Extração de Dados

Paradigmas de Programação

Centro Universitário Senac

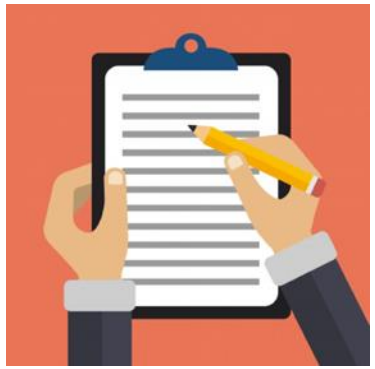
Prof. Celso Crivelaro

celso.vcrivelaro@sp.senac.br

Formas de Extração de Dados

Banco de Dados e Mensageria

Documentos



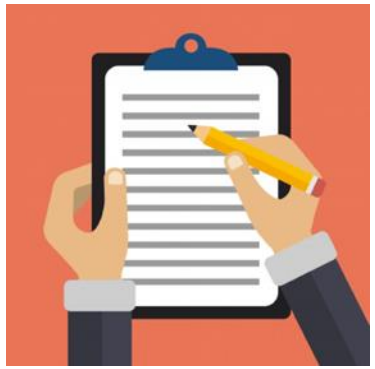
APIs

Scraping (Raspagem)

Formas de Extração de Dados

Banco de Dados e Mensageria

Documentos



APIs

Scraping (Raspagem)

Banco de Dados

Extração de Banco de Dados é a forma mais simples de Integração

Acesso direto ao banco de dados de produção

Muitas vezes é necessário denormalizar e ajustar os dados

Pode ter problemas com sincronismo de dados

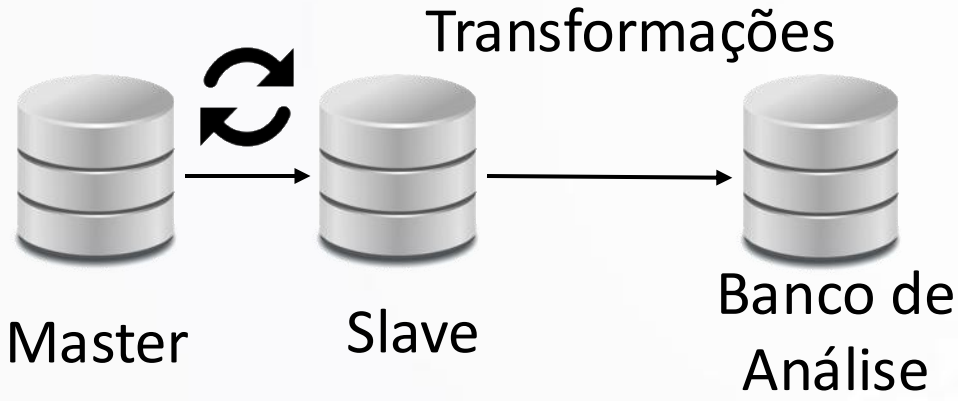
Arquitetura Master-Slave

Master é o banco de produção e Slave de leitura

Evita escrita e acessos indevidos

Evita sobrecarga de leitura no banco de produção

Configuração trivial da maioria dos bancos de dados



Mensageria

Aplicação ativamente envia os dados

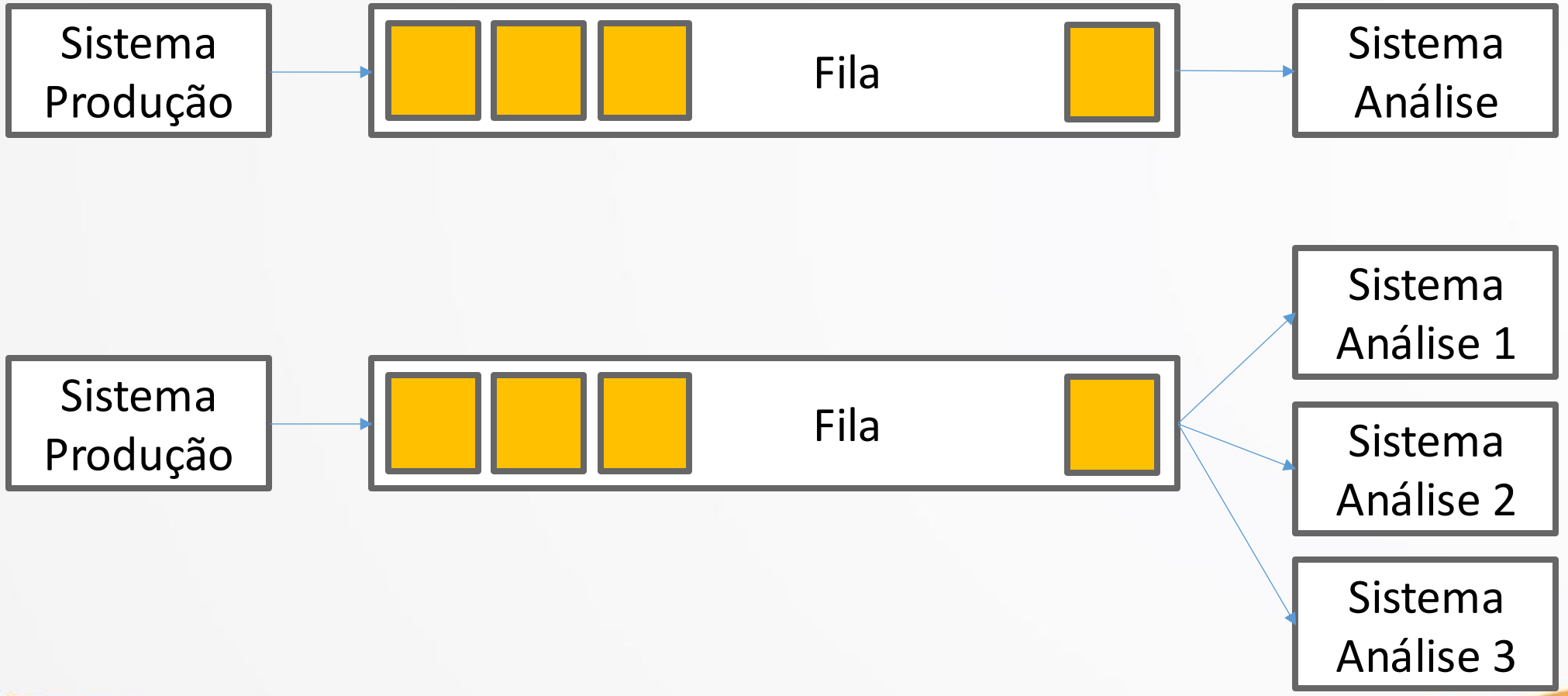
Sistemas de mensageria é responsável por armazenar e enviar ao sistema responsável

