Formação Cientista de Dados

Limpeza e Tratamento de Dados

PROF. FERNANDO AMARAL
www.datascientist.com.br

Um Cientista de Dados

- 80 % do tempo tratamento de dados

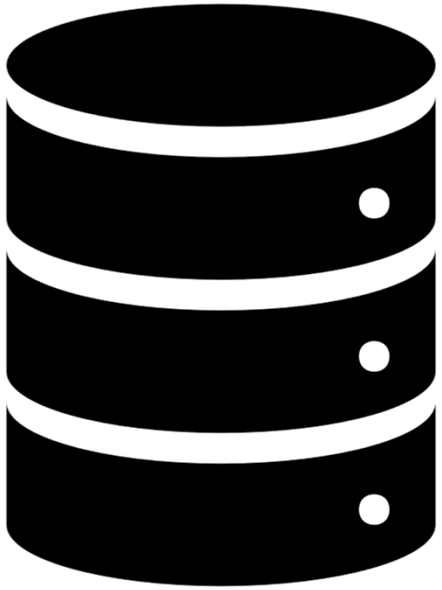

Produção
Financeiro
RH
Compras
Análise
Formação Cientista de Dados
010001
10

Porque dados tem problemas?

- Sistemas de operações e bancos de dados sem restrições de entrada
- Atualizações diretas em bancos de dados
- Sistemas antigos, codificações diferentes (EBCDIC)
- Inconsistência nos processos de carga:
  - Origem da informação é diversa, não padronizada
  - Mudanças no processo
  - Erros no processo

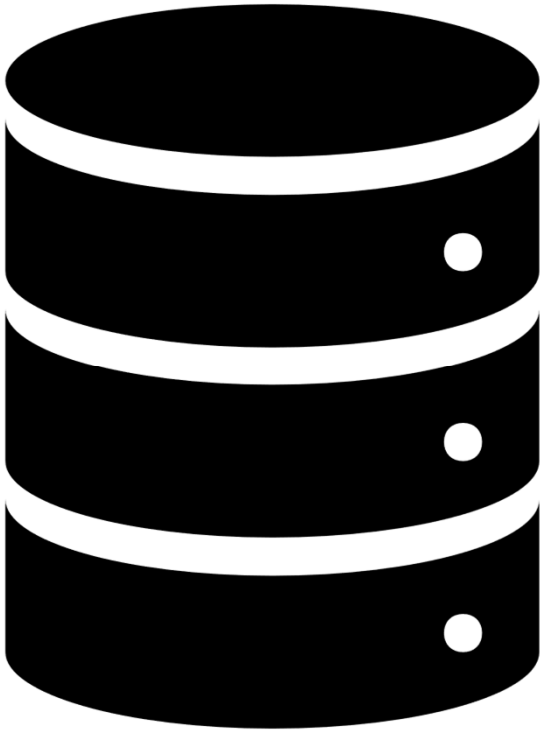
Operação VS Analítico

- Na operação o dado em seu formato individual não pode ser alterado para um valor padrão
- No analítico, o dado não tem valor individual, mas coletivo. Ele pode ser corrigido pelo "bem" do modelo



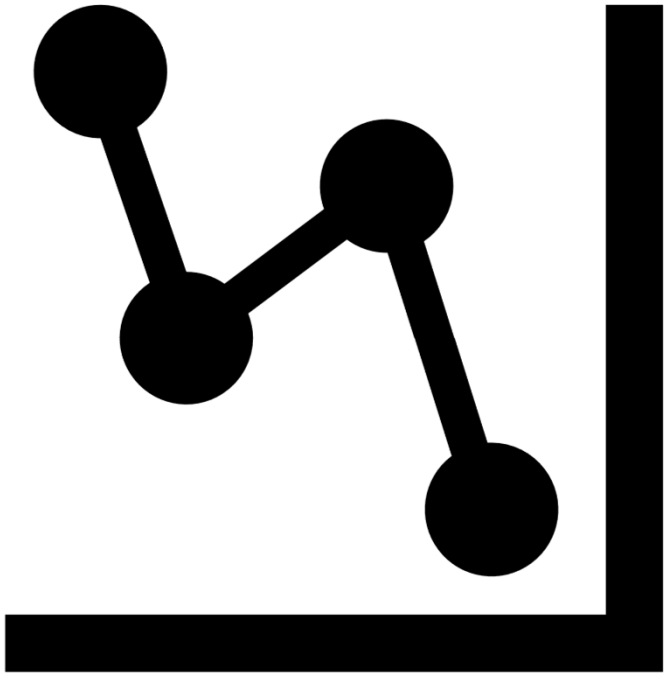
Operação

- Cliente do plano de saúde tem data de nascimento faltante
- Não podemos preencher com a mediana, pois isso influência o valor do plano!



Analítico

- Modelo para prever custo dos clientes para o plano de saúde
- Algoritmo não suporta valores faltantes
- Alterar uma idade faltante para a mediana:
  - Não vai afetar a operação
  - Não vai causar enviesamento no modelo



Problemas encontrados

- Duplicidades
- Consistência
- Completude
- Conformidade
- Integridade

PROF. FERNANDO AMARAL
www.datascientist.com.br