



UNIFESSPA

UNIVERSIDADE FEDERAL DO SUL E SUDESTE DO PARÁ

Aquisição e Preparação de Dados



Cláudio Roberto de Lima Martins

Msc Computação / Prof. IFPA

claudiomartins2000@gmail.com

Treinamento: Aquisição e preparação de dados

■ Objetivo

- Apresentar os problemas mais comuns encontrados na aquisição e preparação de bases de dados para uso acadêmico, e as técnicas usadas para resolvê-los.
- O enfoque é prático e utilizará exemplos e dados envolvidos no projeto LCC, e bases de origem em fontes oficiais (governamentais) e abertas.

Motivação

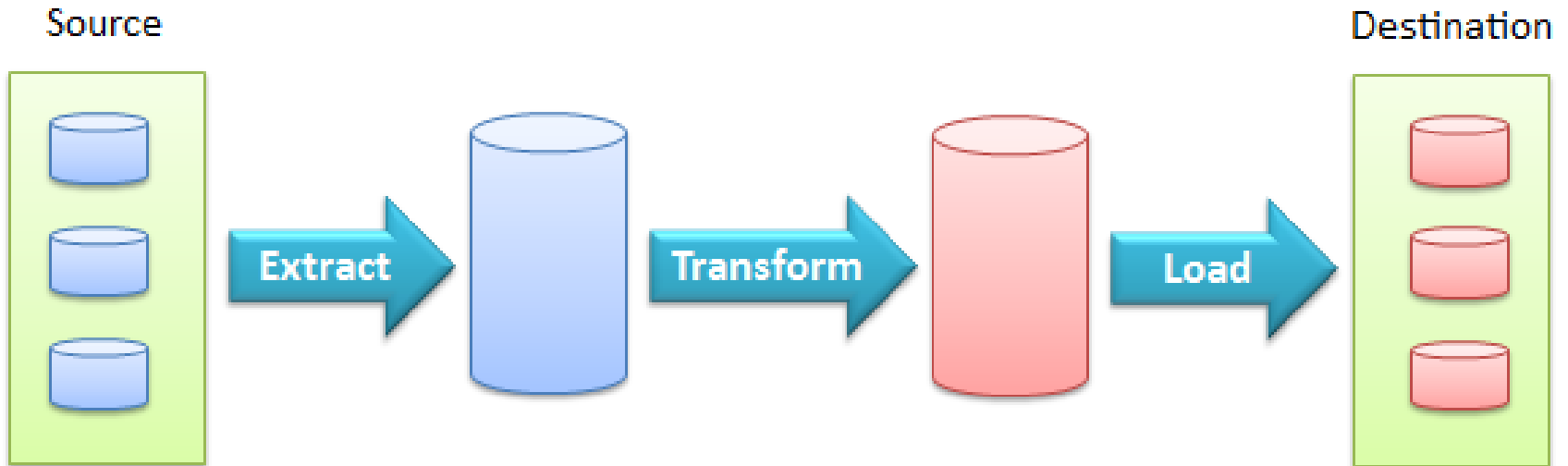
- As atividades de aquisição e preparação de dados fazem parte de um processo chamado de **ETL**.
 - **ETL** (do inglês **Extract, Transform, Load**) = Extração, Transformação, Carga.
 - O ETL visa trabalhar com toda a parte de extração de dados de fontes externas, transformação para atender às necessidades do problema (negócio) e carga dos dados dentro de um banco de dados alvo (Data Warehouse, p.ex).
- O ETL pode ser usado em:
 - Mineração de dados, visando a descoberta de conhecimento (KDD)
 - Na etapa inicial da montagem de um banco de dados para fins específicos (em um sistema de informação), por exemplo:
 - *Data warehousing* (DW)
 - *On-Line Analytical Processing* (OLAP).

ETL – Extract Transform Load

- ETL é um processo de extração de dados a partir de fontes diversas, e carga em uma base de dados destino usando um conjunto de funções de transformação.
 - **Extração:** Responsável pela coleta de dados (source ou origem) e transferência destes dados para o ambiente de pré-processamento.
 - **Transformação:** Responsável em realizar os devidos ajustes, podendo melhorar a qualidade dos dados e consolidar dados de duas ou mais fontes.
 - **Carga:** Responsável por estruturar e carregar os dados para um banco de dados alvo (destino).

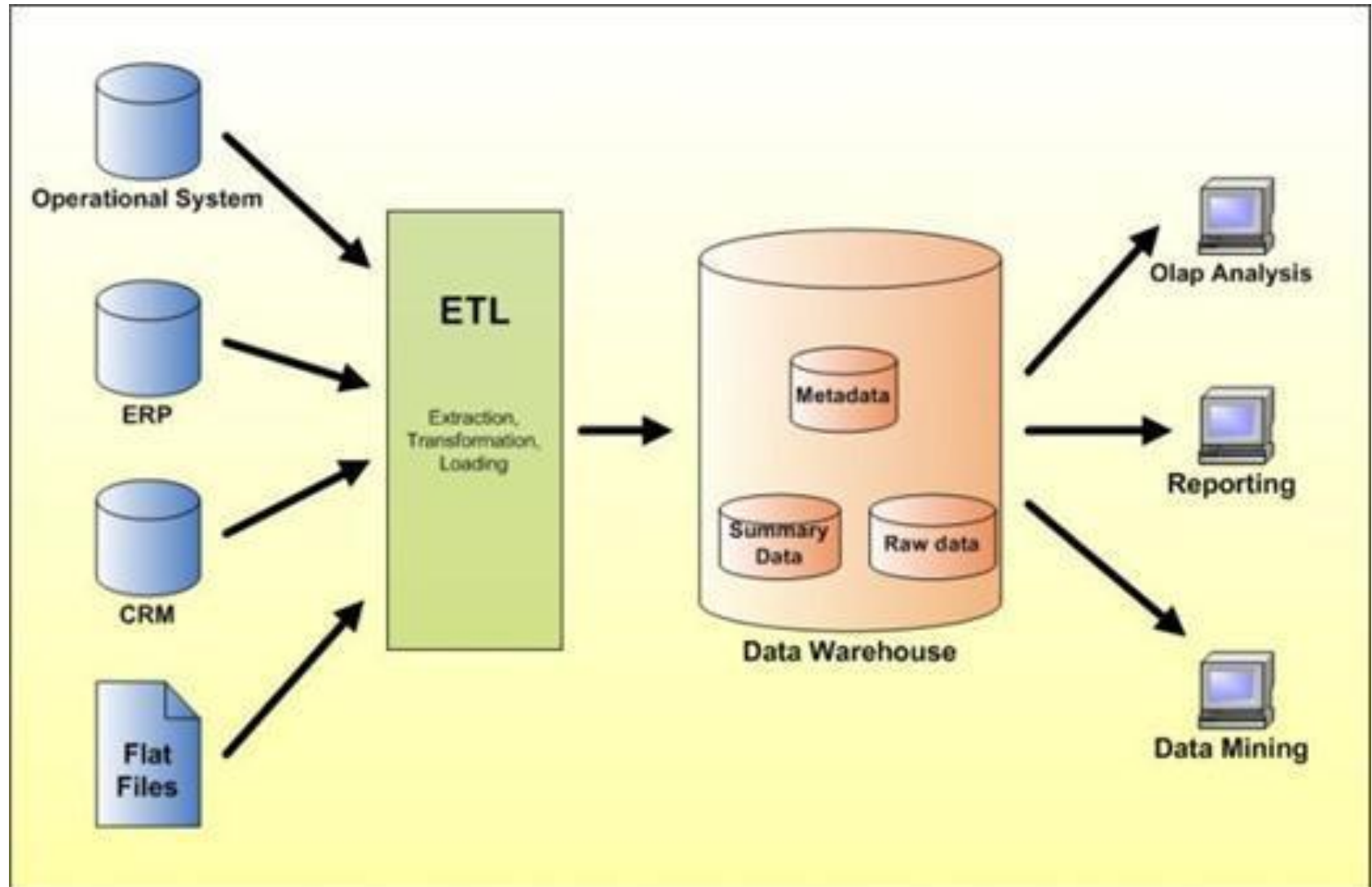
Processo ETL

ETL Process



O sucesso dos projetos de carga de banco de dados ou de mineração de dados depende fortemente da qualidade dos **dados preparados** no processo ETL.

ETL voltado para Data Warehouse



Conceitos e fundamentos

- Meta da unidade:
 - Descrever os diferentes tipos básicos de dados
 - Entender para que cada um destes tipos serve
 - Ser capaz de escolher o tipo adequado para uma determinada necessidade

Conceitos básicos

- **Dado**: Definimos **dado** como fatos, valores , observações, e medidas que não estão contextualizadas ou organizadas.
 - Ex: endereço, uma data qualquer, uma localização geográfica, um nome,...
- **Informação**: fato útil que pode ser extraído direta ou indiretamente a partir dos dados.
 - Dados processados, que foram organizados e interpretados e possivelmente formatados, filtrados, analisados e resumidos
 - Ex: endereço de entrega, idade, localização geográfica de Belém, nome do presidente do Brasil, população do Brasil...
- **Conhecimento**: é um **entendimento**, ou modelo, sobre pessoas, objetos ou eventos, derivado de informações sobre eles.
 - O conhecimento proporciona uma estrutura para interpretar as informações, usualmente incorporando e explicando variações no tempo ou no espaço.
 - Ex: Observando a evolução demográfica dos últimos 20 anos, há uma tendência de envelhecimento da população brasileira.

