INSTITUTO SUPERIOR DE TECNOLOGIAS AVANÇADAS

Ciência de Dados

Licenciatura Engenharia Informática

2º Semestre – 2021/2022

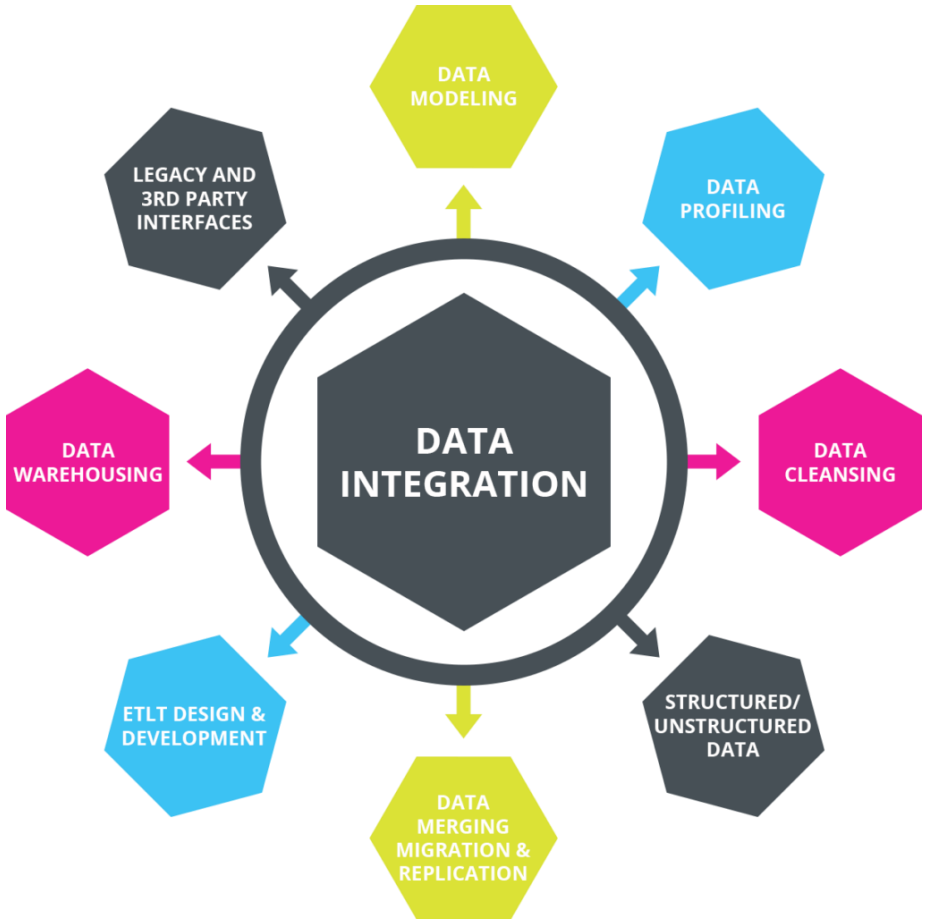
Ricardo Jesus Ferreira

ricardojesus.ferreira@my.istec.pt

Obtenção Dados

- Integração de dados
- ETL vs ELT
- Aquisição de dados

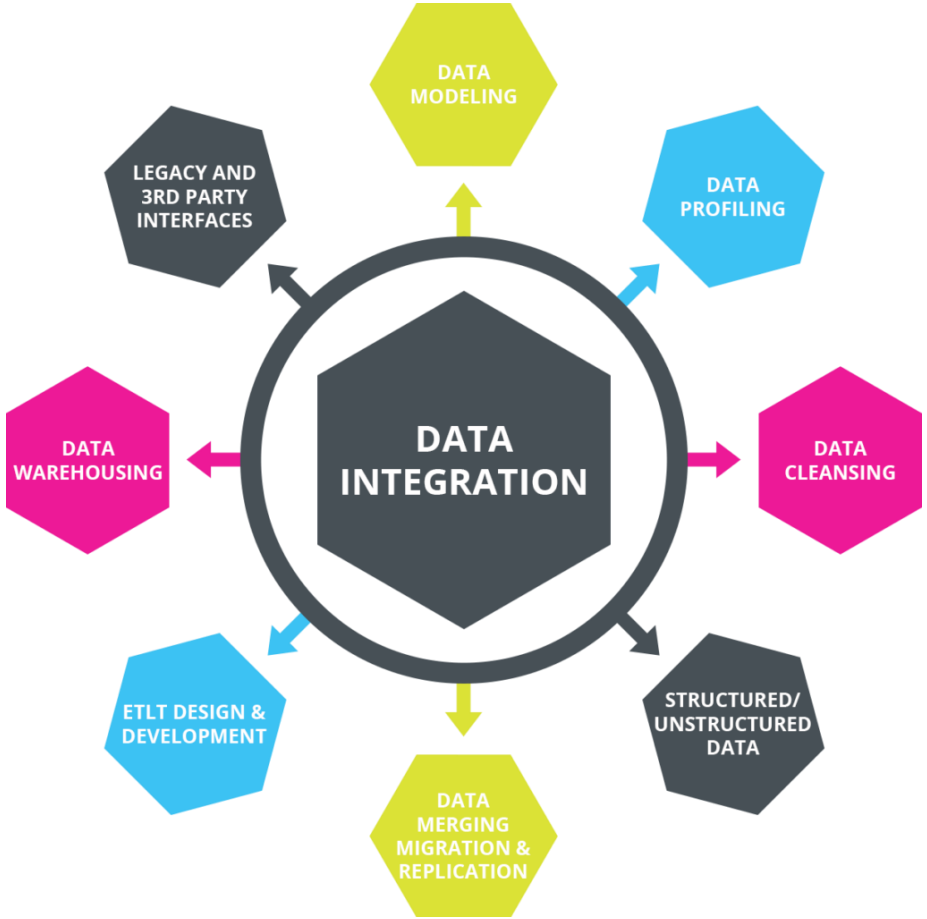
Integração de dados



- Processo de recolha de dados de diferentes fontes com o objetivo de dados de forma, simples, completa, precisa e *up-to-date*

