IGNORÂNCIA ZERO

Curso de Ciência de Dados

Manipulação de Dados

São Paulo, 21 de janeiro de 2022

Pedro Forli

Ignorância Zero

O que são Dados?

São representações computacionais usadas para guardar um conteúdo quantificável associado a uma entidade ou evento

Definições

- Na computação, dados são informações que foram traduzidas em uma forma eficiente para movimentação ou processamento
- Os computadores representam dados, incluindo vídeo, imagens, sons e texto, como valores binários usando padrões de apenas dois números: 1 e 0
- Na era moderna, praticamente todos os sistemas automatizados geram alguma forma de dados ou para fins de diagnóstico ou análise
- Dados, por si só, não expressam significado algum, sendo apenas um conteúdo quantificável associado a alguma entidade ou evento
- Os dados brutos podem ser arbitrários, não estruturados, ou mesmo em um formato que não é imediatamente adequado para processamento automatizado

O que é Mineração de Dados?

É o estudo da coleta, limpeza, processamento, análise e obtenção de insights de dados

- Mineração de dados é a exploração e análise de dados para descobrir padrões ou regras que sejam significativos
- A mineração de dados tem como objetivo:
  - Peneirar muitas informações repetitivas de maneira organizada
  - Extrair informações relevantes e faça o melhor uso delas para melhores resultados
  - Acelerar o ritmo da tomada de decisões bem informada
- Há múltiplos métodos que permitem encontrar padrões de dados como:
  - Reconhecimento Automático de Padrões
  - Previsão dos resultados mais prováveis
  - Destaque os agrupamentos que ocorrem naturalmente

Coleta de Dados

É o processo no qual realizamos a conversão de sinais do mundo real em representações binárias no mundo digital

- Coleta de dados é o processo de amostragem de sinais que medem as condições físicas do mundo real e convertem as amostras resultantes em valores numéricos digitais que podem ser manipulados por um computador
- Os sistemas de aquisição de dados, abreviados pelos inicialismos DAS, DAQ ou DAU, normalmente convertem formas de onda analógicas em valores digitais para processamento
- Os componentes dos sistemas de aquisição de dados incluem:
  - Sensores, para converter parâmetros físicos em sinais elétricos.
  - Circuito de condicionamento de sinal, para converter os sinais do sensor em uma forma que pode ser convertida em valores digitais.
  - Conversores analógico para digital, para converter sinais de sensores condicionados em valores digitais.
- Esta é a etapa mais custosa de todo o processo e exige o dimensionamento adequado desses aparelhos de aquisição, além da construção de um sistema de armazenamento dessa informação

Fontes: https://en.wikipedia.org/wiki/Data_acquisition;

Formato dos Dados

Dados são convertidos para representações binárias de seus elementos fundamentais (texto ou número) e são dispostos em vetores / matrizes para o acesso do usuário

