Data Quality Services (DQS)

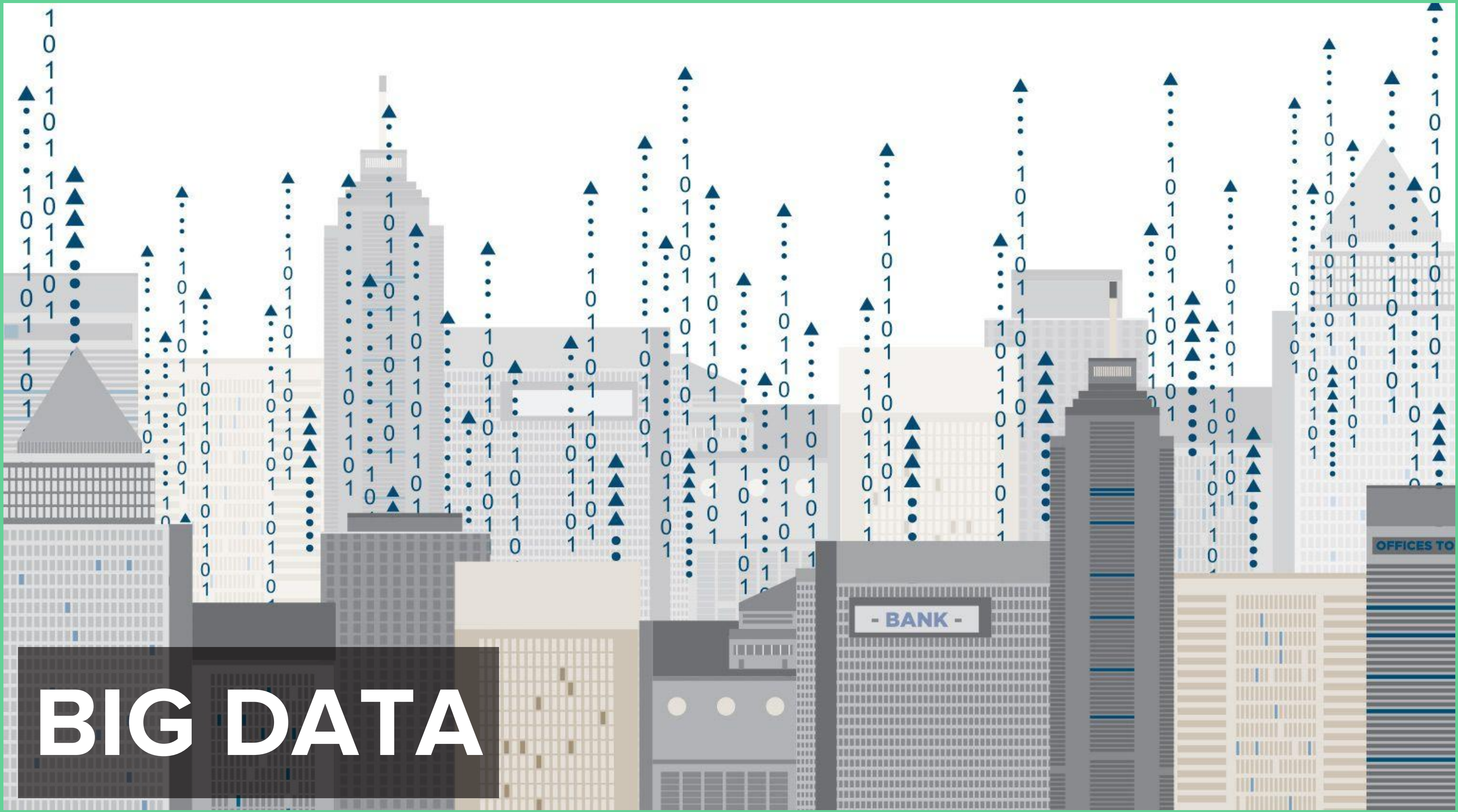
Integração de Dados e Warehousing

Pedro Henrique de Queiroz Lima

Roteiro

- Motivação
- Qualidade de dados
    - O que é?
    - Aspectos
- Data Quality Services
    - Workflow
    - Limpeza de dados
    - Correspondência de dados
    - Profiling
    - Admnistração
- Recapitulando...
- Referências

Motivação



BIG DATA

4.4ZB

1 Zetabyte = 10^{21} bytes

Quantidade de informação digital gerada acumulada até hoje (IDC/Forbes)

44ZB

1.7MB/s/pessoa

Projeção da quantidade de informação digital gerada acumulada até 2020

BUSINESS INTELLIGENCE

US$ 3,1 tri/ano

EUA, 2015

Eram US$600 bi/ano, com dados de 2003 (IDC)

Causados por decisões de negócio baseadas em dados errados, retrabalho em TI e postagem de mala direta (US$ 600 bi nestes dois últimos).

15%

dos executivos confiam na qualidade geral dos dados

27%

dos executivos tem certeza da qualidade dos dados

Qualidade de Dados

O que é?

Qualidade de Dados

Adequação dos dados para as atividades de operações, decisões de negócios e planejamento da instituição.

Dimensões de Qualidade de Dados

Fator	Exemplo
Completude	25% dos registros não contém “Sobrenome”
Conformidade	“Rua” e “r.”; “avenida” e “Ave.” ou “Av.”
Consistência	Sexo representado por “0/1” e “M/F”
Precisão	Casas decimais necessárias ou dado atual
Validade	Altura: -15,5m
Unicidade	Registros devem representar entidades únicas

Dificuldades em Qualidade de Dados

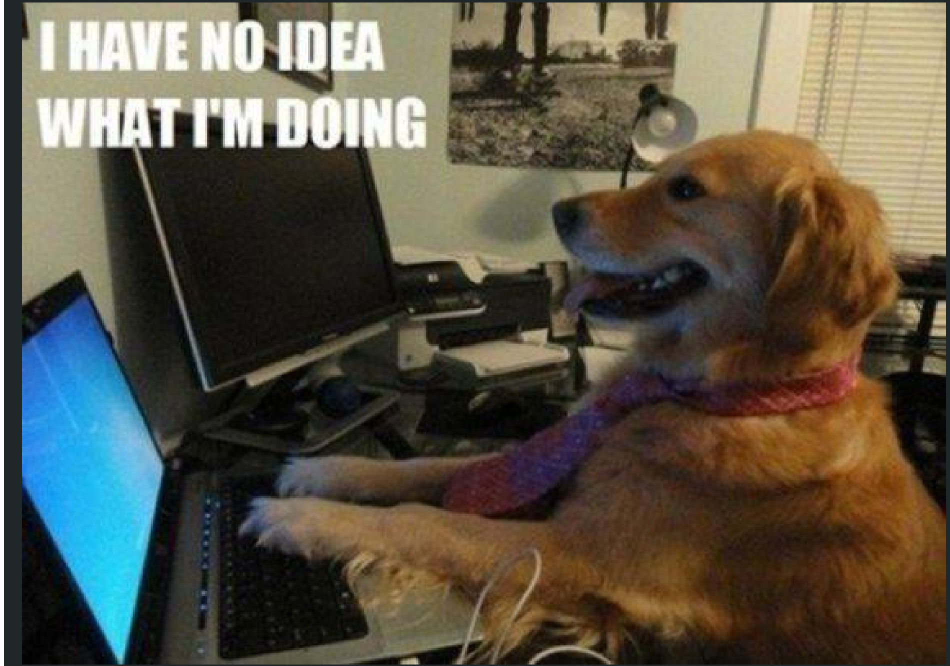
“Difusividade”