Dados, Informação e Conhecimento

Dados	Informação	Conhecimento
<ul style="list-style-type: none">• Simples observações sobre o estado do mundo	<ul style="list-style-type: none">• Dados dotados de relevância e propósito	<ul style="list-style-type: none">• Informação valiosa da mente humana• Inclui reflexão, síntese, contexto
<ul style="list-style-type: none">• Facilmente estruturado• Facilmente obtido por máquinas• Frequentemente quantificado• Facilmente transferível	<ul style="list-style-type: none">• Requer unidade de análise• Exige consenso em relação ao significado• Exige necessariamente a mediação humana	<ul style="list-style-type: none">• De difícil estruturação• De difícil captura em máquinas• Frequentemente tácito (não formalmente expresso)• De difícil transferência
FONTE: DAVENPORT, T. H., PRUSAK, L.. Conhecimento empresarial. Rio de Janeiro: Campus, 1998.		

Conceitos básicos

■ Banco de Dados (BD, *database*):

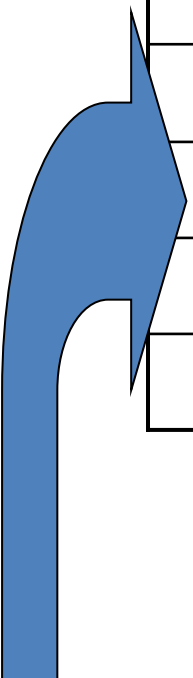
- Uma coleção organizada de dados.
- Coleções de dados interrelacionados e persistentes que representa um sub-conjunto dos fatos presentes em um domínio de aplicação (universo de discurso).

■ Dataset:


- Normalmente se refere aos **dados selecionados e organizados em forma tabular (em colunas)** para processamento computacional (e estatístico).
- Os dados de um dataset são obtidos de um banco de dados.
- Podem estar disponíveis e materializados em formatos diversos (arquivos texto, CSV, planilhas, JSON, etc).

Dados ou informação “bruta”

- Em mineração de dados os dados são representados e organizados em um **dataset** (conjunto ou tabela de dados):



<i>Nome</i>	<i>Telefone</i>	<i>Peso</i>	<i>País</i>
Obama	1 43 228859	67	EUA
Lula	55 23 591927	78	Brasil
Fidel	34 95 499402	82	Cuba
...



campo ou **atributo** (ou também “variável”, descritor)

Linhas (rows), registros, exemplos, casos ou “**instâncias**” (*instances*)

Medidas, Variáveis (descritores)

■ Medidas

- O que é possível medir sobre as características: meu carro é azul escuro, 2 portas, 6 cilindros, 5 passageiros

■ Variáveis, descritores

- Uma variável representa uma medida que toma um numero particular de valores (domínio), com a possibilidade de valores diferentes para cada observação.
- Dependendo da abordagem e do conceito, recebe o nome de “atributo”, “campo” ou “coluna” quando se refere ao armazenamento em tabelas de banco de dados.

Escalas de Medida (tipos de dados)

- As variáveis possuem níveis de mensuração definidos em uma **escala**.
- **Escala** é um conjunto de símbolos ou números, construído com base numa regra, que se aplica aos dados das variáveis.
 - Representam os “tipos” de dados das variáveis.
- Principais Escalas de Medida:
 - Nominal
 - Ordinal
 - Intervalar
 - Proporcional ou Razão

Escalas Nominal e Ordinal

Escala **Nominal**

Nessa escala os valores são não numéricos e são não ordenados.
Duas instâncias apresentam ou não o mesmo valor.
Ex: Cor, Modelos de Carro, etc

Escala **Ordinal**

Nessa escala os são não numéricos e ordenados. Uma instância pode apresentar um valor comparativamente maior do que uma outra. Ex: Grau de Instrução

Escalas Intervalar e Proporcional (razão)

Escala Intervalar

Nessa escala de valores numéricos, existe não apenas uma ordem entre os valores, mas também existe diferença entre esses valores.

O *zero* é relativo.

Ex: Temperatura em Graus Celsius

Escala Proporcional (razão)

Nessa escala de valores numéricos, além da diferença, tem sentido calcular a proporção entre valores (o *zero* é absoluto).

Ex: Peso, Altura, etc.

Cardinalidade dos atributos das variáveis

Qualitativo / quantitativo

Variáveis **qualitativas** (categóricas):
escalas **nominais** ou **ordinais**

Variáveis **quantitativas** (numéricas):
escalas **intervalares** e **proporcionais**

Variáveis Binárias (Dicotômicas)

Variáveis dicotômicas

Ex: Sexo (M, F)

Variáveis *binárias*

Em geral são codificadas como “0”, “1” (ausente, presente)

- “1”, se a característica de interesse está presente
- “0”, se a característica de interesse não está presente

Ex: Possui doença pré-existente? (Sim , Não)

Variáveis **Binárias** podem ser criadas (“**Dummy Variable**”) para representar outras variáveis explicativas que assumem um de dois valores possíveis.

- Representam características qualitativas, em eventos que tenham apenas 2 resultados possíveis. Utiliza-se em geral os valores “0” ou “1”.

Cardinalidade: Discreto versus Contínuo

Variáveis **Discretas**

Números inteiros, sem frações, como em **contagem**. Constituem um conjunto finito.

Variáveis **binárias**, indicadora (0 ou 1) são variáveis discretas.

Ex.: idade em anos completos; ausência (0) presença (1) de algum atributo

Variáveis **contínuas**

Podem, em princípio, assumir qualquer valor dentro de um Intervalo numérico.

Ex: Peso, altura, salário em Real, nota em um teste (entre 0 e 10)

Exercício

EXERCÍCIO - Classifique as variáveis em:

- **Qualitativas** (categóricas) do tipo **nominal** ou **ordinal**
- **Quantitativas** (numéricas) do tipo **contínua** ou **discreta**:

- **Cor dos olhos.**
 -
- **Índice de liquidez nas indústrias paraenses.**
 -
- **Produção de café no Brasil (em toneladas).**
 -
- **Grau da dor de um paciente (forte, moderada, branda, nenhuma)**
 -
- **Número de defeitos em aparelhos de TV.**
 -
- **Grupo sanguíneo de uma pessoa**
 -
- **Comprimento dos pregos produzidos por uma empresa.**
 -
- **O ponto obtido em cada jogada de um dado.**
 -
- **Estado civil de uma pessoa**
 -

Exercício - RESPOSTAS

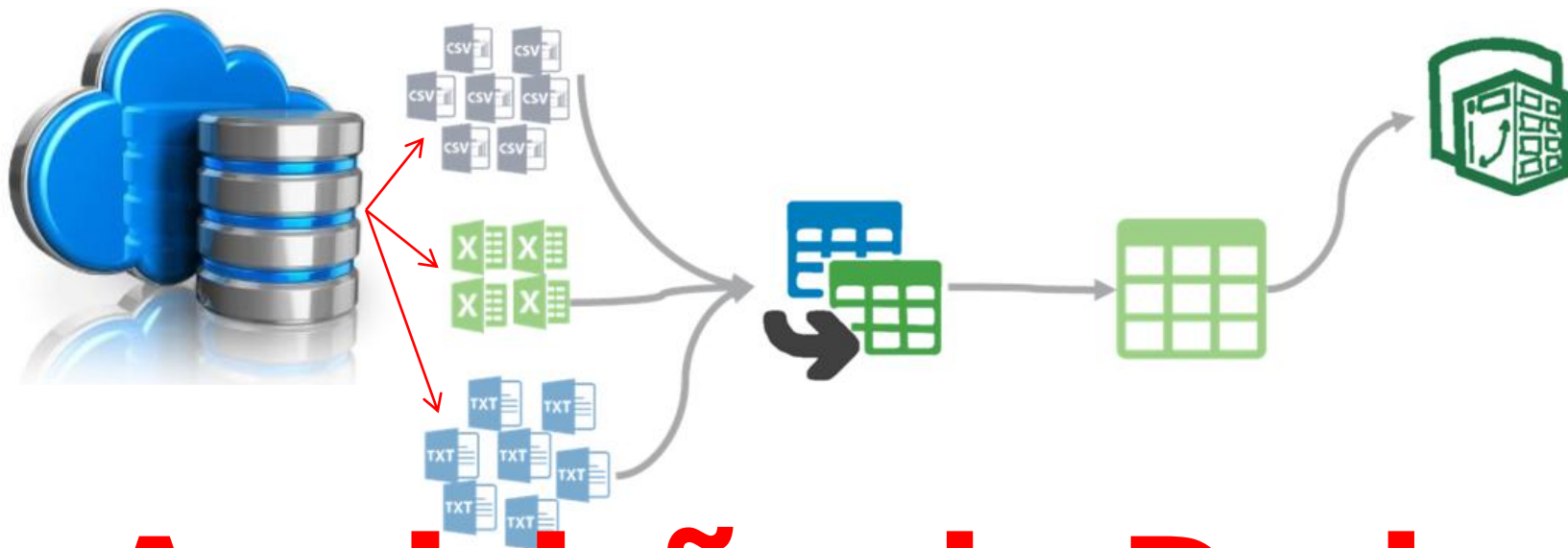
EXERCÍCIO - Classifique as variáveis em:

