Tratamento de Dados

- Data Quality -

DataMotion Tecnologia e Serviços
Rua Gomes de Carvalho, 921 – 1 andar
04547-003 – São Paulo – SP
(11) 3842-2616/3045-9004
www.datamotion.com.br

Tecnologia para Tratamento de Dados

CENÁRIO

As empresas em geral possuem formatos e localizações múltiplas nas quais os dados estão armazenados. Visando suportar a tomada de decisão, aprimorar a performance dos sistemas ou mesmo atualizar sistemas existentes, os dados frequentemente devem ser copiados, movidos, replicados ou mesmo devem sofrer transformações de uma localização para outra. O conceito do DataMotion DataQuality visa justamente oferecer uma solução para essa demanda.

DataMotion é uma tecnologia de componentes específicos para processos de Data Quality, ETL, Validação de Dados Cadastrais e Filtros de Entrada de Dados. Dada sua arquitetura e seus métodos de acesso, o DataMotion pode ser aplicado nos mais diversos cenários, onde necessita-se de higienização e correção de dados, transformação, recodificação e conversão de conteúdo.

Apesar do papel crítico de uma ferramenta de Data Quality & ETL em uma corporação, seu uso tende a ser bastante genérico. O DataMotion, seguindo essa premissa, possui uma arquitetura aberta e flexível para atender as principais demandas do usuário.

Processos de CRM, ERP, SCM, BI, Data Warehouse, Knowledge Management, Data Formation, Cadastros, Cobrança, Anti-Fraudes, enfim, em todas as áreas e segmentos onde a Qualidade dos Dados é ponto focal, o DataMotion pode ser utilizado.

APLICAÇÕES

Exemplos de aplicações e usos típicos do Data Quality:

- ✓ Campanhas de Marketing Direto
- ✓ Validação de Dados Cadastrais
- ✓ Consistência Cadastral em tempo de Entrada de Dados
- ✓ Transformações e validações de conteúdos em tempo de execução
- ✓ Consolidação de diversas bases de dados, com layouts distintos
- ✓ Processamento e conversão de dados para ERP e CRM
- ✓ Integração de arquivos de diferentes fontes em um único repositório
- ✓ Definição de regras "de/para" durante etapas de migração de bases
- ✓ Automação e Encadeamento de processos
- ✓ Preparação dos dados para projetos de Business Intelligence e Data Mining
- ✓ Preparação dos dados para Projetos de Data Mart e Data Warehouse
- ✓ Suporte na Administração e suporte em Pesquisas de Mercado e Enquetes
- ✓ etc

<u>Processos de Validação contemplados pelo DataMotion</u>

Conteúdos Geográficos

- Tratamento de endereços – validação, padronização e separação do logradouro, complemento, bairro, cidade, UF, CEP e Código do IBGE
- Validação e atualização de CEP contra o DNE dos Correios
- Separação e padronização dos componentes do logradouro (tipo de logradouro, logradouro, número, complemento, Bairro, CEP, Cidade e UF)
- Integração com CEPNet dos Correios
- Georreferenciamento a partir do endereço

Nomes

- Atribuição de gênero (sexo)
- Identificação de tipo de pessoa – física ou jurídica
- Identificação de palavrões – nomes incorretamente digitados
- Separação de nome composto, primeiro nome, nome do meio e último nome

Telefones

- Tratamento de Telefones – validação e atualização de DDDs e prefixos

Documentos

- Validação do dígito de controle e formatação de CNPJ, CPF e Inscrição Estadual

E-mail

- Verificação e consistência no conteúdo do campo e-mail
- Ping para verificação da existência do domínio

Campos genéricos

- Padronização e formatação de campos genéricos como Cargos, Tabela de Produtos, Parentesco, Estado Civil, etc...
- Integração com expressões regulares (RegExp)

MatchCode

- Identificação de registros duplicados no cadastro
- Visão única de Cliente
- Householding
- Parametrização e regras de negócios customizáveis
- Criação de MatchKeys Fonéticas
- Geração de arquivo DE/PARA
- Merge & Purge em campanhas de Marketing/CRM

Desempenho e Acuracidade

- O DataMotion tem capacidade para processar mais de 4 milhões de registros por hora, mesmo com o módulo de Debug habilitado.

CARACTERÍSTICAS

Em termos de funcionalidade, o DataMotion é compatível com os mais diversos ambientes, possuindo perfeita integração com os principais gerenciadores de banco de dados, entre eles: SQLServer, Oracle, DB2, MySQL, MSAccess, etc.

A Tecnologia DataMotion é disponibilizada nas versões Batch e Online (Transacional) e pode ser instalada em virtualmente qualquer versão do Windows:

- Através de ferramentas específicas, o usuário terá a disposição, diversos recursos para melhorar sua produtividade em operações envolvendo transformação, migração e pesquisa de dados. A versão Batch é indicada para processos que tratam diretamente as bases de dados, sem necessariamente exigir integração com aplicações internas/externas da empresa.

- A versão Online permite que todo o acervo de funções disponíveis no DataMotion possa ser integrado Online e Realtime com qualquer tipo de aplicação existente na empresa, seja na Internet, ERP, CRM, etc. A versão Online é ideal para validação de processos de entrada de dados, consistências cadastrais, etc.

- Caso a necessidade do usuário seja utilizar os componentes do DataMotion dentro de uma aplicação qualquer, seja ela um processo de validação, de consolidação de base de dados, etc, é disponibilizado também o SDK (Software Development Kit) com acesso as todas funções primitivas do DataMotion.

- Através do "DataMotion Studio", o usuário poderá definir todas as regras e premissas que deverão ser aplicadas durante todo o processo de tratamento de dados.

O DataMotion é totalmente compatível com o MSOffice e o SQLServer. A tecnologia é plug & play com o próprio DTS (Data Transformation Services) ou Integration Services, fazendo com que, por exemplo, um processo de ETL possa acessar diretamente os métodos e classes disponíveis no DataMotion. Toda e qualquer aplicação que possa interagir com componentes COM/DLL da Microsoft, ou possa consumir métodos via XML Web Services, está apta a interagir com o DataMotion.

Facilidades e recursos técnicos específicos do produto:

- Acervo com centenas de métodos voltado a Transformação, Recode e Tratamento de String de Dados
- API integrável com qualquer aplicação Windows e Web IIS
- Explorador de Dados com sofisticados recursos de Query
- Assistente de Importação/Exportação de dados (TXT, MDB, DBF, XLS, SQLServer, Oracle, DB2, etc)
- Recursos avançados de "Procura e Substituição" de conteúdos
- Disponível client (front end) em Excel
- Acionamento automático de objetos
- Envio de mensagens e notificações
- Inteiramente extensível a componentes do usuário (Plug-ins)
- Disponibiliza trace de execução passo a passo (debug);
- Registro de mensagens de log customizado
- Possibilidade de ser executado via linha de comando
- Suporte a XML JSON WebServices
- Gerenciador de Dicionário de Dados
- Interfaces/Customizações via VBScript ou PascalScript
- Uso de Regular Expressions
- Totalmente integrável ao Windows Scheduler
- etc

Deduplicação de Conteúdo e Mecanismo de Busca

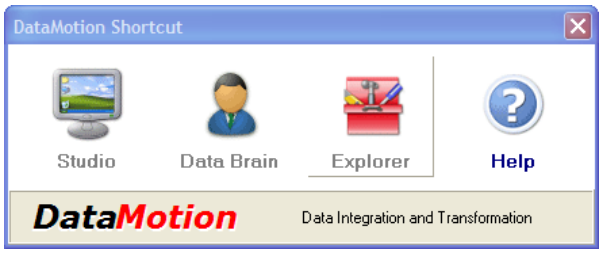
O DataMotion é 100% desenvolvido no Brasil com fonética da língua portuguesa. A interface gráfica, bem como, toda documentação do produto está disponível nos principais idiomas ocidentais.

