

2 **Framework para aprendizado de máquina**

2.1. **Processo, Características e benefícios**

O *framework* onde os algoritmos foram inseridos para validação, testes e posterior utilização tem como objetivo melhorar a qualidade dos dados sob diversos pontos de vista. O resultado deste estudo aplica-se a apenas uma parte do processo estabelecido, a Deduplicação. Mesmo a deduplicação é composta por várias partes, utilizando o resultado deste estudo como pré-processamento dos dados para execução posterior das demais partes.

O *framework* para qualidade de dados terá suas etapas descritas abaixo com maiores detalhes. Algumas destas etapas já utilizam técnicas de *machine learning* enquanto outras estão em fase de estudo e avaliação para implementação.

Em linhas gerais, o processo de *data quality*, realizado pelo *framework*, tem como objetivo garantir integridade e veracidade dos dados. Isto significa que, se o processo não for capaz de identificar, analisar, ponderar, decidir e corrigir duplicidades, redundâncias e inconsistências de dados, tanto em uma mesma base de dados quanto em bases de dados distintas, então não terá atingido os objetivos.

Assim, o processo de *data quality* deve, de forma criteriosa, confiável e rastreável, corrigir resíduos de digitação, validar dados como endereço, CNPJ ou qualquer campo passível de análise (criticável), padronizar campos de preenchimento livre, eliminar inconsistências (informações e estruturas de dados conflitantes) e, por fim, lidar com as duplicidades.

As vantagens comerciais da aplicação desse processo em bases de dados complexas nas quais dados cadastrais e/ou operacionais são não apenas volumosos (seja em termos de quantidade bruta de registros ou em termos de atributos que cada registro pode ter), mas também difusos, podem ser separadas em três categorias principais:

- Aumento de receita devido aos dados em ações comerciais, notadamente através de ações de Marketing de Relacionamento;

- Redução do desperdício em consolidações de dados e comunicações com clientes, *prospects*, e com mercado em geral; e
- Maior segurança operacional em função da redução de resultados errados causados por dados com qualidade fora de padrão.

Aprofundar a discussão sobre as vantagens comerciais do processo de *data quality* não é, entretanto, escopo do presente estudo que limitar-se-á a explicá-lo funcionalmente e detalhar porque ele tem que atender à seguinte definição:

Ser um processo tecnológico capaz de limpar, analisar, validar, comparar, recuperar, padronizar e qualificar dados de bancos de dados em busca de duplicidades, redundâncias, inconsistências e complementaridades, aprimorando seus algoritmos a partir dos próprios resultados obtidos e preparando os dados para eventuais realimentações quer em seus bancos de dados de origem quer em outros aplicativos.

A Figura 2 abaixo representa pictograficamente as etapas e operações envolvidas no processo de *data quality*.



Figura 2: Data quality

2.2. Descrição das etapas do processo

Apresentamos a seguir uma descrição detalhada das etapas do processo de *data quality* representadas na Figura 2. A abordagem aqui utilizada foi acrescentar à descrição de cada etapa o contexto no qual sua demanda é estabelecida.

2.2.1. Limpeza

A primeira etapa é a retirada de resíduos de digitação, sejam eles intencionais ou não. Os resíduos não-intencionais geralmente incluem caracteres especiais e acentuação específica de cada idioma.

É ainda necessário, porém, remover os resíduos intencionais. Estes incluem termos impróprios e/ou obscenos. Eles se encaixam em duas categorias – os previamente cadastrados e aqueles que ainda não foram cadastrados mas que, de alguma maneira, assemelham-se aos cadastrados. Ambos devem ser identificados, mas de maneira diferente, para decisão prévia do administrador sobre como o processo deve proceder em cada caso.

Para os previamente cadastrados, a decisão resume-se a eliminá-los ou não, uma vez que não haverá acesso à lista – ao administrador será permitido apenas incluir novos itens na série histórica.

Para os termos ainda não cadastrados, entretanto, o administrador poderá escolher o grau de proximidade destes com aqueles cadastrados. É necessário que tal distância seja indicada por um único indicador percentual. Estes registros devem ser identificados como “Potencialmente Impróprio e/ou Obsceno” e ter tal potencial indicado em termos percentuais relativos ao termo impróprio e/ou obsceno ao qual faz referência para que o administrador tome a decisão sobre eliminá-los, tanto individual quanto coletivamente.

