

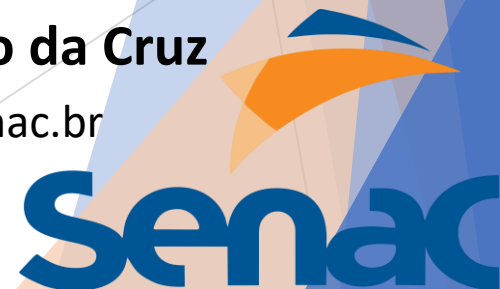
Sistemas de Apoio às Decisões

Aula 06 – Introdução à Engenharia de Dados I

Prof. Esp. Guilherme Jorge Aragão da Cruz

 guilherme.jacruz@sp.senac.br

 linkedin.com/in/guijac

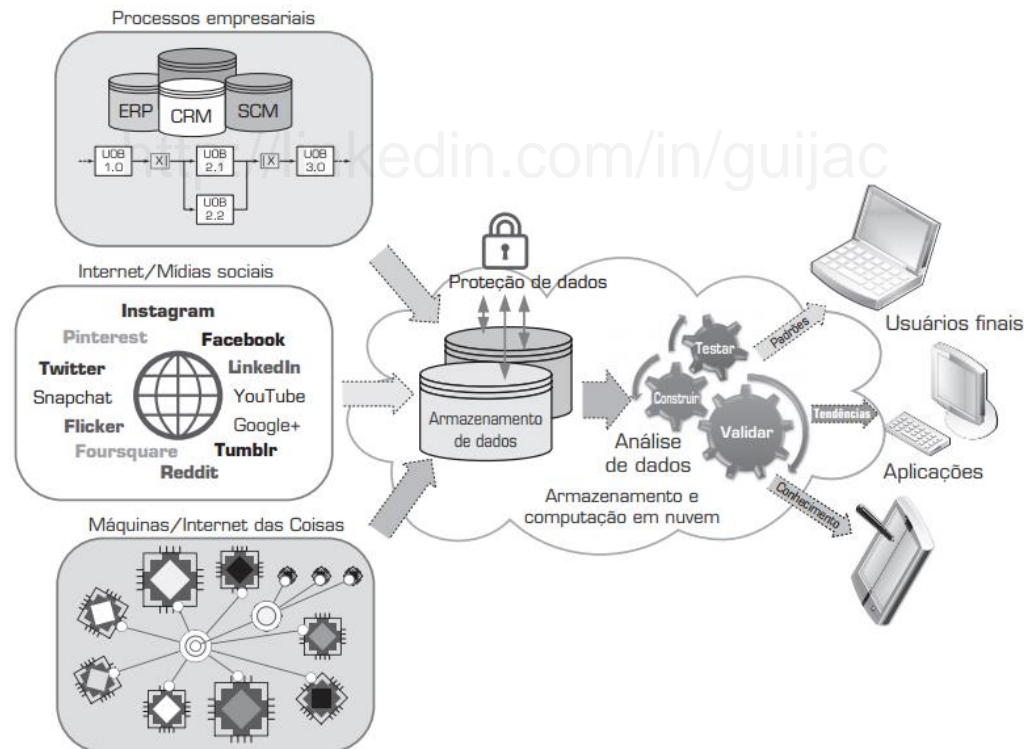


Roteiro

- A Natureza dos Dados;
- Uma Simples Taxonomia dos Dados;
- O Pré-processamento de Dados;
- Tarefas e Métodos em Pré-processamento de Dados;
- O Processo de Extração, Transformação e Carga (ETL);
- Principais Ferramentas ETL;
- Laboratório;
- Referências Bibliográficas.

A Natureza dos Dados

- Dados podem consistir em números, letras, palavras, imagens, gravações de voz e assim por diante;
- Nível mais fundamental de abstração de onde pode-se derivar informações e, então, conhecimento;
- Qualidade dos dados e sua integridade → crucial para a análise de dados



Fonte: SHARDA, DELEN, TURBAN (2019)