- Este processo consiste no processamento, filtragem e modelação dos dados, que posteriormente serão armazenados num DW ou DL

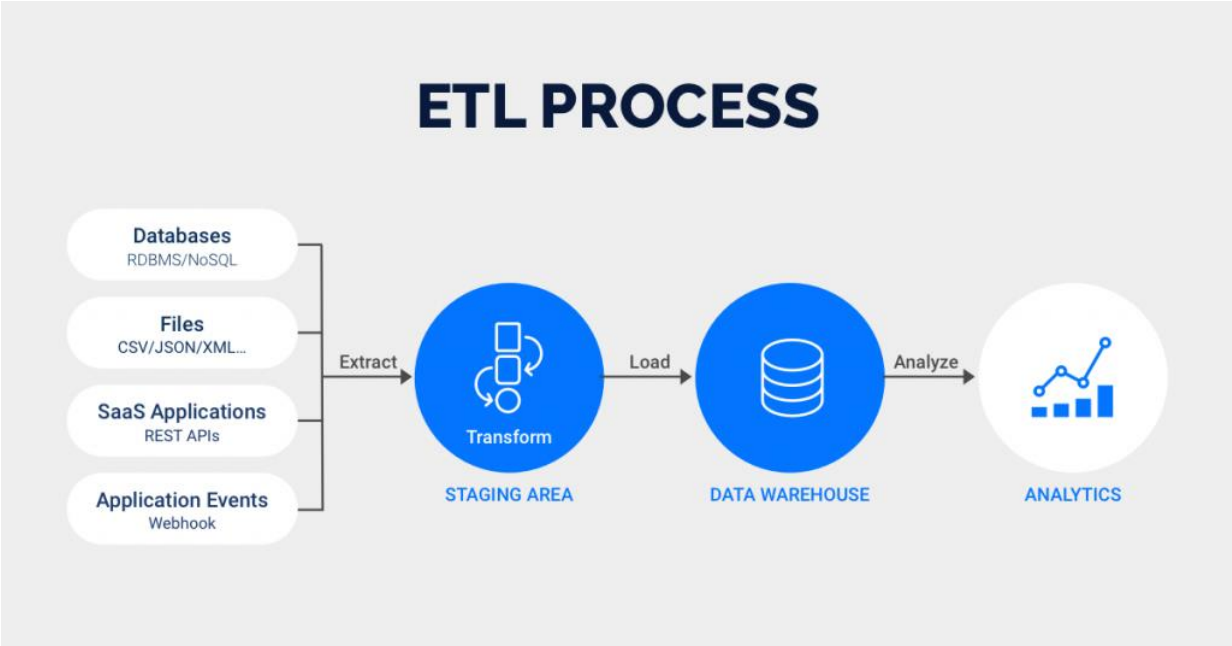
Vantagens



- Facilidade de obter informação de uma fonte com dados consolidados
- Isto permite com maior facilidade a criação de processos e decisões com recurso à analise dos dados
- Aumenta precisão e confiança
- Aumento da eficiência
- *Data-driven actions*