- **Qualitativas** (categóricas) do tipo **nominal** ou **ordinal**
- **Quantitativas** (numéricas) do tipo **contínua** ou **discreta**:

- **Cor dos olhos.**
 - Resp: qualitativa nominal (claros, castanhos, negros).
- **Índice de liquidez nas indústrias paraenses.**
 - Resp: quantitativa contínua.
- **Produção de café no Brasil (em toneladas).**
 - Resp: quantitativa contínua.
- **Grau da dor de um paciente (forte, moderada, branda, nenhuma)**
 - Resp: qualitativa nominal
- **Número de defeitos em aparelhos de TV.**
 - Resp: quantitativa discreta.
- **Grupo sanguíneo de uma pessoa**
 - Resp: Qualitativo Nominal (sangue tipo A, AB, B e O)
- **Comprimento dos pregos produzidos por uma empresa.**
 - Resp: quantitativa contínua.
- **O ponto obtido em cada jogada de um dado.**
 - Resp: quantitativa discreta.
- **Estado civil de uma pessoa**
 - Resp: Qualitativo Nominal (casado/viúvo/solteiro, etc).

Resumo – Tipos de Variáveis

Variáveis	Tipos	Descrição	Exemplos
Qualitativas ou Categóricas	Nominal	Não existe nenhuma ordenação	Cor dos olhos, sexo, estado civil, tipo sanguíneo.
	Ordinal	Existe uma ordenação I, II, III	Nível de escolaridade, estágio da doença, colocação de concurso.
Quantitativas	Discretas	Valor pertence a um conjunto enumerável	Número de filhos por casal, quantidade de leitos
	Contínuas	Quando o valor pertence a um intervalo real	Medidas de altura e peso, taxa de glicose, nível de colesterol.



Aquisição de Dados (*Extração*)

Aquisição de dados

- **Aquisição** ou Coleta (ultimamente usa-se “Colheita”) de dados é a etapa da **Extração** que compreende as atividades relacionadas com o levantamento e os procedimentos para obtenção dos dados disponíveis em um formato aberto.
 - Para o escopo deste curso, os “dados” citados tem origem em bases públicas e abertas.
 - Formato aberto diz respeito a um padrão “open data” (dados primários e processáveis por máquina).
- Para dados alfanuméricos (descritivos), a aquisição de dados obtidos de uma base pública deve observar:
 - Identificação das fontes de dados existentes;
 - Identificação das características técnicas dos dados e metadados;
 - Análises de utilização, considerando aspectos como, atualidade, confiabilidade, escala, disponibilidade, formato, etc.;
 - Análise de compatibilidade semântica (contéudo) com o problema a ser tratado;
 - Coleta dos dados propriamente dita.

Fontes de dados

- Na maioria das vezes, dados primários de pesquisa são obtidos em fontes oficiais e de governo.
- A consulta às fontes governamentais pode oferecer acesso a diversos domínios de dados (demografia, saúde pública, educação, transporte, etc).
- Exemplo de fontes governamentais:
 - **IBGE**: destaca-se a **PNAD** (*Pesquisa Nacional de Amostragem por Domicílios*), realizada anualmente e o **CENSO** brasileiro feito a cada dez anos, que retrata tendências demográficas, indicadores sociais municipais, trabalho e rendimento, características gerais da população, educação, migração, deslocamento, etc.
 - **INEP**: informações sobre **censo escolar**, **ENEM**, etc.

Acesso ao banco de dados Sidra (IBGE)

<http://www.ibge.gov.br/>

The screenshot displays the IBGE website interface. The top navigation bar includes 'Indicadores', 'População', 'Economia', 'Geociências', and 'Canais'. The left sidebar contains a 'Canais' menu with options like 'Banco de Dados', 'Séries Estatísticas', and 'CIDRA'. A red arrow points to the 'Canais' menu, and another points to the 'Banco de dados agregados' link. The main content area features a 'Banco de Dados Agregados' section with a search bar and a list of data series. A table at the bottom shows the 'Índice Nacional de Preços ao Consumidor - fevereiro 2016' for Brazil.

Índice geral e Grupos	Percentual no mês	Percentual
Índice geral	0,95	
Alimentação e bebidas	1,19	
Habituação	-0,18	
Artigos de residência	0,82	
Vestuário	0,34	
Transportes	1,37	
Saúde e cuidados pessoais	1,13	25

INEP (<http://portal.inep.gov.br/>)

Navegação: Inep > Informações Estatísticas > Microdados > Microdados para download

<http://portal.inep.gov.br/basica-levantamentos-acessar>

The screenshot shows the INEP website interface. At the top, there is the INEP logo and the text 'Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira'. To the right of the logo are social media icons for Twitter and RSS, and a button to 'A+ aumentar fonte' and 'A- diminuir fonte'. Below the header is a search bar with the text 'BUSCAR'. The main navigation menu on the left includes 'Página Inicial', 'Microdados', 'Microdados para download', 'Notícias', and 'Fale Conosco'. The 'Microdados' section is active, displaying a list of data sources and their corresponding years. The list includes: Microdados Enade (2004-2014), Microdados Censo Escolar (1995-2006), Microdados Censo da Educação Superior (1995-2006), Microdados Censo dos Profissionais do Magistério (2003), Microdados Saeb (Aneb/Prova Brasil) (1995-2013), Microdados ANA (2014), Microdados Enem (1998-2014), Microdados Provão (1997-2003), and Microdados PNERA (2004).

Microdados	Anos
Microdados Enade	2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Microdados Censo Escolar	1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Microdados Censo da Educação Superior	1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Microdados Censo dos Profissionais do Magistério	2003
Microdados Saeb (Aneb/Prova Brasil)	1995 1997 1999 2001 2003 2005 2011 2013
Microdados ANA	2014
Microdados Enem	1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Microdados Provão	1997 1998 1999 2000 2001 2002 2003
Microdados PNERA	2004

Fontes de dados (mecanismos e origem)

- Os dados podem estar em um sistema de banco de dados SGBD
 - Acesso via ODBC, JDBC ... (ou acesso proprietário, legado)
- Dados podem ser recuperados em arquivos texto (ou flat file, tabela de dados)
- Dados podem vir em forma de planilhas
- Dados podem ser gerados dinamicamente por um serviço web (formato XML, JSON)

Formatos de arquivos - recomendação

- Seguindo os padrões de publicação de **dados abertos**, os dados precisam ser disponibilizados em formato livre, preferencialmente:
 - Arquivo CSV (*Comma-Separated Values*)
 - Planilha (tabular)
 - JSON
 - XML

Principais Princípios de “dados abertos”

- **Completos:** Todos os dados públicos devem ser disponibilizados. Não possui restrições de privacidade, segurança ou outros privilégios.
- **Primários:** com o maior nível possível de granularidade, sem agregação ou modificação.
- **Compreensíveis por máquina:** estruturados de forma que sejam processados automaticamente
 - por exemplo, uma tabela em PDF dificulta o processamento automatizado; já uma tabela em formato estruturado, como CSV ou XML, é processada mais; facilmente por softwares e sistemas.
- **Outros princípios:**
 - atuais, acessíveis, não discriminatório, não proprietário, livre de licença.

Metadados

- Metadados são dados que descrevem os atributos e características de um recurso (conjunto de dados, por ex.).
- Metadados pode descrever:
 - localização, documentação, data da criação, procedência, formato dos dados, etc.
- Metadados fornecem o contexto para entender os dados através do tempo.
- Metadados são dados associados com objetos que ajudam seus usuários a entender a existência e características dos dados.

Codificação de caracteres

- A informação da codificação de caracteres deve ser conhecida (via metadados), ou identificada por uma ferramenta de reconhecimento de codificação.
- Usar dados que não estão com o mesmo formato de codificação da base alvo (destino) pode inviabilizar ou criar problemas de qualidade na carga dos dados.
- Normalmente, a codificação dos dados originais assume a tabela adotada no sistema operacional:
 - O Windows geralmente usa a codificação **ISO-8859-1**
 - GNU/Linux pode usar frequentemente **UTF-8**.
- Para os objetivos do curso, assumimos o padrão **UTF-8**.

Ferramentas para Codificação de arquivos

As melhores ferramentas rodam em Linux.

- Para descobrir a codificação de origem pode-se usar a ferramenta **file**, exemplo:
file [arquivo]
- Para converter o conteúdo em texto dos arquivos você pode usar a ferramenta **iconv**, cuja sintaxe é:
iconv -f [codificação-origem]
-t [codificação-destino]
arquivo > arquivo.new

Obs: funciona bem para grandes arquivos.