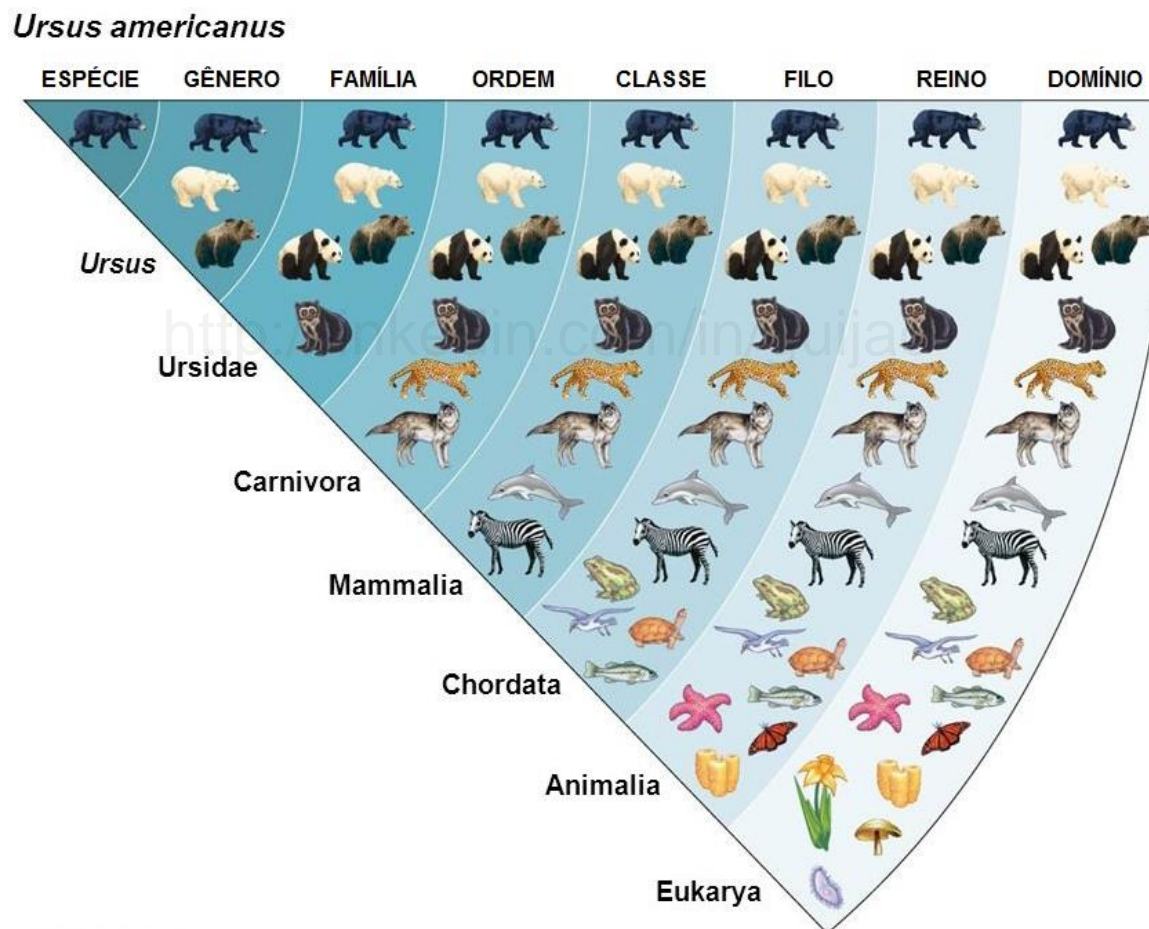
Uma Simples Taxonomia dos Dados

- Taxonomia? 🤔

<http://linkedin.com/in/guijac>

Uma Simples Taxonomia dos Dados

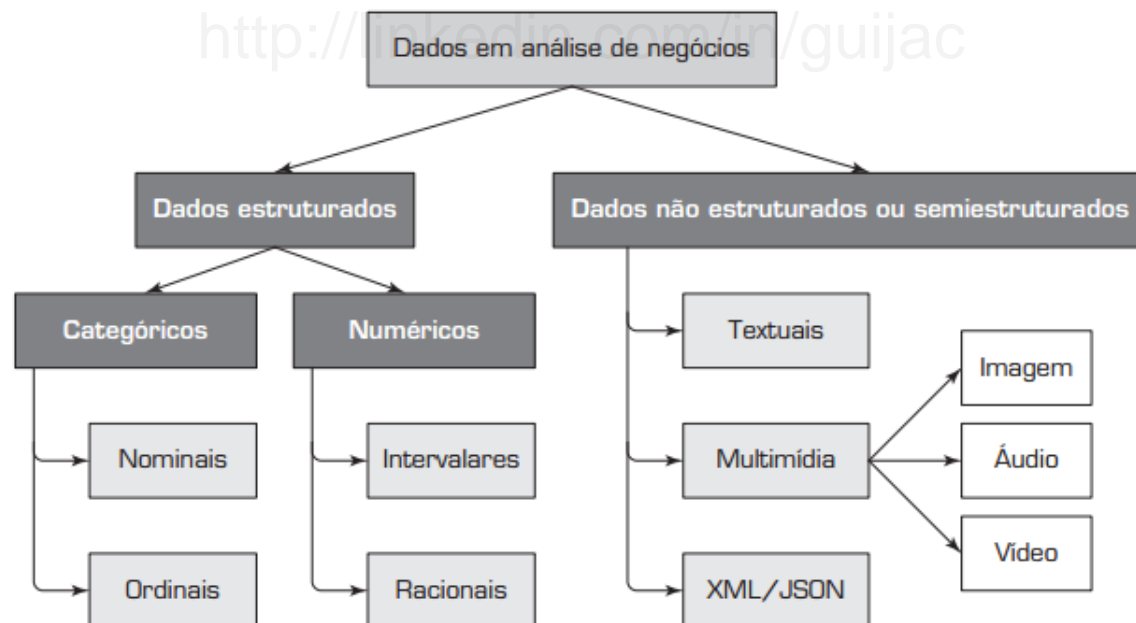
- Taxonomia? 🤔



© 2014 Pearson Education, Inc.

Uma Simples Taxonomia dos Dados

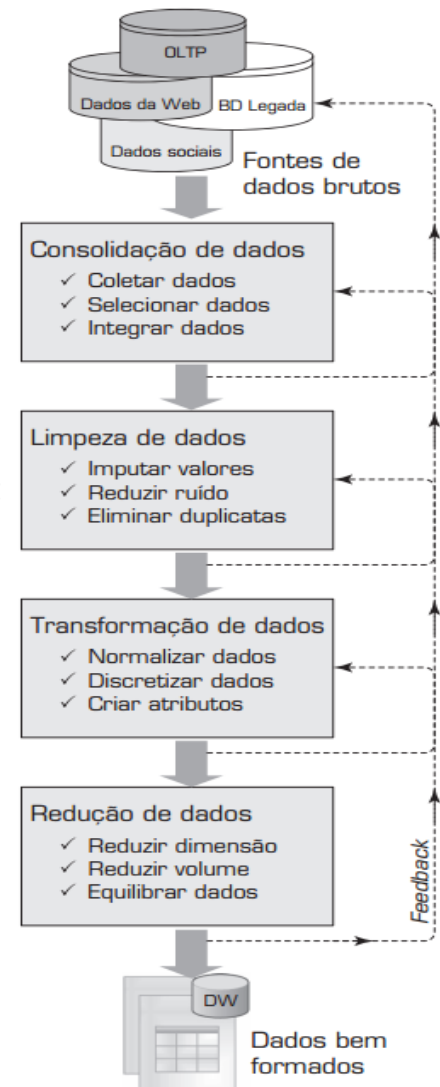
- **Dados estruturados:**
 - Bem organizados para serem processados por computador
- **Dados semiestruturados:**
 - Com identificadores padronizados, para facilitar o processamento por computador.
- **Dados não estruturados:**
 - De maior desafio técnico para serem processados e interpretados por computador.



Fonte: SHARDA, DELEN, TURBAN (2019)

O Pré-processamento de Dados

- No mundo real, os dados costumam se apresentar com problemas variados, mal-formatados, excessivamente complexos e imprecisos
- Não costumam estar prontos para serem usados em tarefas de análise de dados
- É necessário preparar os dados para análise:
 - Consolidação de dados;
 - Limpeza de dados;
 - Transformação de dados;
 - Redução de dados.