Volatilidade

Problemas no design de software

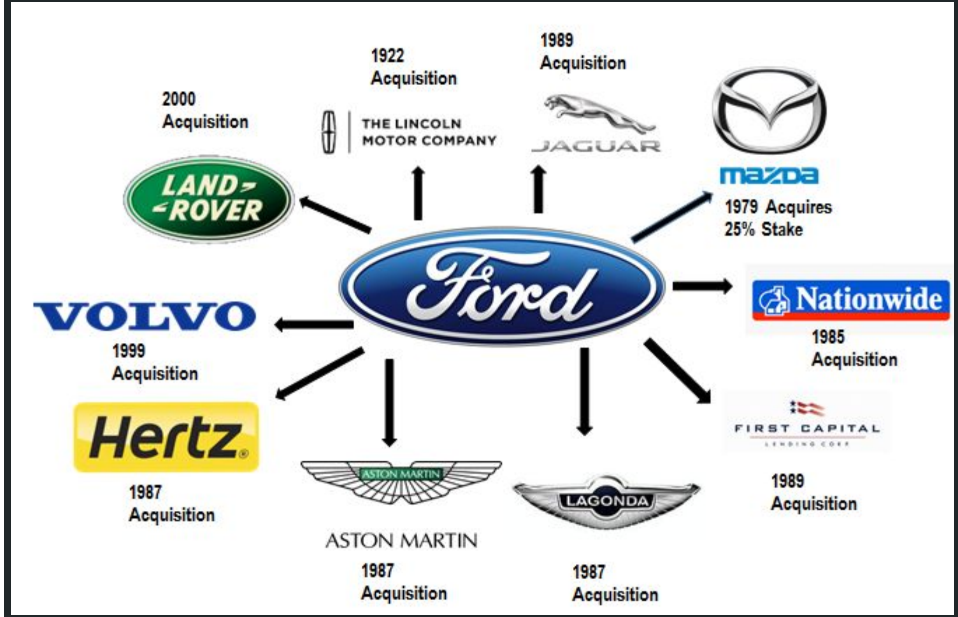
Validação de dados

Erros de entrada podem ser transferidos para a base de dados



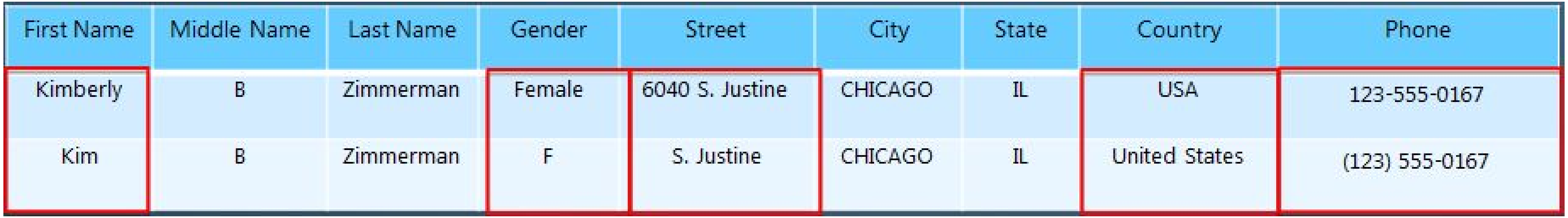
Fusão e aquisição de empresas

Possíveis duplicações, dados em formatos diferentes, incompletude de informações



Formatação

Espaçamentos, abreviações, apelidos

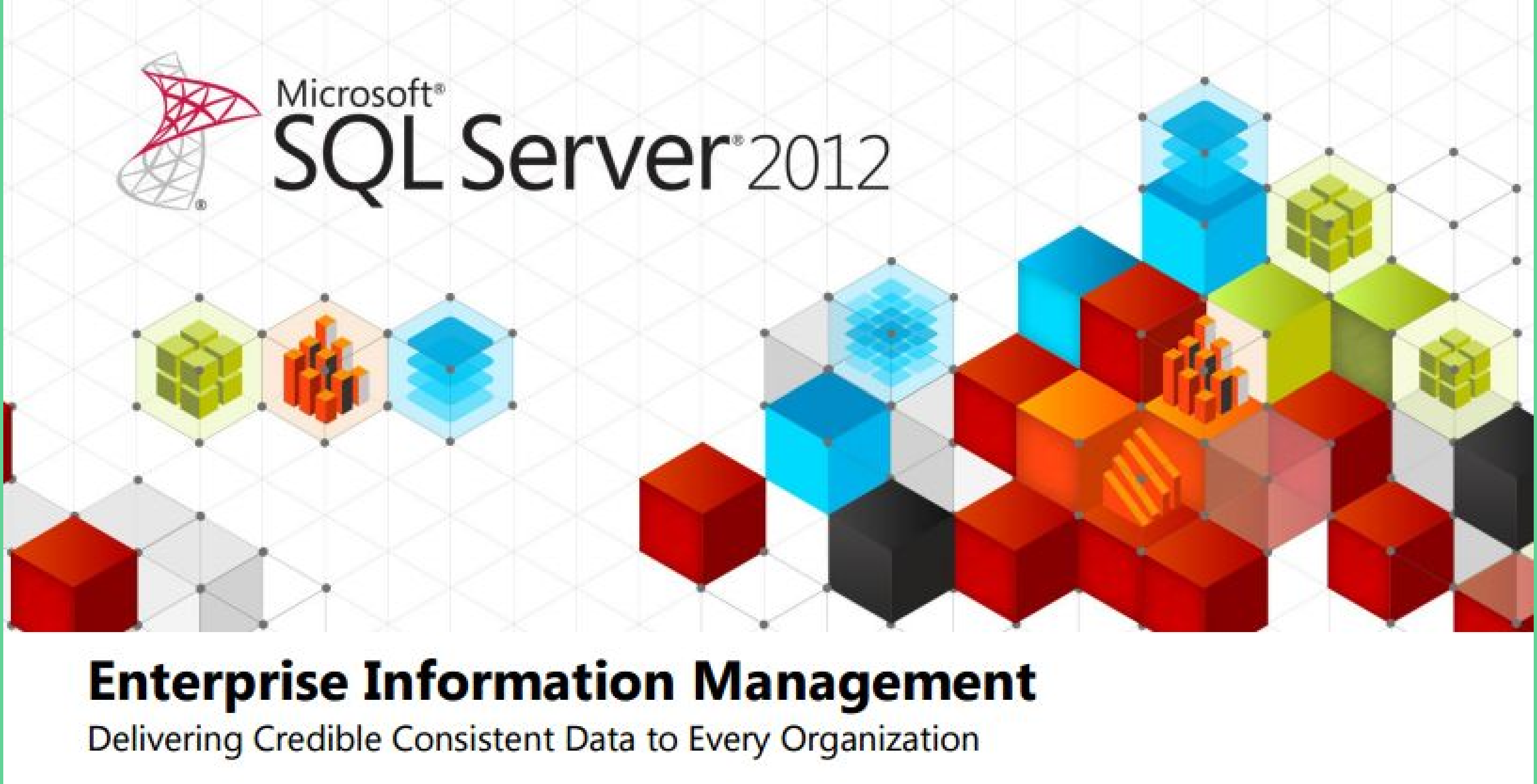


First Name	Middle Name	Last Name	Gender	Street	City	State	Country	Phone
Kimberly	B	Zimmerman	Female	6040 S. Justine	CHICAGO	IL	USA	123-555-0167
Kim	B	Zimmerman	F	S. Justine	CHICAGO	IL	United States	(123) 555-0167

Mudanças de atributos

Alterações de endereços, telefones

Data Quality Services



Microsoft® SQL Server® 2012

Enterprise Information Management

Delivering Credible Consistent Data to Every Organization

Enterprise Information Management

- Master Data Services (MDS)
- SQL Server Integration Services (SSIS)
- Data Quality Services (DQS)

Tira carga de trabalho do fluxo ETL, com interface simplificada

Concorrentes

CRMfusion

DemandTools

DupeBlocker

informatica

Data Quality

cloudingo

RingLead

IBM

InfoSphere

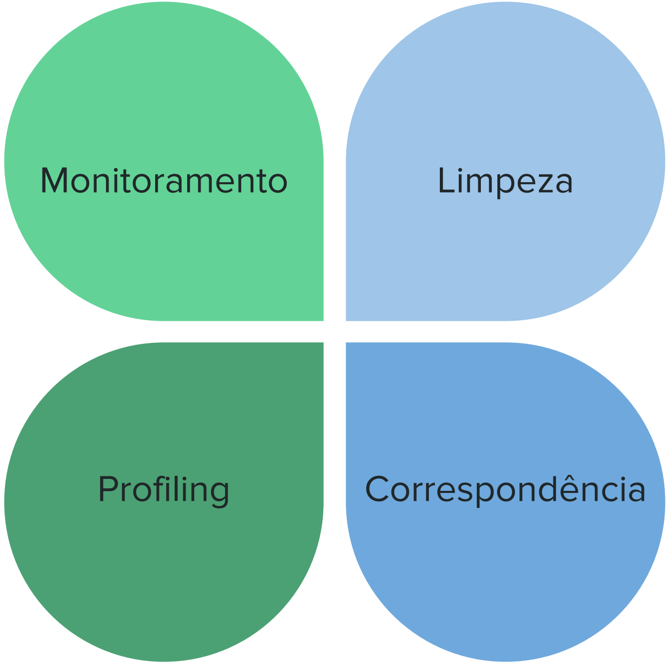
ORACLE

MELISSA DATA

Your Partner in Data Quality

Oceanos

Ferramentas



Monitoramento

Monitoramento e rastreamento dos estados das atividades de qualidade e da qualidade de dados

Limpeza

Correção, remoção ou enriquecimento de dados que estejam incorretos ou incompletos.

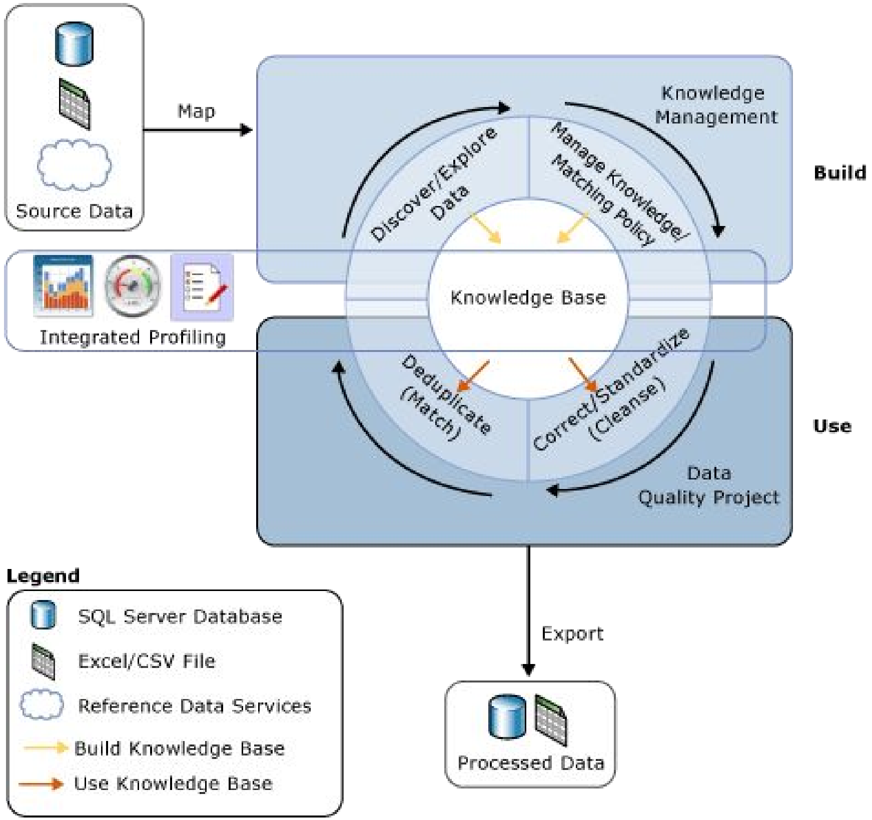
Profiling

Análise dos dados de origem para prover entendimento da qualidade de dados e seus problemas.