Ignorância Zero

Texto

O	L	A		M	U	N	D	O
79	76	65	32	77	85	78	68	79

0000000001001111

Som

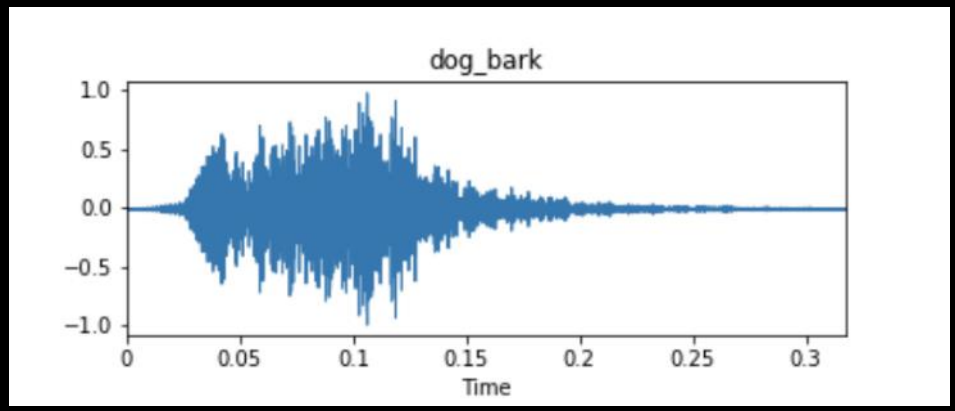
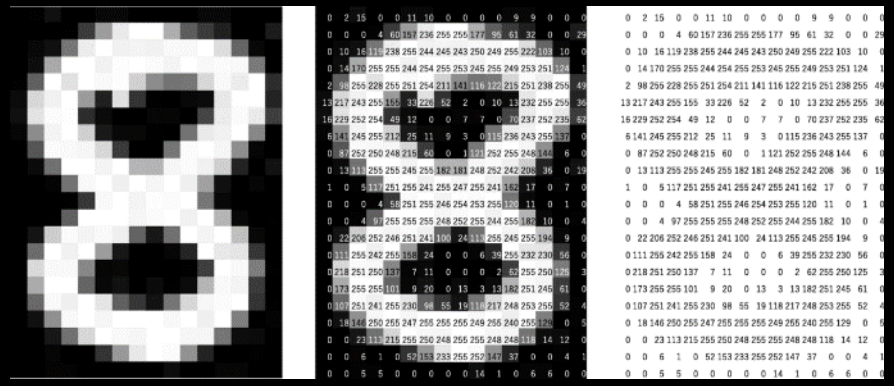
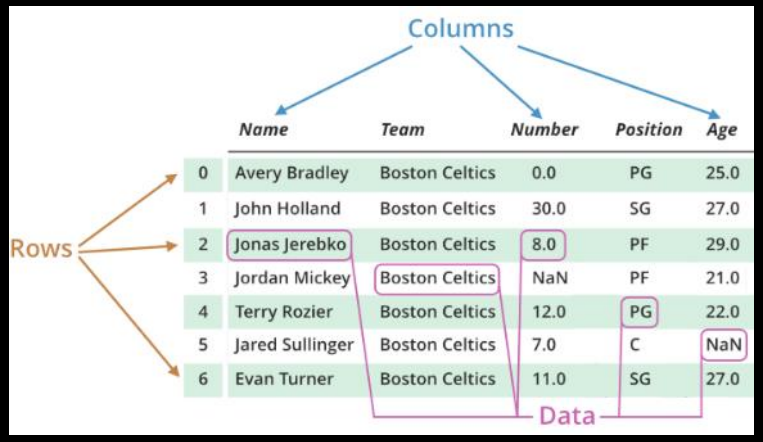


Imagem / Video



Tabela



Classificação de Dados

Os dados tabulares/vetoriais ainda são classificados de acordo com o tipo de entidade descrito e sua relação com outros dados

Quantitativo vs Qualitativo

- **Qualitativos:** Descrevem qualidades ou características
  - Coletado por meio de questionários, entrevistas ou observação e frequentemente aparece na forma de narrativa
  - Ex: Anotações feitas durante um grupo de foco sobre a qualidade da comida no Cafe Mac ou respostas de um questionário aberto
  - Podem ser difíceis de medir e analisar com precisão
  - Podem estar na forma de palavras descritivas que podem ser examinadas quanto a padrões ou significado
- **Quantitativos**: São usados quando está se tentando quantificar um problema ou abordar o "o quê" ou "quantos" aspectos de uma questão de pesquisa
  - São dados que podem ser contados ou comparados em uma escala numérica
  - São geralmente coletados por meio de instrumentos, como um questionário que inclui uma escala de classificação ou um termômetro para coletar dados meteorológicos

Nominal, Ordinal, Discreto e Contínuo

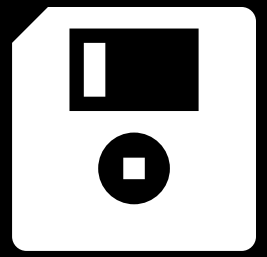
- **Nominal**: Conjunto de valores que não possuem uma ordem natural. Ex: A cor de um objeto
- **Ordinal**: Valores que têm uma ordem natural, embora mantendo sua classe de valores. Ex: O sistema de classificação durante a classificação dos candidatos em um teste em que A+ é melhor do que a nota B
- **Discreto**: Números inteiros são colocados nesta categoria. Ex: O número de alto-falantes no telefone, câmeras, núcleos no processador, o número de sims suportados
- **Contínuo**: Os números fracionários são considerados valores contínuos. Eles podem assumir a forma de frequência operacional dos processadores, a versão Android do telefone, frequência wi-fi, temperatura dos núcleos e assim por diante