Fonte: SHARDA, DELEN, TURBAN (2019)

Tarefas e Métodos em Pré-processamento de Dados

| Tarefa Principal | Subtarefas | Métodos Populares |
|-----------------------|---------------------------------------|---|
| Consolidação de dados | Acesse e colete dados. | Consultas SQL, agentes de software, serviços Web. |
| | Selecione e filtre os dados. | Especialização na área, consultas SQL, testes estatísticos. |
| | Integre e unifique os dados. | Consultas SQL, especialização na área, mapeamento de dados embasado em ontologia. |
| Limpeza de dados | Corrija valores ausentes nos dados. | Preencher valores ausentes com os valores mais apropriados (média, mediana, min/máx, etc); recodificar os valores ausentes com uma constante; remover o registro do valor ausente; deixar como está. |
| | Identifique e reduza ruído nos dados. | Identificar os valores discrepantes nos dados com técnicas estatísticas simples (como médias e desvios padrão) ou com análise de agrupamento; depois de identificados, ou remover os valores discrepantes ou suavizá-los. |

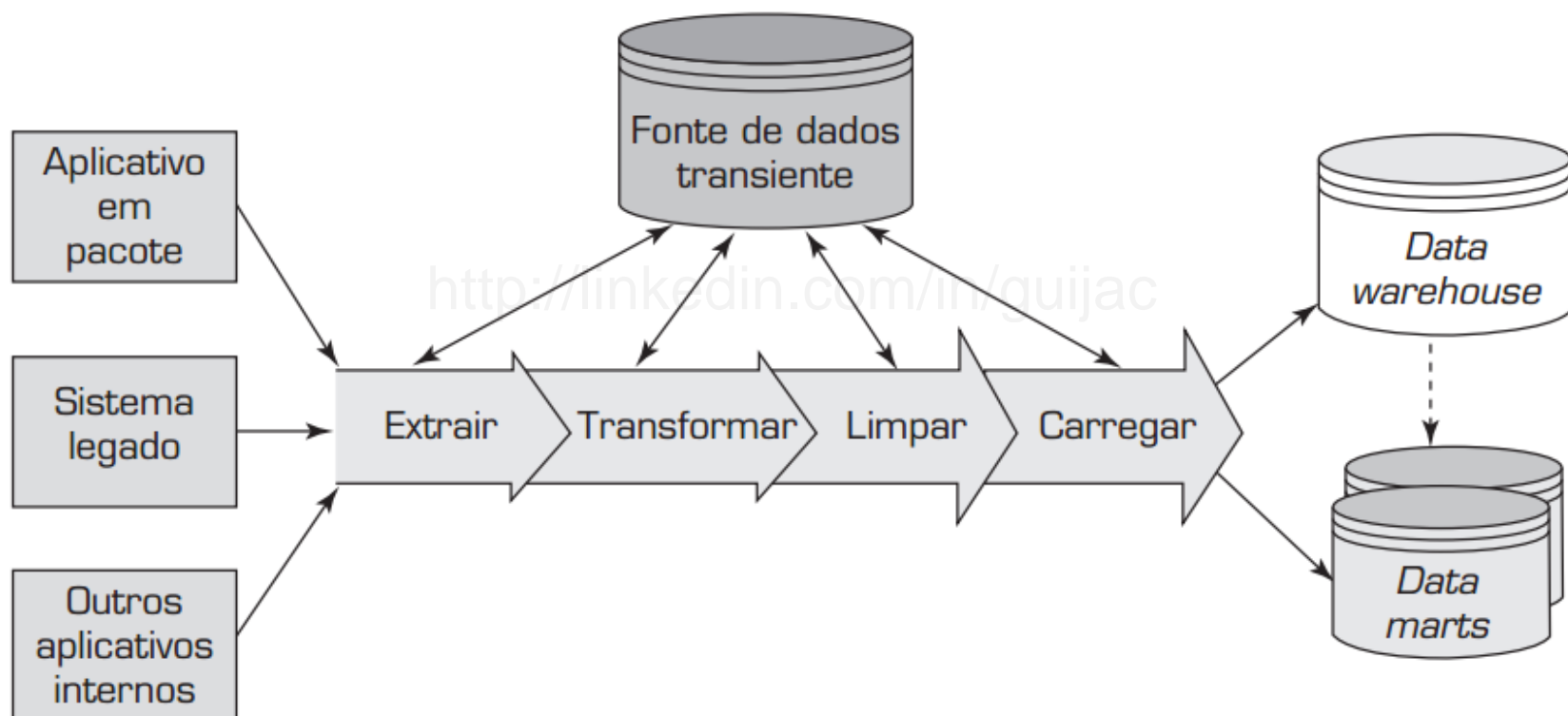
Fonte: Adaptado de SHARDA, DELEN, TURBAN (2019)

Tarefas e Métodos em Pré-processamento de Dados

| Tarefa Principal | Subtarefas | Métodos Populares |
|------------------------|------------------------------------|--|
| Limpeza de dados | Encontre e elimine dados errôneos. | Identificar os valores errôneos nos dados (além das discrepâncias), tais como valores estranhos, designações de classe inconsistentes, distribuições esquisitas; depois de identificados, aplicar especialização na área para corrigir os valores ou remover os registros. |
| Transformação de dados | Normalize os dados. | Reduzir a amplitude de valores em cada variável numérica para uma amplitude-padrão (como de 0 a 1 ou de -1 a $+1$) usando uma variedade de técnicas de normalização ou escala; |
| | Discretize ou agregue os dados. | Caso necessário, converter as variáveis numéricas em representações discretas usando técnicas de segmentação baseadas em amplitude ou frequência; no caso de variáveis categóricas, reduzir a quantidade de valores aplicando hierarquias conceituais apropriadas. |