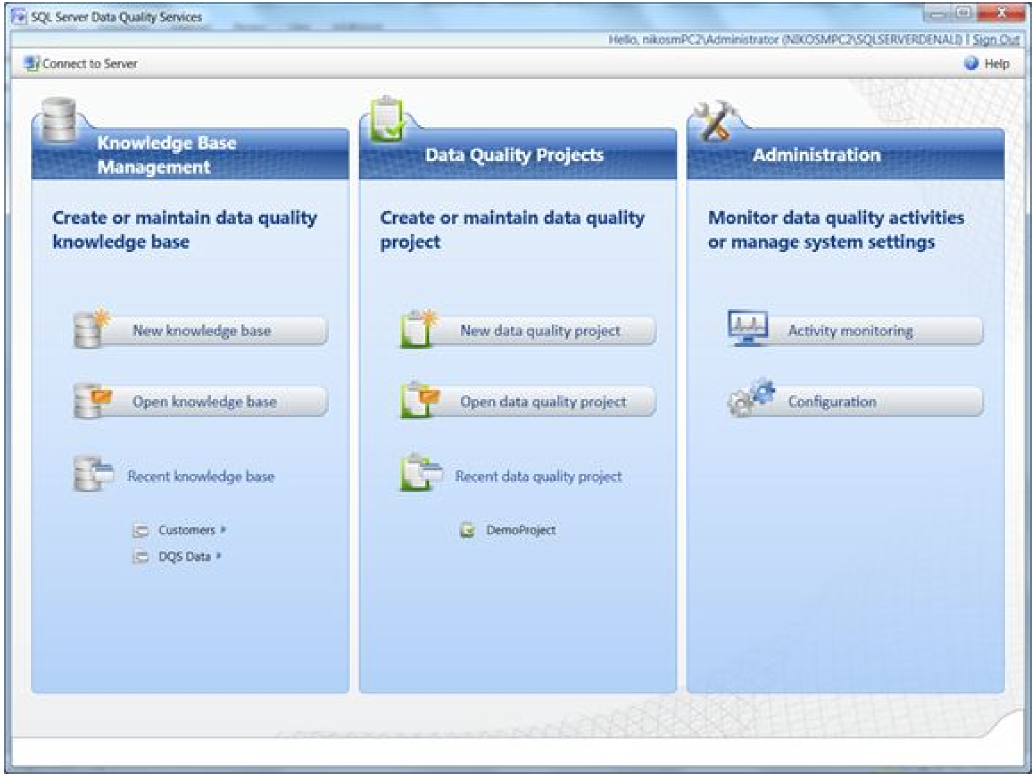
Correspondência

Identificação, ligação ou fusão de registros duplicados entre conjuntos de dados.

Workflow


Source Data
Map
Knowledge Management
Build
Discover/Explore Data
Manage Knowledge/ Matching Policy
Integrated Profiling
Knowledge Base
Deduplicate (Match)
Correct/Standardize (Cleanse)
Use
Data Quality Project
Export
Processed Data
Legend
SQL Server Database
Excel/CSV File
Reference Data Services
Build Knowledge Base
Use Knowledge Base

SQL Server Data Quality Client


SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Connect to Server
Help
Knowledge Base Management
Data Quality Projects
Administration

Create or maintain data quality knowledge base

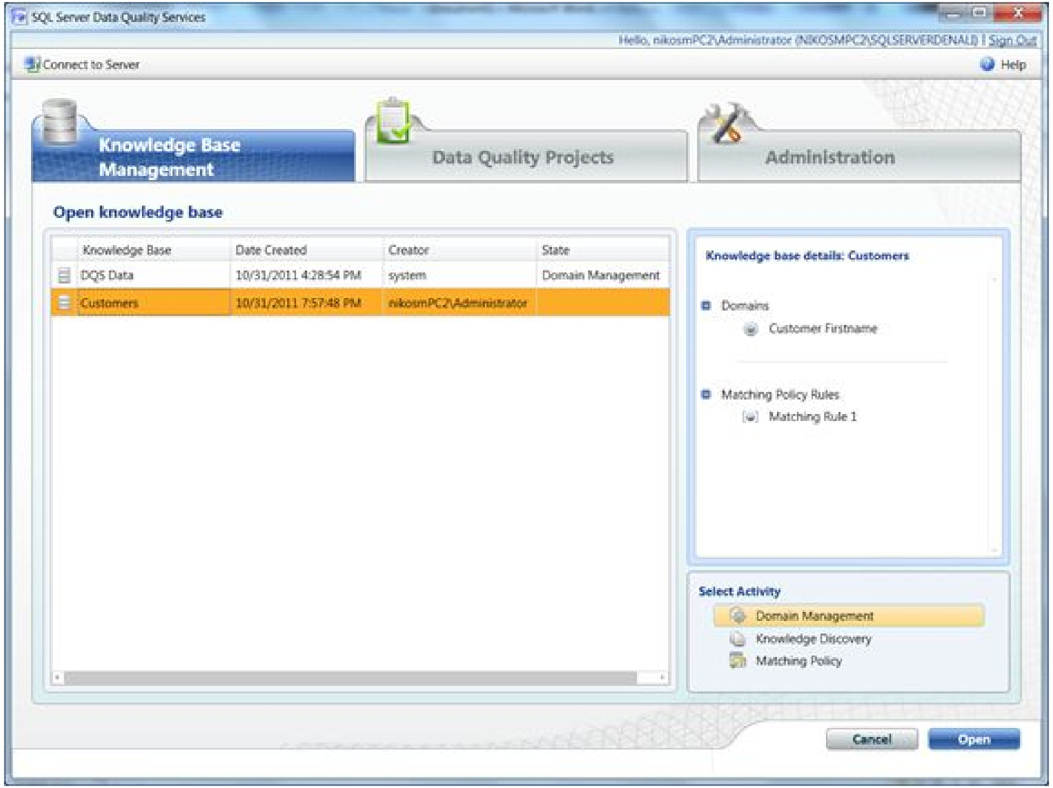
Create or maintain data quality project

Monitor data quality activities or manage system settings

Knowledge Base

- Amostras da base de dados
- Bases de dados de Referência (Azure Marketplace)
- Descoberta de conhecimento
- Domínios
    - Leading Values
    - Regras
- Base de conhecimento cresce com iterações e entradas de usuário

Knowledge Discovery


SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Connect to Server
Help
Knowledge Base Management
Data Quality Projects
Administration
Cancel
Open

Open knowledge base

Knowledge Base	Date Created	Creator	State
DQS Data	10/31/2011 4:28:54 PM	system	Domain Management
Customers	10/31/2011 7:57:48 PM	nikosmPC2\Administrator	