Não dependente vs Dependente

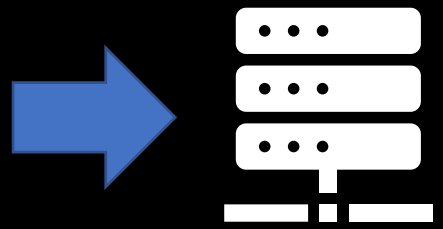
- **Dados orientados para não dependência**: Refere-se a tipos de dados simples, como multi-dimensionais ou dados de texto
  - Mais simples e mais comumente encontrados.
  - Registros de dados não possuem dependências especificadas entre os itens de dados ou os atributos
  - Ex: Conjunto de dados demográficos com registros sobre indivíduos contendo sua idade, sexo e código postal
- **Dados orientados a dependências**: Podem existir relacionamentos implícitos ou explícitos entre itens de dados
  - Ex: conjunto de dados de rede social contém um conjunto de vértices (itens de dados) que são conectados por um conjunto de arestas (relacionamentos).
  - Ex: Série temporal contém dependências implícitas, como dois sucessivos valores coletados de um sensor provavelmente estão relacionados entre si

Armazenamento dos Dados

Dados são armazenados em diferentes formatos em algum local físico e podem ser acessados diretamente ou por meio de protocolos de conexão Web para o caso de armazenamento remoto

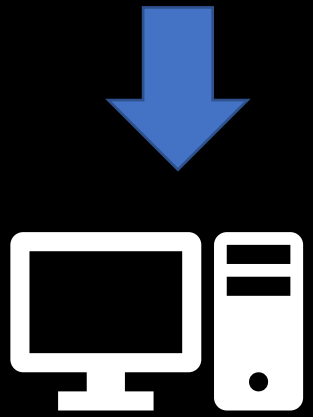


- Som: .wav, .mp3, .ogg, ...
- Imagem: .png, .jpeg, .tif, ...
- Vídeo: .mp4, .mkv, .avi, ...
- Texto: .txt, .docx, .pdf, ...
- Tabela: .csv, .xlsx, .parquet, ...
- Espacial: .shp, .kmv, ...
- Estatístico: .dta, .xport, ...
- Interchange: .json, .ubj, ...
- ...

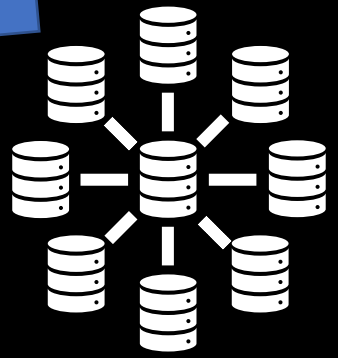


Servidor

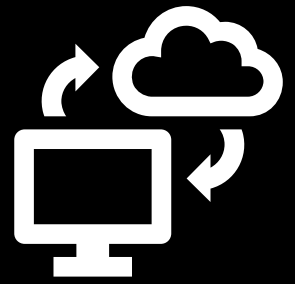
No caso do nosso curso, muitas vezes vamos querer converter o armazenamento remoto em local para realizar o acesso aos dados