A Tabela 2 a seguir apresenta um conjunto exemplo de registros onde podemos observar resíduos de digitação.

Id	Nome Completo	Primeiro Nome	Dia Nascimento	Mes Nascimento	Ano Nascimento	Estado Civil	Sexo	Profissao
90391	DERSON MONTEIRO DO NASCIMENTO	DERSON	26	JAN	1968		M	MILITAR
90401	RONALDO DANILLO	RONALDO	13	NOV	1971		M	MILITAR
90411	== FRAUDE!!! NAO VENDER!!!	EDMILSON	15	JAN	1969		M	MILITAR
90421	CLOVIS DE SOUSA E SILVA	CLOVIS	28	NOV	1959		M	MILITAR
90491	PAULO GOMES	PAULO	28	OCT	1969		M	RELACOES PUBLICAS
90897	RICARDO DUTRA AYDOS	RICARDO	11	AUG	1952		M	PROFISSIONAL LIBERAL
90907	TSSTE	ANTONIO	08	SEP	1951		M	VETERINARIO
90621	ANDRE BESERRA DA SILVA	ANDRE	23	FEB	1974		M	SUPERVISOR VENDAS

Tabela 2: Identificação de termos impróprios e/ou obscenos

Termos impróprios tais como os ilustrados pelos registros da Tabela 2, quando não identificados, podem gerar, além dos evidentes desperdício e desconforto, processos judiciais. Pode-se observar como exemplo o registro Id 90411. Uma carta enviada a este endereço seria um desastre. No lugar do destinatário estaria “FRAUDE” e no corpo da carta, “EDMILSON”.

2.2.2. Validação

Validar é definido aqui como verificar se determinado dado segue uma ordem de formação e/ou consistência. Casos típicos são o CPF, com sua validação através de uma fórmula conhecida como “Código 11”, e o Conjunto Endereço. Por ser mais complexa e envolver diversos elementos, a formação do Conjunto Endereço foi escolhida para especificar como o processo tecnológico deve abordar os casos de validação.

O Conjunto Endereço é formado por: Tipo de logradouro, Logradouro, Número, Complemento, Bairro, Cidade, CEP e Estado. Estes campos podem se apresentar juntos, separados ou ainda parcialmente separados – não há regra definida quanto a isso, e cada administrador escolhe sua maneira de armazenar os dados do Conjunto Endereço.

À exceção da situação em que o Conjunto Endereço se apresenta completamente separado, a primeira ação é separar seus termos. É aplicado um *parser* para reduzir o Conjunto Endereço à suas partes fundamentais.

Essa separação pode ser bastante complexa, notadamente naqueles casos nos quais os elementos não seguem a ordem Tipo de Logradouro, Título, Logradouro, Número e Complemento. Como o processo tecnológico deve, por definição, ser capaz de aprimorar seus algoritmos a partir dos próprios resultados obtidos, tem como principal componente um *parser*.

Este *parser* deve ser capaz de se reprogramar para, após atuar em 80% dos casos, reprocessar os 20% restantes de outra(s) maneira(s) até esgotá-los. O *parser* utiliza diferentes tecnologias para alcançar os objetivos traçados, tanto técnicas de aprendizado de máquina, quanto heurísticas desenvolvidas especificamente para atender à demanda.

Na verdade, isso implica dizer que o processo será capaz de aprender e incorporar a seu processo decisório as particularidades que encontrar.

Após sua separação, o Conjunto Endereço passa por um processo de equalização de grafias com uma base-verdade – a base da Correios S.A. Para que esta comparação seja mais eficiente, cada componente tem sua grafia equalizada, primeiro com uma série histórica, depois comparando com a base-verdade da Correios S.A.. Essencialmente, quanto mais completa for esta equalização de grafias, menor será o grau de esforço tecnológico a ser empregado na validação do Conjunto Endereço por conta de seus elementos básicos.

Por fim, após a padronização, é feita a decisão a respeito da validação do Conjunto Endereço. A validação depende da distância de entre cada um dos termos padronizados e seu equivalente na base-verdade da Correios S.A. Tais distâncias são então ponderadas para alcançar a distância total de cada Conjunto Endereço e gerar um *ranking* dos mais parecidos.