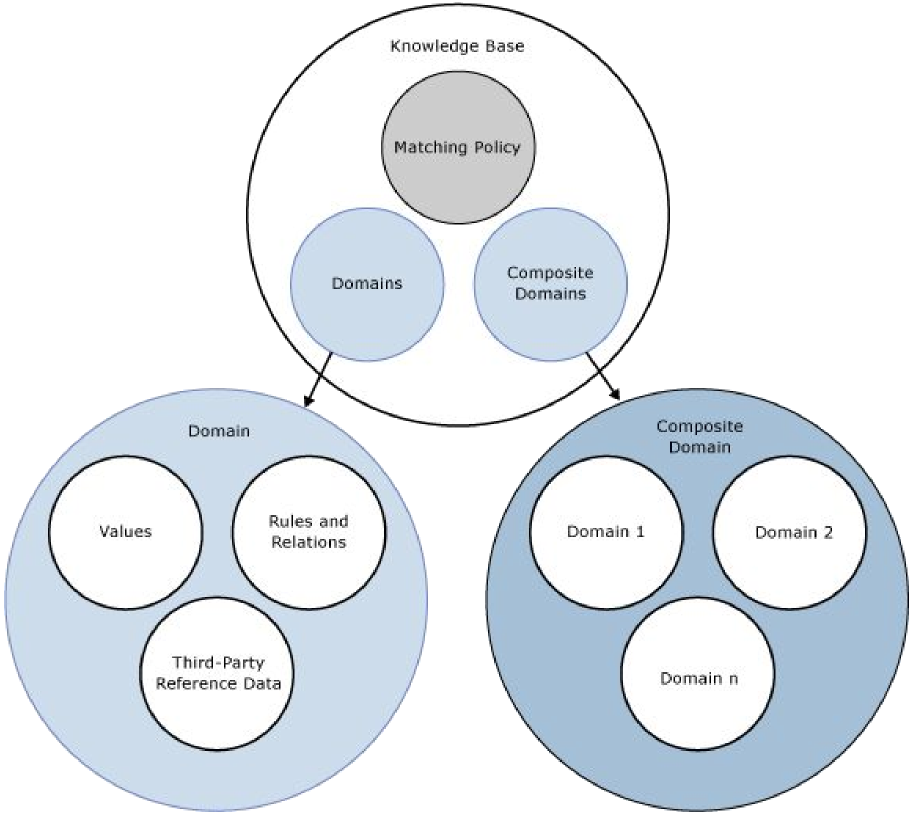
Knowledge base details: Customers

- Domains
  - Customer Firstname
- Matching Policy Rules
  - Matching Rule 1

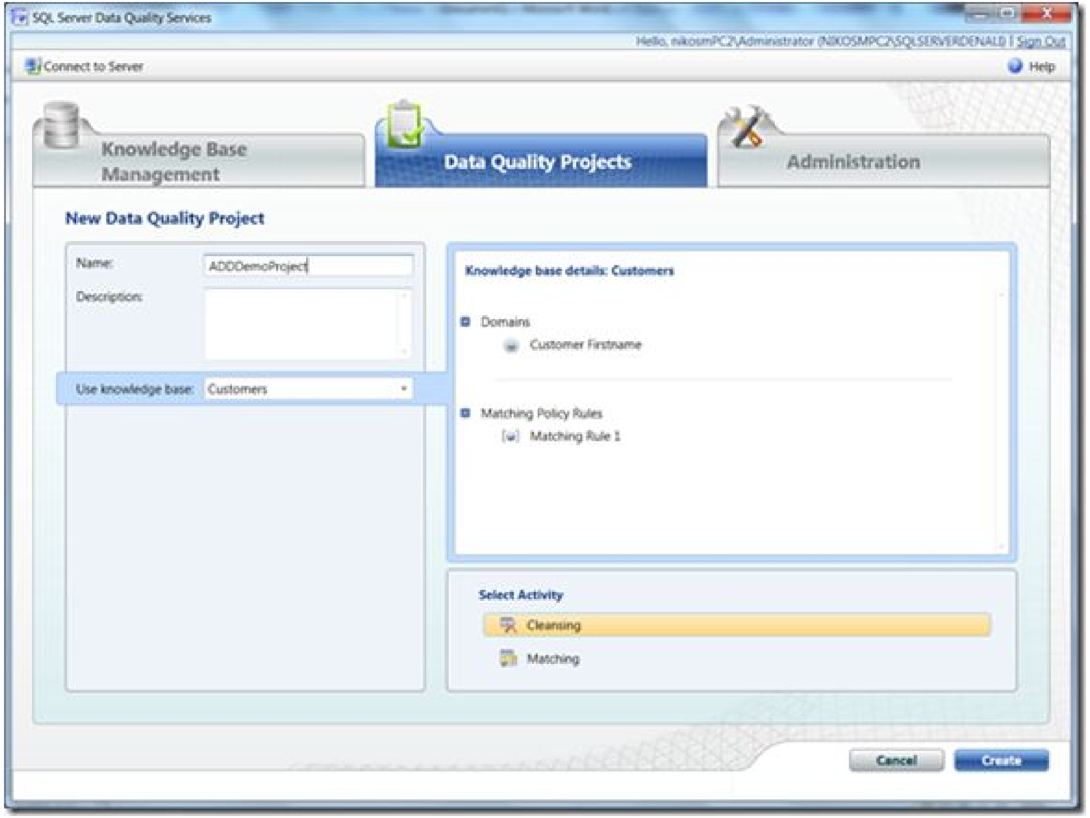
Select Activity

- Domain Management
- Knowledge Discovery
- Matching Policy

Domínios de dados



Projeto de Qualidade de Dados



SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Connect to Server
Help
Knowledge Base Management
Data Quality Projects
Administration

New Data Quality Project

Name: ADDDemoProject

Description:

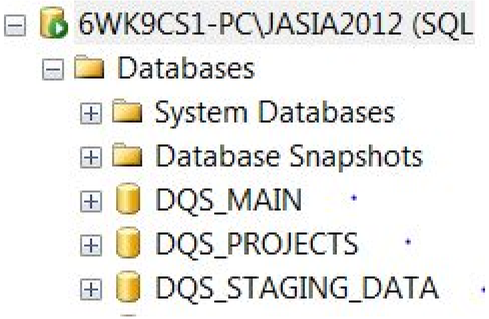
Use knowledge base: Customers

Knowledge base details: Customers

- Domains
  - Customer Firstname
- Matching Policy Rules
  - Matching Rule 1

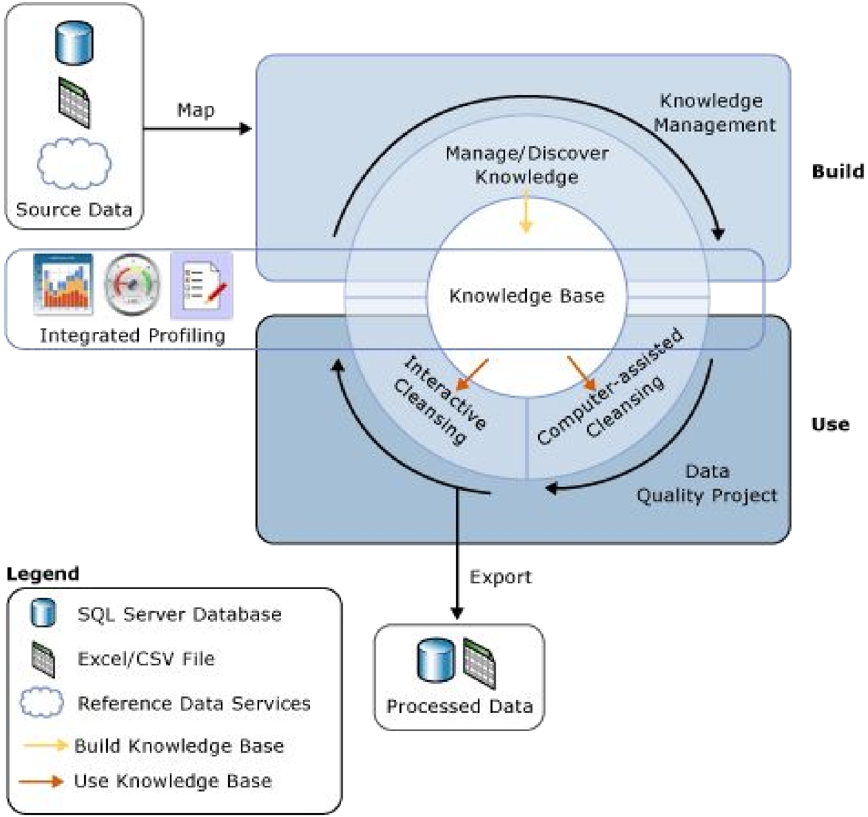
Select Activity

- Cleansing
- Matching

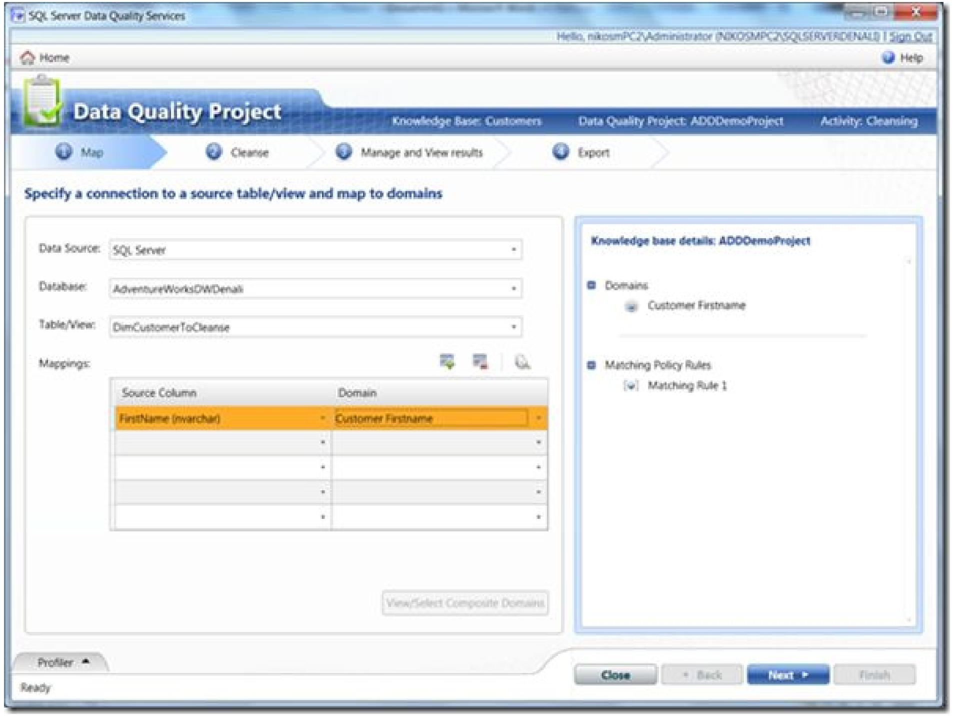


```
⊟ 6WK9CS1-PC\JASIA2012 (SQL
   ⊟ Databases
      ⊞ System Databases
      ⊞ Database Snapshots
      ⊞ DQS_MAIN
      ⊞ DQS_PROJECTS
      ⊞ DQS_STAGING_DATA
```

Limpeza de Dados



Limpeza de Dados



Knowledge Base: Customers   Data Quality Project: ADDDemoProject   Activity: Cleansing

