DQS step-by-step

Domain Management, Knowledge Discovery and Matching Policies

Alexander Karl

SQL Server Öncüleri

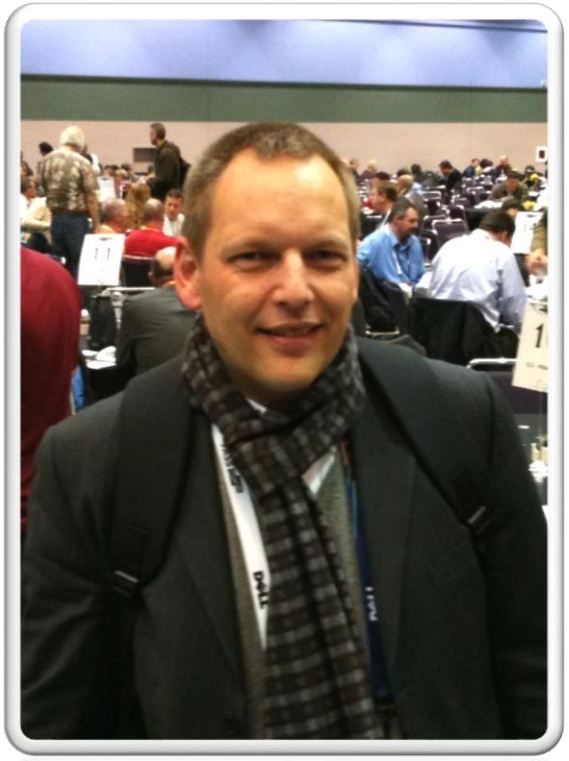
PASS SQL saturday

#359 | ISTANBUL 2014

Sponsors

Main Sponsor	Microsoft
Gold Sponsors	DATADYNAMIC TCM Teknoloji Çözümleri Merkezi MECHSOFT you need, we solve!

About me



Alexander Karl

.net - CDE

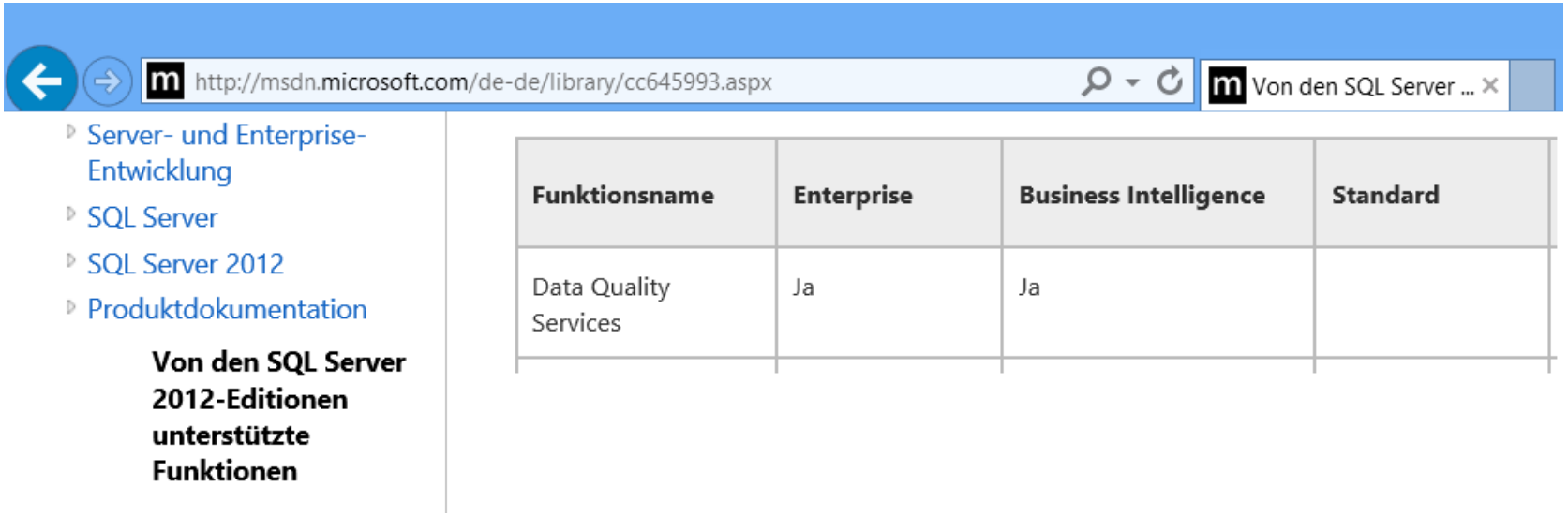
SQL + BI Consultant

Microsoft CERTIFIED Trainer

Microsoft CERTIFIED IT Professional

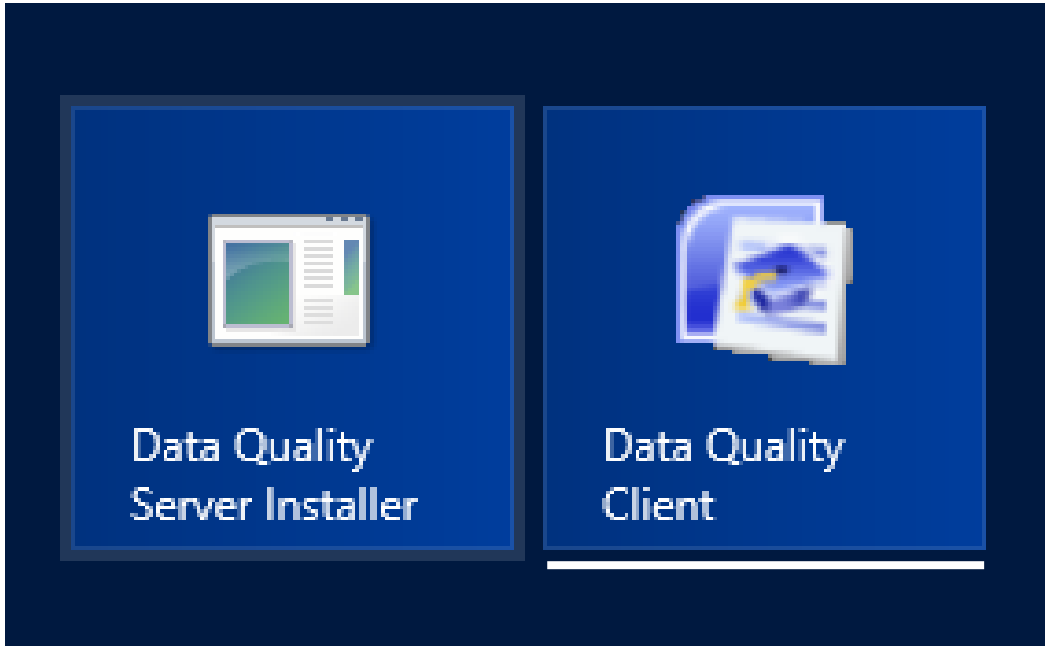
Database Administrator 2008
Server Administrator on Windows Server® 2008
Database Administrator on SQL Server® 2005

SQL Server 2012 editions

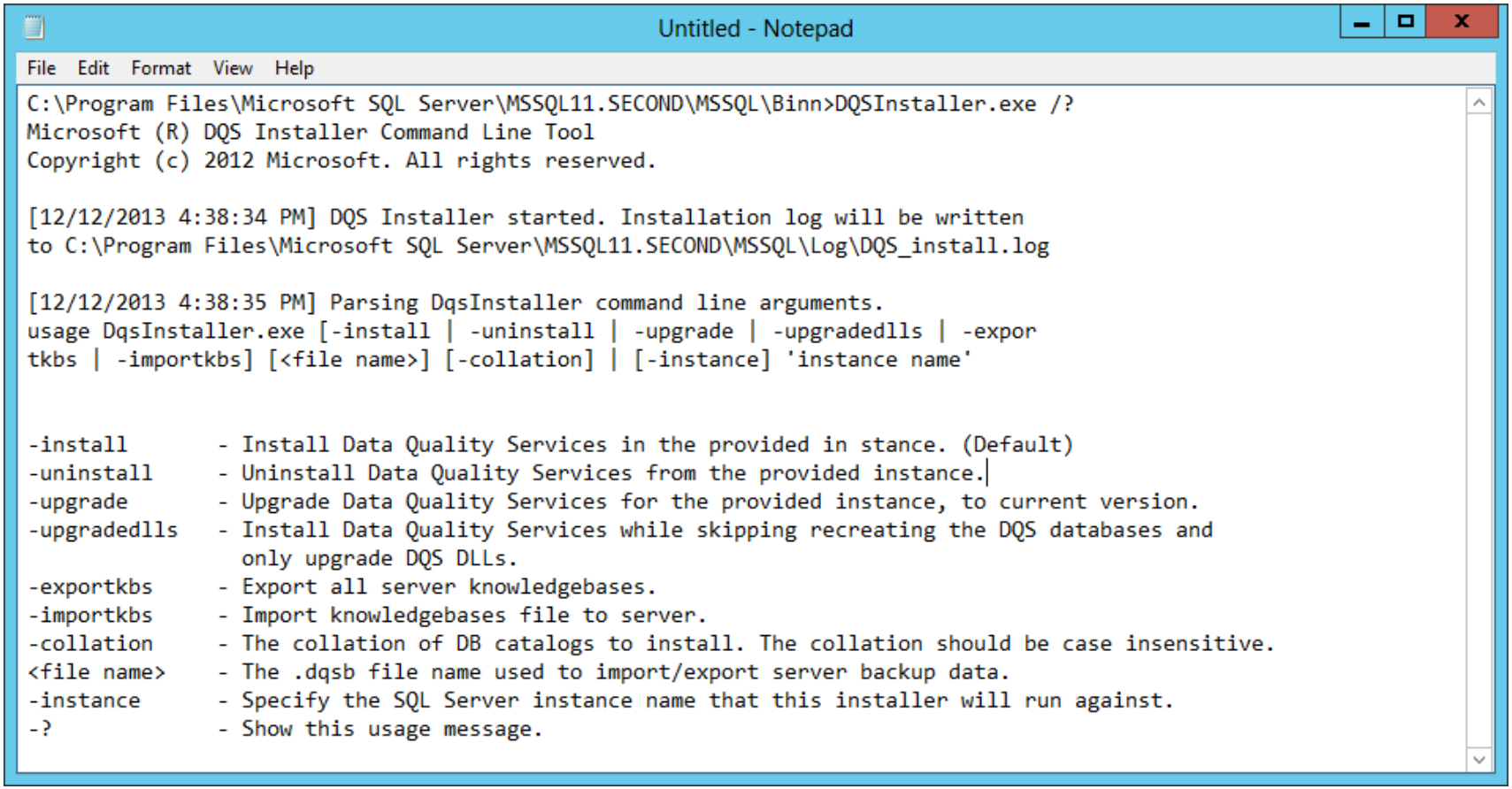


Funktionsname	Enterprise	Business Intelligence	Standard
Data Quality Services	Ja	Ja	