Em função da distância total e dos parâmetros de validação (linha de corte) previamente definidos pelo administrador, cada Conjunto Endereço recebe dois *status* sendo um geral, que indica se ele está pronto para utilização, e outro específico, através do qual cada componente do Conjunto Endereço tem indicadas as transformações que sofreu. A Tabela 3 apresenta um conjunto de registros que devem ser validados pelo processo.

Id	Tp	Logr	Num	Compl	Bairro	Cidade	UF	CEP
1		Avenida Padre Leonel Franca 480, Sala 23 - Gávea - Rio de Janeiro - RJ - 22451-000						
2		Avenida Padre Leonel Franca 480, Sala 23 - Gávea - Rio de Janeiro - RJ						22451-000
3		Avenida Padre Leonel Franca 480, Sala 23 - Gávea - Rio de Janeiro					RJ	22451-000
4		Avenida Padre Leonel Franca 480, Sala 23 - Gávea				Rio de Janeiro	RJ	22451-000
5		Avenida Padre Leonel Franca 480, Sala 23			Gávea	Rio de Janeiro	RJ	22451-000
6		Avenida Padre Leonel Franca	480	Sala 23	Gávea	Rio de Janeiro	RJ	22451-000
7	Avenida	Padre Leonel Franca	480	Sala 23	Gávea	Rio de Janeiro	RJ	22451-000
8	Avenida	Padre Leonel Franca	480	Sala 23	Gávea	Rio de Janeiro	RJ	22451-000

Tabela 3: Evolução da validação dos dados

Um mesmo endereço pode se apresentar em vários momentos de separação. No registro Id 1 do exemplo acima ele aparece totalmente concatenado, ao passo que no Id 8 a separação é total.

Tipicamente os dados estão em estágios intermediários de separação. Há geralmente mais de um padrão em uma mesma base, obrigando que a separação seja um processo adaptável.

A Tabela 4 a seguir mostra um conjunto de endereços já validados após tratamento e separação dos dados. A coluna *status* indica quais transformações cada componente sofreu.

D	CPF	W	ENDEREÇO	NUMERO	COMPLEMENTO	BAIRRO	CIDADE	ESTADO	CEP	STATUS
D	00459409751	T	RUA ALFA	SN	LT 7 QDC	JARDIM ESTRELAS	BELFORD ROXO	RJ	26160160	1
D	00616790031	T	RUA COROADOS	159	CS	VL MTE CARLO	CACHOEIRINHA	RS	94940070	1
D	01717721990	T	AVENIDA REPUBLICA ARGENTINA	394	AP 601	AGUA VERDE	CURITIBA	PR	80240210	1
D	01998387712	T	RUA NICOLINO SOARES	18	CS	ILHA FLORES	VILA VELHA	ES	29115605	1
D	02088181709	T	RUA PORTO PRINCIPE	33	CS	VIGARIO GERAL	RIO DE JANEIRO	RJ	21241210	1
D	02202545832	T	RUA HUELVA	36	B CS	JD S FRANCISCO	SAO PAULO	SP	04918040	1
D	03251312790	T	RUA FILEUTEERPE	329	CS	SAO PEDRO	TERESOPOLIS	RJ	25955100	1
D	03404540751	W	RUA ISMENIA MENEZES	SN	LT 06 QD 08	JARDIM JUREMA	SAO JOAO DE MERITI	RJ	25580231	1
D	03680825803	W	R J D M M COSTA	241	CS	VL A LUIZ	CACAPAVA	SP	12287350	1
D	01074221079	W	RUA ENGENHEIRO THOMAS PAES CUNHA FILHO	236	CS	PRQ RES GOELHO	RIO GRANDE	RS	96202770	1
D	02230393731	T	RUA IARA ROCHA	SN	LT 8 QD 28	MIRIAMI	SAO GONCALO	RJ	24731110	1

Tabela 4: Endereços validados após tratamento

O *status* 1 indica que o Conjunto Endereço passou por validação e/ou padronização de componentes mas não sofreu alteração de CEP pois o original estava correto.