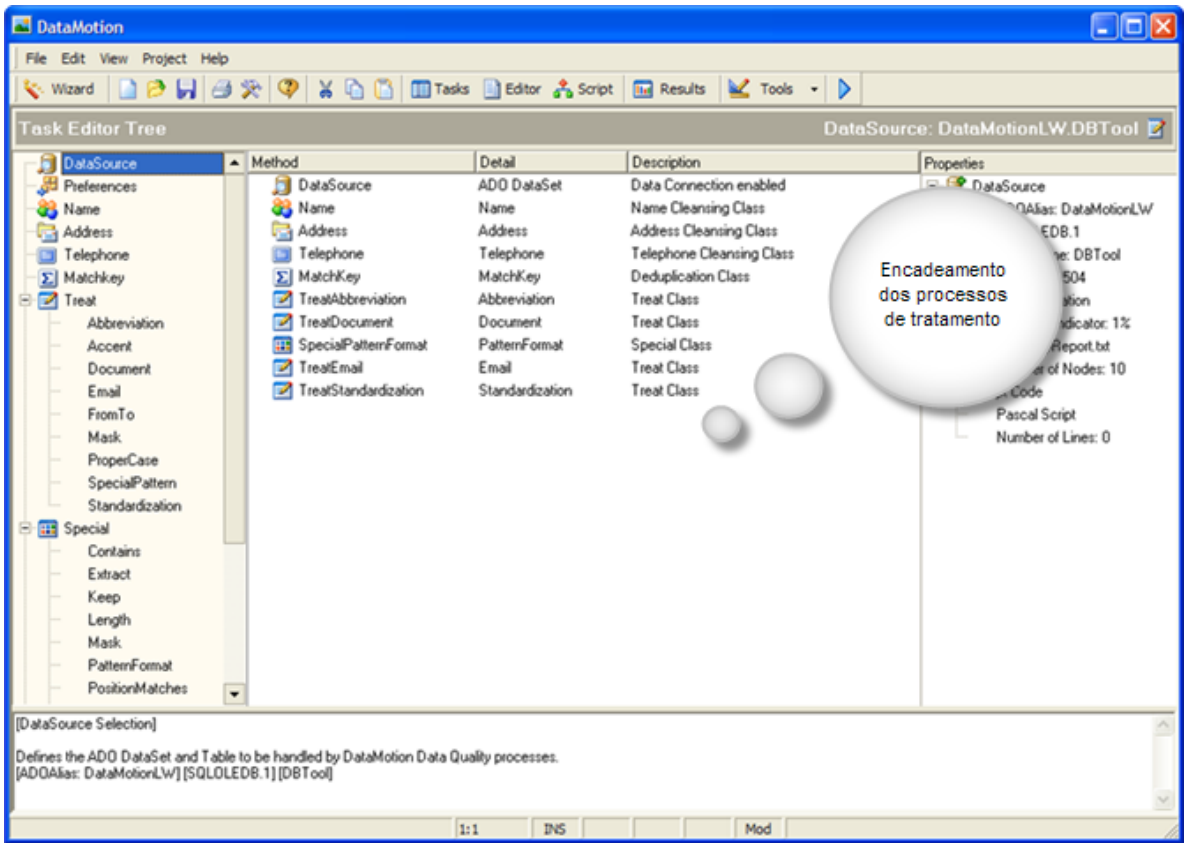
O DataMotion é comercializado também em países da América Latina e nos USA, customizado com suas geografias e fonéticas próprias.

Enriquecimento de dados

O recurso de **DataEnrichment** do DataMotion, permite ao usuário verificar a existência de seus registros nas bases de dados das empresas como CheckExpress, Dun&Bradstreet e MSI para aquisição de registros para atualização de Endereços e Telefones para pessoas físicas e Endereços, Telefones, Qtde de Funcionários, Ramo de Atividade, Executivos, Faturamento e outras informações para pessoas jurídicas.

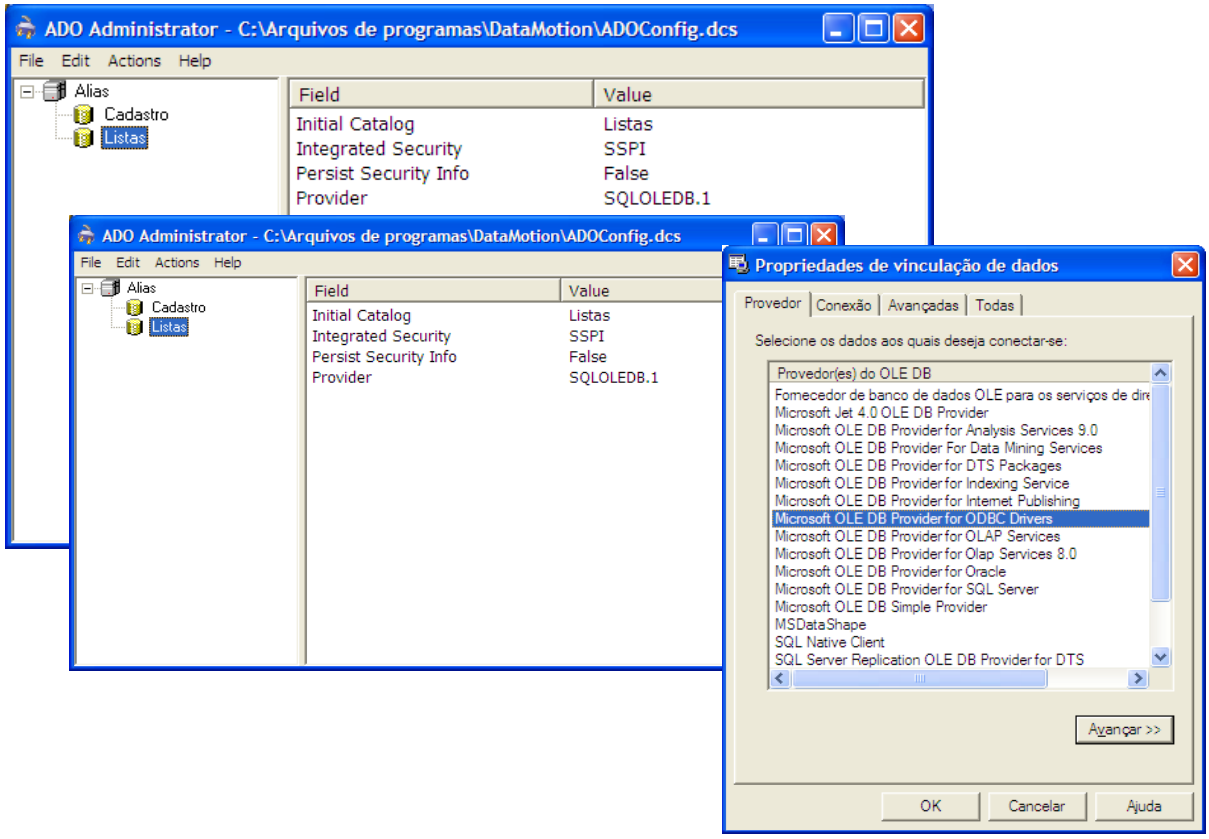
DataMotion Studio: Todas funcionalidades presentes em um único módulo.