Permite fazer broadcast: Entrega para várias aplicações diferentes

Problemas: Sistema diferente para gerenciar

Vantagem: Facilita análises em tempo real

Mensageria


Sistema Produção
Fila
Sistema Análise
Sistema Produção
Fila
Sistema Análise 1
Sistema Análise 2
Sistema Análise 3

Serviços Comuns de Mensageria

amazon
SQS

RabbitMQ by Pivotal

kafka

ØMQ

Formas de Extração de Dados

Banco de Dados e Mensageria

Documentos

APIs

Scraping (Raspagem)

Documentos

Contratos, textos, formulários feitos em Computador

Formato de extração mais difícil. Praticamente cada documento é uma entidade única

Texto em linguagem natural precisa ser tratado e normalizado

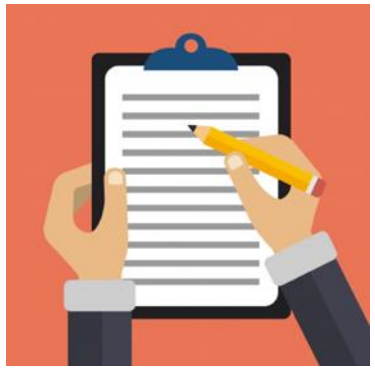
Encontrar padrões para extração de informação e APIs especializadas

Muitos dados podem não ser encontrados ou estarem incompletos

Formas de Extração de Dados

Banco de Dados e Mensageria

Documentos



APIs

Scraping (Raspagem)

Application Programming Interface (API)

Formato programático de acesso à aplicação

Formatos semiestruturados

Permite uma rápida e fácil comunicação Aplicação - Aplicação

Públicas ou Privadas

ÁREA DO DESENVOLVEDOR

Seja bem vindo a área exclusiva que a SPTrans criou para fornecer os dados de transporte público da cidade de São Paulo.

Se você se encaixa no perfil de desenvolvedor de aplicativos que busca oferecer as melhores alternativas para facilitar o uso do transporte público para todo cidadão este é o seu lugar.

// CONHEÇA O GTFS DA SPTRANS

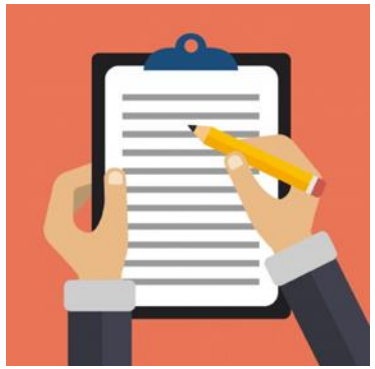
Conheça o GTFS da SPTrans e comece a desenvolver seus aplicativos.

CLIQUE PARA BAIXAR

Formas de Extração de Dados

Banco de Dados e Mensageria

Documentos



APIs

Scraping (Raspagem)

Scraping

Obtenção de dados de páginas web

O robô scraper navega pelas páginas ou faz uma busca específica

Captura de dados que são armazenados em uma base

Problemas legais: Pode não ser permitido a cópia de dados

Velocidade: Scraping da internet é um processo demorado

Tipos de Scraping

Baixo nível

Robô de scraping anda por texto apenas

Busca por tags HTML

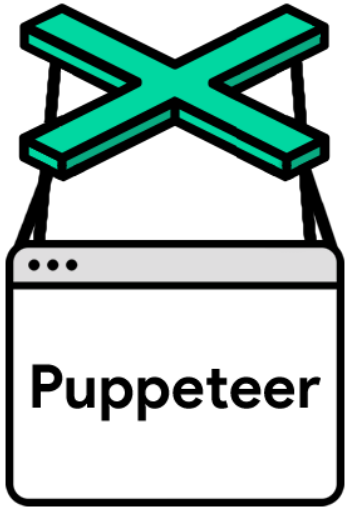
Busca links importantes para navegação

Baixo nível: envia comandos HTTP e trabalha com HTML. Difícil quando a página tem mais iterações com o usuário

Programação mais complexa, mas é veloz

Tipos de Scraping

Alto nível



Robô navega pelas páginas

Alto nível: Simula iteração do usuário: Clique, digita, espera, visualiza

Buscar por trechos, meta informações e recursos visuais da página

Alto nível: envia comandos HTTP e trabalha com HTML

Programação mais intuitiva, porém é mais lento

Muito Obrigado!

Diretoria de Pós-graduação e Pesquisa

Centro Universitário Senac

Prof. Celso Crivelaro

celso.vcrivelaro@sp.senac.br