Exemplo de conversão de ISO-8859-1 para UTF-8:

iconv -f iso-8859-1 -t utf-8 arquivo.txt > arquivo_novo.txt

Convertendo arquivos para UTF-8 (Windows)

- Funciona bem para arquivos com tamanhos pequenos ou médio.
- **Método 1**
 - Use um editor de texto, Ex: TextPad ou NotePad++ (Até mesmo o Notepad)
 - Ao salvar ou no menu “Editar”, escolha o formato UTF-8
- **Método 2**
 - Abra o arquivo como planilha do Libre-Office e salve com opção estendida (veja o código UTF-8).

Outras opções em https://docs.moodle.org/24/en/Converting_files_to_UTF-8

Tipos de arquivo texto para carga de dados

- Quando se fala de importação de dados em formato texto, temos basicamente dois formatos diferentes:
 - Arquivos **de largura fixa**
 - Arquivos com separadores de campos (**CSV**).

Leitura adicional:

<http://imasters.com.br/banco-de-dados/dicas-e-truques-para-importacao-de-arquivos-em-formato-texto/?trace=1519021197&source=single>

Arquivo de largura fixa

- Os arquivos de largura fixa são os formatos mais antigos e usados para intercâmbio de dados.
 - Bases de dados do IBGE (como o censo demográfico) ainda usam esse padrão.
- Neste formato, os campos estão em posições fixas (há uma posição inicial e posição final para cada campo), que têm que ser idênticas para todas as linhas do arquivo.
- Não possuem cabeçalho (com o nome dos campos) na primeira linha, obrigando ter o conhecimento da estrutura dos dados em um dicionário adicional ao arquivo.
- Está em desuso, pois além das dificuldades com a estrutura em formato fixo, há um desperdício de espaço de armazenamento quando contém dados vazios, aumentando em muito o tamanho final do arquivo.

Formato de arquivo: CSV

CSV - Comma Separated Values

- Um **arquivo CSV** é um arquivo texto (pode ser lido como [planilha](#)), que contém dados tabulados, organizados em linhas e os campos (as colunas) são separados por um caractere de separação (em geral, uma vírgula ou um ponto e vírgula).
- Se o conteúdo de uma campo contiver o mesmo delimitador (“;” por ex) usado no arquivo, o conteúdo **deverá** estar entre aspas simples (') ou duplas (").
- Pode conter ou não cabeçalho com a identificação dos nomes dos campos.
- É um formato que se tornou padrão para a maioria das aplicações de intercâmbio de dados.

Exemplos (Fixo e CSV)

Desktops	Alabama	Sales	Jan	2004.0
Desktops	Alabama	Sales	Feb	86957.0
Desktops	Alabama	Sales	Mar	32216.0
Desktops	Alabama	Sales	Apr	94768.0
Random Access Memory	Alabama	Sales	May	0.0
Desktops	Alabama	Sales	Jun	29062.0
Desktops	Rhode Island	Sales Discounts	Jan	5510.0
Desktops	Rhode Island	Sales Discounts	Feb	100620.0
Desktops	Rhode Island	Sales Discounts	Mar	38612.0
Desktops	Rhode Island	Sales Discounts	Apr	0.0

Arquivo Fixo. As linhas vermelhas pontilhadas representam as posições em que terminam um campo e começam o campo seguinte.

```
"Desktops","Alabama","Sales","Jan",2004.0
"Desktops","Alabama","Sales","Feb",86957.0
"Desktops","Alabama","Sales","Mar",32216.0
"Desktops","Alabama","Sales","Apr",94768.0
"Random Access Memory","Alabama","Sales","May",0.0
"Desktops","Alabama","Sales","Jun",29062.0
"Desktops","Rhode Island","Sales Discounts","Jan",5510.0
"Desktops","Rhode Island","Sales Discounts","Feb",100620.0
"Desktops","Rhode Island","Sales Discounts","Mar",38612.0
"Desktops","Rhode Island","Sales Discounts","Apr",0.0
```

Arquivo CSV.

O caractere usado como separador de campo é a vírgula (","), sem cabeçalho.

Atividade prática

Nesta atividade, você fará o papel de um “pesquisador” que vai colher informações sobre o Censo Agropecuário de 1995.

- Abra o banco SIDRA (IBGE) e utilize o link “acervo”, na mensagem:

“Veja outros dados na seção [Pesquisas](#), nas demais seções, nos temas e no [acervo](#).”

- Escolha em “Banco de Dados Agregado” a opção “Pesquisas e tabelas”
- Em seguida, escolha “Censo Agropecuário (672 tabelas) ”
- Na página seguinte, escolha “Tabela 316 - Área dos estabelecimentos por grupos de área total e utilização das terras - Ano 1995”
- Leia as observações sobre os metadados no ícone ao lado da opção
- Em seguida, faça a montagem dos dados a serem gerados em arquivo CSV, seguindo as configurações vistas nas telas seg

Acesso a esse procedimento:

<http://www.sidra.ibge.gov.br/bda/tabela/listabl.asp?c=316&z=t&o=3>

Montando os dados (1)

Montar quadro	Obter ranking	Gerar gráfico	Gerar cartograma
Tabela 316 - Área dos estabelecimentos por grupos de área total e utilização das terras			
Matriz multidimensional (2x17x10x1x5684) com 1.932.560 valores		Veja como montar um quadro	
Variável(2):	Seleção ▼	No cabeçalho ▼	2 ▼
Área dos estabelecimentos agropecuários - decimais:3/3 ▲ Área dos estabelecimentos agropecuários (Percentual) - decimais:5/2 ▼ <small>decimais:x/y - x=nro. de casas em que o valor decimal está disponível; y=nro. padrão de casas para apresentação (pode ser alterado ao final da página)</small>			
Grupos de área total(17):	Seleção ▼	No cabeçalho ▼	3 ▼
Total ▲ Menos de 1 ha 1 a menos de 2 ha 2 a menos de 5 ha 5 a menos de 10 ha 10 a menos de 20 ha 20 a menos de 50 ha 50 a menos de 100 ha ▼			
Utilização das terras(10):	Seleção ▼	Na coluna ▼	4 ▼
Total ▲ Lavouras permanentes Lavouras temporárias Lavouras temporárias em descanso Pastagens naturais Pastagens plantadas Matas e florestas naturais Matas e florestas artificiais ▼			

Montando os dados (2)

Ano(1):
1995

Unidade Territorial(5684):
☐ Exibir código ☒ Exibir nome

☒ **Níveis Territoriais**
Brasil(1):
Grande Região(5): [Fazer seleção avançada](#)
Unidade da Federação(27): [Fazer seleção avançada](#)
Mesorregião Geográfica(137): [Fazer seleção avançada](#)
Microrregião Geográfica(558): [Fazer seleção avançada](#)
Município(4956): Nome:
[Fazer seleção avançada](#)

☐ **Visões Territoriais**
Brasil e Grande Região(6)

[As visões territoriais são c
escolher uma visão obser](#)

Opções de consulta:

☐ **Visualizar** (até 10.000 valores)
☐ Preparar para impressão
☒ Gerar link para consulta posterior

☒ **Gravar** [Veja as gravações a posteriori efetuadas nos últimos 60 dias](#)
Arquivo (não coloque a extensão .csv, .tsv ou .zip, pois será automática)
Formato [Conheça os formatos e como utilizá-los](#)
Modalidade
E-mail (se notificação ou envio por e-mail)
☒ Compressão(.zip)
Apresentar os valores decimais com casas
☒ Dimensões com apenas uma seleção são apresentadas no cabeçalho e as demais nas linhas ou colunas

(Utilize Alt-o como atalho para o OK)

Arquivo gerado (resultado após agendamento de processamento)

Arquivo "agro1995.csv" visualizado no Notepad++

```
1 "Tabela 316 - Área dos estabelecimentos por grupos de área total e utilização das terras"
2 "Variável";"Área dos estabelecimentos agropecuários (Hectares)"
3 "Grupos de área total";"Total"
4 "Ano";"1995"
5 "Brasil e Município";"Utilização das terras"
6 ";Lavouras permanentes";"Lavouras temporárias";"Lavouras temporárias em descanso";"Pastagens natu
7 "Brasil";7541625,591;34252828,911;8310028,686;78048463,080;99652008,615;88897582,416;5396015,930;
8 "Alta Floresta D'Oeste - RO";13713,060;12448,014;2316,125;24080,089;102685,442;208768,389;1497,70
9 "Ariquemes - RO";13668,523;4077,912;3869,383;236,555;110289,081;173012,816;1888,620;3695,856;4619
10 "Cabixi - RO";1162,922;4028,134;1208,022;1730,300;53738,669;89159,416;561,753;1345,316;1644,806
11 "Cacoal - RO";25538,920;4828,592;3822,511;1646,293;122929,558;95818,806;135,520;4316,720;4019,193
12 "Cerejeiras - RO";1063,454;4622,276;1639,471;132957,826;109712,012;137039,858;9849,253;9667,110;7
13 "Colorado do Oeste - RO";870,615;6271,491;1129,932;23810,163;46229,074;50606,805;1200,230;1659,75
14 "Corumbiara - RO";2022,667;6166,945;1971,883;8394,849;110613,552;122412,597;288,741;4228,213;3777
15 "Costa Marques - RO";1581,829;2942,755;489,305;2535,520;13960,813;212975,240;662,140;40917,636;89
16 "Espigão D'Oeste - RO";5740,173;4014,362;2006,283;11748,833;79068,835;117351,741;2608,809;3446,60
```