Computador Pessoal



Data Warehouse



Conexão Web

- HTTP/HTTPS
- FTP
- TCP/IP
- IMAP
- SMTP
- UDP

Ignorância Zero

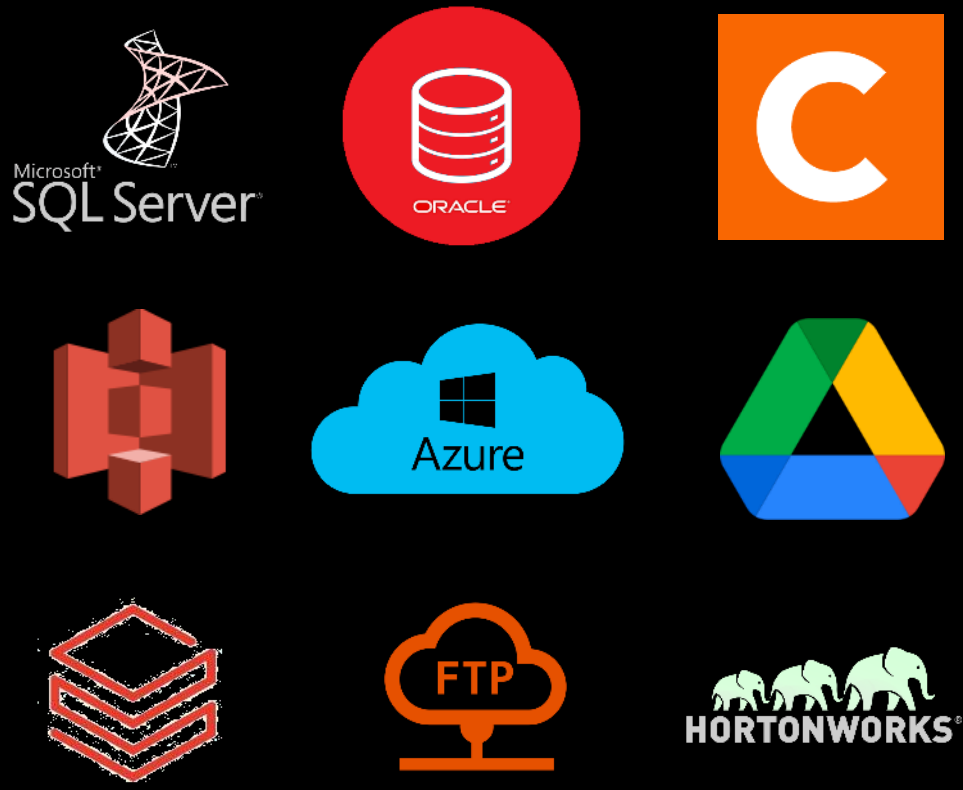
Aquisição de Dados (1/2)

Dentro do contexto de mineração de dados, a aquisição corresponde ao processo de extração dos dados coletados para o sistema que fará a limpeza e processamento desses dados

Fontes Públicas

- Nacionais:
  - https://www.gov.br/pt-br/orgaos-do-governo
  - https://ftp.ibge.gov.br/
  - https://geoftp.ibge.gov.br/
  - https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/inep-data
  - https://www.gov.br/anatel/pt-br/dados
  - https://dadosabertos.bcb.gov.br/
- Internacionais:
  - http://data.worldbank.org/
  - https://ec.europa.eu/eurostat/en/web/lfs/statistics-illustrated
  - https://archive.org/web/
- Empresas:
  - http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/
  - https://www.google.com/publicdata/directory?dl=pt_PT&hl=pt_PT
  - http://aws.amazon.com/public-data-sets/
  - https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75
  - https://www.kaggle.com/datasets

Fontes Proprietárias



Ignorância zero

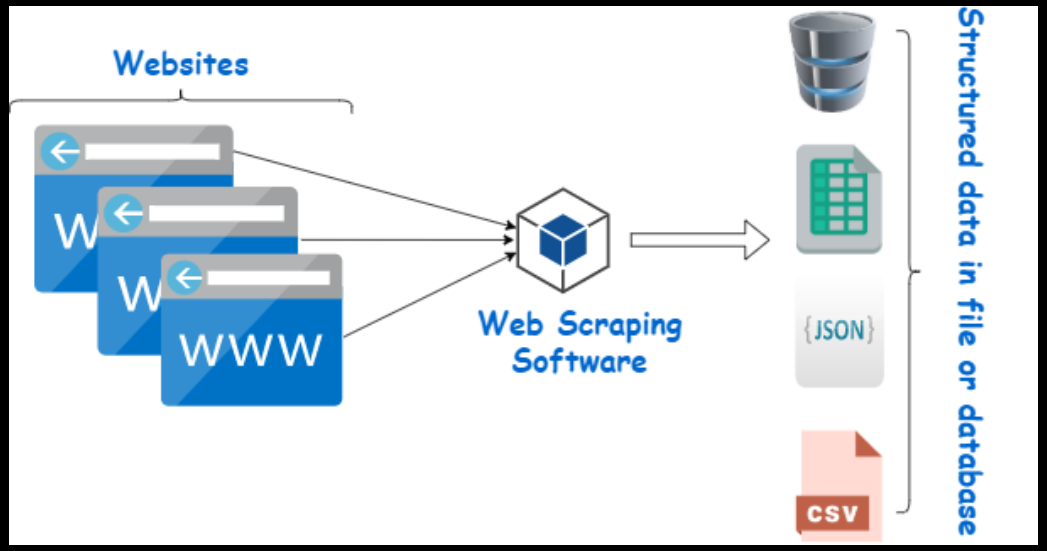
Aquisição de Dados (2/2)

