

Escopo do Scraping PME

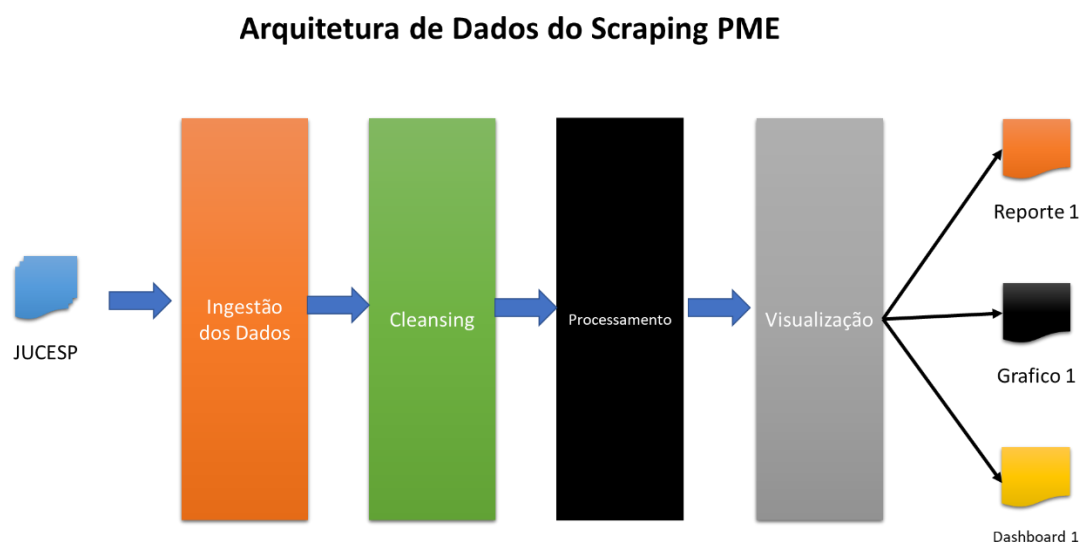
Objetivo: Temos como objetivo extrair informações de empresas do site da Jucesp para posteriormente, extrairmos as informações específicas de uma categoria de empresa e agruparmos as empresas com CNAE semelhantes.

Escopo: Serão consultadas as empresas de determinada cidade e as informações serão extraídas do site para persistir em um banco de dados (NEO4J ou MongoDB).

Deverão ter 3 estágios que serão implementados também no modelo de Dados.

- **Ingestão de dados:** Nesta etapa deverão ser carregados os dados capturados pelo WebScraping para dentro de uma entidade de entrada dos dados.
- **Cleansing de Dados:** Nesta etapa, os dados deverão ser tratados através de uma rotina de equalização dos dados, por exemplo, uma data como string ser transformada em Date. Dados incompletos deverão ser completados com informações padronizadas, por exemplo, se não tiver telefone, completar com (99) 99999-9999.
- **Processamento:** Nesta etapa os dados serão acomodados ou em um banco de dados estruturado ou em um banco não estruturado, o que for mais rápido e fácil de ser trabalhado posteriormente.
- **Visualização:** Nesta etapa, deixaremos Data Marts prontos para consulta por cidade, CNAE ou objeto social, telefones, e-mails, pois é comum as contabilidades colocarem os e-mails e celulares dos próprios escritórios de contabilidade, pois neste caso deveremos realizar busca de informações do verdadeiro proprietário da empresa.

Conforme figura abaixo:



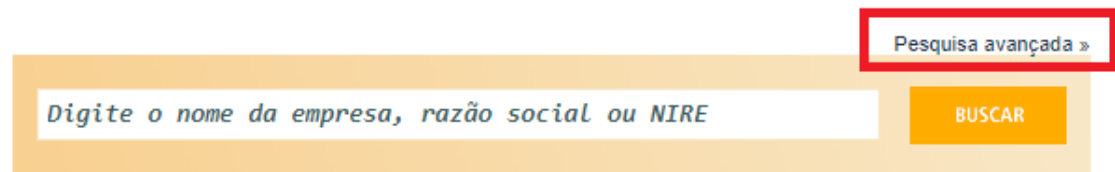
Descrição da Solução

- **Ingestão dos Dados**

O webscraping deverá entrar no site na Jucesp (<https://www.jucesponline.sp.gov.br/>) e clicar sobre o Pesquisa Avançada.

Pesquisar Empresas

Pesquisa no banco de dados da Junta Comercial do Estado de São Paulo.