Specify a connection to a source table/view and map to domains

Data Source: SQL Server

Database: AdventureWorksDWDenali

Table/View: DimCustomerToCleanse

Mappings:

Source Column	Domain
FirstName (nvarchar)	Customer Firstname

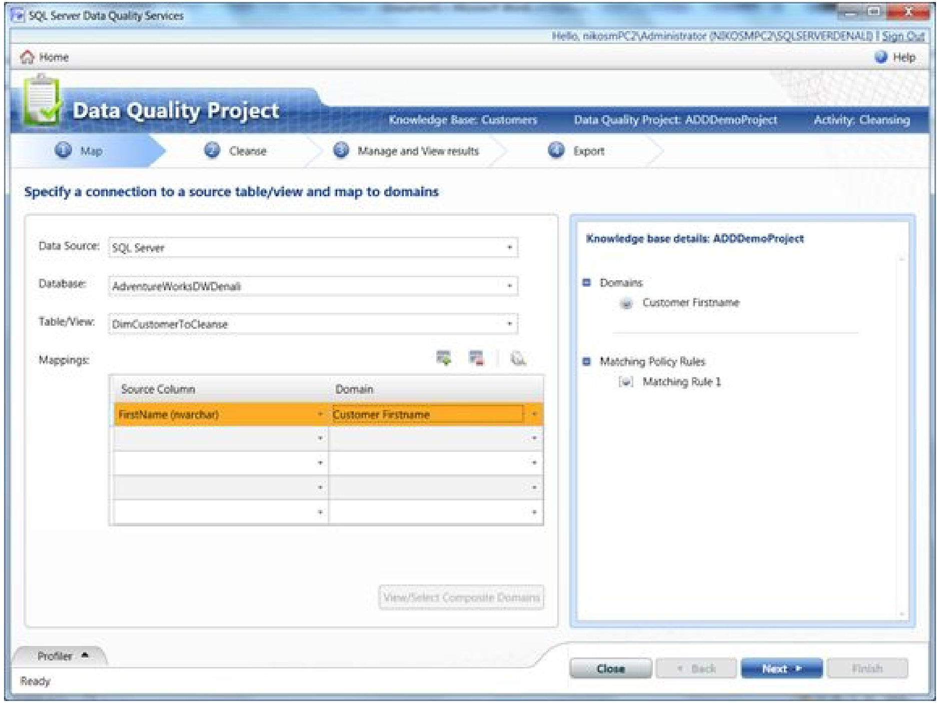
Knowledge base details: ADDDemoProject

- Domains
  - Customer Firstname
- Matching Policy Rules
  - Matching Rule 1

Ready

SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Home
Help
Data Quality Project
Map
Cleanse
Manage and View results
Export
View/Select Composite Domains
Profiler
Close
Back
Next
Finish

Limpeza de Dados


SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Home
Help
Data Quality Project
1 Map
2 Cleanse
3 Manage and View results
4 Export
View/Select Composite Domains
Profiler
Close
Back
Next
Finish
Ready

Knowledge Base: Customers Data Quality Project: ADDDemoProject Activity: Cleansing

Specify a connection to a source table/view and map to domains

Data Source: SQL Server

Database: AdventureWorksDWDenali

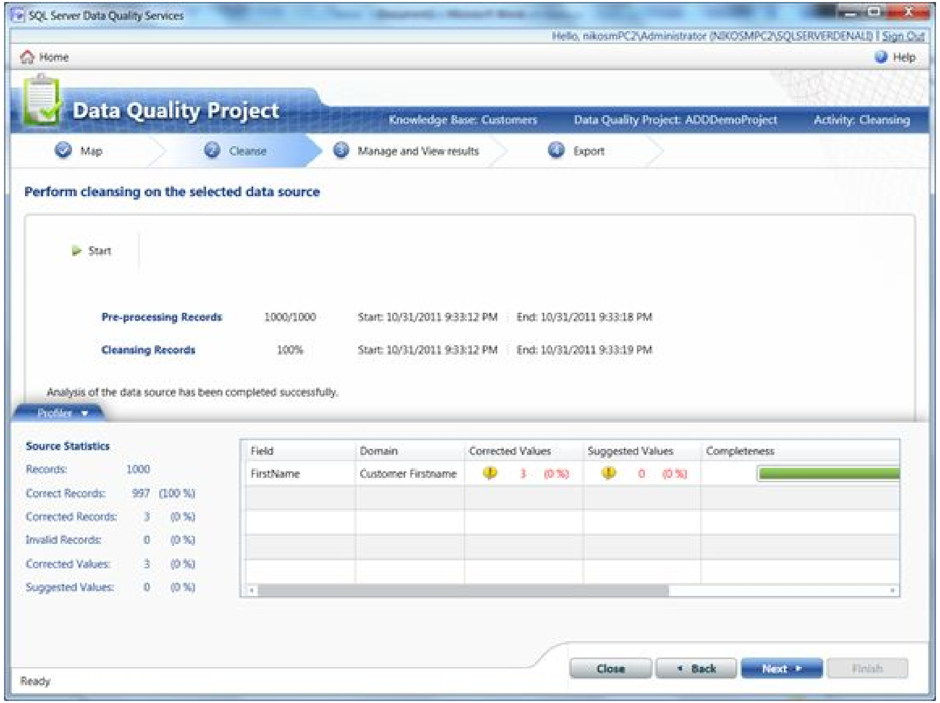
Table/View: DimCustomerToCleanse

Mappings:

Source Column	Domain
FirstName (nvarchar)	Customer Firstname

Knowledge base details: ADDDemoProject

- Domains
  - Customer Firstname
- Matching Policy Rules
  - Matching Rule 1



Knowledge Base: Customers Data Quality Project: ADDDemoProject Activity: Cleansing

Perform cleansing on the selected data source

Pre-processing Records	1000/1000	Start: 10/31/2011 9:33:12 PM	End: 10/31/2011 9:33:18 PM
Cleansing Records	100%	Start: 10/31/2011 9:33:12 PM	End: 10/31/2011 9:33:19 PM

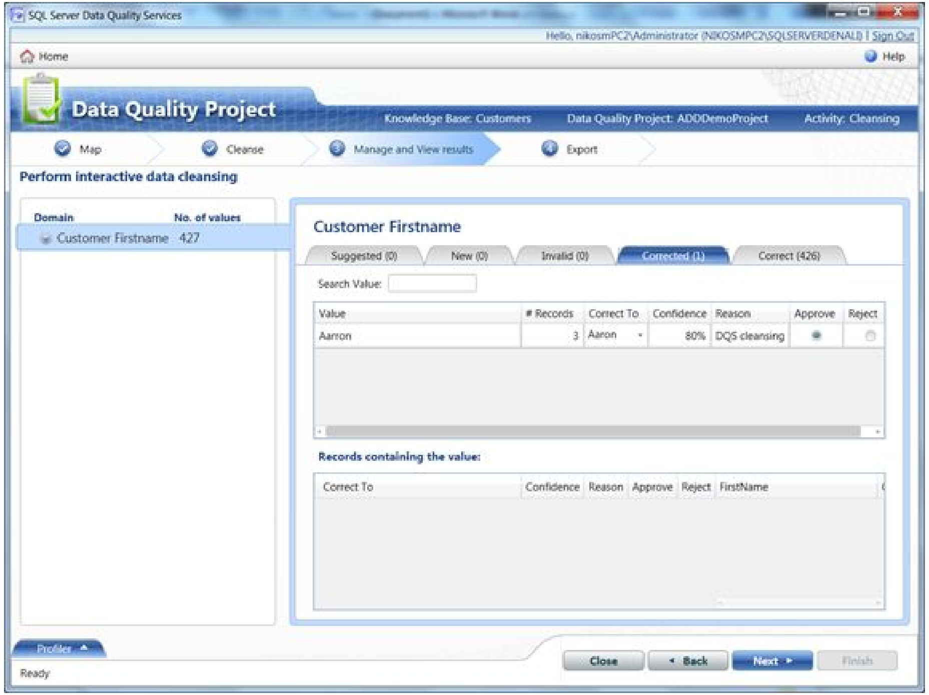
Analysis of the data source has been completed successfully.