Normal text file length: 476821 lines: 4964 Ln: 1 Col: 69 Sel: 0 | 0 Dos\Windows ANSI

Delimitador ";"

Número de linhas

**Codificação no
Windows (ANSI)**

Abrindo o arquivo "agro1995.csv" no Calc

Importação de texto - [agro1995.csv]

Importar

Conjunto de caracteres: Europa ocidental (Windows-1252/WinLatin 1)

Idioma: Padrão - Português (Brasil)

Da linha: 1

Opções de separadores

☐ Largura fixa ☒ Separado por

☐ Tabulação ☐ Vírgula ☒ Ponto-e-vírgula ☐ Espaço ☐ Outro

☐ Mesclar delimitadores

Delimitador de texto: "

Outras opções

☐ Campos entre aspas como texto ☒ Detectar números especiais

Campos

Tipo de coluna: Padrão

	Padrão
1	Tabela 316 - Area dos estabelecimentos por grupos de área total e utilização
2	Variável
3	Grupos de área total
4	Ano
5	Brasil e Município
6	
7	Brasil
8	Alta Floresta D'Oeste - RO

OK Cancelar Ajuda

Codificação (Windows)

Linha inicial para carga

Delimitador ";"

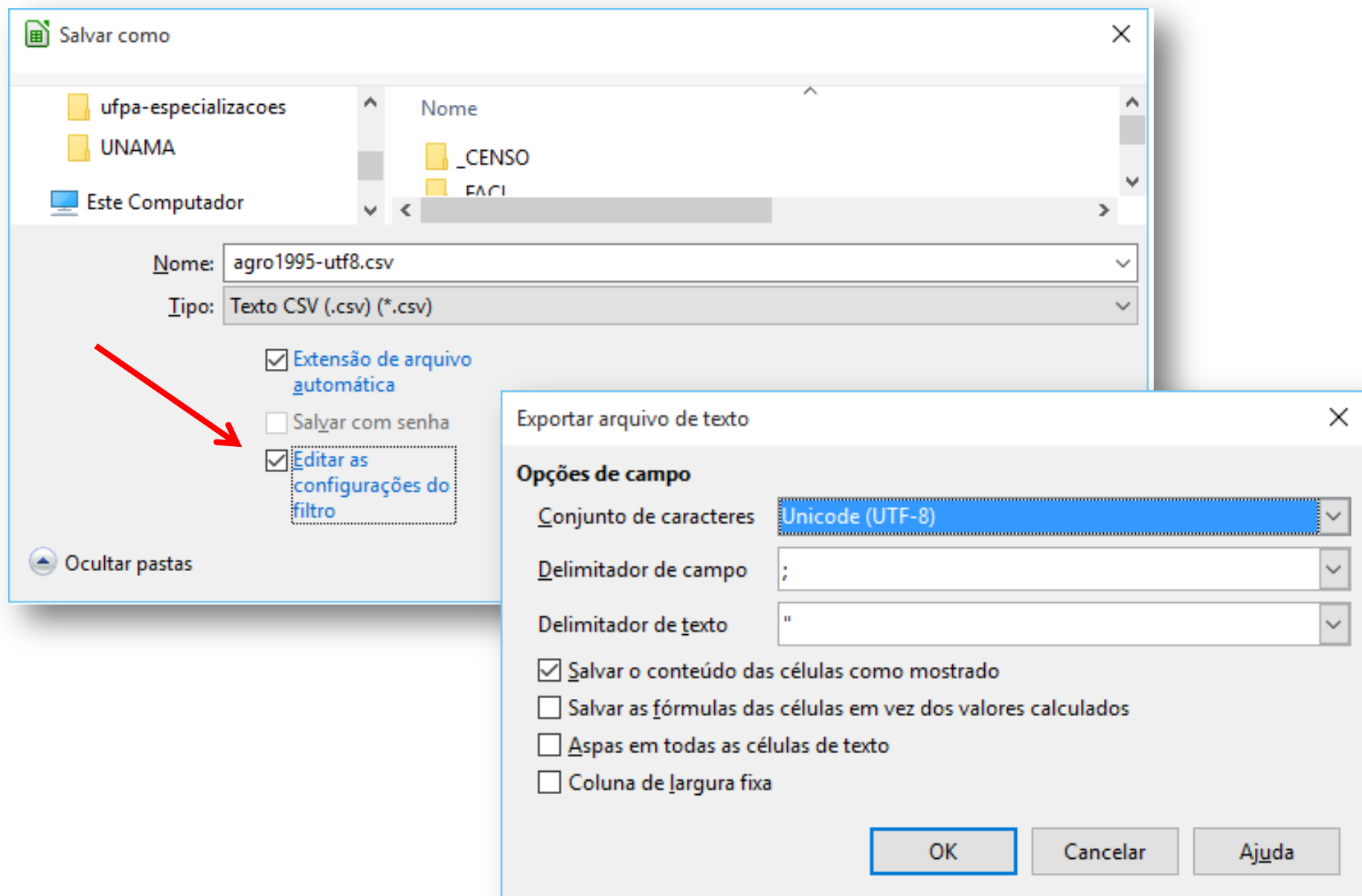
Configuração das
colunas

Visualização do conteúdo e ajustes iniciais

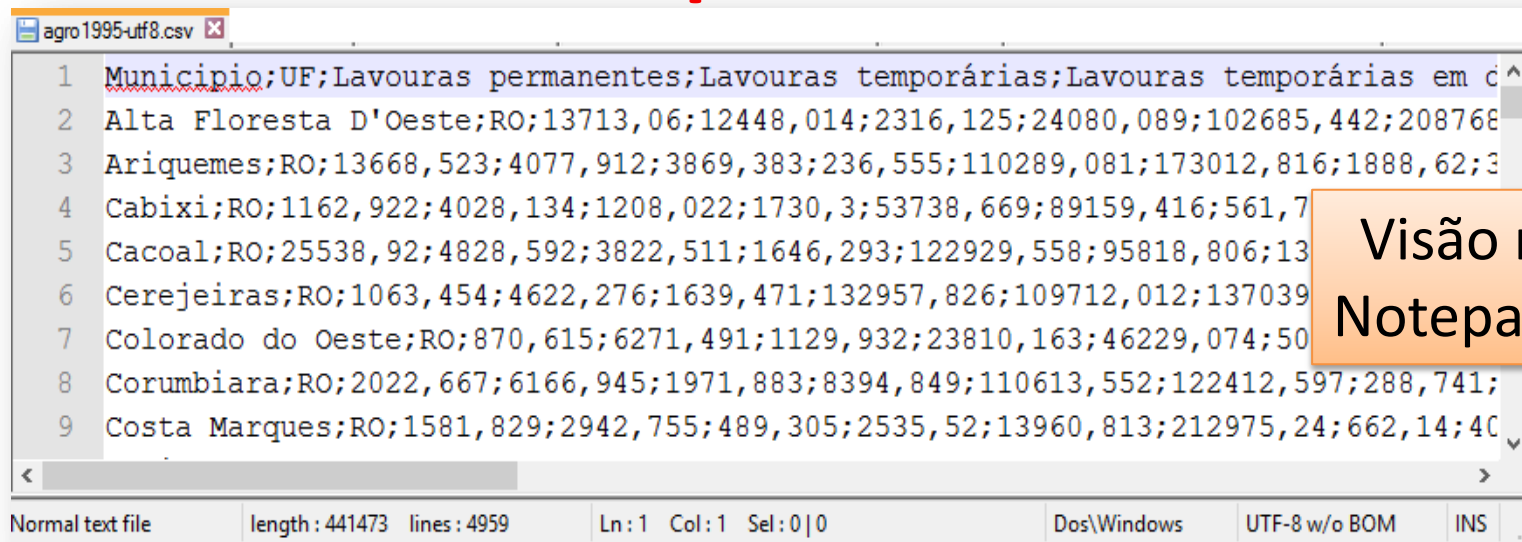
	A	B	C	D	E
1		Lavouras permanentes	Lavouras temporárias	Lavouras temporárias em descanso	Pastagens natur
2	Brasil	7541625,591	34252828,911	8310028,686	7804846
3	Alta Floresta D'Oeste - RO	13713,06	12448,014	2316,125	24080
4	Ariquemes - RO	13668,523	4077,912	3869,383	236
5	Cabixi - RO	1162,922	4028,134	1208,022	17
6	Cacoal - RO	25538,92	4828,592	3822,511	1646
7	Cerejeiras - RO	1063,454	4622,276	1639,471	132957
8	Colorado do Oeste - RO	870,615	6271,491	1129,932	23810
9	Corumbiara - RO	2022,667	6166,945	1971,883	8394
10	Costa Marques - RO	1581,829	2942,755	489,305	253
11	Espigão D'Oeste - RO	5740,173	4014,362	2006,283	11748
12	Guajará-Mirim - RO	727,058	1912,308	1485,098	7550
13	Jaru - RO	9417,439	6994,823	3422,496	20288
14	Ji-Paraná - RO	3968,365	3333,887	1926,97	15372

1. Definir o nome da coluna "A" para "Município"
2. Inserir uma nova coluna antes da coluna B ("Lavouras Permanentes"), onde ficará a sigla da UF.
3. Quebrar a coluna "Município" em duas, uma para "município" e outra para a sigla da Unidade (UF), usando o "|" como delimitador.
Utilize a opção: Dados > Texto para colunas, usando como delimitador "|".
4. Deletar a linha 2 que contém dados totalizados do "Brasil"
5. Por último, salve o arquivo CSV no formato de codificação UTF-8.
Utilize a opção "Salvar Como...", mudando o nome do arquivo (mantém o original).