Pesquisa avançada »

Digite o nome da empresa, razão social ou NIRE

BUSCAR

Para ler os documentos digitais você precisa do leitor de PDF [instalar](#)

Após clicar em Pesquisa Avançada, o campo de Município (único campo obrigatório do formulário) deverá ser preenchido com a palavra Santos (inicialmente, trabalharemos com a cidade de Santos, depois faremos a digitação das outras cidades da Baixada Santista e Cubatão.



Endereço

Logradouro:

CEP:

Bairro:

Município*:

UF: SP ▾

Arquivamentos

Após a digitação do município e teclar Enter (Submit), surgirá um Captcha que deverá ser reconhecido, convertido e preenchido automaticamente pelo robô.

Pesquisar Empresas

Pesquisa no banco de dados da Junta Comercial do Estado de São Paulo.

Pesquisa avançada »

Digite o nome da empresa, razão social ou NIRE

BUSCAR

Para ler os documentos digitais você precisa do leitor de PDF [instalar](#)



Digite o código da imagem

CONTINUAR

Após o preenchimento e teclar Enter (Submit), será carregada a página com os NIREs (Identificador Único), nome da Empresa e Município.

Esses dados serão persistidos no banco de dados para utilização posterior e utilização para consultar o detalhamento.

Nesse momento acontece a primeira carga de dados na estrutura de tabelas da etapa de Ingestão de Dados.

A tabela que acomodará esses dados também deverá conter um atributo se já foram carregados os dados detalhados ou não da empresa. Pois, caso tenha sido carregado os dados detalhados, eles não precisam ser incluídos novamente e deverá ser enviado para uma tabela de expurgo.

No futuro, faremos uma comparação se os dados estão atualizados ou não no estágio de Processamento.

Pesquisar Empresas

Pesquisa no banco de dados da Junta Comercial do Estado de São Paulo.

Pesquisa avançada »

Digite o nome da empresa, razão social ou NIRE

BUSCAR

Para ler os documentos digitais você precisa do leitor de PDF [instalar](#)

Resultados 1 - 15 de 66.939 para a busca avançada (0,0940 segundos)

NIRE	Empresa	Município
35232023807	TRANSVIEIRA TRANSPORTES E LOGISTICAS LTDA	SANTOS
35232146623	AURESEG MANUTENCAO INDUSTRIAL LTDA	SANTOS
35232146721	OLIVEIRA & EIZO SERVICOS DE MOTOBOY LTDA	SANTOS
35232146739	M C GODOI SEMIJOIAS LTDA	SANTOS
35236231706	RADAR TELECOMUNICACOES INOVA LTDA	SANTOS
35232146640	CARGOTEX INSPECOES E SERVICOS LTDA	SANTOS
35232146674	JSA - COMERCIAL IMPORTADORA LTDA	SANTOS
35236235035	BB HOLDING GESTAO E PARTICIPACOES LTDA	SANTOS
35232146644	EMPRESA CEREALISTA ATACADO E VAREJO DE ALIMENTOS E LATICINIOS	SANTOS

Também deverá ter controle sobre as páginas para que todas as informações sejam extraídas adequadamente.

Após a obtenção dos NIREs, o sistema deverá percorrer a tabela de captura de NIREs e realizar a busca das informações detalhadas de cada empresa na URL:

<https://www.jucesponline.sp.gov.br/>.

Deverá ser preenchido o campo de pesquisa e submetido para pesquisa.

desa

Pesquisar Empresas

Pesquisa no banco de dados da Junta Comercial do Estado de São Paulo.

Pesquisa avançada »

Digite o nome da empresa, razão social ou NIRE

BUSCAR

Para ler os documentos digitais você precisa do leitor de PDF [instalar](#)

A Busca trará os detalhes do NIRE pesquisado.

Data de emissão: 13/08/2020 10:48:24

SECRETARIA DO ESTADO DE SÃO PAULO

TRANSVIEIRA TRANSPORTES E LOGISTICAS LTDA

Nire Matriz

35232023807

Tipo de Empresa

SOCIEDADE LIMITADA



[Localizar no Mapa](#)

Data da constituição

12/08/2020

Início de atividade

02/07/2020

CNPJ

38.067.081/0001-10

Inscrição Estadual

Objeto

Transporte rodoviário de carga, exceto produtos perigosos e mudanças, intermunicipal, interestadual e internacional
Transporte rodoviário de carga, exceto produtos perigosos e mudanças, municipal

Capital

R\$ 10.000,00 (Dez Mil Reais)

Logradouro

Avenida Almirante Cochrane

Número

151

Bairro

Embare

Complemento

2 Andar

Município

Santos

CEP

11040-001

UF

SP

Os campos deverão ser criados como atributos em uma tabela no banco de dados e não haverá necessidade de relacionamento, mas o NIRE pesquisado deverá ser guardado nessa tabela também.