SQL server 2012 DQS



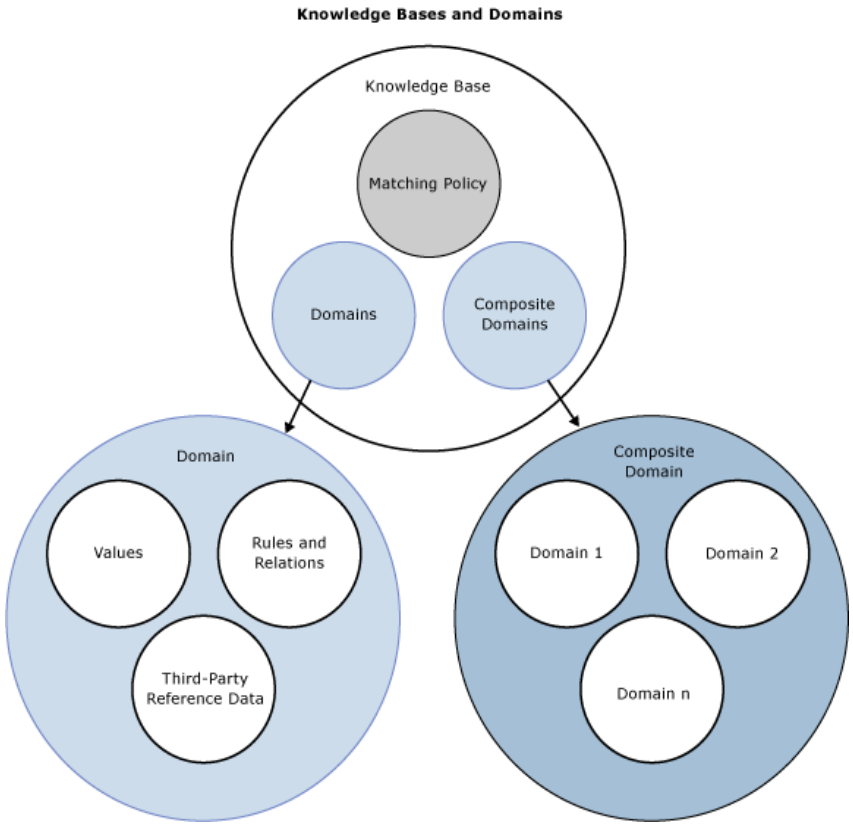
DQS Installer



DQS *"work items"*

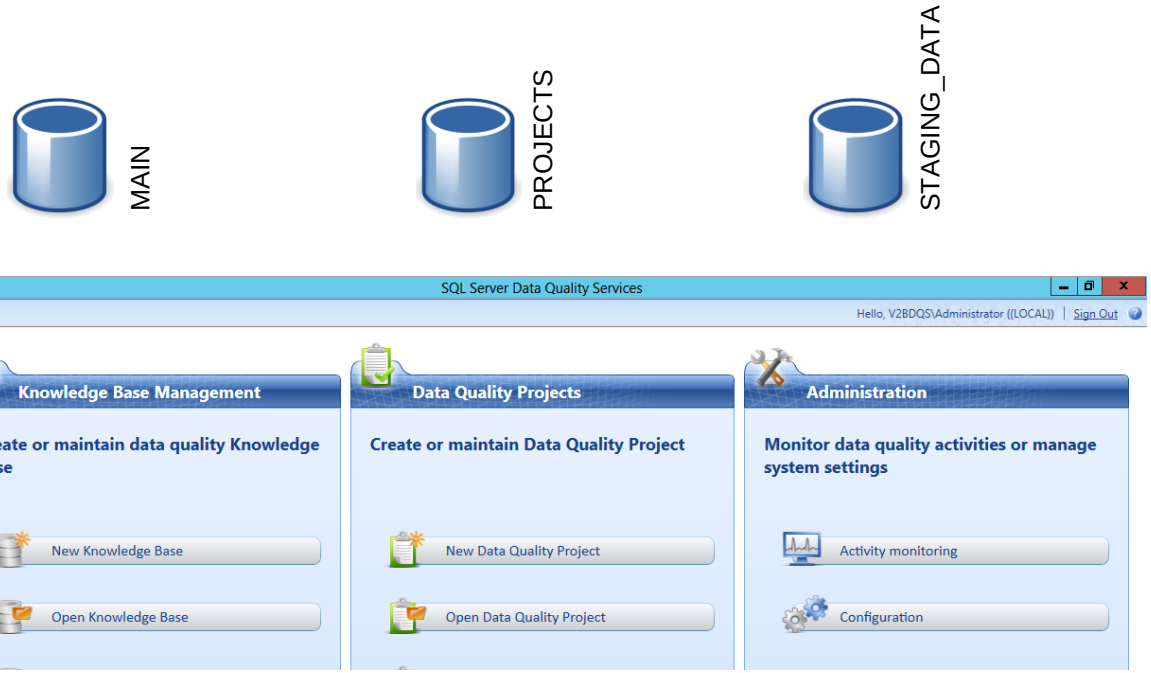
- Domain Management
- Knowledge Discovery
- Matching Policy