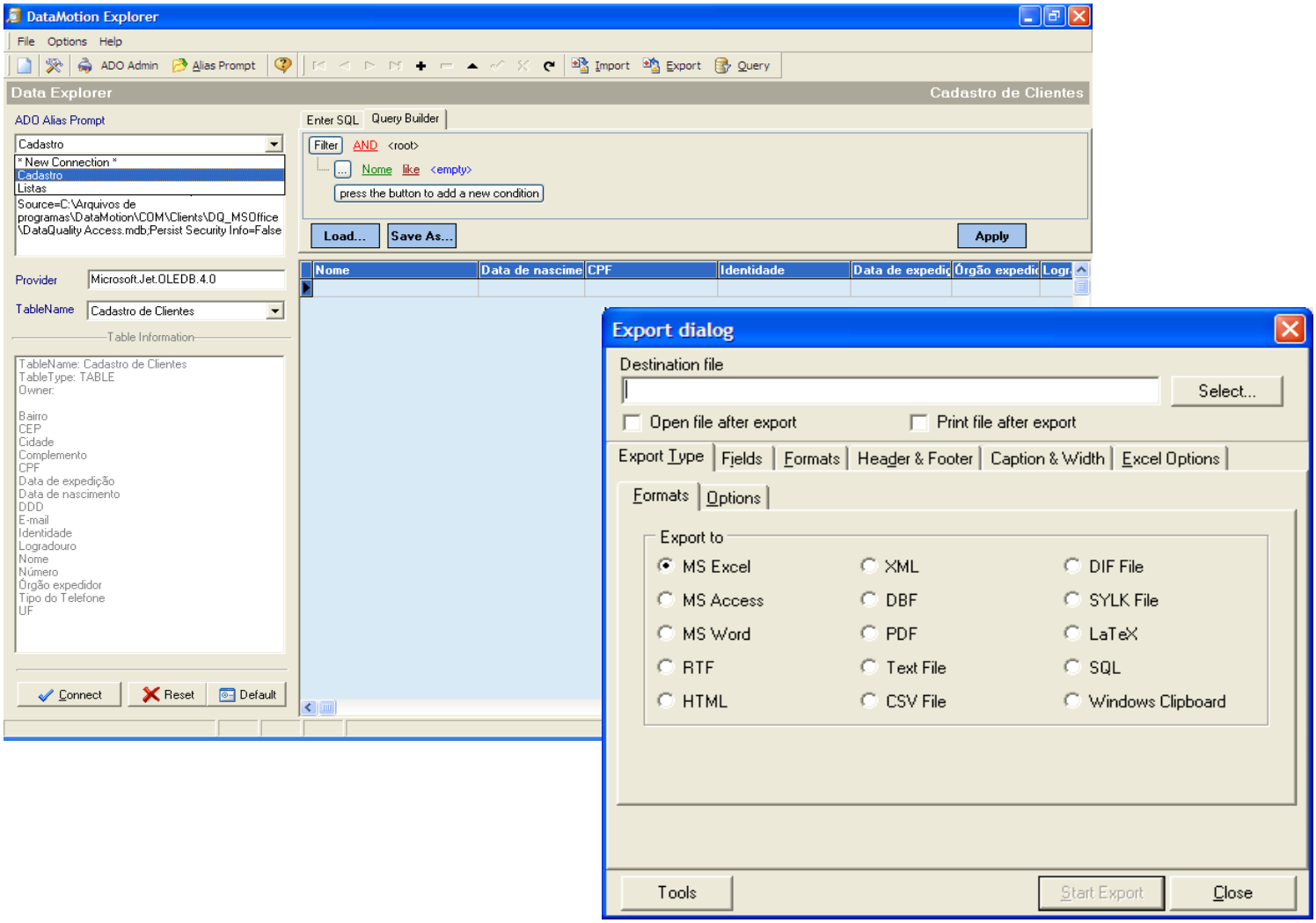


O DataMotion possui diversos módulos para facilitar e permitir que cada etapa do tratamento de dados seja executada da melhor maneira possível.

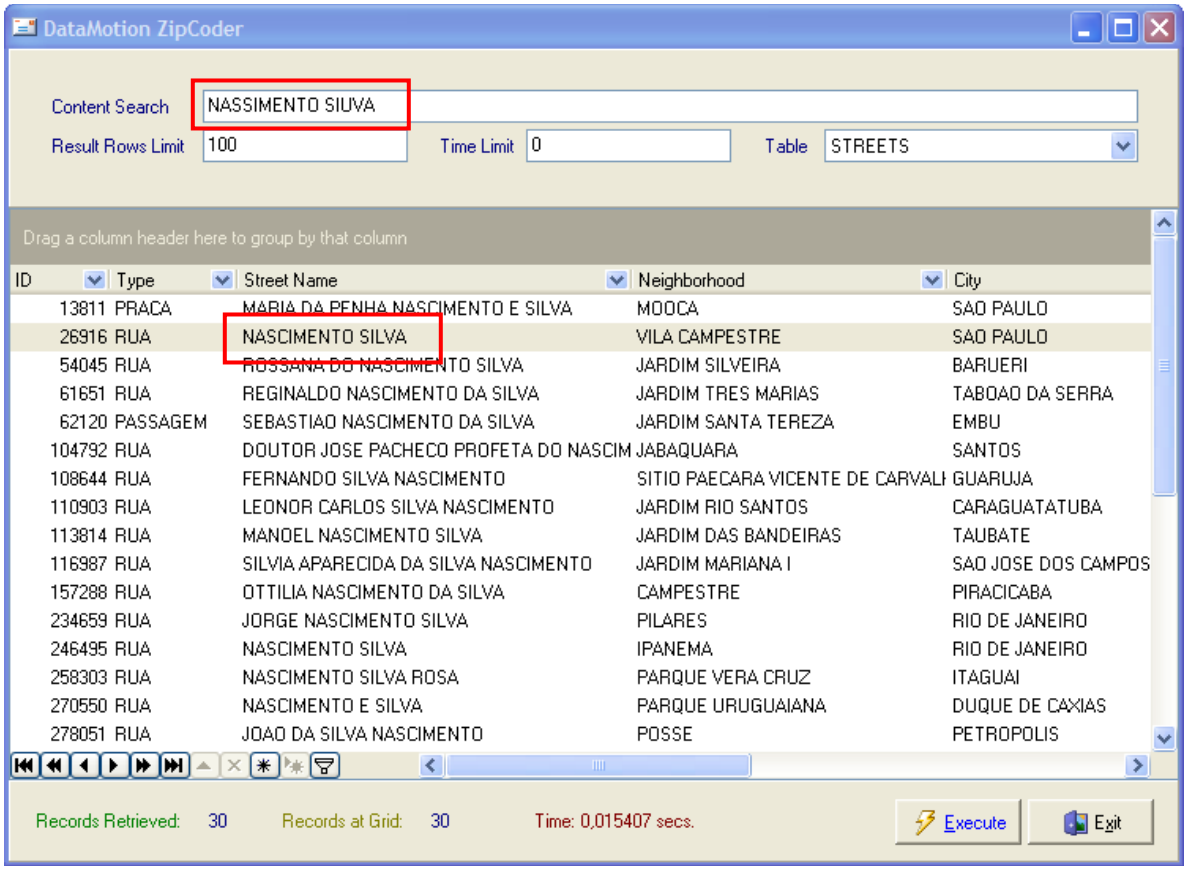
Administrador de ADO : O *ADOAdmin* permite que todas as conexões aos arquivos e banco de dados tratados sejam gerenciadas a partir de uma única tela. Depois de configuradas, as conexões podem ser utilizadas em todos os módulos do DataMotion apenas utilizando-se os *Alias* criados.