Dentro do contexto de mineração de dados, a aquisição corresponde ao processo de extração dos dados coletados para o sistema que fará a limpeza e processamento desses dados

API (application programming interface)

- Google (https://developers.google.com/apis-explorer)
  - Geocoding (https://geopy.readthedocs.io/en/stable/)
  - Search (https://developers.google.com/webmaster-tools/search-console-api-original/v3/quickstart/quickstart-python)
  - Gmail (https://developers.google.com/gmail/api/quickstart/python
- Twitter (https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries)
- Receita Federal (https://www.gov.br/conecta/catalogo/apis/consulta-cnpj/swagger_view)
- Correios (https://pypi.org/project/pycep-correios/)
- Bacen (https://dadosabertos.bcb.gov.br/dataset/sistema-de-registro-de-operacoes-de-credito-com-o-setor-publico-cadip)
- Yahoo Finance (https://pypi.org/project/yfinance/)
- Open Weather (https://rapidapi.com/blog/openweathermap-api-overview/python/)

Web Scrapping



- https://docs.python.org/3/library/urllib.html
- https://docs.python-requests.org/en/master/
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- https://selenium-python.readthedocs.io/

Aquisição de Dados – Nosso Projeto

Como parte do projeto nós criaremos uma ferramenta que fará download, processamento e modelagem de forma automatizada, aos quais você terá acesso a todos os dados por meio do google drive