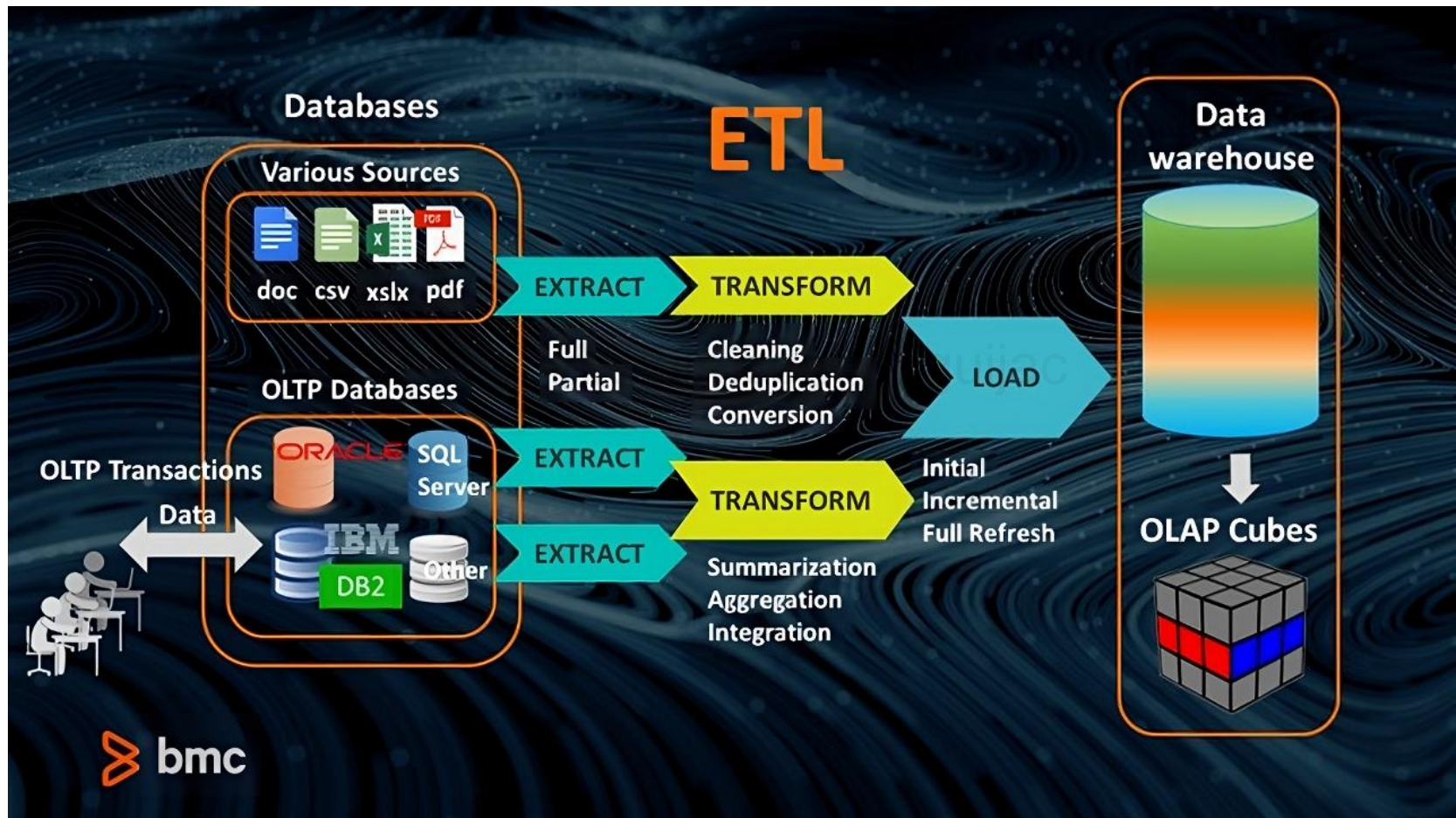
Fonte: Adaptado de SHARDA, DELEN, TURBAN (2019)

O Processo de Extração, Transformação e Carga (ETL)



Fonte: SHARDA, DELEN, TURBAN (2019)

O Processo de Extração, Transformação e Carga (ETL)



Fonte: [Is ETL \(Extract, Transform, Load\) Still Relevant? – BMC Software | Blogs](#)

Principais Ferramentas ETL



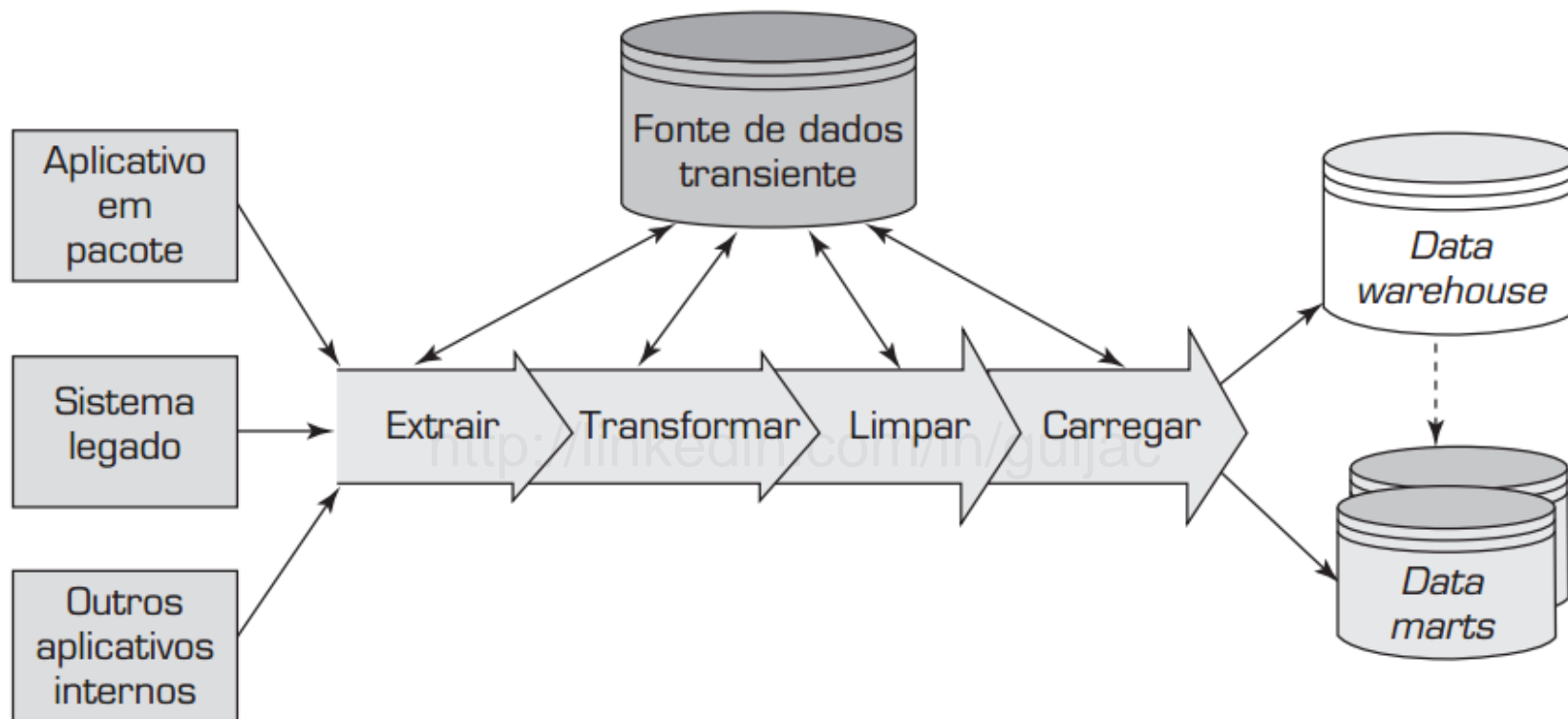
Figure 1: Magic Quadrant for Data Integration Tools¹



