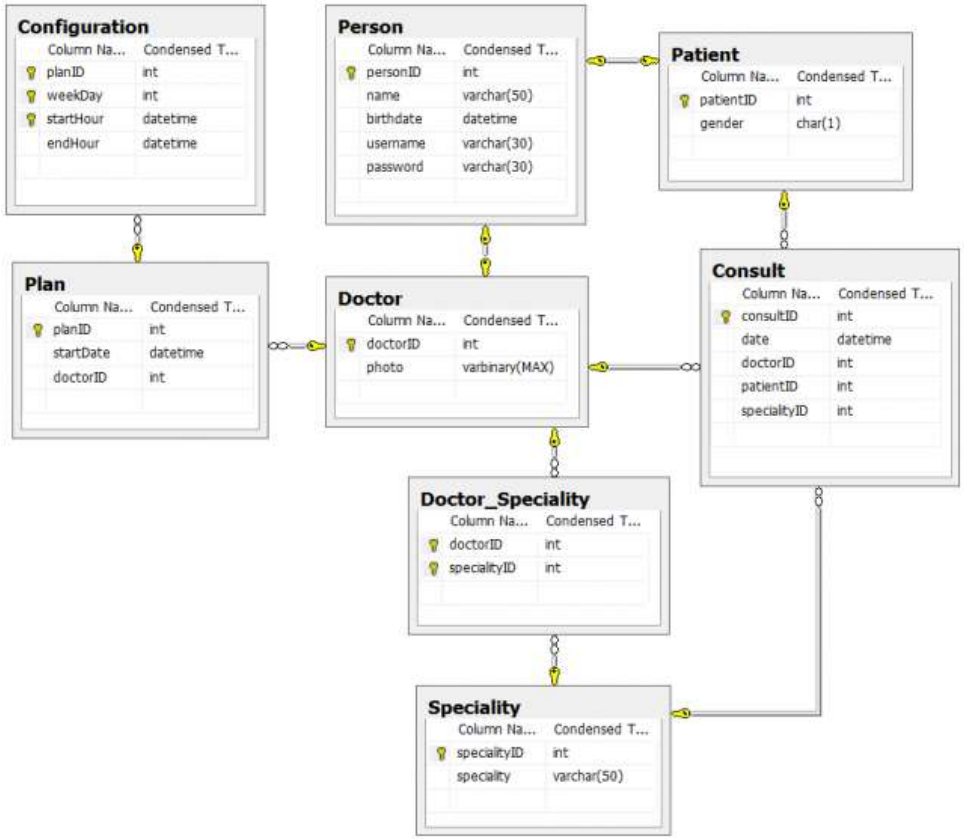
- A linguagem padrão dos Bancos de Dados Relacionais é a Structured Query Language, ou simplesmente SQL, é a mais conhecida

Banco de Dados



- **Tabelas (ou relações, ou entidades)**
- Todos os dados de um banco de dados relacional (RDBM) são armazenados em tabelas
- Uma tabela é uma estrutura de linhas (os registros) e colunas (os campos ou atributos)
- Em uma tabela, cada linha contém um mesmo conjunto de colunas
- 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



- **Colunas ou Atributos**

- As colunas de uma tabela são também chamadas de atributos.

  - Ex.: O campo Nome, ou endereço de uma tabela de um BD

Banco de Dados

- **Registros (ou tuplas)**

- Cada linha formada por uma lista ordenada de colunas representa um **registro**, ou **tupla**

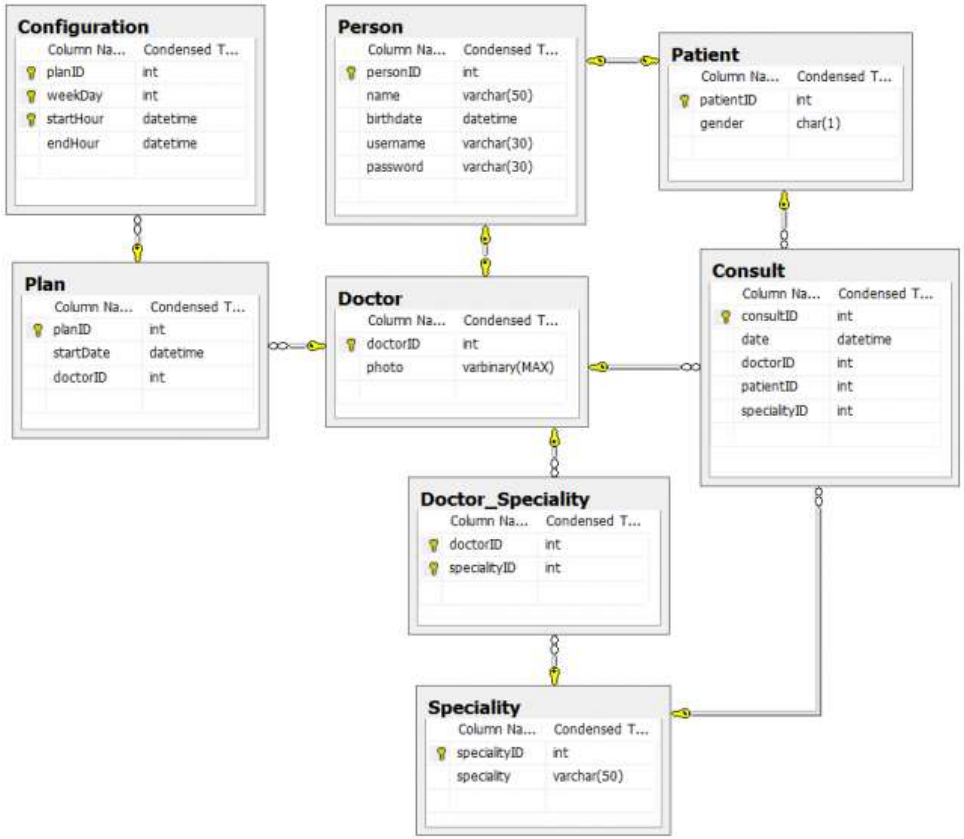
- Os registros não precisam conter informações em todas as colunas, podendo assumir valores nulos quando assim se fizer necessário