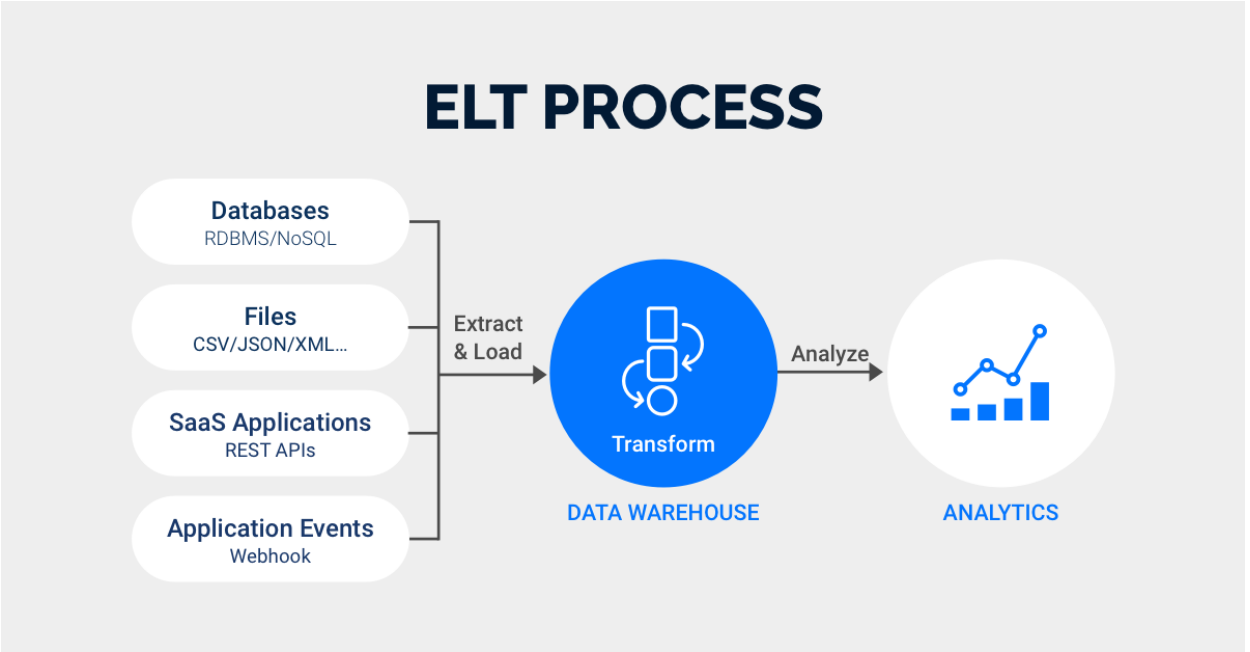
Estratégias de extração - ETL

- Extract, Transform and Load
- Standard utilizado para transformação e introdução dos dados no DW
- Pode ser utilizado:
  - Obter dados de diferentes fontes
  - Atualizar processos ou executar tarefas pré-definidas
  - Obter e transformar os dados de acordo com as necessidades do negocio

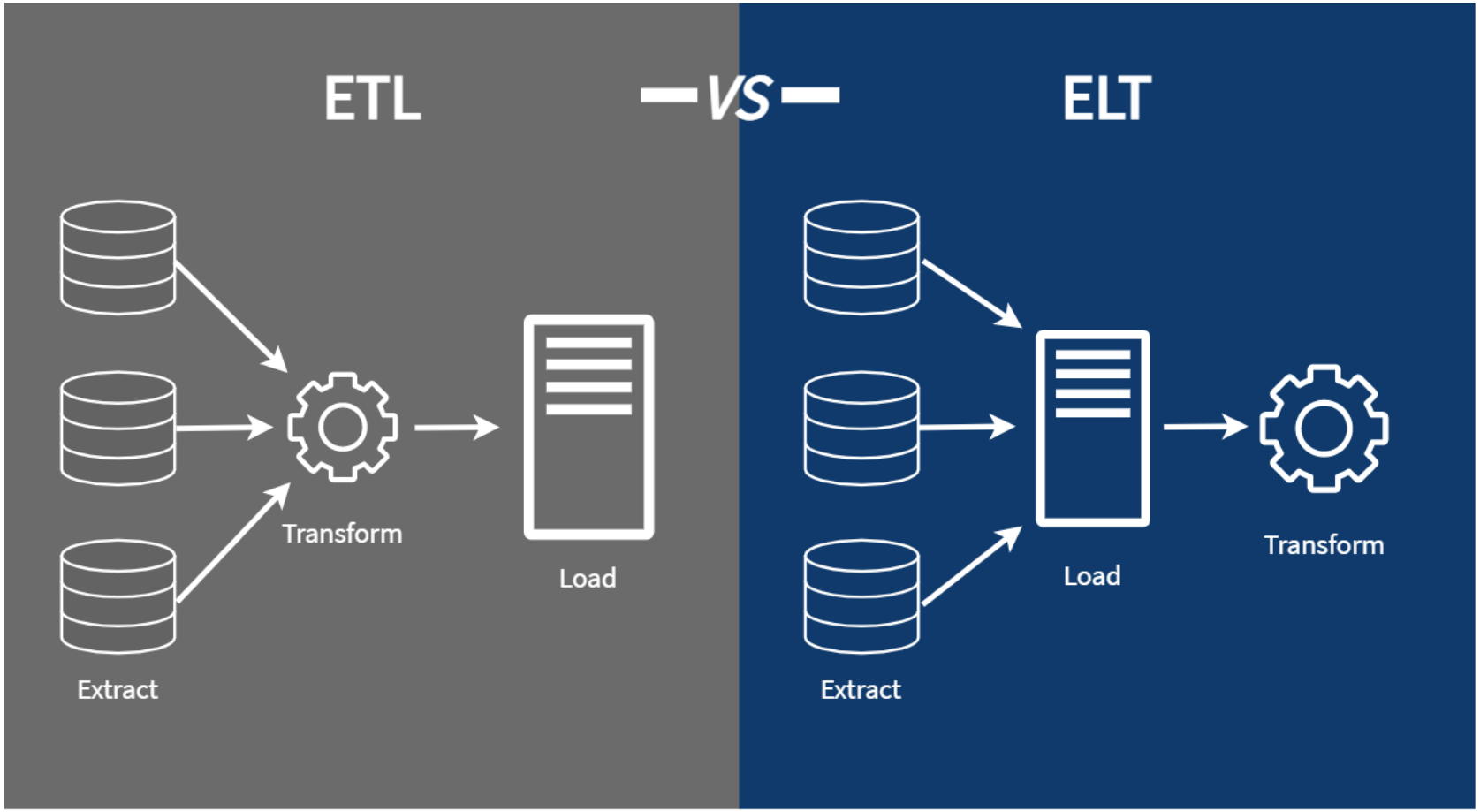


Estratégias de extração - ELT

- Extrat, Load and Transform
- Os dados são carregados e depois transformados de acordo com uma necessidade especifica
- Vantagens:
  - Real-Time
  - Baixo custo de manutenção e implentaçao