Fonte: [2023 Gartner® Magic Quadrant™ for Data Integration Tools](#)

¹ [O que é o Quadrante mágico Gartner e qual a aplicabilidade](#)

Por hoje (de teoria) é só!



Fonte: SHARDA, DELEN, TURBAN (2019)

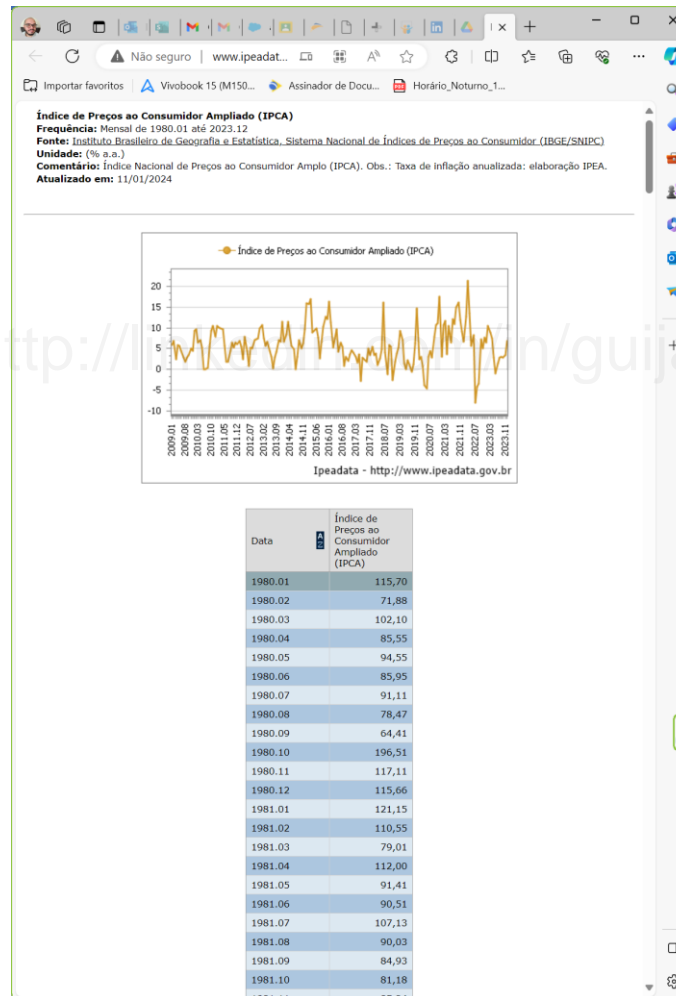
Prof. Esp. Guilherme Jorge Aragão da Cruz

✉ guilherme.jacruz@sp.senac.br

in linkedin.com/in/guijac

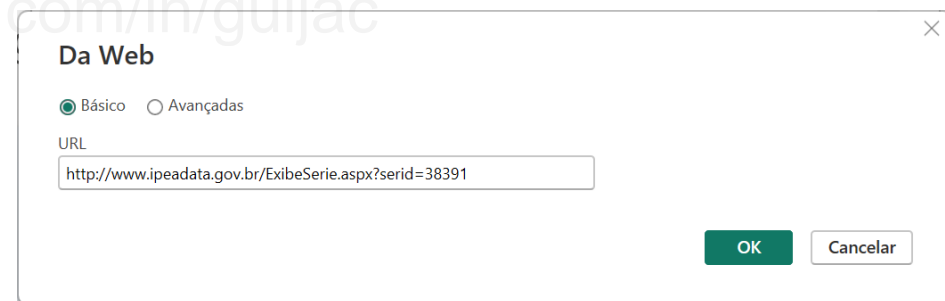
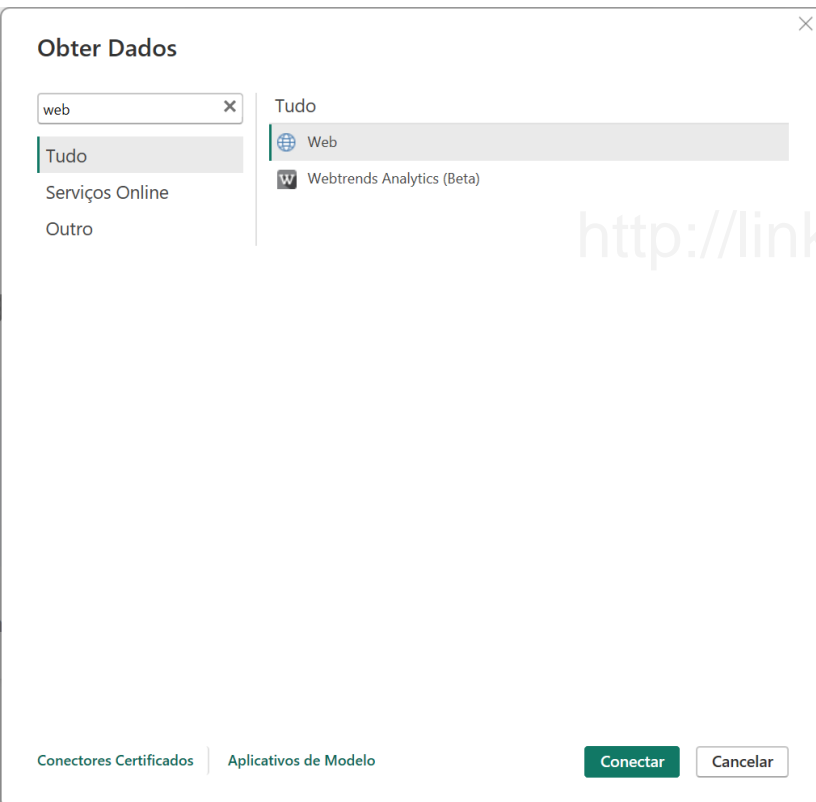
Laboratório