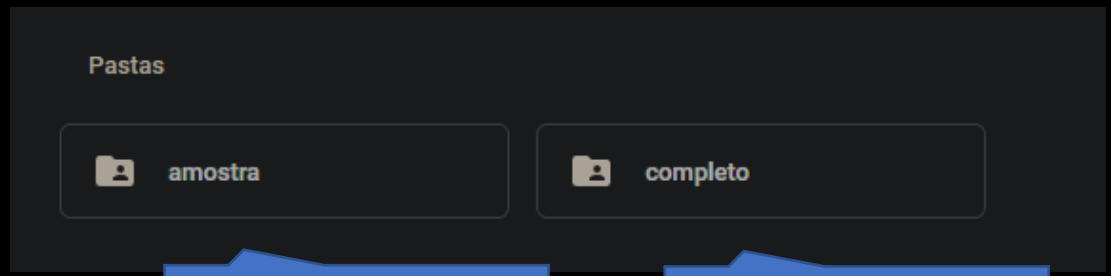
Ignorância Zero

Download dos dados públicos

Processamento e limpeza das bases

Geração de modelos

Escopo da ferramenta

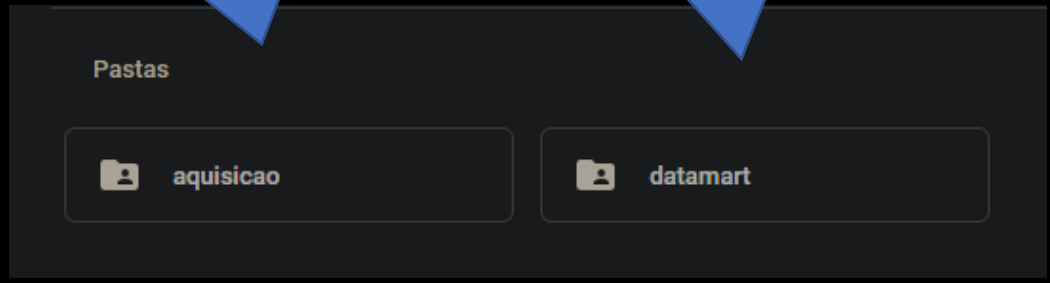


Amostra de escolas para aqueles com menor poder computacional

Dados completos conforme usado na ferramenta

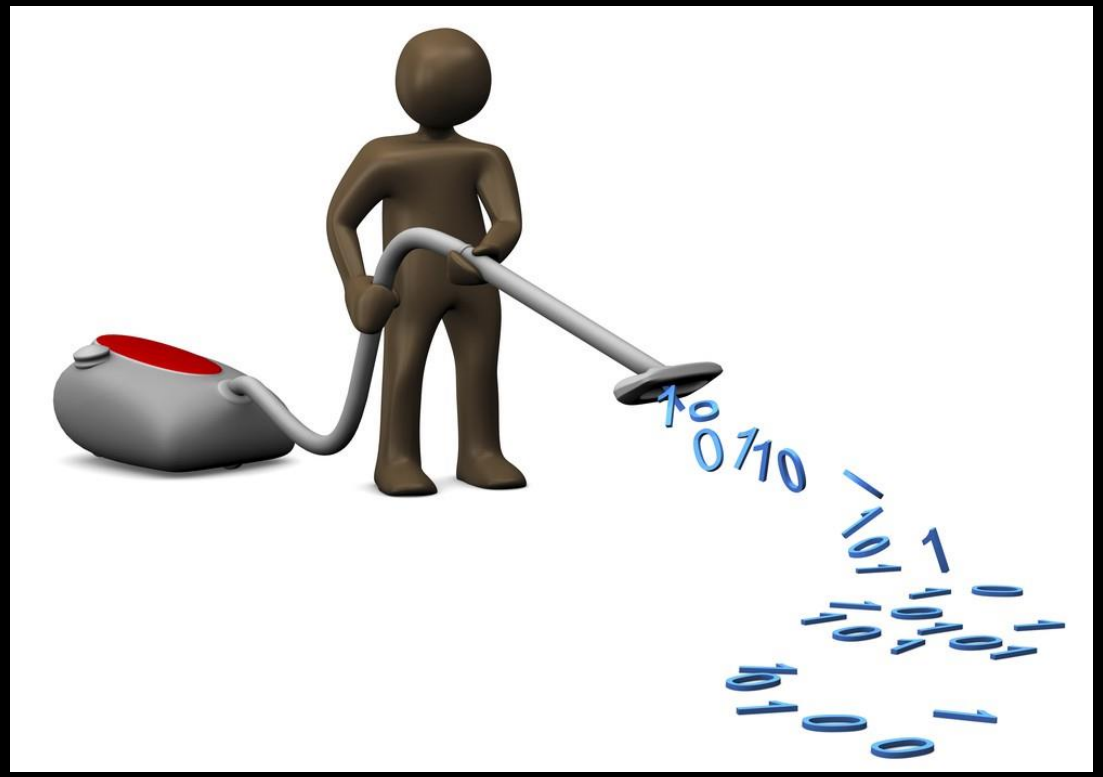
Dados processados pela ferramenta e resultados de modelos

Dados consolidados em diferentes granularidades



Limpeza de Dados

É o processo pelo qual fazemos a correção/remoção de dados e normalizamos variáveis de forma a construir um padrão de estrutura de informação a ser utilizado na ferramenta