A Tabela 5 apresenta registros que não foram validados. O *status* 3 indica que não foi possível recuperar o CEP dentro do intervalo de segurança para o Conjunto Endereço. Isto acontece porque ou há mais de um candidato habilitado e o sistema não tinha elementos suficientes para decidir, ou não havia nenhum candidato no intervalo de segurança.

D	CPF	W	ENDEREÇO	NUMERO	COMPLEMENTO	BAIRRO	CIDADE	ESTADO	CEP	STATUS
D	00114762147	W	GRANJA 143 SADIA	143		NUC RURAL SANTO	PLANALTIMA	DF	73350000	3
D	37026534068	W	RUA INTERNA C ROSA VENTOS	00023	CS	SAO CAETANO	CAXIAS DO SUL	RS	95095390	3
D	56140606691	T	ILHA 16			ILHA	GOV VALADARES	MG	35010000	3
D	56765550097	W	RUA ILHA LEONIDIO	624	CS	ILHA DO LEONIDI	RIO GRANDE	RS	96200970	3
D	84133664787	W	ESTRADA GENIPAPO	6	C 2	MORRO DO CASTRO	NITEROI	RJ	24130670	3
D	00086291700	W	VIA B		3 VL VL DO PINH	MANGUINHOS	RIO DE JANEIRO	RJ	21040000	3
D	00718090785	T	RUA PRINCIPAL	SN	SN CS	SABONETE	CAMPOS DOS GOYTAC	RJ	26225971	3
D	01432220152	W	CHACARA 108	108		NUCLEO RURAL JA	PARANCA	DF	73350000	3
D	81185456963	T	RUA TRAVESA JATAIBA	156		ARMACAO DO PANT	FLORIANOPOLIS	SC	88061130	3
D	82319343991	W	FFFFFFFFFFFFFFFFJJJJJJJJJJJJJJ	12	CS	XXXXXXX	PELOTAS	RS	96001970	3
D	85102647172	T	ELIA	21	346 QD 45 LT 26	SETOR CARAVELAS	GOIANIA	GO	74354644	3
D	98849638787	W	CASA CANTAGALO 226 CASA			CANTAGALO	TRES RIOS	RJ	25800000	3

Tabela 5: Endereços sem validação

2.2.3. Padronização

A Padronização é aplicada em campos de preenchimento livre que não estão sob o jugo de uma regra de formação. Assim, o objetivo de nosso processo é trazer para o mesmo valor registros que têm, essencialmente, a mesma informação e/ou função, embora possuam grafias diferenciadas ou até mesmo fundamentalmente distintas. Este trabalho pode ser definido como a identificação intrínseca de padrões intra-dados. Padrões que não são evidentes à primeira vista são sugeridos.

Através da padronização o administrador pode fazer extrações precisas na base, sem correr o risco de perder informações ou de obter respostas inexatas causadas, por exemplo, por diferentes unidades de medidas e/ou valores utilizados em um mesmo campo.

Há dois tipos de padronização, dependendo da presença ou não de uma base-verdade para onde os dados devem convergir.

Nos casos em que há uma base-verdade, agiremos como na validação e, mais uma vez, mediremos a distância entre cada registro e seus equivalentes na base-verdade. Para tanto, a série histórica é de vital importância. A padronização ou não de cada dado dependerá sempre dos parâmetros de aceitação definidos previamente, lembrando sempre que serão apresentados como percentuais compostos de verossimilhança.

Se para os casos com base-verdade os procedimentos são bastante similares àqueles adotados na validação, para aqueles em que não há base-verdade, o procedimento é mais complexo. Situações deste tipo exigem que o processo tecnológico crie os itens da base-verdade a partir dos dados que serão padronizados, infira as categorias e os dados mais prováveis de cada uma delas e, por fim, apresente sugestões para decisão do administrador, dando a este a opção de aprovar as sugestões integral ou individualmente.

A decisão a respeito da inclusão dos padrões sugeridos na base-verdade é feita pelo administrador. O processo deve permitir tanto o aceite integral quanto individual das sugestões.

Como a ausência de base-verdade torna a padronização mais complexa, tarefas como comparar, classificar e inferir passam a ser críticas. Um mecanismo de *feedback* gera respostas para uma melhoria contínua de desempenho do processo.

A Tabela 6 reúne registros cujos dados foram originados de operações de negócio realizadas em diversos contextos geográficos.