- Extração, Transformação e Carga da série histórica do Índice de Preços ao Consumidor Ampliado (IPCA) através do site [Ipeadata](http://www.ipeadata.gov.br) com Power BI.



Laboratório

- Dentro do Power BI, obtenha os dados de uma fonte “Web” e insira a URL do site do IPEA Data, clique em “OK” e aguarde a leitura do site:



Fonte: Elaboração própria.

Laboratório

- O próprio Power BI irá identificar as estruturas da página HTML, como tabelas e textos;
- Identifique e selecione a tabela que contém os dados da série histórica.

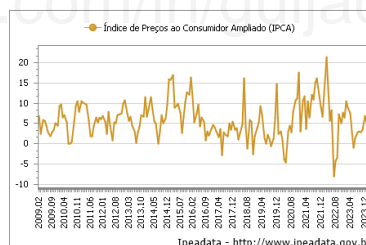
Navegador

Opções de Exibição ▾

- Tabelas HTML [7]
 - ☐ Tabela 1
 - ☐ Tabela 2
 - ☐ Tabela 3
 - ☐ Tabela 4
 - ☐ Tabela 5
 - ☐ Tabela 6
 - ☐ Tabela 7
- Texto [2]
 - ☐ Código HTML
 - ☐ Texto Exibido

Exibição de Tabela | Exibição da Web

Índice de Preços ao Consumidor Ampliado (IPCA)
Frequência: Mensal de 1980.01 até 2024.01
Fonte: Instituto Brasileiro de Geografia e Estatística, Sistema Nacional de Índices de Preços ao Consumidor (IBGE/SNIPC)
Unidade: (% a.a.)
Comentário: Índice Nacional de Preços ao Consumidor Amplo (IPCA). Obs.: Taxa de inflação anualizada: elaboração IPEA.
Atualizado em: 08/02/2024



Ipeadata - <http://www.ipeadata.gov.br>

| Data | Índice de Preços ao Consumidor Ampliado (IPCA) |
|---------|--|
| 1980.01 | 115,70 |
| 1980.02 | 71,88 |
| 1980.03 | 102,10 |
| 1980.04 | 85,55 |
| 1980.05 | 94,55 |
| 1980.06 | 85,95 |
| 1980.07 | 91,11 |
| 1980.08 | 78,47 |
| 1980.09 | 64,41 |

Adicionar a Tabela Usando Exemplos

Carregar Transformar Dados Cancelar

Fonte: Elaboração própria.

Laboratório

- Observe o formato dos atributos da coluna “Data”;
- Será necessário **transformar** este campo para um tipo “data” (Date.Type);
- O tipo pode ser alterado inserindo uma nova etapa no Power Query, identifique uma **substituição** que permita esta transformação

Consultas [1]

Tabela 4

| | Data | Índice de Preços ao Consumidor Ampliado (IPCA) |
|----|------------|--|
| 1 | 01/01/1980 | 115,7 |
| 2 | 01/02/1980 | 71,88 |
| 3 | 01/03/1980 | 102,1 |
| 4 | 01/04/1980 | 85,55 |
| 5 | 01/05/1980 | 94,55 |
| 6 | 01/06/1980 | 85,95 |
| 7 | 01/07/1980 | 91,11 |
| 8 | 01/08/1980 | 78,47 |
| 9 | 01/09/1980 | 64,41 |
| 10 | 01/10/1980 | 196,51 |
| 11 | 01/11/1980 | 117,11 |
| 12 | 01/12/1980 | 115,66 |
| 13 | 01/01/1981 | 121,15 |
| 14 | 01/02/1981 | 110,55 |
| 15 | 01/03/1981 | 79,01 |
| 16 | 01/04/1981 | 112 |

Config. Consulta

PROPRIEDADES

Nome: Tabela 4

Todas as Propriedades

ETAPAS APLICADAS

- Fonte
- Tabela extraída de HTML
- Cabeçalhos Promovidos
- Valor Substituído
- Tipo Alterado

Fonte: Elaboração própria.

Laboratório

- Finalizada a transformação, clique em “Fechar e Aplicar” e crie os gráficos conforme sua necessidade e filtros.