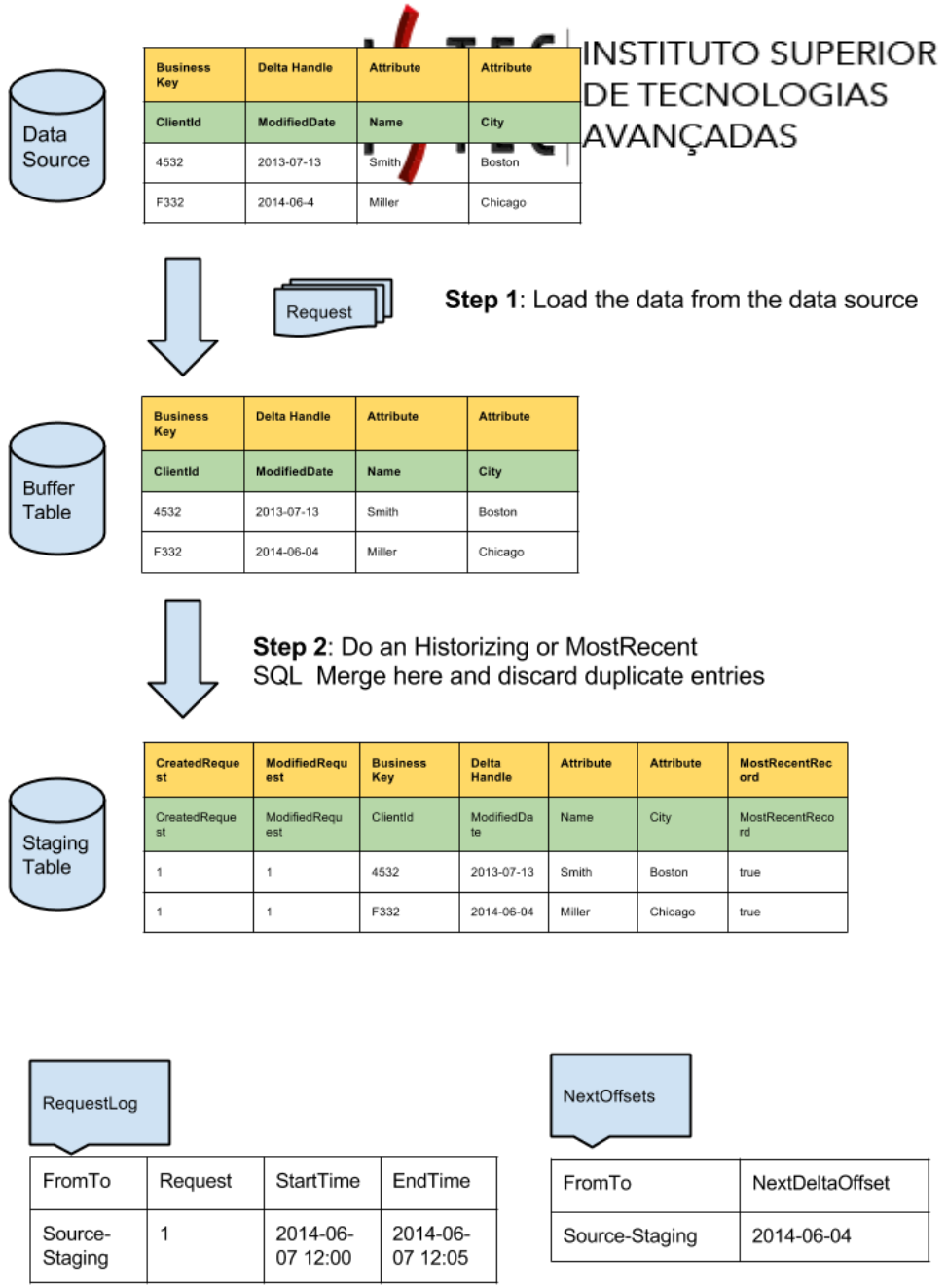
ETL vs ELT



	ETL	ELT
Source data	Support storing structured data from input sources	Can be used for structured, unstructured, and semi-structured data types
Data size	Best suited for smaller amounts of data	Can be used for large amounts of data
Storage type	Can be used for on-premise or cloud storage	Optimised for cloud data warehouses
Latency	High, as transformations need to be completed before storing data	Low, as minimal processing is done before storing in the data warehouse
Flexibility	Low, as data sources and transformations need to be defined at the beginning of the process	High, as transformations need not be defined when integrating new sources
Scalability	Can be low, as the ETL tool should support scaling of operations	High, as ELT tools can be easily configured for changing data sources
Maintenance	May need continuous maintenance in case of changes in data sources or formats	Low maintenance required as usually ELT tools automate the process
Compliance with security protocols	Easy to implement	May need to be supported by data warehouse/ELT tool
Storage requirement	Low as only transformed data is stored	Can be high as raw data is stored

Delta

- Utilizado acesso a dados com maior rapidez
- Não requer uma extração repetida da BD
- Esta estratégia implica o reconhecimento dos dados a extrair e dos dados extraídos