Plant	Volume (Cubic Meters)	Volume (Cubic feet)	Value (Local \$)	Value (US\$)
Brazil	652,118	23,029,221	\$ 6,559,151.50	\$ 3,339,237.06
Australia	1,407,655	49,710,632	\$ 8,739,258.00	\$ 7,208,041.71
Argentina	398,981	14,089,815	\$ 6,371,422.00	\$ 2,043,023.11
Mexico	501,409	17,707,008	\$ 27,805,818.00	\$ 2,567,516.18
China	23,451,220	828,168,109	\$ 921,070,020.00	\$ 120,084,375.76
India	15,567,891	549,772,287	\$ 3,219,442,180.00	\$ 79,716,981.57
Belgium	208,320	7,356,717	\$ 789,596.25	\$ 1,066,723.91
UK	443,078	15,647,078	\$ 1,148,500.20	\$ 2,268,826.31
US	17,554,300	619,921,327	\$ 89,888,592.47	\$ 89,888,592.47
Canada	11,145,321	393,591,438	\$ 62,115,816.00	\$ 57,070,758.58
Total	71,330,293	2,518,993,632	-	\$ 365,254,076.66

Tabela 6: Padronização de unidades monetárias e de medidas

Embora o exemplo ilustrado pela Tabela 6 seja simples, mostra a dificuldade de prover dados úteis aos tomadores de decisão em operações globais. Problemas semelhantes ocorrem com embalagens e até mesmo com nomes de produtos.

A Figura 3 apresenta um problema de padronização causado por diversos tipos de sinonímia para referenciar uma mesma entidade.

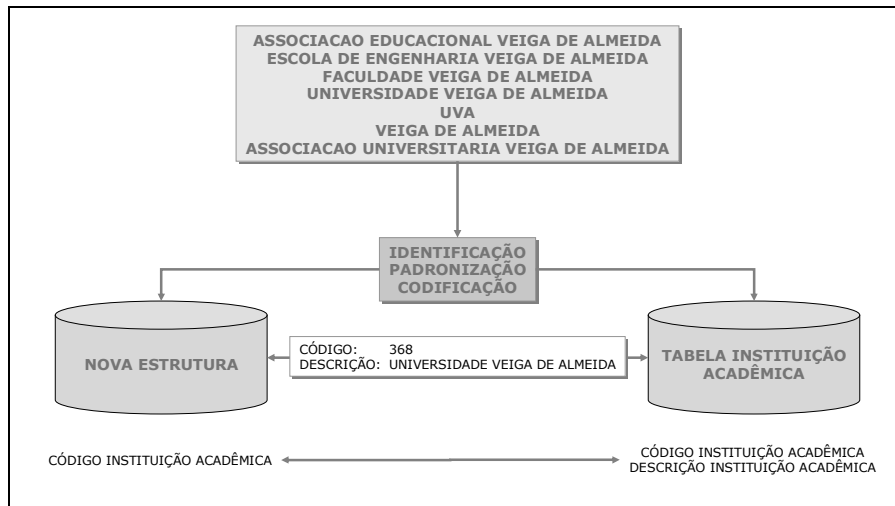


Figura 3: Padronização de nomes de universidades

Aqui, embora todos os nomes façam referência à mesma instituição, grafias diferenciadas levam a resultados diferentes em extrações de dados. O processo de padronização faz com que todos os casos idênticos sejam localizados pelo administrador.