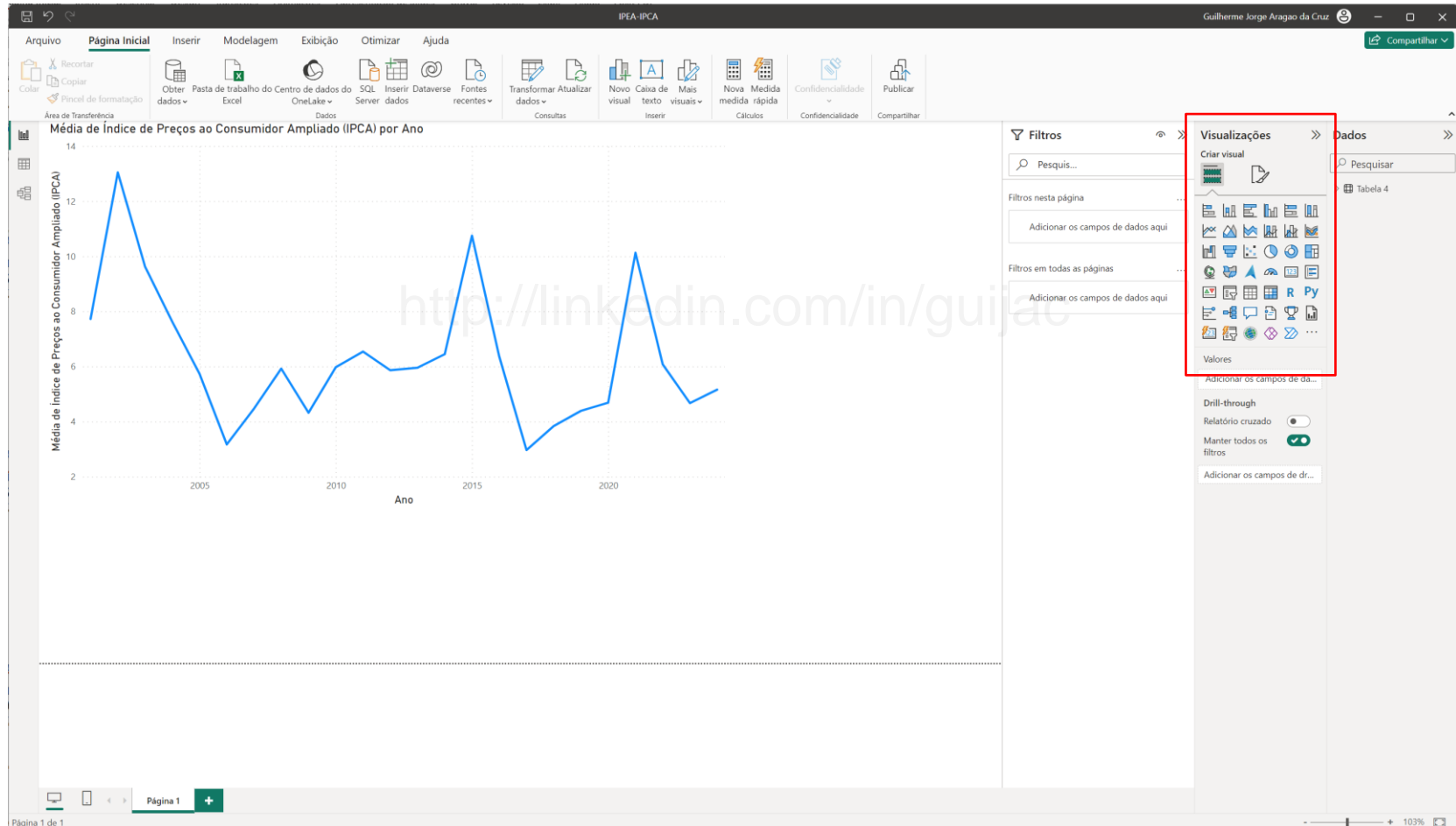
2 COLUNAS, 529 LINHAS Criação de perfil de coluna com base nas primeiras 1000 linhas

VISUALIZAÇÃO BAIXADA À(S) 16:40

Fonte: Elaboração própria.

Laboratório

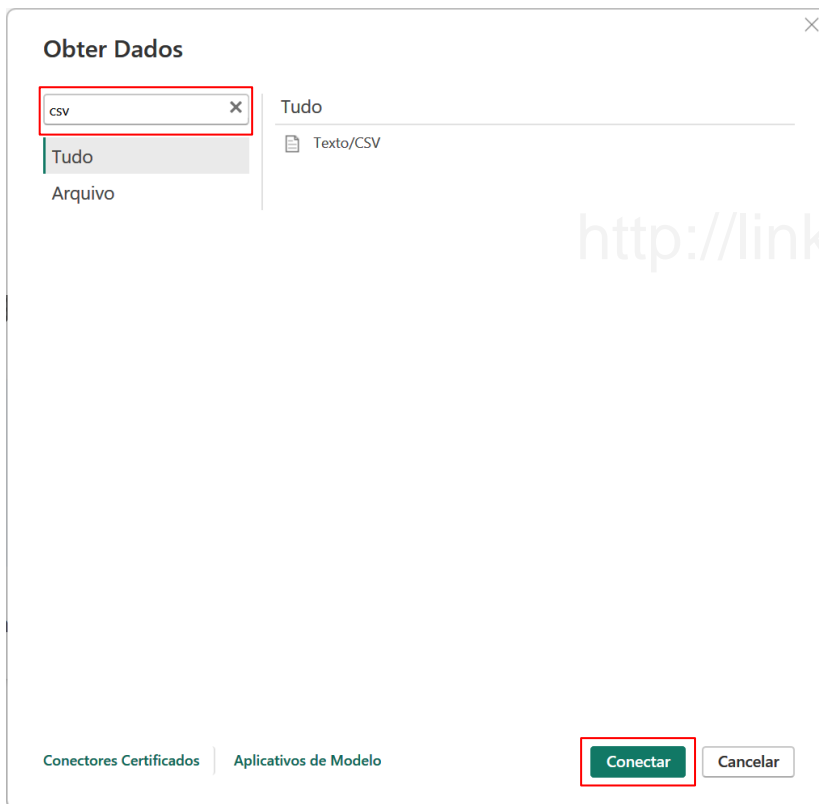
- Finalizada a transformação, clique em “Fechar e Aplicar” e crie os gráficos conforme sua necessidade e filtros.



Fonte: Elaboração própria.