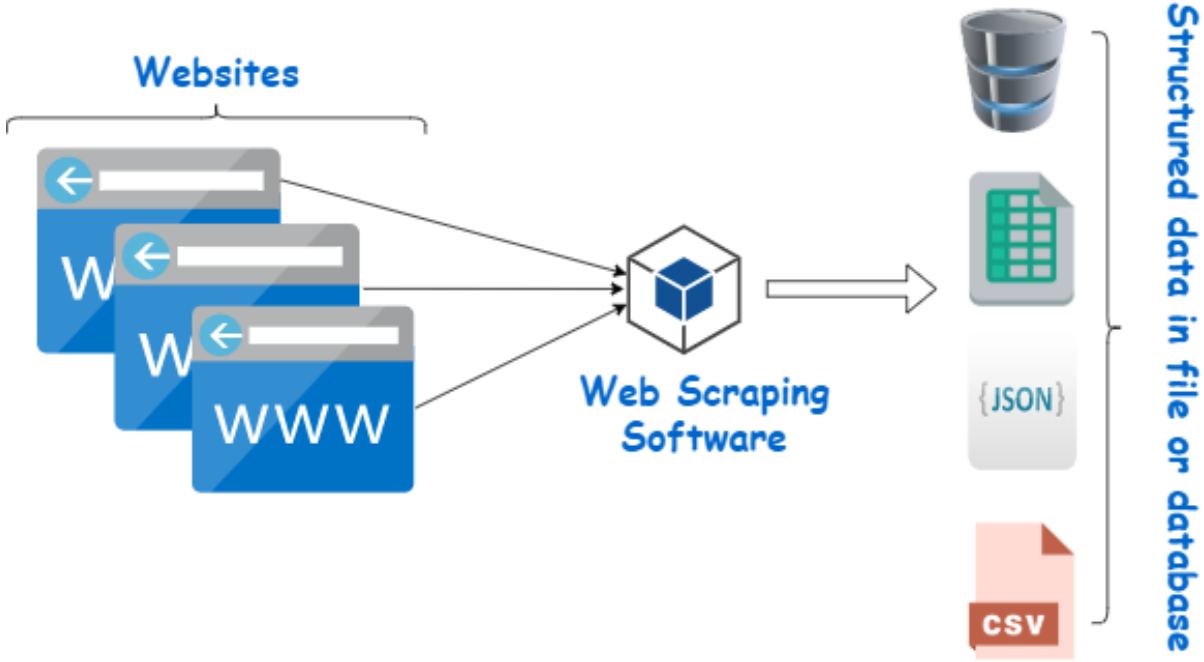


Aquisição de dados

- Recolher os dados certos é crucial!
- Dados incorretos geram informações erradas sobre a realidade, perda de tempo e esforço
- Os métodos de aquisição de dados que iremos abordar providenciam à organização o acesso a informações de qualidade que posteriormente vão gerar decisões de qualidade

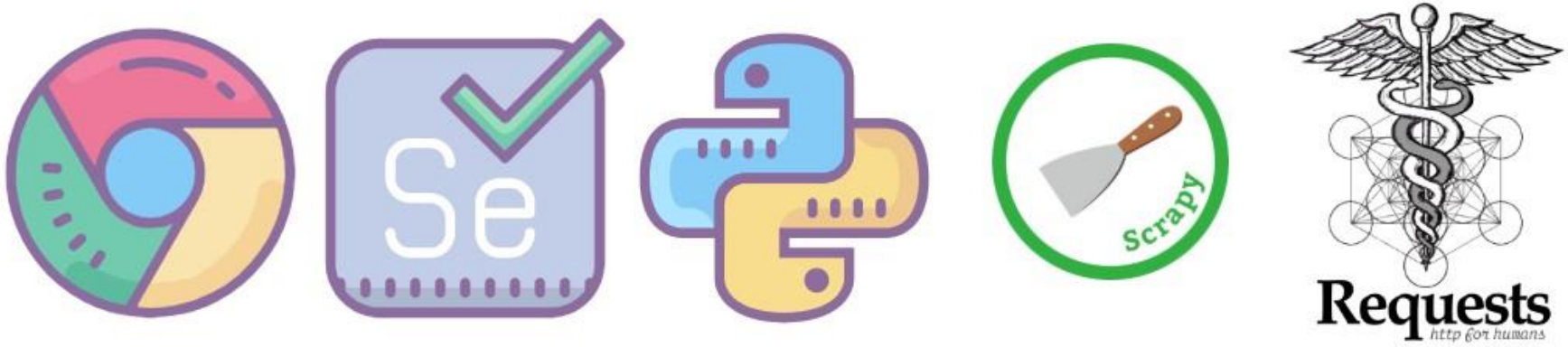
Aquisição de dados – Web Scraping

- Mecanismo de extração de dados de paginas web
- Podem ser extraídos: texto, imagens, vídeos, documentos, tabelas, etc..
- Boas práticas:
  - Não utilizar ou reproduzir conteúdo copyright
  - Respeitar o robots.txt
  - Ter um *crawler* justo
  - Não aceder a páginas privadas