Source Statistics

Records:	1000	
Correct Records:	997	(100 %)
Corrected Records:	3	(0 %)
Invalid Records:	0	(0 %)
Corrected Values:	3	(0 %)
Suggested Values:	0	(0 %)

Field	Domain	Corrected Values	Suggested Values	Completeness
FirstName	Customer Firstname	3 (0 %)	0 (0 %)	🟩

Limpeza de Dados


SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Home
Help
Data Quality Project
Map
Cleanse
Manage and View results
Export
Profiler
Close
Back
Next
Finish
Ready

Knowledge Base: Customers   Data Quality Project: ADDDemoProject   Activity: Cleansing

Perform interactive data cleansing

Domain	No. of values
Customer Firstname	427

Customer Firstname

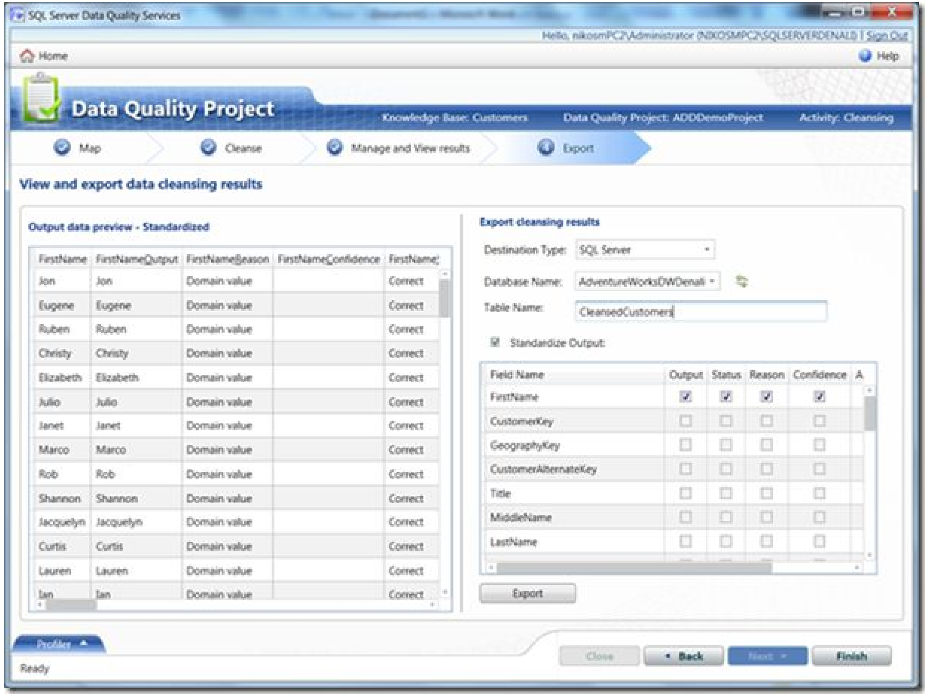
Suggested (0) | New (0) | Invalid (0) | Corrected (1) | Correct (426)

Search Value:

Value	# Records	Correct To	Confidence	Reason	Approve	Reject
Aarron	3	Aaron	80%	DQS cleansing	●	○

Records containing the value:

Correct To	Confidence	Reason	Approve	Reject	FirstName


SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Home
Help
Data Quality Project
Map
Cleanse
Manage and View results
Export
Profiler
Close
Back
Next
Finish
Ready

Knowledge Base: Customers   Data Quality Project: ADDDemoProject   Activity: Cleansing

View and export data cleansing results

Output data preview - Standardized

FirstName	FirstNameOutput	FirstNameReason	FirstNameConfidence	FirstName
Jon	Jon	Domain value		Correct
Eugene	Eugene	Domain value		Correct
Ruben	Ruben	Domain value		Correct
Christy	Christy	Domain value		Correct
Elizabeth	Elizabeth	Domain value		Correct
Julio	Julio	Domain value		Correct
Janet	Janet	Domain value		Correct
Marco	Marco	Domain value		Correct
Rob	Rob	Domain value		Correct
Shannon	Shannon	Domain value		Correct
Jacquelyn	Jacquelyn	Domain value		Correct
Curtis	Curtis	Domain value		Correct
Lauren	Lauren	Domain value		Correct
Ian	Ian	Domain value		Correct

Export cleansing results

Destination Type: SQL Server

Database Name: AdventureWorksDWDenali

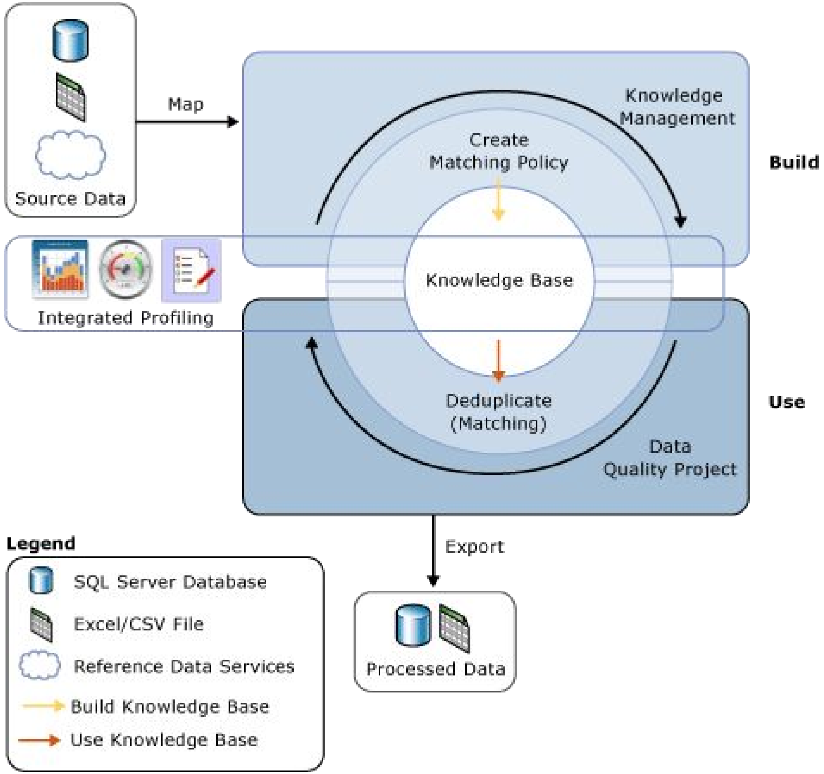
Table Name: CleansedCustomers

☑ Standardize Output:

Field Name	Output	Status	Reason	Confidence	A
FirstName	☑	☑	☑	☑	
CustomerKey	□	□	□	□	
GeographyKey	□	□	□	□	
CustomerAlternateKey	□	□	□	□	
Title	□	□	□	□	
MiddleName	□	□	□	□	
LastName	□	□	□	□	

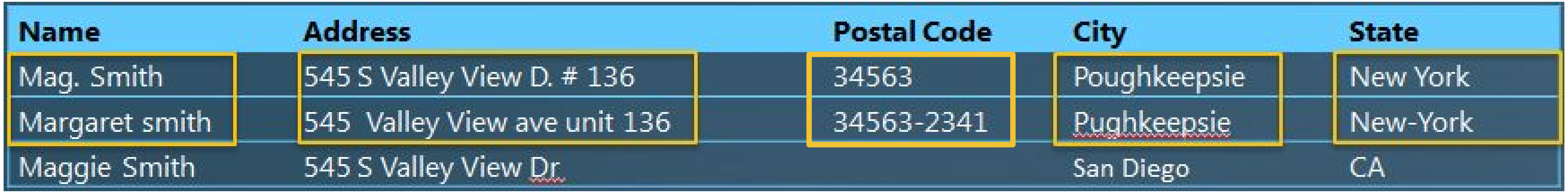
Export

Correspondência (Matching)



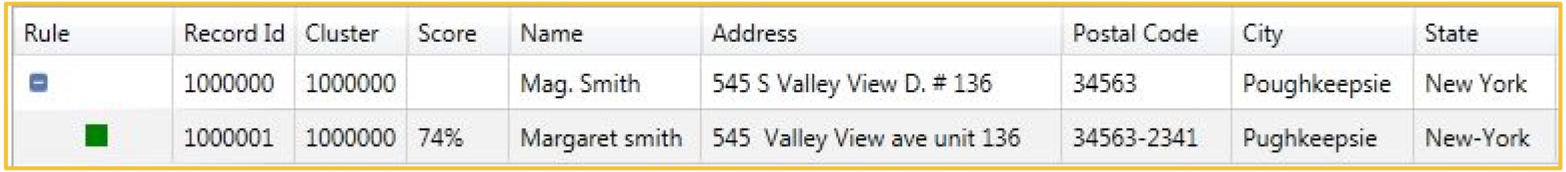
Exemplo Matching

Dados



Name	Address	Postal Code	City	State
Mag. Smith	545 S Valley View D. # 136	34563	Poughkeepsie	New York
Margaret smith	545 Valley View ave unit 136	34563-2341	Pughkeepsie	New-York
Maggie Smith	545 S Valley View Dr		San Diego	CA

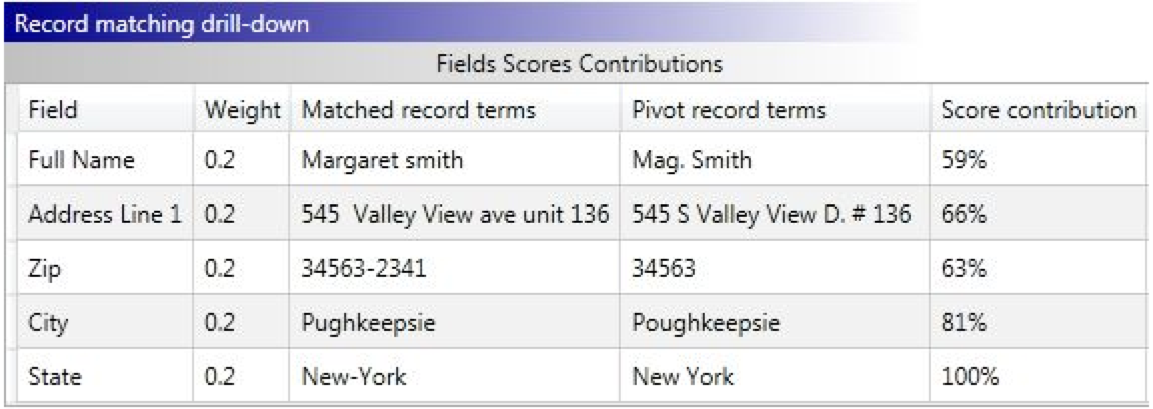
Resultados do matching



Rule	Record Id	Cluster	Score	Name	Address	Postal Code	City	State
⊟	1000000	1000000		Mag. Smith	545 S Valley View D. # 136	34563	Poughkeepsie	New York
🟩	1000001	1000000	74%	Margaret smith	545 Valley View ave unit 136	34563-2341	Pughkeepsie	New-York