Domain Management

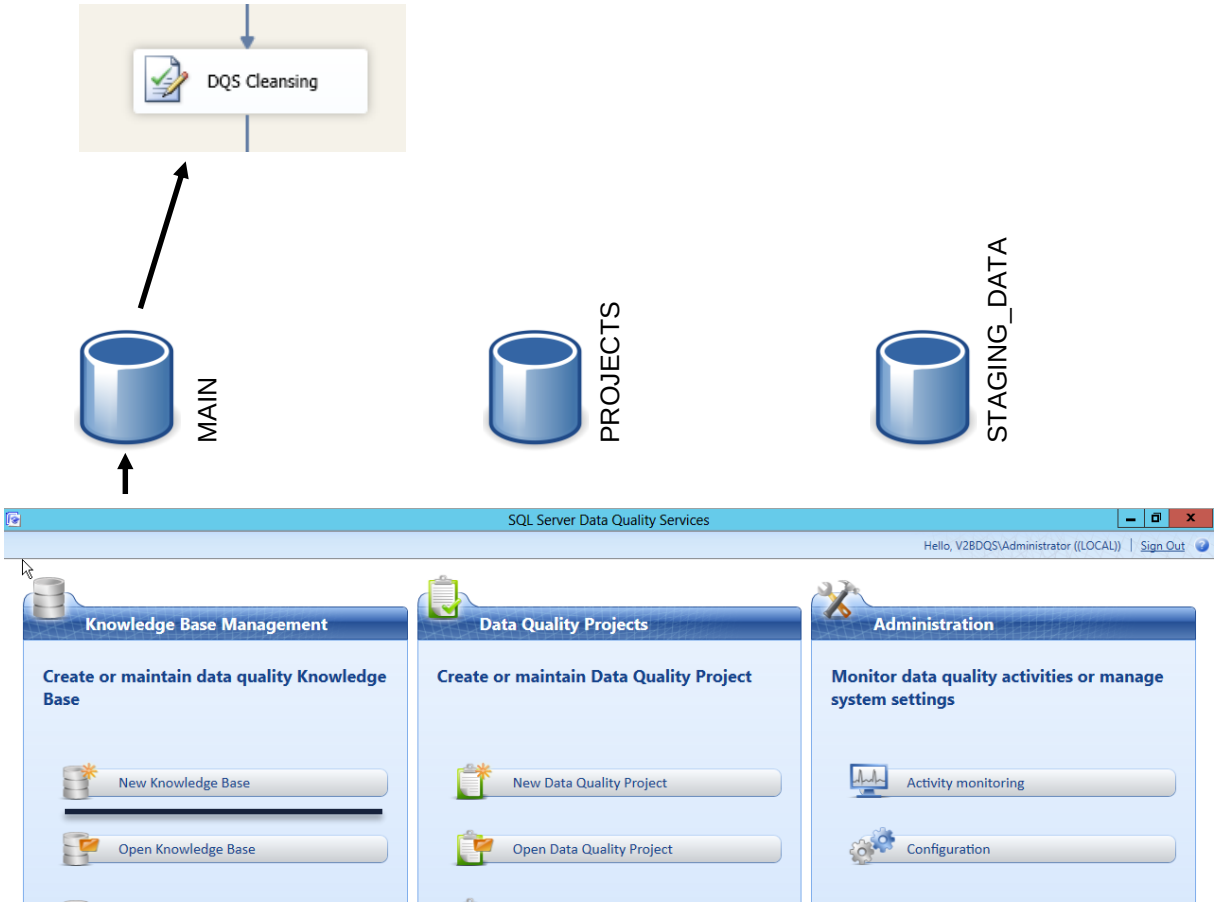


http://technet.microsoft.com/en-us/library/gg524799.aspx

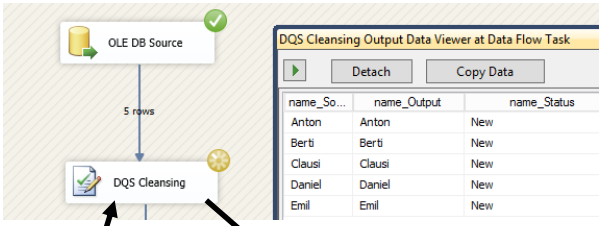
DQS workflow

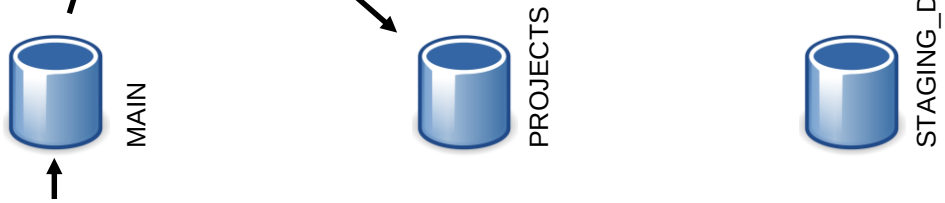


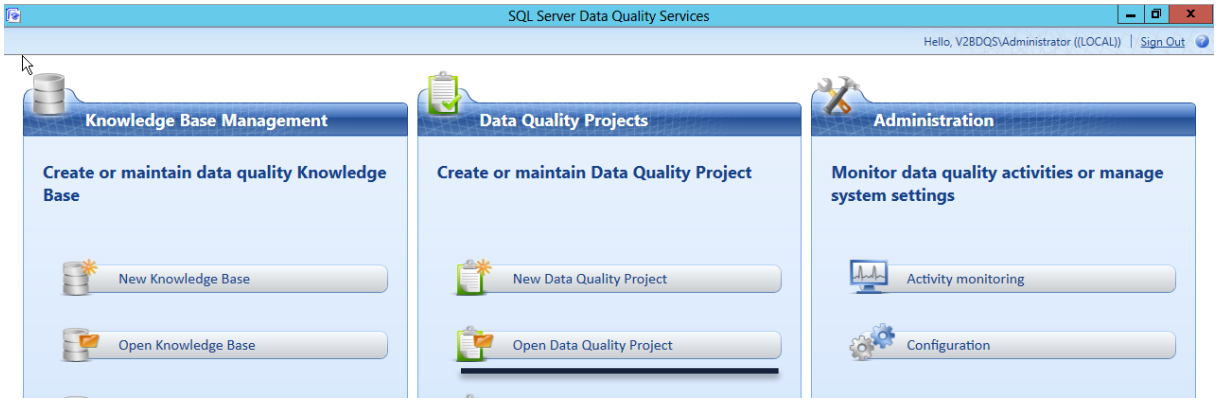
DQS workflow



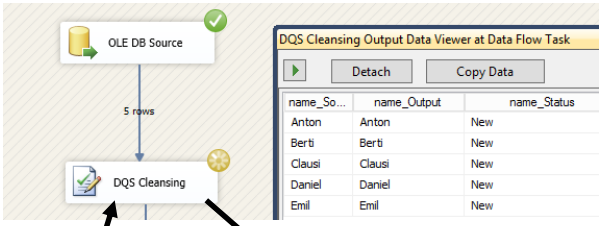
DQS workflow

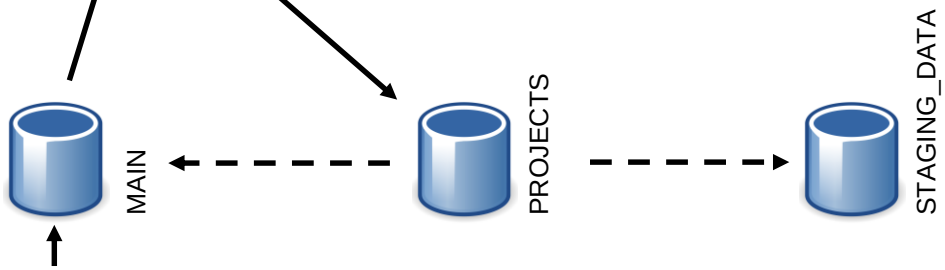


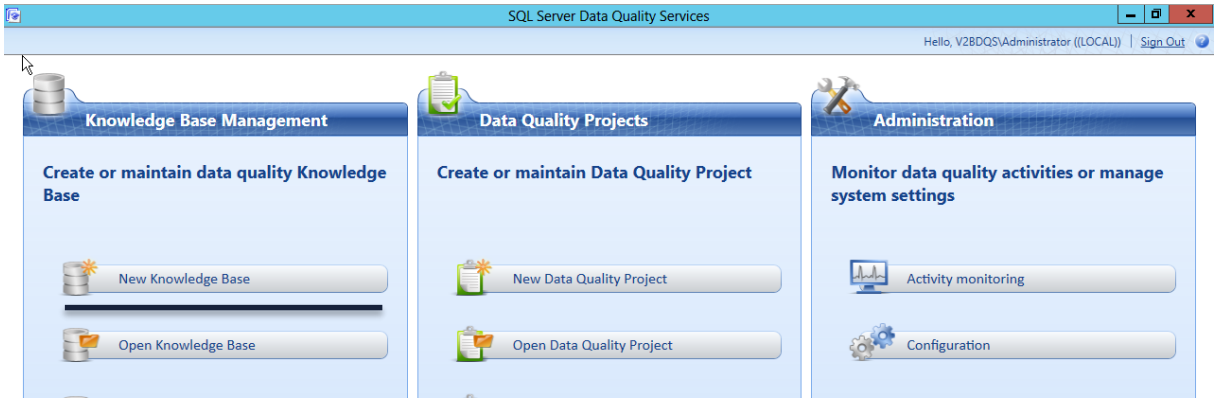




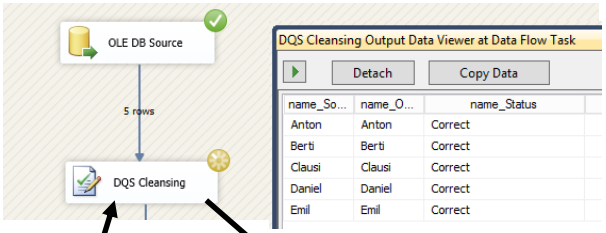
DQS workflow

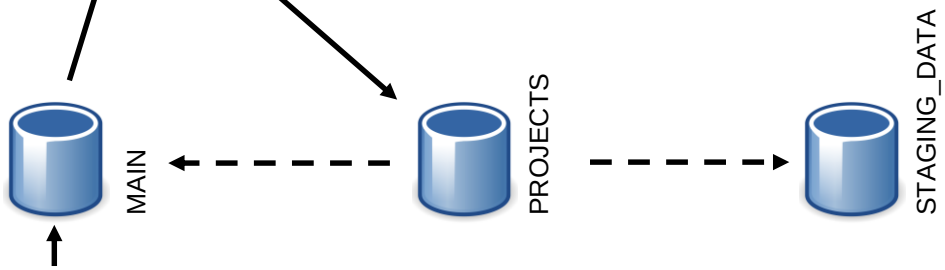


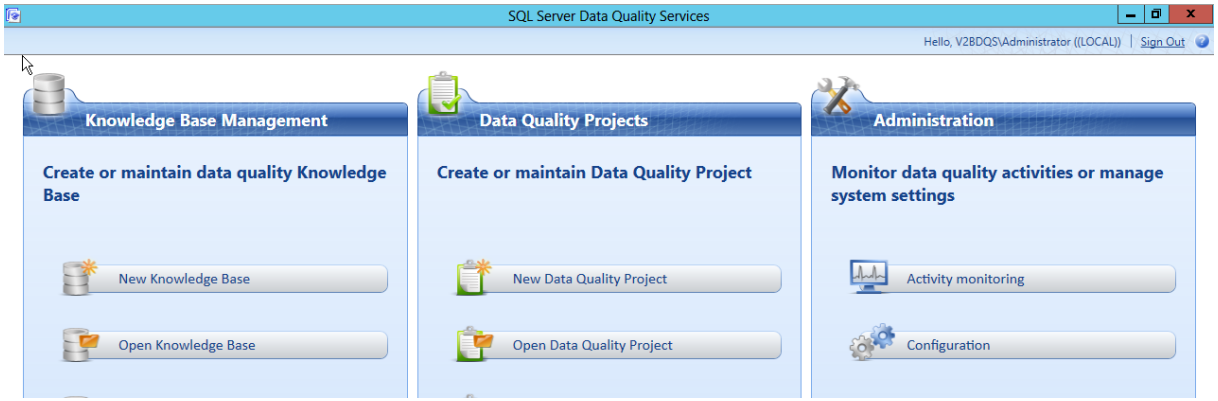




DQS workflow

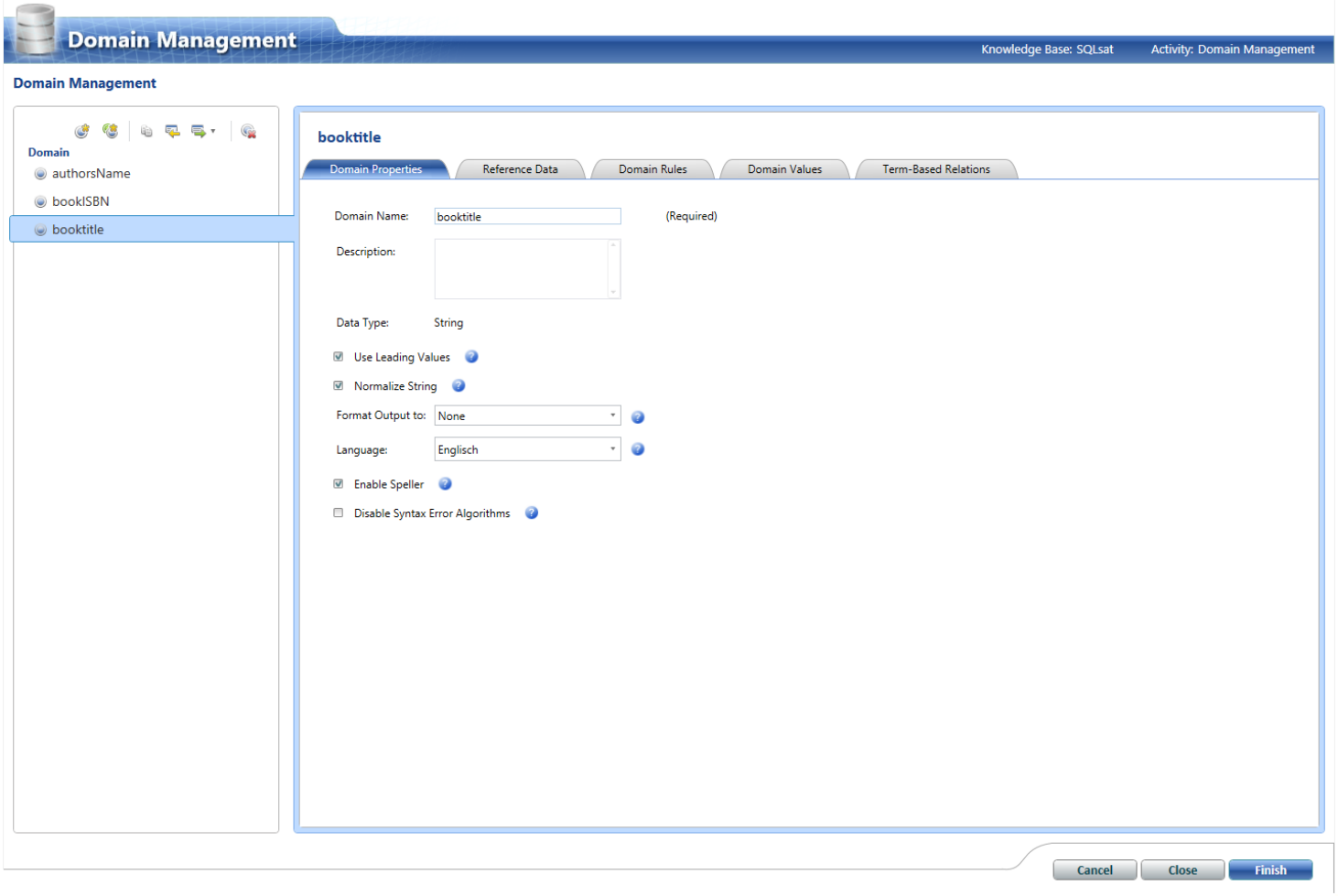




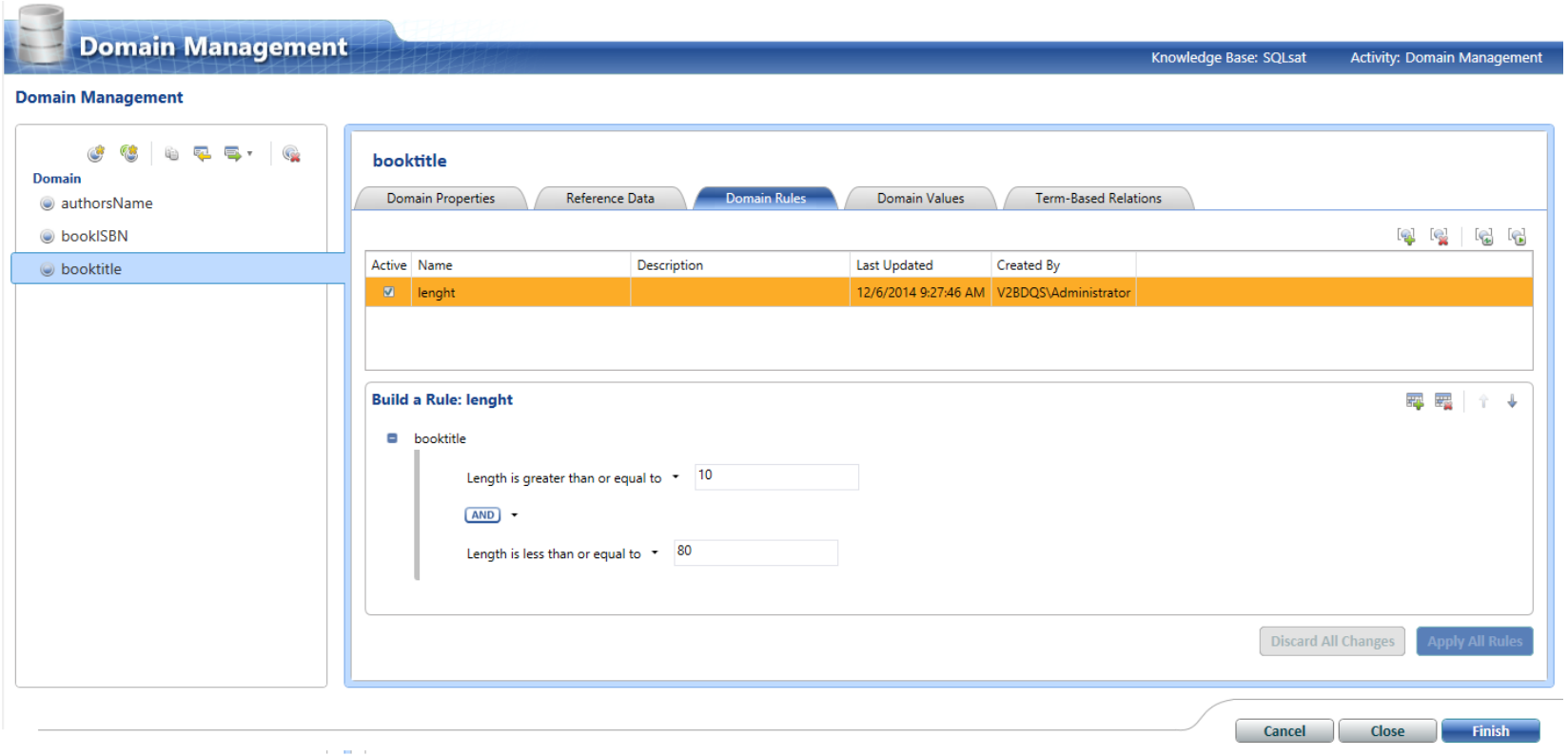


DEMO

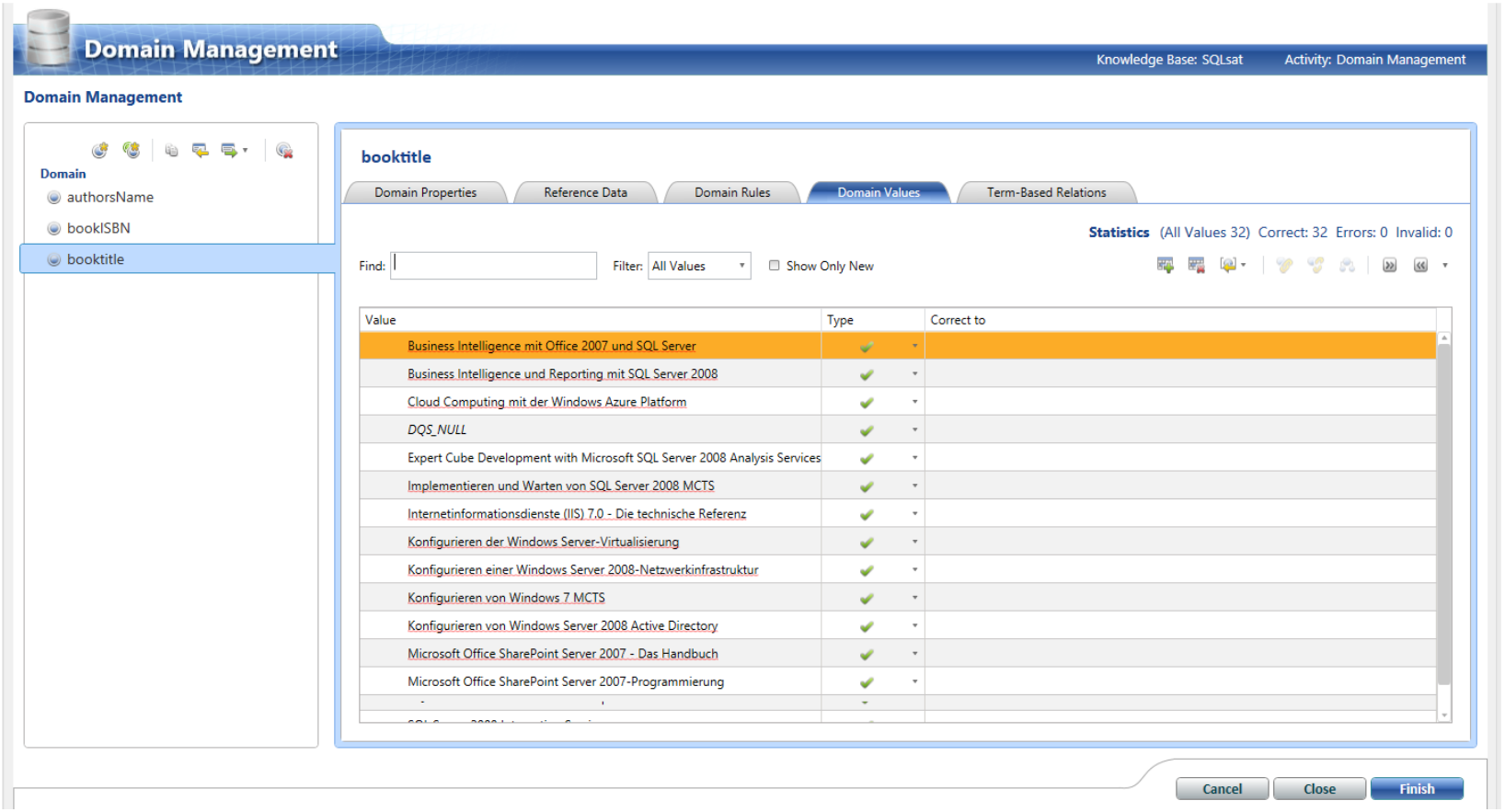
Demo result


Domain Management
Knowledge Base: SQLsat
Activity: Domain Management
Domain Management
Domain
authorsName
bookISBN
booktitle
booktitle
Domain Properties
Reference Data
Domain Rules
Domain Values
Term-Based Relations
Domain Name:
booktitle
(Required)
Description:
Data Type:
String
Use Leading Values
Normalize String
Format Output to:
None
Language:
Englisch
Enable Speller
Disable Syntax Error Algorithms
Cancel
Close
Finish

Demo result


Domain Management
Knowledge Base: SQLsat
Activity: Domain Management
Domain Management
Domain
authorsName
bookISBN
booktitle
booktitle
Domain Properties
Reference Data
Domain Rules
Domain Values