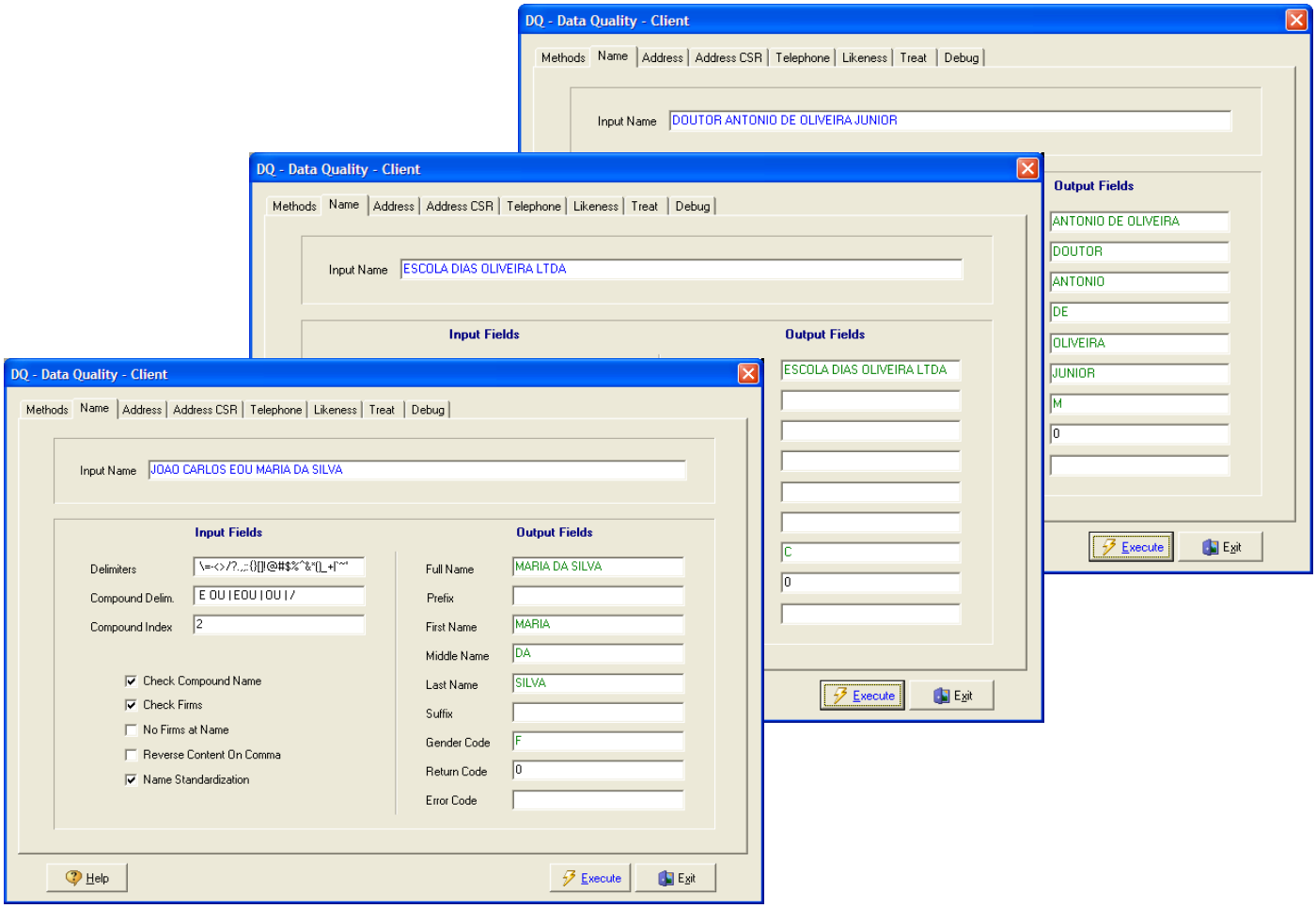
Explorador de arquivos : Utilizando as conexões (*Alias*) criadas no Administrador de ADO, os usuários podem acessar as tabelas ou arquivos para a realização de consultas, queries, exportações, importações de mais dados e manipulações diversas.



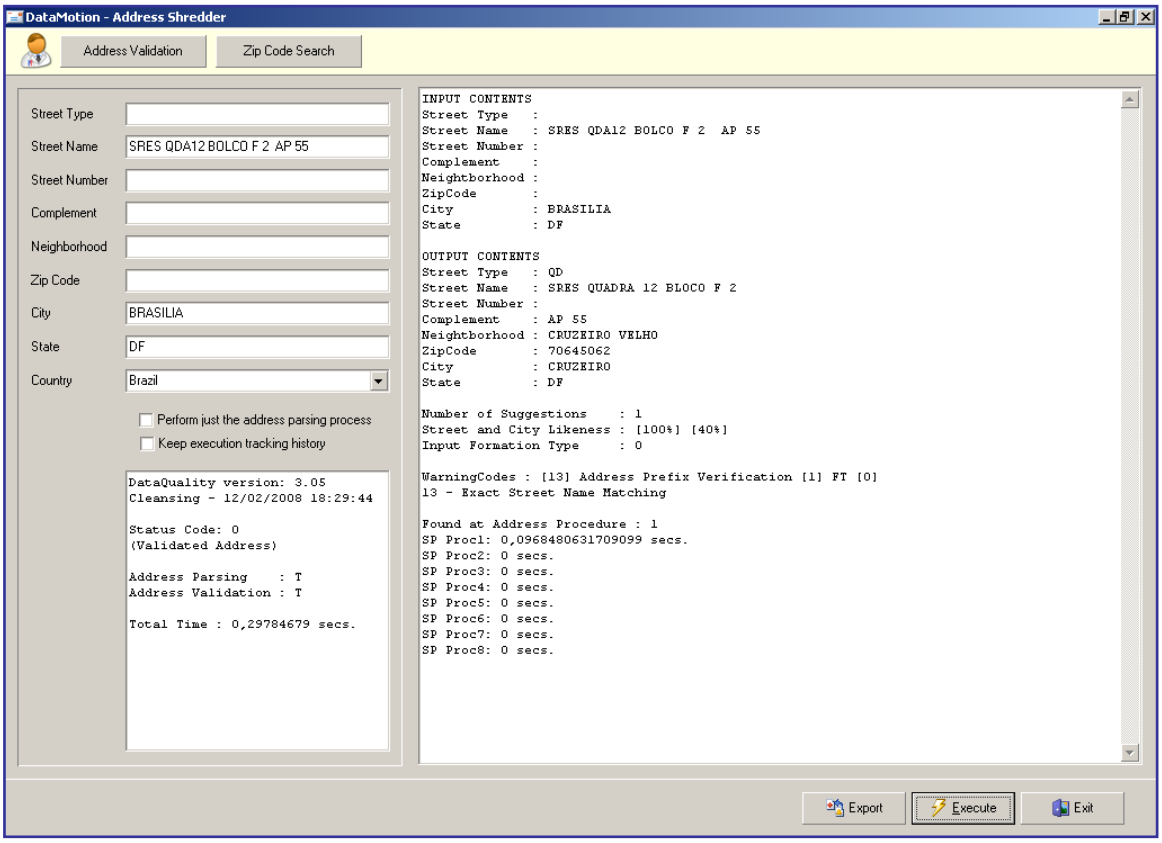
Content Search and Retrieval : Poderoso mecanismo de busca de conteúdos em arquivos ou tabelas. Todo dado é fonetizado e padronizado de modo que o componente retorne o maior número de opções possíveis. No exemplo abaixo, o componente é utilizado em uma tabela de endereços, mas pode ser utilizado em qualquer tipo de conteúdo como observações, produtos, contratos, textos diversos, etc ...



Tratamento de Nomes : Identificação do tipo de pessoa (PF ou PJ), atribuição do sexo, identificação e separação de vários nomes dentro de um mesmo campo.



Tratamento de endereços : Padronização e correção de todos os componentes que compõem um endereço.