Exemplo Matching

Resolução de Correspondências



Record matching drill-down

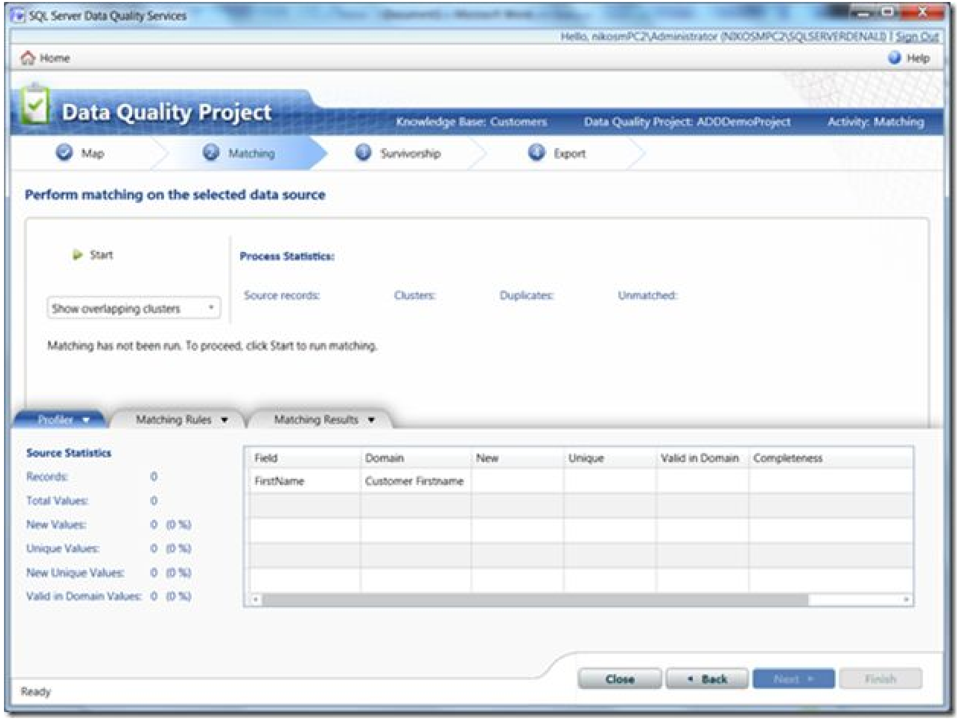
Fields Scores Contributions

Field	Weight	Matched record terms	Pivot record terms	Score contribution
Full Name	0.2	Margaret smith	Mag. Smith	59%
Address Line 1	0.2	545 Valley View ave unit 136	545 S Valley View D. # 136	66%
Zip	0.2	34563-2341	34563	63%
City	0.2	Pughkeepsie	Poughkeepsie	81%
State	0.2	New-York	New York	100%

Propriedades de correspondência

- Similaridade
- Peso
- Pré-requisito
- Pontuação mínima de correspondência

Correspondência de dados


SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Home
Help
Data Quality Project
Knowledge Base: Customers
Data Quality Project: ADDDemoProject
Activity: Matching
Map
Matching
Survivorship
Export
Start
Show overlapping clusters
Profiler
Matching Rules
Matching Results
Close
Back
Next
Finish

Perform matching on the selected data source

Process Statistics:

Source records: Clusters: Duplicates: Unmatched:

Matching has not been run. To proceed, click Start to run matching.

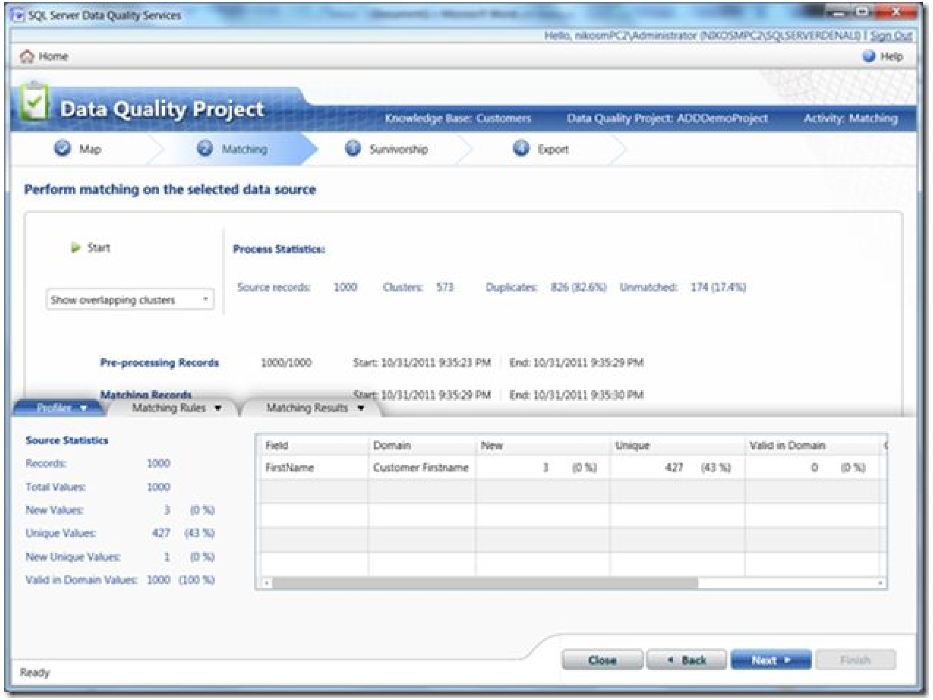
Source Statistics

Records:	0	
Total Values:	0	
New Values:	0	(0 %)
Unique Values:	0	(0 %)
New Unique Values:	0	(0 %)
Valid in Domain Values:	0	(0 %)

Field	Domain	New	Unique	Valid in Domain	Completeness
FirstName	Customer Firstname				

Ready

Correspondência de dados



SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Home
Help
Data Quality Project
Map
Matching
Survivorship
Export
Start
Show overlapping clusters
Profiler
Matching Rules
Matching Results
Close
Back
Next
Finish
Ready

Knowledge Base: Customers Data Quality Project: ADDDemoProject Activity: Matching

Perform matching on the selected data source

Process Statistics:

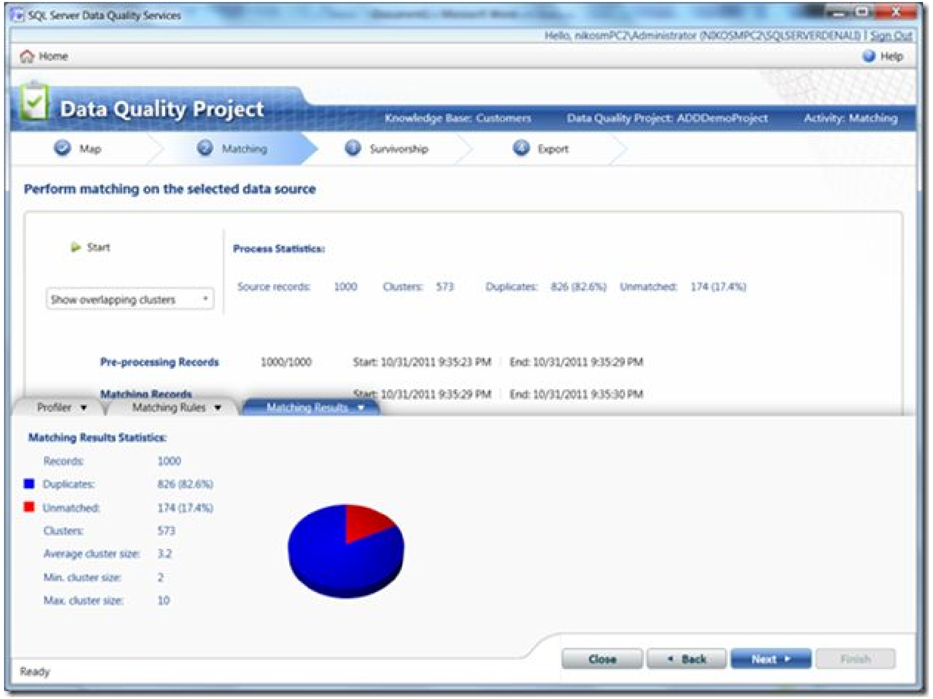
Source records: 1000 Clusters: 573 Duplicates: 826 (82.6%) Unmatched: 174 (17.4%)

Pre-processing Records	1000/1000	Start: 10/31/2011 9:35:23 PM	End: 10/31/2011 9:35:29 PM
Matching Records		Start: 10/31/2011 9:35:29 PM	End: 10/31/2011 9:35:30 PM

Source Statistics

Records:	1000	
Total Values:	1000	
New Values:	3	(0 %)
Unique Values:	427	(43 %)
New Unique Values:	1	(0 %)
Valid in Domain Values:	1000	(100 %)

Field	Domain	New	Unique	Valid in Domain
FirstName	Customer Firstname	3 (0 %)	427 (43 %)	0 (0 %)



SQL Server Data Quality Services
Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out
Home
Help
Data Quality Project
Map
Matching
Survivorship
Export
Start
Show overlapping clusters
Profiler
Matching Rules
Matching Results
Close
Back
Next
Finish
Ready