Term-Based Relations
Active
Name
Description
Last Updated
Created By
lenght
12/6/2014 9:27:46 AM
V2BDQS\Administrator
Build a Rule: lenght
booktitle
Length is greater than or equal to
10
AND
Length is less than or equal to
80
Discard All Changes
Apply All Rules
Cancel
Close
Finish

Demo result

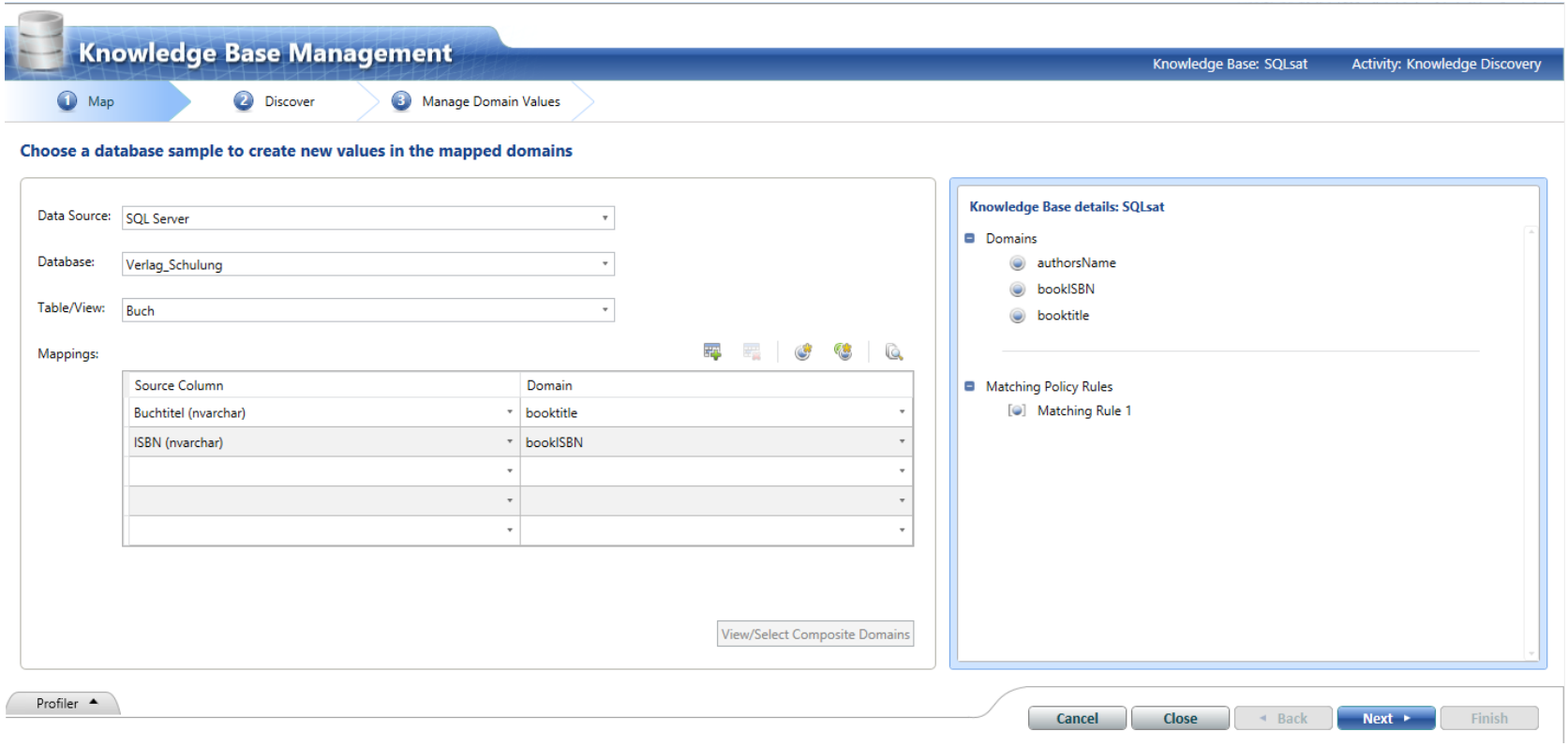

Domain Management
Knowledge Base: SQLsat
Activity: Domain Management
Domain Management
Domain
authorsName
bookISBN
booktitle
booktitle
Domain Properties
Reference Data
Domain Rules
Domain Values
Term-Based Relations
Statistics (All Values 32) Correct: 32 Errors: 0 Invalid: 0
Find:
Filter: All Values
Show Only New
Value
Type
Correct to
Business Intelligence mit Office 2007 und SQL Server
Business Intelligence und Reporting mit SQL Server 2008
Cloud Computing mit der Windows Azure Platform
DQS_NULL
Expert Cube Development with Microsoft SQL Server 2008 Analysis Services
Implementieren und Warten von SQL Server 2008 MCTS
Internetinformationsdienste (IIS) 7.0 - Die technische Referenz
Konfigurieren der Windows Server-Virtualisierung
Konfigurieren einer Windows Server 2008-Netzwerkinfrastruktur
Konfigurieren von Windows 7 MCTS
Konfigurieren von Windows Server 2008 Active Directory
Microsoft Office SharePoint Server 2007 - Das Handbuch
Microsoft Office SharePoint Server 2007-Programmierung
Cancel
Close
Finish