Salvando CSV em UTF-8



Resultado do arquivo CSV formatado



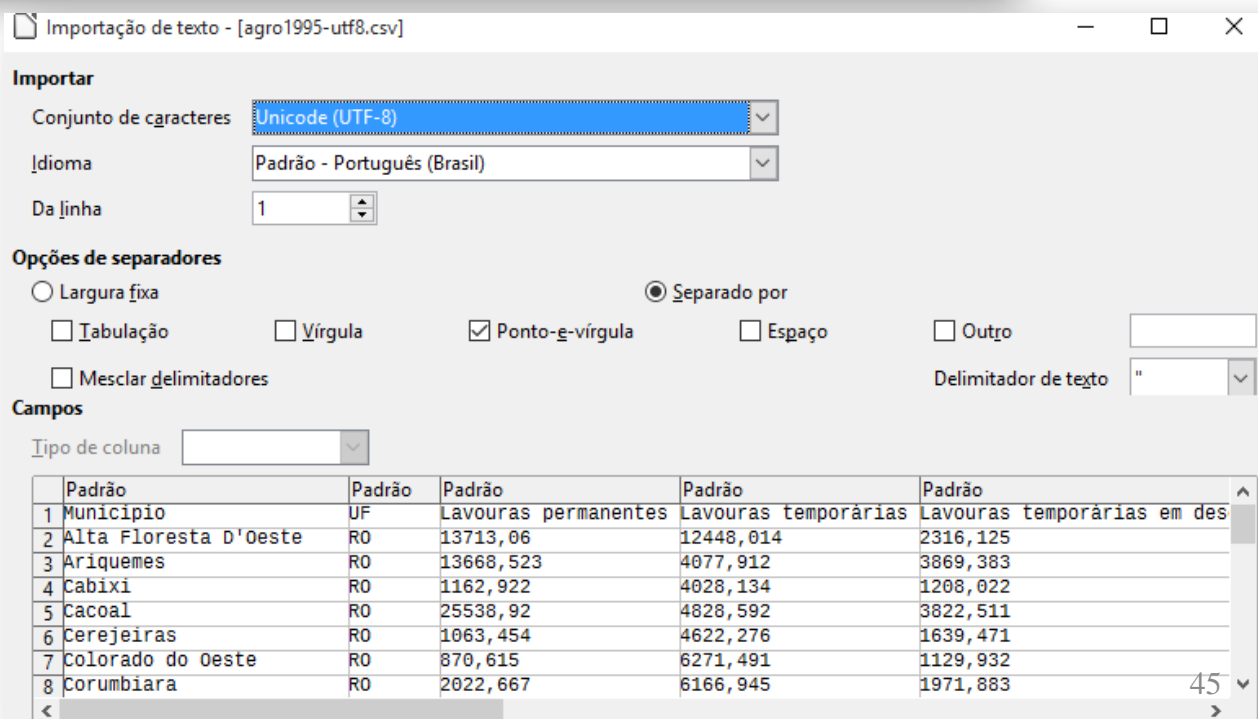
agro1995-utf8.csv

```
1 Municipio;UF;Lavouras permanentes;Lavouras temporárias;Lavouras temporárias em d
2 Alta Floresta D'Oeste;RO;13713,06;12448,014;2316,125;24080,089;102685,442;20876
3 Ariquemes;RO;13668,523;4077,912;3869,383;236,555;110289,081;173012,816;1888,62;3
4 Cabixi;RO;1162,922;4028,134;1208,022;1730,3;53738,669;89159,416;561,7
5 Cacoal;RO;25538,92;4828,592;3822,511;1646,293;122929,558;95818,806;13
6 Cerejeiras;RO;1063,454;4622,276;1639,471;132957,826;109712,012;137039
7 Colorado do Oeste;RO;870,615;6271,491;1129,932;23810,163;46229,074;50
8 Corumbiara;RO;2022,667;6166,945;1971,883;8394,849;110613,552;122412,597;288,741;
9 Costa Marques;RO;1581,829;2942,755;489,305;2535,52;13960,813;212975,24;662,14;40
```

Normal text file | length : 441473 lines : 4959 | Ln: 1 Col: 1 Sel: 0 | 0 | Dos\Windows | UTF-8 w/o BOM | INS

Visão no
Notepad++

Importação no
Calc



Importação de texto - [agro1995-utf8.csv]

Importar

Conjunto de caracteres: Unicode (UTF-8)

Idioma: Padrão - Português (Brasil)

Da linha: 1

Opções de separadores

☐ Largura fixa ☒ Separado por

☐ Tabulação ☐ Vírgula ☒ Ponto-e-vírgula ☐ Espaço ☐ Outro

☐ Mesclar delimitadores

Delimitador de texto: "

Campos

Tipo de coluna:

	Padrão	Padrão	Padrão	Padrão	Padrão
1	Município	UF	Lavouras permanentes	Lavouras temporárias	Lavouras temporárias em des
2	Alta Floresta D'Oeste	RO	13713,06	12448,014	2316,125
3	Ariquemes	RO	13668,523	4077,912	3869,383
4	Cabixi	RO	1162,922	4028,134	1208,022
5	Cacoal	RO	25538,92	4828,592	3822,511
6	Cerejeiras	RO	1063,454	4622,276	1639,471
7	Colorado do Oeste	RO	870,615	6271,491	1129,932
8	Corumbiara	RO	2022,667	6166,945	1971,883

Transformação de Dados

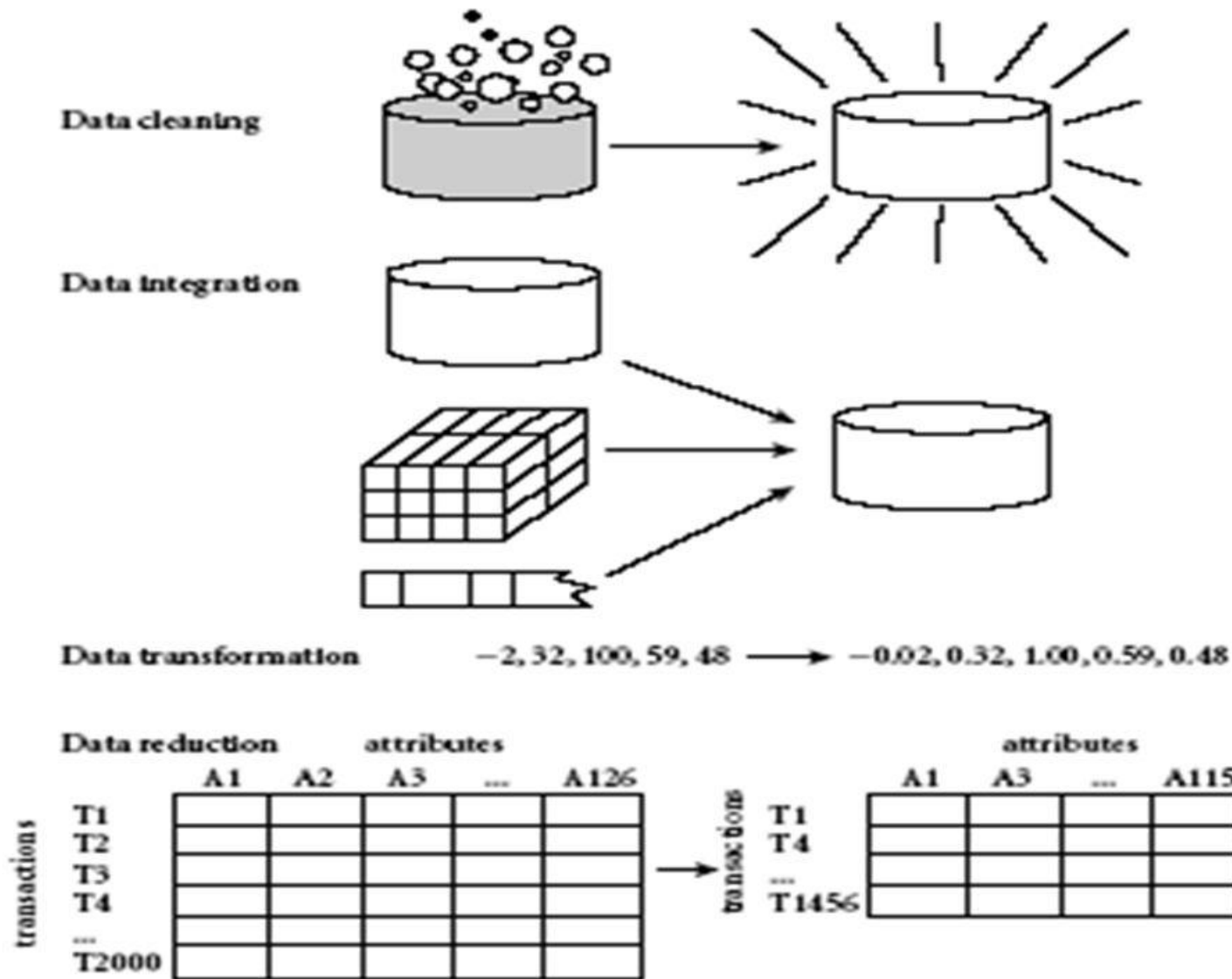
(Limpeza, Ajustes e Consolidação)

Limpeza, Ajustes e Consolidação (transformação)

É nesta etapa que realizamos os devidos ajustes, podendo assim melhorar a qualidade dos dados e consolidar dados de duas ou mais fontes.