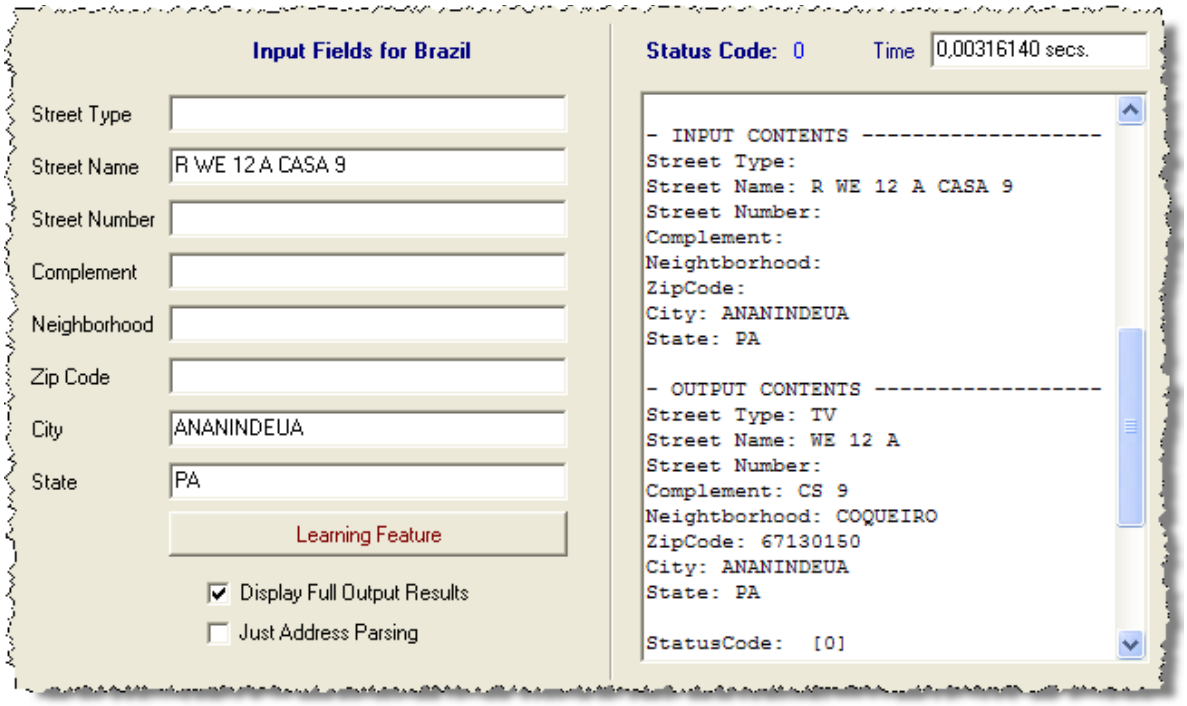


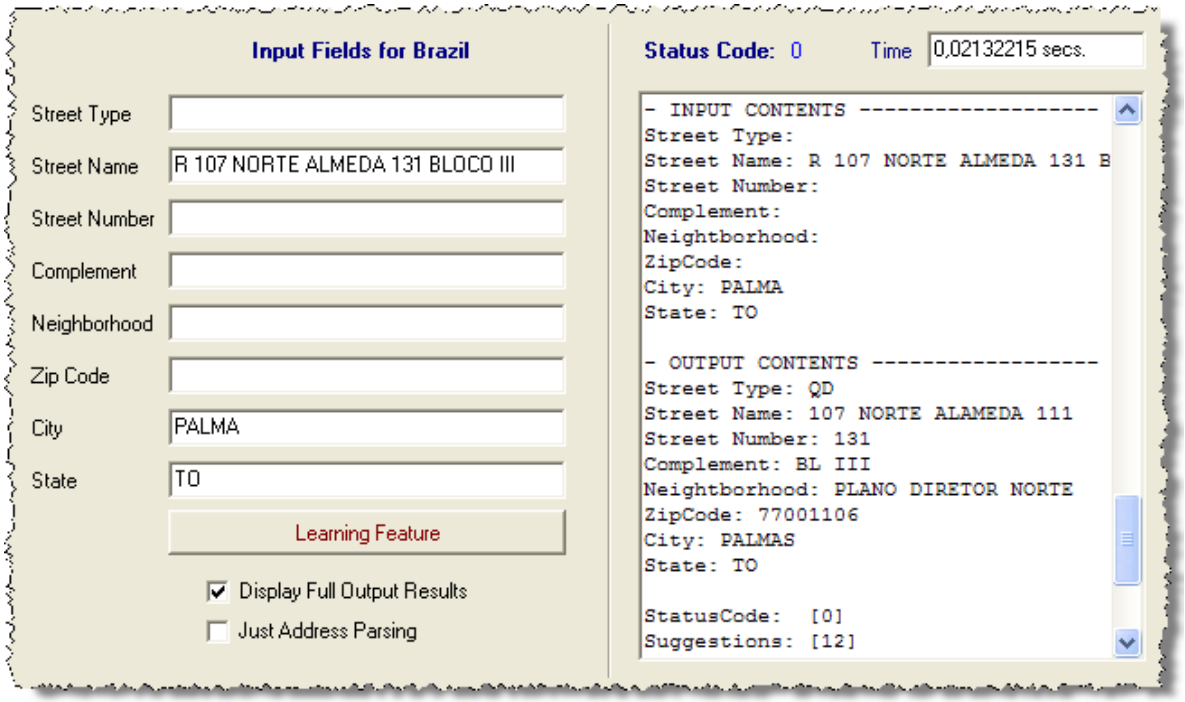
O resultado do tratamento pode ser em campos específicos de Tipo de Logradouro, Logradouro, Número e Complemento ou todos juntos ou parcialmente juntos.

Além de endereços o DataMotion também trata Telefones, Documentos (CPF, CNPJ e Inscrição Estadual), Cargos, Emails, etc ...

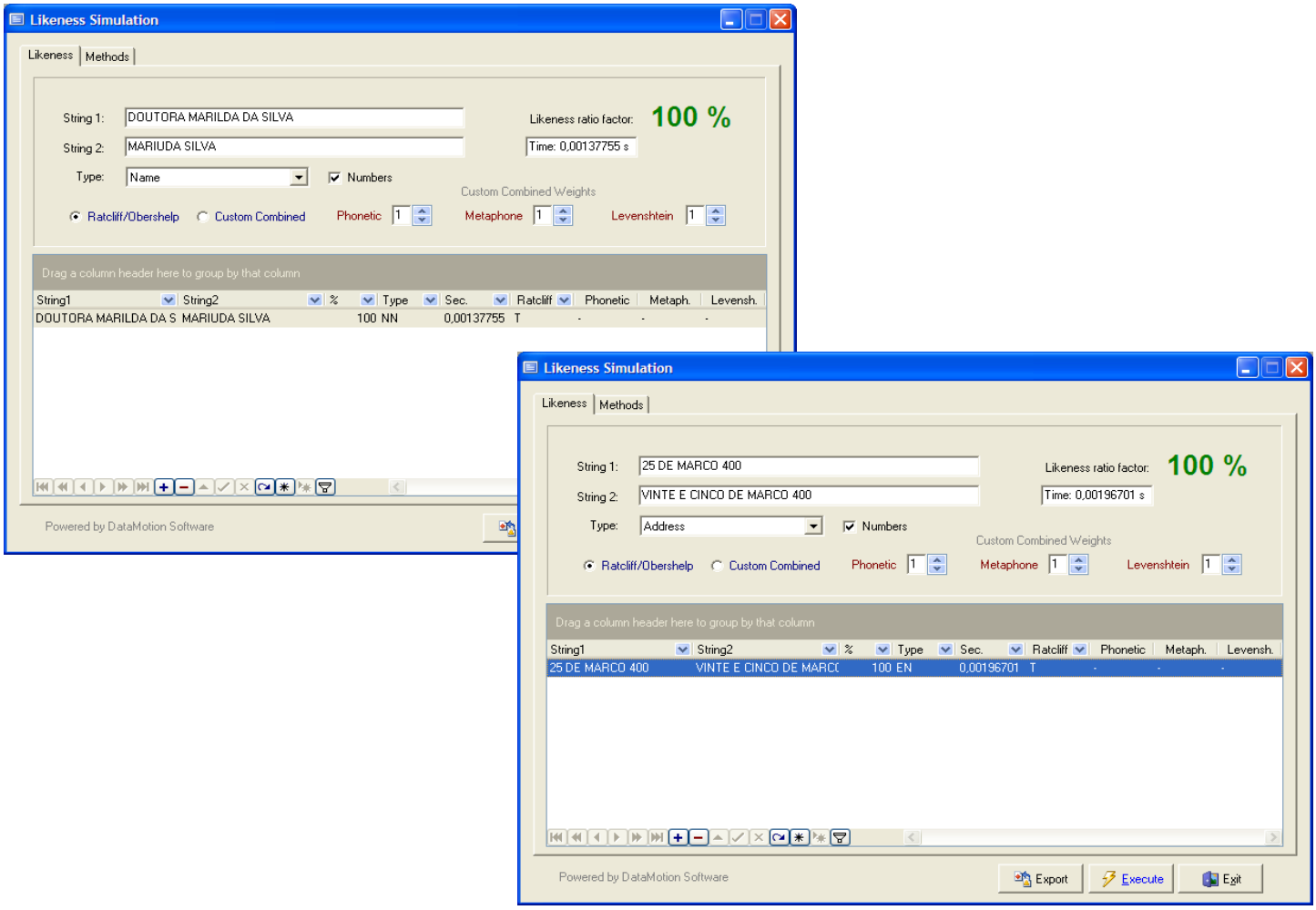
Mais de 20 códigos de retorno para o usuário saber exatamente o que aconteceu com o dado, como o dado estava originalmente, como ficou e o que foi atualizado.