Veículo - DE	Veículo - PARA MARCA	Veículo - PARA MODELO	Veículo - PARA VERSAO	Veículo - PARA LP	Veículo - PARA MICRO-SEGMENTO
VOLKSWAGEN PASSAT - 2.0 VARIANT GL (95)	VOLKSWAGEN	PASSAT		TC	PERFORMANCE H
VOLKSWAGEN PASSAT - 2.8 VARIANT VR6 (95)	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT - 2.8 VR6 - (95)	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT 1.8	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT 2.8 V6	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT 2.8 VR6	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT 2P	VOLKSWAGEN	PASSAT	2P	TC	CORE MARKET COUNTRY B
VOLKSWAGEN PASSAT 4P 1.8 TURBO	VOLKSWAGEN	PASSAT	4P 1.8 TURBO	TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT 4P 2.0	VOLKSWAGEN	PASSAT	4P 2.0	TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT GTS POINTER	VOLKSWAGEN	PASSAT	GTS POINTER	TC	ESPORTY COSMETIC
VOLKSWAGEN PASSAT L 2P	VOLKSWAGEN	PASSAT	L 2P	TC	CORE MARKET COUNTRY B
VOLKSWAGEN PASSAT LS 3P	VOLKSWAGEN	PASSAT	LS 3P	TC	CORE MARKET COUNTRY B
VOLKSWAGEN PASSAT LS 4P	VOLKSWAGEN	PASSAT	LS 4P	TC	CORE MARKET COUNTRY B
VOLKSWAGEN PASSAT TURBO 1.8	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT VARIANT 1.8 / 1.8 TURBO	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT VARIANT 2.0	VOLKSWAGEN	PASSAT		TC	PERFORMANCE H
VOLKSWAGEN PASSAT VARIANT 2.8 VR6	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT VARIANT 4P 1.8 TURBO	VOLKSWAGEN	PASSAT	VARIANT 4P 1.8 TURBO	TC	HIGH PERFO HP LC
VOLKSWAGEN PASSAT VARIANT 4P 2.8 VR6	VOLKSWAGEN	PASSAT	VARIANT 4P 2.8 VR6	TC	HIGH PERFO HP LC
VOLKSWAGEN VARIANT TURBO	VOLKSWAGEN	PASSAT		TC	HIGH PERFO HP LC
VW PASSAT VARIANT 1.8 T	VOLKSWAGEN	PASSAT	VARIANT 4P 1.8 TURBO	TC	HIGH PERFO HP LC

Tabela 7: Padronização de veículos

No caso exemplificado pela Tabela 7, há uma base-verdade na qual cada registro deve possuir um equivalente. Antes da Padronização os dados passaram por uma separação efetuada pelo *parser*.

O próximo exemplo trata de uma base de cadastro de funcionários e padroniza os cargos aos quais estes registros fazem referência.

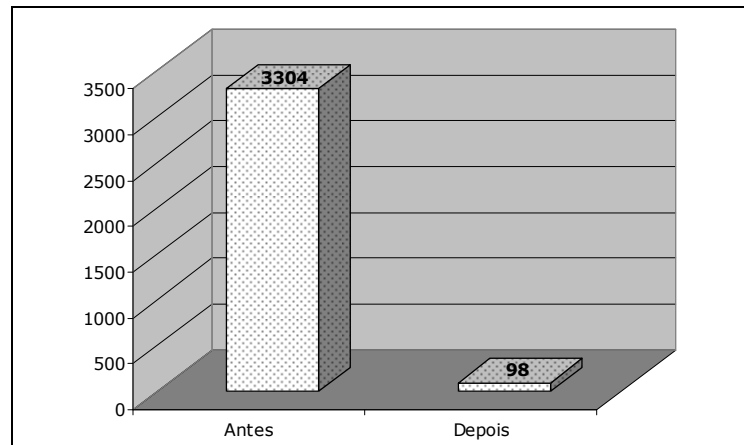


Figura 4: Resultado da Padronização de cargos

A Figura 4 mostra os resultados da padronização em uma base de dados cuja quantidade de cargos foi reduzida de mais de três mil para menos de cem, ou seja, menos de 3% do original.

2.2.4. Qualificação

Qualificar é aumentar o estoque de informações a partir dos próprios dados. Isto é possível porque muitos dados trazem indicações que podem e devem ser usadas para qualificá-los. Usualmente, há um tipo de dados que é qualificável por si. Trata-se do nome.

O processo será capaz de, a partir do nome, tomar três ações diretas. A primeira é a identificação da natureza jurídica. Em muitos bancos de dados, pessoas físicas e jurídicas não são separadas, o que dificulta diversas ações como classificação de clientes ou atualização de contatos. Para tanto, temos que identificar a natureza jurídica de cada nome. Fundamentalmente, isto é feito buscando marcadores de pessoa jurídica, como Ltda., ME, Cia. etc. no texto. Este processo, embora não seja exatamente complexo, é de grande valor para o administrador.