- O estágio de transformação aplica um série de regras ou funções aos dados extraídos para ajustar os dados a serem carregados.
- Algumas fontes de dados podem requisitar algumas transformações, por exemplo:
 - **Limpeza** de dados ausentes e inconsistentes;
 - **Junção** de dados provenientes de diversas fontes (integração);
 - **Seleção** de apenas determinadas colunas (redução de dimensão);
 - **Tradução** de valores codificados (mapeamento).
 - Ex: Se a fonte de dados armazena “1” para sexo *masculino* e 2 para *feminino*, no banco alvo estes valores serão “M” para masculino e “F” para feminino.

Transformações no Pré-processamento



85

Limpeza dos dados

Em que consiste a “limpeza” dos dados?

- preencher dados ausentes
- identificar valores aberrantes (*outliers*)
- Identificar inconsistências
- identificar e “alisar” ruídos
- etc

Valores ausentes

Valores ausentes

- Um valor ausente é aquele ausente no conjunto de dados de uma variável, mas existente no contexto em que a medida foi realizada
Ex: renda da família em determinadas linhas do censo (PNAD)
- Numa base de dados eles podem ser indicados por **valores negativos** ou **nulos** em atributos numéricos.
- Em atributos **não numéricos** por **brancos** ou **traços** (ou qualquer símbolo convencionalizado para isso, por exemplo “<NULL>”).

Valores inaplicáveis

Valores inaplicáveis

Um valor inaplicável é um valor ausente e inexistente no contexto em que a medida foi realizada.
Envolve a dependência entre variáveis.

Ex: Uma base de dados sobre partos (em mulheres grávidas)
Sexo = **Masculino** e Número de Partos = **null**
Sexo = Feminino e Número de Partos = **0**

Valores ausentes e “vazios”

- A diferenciação entre valores **ausentes** e valores **inaplicáveis** é importante, mas não se dispõe de técnicas automáticas para fazer isso.
 - Deve-se fazê-lo manualmente.
- Há algumas técnicas para tratar dados ausentes, por exemplo:
 - ignorá-los, atribuir um valor fixo aos valores ausentes ou estimar os valores ausentes à partir de outras variáveis.
- Em algumas situações os dados ausentes são altamente informativos e ao serem tratados perde-se essa informação.

Tratamento para Valores ausentes

Quais os tratamentos usuais para valores ausentes?

- Ignorar a linha (a instância) do indivíduo ou mesmo eliminar o descritor (remover a variável, coluna);
- Preencher os valores ausentes manualmente;
- Usar uma constante global para representar os valores ausentes (não recomendado, pois o sistema pode identificar esse valor como um conceito);
- Usar a média (ou a moda);
- Usar a média (ou a moda) por classe
- Usar o valor mais provável segundo um modelo (regressão, regra de Bayes, árvores de decisão)

Observações Atípicas, dados aberrantes (Outliers)

- Ocorre quando uma instância possui valores de atributos identificáveis como sendo diferente das outras instâncias.
 - O valor extrapola (para mais ou menos) os valores esperados para o domínio dos dados.
- As instâncias atípicas não podem ser categoricamente caracterizadas como benéficas ou problemáticas
- Devem ser vistas no contexto da análise e avaliadas pelos tipos de informação que possam fornecer
 - **Benéficas:** podem ser indicativas de características da população que não seriam descobertas no curso normal da análise
 - **Problemáticas:** não são representativas da população, são contrárias aos objetivos da análise e podem confundir os algoritmos de aprendizado

Dados com ruído e /ou valores aberrantes

Alisamento: consiste em distribuir dados ordenados em “caixas”, tendo como referência os seus vizinhos.

Ordenação: 1, 1, 2, 3, 3, 3, 4, 5, 5, 7

Particionamento em “caixas”