Abaixo alguns exemplos de tratamento de endereço:





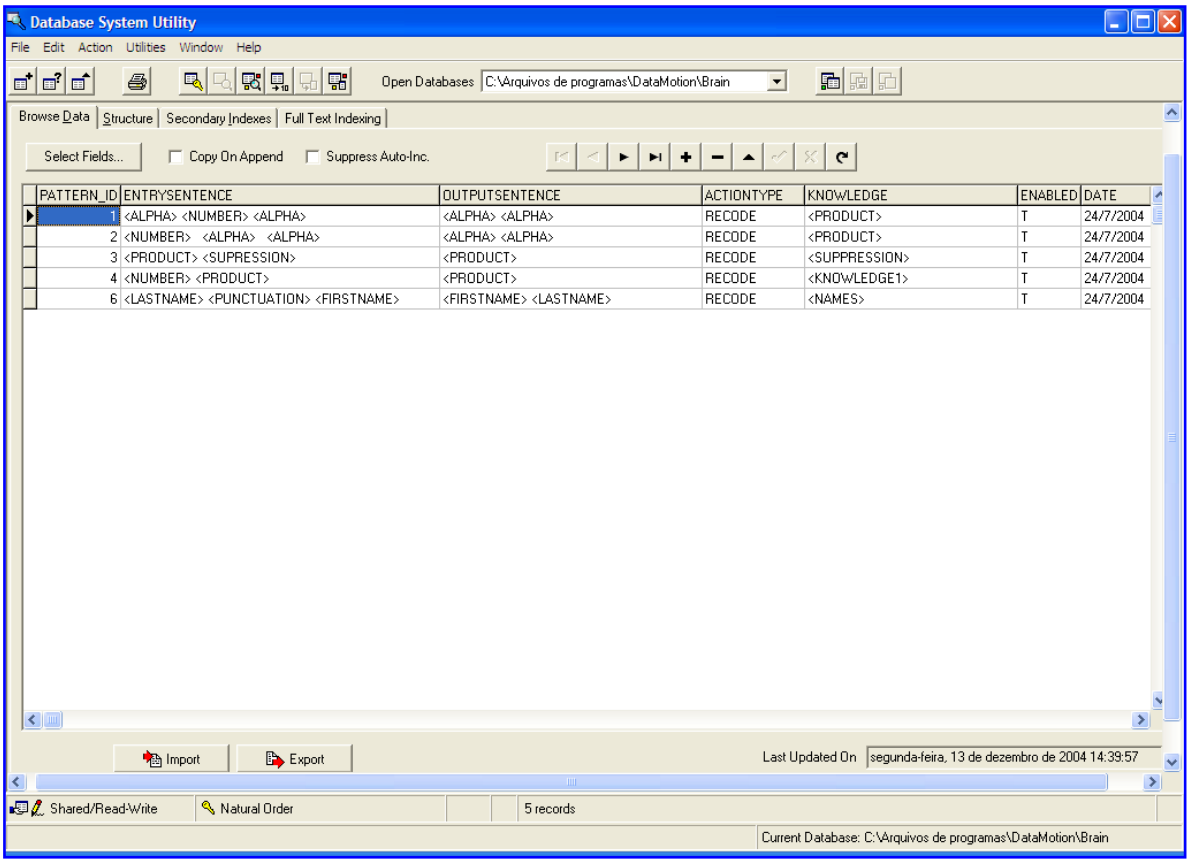
Likeness : Para identificação da similaridade de possíveis duplicidades o DataMotion utiliza várias customizações de algoritmos clássicos e consagrados, tais como Ratcliff Oberschelp Pattern Matching, Daitch Mokotoff Soundex, Metaphone e Levenshtein Distance. Estes algoritmos analisam variações ortográficas e fonéticas atribuindo um percentual de semelhança entre as comparações levando em consideração padrões específicos para strings do tipo Nome, Endereço, Razão Social e campos gerais. As comparações podem ser feitas entre os conteúdos dos campos ou entre MatchKeys criadas a partir dos conteúdos mais significativos dos campos.



Exemplos de nomes após a criação de chaves fonéticas (note que os nomes são diferentes mas as chaves fonetizadas são iguais).

NOME	MATCHKEY	NOME	MATCHKEY
GRASIELE	NYbOqYY}	AIRTOM	reC_IN
GRASIELLE	NYbOqYY}	AIRTON	reC_IN
GRAZIELE	NYbOqYY}	AIRTTON	reC_IN
GRAZIELI	NYbOqYY}	AYRTOM	reC_IN
GRAZIELLE	NYbOqYY}	AYRTON	reC_IN
GRAZIELY	NYbOqYY}	HAIRTON	reC_IN
		HAYRTON	reC_IN

Brain : Com este módulo o usuário pode ensinar o DataMotion a realizar novos tipos de padrões e correções no conteúdo e no formato dos dados. Este componente pode ser utilizado para identificação e correção de formas de escrita em campos genéricos, como por exemplo, tabela de produtos. Se o conteúdo de algum campo não seguir uma ordem correta de digitação ou tiver alguma palavra, sigla ou número inválidos o DataMotion pode indicar que a forma de preenchimento está incorreta como também pode corrigi-la se tiver o padrão correto.



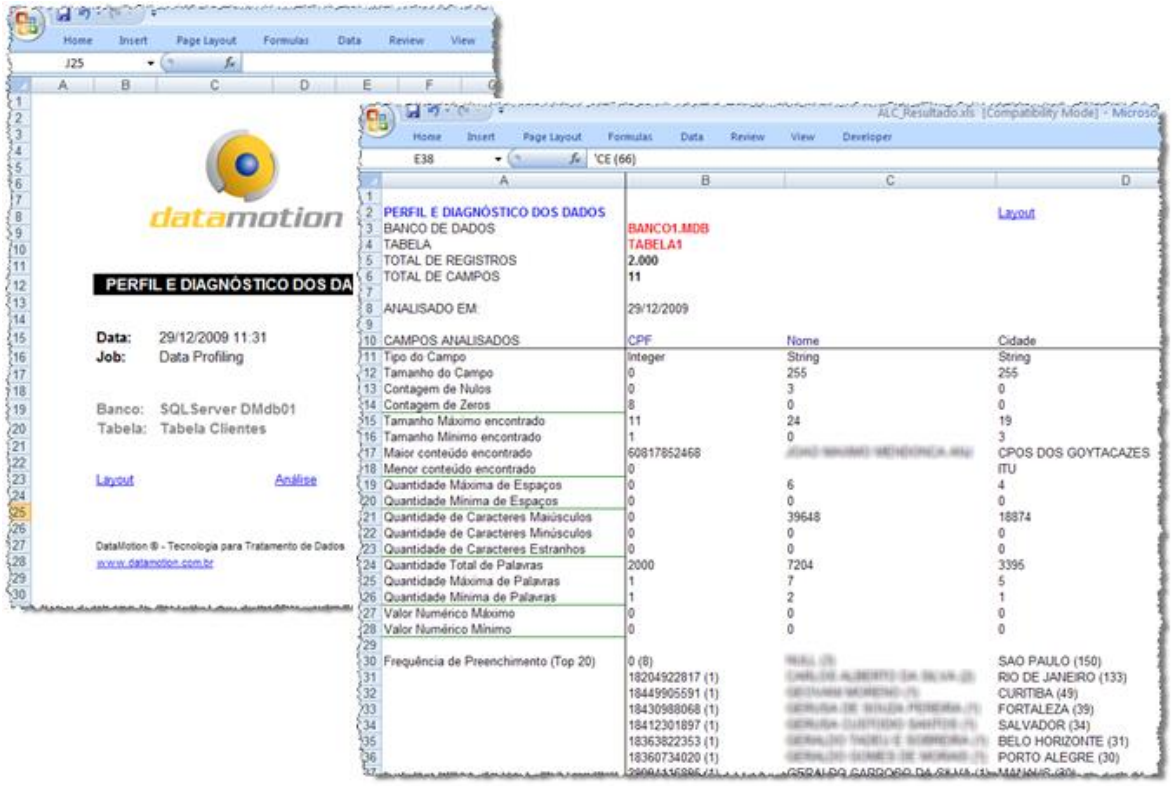
Data Profiling : Segundo estimativas da DMReview, de 20 a 50% das empresas no mundo têm problemas de qualidade em seus dados. Através de avaliação e do monitoramento preventivo é possível fazer com que não conformidades que estejam acontecendo não ocorram mais, bem como, é possível também prever futuras inconsistências cadastrais ou transacionais.

Todas as etapas relacionadas a análise do conteúdo e estrutura dos dados são definidas pelos processos de Data Profiling. Através do DataMotion, diversas e poderosas funcionalidades de diagnóstico e análise de conteúdo cadastral são disponibilizadas ao usuário.