DQS *"work items"*

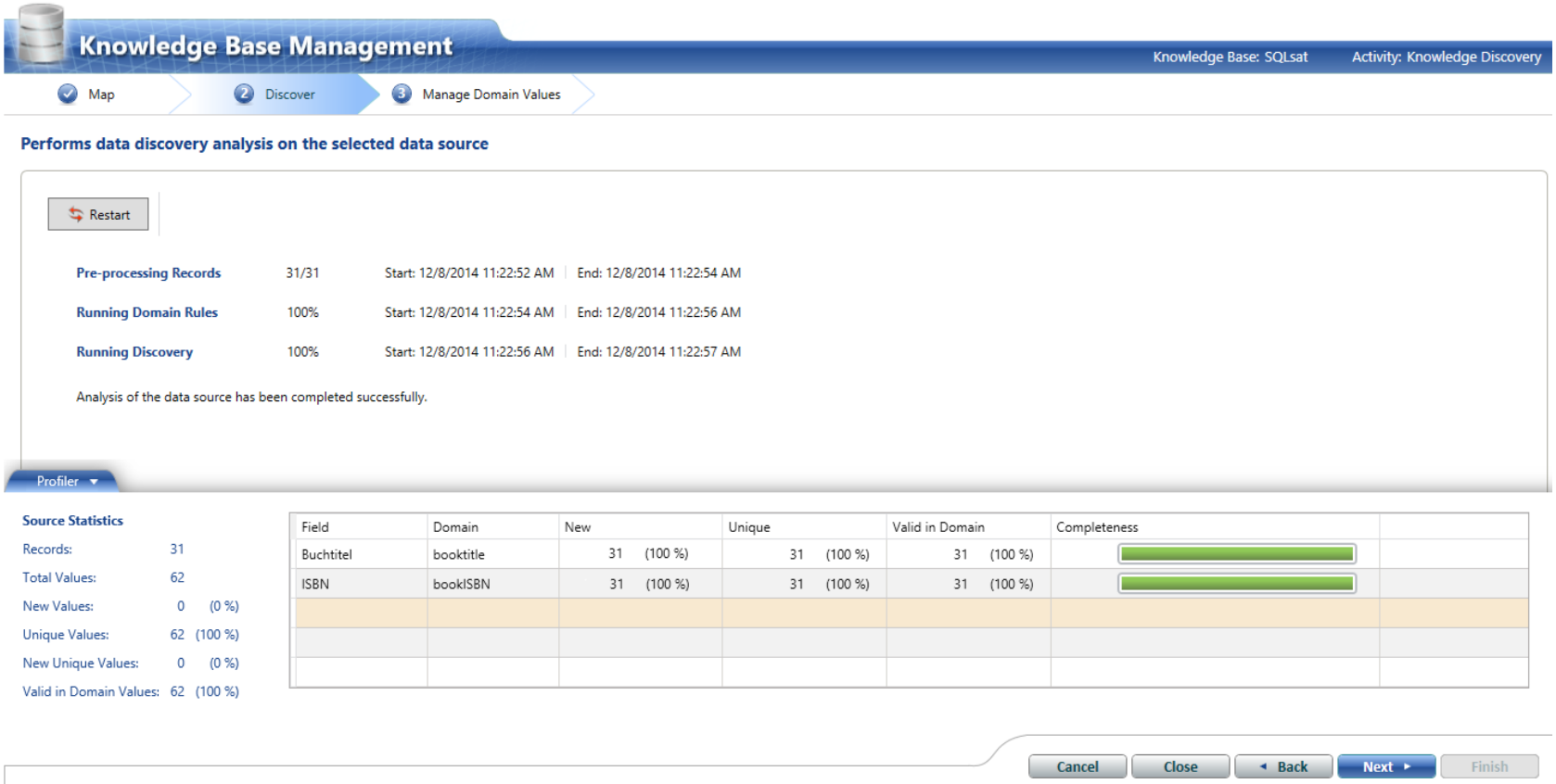
- Domain Management
- Knowledge Discovery
- Matching Policy

DEMO

Demo result


Knowledge Base Management
Knowledge Base: SQLsat
Activity: Knowledge Discovery
1 Map
2 Discover
3 Manage Domain Values
Choose a database sample to create new values in the mapped domains
Data Source: SQL Server
Database: Verlag_Schulung
Table/View: Buch
Mappings:
Source Column
Domain
Buchtitel (nvarchar)
booktitle
ISBN (nvarchar)
bookISBN
View/Select Composite Domains
Knowledge Base details: SQLsat
Domains
authorsName
bookISBN
booktitle
Matching Policy Rules
Matching Rule 1
Profiler
Cancel
Close
Back
Next
Finish

Demo result

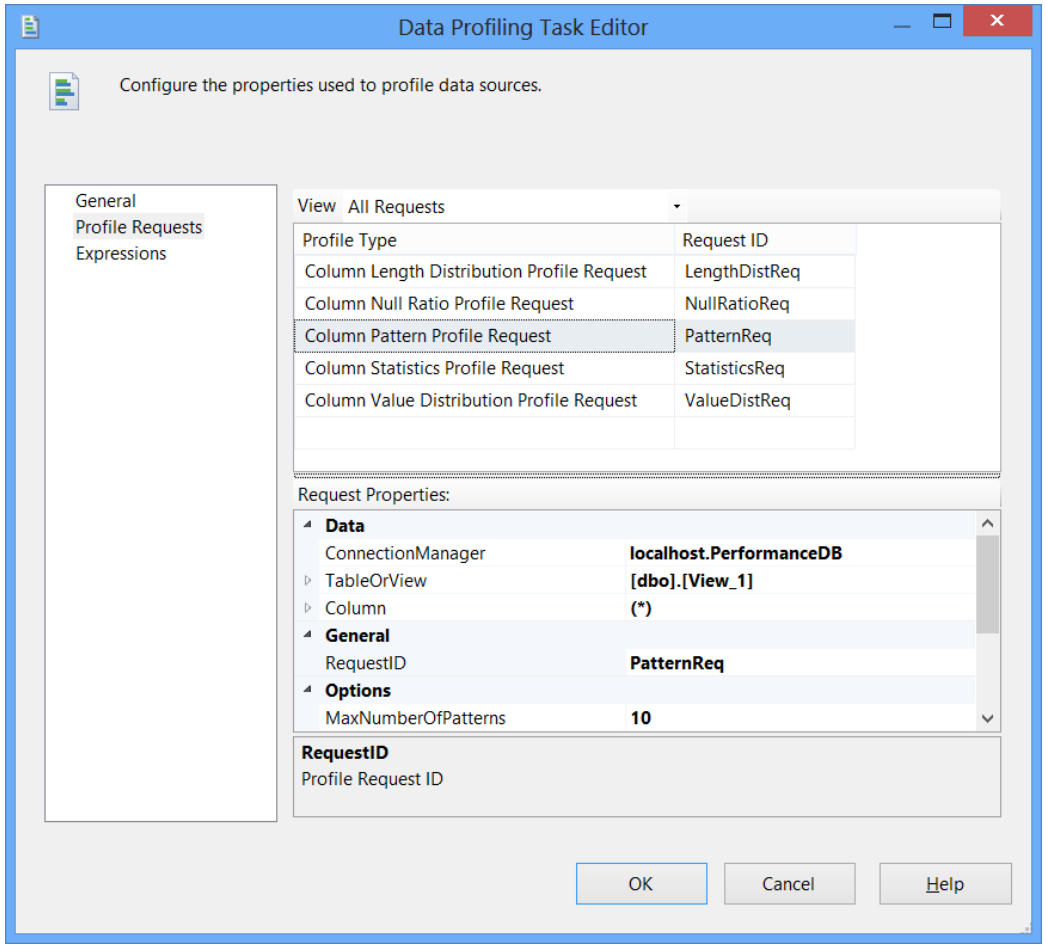

Knowledge Base Management
Knowledge Base: SQLsat
Activity: Knowledge Discovery
Map
2 Discover
3 Manage Domain Values
Performs data discovery analysis on the selected data source
Restart
Pre-processing Records 31/31 Start: 12/8/2014 11:22:52 AM End: 12/8/2014 11:22:54 AM
Running Domain Rules 100% Start: 12/8/2014 11:22:54 AM End: 12/8/2014 11:22:56 AM
Running Discovery 100% Start: 12/8/2014 11:22:56 AM End: 12/8/2014 11:22:57 AM
Analysis of the data source has been completed successfully.
Profiler
Source Statistics
Records: 31
Total Values: 62
New Values: 0 (0 %)
Unique Values: 62 (100 %)
New Unique Values: 0 (0 %)
Valid in Domain Values: 62 (100 %)
Field Domain New Unique Valid in Domain Completeness
Buchtitel booktitle 31 (100 %) 31 (100 %) 31 (100 %)
ISBN bookISBN 31 (100 %) 31 (100 %) 31 (100 %)
Cancel
Close
Back
Next
Finish

SSIS

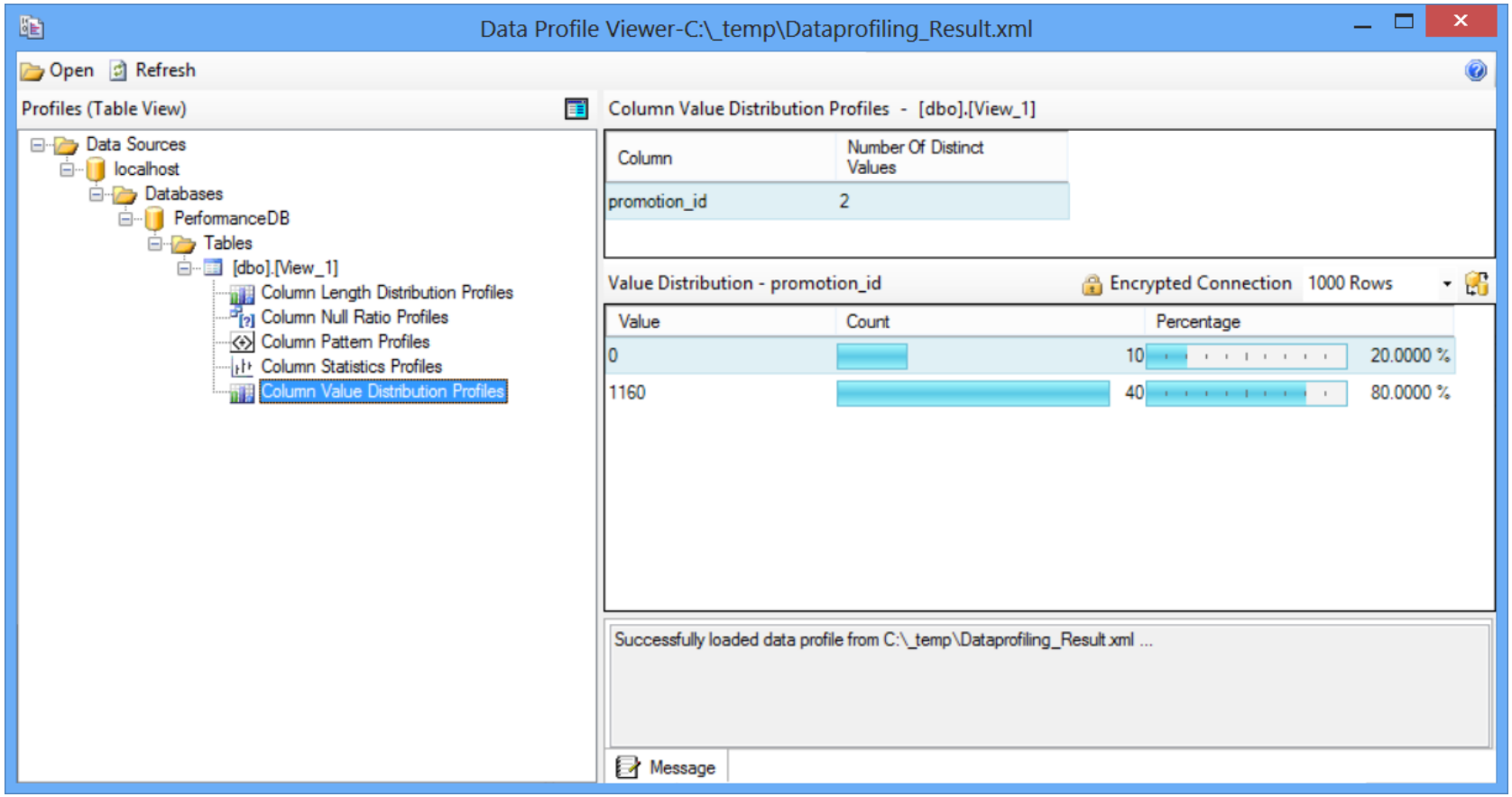
- <xml> output
- Data Profiling Viewer
- Xquery für „handmade Analyse“