- Um registro é uma instância de uma tabela, ou entidade.

- Uma entidade é uma representação de um conjunto de informações sobre determinado conceito do sistema. Toda entidade possui ATRIBUTOS, que são as informações que referenciam a entidade.

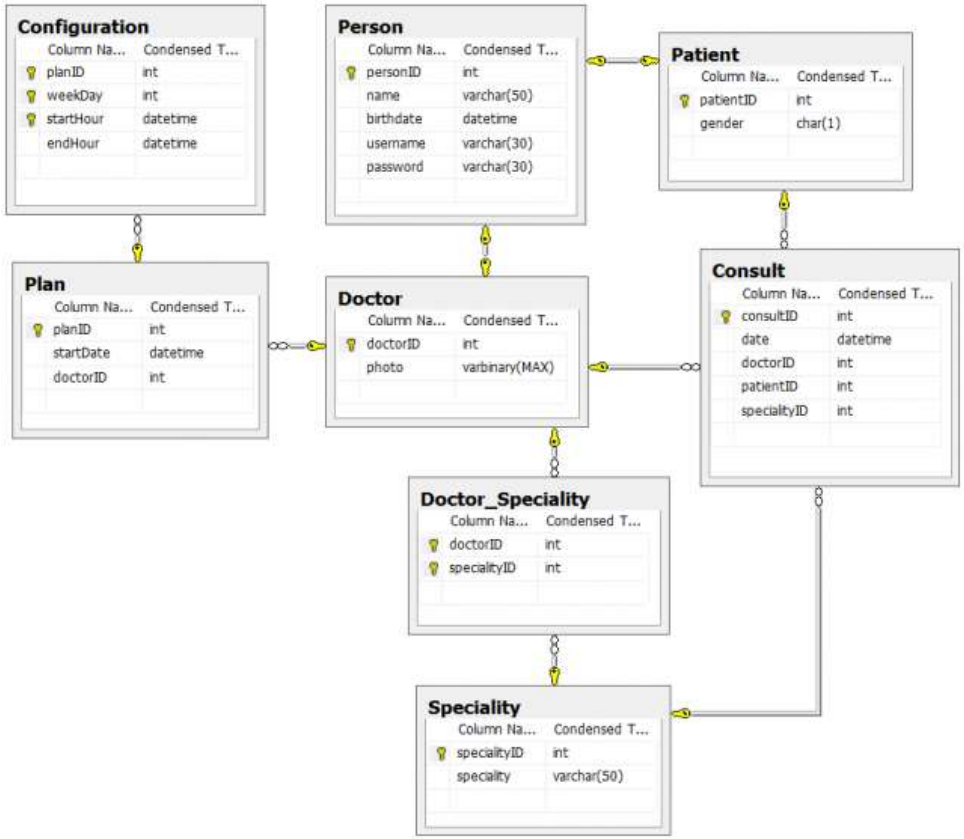
	CO_ANO	CO_MES	CO_NCM	CO_UNID	CO_PAIS	SG_UF_NCM	CO_VIA	CO_URF	QT_ESTAT	KG_LIQUIDO	VL_FOB
1	1997	3	41043911	15	149	RS	1	1010500	3987	4150	16725
2	1997	5	63019000	10	97	MG	7	145200	0	1002	8420
3	1997	6	87168000	11	586	RS	7	145300	48	153	915

Banco de Dados



- **Chave**
- As tabelas relacionam-se umas as outras através de chaves
- Uma chave é um conjunto de um ou mais atributos que determinam a unicidade de cada registro.
- A unicidade dos registros, determinada por sua chave, também é fundamental para a criação dos índices.

Banco de Dados



- Temos dois tipos de chaves:

- Chave primária: (PK - Primary Key) é um identificador exclusivo de todas as informações de cada registro dando-lhe unicidade
  - 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