A segunda ação é a determinação de sexo. Pelo prenome, são atribuídas as categorias de gênero, a saber: Masculino, Feminino, Indeterminado e Desconhecido. Em função disso, produtos são segmentados e comunicações são personalizadas, resultando em maior acurácia e foco nas ações comerciais que os administradores porventura venham a tomar.

Por fim, há a identificação de nomes compostos. O objetivo aqui é que o tratamento seja feito com o nome completo, não apenas com o primeiro nome.

Esta pequena inteligência de localizar no nome um segundo prenome (nome composto) evita chamar “João Carlos” de “João”, “Maria Claudia” de “Maria”, etc.

ID	NOME	TIPO	SEXO	PRENOME
101857	JULIO BRAMMART	PF	M	JULIO
101861	RODRIGO SALLES SCHULTZ	PF	M	RODRIGO
101876	NILTON EDUARDO SANTOS	PF	M	NILTON EDUARDO
103894	COMPROCRED FOMENTO MERCANTIL	PJ	I	
103939	CENTRAL DE AR COMPRIMIDO LT	PJ	I	
102082	ALINE VIEIRA	PF	F	ALINE
103728	CENTRO DE FORTMACAO C JANAINA	PJ	I	
102183	FRANCISCA NEVES	PF	M	FRANCISCA
102413	ARLINDO FRANCISCO CASTRO FILHO	PF	M	ARLINDO FRANCISCO
102421	RONALDO LIMA COUTO	PF	M	RONALDO
103441	CAMBRAIA E ROSA COM VEICULOS	PJ	I	
104008	JOAO CARLOS SAMPAIO FILHO	PF	M	JOAO CARLOS
104016	CRIMAC COM DE PROD GRAF	PJ	I	

Tabela 8: Qualificação

Nos exemplos ilustrados pela Tabela 8 foram identificados natureza jurídica, sexo e nomes compostos. Todos estes campos foram derivados automaticamente do nome.

2.2.5. Segmentação

Há duas segmentações principais. A primeira é a análise de correlações que consiste em uma busca de evidências de comportamento relacionado. Isto significa procurar características comuns em registros que tiveram determinado comportamento para buscar na base registros que, embora possuam as mesmas características, não apresentam o mesmo comportamento, mas têm propensão a fazê-lo. De posse destes dados, são feitas inferências probabilísticas e, com base nos resultados desta análise matemática, ações são tomadas.

Outra análise é a de valoração ponderada. Aqui, o objetivo é elaborar *rankings* de registros. Como há muitas segmentações possíveis, apresentaremos as três mais conhecidas para descrever o que o processo pode gerar. São as classificações RFM (*Recency, Frequency, and Monetary Value*), RFD (*Recency, Frequency, and Duration*) e LTV (*LifeTime Value*). A seguir é apresentada uma breve explicação de cada uma delas seguida de um exemplo.

A classificação RFM elabora *ranking* de registros a partir de atributos definidos de recência (quando foi a última interação?), frequência (quantas interações foram feitas?) e valor (quanto custou ou, alternativamente, quanto rendeu cada interação?). Através de uma análise CHAID (*CHI-squared Automatic Interaction Detector*), até 10 categorias são indicadas por item.

Por exemplo, para *Recency* de interações em uma montadora de automóveis, os clientes podem ser divididos por recência de compra em “Até 1 semana”, “De 1 a 2 semanas”, “De 2 semanas a 1 mês”, “De 1 a 3 meses”, “De 3 a 6 meses”, “De 6 meses a 1 ano”, “De 1 a 2 anos”, “De 2 a 5 anos”, “De 5 a 10 anos” e “Mais de 10 anos”.

Para cada grupo, uma ação diferente é adotada. Ainda em nosso exemplo, quem comprou há menos de 1 semana recebe um *kit* de boas vindas, quem comprou de 2 a 5 anos recebe propostas de desconto na troca por um modelo mais novo etc.

Assim, combinando as 10 categorias dos 3 itens (*Recency*, *Frequency*, and *Monetary Value*) chegamos a 720 possíveis grupos. As definições, tanto do número de categorias quanto dos respectivos escopos são customizáveis, e o administrador pode tanto aceitar as sugestões do CHAID quanto utilizar suas próprias categorias.