1,1,2	3,3,3	4,5,5,7
caixa1	caixa2	caixa3

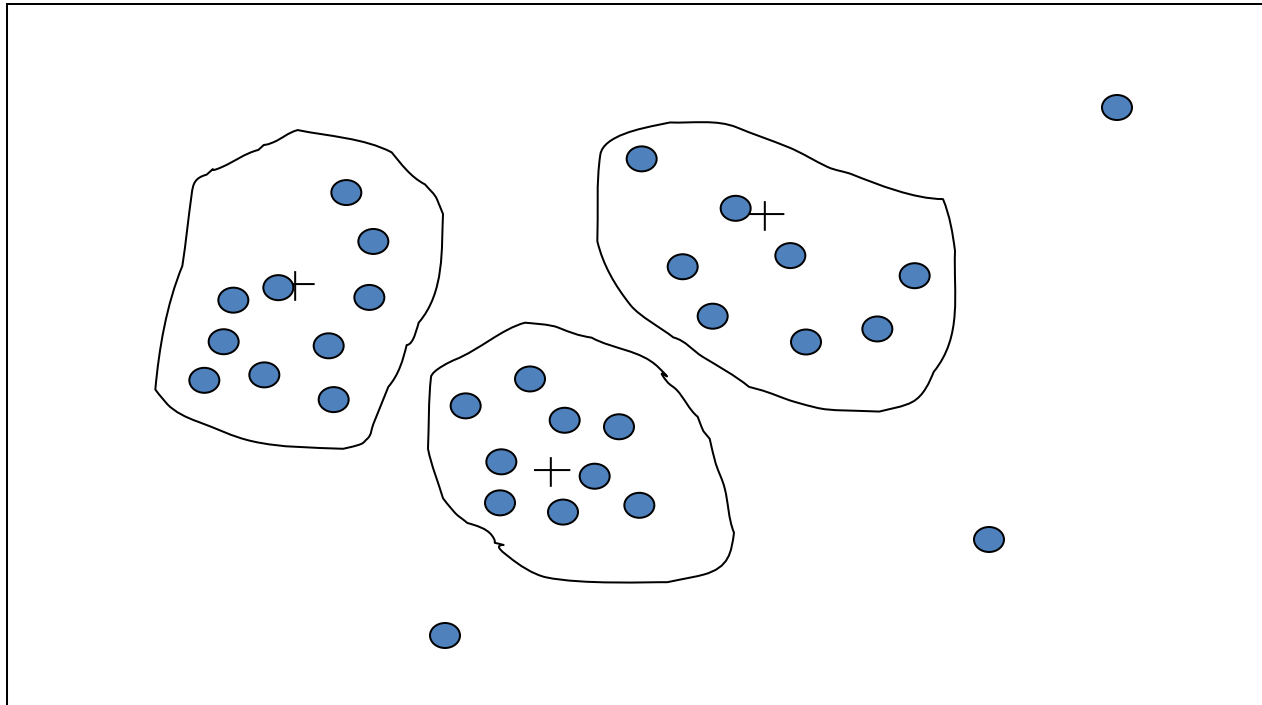
Alisamento pela mediana

1,1,1	3,3,3	5,5,5,5
caixa1	caixa2	caixa3

Outras alternativas: média, fronteiras

Dados com ruído e /ou valores aberrantes

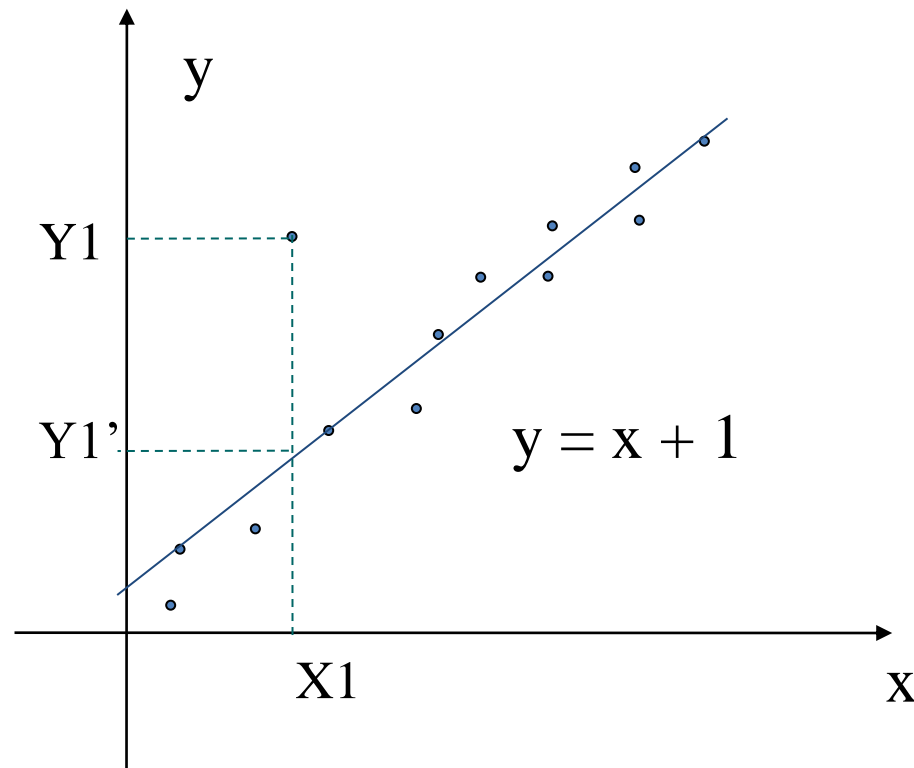
- **Clustering:** detecção e remoção de valores aberrantes.
 - os valores são organizados em grupos;
 - os valores isolados podem ser considerados aberrantes;



Dados com ruído e /ou valores aberrantes

- **Regressão:**

- os dados podem ser alisados pelo ajustamento a uma função (regressão linear, por exemplo);



Dados Inconsistentes

- Erros no momento de introdução dos dados
- Erros oriundos da integração de várias bases de dados
 - mesmo atributo com diferentes codificações;
 - duplicação de objetos
- Etc
- Ex: Nomes de municípios diferentes para o mesmo código de identificação na UF (pode ocorrer mudança de nome ao longo do tempo).

Integração de Dados

Esquema em bases de dados relacionais

- identificação das mesmas entidades do mundo real a partir de múltiplas fontes de dados
- Integração dos metadados de diferentes fontes

Redundância

Dados redundantes ocorrem quando da integração de bases de dados

- Diferentes nomes para um mesmo atributo;
- Um atributo pode ser derivado diretamente de outro;

Análise de correlação: instrumento para a detecção de redundâncias

Duplicação de objetos;

Integração de Dados

Detecção e resolução de conflitos

Os valores de um mesmo atributo pode diferir segundo as diversas fontes

Isso pode acontecer devido a diferenças na representação, escala ou codificação.

Ex:

- Peso (em libras ou em quilos)
- Altura (valor numérico ou categórico (médio, pequeno...))
- Preço (pode indicar serviços diferentes)
- Área espacial (Hectares ou Km²)

Transformações

- Generalização (Hierarquia de Conceitos)
- Mudança de escala (Mapeamento)
 - Numérico p/ Categórico
 - Categórico p/ Numérico

Generalização (Hierarquias de conceitos)

Utilizado quando os dados são muito esparsos e não se consegue bons resultados .

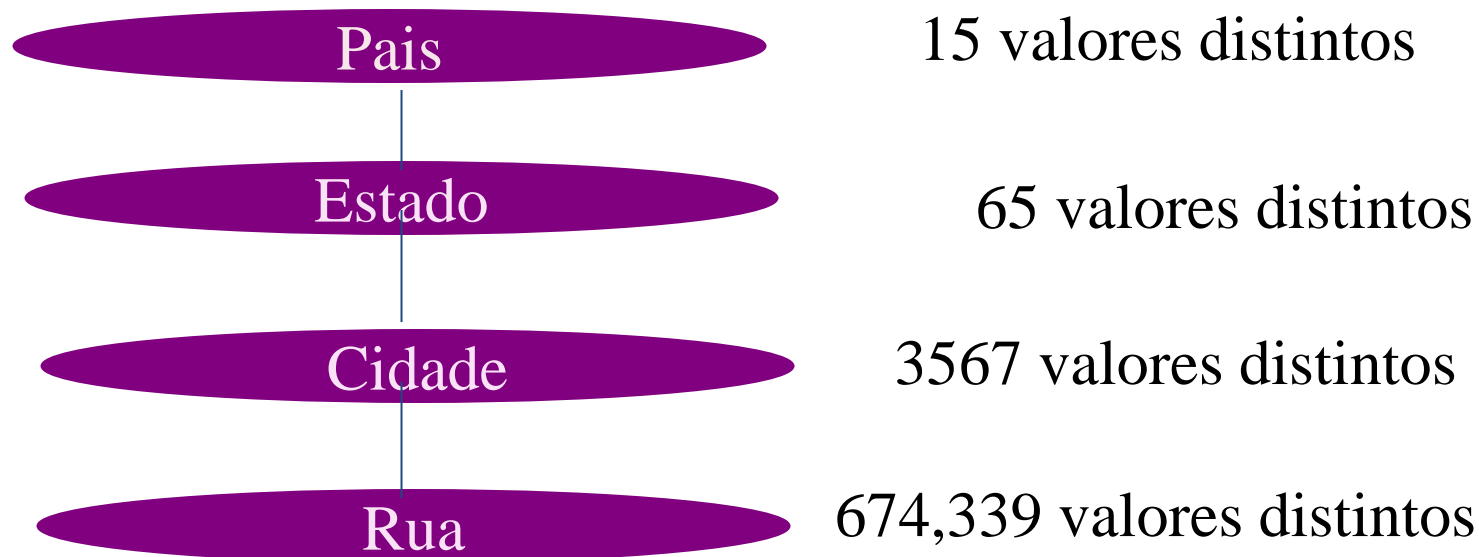
Então, dados primitivos são substituídos por conceitos de ordem superior via uma hierarquia de conceitos.

Exemplo:

- *calça, blusa, saia*, etc. são substituídos por *roupa*
- nomes de cidades são substituídas pelo nome do estado ao qual pertencem

Hierarquias de conceitos para dados categóricos

- Hierarquia conceitual pode ser gerada automaticamente com base no número de valores distintos por atributo. O atributo com o maior número de valores distintos é colocado no nível mais baixo da hierarquia.



Transformação numérico → categórico

Objetivo: transformação de valores numéricos para categóricos ou discretos

- Mapeamento direto
- Mapeamento em intervalos (discretização)

Transformação numérico → categórico

Mapeamento direto

- Objetivo: substituição de valores numéricos por valores categóricos

Exemplo: sexo

1 → M

0 → F

Transformação numérico → categórico

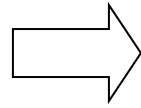
Mapeamento em intervalos (discretização)

- Objetivo: substituição de valores dentro de um intervalo por um identificador
- Identificador de intervalo:
 - Categórico: nome (sugestão: mneumônico)
 - Numérico
- Exemplo: número de dependentes

Num_Dep:	0 a 1	2 a 5	6 a 99
categórico	poucos_dep	media_dep	muitos_dep
numérico	0	1	2

Ex Mudança de Escala (Intervalo para Ordinal)

Intervalar



Ordinal

Ex: Idade $\mathbf{O} = [0, 150]$

0-20: *jovem*;

20-60: *adulto*;

>60: *idoso*

$\mathbf{O}' = \{\text{jovem, adulto, idoso}\}$

Trata-se de subdividir \mathbf{O} em subintervalos contíguos e associar a cada um deles uma modalidade

Transformação categórico → numérico

Objetivo: transformação de valores categóricos em numéricos

- Mapeamento direto
- Representação binária 1-de-N

Transformação categórico → numérico

Mapeamento direto

Mapeamento em valores de 1 a N

Est_Civil	mapeamento
Casado	1
Solteiro	2
Viúvo	3
Divorciado	4
Outro	5

Transformação categórico → numérico

Mapeamento direto

Quando o atributo categórico for **ordinal**, é importante que os valores numéricos sigam a mesma ordem

conceito	mapeamento
Ruim	1
Regular	2
Bom	3
Ótimo	4

Transformação categórico → numérico

Representação binária 1-de-N

- Mapeamento em número cuja representação binária tenha N dígitos
 - Somente um dígito é “1”

Est_Civil	Representação binária 1-de-N
Casado	00001
Solteiro	00010
Viúvo	00100
Divorciado	01000
Outro	10000

Redução de dimensão

Significa a redução do escopo dos dados em análise, além de mudar a ordem das dimensões, mudando desta forma a orientação segundo a qual os dados são visualizados.

Em DW é chamado de “Slice e Dice”.

Exemplo:

Volume de Produção (em milhares)		1999			
		Telefone Celular		Pagers	
		1001	1002	2001	2002
RS	Canoas	13	4	2	5
	Porto Alegre	20	8	6	7

Mudando a perspectiva para modelo de produto, produto, ano, estado e cidade ...

Volume de Produção (em milhares)		1999	
		RS	
		Canoas	Porto Alegre
Telefone Celulares	1001	13	2
	1002	4	5
Pagers	2001	2	6
	2002	5	7

Atividades



...