Knowledge Base: Customers Data Quality Project: ADDDemoProject Activity: Matching

Perform matching on the selected data source

Process Statistics:

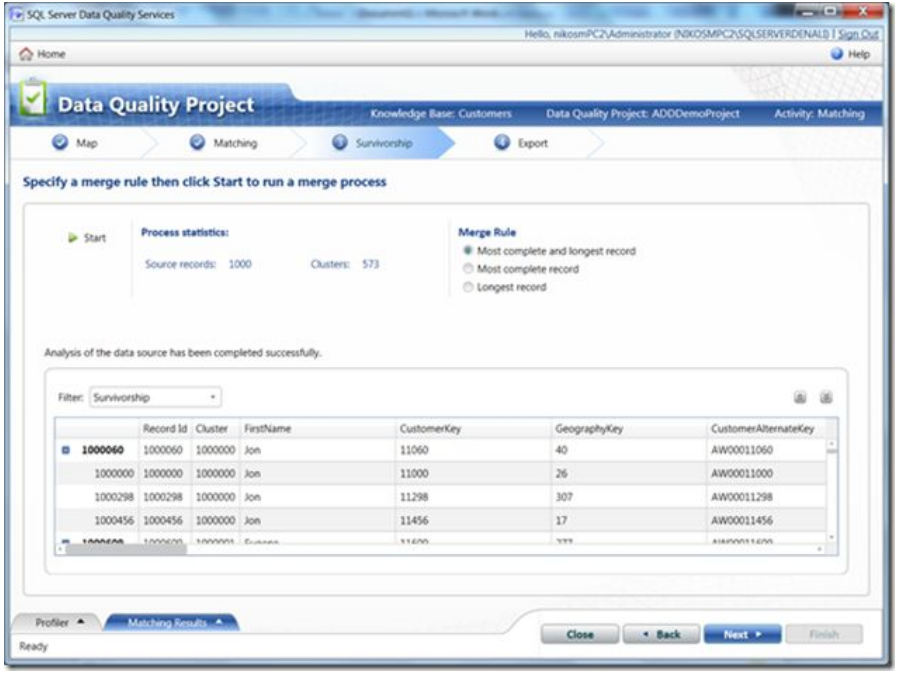
Source records: 1000 Clusters: 573 Duplicates: 826 (82.6%) Unmatched: 174 (17.4%)

Pre-processing Records	1000/1000	Start: 10/31/2011 9:35:23 PM	End: 10/31/2011 9:35:29 PM
Matching Records		Start: 10/31/2011 9:35:29 PM	End: 10/31/2011 9:35:30 PM

Matching Results Statistics:

Records:	1000
🟦 Duplicates:	826 (82.6%)
🟥 Unmatched:	174 (17.4%)
Clusters:	573
Average cluster size:	3.2
Min. cluster size:	2
Max. cluster size:	10

Correspondência de dados



SQL Server Data Quality Services

Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out

Home — Help

Data Quality Project

Knowledge Base: Customers — Data Quality Project: ADDDemoProject — Activity: Matching

Map — Matching — Survivorship — Export

Specify a merge rule then click Start to run a merge process

Start

Process statistics:

Source records: 1000 Clusters: 573

Merge Rule

● Most complete and longest record

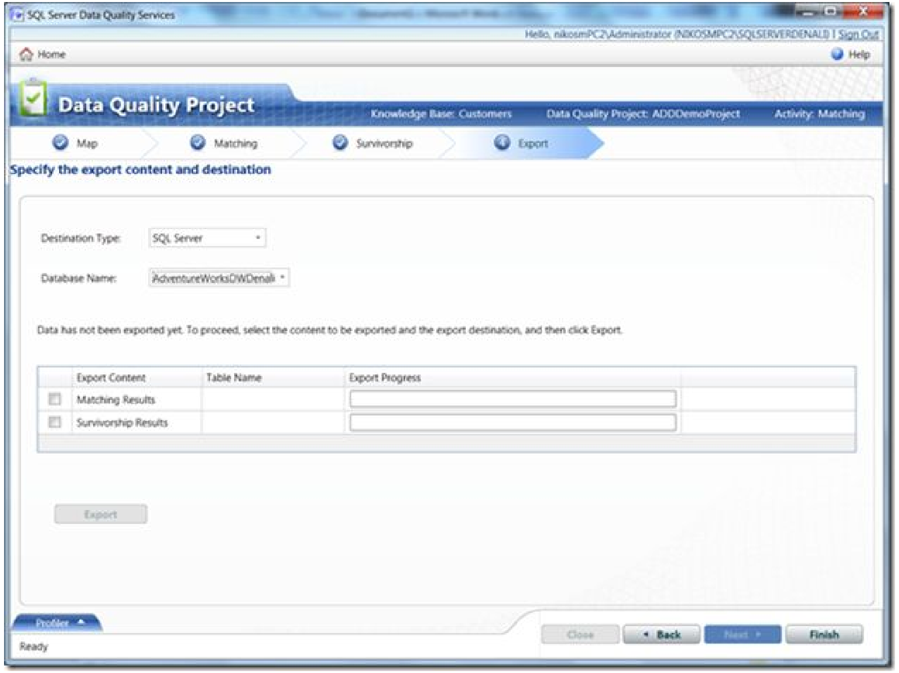
○ Most complete record

○ Longest record

Analysis of the data source has been completed successfully.

Filter: Survivorship

	Record Id	Cluster	FirstName	CustomerKey	GeographyKey	CustomerAlternateKey
1000060	1000060	1000000	Jon	11060	40	AW00011060
1000000	1000000	1000000	Jon	11000	26	AW00011000
1000298	1000298	1000000	Jon	11298	307	AW00011298
1000456	1000456	1000000	Jon	11456	17	AW00011456
1000600	1000600	1000001	[illegible]	11600	[illegible]	AW00011600



SQL Server Data Quality Services

Hello, nikosmPC2\Administrator (NIKOSMPC2\SQLSERVERDENALI) | Sign Out

Home — Help

Data Quality Project

Knowledge Base: Customers — Data Quality Project: ADDDemoProject — Activity: Matching

Map — Matching — Survivorship — Export

Specify the export content and destination

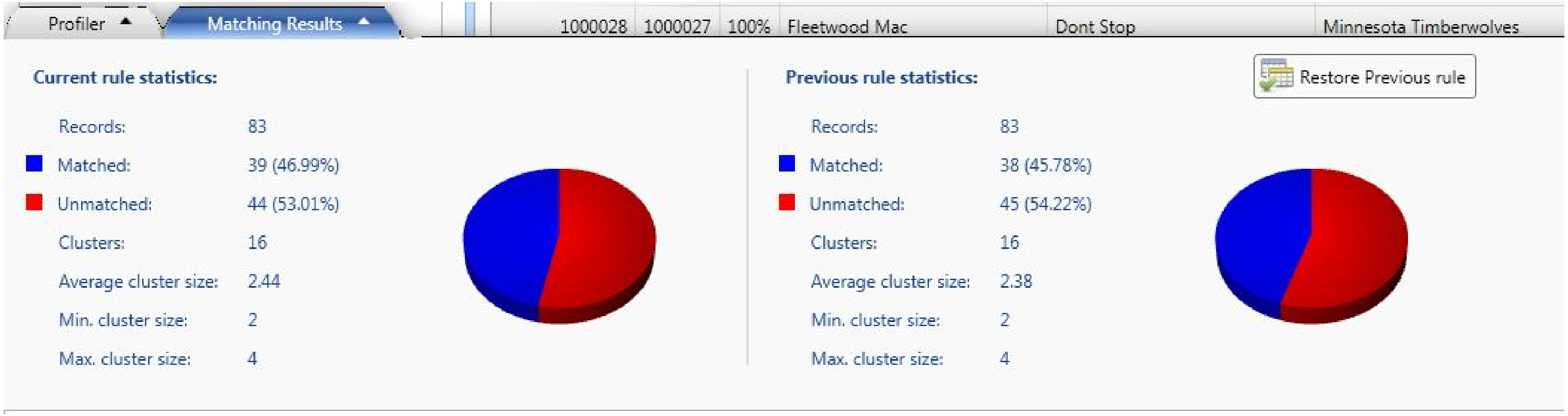
Destination Type: SQL Server

Database Name: AdventureWorksDWDenali

Data has not been exported yet. To proceed, select the content to be exported and the export destination, and then click Export.

	Export Content	Table Name	Export Progress
□	Matching Results		
□	Survivorship Results		

Correspondência de dados



Current rule statistics:

	Records:	83
🟦	Matched:	39 (46.99%)
🟥	Unmatched:	44 (53.01%)
	Clusters:	16
	Average cluster size:	2.44
	Min. cluster size:	2
	Max. cluster size:	4

Previous rule statistics:

	Records:	83
🟦	Matched:	38 (45.78%)
🟥	Unmatched:	45 (54.22%)
	Clusters:	16
	Average cluster size:	2.38
	Min. cluster size:	2
	Max. cluster size:	4

Recapitulando...

- Remove carga do processo ETL
- Oferece ferramentas de validação e estatísticas sobre regras de domínio
- Pode limpar e remover duplicações
- Pode ser integrada ao SSIS com o DQS Transform
- Baseado em base de conhecimento
- Pode usar dados de terceiros para conhecimento

Referências

1. Big Data: 20 Mind-Boggling Facts Everyone Must Read
2. Integrating and governing Big Data
3. Extracting Value from Chaos
4. The Cost of Poor Data Quality
5. Data Quality Services Demo