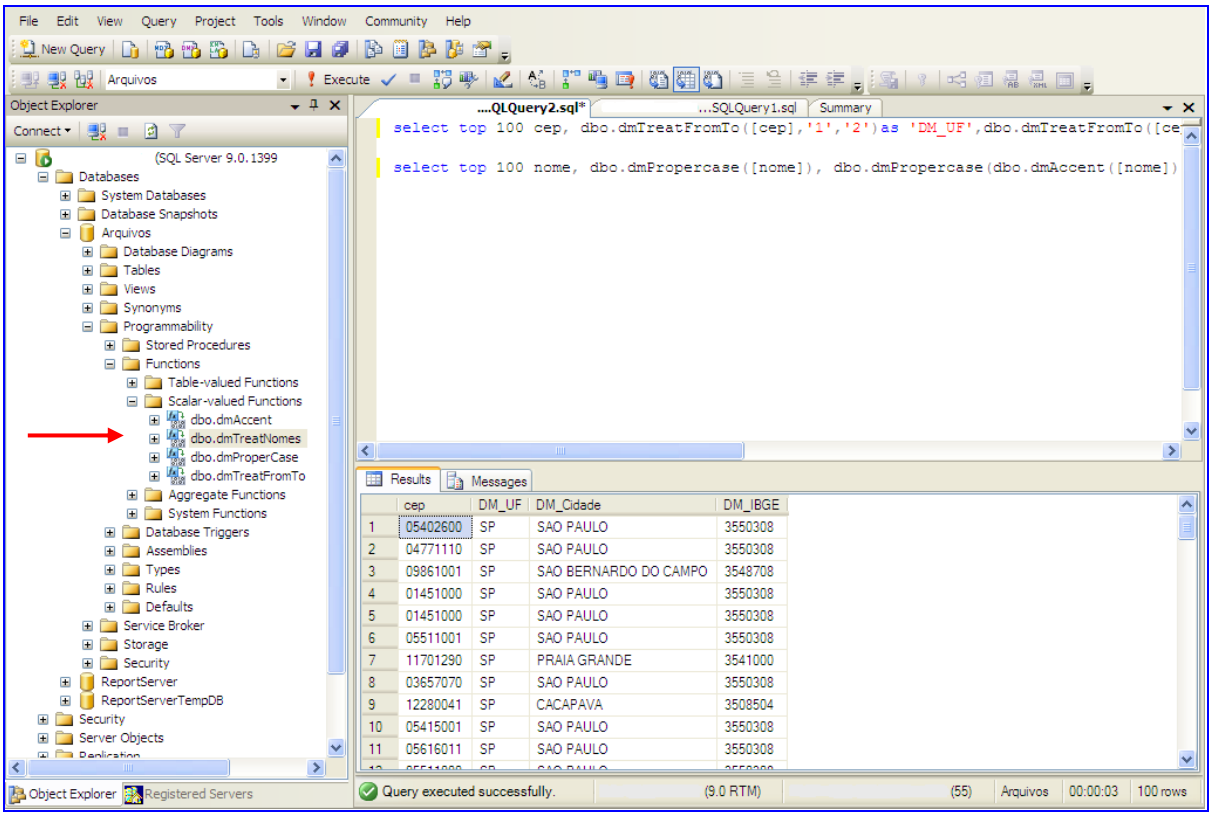
O Data Profiling DataMotion possui as seguintes funcionalidades integradas:

- Análise de praticamente qualquer tipo de arquivo;
- Suporte a arquivos de virtualmente qualquer tamanho;
- Mapeamento do layout dos campos;
- Análise Estrutural da base de dados;
- Análise de Missings;
- Tabulação de conteúdo e de padrão de preenchimento dos campos;
- Identificação de possíveis anormalidades cadastrais;
- Relatórios gerenciais em planilhas Excel;
- Etc.

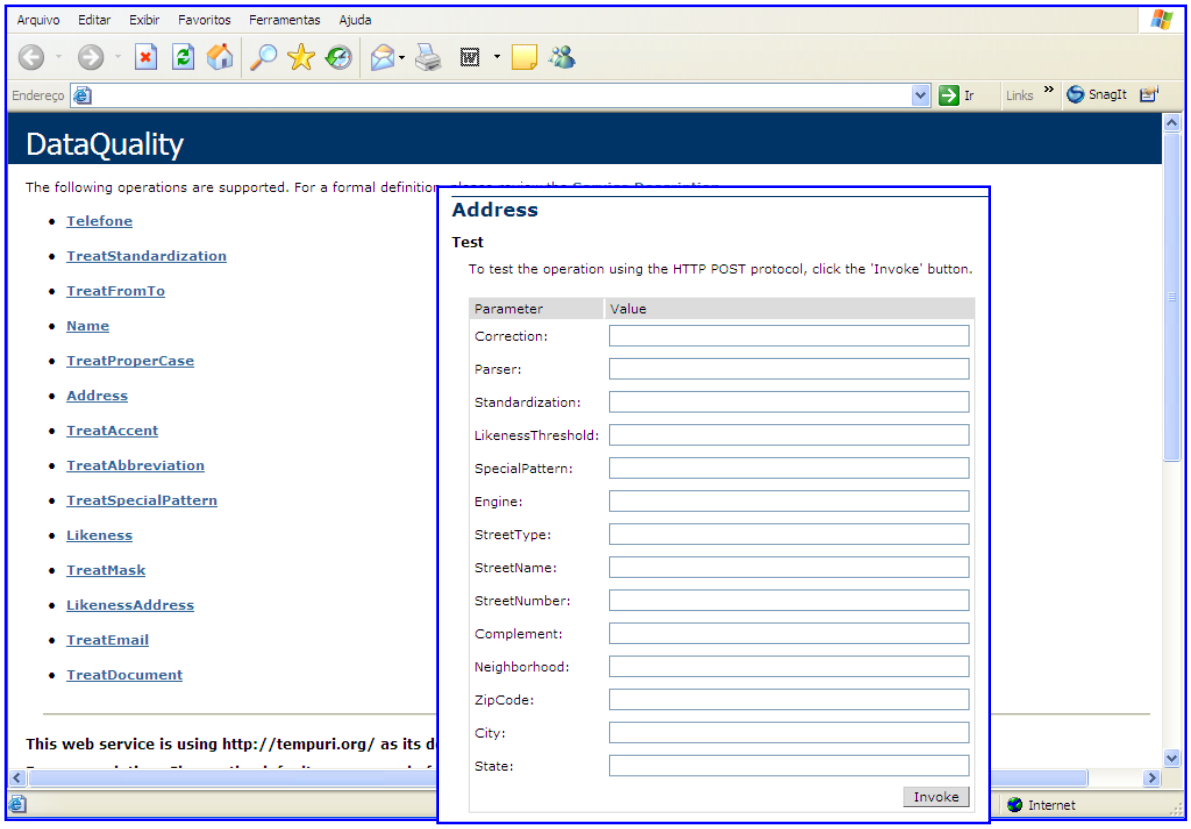
Segue abaixo telas produzidas pelo módulo:



Integração no SQL Server : Integração da API do DataMotion no SQL Server para utilização dos métodos através de Functions ou Stored Procedures.



Web Service : Disponibilização de todos os componentes do DataMotion para serem utilizados em XML JSON Web Services, permitindo acesso multiplataforma aos recursos de tratamento de dados.



Arquitetura e Glossário

METHODS – Conjunto de funções (Primitivas) que formam o Kernel (SDK) do DataMotion. As primitivas são basicamente os **Métodos**. Um conjunto de Métodos com a mesma afinidade e característica é denominado uma **Classe** (Class).

TASKS – Uma Tarefa (Task) é um processo de transformação que pode reunir um ou mais métodos. As Tasks são definidas pelo usuário utilizando-se o DataMotion Studio. Através do DataMotion Studio, o usuário terá acesso ao editor de TaskCode.

TASKCODE – Código que descreve quais e como as Tarefas serão executadas, compondo assim um Package.

BRAIN – (Gerenciador do Metadados) Armazena todos os conhecimento do DataMotion relativo aos dicionários de dados, translate tables e métricas de negócio. O Brain poderá ter vários 'Knowledges' e cada 'Knowledge' poderá ter várias regras. O conteúdo - ou knowledge - do Brain poderá ser editado, importado ou exportado, podendo-se assim, transferir conhecimentos em diferentes versões do DataMotion ou diferentes versões de banco de dados.