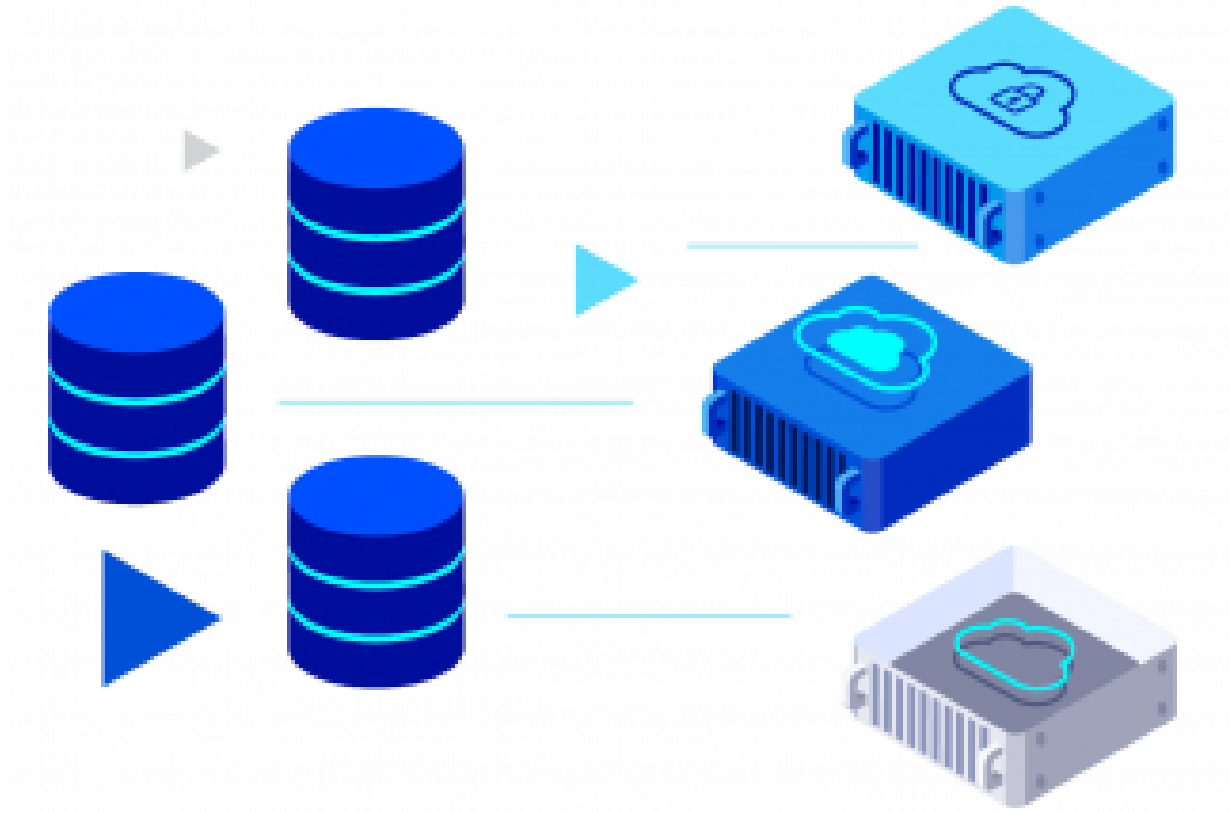
Aquisição de dados – Web Scraping

- CommonCrawl
- Beutiful Soup
- Requests
- Mechanize
- Scrapy
- Selenium

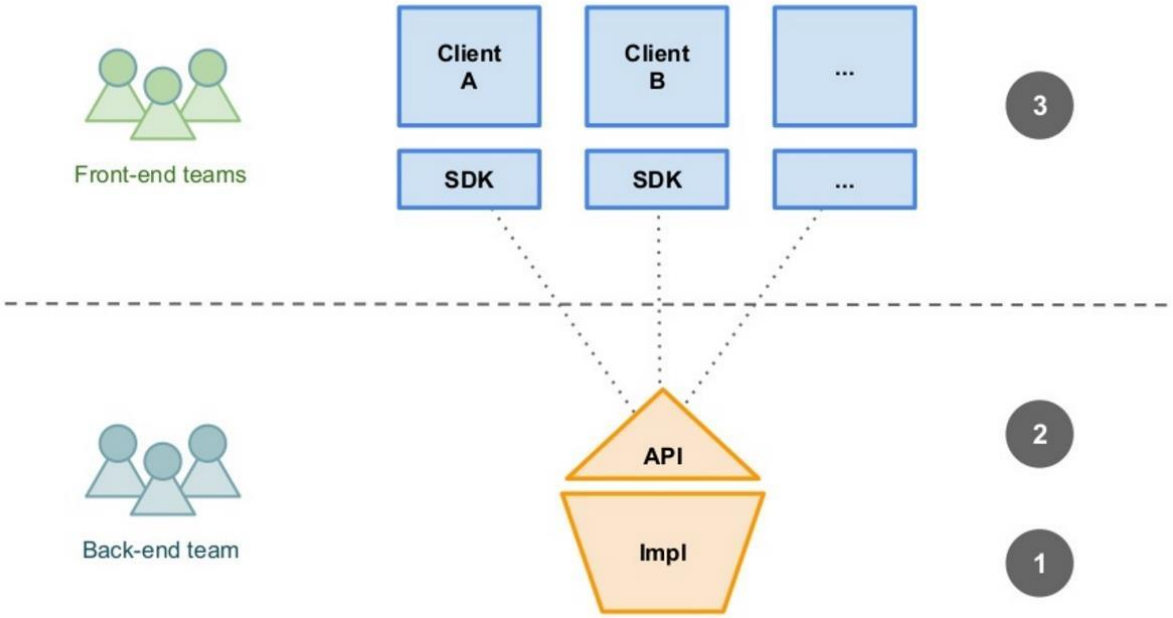


Aquisição de dados – BD Públicas

- Obtenção de dados não é tarefa fácil
- O acesso a dados públicos pode ser uma alternativa
- Por definição: Dados públicos podem ser utilizados, reutilizados ou distribuídos de forma gratuita