SSIS

Data Profiling Task



Demo result


Data Profile Viewer-C:_temp\Dataprofiling_Result.xml
Open
Refresh
Profiles (Table View)
Data Sources
localhost
Databases
PerformanceDB
Tables
[dbo].[View_1]
Column Length Distribution Profiles
Column Null Ratio Profiles
Column Pattern Profiles
Column Statistics Profiles
Column Value Distribution Profiles
Column Value Distribution Profiles - [dbo].[View_1]
Column
Number Of Distinct Values
promotion_id
2
Value Distribution - promotion_id
Encrypted Connection
1000 Rows
Value
Count
Percentage
0
10
20.0000 %
1160
40
80.0000 %
Successfully loaded data profile from C:_temp\Dataprofiling_Result.xml ...
Message

Demo result

Dataprofiling_Result.xml

```
1   <?xml version="1.0"?>
2   <DataProfile xmlns:xsd="http://www.w3.org/2001/XMLSchema"
3                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4                xmlns="http://schemas.microsoft.com/sqlserver/2008/DataDebugger/">
5     <ProfileVersion>1.0</ProfileVersion>
6     <DataSources>...</DataSources>
14    <DataProfileInput>...</DataProfileInput>
55    <DataProfileOutput>
56      <Profiles>
57        <ColumnValueDistributionProfile IsExact="true" ProfileRequestID="ValueDistReq">
58          <DataSourceID>{33E8D61B-6415-4970-8B54-FAFD156C27DD}</DataSourceID>
59          <Table DataSource="localhost" Database="PerformanceDB" Schema="dbo" Table="View_1" RowCount="50" />
60          <Column Name="promotion_id" SqlDbType="Int" MaxLength="0" Precision="10" Scale="0" LCID="-1"
61                  CodePage="0" IsNullable="true" StringCompareOptions="32768" />
62          <NumberOfDistinctValues>2</NumberOfDistinctValues>
63          <ValueDistribution>
64            <ValueDistributionItem>
65              <Value>0</Value>
66              <Count>10</Count>
67            </ValueDistributionItem>
68            <ValueDistributionItem>
69              <Value>1160</Value>
70              <Count>40</Count>
71            </ValueDistributionItem>
72          </ValueDistribution>
73        </ColumnValueDistributionProfile>
74        <ColumnLengthDistributionProfile ProfileRequestID="LengthDistReq" IsExact="true">
```

Demo result

```
1  -- Dataprofiling
2  -- ValueDistributionProfile
3
4  Declare @file    varChar(255)
5  Set     @file    = 'C:\ <folder> \Dataprofiling_Result.xml'
6
7  Declare @charVar varChar(max)
8        , @nameSp  varChar(400)
9        , @sqlCmd  varChar(400)
10       , @xmlVar  xml
11
12 Declare @tmpTable Table (col1 varchar(max))
13
14 Set @nameSp  = ' xmlns:xsd="http://www.w3.org/2001/XMLSchema"
15                 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
16                 xmlns="http://schemas.microsoft.com/sqlserver/2008/DataDebugger/"'
17
18 Set @sqlCmd  = ' Select * From OPENROWSET ( BULK ''' + @file + ''', SINGLE_BLOB ) AS x ';
19
20 Insert Into @tmpTable
21 exec( @sqlCmd )
22
23 Set  @xmlVar = ( select Top(1) CAST( Replace( col1, @nameSp, '' )   as xml ) from @tmpTable );
24 ---- select @xmlVar
```

Demo result

```
26  Declare @i int;  -- idoc
27  Execute sp_xml_preparedocument @i OutPut
28                              , @xmlVar
29
30  SELECT *
31  FROM   OpenXML ( @i, '/DataProfile/DataProfileOutput/Profiles/ColumnValueDistributionProfile
32                                                    /ValueDistribution/ValueDistributionItem' )
33  WITH  ( ProfileRequest  nvarChar(100)  '../../@ProfileRequestID'
34        , SchemaName          nvarchar(100)  '../../Table/@Schema'
35        , TableName           nvarChar(100)  '../../Table/@Table'
36        , RowCnt              nvarChar(100)  '../../Table/@RowCount'
37        , ColumnName          nvarChar(100)  '../../Column/@Name'
38        , ColumnType          nvarChar(100)  '../../Column/@SqlDbType'
39        , DistinctValues      nvarChar(100)  '../../NumberOfDistinctValues'
40        , Value_item          nvarChar(100)  'Value'
41        , Count_item          nvarChar(100)  'Count'
42        )
43
44  Execute sp_xml_removedocument @i
```

100 %

Results | Messages

	ProfileRequest	SchemaName	TableName	RowCnt	ColumnName	ColumnType	DistinctValues	Value_item	Count_item
1	ValueDistReq	dbo	View_1	50	promotion_id	Int	2	0	10
2	ValueDistReq	dbo	View_1	50	promotion_id	Int	2	1160	40

DQS *"work items"*

- Domain Management
- Knowledge Discovery
- Matching Policy

DEMO

Demo result

```
24  -- Demo 01 TSQL
25  -->> gibt es phonetisch ähnliche Einträge
26
27  SELECT b.Buchtitel, SoundEx(b.Buchtitel) as 'SoundEx_Buchtitel'
28  FROM   dbo.Buch b
29  ORDER  by 2 DESC
30
31  SELECT Difference( 'Microsoft Windows Server 2008 Serveradministration'
32                   , 'Microsoft Windows Server 2008 Unternehmensadministration'
33                   ) as 'Difference'
34
35  --  http://support.microsoft.com/kb/100365
36
```