DMI - DATAMOTION INTERPRETER : Recurso utilizado para execução do DataMotion em modo batch. Através do DMI o usuário pode criar um script de comandos para execuções em lotes.

Customização

Conforme mencionado, o uso do Data Quality & ETL nas organizações é feito de uma maneira bastante genérica. Sua aplicação é vista em diversos tipos de sistemas e operações, fazendo com que seja necessário que o DataMotion – enquanto solução - possua uma arquitetura aberta para atender as principais demandas do usuário. Dado esse escopo, o DataMotion foi desenhado para poder ser 'customizado' através do uso de Métodos e/ou outros componentes (Plug Ins).

Por outro lado, a cada novo release, novos Métodos poderão ser implementados nativamente no DataMotion. Periodicamente, novas Classes serão disponibilizadas.

Os Plug Ins (ou AddOns) poderão ser desenvolvidos sob medida as necessidades ou regras de negócios dos usuários, e ainda assim terem completa integração com o próprio kernel do DataMotion. Vale ressaltar que esses Plug Ins serão Métodos pontuais criados a partir de uma necessidade, e que poderão ser integrados a qualquer Package ou Task DataMotion.

Portanto, o DataMotion pode ser customizado para resolver qualquer tipo de problema que relacione-se com Data Quality, Data Transformation, Recoding e Conversão de Dados.

www.datamotion.com.br

Registro INPI : 018060114531

Anexo – Metodologia de Tratamento, Padronização e Enriquecimento dos Endereços

Todos elementos que compõem o endereço são detalhadamente avaliados e auditados por rotinas de validação cadastral. O processo que antecede o tratamento e a padronização é o Diagnóstico. Nesta etapa, cada um dos elementos, assim como cada string de endereçamento é investigada quanto a forma e padrão de preenchimento. A identificação do melhor algoritmo de tratamento a ser implementado, depende das conclusões da etapa de diagnóstico.

As partículas de Tipo de Logradouro, Logradouro, Número, Complemento, Bairro, CEP, Cidade e UF, são então dispostas em forma de uma grande equação simultânea, onde diversas regras de negócio são aplicadas. Nesse momento, o algoritmo de tratamento e padronização entra em ação. Nomes oficiais de logradouros, tratamento de acentuações, recodificações de prefixos e sufixos, etc, enfim todo tipo de validação é executada de forma sincronizada durante os processos de tratamento dos logradouros.

Durante o processamento, toda consistência e enriquecimento de endereços é feita com base no Diretório Nacional de Endereços (DNE dos Correios), assim como, em base de conhecimento própria, que ao todo, compõem um gigantesco banco de dados, com milhões de referência sobre todos os logradouros do país, seus respectivos códigos de endereçamento postal, municípios de CEP único, etc. Endereços com problemas terão seus componentes inconsistentes adequadamente substituídos por conteúdos corretos, utilizando-se as premissas e padrões oficiais dos Correios.

Validações de combinações exatas e aproximadas de endereços, endereços sem números, eliminação de títulos, preposições, conectivos ou então de palavras intermediárias, abreviaturas e erros de sílabas são exaustivamente tratadas, utilizando-se sempre um gestor de similaridade fonética para enriquecimento dos elementos que compõem o endereço.

A cada registro processado são gerados diversos tipos de códigos de retorno, com informações detalhadas sobre o resultado do tratamento. Para cada método aplicado na tentativa de validação cadastral é gerado uma pontuação com a respectiva avaliação.

Níveis de sensibilidade de Match

Os níveis de sensibilidade dos processos de Match, são orientados ao diagnóstico do conteúdo a ser processado. Essa orientação permite que a aplicação possa ser parametrizada pelos mais diversos critérios, tais como: fonéticos e ortográficos.

A parametrização do Match possui todos os elementos necessários, para que seja feito a melhor validação cadastral possível. Fazem parte dos parâmetros de Match os seguintes elementos:

- Tipo de conteúdo a ser analisado
- Formato e Tamanho da Chave de Match
- Rotina Fonética
- Algoritmo de similaridade
- Grau de corte da similaridade

1) Análise dos layouts e campos

Levantamento dos diversos layouts dos arquivos com a finalidade de se obter o melhor layout para o arquivo final (cadastro único). Nesta etapa os campos são analisados levando em consideração o tipo, tamanho e conteúdo.

Análise de freqüência de conteúdos são utilizadas para que seja determinado o conteúdo final dos campos. Ex: Em um arquivo podemos ter o campo GÊNERO com conteúdos "M","F" e em outro arquivo "1","2".

Análises de padrões também podem ajudar a entender melhor determinados conteúdos, como no caso dos Telefones. Ex: (99) 9999-9999, (0xx99) 99999999, etc ...

Arquivo 1	Arquivo 2	Arquivo N
GENERO	**GENERO**	**GENERO**
Type Char (1)	Type Integer	...
Freqüência	Freqüência	...
F	1	
M	2	
Máscaras	Máscaras	...
X	9	
TELEFONE	**TELEFONE**	**TELEFONE**
Type Integer	Type Char (20)	...
Máscaras	Máscaras	...
99999999	(99) 9999-9999	
9999999	(99) 99999999	
	99 - 99999999	
	99 9999999	
	(9xx99) 9999-9999	
	...	
DDD		**DDD**
Type Integer		...
Máscaras		...
999		
99		

2) Padronização e Parser

Aplicação de tabelas de máscaras e padrões de escrita para correção e separação de determinados tipos de campos

Telefone	DDD	Número Telefone
(11) 3842-2616	11	38422616
(99) 9999-9999		
(0xx11) 38422616	11	38422616
(9xx99) 99999999		

Razão Social e Nome	
Empresa XPTO S A	Empresa XPTO S.A.
Empresa ABC Ltd	Empresa ABC Ltda
Antonio Fco da Silva	Antonio Francisco da Silva
Ma Aparecida Oliveira	Maria Aparecida Oliveira

3) Correção de endereços e outros campos

Através do cruzamento do campo endereço com o DNE (Diretório Nacional de Endereços) podemos validar/corrigir os logradouros. Além de endereços podemos validar DDDs, prefixos telefônicos, domínios de emails, dígitos de controle de documentos (CPF, CNPJ e IE), etc ...

4) Identificação dos registros duplicados

Com o dado padronizado e validado/corrigido o próximo passo é a definição das chaves de match e a execução do processo de escolha dos registros sobreviventes. A escolha do registro sobrevivente pode ser pela recência, utilizando-se uma data de cadastramento (ou alteração), por quantidade de campos com melhor preenchimento ou qualquer outro critério necessário. A escolha das melhores chaves de match dependem de análise previa dos conteúdos dos campos.

Código	1	2	3	4
Nome	João da Silva Nascimento	João Siuva Nascimento	João S Nassimento	J. S. Nascimento
Data Nascimento	20/12/1966		20/12/1966	
CPF		0.99.111.222.-33	0.99.111.222-33	
Telefone	(11) 3842-2616			(11) 3842-2616
Estado Civil	Casado			
Email	joao@datamotion.com.br			joao@datamotion.com.br
Data Cadastramento	01/01/2000	05/03/2002	30/10/2005	20/11/2004

No exemplo acima o registro sobrevivente, tendo como critério a recência, é o registro de código 3

5) Fusão dos registros duplicados

Preenchimento das variáveis que o registro sobrevivente não possui, mas os registros duplicados possuem. No caso abaixo, o registro de código 3 passa a conter o conteúdo gerado pela fusão dos outros registros.

Código	1	2	3	4
Nome	João da Silva Nascimento	João Siuva Nascimento	**João da Silva Nascimento**	J. S. Nascimento
Data Nascimento	20/12/1966		**20/12/1966**	
CPF		**0.99.111.222.-33**	**0.99.111.222-33**	
Telefone	(11) 3842-2616		**(11) 3842-2616**	**(11) 3842-2616**
Estado Civil	Casado		Casado	
Email	joao@datamotion.com.br		joao@datamotion.com.br	joao@datamotion.com.br
Data Cadastramento	01/01/2000	05/03/2002	30/10/2005	20/11/2004