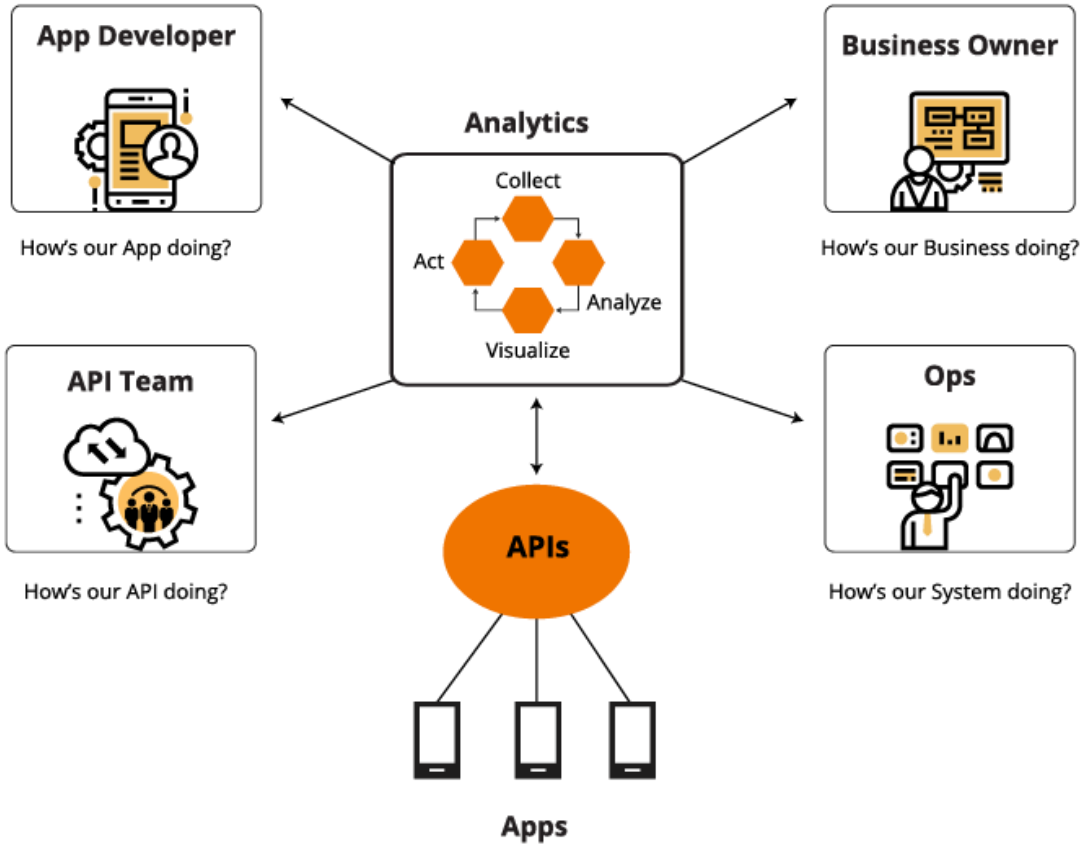


Aquisição de dados - APIs



- *Application Programming Interfaces* (APIs)
- Serve para garantir a integração, disponibilidade e partilha de informação e dados
- Funcionalidades que a camada aplicacional disponibiliza para interação/integração

Aquisição de dados - APIs

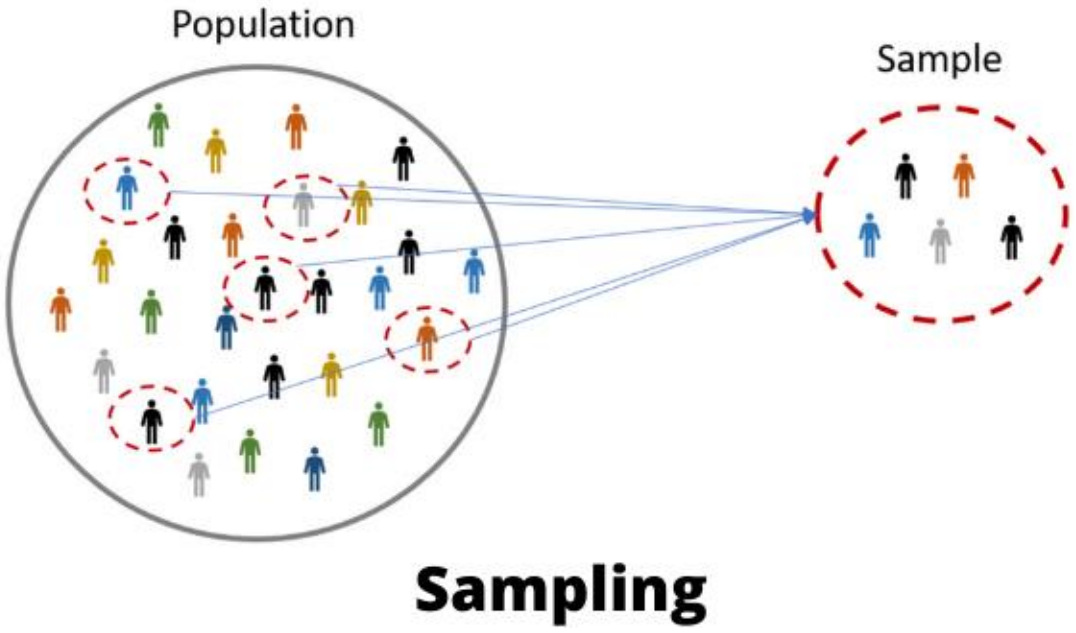


- Podem ser utilizadas para:
  - Partilha/consulta de informação, dados ou serviços
  - Colaboração entre equipas de desenvolvimento aplicacional
  - Integração entre várias aplicações
- Disponibilizam informação em diversos formatos, o mais conhecido é JSON