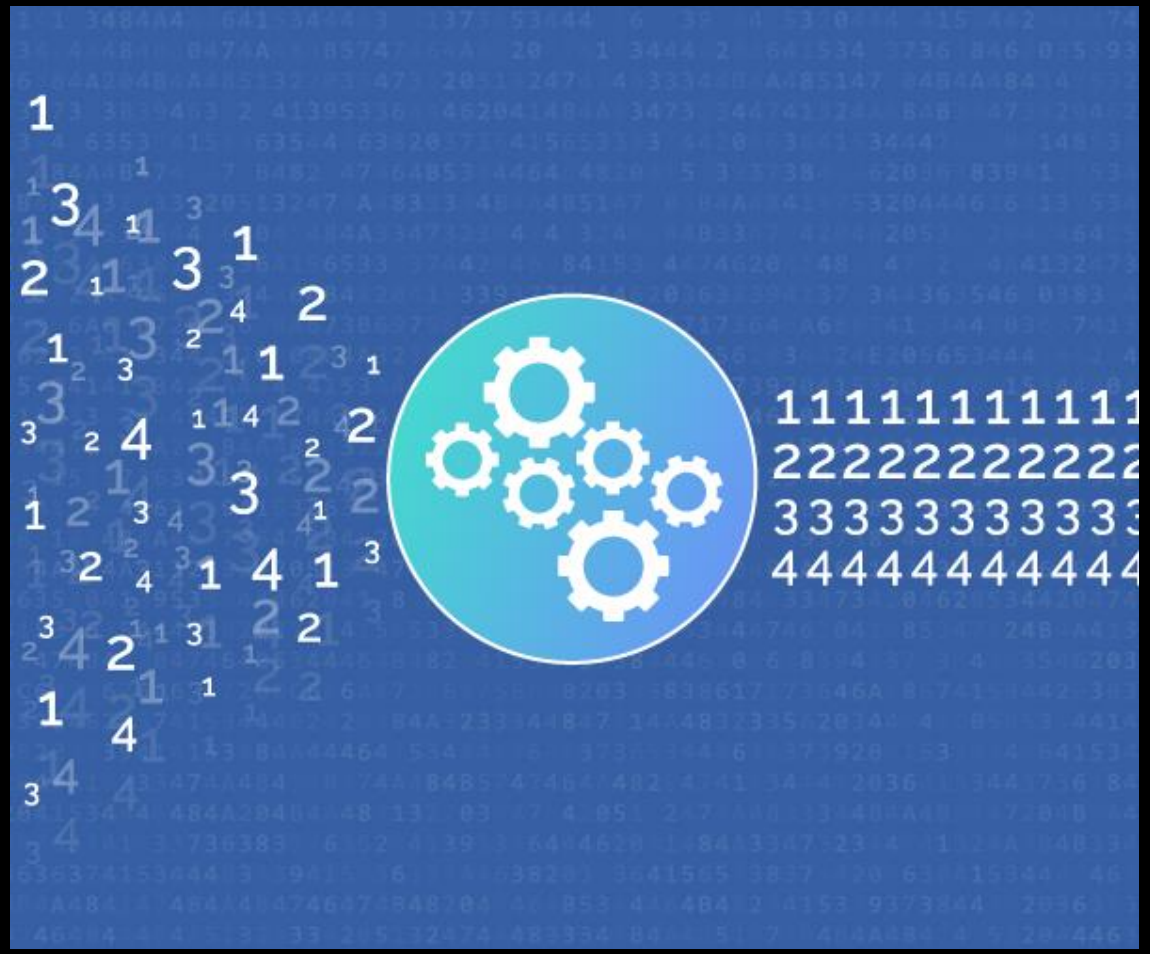


- A limpeza de dados é o processo de corrigir ou remover dados incorretos, corrompidos, formatados incorretamente, duplicados ou incompletos em um conjunto de dados
- Ao combinar várias fontes de dados, existem muitas oportunidades para os dados serem duplicados ou rotulados incorretamente
- Se os dados estiverem incorretos, os resultados e algoritmos não são confiáveis, embora possam parecer corretos
- Não existe uma maneira absoluta de prescrever as etapas exatas no processo de limpeza de dados porque os processos variam de conjunto de dados para conjunto de dados
- Etapas comuns do processo:
  - Remoção de duplicatas
  - Remoção de observações irrelevantes
  - Correção de erros estruturais
  - Remoção de outliers
  - Ajuste de valores nulos

Fonte: https://www.tableau.com/learn/articles/what-is-data-cleaning

Processamento / Transformação de Dados

Consiste no processo de transformação e conversão de dados entre diferentes estruturas e formatos, com o objetivo de facilitar o acesso e homogeneizar o conceito de determinados campos



- A limpeza de dados é o processo que remove dados que não pertencem ao seu conjunto de dados, enquanto que a transformação de dados é o processo de conversão de dados de um formato ou estrutura em outro
- As organizações que usam data warehouses locais geralmente usam um processo ETL (extrair, transformar, carregar)
- A transformação de dados pode ser construtiva (adicionar, copiar e replicar dados), destrutiva (excluir campos e registros), estética (padronizar saudações ou nomes de ruas) ou estrutural (renomear, mover e combinar colunas em um banco de dados)
- Benefícios:
  - Os dados transformados podem ser mais fáceis de usar tanto para humanos quanto para computadores
  - Os dados devidamente formatados e validados melhoram a qualidade dos dados e protegem os aplicativos de possíveis minas terrestres, como valores nulos, duplicatas inesperadas, indexação incorreta e formatos incompatíveis
  - A transformação de dados facilita a compatibilidade entre aplicativos, sistemas e tipos de dados

Fonte: https://www.tableau.com/learn/articles/what-is-data-cleaning; https://www.stitchdata.com/resources/data-transformation/

Ferramentas de Limpeza e Processamento

Há uma imensa quantidade de ferramentas, entretanto focaremos primeiro em pandas, depois ao longo do curso evoluiremos para Hive e Spark

Ignorância Zero

“Education is the passport to the future, for tomorrow belongs to those who prepare for it today”

- Malcolm X