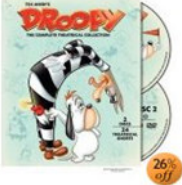
A classificação RFD é uma variante da RFM. A diferença reside no fato que a RFD não avalia quanto custou e/ou rendeu cada interação, e sim quanto tempo durou cada uma delas. Esta classificação é particularmente útil para dimensionar e monitorar estruturas de atendimento (*call centers*, sites etc.).

Para encerrar o exemplo, o LTV, onde a segmentação é feita a partir do retorno líquido de cada registro ao longo de um determinado período. As receitas (retorno em cada interação, normalmente valor das vendas) e as despesas (custo em cada interação, normalmente diretos como custo do produto ou serviço vendido e indiretos como custo de atendimento e manutenção) são apuradas e descontadas para obter um único valor. E a partir deste valor os registros são segmentados de acordo com o padrão 80-20.

amazon.com.

Dear Amazon.com Customer,

We've noticed that customers who have expressed interest in [Looney Tunes - Golden Collection, Volume Three](#) have also ordered Tex Avery's Droopy - The Complete Theatrical Collection on DVD. For this reason, you might like to know that Tex Avery's Droopy - The Complete Theatrical Collection will be released on DVD on May 15, 2007. You can pre-order your copy at a savings of \$6.99 by following the link below.



[Tex Avery's Droopy - The Complete Theatrical Collection](#)
Bill Thompson

List Price:	\$26.98
Price:	\$19.99
You Save:	\$6.99 (26%)

Release Date: May 15, 2007
Other Versions and Languages
VHS Tape
[Adventures of Droopy](#)




Figura 5: Ação segmentada de venda

A Figura 5 apresenta um exemplo onde a empresa decide ofertar o produto com desconto para clientes que tiveram o mesmo comportamento inicial.

A segmentação é um processo fortemente baseado em séries históricas. Nesse contexto, quanto mais dados acumulados, mais eficiente será a segmentação.

2.2.6. Deduplicação

Deduplicar é o processo de comparar dados, completos ou não, e definir se os mesmos são redundantes ou conflitantes para separar apenas os que estão de acordo com os parâmetros definidos, eliminando assim registros duplicados e garantindo que os remanescentes possuam todo e qualquer dado relevante, seja nativo ou oriundo de outro registro definido como redundante ou conflitante.

A deduplicação consiste em (a) identificação de registros duplicados; (b) correção dos mesmos com base no conjunto das informações das duplicatas e (c) criação de uma nova base de dados como resultado do processo.

Para os administradores, deduplicação é um desafio imperativo em empresas que possuem informações espalhadas em mais de um banco de dados, onde a ocorrência de duplicidades e dados conflitantes reduz a confiabilidade das decisões, afeta resultados, aumenta o desperdício e, em função da pouca eficiência decorrente, inibe projetos e investimentos.

Por fim, a pré-condição para a deduplicação é os dados terem passado pelos processos de limpeza, validação e padronização. Para deduplicar, é necessário que o administrador defina a ordem de prevalência de bases e campos. A partir disso, são identificados registros duplicados e estes são então unificados e, havendo sobreposição e/ou conflito de dados, aplica-se o critério de prevalências.

NOME	CPF	NASCIM	PROFISSÃO	EMPRESA	UNIVERSIDADE	CURSO
CLAUDIA J. ALMEIDA	00497821880	12/12/97	ANAL. SISTEMA	BB		SISTEMAS
CLAUDIA J. ALMEIDA	00497821880			BANCO BRASIL	VEIGA DE ALMEIDA	
CLAUDIA J. ALMEIDA		12/12	ANALISTA SIST.			T.P.D
CLAUDIA ALMEIDA	00497821881		A SISTEMAS		UYA	
CLAUDIA ALMEIDA		12/12/67		BBRASIL	UNIV V ALMEIDA	ANAL. SIST.
↓	↓	↓	↓	↓	↓	↓
CLAUDIA J. ALMEIDA	00497821880	12/12/67	ANALISTA DE SISTEMAS	BANCO DO BRASIL	UNIVERSIDADE VEIGA DE ALMEIDA	TECNÓLOGO EM PROCESSAMENTO DE DADOS

Tabela 9: Deduplicação

O exemplo explicitado na Tabela 9 mostra um registro gerado ao final do processo. A ordem de prevalência e a qualidade dos dados impactam diretamente no resultado.