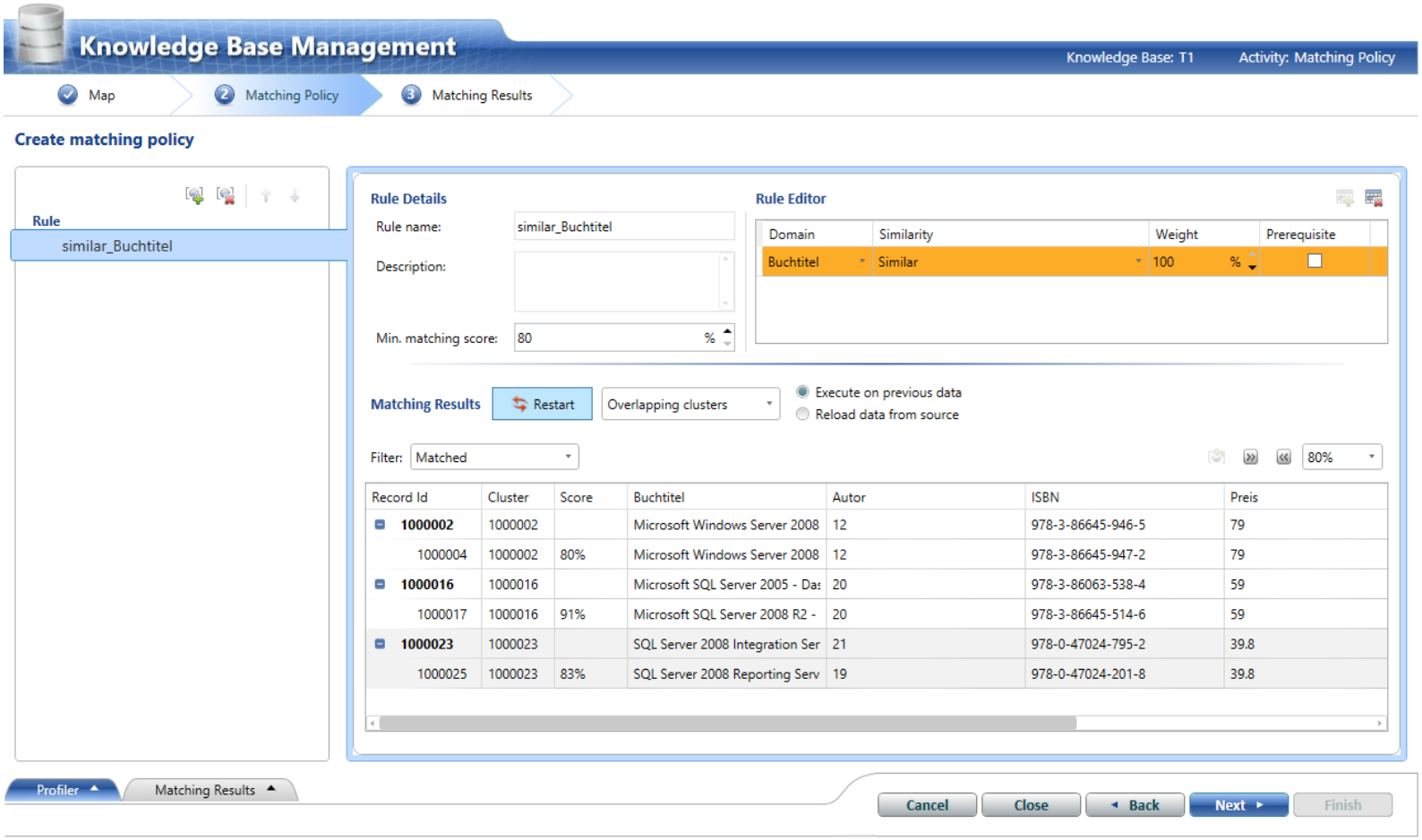
100 %

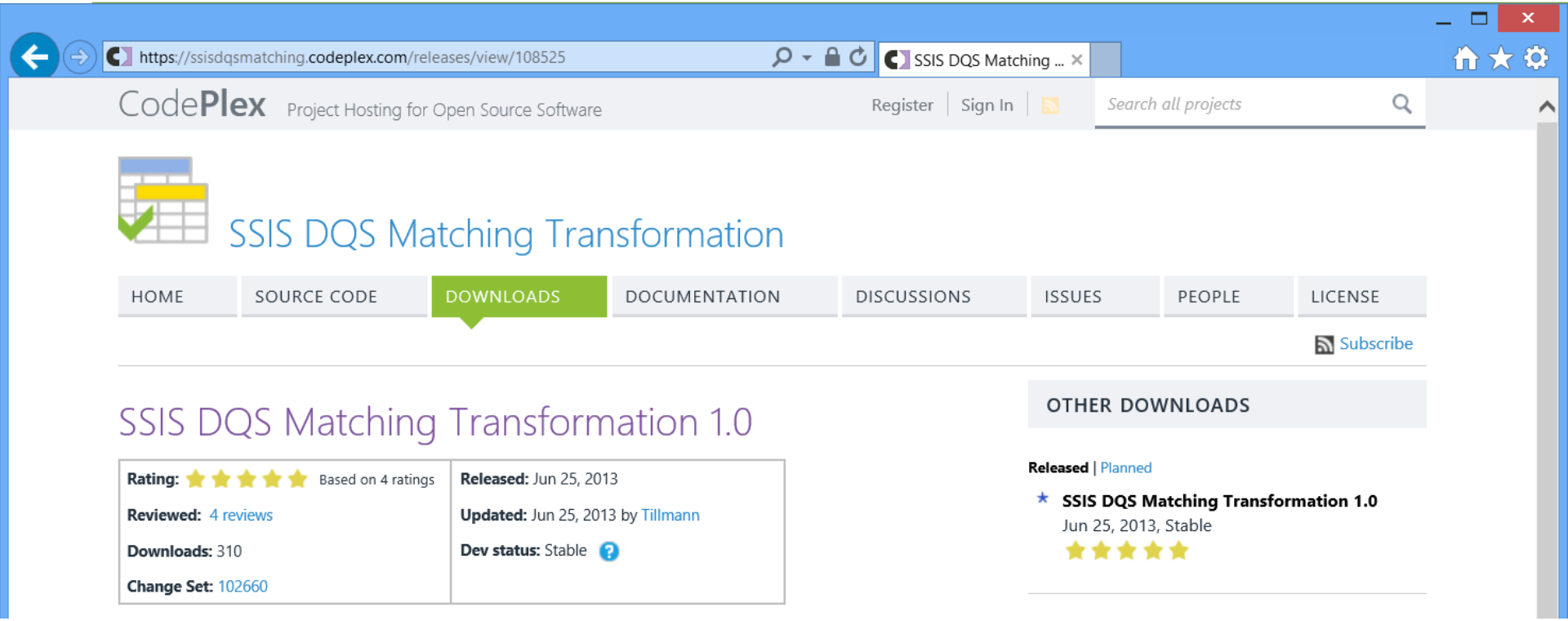
Results | Messages

	Buchtitel	SoundEx_Buchti...
14	SQL Server 2008-Programmierung mit der CLR und .NET	S240
15	Microsoft Windows Server 2008 Serveradministration	M262
16	Microsoft Windows Server 2008 Unternehmensadministration	M262
17	Microsoft Office SharePoint Server 2007 - Das Handbuch	M262
18	Microsoft Office SharePoint Server 2007-Programmierung	M262
19	Microsoft Windows Server 2008 Hyper-V - Die technische Referenz	M262
20	Microsoft SQL Server 2005 - Das Entwicklerbuch	M262

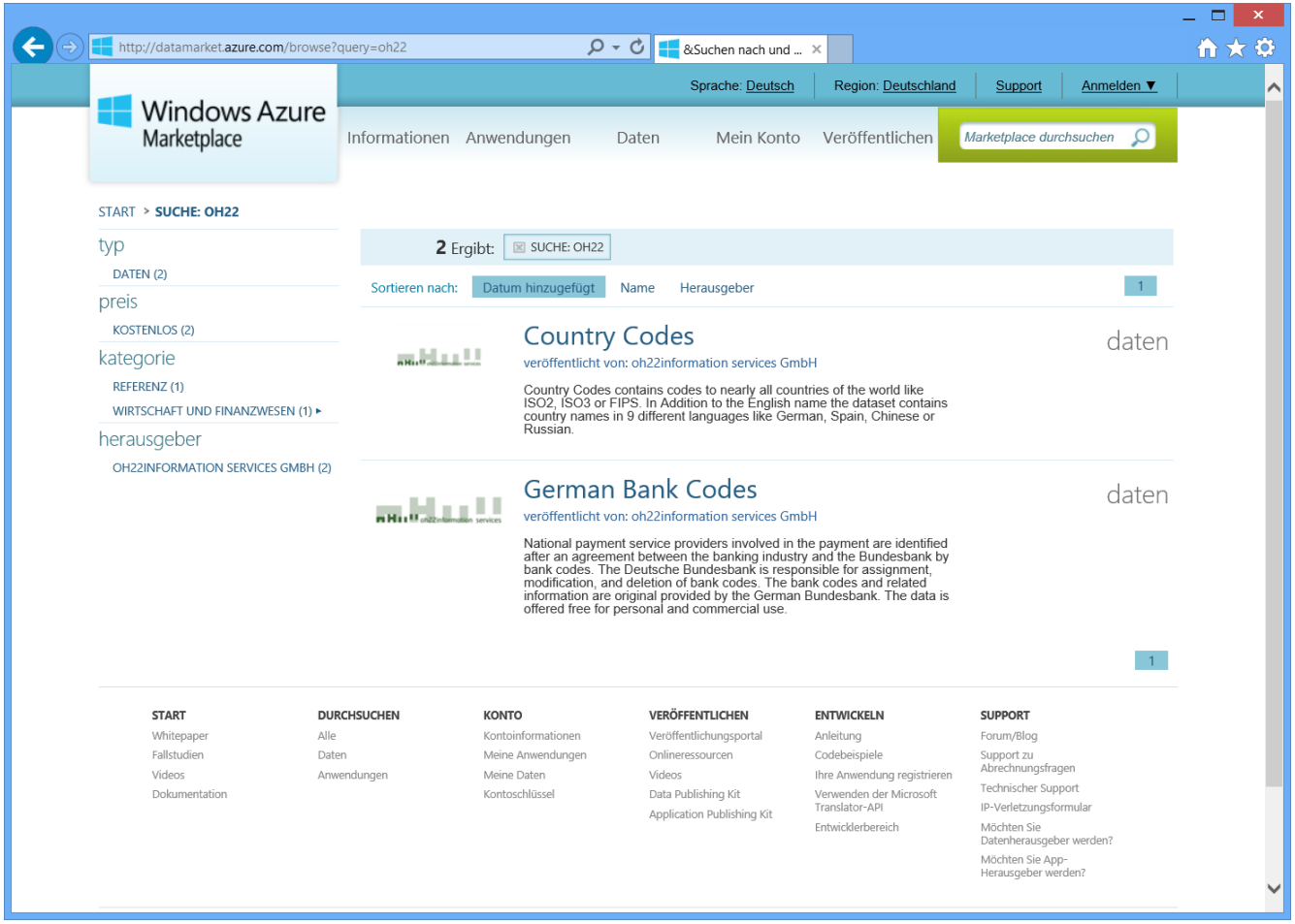
	Difference
1	4

Demo result

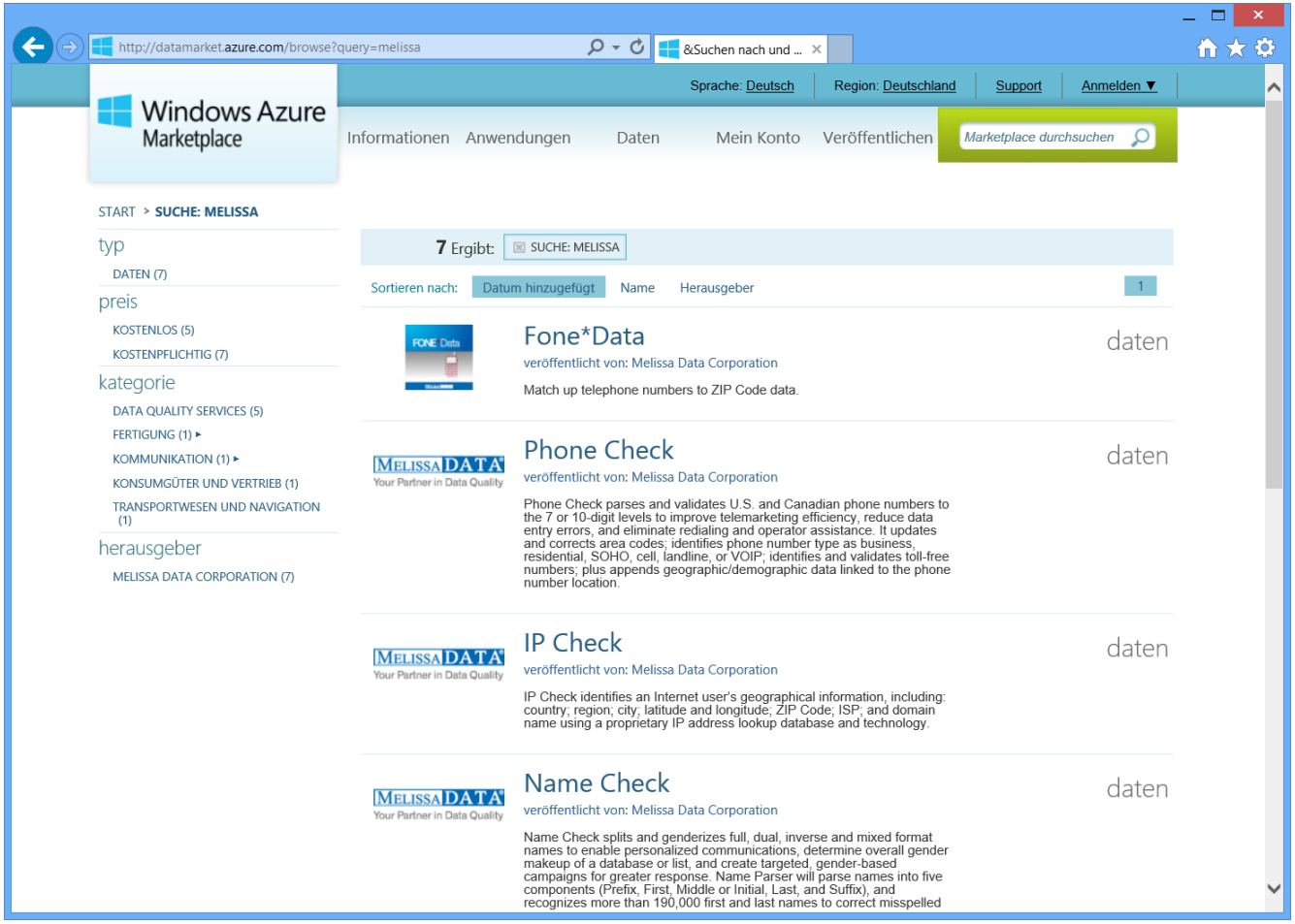

Knowledge Base Management
Knowledge Base: T1
Activity: Matching Policy
Map
2 Matching Policy
3 Matching Results
Create matching policy
Rule
similar_Buchtitel
Rule Details
Rule name: similar_Buchtitel
Description:
Min. matching score: 80 %
Rule Editor
Domain | Similarity | Weight | Prerequisite
Buchtitel | Similar | 100 % |
Matching Results
Restart
Overlapping clusters
Execute on previous data
Reload data from source
Filter: Matched
80%
Record Id | Cluster | Score | Buchtitel | Autor | ISBN | Preis
1000002 | 1000002 | | Microsoft Windows Server 2008 | 12 | 978-3-86645-946-5 | 79
1000004 | 1000002 | 80% | Microsoft Windows Server 2008 | 12 | 978-3-86645-947-2 | 79
1000016 | 1000016 | | Microsoft SQL Server 2005 - Das | 20 | 978-3-86063-538-4 | 59
1000017 | 1000016 | 91% | Microsoft SQL Server 2008 R2 - | 20 | 978-3-86645-514-6 | 59
1000023 | 1000023 | | SQL Server 2008 Integration Ser | 21 | 978-0-47024-795-2 | 39.8
1000025 | 1000023 | 83% | SQL Server 2008 Reporting Serv | 19 | 978-0-47024-201-8 | 39.8
Profiler
Matching Results
Cancel
Close
Back
Next
Finish