Laboratório

- Agora faça o mesmo procedimento para a sua base de dados e avalie quais serão as transformações necessárias para trabalhar com o Power BI.



archinsurance-base-new.csv

Origem do Arquivo: 1252: Europeu Ocidental (Windows) Delimitador: Vírgula Detecção de Tipo de Dados: Com base nas primeiras 200 linhas

| customer-id | date | premium-paid | company | country | channel |
|-------------|------------|--------------|-------------|--------------|---------|
| 109 | 09/02/2019 | 19105 | Home & Away | Chile | site |
| 601 | 22/05/2019 | 16086 | Home & Away | South Africa | site |
| 364 | 21/10/2019 | 19461 | Home & Away | Argentina | site |
| 954 | 18/09/2019 | 11614 | Home & Away | Brazil | site |
| 71 | 21/05/2019 | 19419 | Home & Away | South Africa | site |
| 911 | 23/10/2019 | 10633 | Home & Away | Chile | site |
| 465 | 08/02/2019 | 17058 | Home & Away | South Africa | site |
| 438 | 17/01/2019 | 13461 | Home & Away | Argentina | site |
| 890 | 26/10/2019 | 11634 | Home & Away | Peru | site |
| 35 | 29/08/2019 | 6022 | Home & Away | Argentina | site |
| 610 | 07/08/2019 | 7034 | Home & Away | Argentina | site |
| 780 | 13/07/2019 | 17132 | Home & Away | Peru | site |
| 6 | 28/06/2019 | 15251 | Home & Away | Brazil | site |
| 948 | 25/06/2019 | 19948 | Home & Away | Chile | site |
| 1000 | 05/12/2019 | 112 | Home & Away | Brazil | site |
| 264 | 06/03/2019 | 18536 | Home & Away | Brazil | site |
| 818 | 30/03/2019 | 9947 | Home & Away | Mexico | site |
| 682 | 14/09/2019 | 16361 | Home & Away | Peru | site |
| 294 | 30/05/2019 | 7536 | Home & Away | Peru | site |
| 638 | 23/08/2019 | 7548 | Home & Away | South Africa | site |

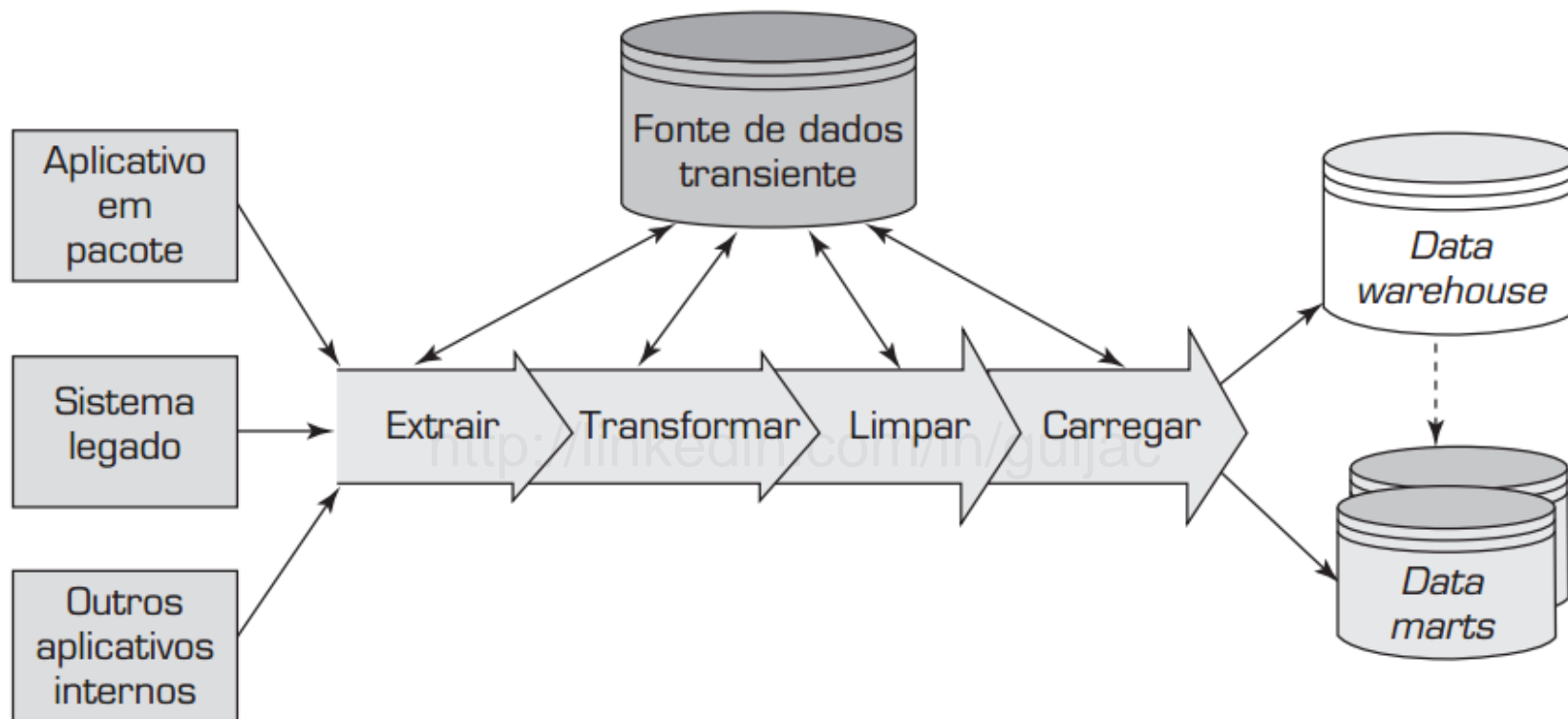
Os dados na visualização foram truncados devido ao limite de tamanho.

Extrair a Tabela Usando Exemplos

Carregar **Transformar Dados** Cancelar

Fonte: Elaboração própria.

Por hoje (agora sim) é só!



Fonte: SHARDA, DELEN, TURBAN (2019)

Prof. Esp. Guilherme Jorge Aragão da Cruz

 guilherme.jacruz@sp.senac.br

 linkedin.com/in/guijac

Referências Bibliográficas

- LAUDON, K. C.; LAUDON, J. P. **Sistemas De Informações Gerenciais**. 17. ed. Porto Alegre: Bookman, 2019;
- MICROSOFT LEARN. **Documentação do Power Query**. Disponível em <https://learn.microsoft.com/pt-br/power-query/>. Acesso em 06 fev 2024;
- SHARDA, R. ; DELEN, D. ; TURBAN, E. **Business intelligence e análise de dados para gestão do negócio**. 4. ed. Porto Alegre: Bookman, 2019.