Aquisição de dados – Questionário

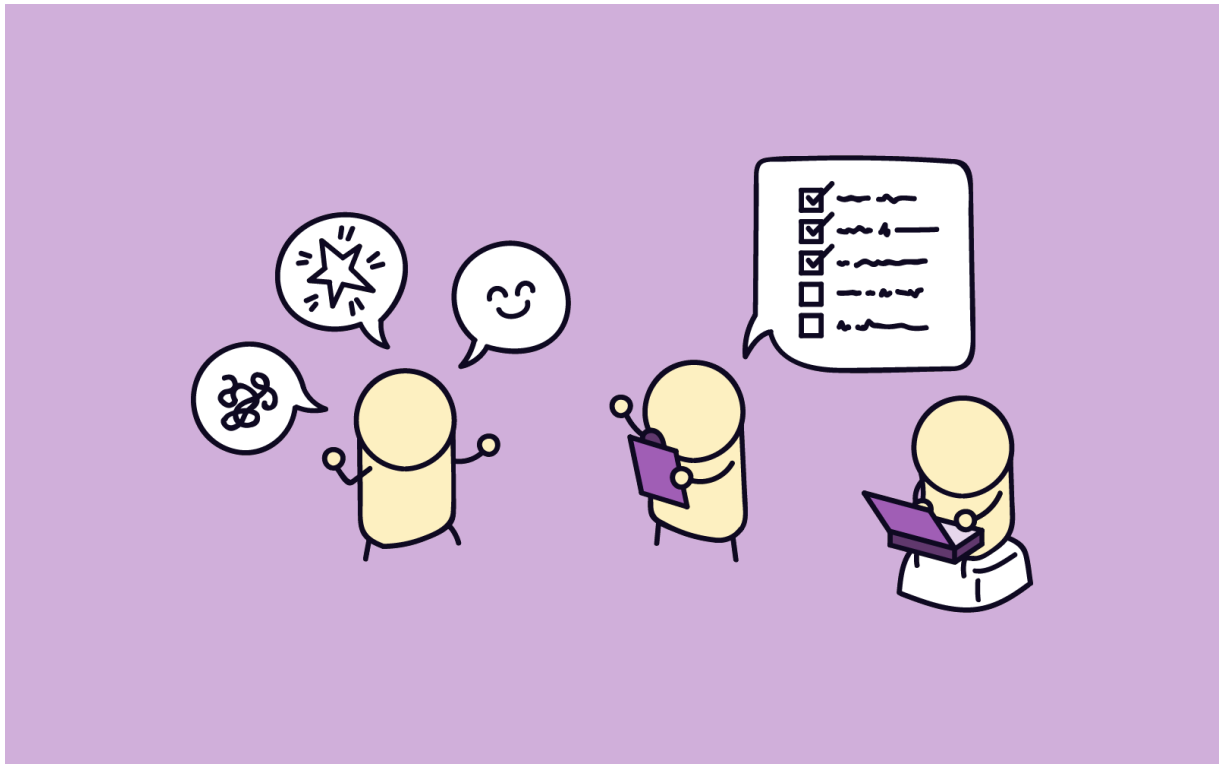
- Dados recolhidos por intermédio de um questionário ou inquérito
- Estes dados contêm informação especifica do grupo em estudo/análise
- Podem ser realizados:
  - Online
  - Telefone
  - Papel
  - Etc..
- Vantagens
  - Versátil
  - Eficaz (não recolhemos dados que não queremos)
  - Ferramenta por excelência para recolha de dados
- Desvantagens
  - Respostas demoram tempo
  - Custo

Aquisição de dados - *Sampling*



- Amostra, parte ou subconjunto da população total de dados
- Solução de IoT gera por dia 800000M de registos por dia
  - A manipulação de 800000M de registos é pesada por isso optamos por analisar 1 hora
- Tipos de amostras:
  - Baseadas em probabilidade
  - Baseadas na ausência de probabilidade

Aquisição de dados - Observação



- Obter dados pela observação de um evento ou comportamento
- Esta observação pode ser conhecida ou desconhecida
- Utilizado para:
  - relatar um problema
  - Obter informação sobre o comportamento de um individuo
  - Quando não conseguimos outro método

Bibliografia

- B. Gomez,(2020) “Resolviendo problemas de Big Data”, Alfaomega.
- D. Insua, (2019)“Big data: Conceptos, tecnologías y aplicaciones”, CSIC.
- H. Jones, (2019)“Analítica de datos”, HJ,.
- J. Somed, (2020)“Big Data Analytics”, JLC.
- D. Petković (2020)“Microsoft® SQL Server® 2019 A Beginner’s Guide - Seventh Edition”, McGraw Hill.