https://ssisdqsmatching.codeplex.com/releases/view/108525
SSIS DQS Matching ...
CodePlex
Project Hosting for Open Source Software
Register
Sign In
Search all projects
SSIS DQS Matching Transformation
HOME
SOURCE CODE
DOWNLOADS
DOCUMENTATION
DISCUSSIONS
ISSUES
PEOPLE
LICENSE
Subscribe
SSIS DQS Matching Transformation 1.0
Rating: Based on 4 ratings
Reviewed: 4 reviews
Downloads: 310
Change Set: 102660
Released: Jun 25, 2013
Updated: Jun 25, 2013 by Tillmann
Dev status: Stable
OTHER DOWNLOADS
Released | Planned
SSIS DQS Matching Transformation 1.0
Jun 25, 2013, Stable

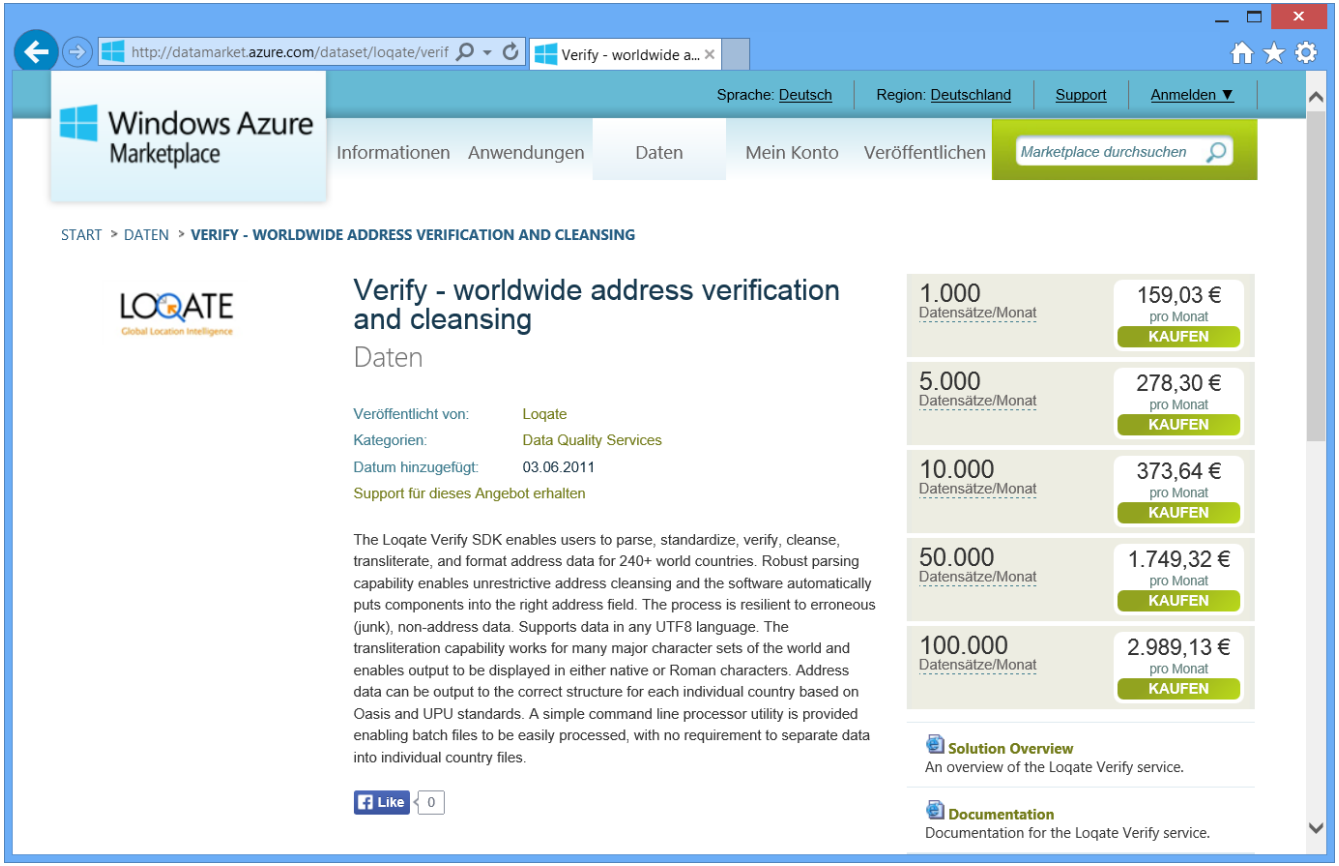
Azure Datamarket



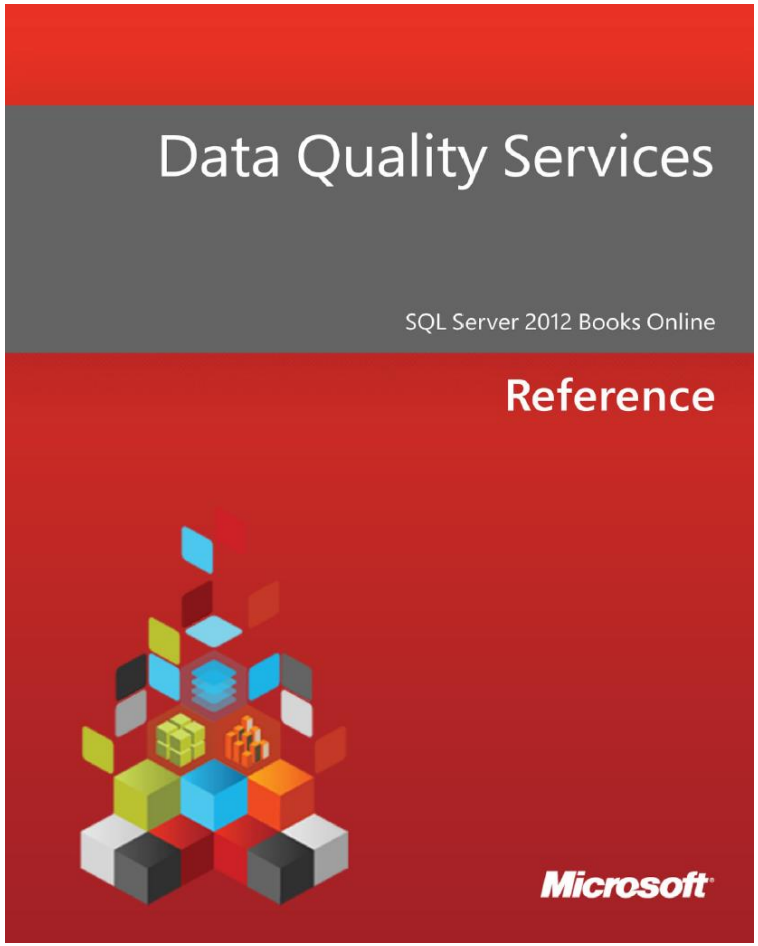
Azure Datamarket



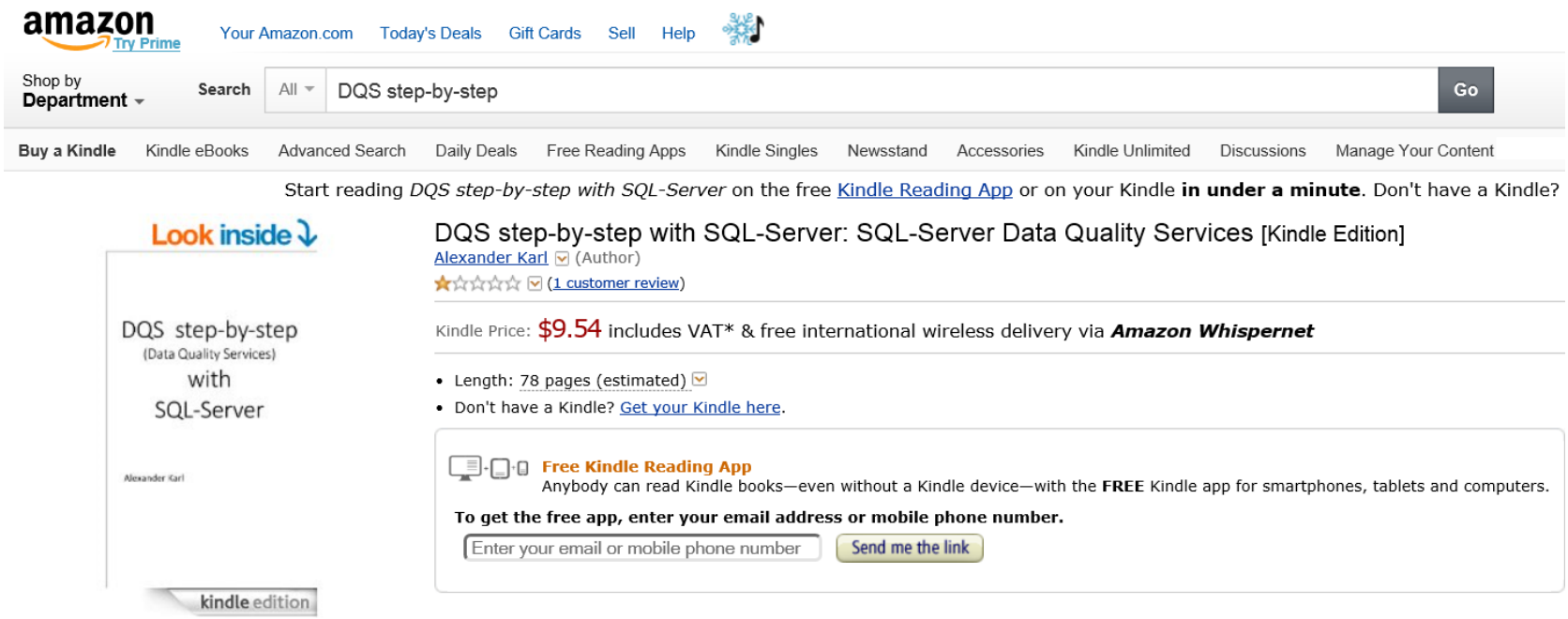
Azure Datamarket



DQS recommended literature



DQS recommended literature



???

Thank you

SQL Server Öncüleri

SQL Server konusunda bilmedikleriniz ve merak ettikleriniz için!