Ainda deverá ser realizada a Pesquisa Simples para abrir um arquivo no formato PDF para obtenção dos nomes dos sócios.

Selecione o documento ou o serviço desejado

- ☐ Ficha Cadastral Completa (dados a partir de 1992)
- ☒ **Ficha Cadastral Simplificada (dados atuais da empresa)**
- ☐ Cópia Digitalizada de Documentos Arquivados (cópia simples - não tem valor jurídico de certidão)
- ☐ Certidão Simplificada
- ☐ Certidão Específica Pré-formatada
- ☐ Certidão Específica com Teor Solicitado
- ☐ Certidão Específica com Teor Solicitado - Registro de Livros
- ☐ Certidão de Inteiro Teor
- ☐ Solicitação de Correção de Dados Cadastrais

OK

Abrirá o arquivo PDF, que deverá ser realizado o download para posterior extração dos nomes e dados dos sócios, conforme imagem abaixo:

OBJETO SOCIAL
TRANSPORTE RODOVIÁRIO DE CARGA, EXCETO PRODUTOS PERIGOSOS E MUDANÇAS, INTERMUNICIPAL, INTERESTADUAL E INTERNACIONAL
TRANSPORTE RODOVIÁRIO DE CARGA, EXCETO PRODUTOS PERIGOSOS E MUDANÇAS, MUNICIPAL

TITULAR / SÓCIOS / DIRETORIA
ANTONIO JOSE VIEIRA, CUTIS: NÃO INF., NACIONALIDADE BRASILEIRA, CPF: 044.531.408-78, RG/RNE: 22591035 - SP, RESIDENTE À AVENIDA ALMIRANTE COCHRANE, 83, APT. 131, EMBARE, SANTOS - SP, CEP 11040-001, NA SITUAÇÃO DE SÓCIO. COM VALOR DE PARTICIPAÇÃO NA SOCIEDADE DE \$ 3.000,00
CARLA DE CASSIA GONCALVES MACHADO, CUTIS: NÃO INF., NACIONALIDADE BRASILEIRA, CPF: 108.284.888-35, RG/RNE: 189378347 - SP, RESIDENTE À RUA RICARDO PINTO DE OLIVEIRA, 07, AREIA BRANCA, SANTOS - SP, CEP 11086-120, NA SITUAÇÃO DE SÓCIO E ADMINISTRADOR, ASSINANDO PELA EMPRESA. COM VALOR DE PARTICIPAÇÃO NA SOCIEDADE DE \$ 3.000,00.
CHRISTIAN DA SILVA LOPES, CUTIS: NÃO INF., NACIONALIDADE BRASILEIRA, CPF: 169.541.888-32, RG/RNE: 372090965 - SP, RESIDENTE À AVENIDA REI ALBERTO I, 275, CASA, PONTA DA PRAIA, SANTOS - SP, CEP 11030-381, NA SITUAÇÃO DE SÓCIO. COM VALOR DE PARTICIPAÇÃO NA SOCIEDADE DE \$ 4.000,00

Documento Gratuito
Proibida a Comercialização

Página 1 de 2


• Cleansing

Nesta etapa do processo, os campos quando estiverem vazios, deverão ser preenchidos com o valor de "NA" para campos de caracteres e "0,00" para campos numéricos.

O único campo que não poderá ser preenchido será o "NIRE", pois trata-se de um identificador único.

• Processamento

Nesta etapa, cada parágrafo do campo "Objeto" deverá ser desmembrado em atributos na tabela de dados, como abaixo:

	Objeto
	Transporte rodoviário de carga, exceto produtos perigosos e mudanças, intermunicipal, interestadual e internacional
	Transporte rodoviário de carga, exceto produtos perigosos e mudanças, municipal
	Capital
	R\$ 10 000 00 (Dez Mil Reais)

Serão desmembrados em atividade principal e atividade secundária 1, atividade secundária 2, atividade secundária 3, em diante, até a atividade secundária 10, sendo 11 campos o limitador (1 atividade principal + 10 atividades secundárias).

Os dados deverão ser persistidos na estrutura de dados final para montagem dos Data Marts, contendo as visões prontas para consulta pelos usuários.

- **Visualização**

<Detalhar a visualização>